Jan H. van Schuppen

Control and System Theory of Discrete-Time Stochastic Systems

June 23, 2021

Springer

Preface

Researchers and students who want to learn about control and system theory of discrete-time stochastic systems at the master level or the Ph.D. level in engineering or in applied mathematics, may benefit from reading this book. Control of stochastic systems requires concepts not usually covered in courses of control of deterministic systems. The emphasis of the book is on control theory with system theory.

Motivation of Control and Filtering of Stochastic Systems

In control engineering there exist control problems for which the observed signals fluctuate irregularly. Examples of phenomena are the radio signals of early times, the signals of modern communication networks, the power produced by a wind turbine or by a photo-voltaic panel, the actual production rate of a manufacturing system, the course of ship at sea, the call requests arriving at a telephone switch, the prices of stocks on the market, etc. A mathematical model for a phenomenon with fluctuations is a stochastic control system. System theory of stochastic systems provides concepts and theorems for the formulation of stochastic control systems.

Control engineering and related research areas have problems of control and of filtering including prediction. A control problem is to determine a control law for a control system such that the closed-loop system achieves the control objectives of stability, of performance minimization, and of robustness. In filtering and prediction one needs to estimate or to predict the state and the output of a control system such as for the power demand of consumers in the next hour. The book provides concepts and theory how to solve these problems.

About the Book

The book is intended for students in engineering and in applied mathematics with an interest in control theory. In particular, to students in control, in mathematics, and in communication, information theory, mechanical, aeronautical, and computer engineering. The book can be used at the master and at the Ph.D. level depending on the background of the students and with an appropriate selection of topics.

The scope of the book is control and filtering of discrete-time stochastic systems together with stochastic realization theory. A broad spectrum of stochastic systems is treated. A probabilist ones told the author that in control theory one only formulates problems for stationary Gaussian stochastic processes. Partly this onesided attention is due to the overwhelming influence of control researchers with a background in linear systems. The scope of this book is therefore on engineering phenomena not only for stationary Gaussian processes but output processes taking values in various subsets of the real numbers. This choice distinguishes this book from many other books on control of stochastic systems.

The scope does neither include decentralized control, nor control of networked stochastic systems, nor system identification, nor adaptive control, nor stochastic systems for images called $2D$ systems, nor continuous-time stochastic systems. The space for those topics is not available in this book.

System theory treats the realization problem for obtaining from input-output trajectories or from an input-output map, a finite or finite-dimensional control system which represents those trajectories or maps and this system is then called a realization; characterizes systems which are minimal realizations in a specified sense; and relates such minimal realizations. The main concepts are controllability, observability, minimality of a realization, and input-output-map equivalence of minimal realizations.

Control theory aims at providing for control problems sufficient and necessary conditions for the existence of a control law achieving prespecified control objectives of stability, performance, robustness, etc. Sufficient and necessary conditions for existence of a control law are often controllability and observability of the control system.

The style of the book is mathematically a little formal. Concepts are stated in definitions for future reference. Theorems then state that particular concepts are sufficient and necessary conditions for the existence of a stochastic realization, for the existence of a control law, or for the existence of a filter system. Only if in a theorem the list of conditions is fully specified, is it clear on which conditions the result actually depends.

Chapter Relations

The relations of the chapters of the book are described in Table 0.1. The main lines of the book are listed in the columns 2 and 5 of the table. The chapters of the last column deals with Gaussian stochastic systems while the chapters of the second column deal with general stochastic systems. The chapters of the third and fourth column provided background information for the chapters listed on the same line; those in the third column for the chapters of the second column on the same line and those in the fourth column for the chapters in the fifth column on the same line.

Topic	Chapters General	Appendices General	Appendices Gaussian	Chapters Gaussian
Motivation	1			
Probability and stoc. processes	2, 3	17, 19, 20		
Stoc. systems	5	18	21, 22	4
Stoc. realization	7	19	22, 23, 24	6
Stoc. filtering	9	19	22	8
Stoc. control system	10		21	10 (partly)
Stoc. control problem	11			11 (partly)
Stoc. control complete observations	12, 13	17	22	12, 13 (partly)
Stoc. control partial observations	14, 15	17	22	14, 15 (partly)
Stoc. control general	16			

Table 0.1 Table relating chapters to book topics.

An introduction to control problems, to probability, and to stochastic processes is provided in the Chapters 1, 2, and 3.

The concepts of a stochastic system, of stochastic observability, and stochastic co-observability are presented in the Chapters 4 and 5.

Stochastic realization theory is treated in the Chapters 6 and 7.

Filtering problems are formulated and solved in the Chapters 8 and 9. The filter problem for a stochastic system is to determine the conditional distribution of the next state conditioned on the past outputs.

Control problems of stochastic systems are formulated and solved in the Chapters 10 – 16. The concept of a stochastic control system is defined and its properties are derived, the concept of stochastic controllability and co-controllability are formulated and characterized. Necessary and sufficient conditions for the existence of optimal stochastic control laws are established.

The Chapters 17-24 deal with mathematical and control theoretic concepts which are not part of control and system theory of stochastic systems. At the request of the publisher, the Chapters 17 – 24 are labelled according to the format
Appendix A – TITLE, etc. because they provide background information for the body of the book. The appendices are best used as an enclopedia, to be read when needed.

At the end of each chapter there are several special sections. There may be a section *Computational Issues*. Several chapters have a section *Exercises* for students. Every chapter has a section titled *Further Reading*. Those sections often start with a history of the subject of the chapter, which comments are written at the explicit request of students and of colleagues. Finally the list of *References* is stated.

To the Reader

The reader of the book is expected to have basic knowledge of linear algebra, analysis, and of elementary probability theory and stochastic processes as taught in undergraduate course programs of engineering or of mathematics. Prior knowledge of measure theory and of stochastic processes is not required but, when known, it will be useful. To a reader without this knowledge the book offers a gentle introduction to those topics in the Chapters 2 and 3. Readers without prior knowledge of linear systems as taught in control engineering may benefit from reading the basic concepts as described in Chapter 21.

The electronic version of the book has been provided facilities for navigation and for relations. The *Table of Contents* has clickable entries. Any cite command is clickable by which one jumps to the relevant entry of the reference list at the end of the chapter. At that reference, there is a clickable item with the page number which brings one back to the page from which has come.

To the Teacher

A basic course in control and system theory of discrete-time stochastic systems may start with parts of the Chapters 1, 2, and 3 which provide a motivation by control problems and a background in probability theory and in stochastic processes.

Stochastic systems are treated in the Chapters 4 and 5. The weak stochastic realization problem for stationary Gaussian processes can be presented at an elementary level, see parts of Chapter 6. The Kalman filter presented in Chapter 8 is directly related to stochastic realization.

Stochastic control problems with complete observations are treated using dynamic programming in Chapter 12. The infinite-horizon case is best treated for the average cost function with Chapter 13. Stochastic control problems with partial observations, Chapter 14 and Chapter 15, are best presented in a first course because then students will become aware of the problem and of its subtle research issues.

Concluding Remarks

A serious effort has been made to credit concepts and results to papers and to books of the literature. Because the literature is so vast, it is quite likely that due credits have been overlooked. Therefore the author apologizes in advance to all other authors whose results are not properly cited.

An effort has also been made to produce a book without errors. There is no guarantee that this ideal has been attained. Therefore the author appreciates messages with notifications of errors or with questions about the text. At the website
`http://ta.twi.tudelft.nl/mf/users/schuppen`
`/bookcontrolstocdt/bookerrata`
the reader will find a list of changes of the book, if any, in an electronic and in a printable form.

Amsterdam, *Jan H. van Schuppen*
28 September 2020

Acknowledgements

The author has learned about the topic of this book over a long period. He acknowledges the major influence by his teachers at the University of California at Berkeley, Charles A. Desoer, Pravin Varaiya, and Eugene Wong; and by his teacher H. Kwakernaak of, at that time, Delft University of Technology.

He acknowledges learning concepts and methods from the publications of: D.P. Bertsekas, René K. Boel, Charalambos D. Charalambous, Bruce Hajek, M.L.J. Hautus, Joe Hibey, R.E. Kalman, P.R. Kumar, H.J. Kushner, H. Kwakernaak, Steve Marcus, Giorgio Picci, S.E. Shreve, C. Striebel, P.M. Van Dooren, and J.C. Willems.

The author thanks several of his former Ph.D. students and former post-docs for their cooperation on theory used in the book: Peter R. de Waal, Ger M. Koole, Jana Němcová, Mihály Petreczky, Stef A. Smulders, Anton A. Stoorvogel, César Uribe, Jacqueline M. van den Hof, Cees van Putten, Kaihua Xi, and Yashar Zeinaly.

Former students of the course have provided comments on earlier drafts of this book of which are mentioned in particular Wouter Jongeneel and César Uribe. Thanked are also several colleagues for comments on the draft of the book.

The CWI secretary Lieke Schultze has typed part of the text of this book in the initial phase of the writing, for which the author is very grateful. The secretary Dorothée Engering has assisted the author which efforts are much appreciated.

My parents have supported me throughout my working life for which I am very grateful.

Several research institutes and universities have supported the author during the writing of this book and the lecture notes on which it is based, and their support is gratefully acknowledged. The names of these organizations in the historical order is: Washington University in St. Louis, Missouri, U.S.A.; the Centrum voor Wiskunde en Informatica (CWI) in Amsterdam; the Dutch Netwerk Systeem- en Regeltheorie, University of Groningen, The Netherlands; the Department of Mathematics of the Vrije Universiteit in Amsterdam; the Department of Electrical and Computer Engineering of the University of Illinois at Urbana-Champaign, Urbana, Illinois, U.S.A.; and the Department of Applied Mathematics of Delft University of Technology.

Several members of Springer Nature are thanked by the author for their professional support with the production of this book, in particular, Oliver Jackson.

Contents

Acronyms and Symbols

Acronyms

ac	average cost
ADPECO	Additive-conditional dynamic programming equation
AR	Autoregressive representation of a Gaussian system
ARMA	Autoregressive-moving-average representation of a Gaussian system
ARMAX	Autoregressive-moving-average representation of a Gaussian control system
as	almost surely
CARE	Control algebraic Riccati equation
CG	Conditionally Gaussian
CI	Conditional independence relation
CIG	Gaussian conditional independence relation
CIG_{min}	Gaussian conditional independence relation with minimality condition
dc	discounted cost
DP	Dynamic programming
DPE	Dynamic programming equation
DPECO	Conditional dynamic programming equation
EUR	Euro, currency of the Euro countries of the European Union
FARE	Filter algebraic Riccati equation
FR	Filter Riccati
FSCS	Finite stochastic control system
FSCS-CO-AC	Finite stochastic control system on an infinite-horizon with average cost function and its corresponding control law
FStocS	Finite stochastic system
GStocS	Gaussian stochastic system
GStocSP	Gaussian stochastic system parameters

$GStocSP_{s,d}$	Gaussian stochastic system parameters with stabilizable and detectable pairs of system matrices
GStocConS	Gaussian stochastic control system
GStocConSP	Gaussian stochastic control system parameters
KF	Kalman filter
LEQG	Linear-Exponential-Quadratic-Gaussian control problem and control law
LEQG-CO-FH	Linear-Exponential-Quadratic-Gaussian control problem on a finite-horizon and the corresponding control law
LQG	Linear-Quadratic-Gaussian control problem or control law
LQG-CO-AC	Linear-Quadratic-Gaussian control problem with complete observations on an infinite-horizon with average cost or the corresponding control law
LQG-CO-DC	Linear-Quadratic-Gaussian control problem with complete observations on an infinite-horizon with discounted cost or the corresponding control law
LQG-CO-FH	Linear-Quadratic-Gaussian control problem with complete observations on a finite-horizon or the corresponding control law
LQG-PO-FH	Linear-Quadratic-Gaussian control problem with partial-observations on a finite-horizon or the corresponding control law
LS	Linear system
LSP	Linear system parameters
$LSP_{c,o}$	Linear system parameters with controllable and observable tuples of system matrices
$LSP_{s,d}$	Linear system parameters with stabilizable and detectable tuples of system matrices
LSP_{min}	Linear system parameters of a minimal realization
LSRE	Equivalence relation of linear systems based on realization equivalence
$LSRE_{q_{min}}$	equivalence relation of linear systems based on minimal realization equivalence
pdf	probability distribution function
RARE	Realization algebraic Riccati equation
RARED	Dual-realization algebraic Riccati equation
SGSR	Strong Gaussian stochastic realization
SGSRP	Strong Gaussian stochastic realization parameters
$SGSRP_{min}$	Strong Gaussian stochastic realization parameters of a minimal realization
si	spectral index
StocConS	Stochastic control system
StocS	Stochastic system
tc	total cost
VOC	Verenigde Oost-Indische Compagnie (East-India Company established in 1602 in Amsterdam, The Netherlands)
WGSR	Weak Gaussian stochastic realization

WGSRP	Weak Gaussian stochastic realization parameters
$WGSRP_{min}$	Weak Gaussian stochastic realization parameters of a minimal realization

Symbols

Vectors and matrices are in roman font, not in bold font.

Logic symbols

$\exists$	there exists
$\forall$	for all
$\vee$	or
$\wedge$	and
$\Rightarrow$	implies
$\Leftarrow$	is implied by
$\Leftrightarrow$	is equivalent with

Set theory

$\{e_1, e_2, \ldots\}$	set with objects $e_1,\ e_2$ etc.
$\{x \in X \| x\ \text{ has property } P\}$	set of all x in the set X having property P
$\text{Pwrset}(X)$	set of all subsets of the set X
$\emptyset$	empty set
$\in$	member of
$\notin$	not a member of
A^c	complement of the set A
$\subseteq$	subset of
$\supseteq$	superset of
$\subset$	subset of but not equal to
$\supset$	superset of but not equal to
$\cup$	union
$\cap$	intersection
$\otimes$	σ-algebra generated by the measurable rectangles of two σ-algebras

Integers

$\mathbb{Z}$	set of the integers
$\mathbb{Z}_+ = \{1, 2, \ldots\}$	set of the positive integers
$\mathbb{Z}_n = \{1, 2, \ldots, n\}$	set of the first n integers
$\mathbb{N} = \{0, 1, 2, \ldots\}$	set of the natural numbers
$\mathbb{N}_n = \{0, 1, 2, \ldots, n\}$	set of the first $n+1$ natural numbers

Functions and maps

$f : X \to Y$	function f maps elements from the set X (domain) to the set Y (image)
$x \mapsto^f y$	function f maps x to y
$\text{Dom}(f) \subseteq X$	domain of the function f; is X if $f : X \to Y$
$\text{Range}(f) \subseteq Y$	range of the function f; equals Y if $f : X \to Y$
$\text{Im}(f) = \{f(x) \in Y \| \ \forall\, x \in X\}$	image of the function f
$\ker(f) = \{x \in X \| \ f(x) = 0\}$	kernel of the function $f : X \to Y$

$f^{-1}: Y \to X$	inverse of the function $f: X \to Y$, may be a point to set map

Real and complex analysis

$\mathbb{R}$	set of the real numbers
$\mathbb{R}_+ = [0,\infty)$	set of the positive real numbers
$\mathbb{R}_{s+} = (0,\infty)$	set of the strictly-positive real numbers
$\mathbb{R}_- = (-\infty,0]$	set of the negative real numbers
$\mathbb{R}_{s-} = (-\infty,0)$	set of the strictly-negative real numbers
$(a,b) \subset \mathbb{R}$	open interval for $a,b \in \mathbb{R}$, $a < b$
$[a,b] \subset \mathbb{R}$	closed interval
$(a,b] \subset \mathbb{R}$	interval that is left open and right closed
∞	infinity
$B(X)$	Borel σ-algebra of subsets of $X \subseteq \mathbb{R}$ or of $X \subseteq \mathbb{R}^n$ for $n \in \mathbb{Z}_+$
$\mathbb{C}$	set of the complex numbers
$\mathrm{D}_o = \{c \in \mathbb{C} \vert \vert c\vert < 1\}$	the open unit disc
$\mathbb{D}_c = \{c \in \mathbb{C} \vert \vert c\vert \leq 1\}$	the closed unit disc
$\approx$	is approximately equal to
$\times$	times
∇	gradient
∂	partial derivative or the boundary of a set
$\int$	integral sign

Vector spaces

$\mathbb{R}^n$	vector space consisting of n copies of the real numbers
$\mathbb{C}^n$	vector space consisting of n copies of the complex numbers
$\vert.\vert, \Vert.\Vert$	norm on a vector space
$(.,.)$	inner product

Matrices

$\mathbb{R}^{n \times m}$	set of matrices with entries from $\mathbb{R}$ and with n rows and m columns
$\mathbb{C}^{n \times m}$	set of matrices with entries from $\mathbb{C}$, and n rows and m columns
$\mathbb{R}^{n \times n}_{nsng}$	nonsingular square matrices
$\mathbb{R}^{n \times n}_{diag}$	diagonal matrices
$\mathbb{R}^{n \times n}_{ortg}$	orthogonal matrices satisfying $UU^T = I = U^T U$
$\mathbb{R}^{n \times n}_{spd}$	symmetric and positive-definite matrices
$Q \in \mathbb{R}^{n \times n}_{spd}$	symmetric and positive-definite matrix
$Q \succcurlyeq 0$	positive-definite matrix: if for all $w \in \mathbb{R}^n$, $w^T Q w \geq 0$
$\mathbb{R}^{n \times n}_{sspd}$	symmetric and strictly-positive-definite matrices
$Q \in \mathbb{R}^{n \times n}_{sspd}$	symmetric and strictly-positive definite matrix

$Q \succ 0$	strictly-positive-definite matrix:
$\mathrm{Diag}(a_1,\ldots,a_n)$	diagonal matrix with on the diagonal the elements $a_1,\ldots,a_n$
$\mathrm{Diag}(x) \in \mathbb{R}^{n\times n}$	diagonal matrix with on the diagonal the vector $x \in \mathbb{R}^n$
$\mathrm{Diag}(x) \in \mathbb{R}_+^{n\times n}$	positive diagonal matrix with on the diagonal the positive vector $x \in \mathbb{R}_+^n$
$\mathrm{Block}-\mathrm{diag}(A_1,\ldots,A_n)$	block diagonal matrix with on the diagonal the square matrices $A_1,\ldots,A_n$
$col(b_1,\ldots,b_m)$	the column vector with the elements $b_1,\ldots,b_m$
A^T	transpose of the matrix A
det	determinant
tr	trace of a matrix
	if for all $w \in \mathbb{R}^n\backslash\{0\}$, $w^TQw > 0$

Probability

Ω	sample space
(Ω,F)	measurable space consisting of the set Ω and the σ-algebra F
(Ω,F,P)	probability space consisting of the measurable space (Ω,F) and the probability measure P
$E[.]$	expectation operator
CI	conditional independence relation
$E[.\vert G]$	conditional expectation operator with respect to a σ-algebra G
$\ll$	absolute continuity of two probability measures
$\sim$	mutual absolute continuity of two probability measures
$\perp$	singularity of two probability measures

Function spaces

$L(X,Y)$	the set of Lebesgue measurable functions $f: X \rightarrow Y$ with $X \subseteq \mathbb{R}^n$
$L(\Omega,\mathbb{R}_+)$	$= \{x : \Omega \rightarrow \mathbb{R}_+ \vert\; x$ random variable$\}$
$L^k(\Omega,\mathbb{R}^n)$	set of random variables $x : \Omega \rightarrow \mathbb{R}^n$ for which $E\vert x\vert^k < \infty$
$C^0(X,Y)$	set of continuous functions $f : X \rightarrow Y$
$C^k(X,Y)$	set of continuous functions $f : X \rightarrow Y$ that are k times differentiable and of which the k-th differential is a continuous function
$C^\infty(X,Y)$	set of continuous functions $f : X \rightarrow Y$ that are infinitely differentiable

Chapter 1
Control Problems

Abstract Several examples of engineering control problems are described for which control of stochastic systems has been developed. Examples treated include: control of a mooring tanker, control of freeway traffic flow, and control of shock absorbers. A list of additional control problems is provided.

Key words: Control problems. Stochastic control systems.

Three examples of control engineering control problems are described for which control of stochastic systems has been used. The descriptions make use of concepts and results that are formulated only in the subsequent chapters of this book.

The examples are all formulated in terms of continuous-time stochastic systems. That was the original formulation of the papers from which those examples originate. Note that this book is about discrete-time stochastic control systems. One may reformulate the systems of the examples of this chapter as discrete-time stochastic systems. The study of continuous-time stochastic systems requires a theoretically deeper background in analysis and stochastic differential equations than expected from the readers of this book. Hence the restriction to discrete-time stochastic systems in the book. But the motivating examples are of continuous-time stochastic systems.

1.1 Control of a Mooring Tanker

Example 1.1.1. *Control of single-point moored large tanker*. This example describes the design of a controller who's objective is to reduce slow oscillations which occur when a large tanker is connected to an offshore loading terminal. The project was carried out by researchers of the University of Padova in Italy, G. Di Masi, L. Finesso, and G. Picci.

Single point mooring is used for most loading systems in deep sea. It consists of connecting the ships bow to the loading for terminal through a single elastic mooring line that is called a *hawser*. This connection allows the ship to position itself according to the prevailing weather and sea conditions. This way the ship will remain moored with a minimum of mooring force when the direction of the environmental forces change. Experience shows that offshore loading still becomes increasingly difficult when the environmental forces increase. This is due mainly to slowly varying drift forces caused by the nonlinear ship-wave hydrodynamic interaction. These forces excite the system near to resonance frequencies producing large oscillations and high mooring forces. These effects reduce the regularity and reliability of loading operations and can be dangerous for the structure of loading platforms.

To reduce the low frequency oscillations in single-point moored systems, dynamic positioning has been proposed. This technique consists of maintaining ship position and heading by means of a control system which commands the main propeller thrust and a bank of additional lateral thrusters placed in the ship's bow and astern positions.

Modeling A mathematical model for the dynamic behaviour of the ship and the monopile (deep sea loading structure) will be described. For details the reader is referred to the original paper that is mentioned in the section titled *Further reading*. Only motions in a plane are of interest. In a two-dimensional plane, positions and angles are indicated. The forces and momenta acting on the ship are:

1. Inertial forces described by Newton's laws of mechanics.
2. Ideal hydrodynamic reaction forces.
3. First order hydrodynamic forces which act at the wave frequency and do not affect the motion in a significant way.
4. Second order hydrodynamic forces. They are described by a simple noise process.
5. Viscous forces.
6. Mooring forces of the elastic restoring action in the monopile-hawserline.
7. Wind forces.
8. Manoeuvering forces due to control actuators.

Proper modelling of these forces leads to a continuous-time nonlinear stochastic system. The state variables may be regarded as slowly varying in time. A simple discrete time approximation yields a realistic discrete-time nonlinear stochastic system. Simulation of this stochastic system results in a behaviour of the model which is regarded as realistic by the researchers.

The available measurements are: (1) The bow-mooring-point distance. (2) The hawser tension. (3) The ship-hawserline relative angle. (4) The absolute heading. The measurements contain disturbances. High frequency wave forces act on the loading terminal. The high-frequency wave forces must therefore be modelled and be filtered out before synthesizing a control law.

The wave forces have been modelled empirically. Engineers have taken measurements and from these measurements constructed a spectrum of the wave forces, called the *Pierson-Maskowitz spectrum*. The spectrum is analytically described by the formula,

$$q(f) = \frac{a^2}{f^5} exp(-k/f^4),\ a,\ k \in \mathbb{R}_{s+},$$

that is appropriate for a fully developed sea. This spectrum is modified by the transfer function of the elastic mooring system. The combined spectrum is approximated by a second order shaping filter of the form

$$d\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -a_1 & -a_2 \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} dt + \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} dv(t),$$
$$dw(t) = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} dt + q_v^{1/2} dv(t),\ w(0) = 0,$$

where v is a standard Brownian motion and w denotes the noise process being shaped. The process w affects the measurements as noise. The parameters of the stochastic system described by the above equations may be estimated from data.

Control The control objective for mooring is to regulate the ship around a slowly varying equilibrium. The ship should be left free to rotate around the terminal and to position itself in such a way that the actions of the hawser tension, wind, current, and hydrodynamic drift forces are equalized. This way the controller eliminates dangerous oscillations of the ship and of the hawser tension.

Control design will be based on a simple regulator that keeps the ship in an equilibrium position. The quasi-static equilibrium position of the ship is estimated on the basis of environmental and slowly varying elastic forces. The filter is essential in estimating this equilibrium. The regulator increases damping and reduces amplitude excursions.

Synthesis of a control law is based on a procedure described in Chapter 14. According to that procedure one first determines a second stochastic control system which is measurable with respect to the observations and subsequently synthesizes an optimal control law as a function of the state of the second stochastic control system.

In the first stage of this synthesis procedure a filter system must be derived. Because the stochastic system is nonlinear of high dimension the performance of an extended Kalman filter is likely to run into trouble on numerical accuracy and computation time. Therefore the stochastic system is decoupled into the surge motion, that in the longitudinal direction of the ship, and that for the sway motion, the sideway direction. The filter that was designed according to this decomposition, had a satisfactorily performance.

Two separate control loops will be designed, one for surge motion and one for the sway-yaw motions (transversal to surge). The control design will be based on linearized models of the two motions.

The design of a control law for the surge motion is described below. The linearized model for surge motion consists of a third-order Gaussian stochastic control system. After elimination of the constraints imposed by the quasi-static steady state, a second-order such system is obtained.

$$dx_1(t) = x_2(t)dt, \tag{1.1}$$

$$dx_2(t) = -a_1x_1(t)dt - a_2x_2(t)dt + \frac{1}{m_x}u(t)dt + \frac{1}{m_x}(dv_1(t) + dv_2(t)), \tag{1.2}$$

where x_1, x_2 are state variables, u is an input, v_1 is the disturbance of the wind and v_2 is a high-frequency noise disturbance caused by the waves.

The effect of wind disturbances is taken care of by a feedforward compensation according to $u(t) = u_1(t) + s(t)$ where s is computed according to an averaging filter fed by wind speed measurements and u_1 is another input to be determined.

The control problem of regulation can now be formulated as an optimal stochastic control for the system described by the Equations (1.1,1.2) and a cost function of the form,

$$J(g) = E\left[\int_{t_0}^{t_1} \begin{pmatrix} x(s) \\ u_1(s) \end{pmatrix} Q_{cr} \begin{pmatrix} x(s) \\ u_1(s) \end{pmatrix} ds\right],\ x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix},$$

where Q_{cr} is a weighting matrix representing the cost rate which is a parameter of the control design. This optimal stochastic control problem can be solved as will be described in Chapter 15. Algorithms for the computation of the optimal control law then allow implementation. The optimal control law has the form,

$$u_1(t) = f_1\hat{x}_1(t) + f_2\hat{x}_2(t)$$

for $f_1, f_2 \in \mathbb{R}$ where $\hat{x}_1$, and $\hat{x}_2$ are produced by a filter.

Simulations show that the proposed controller performs quite well. The low-frequency oscillations almost disappear when the control action is switched on. The resulting hawser tension show little low frequency oscillation and dangerous tension peaks are eliminated. The results of this investigation may be used by control engineers for testing of the controller on a tanker. Further design is necessary to obtain an implementation of the controller.

1.2 Control of Freeway Traffic Flow

Example 1.2.1. *Control of freeway traffic flow.* The steady increase of traffic demand on freeways during the last decades of the 20th century in The Netherlands has led to a high rate of congestion. This not only meant long delays for individual drivers but also resulted in a high accident rate. As it turns out the accident rate on a congested freeway is about twice the rate on a freeway with freely flowing traffic. Increasing the capacity of the freeway by increasing the number of lanes is a solution which is not always acceptable or even preferable to alternative approaches.

An alternative approach consists of exerting a kind of control over the flow of vehicles by means of signals, with the objective of avoiding unnecessary congestions and reducing the accident rate. Several control methods have been considered in the literature and have been applied in practice. Most of the research was initially concentrated upon *on-ramp control*, regulating the rate with which vehicles enter the freeway by means of traffic lights. This allows direct control over the number of vehicles on the freeway. By maintaining the traffic volume at a certain level, below capacity, one may reduce the probability of congestion. Specific control laws have been proposed in the literature. Another approach is concerned with *rerouting* traffic through a network of freeways and/or secondary roads.

Control by means of *variable speed limits* is not considered very often in the literature either. In this example a set-up is described for control of freeway traffic by means of variable speed signs meant to avoid or at least to postpone congestion.

Modelling. In The Netherlands a freeway control and signalling system has been installed on several freeways. The system consists of measuring loops embedded in the road surface, matrix signal boards above the road mounted on gantries, and computer and communication hardware. The measuring locations are spaced approximately 500 meters apart. At each of these locations there is one pair of loops per lane to allow detection of vehicles passing and measurement of their speed. The measurements are sent to a control centre from where the matrix boards can be controlled. The matrix board gantries are spaced approximately 500 to 1000 meters apart. The boards can display advisory speed signals, lane arrows, a red cross, and a road clear signal.

The model of freeway traffic flow is based on an analogy with fluid flow. The freeway is divided into sections of approximately 500 meters long with the measuring loops at the boundaries of each section. For section i one defines the variables:

$\rho_i(t)$ the density in section i at time $t \in T,$ in veh/km.lane,

$v_i(t)$ the average speed of the vehicles in section i at time $t \in T$ in km/h.

The behaviour may then be described by a set of stochastic differential equations driven by martingales.

In this example attention is limited to one freeway section only. The nonlinear stochastic system describing the behaviour of traffic in one section allows a control analysis.

Therefore consider the model for freeway traffic flow in one section described by the equations,

$$d\rho(t) = \frac{1}{Ll}[\lambda_0 - l\rho(t)v(t)]dt + \sigma_1 dw(t), \tag{1.3}$$

$$dv(t) = -\frac{1}{t_r}[v(t) - v_e(\rho(t)]dt + \sigma_2 dz(t), \tag{1.4}$$

$$x(t) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \rho(t) \\ v(t) \end{pmatrix},\ f(x) = \begin{pmatrix} [\lambda_0 - l\rho v]/lL \\ -[v - v_e(\rho)]/t_r \end{pmatrix}, \tag{1.5}$$

where L is the length of the section, l is the number of lanes in the section, t_r is the relaxation time, λ_0 is the intensity of the traffic flow into the section, $v_e : \mathbb{R}_+ \to \mathbb{R}_+$ is the equilibrium relation between density and average speed, σ_1, σ_2 are standard deviations of the noise processes, and $w, z : \Omega \times T \to \mathbb{R}$ are standard Brownian motion processes. It will be assumed that the processes w and z are independent.

The nonlinear stochastic system described by the above equations has an interesting stability behaviour. Consider first the associated deterministic system that may be obtained by setting the standard deviations of the noise processes to zero. There are two steady states, where $f(x)=0$, denoted by $(\rho_0^s, v^e(\rho_0^s))$ and $(\rho_0^u, v^e(\rho_0^u))$. Linearization around these steady states shows that the first one is stable and the second one unstable. Determining the *domain of attraction* of the stable steady state, one discovers that there is a line, called the *separator*, that divides the state space into two connected components and that forms the boundary of the two domains of attraction.

The stability behaviour of the nonlinear stochastic system requires additional analysis. The dynamic behaviour will in general consist of fluctuations around the stable equilibrium till the trajectory leaves the domain of attraction of this equilibrium and then rapidly moves to full saturation at $(\rho, v) = (110, 0)$. One performance measure for stability is the mean time from starting in the stable equilibrium to the point where traffic comes to a stand still, at high density and zero average speed.

Control. The effect of homogenising control will now be incorporated in the model. A headway is the time that elapses between the passing of two subsequent cars in a lane at a location. The conclusion from experiments performed with and without the traffic control and signalling system are, that on the left lane the percentage of short headways decreased significantly when control is applied compared with when control is not applied. This leads to a smaller variance of these headways. The variance of density increments over small time intervals (15 sec.) then also decreases. On the right lane the mean time headway decreased, implying a larger density. On both lanes the mean speed decreased slightly.

The effect of control is modelled as follows. There are two regimes, one with and one without control. Control affects the equilibrium relation between density and average speed, and the variance of the process that affects the density equation. Let $U = \{0,1\}$ where 0 represents no speed limit and 1 represents display of a speed limit. The nonlinear stochastic control system is then described by the equations,

$$\begin{aligned} d\rho(t) &= \frac{1}{Ll}[\lambda_0 - l\rho(t)v_e(\rho(t), u(t))]dt + \sigma_1(u(t))dw(t), \\ dv(t) &= \frac{1}{t_r}[v(t) - v_e(\rho(t), u(t))]dt + \sigma_2 dz(t). \end{aligned}$$

One may then evaluate the stability properties of the nonlinear system with and without control.

The control objective is to reduce the instabilities present in traffic flow when intensity approaches capacity. As mentioned before, one cannot avoid congestion but only aim to postpone it. One may expect to reduce the probability that congestion occurs within a time interval. As in practice intensity will stay at high level for a limited time only, control may actually lead to avoidance of serious congestion.

Synthesis of a control law proceeds via optimal stochastic control. It turns out to be advantageous to control primarily on density. Thus the control is switched on if the density exceeds a set value, and switched off if it falls below this value. For details on this problem see the references in the section titled *Further Reading*.

1.3 Control of a Shock Absorber

Example 1.3.1. *Control of Shock Absorbers*. To make the drive with a car comfortable for the passengers, car manufacturers have installed shock absorbers in cars. Shock absorbers have been investigated in regard to how the degree of damping can be electronically controlled. There is thus a design problem for control of shock absorbers. The problem was treated by F. Campillo and co-workers at the research institute INRIA at the location of Sophia Antipolis in France.

Consider the following simple mechanical model of a damped mass,

$$m\ddot{y}(t) + u(t)\dot{y}(t) + ky(t) + F\text{sign}(\dot{y}(t)) = -m\ddot{e}(t),$$

y	displacement from equilibrium of the car frame with respect to a reference level
e	displacement of the bottom of the tire from an imaginary plane due to road irregularities
m	mass of the car
k, F	coefficients of the damping system
u	the damping coefficient that can be controlled

Suppose that the profile of the road can be described by a standard Brownian motion process w and a standard deviation $\sigma \in (0,\infty)$ according to

$$d\dot{e}(t) = -\sigma dw(t).$$

This model is not so realistic but it can be made more realistic at the cost of being more complex.

With the definitions $X = \mathbb{R}^2$, $x_1(t) = y(t)$, and $x_2(t) = \dot{y}(t)$ one obtains the continuous-time nonlinear stochastic system described by the stochastic differential equation,

$$\begin{aligned} dx(t) &= f(x(t),u(t))dt + Mdw(t), \\ f(x,u) &= \begin{pmatrix} f_1(x,u) \\ f_2(x,u) \end{pmatrix} = \begin{pmatrix} x_2 \\ -\frac{1}{m}[ux_2 + kx_1 + F\text{sign}(x_2)] \end{pmatrix},\ M = \begin{pmatrix} 0 \\ \sigma \end{pmatrix}. \end{aligned}$$

The control objective is to minimize a norm of the acceleration of the car mass,

$$\ddot{y}(t) + \ddot{e}(t) = -\frac{1}{m}[ux_2 + kx_1 + F\text{sign}(x_2)] = f_2(x,u).$$

Let $g : X \to \mathbb{R}$ be a control law. The closed-loop system can then be written as

$$\begin{aligned} dx^g(t) &= f(x^g(t), g(x^g(t)))dt + Mdw(t),\ x^g(0) = x_0, \\ u(t) &= g(x^g(t)). \end{aligned}$$

The cost function to be minimized is taken to be the average cost on the infinite horizon,

$$J(g) = \limsup_{t\to\infty} \frac{1}{t} E[\int_0^t f_2(x^g(s), u^g(s))^2 ds].$$

The optimal stochastic control problem is then to determine the optimal control law $g^* \in G$ and the value J^* satisfying

$$J^* = \inf_{g\in G} J(g) = J(g^*).$$

This optimal stochastic control problem could not be solved analytically as is the case with many similar problems. For this problem a numerical approximation of the optimal control law has been used to synthesize a practically useful control law. This approach is also suitable to many other problems.

The synthesis approach of numerical approximation will be described elsewhere in these notes. According to this approach one delimits a finite interval of the state space and discretizes that space. The nonlinear stochastic system is then approximated by a controlled state-finite Markov process. The cost criterion is similarly converted. Dynamic programming then leads to an equation for the value function which denotes the minimal cost as function of the initial state. The solution of this equation can be numerically approximated. A byproduct of this solution is the optimal control law.

An advantage of the synthesis approach of numerical approximation to stochastic control is that it makes possible a numerical comparison of the performance measure of the optimal control law with any other control law. Consider for example the control law

$$g_1(x) = [-kx_1 - F\text{sign}(x_2)]/x_2, \;\; k, F \in \mathbb{R}, \tag{1.6}$$

which is obtained by minimizing the instantaneous cost, see the system dynamics as described above. This control law should be modified further to account for the practical constraint that $0 \leq g_1(x) \leq \bar{u}$ where the upper bound is assumed to be specified. One may also consider a class of control laws of the form

$$g_2(x) = [a + bx_1 \text{sign}(x_2)]^+, \;\; a, b \in \mathbb{R}. \tag{1.7}$$

One may then select an element in this class for which the average cost is minimal.

The sub-optimal control law may then be acceptable to engineers because the trade-off between the minimization of the performance criterion and the minimization of the complexity of the control law is considered satisfactory.

1.4 Further Reading

Examples. The example of control of a mooring tanker was developed by G. Di Masi, L. Finesso, and G. Picci, [12], where the full story is described. A model for a stochastic process of a fully developed sea is provided in [14]. The example of control of freeway traffic flow is based on the research of S.A. Smulders, and the detailed model is described in [18, 20]. The example of control of shock absorbers is due to F. Campillo and colleagues, [3], where the full story is described.

Estimation of probability distributions. Algorithms and theory for estimation of probability distribution and density functions from data are described in the books [8, 16].

System identification. The research area of system identification treats the problem of how to obtain from observations a realistic model in the form of a dynamic system. The problem of system identification involves modeling, selection of a class of dynamic systems, input design, a check on identifiability, the approximation problem (including parameter estimation), and the evaluation problem. For system identification of a time-invariant Gaussian control system the subspace identification algorithm is most effective. For other sets of stochastic control systems, there are few satisfactory results. Books on system identification include [5, 11, 23, 28].

Approximations of stochastic systems. Transformation of a continuous-time stochastic system to a discrete-time stochastic system. See the book of K.J. Aström and B. Wittenmark, [2], and that of H. Kushner and P. Dupuis, [10].

Books with models of time series, including stochastic systems, occuring in engineering and physics include [6, 15, 24, 26].

A List of Books and Papers on other Stochastic Systems. Noise in communication systems [9]. Counting and jump systems [22]. Industrial process control [1]. Hydropower production [7]. Maintenance. Communication networks [27]. Freeway traffic flow [17, 20, 19, 21]. Physics [25]. Stochastic mechanics [4, 13].

There follows a list of research areas in which problems of stochastic control or of modeling and of realization of stochastic systems arise: (1) control engineering, system identification, machine learning and reinforcement learning, information theory, communication theory; (2) electrical engineering, electric power systems, control of motorway traffic, control of urban traffic, civil engineering, hydrology, weather prediction; (3) compartmental systems and physiological models, biochemical reaction systems, chemical engineering.

Current and future problems for which modeling, filtering, and control of stochastic systems is useful are likely to arise in: the effect of wind turbines and wind parks on power systems. The modeling and control of electric devices in large buildings and in homes. The use of batteries in mobile devices facing uncertain power demand and control of power charging.

The processing of messages and packets in communication networks where the finite capacity of buffers and of lines is a serious restriction.

Modeling and control of the operations of economic organizations like banks and service companies.

References

1. K.J. Aström. Computer control of a paper machine - An application of linear stochastic control theory. *IBM J. Res. & Developm.*, 11:389–405, 1967. 9, 78, 120, 522, 575, 596

2. K.J. Aström and B. Wittenmark. *Computer controlled systems*. Prentice-Hall Inc., Englewood Cliffs, N.J., 1984. 9
3. S. Bellizzi, R. Bouc, F. Campillo, and E. Pardoux. Contrôle optimal semi-actif de suspension de véhicule. In A. Bensoussan and J.L. Lions, editors, *Analysis and optimization of systems*, volume 111 of *Lecture Notes in Control and Information Sciences*, pages 689–699. Springer-Verlag, Berlin, 1988. 9, 377, 468
4. J.-M. Bismut. *Mécanique aléatoire*, volume 866 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1981. 9
5. S. Bittanti. *Model identification and data analysis*. John Wiley & Sons, Inc., New York, 2019. 9
6. W.S. Burdic. *Underwater acoustic system analysis*. Prentice-Hall, Englewood Cliffs, NJ, 1984. 9
7. F. Delebecque and J.-P. Quadrat. Contribution of stochastic control singular perturbation averaging and team theories to an example of large-scale systems: Management of hydropower production. *IEEE Trans. Automatic Control*, 23:209–221, 1978. 9
8. L. Devroye. *A course in density estimation*. Birkhäuser Verlag, Basel, 1987. 9, 742
9. W.B. Davenport Jr. and W.L. Root. *An introduction to the theory of random signals and noise*. McGraw-Hill Book Co., New York, 1958. 9, 72, 310
10. H.J. Kushner and P.G. Dupuis. *Numerical methods for stochastic control problems in continuous time (2nd Ed.)*. Number 24 in Applications of Mathematics. Springer, New York, 2001. 9, 376, 377, 466
11. L. Ljung. *System identification: Theory for the user*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1987. 9
12. G.B. Di Masi, L. Finesso, and G. Picci. Design of LQG controller for single point moored large tankers. *Automatica J.-IFAC*, 22:155–169, 1986. 8, 78, 575
13. E. Nelson. *Dynamical theories of Brownian motion*. Princeton University Press, Princeton, 1967. 9
14. W.J. Pierson Jr. and L. Moskowitz. A proposed spectral form for fully developed wind seas based on the similarity theory of S.A. Kitaigorodskii. *J. Geophysical Research*, 69:5181–5190, 1964. 8
15. E.A. Robinson and S. Treitel. *Geophysical signal analysis*. Prentice-Hall, Englewood Cliffs, NJ, 1980. 9
16. B.W. Silverman. *Density estimation for statistics and data analysis*. Chapman and Hall, London, 1986. 9
17. S.A. Smulders. Modelling and filtering of freeway traffic flow. In N.H. Gartner and N.H.M. Wilson, editors, *Proceedings of the 10th International Symposium on Transportation and Traffic Theory*, pages 139–158, New York, 1987. Elsevier. 9
18. S.A. Smulders. Control of freeway traffic flow. Report OS-R8817, Centrum voor Wiskunde en Informatica, Amsterdam, 1988. 9, 78, 169
19. S.A. Smulders. *Application of stochastic control concepts to a freeway traffic control problem*, pages 295–301. X, XX, 1989. 9
20. S.A. Smulders. *Control of freeway traffic flow*. PhD thesis, University of Twente, Enschede, 1989. 9, 169
21. S.A. Smulders. Control of freeway traffic flow by variable speed signs. *Transpn. Res.-B*, 24B:111–132, 1990. 9
22. D.L. Snyder. Random point processes. *John Wiley & Sons*, 1975. 9
23. T. Söderström and P. Stoica. *System identification*. Prentice Hall, New York, 1989. 9
24. M.D. Srinath and P.K. Rajasekaran. *An introduction to statistical signal processing with applications*. Wiley, 1979. 9
25. N.G. van Kampen. *Stochastic processes in physics and chemistry*. North-Holland, Amsterdam, 1983. 9
26. H.L. van Trees. *Detection, estimation, and modulation theory*. Wiley, 1968. 9, 72
27. J. Walrand and P.P. Varaiya. *High-performance communication networks*. Kaufmann, San Francisco, 1996. 9, 468
28. Peter C. Young. *Recursive identification and time-series analysis*. Springer-Verlag, Berlin, 2nd ed. edition, 2011. 9

Chapter 2
Probability

Abstract The study of control of stochastic systems requires knowledge of probability and of stochastic processes. Probability is summarized in this chapter in a way which is sufficient for studying the control and system theory of the subsequent chapters. Additional concepts and results of probability theory are provided in an appendix labelled Chapter 20. The main concepts of probability needed for the reading of this book are probability distribution functions, probability measures, random variables, Gaussian random variables, conditional expectation, and conditional independence.

Key words: Probability. Random variable. Conditional expectation.

The reader finds in this chapter an introduction to measuretheoretic probability theory. Major concepts to learn from this chapter are probability distributions, probability measures, random variables, the conditional expectation of a random variable conditioned on a σ-algebra, and the conditional independence relation.

For many students the theory is new and it takes much time to became familiar with it. The chapter is a summary of part of a specialized course on the subject. But the concepts of this chapter are the basic building blocks for the subsequent chapters. An approach students have used is to go directly to the exercises and then read relevant parts of the chapter for your understanding.

2.1 Probability Distribution Functions

The subject of probability started with the modeling of phenomena exhibiting uncertainty in data. The reader may think of such phenomena as: the outcomes of tosses of a coin; the length of humans, separately for men and for women; and the number of messages arriving at an email server. Based on the data observed, researchers formulated a model in the form of a probability distribution function. Subsequently several special probability distribution functions were formulated as analytic functions which realistically approximate the observed data distributions. The so obtained analytic probability distributions, called *special probability distribution functions*, were the starting point of probability theory.

The reader finds in this section the concept of a probability distribution function and definitions of several such functions.

Definition 2.1.1. A *probability distribution function*, abbreviated as pdf, on the real numbers $\mathbb{R}$ is a function $f : \mathbb{R} \to \mathbb{R}_+$ that:

1. is *increasing*: $u < v$ implies that $f(u) \leq f(v)$;
2. has limits at $-\infty$ and $+\infty$: $\lim_{u \to -\infty} f(u) = 0$, $\lim_{u \to \infty} f(u) = 1$;
3. is *right continuous*: for any $u \in \mathbb{R}$, $f(u_-) \leq f(u) = f(u_+)$, where $f(u_-) = \lim_{v \uparrow u} f(v)$ and $f(u_+) = \lim_{v \downarrow u} f(v)$.

Because a pdf is monotonically increasing and bounded from below and above, the indicated limits always exist.

The definition of a pdf on $\mathbb{R}^n$ involves more conditions than that of a pdf on $\mathbb{R}$, see [2, Section 1.4] for a lengthy definition. Because a probability distribution function on the real numbers is an increasing function, it is a Borel measurable function.

Definition 2.1.2. *Subsets of probability distribution functions.*

(a) A *discrete pdf* is a pdf on the real numbers of the form,

$$f(u) = \sum_{n \in \mathbb{Z}} p(n) I_{[u_n, \infty)}(u),$$

where $p : \mathbb{Z} \to [0, \infty)$ satisfying $\sum_{n \in \mathbb{Z}} p(n) = 1$ is called the associated *probability frequency function* and where $\{u_n \in \mathbb{R}, n \in \mathbb{Z}\}$ is a strictly-increasing finite or countable sequence of elements of $\mathbb{R}$.

(b) An *absolutely continuous pdf* is a pdf on the real numbers of the form,

$$f(u) = \int_{-\infty}^{u} p(v)\, dv, \text{ where } p : \mathbb{R} \to \mathbb{R}_+ \text{ satisfies } 1 = \int_{-\infty}^{\infty} p(v) dv,$$

and the function p is called the associated *probability density function.* The expression *absolutely continuous pdf* can be understood in terms of absolute continuity of the measure defined for the pdf with respect to Lebesgue measure, see Section 19.9.

(c) A *singular pdf* is a pdf on the real numbers which is a continuous function while its derivative exists and is zero almost everywhere.

The set of discontinuities of an arbitrary pdf has been proven to be countable. It can be shown that an arbitrary pdf has a unique decomposition as a convex combination of a discrete, an absolute continuous, and a singular continuous pdf. For an example of a pdf that is singular see [13, p. 12].

Below follow examples of well known frequency functions and density functions. The names of these functions start with a capital letter due to the particular word used. Several of these names are the family names of mathematicians. In the examples of this book it will be described for which situations these distributions are suitable mathematical models.

Definition 2.1.3. The *Bernoulli* pdf on the set $\{0,1\} \subset \mathbb{R}$ with parameter $q \in [0,1] \subset \mathbb{R}$ is a discrete pdf with the frequency function,

$$p(1) = q,\ \ p(0) = 1 - q,\ p : \{0,\ 1\} \to \mathbb{R}_+,$$
$$f(w) = \begin{cases} 0, & w < 0, \\ 1-q, & 0 \le w < 1, \\ 1, & 1 \le w. \end{cases}$$

It is named after Jacob Bernoulli (1655–1705), to be distinguished from other mathematicians of the Bernoulli family.

Definition 2.1.4. The *Binomial* pdf on $\mathbb{N}_n = \{0,1,\ldots,n\} \subset \mathbb{Z}$ with parameters $(n,\ q) \in \mathbb{Z}_+ \times [0,1]$ is a discrete pdf with the frequency function,

$$p(k) = \binom{n}{k} q^k (1-q)^{n-k},\ \text{where}\ \binom{n}{k} = \frac{n!}{k!(n-k)!},\ p : \mathbb{N}_n \to \mathbb{R}_+.$$

Definition 2.1.5. The *Poisson* pdf on $\mathbb{N} = \{0,1,...\}$ with parameter $\lambda \in \mathbb{R}_+ = (0,\infty)$ is a discrete pdf with the Poisson frequency function,

$$p(k) = \lambda^k exp(-\lambda)/k!,\ \ p : \mathbb{N} \to \mathbb{R}_+;$$
$$\sum_{k=0}^{\infty} p(k) = \sum \lambda^k/k!\ \exp(-\lambda) = \exp(\lambda - \lambda) = \exp(0) = 1.$$

It is named after the mathematician Siméon Denis Poisson (1781–1840) who was born in France.

Definition 2.1.6. The *uniform* pdf on the interval $[0,1] \subset \mathbb{R}$ is an absolutely continuous pdf with the density function,

$$p(v) = \begin{cases} 1, \ v \in [0,1), \\ 0, \ elsewhere. \end{cases} \ \ p : \mathbb{R} \to \mathbb{R}_+,$$
$$f(w) = \begin{cases} 0, \ u < 0, \\ w, \ 0 \le w < 1, \\ 1, \ 1 \le w. \end{cases}$$

Definition 2.1.7. The *Beta* pdf on $(0,1) \subset \mathbb{R}_+$ with parameters
$(\beta_1,\beta_2) \in \mathbb{R}^2_{s+} = (0,\infty)^2$ is an absolutely continuous pdf with the density function,

$$p(v) = v^{\beta_1 - 1}(1-v)^{\beta_2 - 1}/B(\beta_1,\beta_2),\ \ p : [0,1] \to \mathbb{R}_+,$$
$$B(\beta_1,\beta_2) = \int_0^1 v^{\beta_1 - 1}(1-v)^{\beta_2 - 1} dv = \frac{\Gamma(\beta_1)\Gamma(\beta_2)}{\Gamma(\beta_1+\beta_2)}. \tag{2.1}$$
$$B(k_1,k_2) = \frac{(k_1-1)!(k_2-1)!}{(k_1+k_2-1)!},\ \ \forall\, k_1,k_2 \in \mathbb{Z}_+, \tag{2.2}$$

where B is called the *Beta function* with the parameters (β_1,β_2). In the above statements, Γ is called the Gamma function which is defined in Def. 2.1.8.

In this book the term Beta pdf is written with a capital of Beta because this is a particular term. Similarly for the Gamma pdf etc.

If in the Beta pdf the parameter values are $(\beta_1, \beta_2) = (1,1)$ then the Beta pdf is the uniform distribution on the interval $[0,1]$. If $(\beta_1, \beta_2) \in (1,\infty)^2$ then the Beta density function is cap-shaped while if $(\beta_1, \beta_2) \in (0,1)^2$ then the Beta density function is cup-shaped.

The beta distribution arises from a quotient formula. Below use is made of the Chi-Square pdf for which see Def. 2.1.9. Consider two random variables,

$x_1^2 : \Omega \rightarrow \mathbb{R}_+,$ with the Chi-Square probability distribution
with v_1-degrees of freedom,
$x_2^2 : \Omega \rightarrow \mathbb{R}_+,$ with the Chi-Square probability distribution
with v_2-degrees of freedom, x_1^2, x_2^2 are independent random variables,

$$y^2 = \frac{x_1^2}{x_1^2 + x_2^2} : \Omega \rightarrow (0,1).$$

Then y^2 has a beta probability distribution function with parameter values $(\beta_1, \beta_2) = (v_1/2, v_2/2)$.

Definition 2.1.8. The *Gamma* pdf on $\mathbb{R}_+$ with parameters $(\gamma_1, \gamma_2) \in \mathbb{R}_{s+}^2 = (0,\infty)^2$, is an absolutely continuous pdf with the probability density function,

$$p(v) = v^{\gamma_1 - 1} \exp(-v/\gamma_2) \gamma_2^{-\gamma_1} / \Gamma(\gamma_1), \;\; p : \mathbb{R} \rightarrow \mathbb{R}_+, \tag{2.3}$$

$$\Gamma(\gamma_1) = \int_0^\infty v^{\gamma_1 - 1} \exp(-v) dv, \;\; \Gamma : \mathbb{R}_+ \rightarrow \mathbb{R}_+, \tag{2.4}$$

is called the *Gamma function*. It is proven, using integration by parts, that $\Gamma(\gamma_1 + 1) = \gamma_1 \Gamma(\gamma_1)$. Consequently, $n \in \mathbb{Z}_+$ implies that $\Gamma(n+1) = n!$ It is recommended to use in examples a value for the parameter γ_1 which is a positive integer.

The case of $\gamma_1 = 1$ with $p(v) = \exp(-v/\gamma_2)/\gamma_2$ is called the *exponential* pdf with parameter $\gamma_2 \in \mathbb{R}_{s+} = (0,\infty)$.

The reader should note that in other references the parameters of the Gamma probability density function can be defined differently, for example by (γ_1, γ_3) with $\gamma_3 = 1/\gamma_2$.

The calculations for the Gamma density function are simplified by the following formula,

$$\int_0^\infty v^{\gamma_1 - 1} \exp(-v/\gamma_2) dv = \int_0^\infty w^{\gamma_1 - 1} \exp(-w) dw \; \gamma_2^{\gamma_1} = \gamma_2^{\gamma_1} \Gamma(\gamma_1), \tag{2.5}$$

by the substitution $w = v/\gamma_2$. That the function p is indeed a density function follows directly from the above formula.

Definition 2.1.9. The *Chi-Square* pdf on $\mathbb{R}_+$ with parameter $v \in \mathbb{R}_{s+}$ called the *degrees of freedom*, denoted by χ_v^2, is an absolution continuous pdf with the density function,

$$p_{\chi_v^2}(w) = w^{v/2} \exp(-w/2) / (2^{v/2} \Gamma(v/2)), \; p : \mathbb{R}_+ \rightarrow \mathbb{R}_+.$$

The parameter in the early literature took only integer values, $v \in \mathbb{Z}_+$ and was later generalized to $v \in \mathbb{R}_{s+}$.

The Chi-square density function with ν degrees of freedom equals the Gamma probability density function with the parameters $(\nu/2,2)$.

If one has $n \in \mathbb{Z}_+$ independent Gaussian distributed random variables $\{x_i : \Omega \to \mathbb{R},\ i \in \mathbb{Z}_n\}$ then the sum $\sum_{i=1}^n x_i^2$ is a random variable which has a Chi-Square probability density function with n degrees of freedom.

Definition 2.1.10. The *Gaussian* pdf on the real numbers $\mathbb{R}$ with parameters $(m,q) \in (\mathbb{R} \times \mathbb{R}_{s+})$, is an absolutely continuous pdf with the density function $p : \mathbb{R} \to \mathbb{R}_+$,

$$p(v) = \exp(-(v-m)^2/2q)/\sqrt{2\pi q}.$$

The *Gaussian* pdf on the tuples of the real numbers $\mathbb{R}^n$ with parameters (m,Q) is defined by the density function,

$$p : \mathbb{R}^n \to \mathbb{R}_+,\ (m,Q) \in \mathbb{R}^n \times \mathbb{R}^n_{spds},$$
$$p(v_1,\ldots,v_n) = exp(-1/2(v-m)^T Q^{-1}(v-m)) \times \ [(2\pi)^n \det(Q)]^{-1/2}.$$

It is named after the German-born mathematician Carl Friedrich Gauss (1777–1855).

There exists a probability distribution function of which the expectation is not finite! It was constructed by Baron Louis-Augustin Cauchy (1789 - 1857) born in France, specifically for the purpose to disprove the Central Limit Theorem.

Definition 2.1.11. The *Cauchy probability density function* with parameters $a \in \mathbb{R}$ and $b \in (0,\infty)$ is defined as the function

$$p : \mathbb{R} \to \mathbb{R}_+,\ p(v) = \frac{1}{b\pi}\frac{1}{1+\frac{(v-a)^2}{b^2}}.$$

Proposition 2.1.12. *Consider the Cauchy probability density function.*

(a)This function is indeed a probability density function.
(b)If $x : \Omega \to \mathbb{R}$ is a random variable with the Cauchy probability density function with parameters $(a,b) \in \mathbb{R} \times (0,\infty)$ then $E|x| = \infty$.

Proof. (a) Note that,

$$r = \frac{v-a}{b},\ \int_a^\infty \frac{1}{b\pi}\frac{1}{1+\frac{(v-a)^2}{b^2}}dv = \frac{1}{\pi}\int_0^\infty \frac{1}{1+r^2}dr = \frac{1}{\pi}ctg(r)|_0^\infty = \frac{1}{2},$$
$$\int_{-\infty}^a \frac{1}{b\pi}\frac{1}{1+\frac{(v-a)^2}{b^2}}dv = \frac{1}{2}.$$

(b)

$$r = \frac{v-a}{b},\ \int_a^\infty (v-a)p(v)dv = \int_a^\infty \frac{1}{b\pi}(v-a)\frac{1}{1+\frac{(v-a)^2}{b^2}}dv$$
$$= \frac{b}{\pi}\int_0^\infty r\frac{1}{1+r^2}dr \ = \frac{b}{2\pi}\ln(1+r^2)|_{r=0}^{r=\infty} = \infty,$$
$$E|x| = E[(x-a)^+] + E[(x-a)^-] = \infty.$$

□

The reader finds in Section 19.10 the set of expontial probability distribution functions which includes several of the above special pdfs.

2.2 Motivation of the Concept of a Probability Measure

There follows a motivation of the concept of a probability measure.

An early attempt to formulate an axiomatic definition of a probability measure on a set Ω proceeded as follows. Consider the power set $\text{Pwrset}(\Omega) = 2^{\Omega}$ of all subsets of the set Ω. A probability measure may then be defined as a function $f : \text{Pwrset}(\Omega) \to \mathbb{R}_+$ satisfying conditions. After a time it was discovered that a function satisfying all the required conditions did not exist. The initial definition of a probability measure on Ω was therefore adjusted, it was to be defined not for all subsets of Ω but only for a smaller set of subsets, while also the required conditions were relaxed. The subset of the power set for which a probability measure may be shown to exist, is a σ-algebra. Therefore it is useful to introduce the concept of a σ-algebra.

Below concepts of set theory are stated, including the definition of a σ-algebra. In a subsequent section a definition of a probability measure is presented.

2.3 Sets and Sigma-Algebras

Objects are things, numbers etc. A *set* is a collection of objects.

Example 2.3.1. There follow several examples of sets which are frequently used in this book.

The *binary set* $\{0,\ 1\} \subset \mathbb{N}$. The *finite set of the first $n \in \mathbb{Z}_+$ integers*, denoted by $\mathbb{Z}_n = \{1,\ 2,\ \ldots,\ n\}$.

The *strictly positive integers* denoted by $\mathbb{Z}_+ = \{1,\ 2,\ \ldots\} \subset \mathbb{R}$. The set of the *natural numbers* denoted by $\mathbb{N} = \{0,\ 1,\ \ldots\} \subset \mathbb{R}$. The set of the *integers* denoted by $\mathbb{Z} = \{\ldots,\ -1,\ 0,\ +1,\ldots\} \subset \mathbb{R}$.

The set of the *real numbers* denoted by $\mathbb{R}$. The set of the *positive real numbers* denoted by $\mathbb{R}_+ = [0,\infty)$. The set of the *strictly-positive real numbers* denoted by $\mathbb{R}_{s+} = (0,\infty) \subset \mathbb{R}_+$.

The set of *n-tuples of the real numbers* for a strictly positive integer $n \in \mathbb{Z}_+$ denoted by $\mathbb{R}^n$. The set of *n-tuples of the positive real numbers* for a strictly positive integer $n \in \mathbb{Z}_+$ denoted by $\mathbb{R}^n_+$.

Let Ω be a set. The notation $\omega \in \Omega_1$ denotes that the object ω is an element of the subset $\Omega_1 \subseteq \Omega$. Its negation is $\omega \notin \Omega_1$ which denotes that ω is not a member of the set $\Omega_1 \subseteq \Omega$ but is an element of the set Ω. The set that has no elements is denoted by $\emptyset$ and it is called the *empty set*.

Operations on Sets

Given two sets A,B, one says that A is *included* in B if every element of A is also an element of B. This is denoted by $A \subseteq B$, and one may also say that the set A is a *subset* of the set B. If $A \supseteq B$ then one calls A a *superset* of B. One says that A and B are *disjoint* if they have no element in common.

A *family of sets* is a set whose elements are also sets. An example is the family of all subsets of a set. For the set Ω, the family of all its subsets is called the *power set* and it is denoted by,

$$\text{Pwrset}(\Omega) = \{A \subseteq \Omega | \; \forall \, A \subseteq \Omega\}.$$

The family may be indexed, in which case the notation $\{A_i \subseteq \Omega, i \in I\}$ may be used. Call then I the *index set* of the family. The family is said to be *finite* if there exists a $n \in \mathbb{Z}_+$ such that $I = \mathbb{Z}_n = \{1,2,...,n\}$ or is bijectively related to $\mathbb{Z}_n$. The family is said to be *countable* if the index set I equals or is bijectively related to the set of the strictly positive integers, $\mathbb{Z}$.

The *Cartesian product* or just the *product* $X \times Y$ of two sets X,Y is defined by

$$X \times Y \; := \; \{(x,y) \, | \, x \in X, y \in Y\}.$$

This definition generalizes to finite products. For $n \in \mathbb{Z}_+$ denote the multiple product set by $X^n \; = \; X \times X \times ... \times X$, which is called the *n-fold Cartesian product* of X.

Let Ω be a set, and $A,B \subseteq \Omega$ and $\{A_i \subseteq \Omega, i \in I\}$. Define the following operations on these sets:

- *Complementation* which results in the *complement* of a set, $A^c = \{\omega \in \Omega | \; \omega \notin A\}$.
- *Binary union* which results in the *union* of the two corresponding sets, $A \cup B = \{\omega \in \Omega | \; \omega \in A \text{ or } \omega \in B\}$.
- *Binary intersection* which results in the *intersection* of the corresponding sets, $A \cap B = \{\omega \in \Omega | \; \omega \in A \text{ and } \omega \in B\}$.
- *Union* of an arbitrary collection of subsets of Ω (not necessarily countable) which results in the set, $\cup_{i \in I} A_i = \{\omega \in \Omega | \; \exists i \in I \text{ such that } \omega \in A_i\}$.
- *Intersection* of a collection of subsets of Ω which results in the set, $\cap_{i \in I} A_i = \{\omega \in \Omega | \; \forall \, i \in I, \; \omega \in A_i\}$.

The binary union and binary intersection are seen to be consistent with the union and intersection operations.

The following set equalities hold:

- *Commutativity,* $A \cup B = B \cup A$, $A \cap B = B \cap A$.
- *Associativity,* $(A \cup B) \cup C = A \cup (B \cup C)$, $(A \cap B) \cap C = A \cap (B \cap C)$.
- *Distributivity,* $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$, $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.
- *De Morgan laws,* $(A \cup B)^c = A^c \cap B^c$, $(A \cap B)^c = A^c \cup B^c$.

The properties of distributivity and the De Morgan laws extend from a binary union to a finite union and from a binary intersection to a finite intersection.

The family $\{A_n \subseteq \Omega, n \in I\}$ with I a countable set is said to be *disjoint* if for all $n,\ m \in I$, $n \neq m$ implies that the sets A_n and A_m are disjoint. The disjointness can now be written as $A_n \cap A_m = \emptyset$. The family $\{A_n \subseteq \Omega, n \in \mathbb{Z}_+\}$ is said to be a *partition* of Ω if it is disjoint and if $\cup_{n\in\mathbb{Z}_+} A_n = \Omega$. If in addition the index set I is finite then it is called a *finite partition* of Ω.

The family $\{A_n, n \in \mathbb{Z}_+\}$ is said to be: *increasing* if for all $n \in \mathbb{Z}_+$, $A_n \subseteq A_{n+1}$; *decreasing* if for all $n \in \mathbb{Z}_+$, $A_{n+1} \subseteq A_n$; and *monotone* if it is either increasing or decreasing.

Definition 2.3.2. The *limit* of a countable family of subsets of Ω, $\{A_n, n \in \mathbb{Z}_+\}$, is defined as follows.

- If the family is increasing then $\lim_{n\to\infty} A_n = \sup_{n\in\mathbb{Z}_+} A_n = \cup_{n\in\mathbb{Z}_+} A_n$.
- If the family is decreasing then $\lim_{n\to\infty} A_n = \inf_{n\in Z_+} A_n = \cap_{n\in Z_+} A_n$.
- If the family is not monotone then define

$$\limsup_{n\in\mathbb{Z}_+} A_n \ = \cap_{n\in\mathbb{Z}_+} \cup_{k=n}^{\infty} A_k, \ \ \liminf_{n\in\mathbb{Z}_+} A_n = \cup_{n\in\mathbb{Z}_+} \ \cap_{k=n}^{\infty} A_k.$$

Note that the family $\{\cup_{k=n}^{\infty} A_k \subseteq \Omega, n \in \mathbb{Z}_+\}$ is a decreasing sequence in $n \in \mathbb{Z}_+$, and, correspondingly, $\{\cap_{k=n}^{\infty} A_k \subseteq \Omega, n \in \mathbb{Z}_+\}$ is a increasing sequence in $n \in \mathbb{Z}_+$ hence the expressions of $\limsup A_n$ and $\liminf A_n$ are well defined. In general $\liminf A_n \subseteq \limsup A_n$. One says that $\lim\ A_n$ *exists* if $\limsup A_n = \liminf A_n$ and in that case one defines, $\lim\ A_n = \limsup\ A_n = \liminf\ A_n$.

σ-Algebras

As mentioned before, a probability measure is defined on a σ-algebra. The concept of a σ-algebra is introduced below. A family of sets is said to be *closed* with respect to a set operation if the set operation performed on sets of the family produces again a set of the family.

Definition 2.3.3. An *algebra* F_0 (or *field*) on Ω is a collection of subsets of Ω such that:

(1)$\Omega \in F_0$;
(2)if $A \in F_0$ then $A^c \in F_0$; or, equivalently, F_0 is closed with respect to complementation;
(3)if $A, B \in F_0$ then $(A \cup B) \in F_0$; or, equivalently, F_0 is closed with respect to binary unions.

From (2) and (3) above it follows that F_0 is also closed with respect to binary intersections because, $\forall\ A,\ B \in F, A \cap B = (A^c \cup B^c)^c \in F$.

Definition 2.3.4. A *σ-algebra F* (or a *σ-field*) on Ω is a collection of subsets of Ω such that:

(1)$\Omega \in F$; and
(2)if $A \in F$ then $A^c \in F$; or, equivalently, F is closed with respect to taking complements;
(3)if $\{A_n \in F,\ n \in \mathbb{Z}_+\}$ is a countable collection of sets in F then $(\cup_{n\in\mathbb{Z}_+} A_n) \in F$; or, equivalently, F is closed with respect to countable unions.

The ordered tuple (Ω, F) consisting of a set Ω and a σ-algebra F will be called a *measurable space.*

A subset $G \subseteq \text{Pwrset}(\Omega)$ of subsets of a measurable space (Ω, F) is said to be a *sub-σ-algebra* of F if (1) G is a σ-algebra of subsets of Ω and (2) $G \subseteq F$.

The definition of a σ-algebra is a restriction on the earlier attempt to define a probability measure because by condition (3) the closure with respect to unions is to hold only for countable unions and not for arbitrary unions.

It follows from the definition that a σ-algebra is also closed with respect to countable intersections due to Condition (2) and the formula of the De Morgan laws according to $\cap_{i\in\mathbb{Z}_+} A_i = (\cup_{i\in\mathbb{Z}_+} A_i^c)^c \in F$.

Example 2.3.5. Examples of σ-algebras on a set Ω are: (1) $\{\emptyset, \Omega\}$.
(2) $\{\emptyset, A, A^c, \Omega\}$ for any subset A of Ω. Other examples are constructed below.

Proposition 2.3.6. *Consider a probability space (Ω, F) and a family of sets $\{A_i \in F,\ i \in I\}$, there exists a smallest σ-algebra, denoted by $F(\{A_i, i \in I\})$, such that for all $i \in I$, $A_i \in F(\{A_i, i \in I\})$. It will be called the* σ-algebra generated by the family of subsets $\{A_i \subseteq \Omega, i \in I\}$. *Note that the index set I need not be countable.*

Proof. Define the set F_s of sub-σ-algebras each of which contains the family of sets $\{A_i \in F,\ i \in I\}$. Note that $F \in F_s$ hence the set F_s is not empty. Define then the intersection,

$$F_c = \cap_{F_e \in F_s} F_e.$$

Because $F \in F_s$, the intersection is not empty and the σ-algebra F_c is well defined. It is then an exercise that prove F_c is a sub-σ-algebra and that it contains the family of sets $\{A_i \subseteq \Omega, i \in I\}$. By definition of F_c as an intersection over the set of sub-σ-algebras F_s, it is the smallest such sub-σ-algebra. □

Example 2.3.7. If $A \subseteq \Omega$ then $F(A) = \{\emptyset, A, A^c, \Omega\}$.

If F_1, F_2 are two σ-algebras on Ω then the σ-algebra generated by F_1, F_2 is denoted by $F_1 \vee F_2$. Similarly, if $\{F_i, i \in I\}$ is a family of σ-algebras then one denotes by $\vee_{i\in I} F_i$ the σ-algebra generated by this family according to Proposition 2.3.6.

The Borel Sigma-Algebra of the Real Numbers

The set of the real numbers and the vector space of tuples of the real numbers are frequently used in this book. Therefore probability theory on those sets is discussed in detail.

Definition 2.3.8. *Borel σ-algebra.* Consider the set of the real numbers $\mathbb{R}$. Define the *Borel σ-algebra* on the set of the real numbers, denoted by $B(\mathbb{R})$, as the smallest σ-algebra of subsets of $\mathbb{R}$ which contains all open intervals of the form $(c,+\infty) \subset \mathbb{R}$ for all $c \in \mathbb{R}$. Call any element of $B(\mathbb{R})$ a *Borel set.*

For any Borel set $X \subseteq \mathbb{R}$, define the *Borel σ-algebra* of subsets of X as the smallest σ-algebra which contains all open subsets of X. Denote this σ-algebra by $B(X)$.

Consider the set $\mathbb{R}^n$ of n-tuples of the real numbers for an integer $n \in \mathbb{Z}_+$. Define the *Borel σ-algebra* of $\mathbb{R}^n$ as the smallest σ-algebra which contains all open intervals of the form,

$$\forall\, c \in \mathbb{R}^n,\ \prod_{i=1}^{n} (c_i,+\infty) = \{x \in \mathbb{R}^n |\ \forall\, i \in \mathbb{Z}_n,\ c_i < x_i\}.$$

Denote this σ-algebra by $B(\mathbb{R}^n)$. The construction used is similar to the construction of a σ-algebra on a product space.

On a topological set X (a topology consists of subsets of X which are either open or closed), define the *Borel σ-algebra* as the smallest σ-algebra generated by all open subsets of X. Denote the corresponding σ-algebra by $B(X)$.

2.4 Probability Measures

Definition 2.4.1. A *probability measure* on a measurable space (Ω,F) is a function $P : F \to \mathbb{R}_+$ for which the following conditions all hold:

1. *mass one*: $P(\Omega) = 1$;
2. *finite additivity*: if $\{A_k \in F,\ k \in \mathbb{Z}_n\}$ with $\mathbb{Z}_n = \{1,2,\ldots,n\}$, is a finite familiy of disjoint measureable sets then,
$$P(\cup_{k\in\mathbb{Z}_n} A_k) = \sum_{k\in\mathbb{Z}_n} P(A_k);$$
3. *monotone sequential continuity at the empty set*: if $\{A_k \in F,\ k \in \mathbb{Z}_+\}$ is a decreasing sequence of measurable sets such that $\lim_{k\to\infty} A_k = \cap A_k = \emptyset$ then $\lim_{k\to\infty} P(A_k) = 0$.

A function $M : F \to \mathbb{R}_+$ is called a *σ-finite measure* if the following conditions all hold: (1) the condition of finite additivity, (2) of monotone sequential continuity at the empty set, and (3) if there exist a countable collection of sets $\{A_i \in F,\ \forall\, i \in I \subseteq \mathbb{Z}_+\}$ (generally take to be disjoint) such that $\Omega = \cup_{i\in I} A_i$ and for all $i \in I, M(A_i) < \infty$.

Example 2.4.2. Consider the set of the natural numbers $\mathbb{N} = \{0,1,\ldots\}$. Define the *Poisson probability measure* with parameter $\lambda \in (0,\infty)$ on this space by the formula, $P : \text{Pwrset}(\mathbb{N}) \to \mathbb{R}_+$, $P(A) = \sum_{k\in A}\ p(k)$, where $p(k) = \lambda^k \exp(-\lambda)/k!$. Then P is a probability measure.

Example 2.4.3. Consider Lebesgue measure L on the set of the positive real numbers $(\mathbb{R}_+, B(\mathbb{R}_+)$. Because the left-closed and right-open intervals $\{[i,i+1) \subseteq \mathbb{R}_+,\ i \in \mathbb{N}\}$, cover the positive real numbers, $\mathbb{R}_+ = \cup_{i\in\mathbb{N}}\ [i,i+1)$, and for all $i \in \mathbb{N}$, $L([i,i+1)) = 1$, Lebesgue measure is a σ-finite measure on the positive real numbers.

Proposition 2.4.4. *Consider a probability measure P on a measurable space (Ω, F). Then the following properties all hold.*

(a) $P(\emptyset) = 0$.
(b) Monotonicity. *If $A_1,\ A_2 \in$ such that $A_2 \subseteq A_1$ then $P(A_2) \leq P(A_1)$.*
(c) Range. *For all $A \in F$, $P(A) \in [0,1]$.*
(d) Strong binary additivity. *If $A_1,\ A_2 \in F$ then*

$$P(A_1) + P(A_2) = P(A_1 \cup A_2) + P(A_1 \cap A_2).$$

(e) Finite subadditivity. *If $\{A_k \in F, k \in \mathbb{Z}_n\}$ then $P(\cup_{k\in\mathbb{Z}_n}) \leq \sum_{k\in\mathbb{Z}_n} P(A_k)$.*

Proof. (a) $1 = P(\Omega) = P(\Omega \cup \emptyset) = P(\Omega) + P(\emptyset) = 1 + P(\emptyset)$ implies $P(\emptyset) = 0$.
(b)

$$\begin{aligned} A_1 &= (A_1 \cap A_2) \cup (A_1 \cap A_2^c) = A_2 \cup (A_1 \cap A_2^c), \\ P(A_1) &= P(A_2) + P(A_1 \cap A_2^c) \geq P(A_2), \text{ because } P : F \to \mathbb{R}_+. \end{aligned}$$

(c) $A \in F$ and $\emptyset \subseteq A \subseteq \Omega$ imply by (b) that $0 = P(\emptyset) \leq P(A) \leq P(\Omega) = 1$.
(d)

$$\begin{aligned} A_1 \cup A_2 &= (A_1 \cap A_2^c) \cup (A_1 \cap A_2) \cup (A_1^c \cap A_2), \text{ hence,} \\ P(A_1 \cup A_2) &= P(A_1 \cap A_2^c) + P(A_1 \cap A_2) + P(A_1^c \cap A_2), \text{ by disjointness,} \\ &= [P(A_1 \cap A_2^c) + P(A_1 \cap A_2)] + \\ &\quad + [P(A_1^c \cap A_2) + P(A_1 \cap A_2)] - P(A_1 \cap A_2) \\ &= P(A_1) + P(A_2) - P(A_1 \cap A_2). \end{aligned}$$

(e) From (c) and (d) follows that,

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq P(A_1) + P(A_2).$$

from which the general finite case follows by induction. □

Proposition 2.4.5. *Consider a probability space (Ω, F, P).*

(a) monotone sequential continuity at a set. *If $A \in F$ and $\{A_k \in F,\ k \in \mathbb{Z}_+\}$ is a decreasing sequence of sets such that $\lim_{k\to\infty} A_k = A$ then $\lim_{k\to\infty} P(A_k) = P(A)$. A corresponding result holds for increasing sequences.*
(b) Characterization of a probability measure. *The function $P : F \to [0,1]$ is a probability measure on (Ω, F) if and only if*
(b.1) $P(\Omega) = 1$; and
(b.2) if $\{A_k \in F,\ k \in \mathbb{Z}_+\}$ is a family of pairwise disjoint sets then $P(\cup_{k\in\mathbb{Z}_+}) = \sum_{k\in\mathbb{Z}_+} P(A_k)$.

(c) σ-subadditivity. *If* $\{A_k \in F,\ k \in \mathbb{Z}_+\}$ *is a family of measurable sets (disjointness is not imposed) such that* $\cup_{k\in\mathbb{Z}_+} A_k \in F$ *then*

$$P(\cup_{k\in\mathbb{Z}_+} A_k) \leq \sum_{k\in\mathbb{Z}_+} P(A_k).$$

Proof. (a)

$$\begin{aligned} A_k &= (A_k \cap A^c) \cup (A_k \cap A) = (A_k \cap A^c) \cup A, \\ &\{A_k \cap A^c),\ k \in \mathbb{Z}_k\},\ \text{is a decreasing family,}\ \lim_{k\to\infty}(A_k \cap A^c) = \emptyset, \\ \lim_{k\to\infty} P(A_k) &= \lim[P(A_k \cap A^c) + P(A)] = P(\emptyset) + P(A) = P(A). \end{aligned}$$

□

Proposition 2.4.6. *Consider a probability space* (Ω, F, P). *Recall the notation of* $\liminf A_k$ *and* $\limsup A_k$ *of Def. 2.3.2.*

(a)Then,

$$P(\liminf A_k) \leq \liminf P(A_k) \leq \limsup P(A_k) \leq P(\limsup A_k).$$

(b)If $\lim_{k\to\infty} A_k$ *exists, hence* $\liminf_{k\to\infty} A_k = \limsup_{k\to\infty} A_k$*, then*

$$\lim_{k\to\infty} P(A_k) = P(\lim_{k\to\infty} A_k).$$

Proof. Note that $\{\cap_{k\geq m} A_k \in F,\ m \in \mathbb{Z}_+\}$ is an increasing sequence while $\{\cup_{k\geq m} A_k \in F,\ m \in \mathbb{Z}_+\}$ is a decreasing sequence. It then follows from monotone sequential continuity, and monotonicity that,

$$\begin{aligned} P(\liminf_{k\to\infty} A_k) &= \lim_{m\to\infty} P(\cap_{k\geq m} A_k) \leq \liminf_{m\to\infty} P(A_m) \\ &\leq \limsup_{m\to\infty} P(A_m) \leq \lim_{m\to\infty} P(\cup_{k\geq m} A_k) = P(\limsup_{m\to\infty} A_m). \end{aligned}$$

□

The following lemma is extremely useful for computing probabilities of the limit of a sequence of sets.

Lemma 2.4.7 (Borel-Cantelli lemma, first half). *Let* $\{A_n \in F, n \in \mathbb{Z}_+\}$. *Then,*

$$\sum_{n\in\mathbb{Z}_+} P(A_n) < \infty \Rightarrow P(\{\limsup_{n\to\infty} A_n\}) = 0.$$

Proof.

$$\begin{aligned} &P(\limsup_{n\to\infty} A_n) = P(\lim_{n\to\infty} \cup_{k\geq n} A_k),\ \text{by def. of}\ \limsup A_n, \\ &= \lim_{n\to\infty} P(\cup_{k\geq n} A_k), \text{by decreasing sequential continuity,} \\ &\leq \lim_{n\to\infty} \sum_{k\geq n} P(A_k), \text{by sub-additivity,} \\ &= 0, \text{by assumption.} \end{aligned}$$

□

Probability Measures on the Real Numbers

In a course on real analysis one may learn about the Lebesgue measure m on the set of the real numbers $\mathbb{R}$ and the construction of the σ-algebra $M \subset \text{Pwrset}(\mathbb{R})$ of Lebesgue measureable sets. The measure and the σ-algebra of Lebesgue-measureable sets are related.

Each Borel set is a Lebesgue-measureable set but the converse does not hold. The concept of a Borel space was introduced by D. Blackwell.

Definition 2.4.8. A *Borel space* is a Borel subset of a complete separable metric space. Alternatively, it is bijectively related to a Borel subset of a complete seperable metric space.

An example of a complete separable metric space is the set of the real numbers and also the set of n-tuples of the real numbers, for $n \in \mathbb{Z}_+$.

The following result establishes the correspondence between probability measures on the real numbers and probability distribution functions. The integral associated with the probability distribution function is a Lebesgue integral associated with the Lebesgue measure.

Theorem 2.4.9. *P is a probability measure on $(\mathbb{R}, B(\mathbb{R}))$ if and only if there exists a pdf $f : \mathbb{R} \to \mathbb{R}_+$ such that P and f are related by*

$$P((a,b]) = f(b) - f(a),\ \forall\ a,b \in \overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\} \cup \{-\infty\},\ a < b.$$

Then $(\mathbb{R},\ B(\mathbb{R}),\ P)$ is a probability space. Given P, f is unique in the class of pdf's as defined in Def. 2.1.1.

Similarly, any probability distribution function f on $\mathbb{R}^n$ induces a probability measure P by the extension of the above formula to $\mathbb{R}^n$ and then $(\mathbb{R},\ B(\mathbb{R}^n),\ P_f)$ is a probability space.

Example 2.4.10. If f is a Gaussian probability distribution function on $\mathbb{R}^n$ for a strictly positive integer $n \in \mathbb{Z}_+$ then the formula of Thm. 2.4.9 induces a probability measure P such that $(\mathbb{R}^n,\ B(\mathbb{R}^n),\ P)$ is a probability space.

Finite Probability Spaces

Examples of finite probability spaces are used throughout the book to illustrate various properties. An introduction to the terminology and to the notation of finite probability spaces follows.

A *finite probability space* consists of the probability space (Ω, F, P) if there exists an integer $n \in \mathbb{Z}_+$, such that $\Omega = \mathbb{Z}_n = \{1, 2, \ldots, n\}$, $F \subseteq \text{Pwrset}(\Omega)$ is the set consisting of all subsets of Ω, and $P : F \to [0,1]$ is a probability measure. Call P a *uniform probability measure* of the finite probability space if $P(\{i\}) = 1/n$ for all $i \in \Omega$.

For any sub-σ-algebra $G \subseteq F$, call the smallest collection $A_G = \{A_1, \ldots, A_{k_G}\} \subseteq F$ such that (1) $G = \sigma(A_G)$ and (2) $\forall\, A_i \in A_G$, $P(A_i) > 0$, the *set of atoms* of the σ-algebra G and call each $A_i \in A_G$ an *atom* of G. The empty set is by definition omitted from inclusion in the set of atoms A_G. This definition implies that $\Omega = \cup_{i=1}^{k_G} A_i$ and for all $i,\ j \in \Omega$ with $i \neq j$, $A_i \cap A_j = \emptyset$.

Example 2.4.11. Consider the finite probability space (Ω, F, P) with $\Omega = \mathbb{Z}_7 = \{1, 2, \ldots, 7\}$ and a uniform probability measure $P : F \to [0,1]$, hence for all $i \in \Omega$, $P(\{i\}) = 1/7$.

Three sub-σ-algebras of this probability space are displayed in Fig. 2.1, in each case by their atoms.

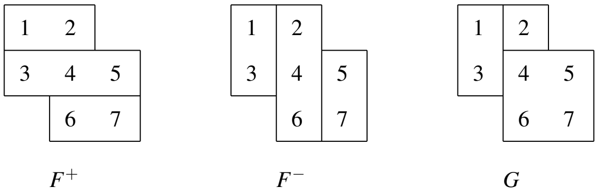


Fig. 2.1 Diagrams of several σ-algebras of the finite probability space of Example 2.4.11.

Independence

The reader may know the concept of independence from ordinary conversation. The concept of independence used in probability theory has an analogous meaning. Independence is a relation of σ-algebras jointly with a probability measure, rather than a relation of measurable sets.

Definition 2.4.12. (a) A finite family of σ-algebras $\{F_k,\ k \in \mathbb{Z}_n\}$ is said to be *independent* with respect to the probability measure P if for any finite collection,

$$(\forall\, A_k \in F_k,\ k \in \mathbb{Z}_n),\ P(\cap_{k=1}^{n} A_k) = \prod_{k=1}^{n} P(A_k).$$

(b) Any countably-infinite family of sub-σ-algebras $\{F_n \subseteq F,\ n \in \mathbb{Z}_+\}$ is said to be *independent* with respect to P if any finite subfamily is independent with respect to P.

(c) A countable family of sets $\{A_n, n \in \mathbb{Z}_+\}$ is said to be *independent* with respect to P if the associated family of σ-algebras $\{F(A_n) \subseteq F,\ n \in \mathbb{Z}_+\}$ generated by these sets is independent with respect to P.

There follows a sufficient condition for the independence of a family of sets.

Proposition 2.4.13. *Let $\{G_i \subseteq \Omega, i \in I\}$ be a family of non-empty subsets of Ω such that:*

(1)for all $i \in I$, G_i is closed with respect to intersection;
(2)the family $\{G_i, i \in I\}$ is such that for any finite set $J \subseteq I$ and,

$$\forall \, \{A_j \in G_j, \; j \in J\}, \;\; P(\cap_{j \in J} A_j) \; = \; \prod_{j \in J} P(A_j).$$

Then the family of σ-algebra's $\{F(G_i), i \in I\}$ is independent with respect to P.

2.5 Random Variables

In daily life one has the need to speak about the probability that a variable is lower than a specified value, for example the temperature. There is thus a need to compute the probability of such a relation for a variable. But, to define that probability, it is necessary that the set or event belongs to the σ-algebra F, $\{\omega \in \Omega \mid x(\omega) \leq c\} \in F$ for all $c \in \mathbb{R}$. A variable should therefore have this property. The following definition is now motivated.

Definition 2.5.1. A *random variable* defined on the measurable space (Ω, F) and taking values in the measurable space (X, G) is a function $x : \Omega \to X$ such that,

$$\forall \, A \in G, \quad x^{-1}(A) = \{\omega \in \Omega | x(\omega) \in A\} \in F.$$

If $(X, G) = (\mathbb{R}, B(\mathbb{R}))$, where $B(\mathbb{R})$ denotes the Borel σ-algebra, then $x : \Omega \to \mathbb{R}$ is a random variable if and only if,

$$\forall \, c \in \mathbb{R}, \;\; x^{-1}((-\infty, c]) \; = \; \{\omega \in \Omega | x(\omega) \leq c\} \in F.$$

This characterization of a real-valued random variable follows directly from a characterization of the Borel σ-algebra as the σ-algebra generated by the intervals.

Examples of random variables follow. The *indicator function* of a subset $A \subseteq \Omega$ is defined by,

$$I_A(\omega) \; = \; \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \notin A. \end{cases}$$

An indicator function of a set $A \subseteq \Omega$ is a random variable if and only if $A \in F$.

The following elementary results are easily deduced: For subsets A, B of Ω the following relations hold: $I_A I_B = I_{A \cap B}$; $I_{A \cup B} = I_A + I_B - I_{A \cap B}$; $A = B^c$ if and only if $I_A I_B = 0$ and $I_A + I_B = 1$; $A \subseteq B$ if and only if $I_A \leq I_B$.

A *simple* random variable x is defined as a finite linear combination of indicator functions of measurable sets,

$$\exists \, n \in \mathbb{Z}_+, \; \exists \, \{c_k \in \mathbb{R}, \; k \in \mathbb{Z}_n\}, \; \exists \, \{A_k \in F, k \in \mathbb{Z}_n\}, \; \text{such that, } x = \sum_{k=1}^{n} \; c_k \, I_{A_k}.$$

It can be proven that for any simple random variable there exists a representation of this sum with a disjoint family $\{A_k, k \in Z_n\}$, hence any such family can be assumed to be a finite partition of Ω.

If x, y are real-valued random variables, then so are $x+y$, $x-y$, xy, $x \wedge y = \min\{x,\ y\}$, and $x \vee y = \max\{x,\ y\}$. Also $xy^{-1}$is a random variable if on the set $\{\omega \in \Omega | y(\omega) = 0\}$ it is assigned to an arbitrarily-chosen real number.

Define for a real-valued random variable $x : \Omega \to \mathbb{R}$ its *positive-negative* decomposition by,

$$x_+(\omega) = \max\{x(\omega), 0\},\ x_-(\omega) = \max\{-x(\omega), 0\} = \begin{cases} -x(\omega), & \text{if } x(\omega) \leq 0, \\ 0, & \text{else;} \end{cases}$$

$$\text{then } x = x_+ - x_-.$$

Two random variables taking values in the same set, say $x, y : \Omega \to \mathbb{R}$, are said to be *equal almost surely* if $P(\{\omega \in \Omega | x(\omega) = y(\omega)\}) = 1$. This property is denoted by $x = y$ a.s. One also says that y is an *almost sure modification* of x.

A function $f : X \to Y$ on the measurable spaces (X, G) and (Y, H) is called a *measurable function* if,

$$\forall\, A \in H,\ f^{-1}(A) = \{x \in X | f(x) \in A\} \in G.$$

If (X, G), (Y, H), and (Z, I) are measurable spaces and $f : X \to Y$, $g : Y \to Z$ are measurable functions then $g \circ f : X \to Z$ is a measurable function. Thus if $x : \Omega \to X$ is a random variable and $f : X \to Y$ is a measurable function, then $f(x) : \Omega \to Y$ is also a random variable.

Real-Valued Measurable Functions

Definition 2.5.2. The real-valued function $f : \mathbb{R}^m \to \mathbb{R}^n$ for $n,\ m \in \mathbb{Z}_+$ is called *Borel measureable* if

$$\forall\, A \in B(\mathbb{R}^n),\ f^{-1}(A) = \{x \in \mathbb{R}^m |\ f(x) \in A\} \in B(\mathbb{R}^m).$$

The function is called *Lebesgue measureable* if the corresponding definition holds for the σ-algebras of Lebesgue-measureable sets.

Example 2.5.3. For any constant $c \in \mathbb{R}$, the function $f : \mathbb{R} \to \mathbb{R}\ f(x) = x + c$ is Lebesgue-measureable and so is $g : \mathbb{R} \to \mathbb{R}$, $g(x) = cx$. If the functions $f,\ g : \mathbb{R} \to \mathbb{R}$ are Lebesgue-measureable, then so are fg, $f+g$, $g-f$, $f \vee g = \max\{f, g\}$, and $f \wedge g = \min\{f, g\}$. Corresponding results hold for multivariable Lebesgue-measureable functions.

If $\{f_k : \mathbb{R} \to \mathbb{R},\ \forall\, k \in \mathbb{Z}_+\}$ is a sequence of Lebesgue-measureable functions then so are: $\inf_{k \in \mathbb{Z}_+} f_k$, $\sup_{k \in \mathbb{Z}_+} f_k$, $\liminf_{k \in \mathbb{Z}_+} f_k$, and $\limsup_{k \in \mathbb{Z}_+} f_k$.

Consider a measurable function $f : \mathbb{R}^n \to \mathbb{R}^m$. If f is injective then there exists an inverse function $f^{-1} : f(X) \subseteq \mathbb{R}^m \to \mathbb{R}^n$. If the inverse function is also a measureable function then call the function f *bimeasurable*.

If a random variable $x : \Omega \to \mathbb{R}^n$ has a probability distribution pdf_x and if $f : \mathbb{R}^n \to \mathbb{R}^n$ is a continuously-differentiable and bijective map then the random variable $f(x) : \Omega \to \mathbb{R}^n$ has a probability distribution function which can be obtained by the well known change-of-variable formula.

Transformation of a Probability Space by a Random Variable

Theorem 2.5.4. Transformation of a probability measure by a random variable. *Consider a probability space (Ω,F,P), a measurable space (X,G), and a random variable $x : \Omega \to X$.*

Define on the measureable space (X,G), the function $P_x : G \to \mathbb{R}_+$ by $P_x(A) = P(\{\omega \in \Omega | \, x(\omega) \in A\})$. Then P_x is a probability measure and (X,G,P_x) is a probability space.

Proof.

$$\begin{aligned}
P_x(X) &= P(\{\omega \in \Omega | \, x(\omega) \in X\}) = 1, \\
&\quad \forall\, A_1,\, A_2 \in G,\, A_1 \cap A_2 = \emptyset,\ \ P_x(A_1) + P_x(A_2) \\
&= P(\{\omega \in \Omega | \, x(\omega) \in A_1\}) + P(\{\omega \in \Omega | \, x(\omega) \in A_2\}) \\
&= P(\{\omega \in \Omega | \, x(\omega) \in A_1 \cup A_2\}) = P_x(A_1 \cup A_2), \\
&\quad \text{hence finite additivity holds;} \\
&\quad \{A_k \in G,\ \forall\, k \in \mathbb{Z}_+\},\ \text{a decreasing sequence such that} \lim_{k\to\infty} A_k = \emptyset, \\
&\quad \lim_{k\to\infty} P_x(A_k) = \lim\ P(\{\omega \in \Omega | \, x(\omega) \in A_k\}) = 0, \\
&\quad \text{because}\ \{\{\omega \in \Omega | \, x(\omega) \in A_k\} \in F,\ \forall\, k \in \mathbb{Z}_+\},
\end{aligned}$$

is a decreasing sequence and because monotone sequential continuity holds for P. Thus P_x is a probability measure and (X,G,P_x) is a probability space. □

Example 2.5.5. Consider a probability space (Ω,F,P) and a random variable $x : \Omega \to \mathbb{N}$ having a Poisson probability distribution with parameter $\lambda \in (0,\infty)$. Then it follows from Theorem 2.5.4 that x induces a probability measure P_x on $(\mathbb{N},B(\mathbb{N}))$ where the σ-algebra on the countable set $\mathbb{N}$ is not needed. Note that then, $P_x(k) = P(\{\omega \in \Omega | \, x(\omega) = k\}) = \lambda^k \exp(-\lambda)/k!$ hence P_x is the probability frequency function of x.

Example 2.5.6. Consider a probability space (Ω,F,P) and a random variable $x : \Omega \to \mathbb{R}^n$ having a Gaussian probability distribution function with density p_x. Then it follows from Theorem 2.5.4 that x induces a probability measure P_x on $(\mathbb{R}^n,B(\mathbb{R}^n))$. Note that then, P_x has an associated probability distribution function with the probability density p_x, for all $u \in \mathbb{R}^n$,

$$P_x((-\infty,u)) = P(\{\omega \in \Omega | \, x(\omega) \in \prod_{i=1}^{n} (-\infty,u_i]\}) = \int_{-\infty}^{u_1} \ldots \int_{-\infty}^{u_n} p_x(v) dv.$$

Support of a Measure of a Random Variable

Definition 2.5.7. (a)Consider a probability space (Ω, F, P) and a finite-valued or countably-valued random variable $x : \Omega \rightarrow X \subset \mathbb{Z}^n$. Denote the probability frequency function of the random variable x by $p_x : X \rightarrow \mathbb{R}_+$, $p_x(k) = P(\{x(\omega) = k\})$. Define the *support set* of the probability frequency function p_x as the set, $X_{supp} = \{k \in X |\ p_x(k) > 0\}$. One says that the random variable x has *full support* on X if $X_{supp} = X$.

(b)Consider a probability space (Ω, F, P), a random variable $x : \Omega \rightarrow X = \mathbb{R}^n$ with $n \in \mathbb{Z}_+$ for which there exists a probability density function $p_x : \mathbb{R}^n \rightarrow \mathbb{R}_+$.
Define the *support set* of the random variable x as the set
$X_{supp} = \{v \in X = \mathbb{R}^n |\ p_x(v) > 0\}$. One says that the random variable x has *full support* on X if $X_{supp} = X$.

The interest in the concept of a support set is to determine whether or not the support set X_{supp} equals the entire state set X. If $X_{supp} \subsetneqq X$ then part of the state set is not used and the state set can therefore be reduced in size.

An example of the support of a random variable with a conditional density function is provided in Example 2.7.8. An example of the support of a countably-valued random variable follows.

Example 2.5.8. *The support of a Poisson measure*. Consider a probability space (Ω, F, P), a random variable $x : \Omega \rightarrow \mathbb{N} = X$ with values in the natural numbers, the probability frequency function of Poisson type with parameter $\lambda \in \mathbb{R}_{s+}$, $p_x : \mathbb{N} \rightarrow \mathbb{R}_+$, with $p_x(k) = \lambda^k \exp(-\lambda)/k!$, and the induced probability space on the range space $(\mathbb{N}, B(\mathbb{N}), P_x)$.

The support set of the random variable is then the set,
$X_{supp} = \{k \in X = \mathbb{N} |\ p_x(k) > 0\}$. It then follows that $X_{supp} = X = \mathbb{N}$ hence this random variable has full support on the state set $X = \mathbb{N}$.

If the random variable x had in stead been defined on the set of the integers $X = \mathbb{Z}$ with the Poisson measure as defined above, then of course $X_{supp} = \mathbb{N} \subsetneqq \mathbb{Z} = X$.

Finite-Valued Random Variables

In the book finite-valued random variables are used as illustration of properties. Such random variables have a particular representation defined next.

Definition 2.5.9. A real-valued or a $\mathbb{R}^n$-valued random variable is said to be a *finite-valued random variable* of the space in which it takes values is finite. For example, the random variable $x : \Omega \rightarrow \mathbb{Z}_3 = \{1, 2, 3\}$ is finite valued.

For a finite-valued random variable, define its *indicator representation* by the formulas,

$$
\begin{aligned}
&\quad x : \Omega \to X = \{C_1, C_2, \ldots, C_{k_x}\} \subseteq \mathbb{R}^n,\ n,\ k_x \in \mathbb{Z}_+,\\
&x = C_x\, x_I,\\
&\quad C_x \in \mathbb{R}_+^{n \times k_x},\ x_I : \Omega \to \{0,1\}^{k_x},\ C_x = \begin{pmatrix} C_1 & C_2 & \ldots & C_{k_x-1} & C_{k_x} \end{pmatrix} \in \mathbb{R}_+^{n \times k_x},\\
&x_I = \begin{pmatrix} x_{I,1} \\ x_{I,2} \\ \vdots \\ x_{I,k_x} \end{pmatrix},\ x_{I,i}(\omega) = I_{\{x(\omega)=C_i\}}(\omega).
\end{aligned}
$$

Call x_I the vector *indicator representation* of the finite-valued random variable x and C_x the *value-matrix* of x. In the remainder of the book, one often denotes the random variable x_I by only x with the understanding it is in the indicator representation.

Definition 2.5.10. Consider a tuple of finite-valued random variables denoted by,

$$
\begin{aligned}
&x : \Omega \to \mathbb{R}^{n_x},\ y : \Omega \to \mathbb{R}^{n_y},\ n_x,\ n_y \in \mathbb{Z}_+;\\
&Q_{x,y} \in \mathbb{R}_+^{n_x \times n_y},\ Q_{x,y,(i,j)} = E[x_i y_j] = E[I_{A_i} I_{B_j}] = P(A_i \cap B_j),\ \forall\, i \in \mathbb{Z}_{n_x},\ j \in \mathbb{Z}_{n_y};\\
&p_x = E[x],\ p_y = E[y].
\end{aligned}
$$

For a tuple of finite-valued random variables one works with a multiple array according to,

$$
\begin{aligned}
&Q_{x_1,\ldots,x_k} \in \mathbb{R}_+^{n_{x_1} \times \ldots n_{x_k}},\ k,\ n_{x_1},\ \ldots,\ n_{x_k} \in \mathbb{Z}_+,\\
&Q_{x_1,\ldots,x_k,(i_1,\ldots,i_k)} = E[x_{1,i_1} \ldots x_{x_k,i_k}],\ \forall\, (i_1,\ldots,i_k) \in \mathbb{Z}_{n_{x_1}} \times \ldots \mathbb{Z}_{n_{x_k}}.
\end{aligned}
$$

Proposition 2.5.11. *Elementary propertes of a tuple of finite-valued random variables each in the indicator representation, are stated for further reference.*

(a)$p_x = Q_{x,y} 1_{n_y}$ *and* $p_y = Q_{x,y}^T 1_{n_x}$.
(b)$F^x,\ F^y$ *are independent if and only if* $Q_{x,y} = p_x p_y^T$.

The proofs of the above properties follows directly from the corresponding definitions.

Random Variables and σ-Algebras

The viewpoint in which one considers the spaces that variables generate rather than the variables themselves, is called the *geometric approach* to random variables. For random variables, the associated spaces are σ-algebras.

Questions in this context are: Given a σ-algebra generated by a random variable, which transformations on the random variable leave the space invariant? What is a canonical form for the generating variable of a σ-algebra?

Let $x : \Omega \to X$ be a random variable taking values in the measurable space (X, G). Define,

$$x^{-1}(A) = \{\omega \in \Omega |\, x(\omega) \in A\}, \;\; \forall\, A \in G,$$
$$x^{-1}(G) = = \{x^{-1}(A) \in F |\, \forall A \in G\},$$

and note the abuse of notation for $x^{-1}(A)$ and for $x^{-1}(G)$. It may be proven that $x^{-1}(G)$ is a σ-algebra. Denote $F^x = F(x) = x^{-1}(G)$. If $x = \sum_{i=1}^{n} c_i I_{A_i}$ is a simple random variable with the associated family of subsets $\{A_i \in F, i \in Z_n\}$ being a partition of Ω then $x^{-1}(B(\mathbb{R})) = F(\{A_i, i \in Z_n\})$ is the σ-algebra generated by the partition $\{A_i \in F, i \in I\}$.

A random variable $x : \Omega \to X$ on (Ω, F) and (X, G) is said to be *measurable* with respect to the sub-σ-algebra $H \subseteq F$ if for all $A \in G$, $x^{-1}(A) \in H$. If $x : \Omega \to X$ is a random variable and $f : X \to Y$ is a measurable function then $f(x) : \Omega \to Y$ is a random variable measurable with respect to F^x. Hence $F^{f(x)} \subseteq F^x$.

The concept of independence can now be easily extended from σ-algebras to random variables.

Definition 2.5.12. A finite or a countable family of random variables $\{x_n,\ n \in I\}$ is said to be *independent* with respect to P if the σ-algebra family $\{F^{x_n},\ n \in I\}$ generated by it, is independent with respect to P.

Proposition 2.5.13. *Let $x : \Omega \to \mathbb{R}$, $y : \Omega \to \mathbb{R}$ be two random variables. Assume that y is measurable with respect to F^x, or $y^{-1}(B(\mathbb{R})) \subseteq F^x$. Then there exists a Borel measurable function $h : \mathbb{R} \to \mathbb{R}$ such that $y = h(x)$.*

This result together with the paragraph above relating a random variable with a σ-algebra, establishes that any real-valued random variable y measurable with respect to F^x may be represented as $y = h(x)$, and conversely. This remark may help the reader to interpret the σ-algebra F^x.

Proposition 2.5.14. *Consider the random variable $x : \Omega \to \mathbb{R}^n$ and consider the Borel measurable function $f : \mathbb{R}^n \to \mathbb{R}^n$. Assume that* Range$(f) \in B(\mathbb{R}^n)$.

Then $F^{f(x)} = F^x$ if and only if f is an injective function.

It seems that the condition that Range$(f) \in B(\mathbb{R}^n)$ cannot be dispensed of. From [12, Exercise 10.9.3 and 4] follows that there exists a Lebesgue measurable subset $A \subseteq [0,1]$ and a continuous function $f : [0,1] \to \mathbb{R}$ such that $f(A)$ is not a Lebesgue measurable set. The proof makes use of the Canter set and the Cantor function. However, the statement above uses only Borel measurability.

Proof. Proof of Proposition 2.5.14. Define the random variable $y : \Omega \to \mathbb{R}^n$, $y = f(x)$. It follows from the statements earlier in this section that then $F^y = F^{f(x)} \subseteq F^x$. $(\Leftarrow)$ Define the range of the function f as the set,

$$\text{Range}(f) = \{f(w) \in \mathbb{R}^n |\, \forall\, w \in \mathbb{R}^n\}.$$

By assumption Range$(f) \in B(\mathbb{R}^n)$. If $v \in$ Range(f) then it follows from the definition of Range(f) that there exists a $w \in \mathbb{R}^n$ such that $v = f(w)$. Suppose there exist $w_1,\ w_2 \in \mathbb{R}^n$ such that $f(w_1) = f(w_2)$. Because the function f is injective, it follows that $w_1 = w_2$. Thus for any $v \in$ Range(f) there exists a unique $w \in \mathbb{R}^n$ such that $v = f(w)$. Therefore one may define the function $g :$ Range$(f) \to \mathbb{R}^n$, $g(v) = w$ if $v = f(w)$. This function is then well defined.

Note that for all $w \in \mathbb{R}^n$ with $v = f(w)$, hence $w = g(v)$, $g(f(w)) = g(v) = w$ so $g(f(.)) = i_{\mathbb{R}^n}$ is an identity function; and for all $v \in$ Range(f), $f(g(v)) = f(w) = v$, hence $f(g(.)) = i_{\text{Range}(f)}$ is also an identity function. Thus $g = f^{-1}$ and $f = g^{-1}$.

It has to be proven that the function g is a measurable function. Let $A \in B(\mathbb{R}^n)$. It has to be proven that, for all $A \in G$, $\{v \in \text{Range}(f) | \ g(v) \in A\} \in B(\mathbb{R}^n \cap \text{Range}(f))$. Note that the fact that f is a measurable function implies that, for all $A \in G$, $\{w \in \mathbb{R}^n | f(w) \in A\} \in B(\mathbb{R}^n)$. Note that,

$$\begin{aligned}
&\{v \in \text{Range}(f) | \ g(v) \in A\} \\
= &\{v \in \text{Range}(f) | \ f(g(v)) \in f(A)\}, \text{ because } f \text{ is injective,} \\
= &\{v \in \text{Range}(f) | \ v \in f(A)\} \in B(\mathbb{R}^n) \cap \text{Range}(f), \\
&\text{because } v = f(g(v)) \text{ and because } f \text{ is a measurable function.}
\end{aligned}$$

Thus g is a measurable function. Then $y = f(x)$ and $x = g(f(x)) = g(y)$. Hence $F^x = F^{g(y)} \subseteq F^y = F^{f(x)}$ and with the first paragraph of the proof it follows that $F^{f(x)} = F^x$.
$(\Rightarrow)$ By assumption, $F^y = F^{f(x)} = F^x$. From that relation and from Proposition 2.5.13 follows that there exists a measureable function $g : \mathbb{R}^n \to \mathbb{R}^n$ such that $x = g(y)$. Hence for all $\omega \in \Omega$, $x(\omega) = g(y(\omega)) = g(f(x(\omega)))$. Let $w_1, \ w_2 \in \mathbb{R}^n$ be such that $f(w_1) = f(w_2)$. This implies that $w_1 = g(f(w_1)) = g(f(w_2)) = w_2$. Thus f is an injective function. □

The following question has been posed. Let (Ω, F) be a measurable space and $G \subseteq F$ be a σ-algebra. Does there exists a random variable $x : \Omega \to \mathbb{R}$ such that $G = F(x)$? The answer to this question is negative, [3].

2.6 Expectation and the Characteristic Function

The concept of expectation of a random variable is so well known that the definition and the properties of the expectation operator have been removed from the book. The theory of expectation is based on integration theory. Retained has only been a the following proposition on inequalties using expectation.

Theorem 2.6.1. *Let $f : \mathbb{R} \to \mathbb{R}_+$ be a function that is symmetric, $f(u) = f(-u)$ for all $u \in R_+$, and strictly increasing on $(0, \infty)$ ($u < v \ \Rightarrow \ f(u) < f(v)$). Let $x : \Omega \to \mathbb{R}$ be a random variable such that $f(x)$ is integrable.*

(a)Then

$$P(|x| > u) \leq E[f(x)]/f(u), \ \ \forall u \in (0, \infty).$$

If $f(0) > 0$ then this inequality also holds for $u = 0$.

(b)The special case of (a) with f being the absolute value function to an integer power is known as the Markov inequality,

$$P(|x| \geq c) \leq c^{-k} E[|x|^k], \forall\, c \in (0, \infty), \forall\, k \in \mathbb{Z}_+.$$

The special case of the Markov inequality for $k = 2$ is known as the Chebyshev inequality.

Proof. (a) Because $f(x)$ is integrable and the function f is strictly increasing, it follows that $\infty > E[f(x)] \geq E[I_{(|x| \geq u)} f(x)] \geq f(u) P(|\, x \,| \geq u)$.
(b) Take for $k \in \mathbb{Z}_+$, $f(u) = |u|^k$. □

The Characteristic Function

The characteristic function of a real-valued random variable is the Fourier transform of its probability distribution function. Computations of the expectation of a random variable are in specific cases much simpler when use is made of the associated characteristic function.

Definition 2.6.2. Let $x : \Omega \to \mathbb{R}^n$ be a random variable. The *characteristic function* of x is defined as the function,

$$c_x : \mathbb{R}^n \to \mathbb{C}, \;\; c_x(w) = E[\exp(i w^T\, x)].$$

The above expression is well defined because $|\, \exp(i w^T\, x) \,| = 1$. There is a bijective correspondence between characteristic functions and pdf's. Examples of characteristic functions are provided below for several probability distribution functions.

There follows a useful relation between a characteristic function and moments of the associated random variable.

Proposition 2.6.3. *Consider a probability space (Ω, F, P) and a real-valued random variable $x : \Omega \to \mathbb{R}$.*

(a)Assume that the random variable x is integrable, hence $E|x| < \infty$. Then,

$$D_w E[\exp(i w\, x)] = E[\exp(i w\, x)\; i\, x], \; \forall w \in \mathbb{R}, \; where \; D_w = \frac{d}{dw};$$
$$D_w E[\exp(i w\, x)]|_{w=0} = i\, E[x].$$

(b)Assume that there exists an integer $m \in \mathbb{Z}_+$ such that $E|x|^m < \infty$. Then for all $k \in \mathbb{Z}_m$ there holds,

$$\frac{d^k}{dw^k} E[\exp(i w\, x)]|_{w=0} = E[(i\, x)^k].$$

Proof. (a) Consider $w \in \mathbb{R}$ and a sequence $\{w_n \in \mathbb{R},\, n \in \mathbb{Z}_+\}$ satisfying $\lim_{n \to \infty}\, w_n = 0$. Then

$$\begin{aligned}
D_w E[\exp(iw\,x)] &= \lim_{n\to\infty} \frac{E[\exp(i(w+w_n)\,x)] - E[\exp(iw\,x)]}{(w+w_n)-w} \\
&= \lim E[\exp(iw\,x)\frac{\exp(iw_n\,x)-1}{w_n}].
\end{aligned}$$

The dominated convergence theorem will be applied. Note that

$$\left|\exp(iw\,x)\frac{\exp(iw_n\,x)-1}{w_n}\right| = \left|\int_0^{w_n}\frac{i\,x}{w_n}\exp(iv\,x)dv\right| \leq \int_0^{|w_n|}\frac{|i\,x|}{|w_n|}dv = |x|.$$

By assumption $E|x| < \infty$. The dominated convergence theorem can then be applied. There follows,

$$\begin{aligned}
D_w E[\exp(iw\,x)] &= \lim_{n\to\infty} E[\exp(iw\,x)\frac{\exp(iw_n\,x)-1}{w_n}] \\
&= E[\exp(iw\,x) \lim_{n\to\infty} \frac{\exp(iw_n\,x)-1}{w_n}] \\
&= E[\exp(iw\,x)\; D_w \exp(iw\,x)|_{w=0}] = E[\exp(iw\,x)\; i\,x].
\end{aligned}$$

(b) This results follows by induction of which the steps are clear from Part (a). □

Expectations of Several Probability Distributions

The proofs of the subsequent propositions are simple and are therefore omitted. The reader can find a reference to proofs in the section *Further Reading* of this chapter. The proof for a random variable with the Gamma probability distribution is provided to show the calculations involved.

Proposition 2.6.4. *Consider a random variable* $x : \Omega \to \mathbb{N}_n = \{0,1,\ldots,n\}$ *having a Binomial pdf function with the parameters* $n \in \mathbb{Z}_+$ *and* $r \in (0,1)$*, see Def. 2.1.4. Then the mean, the variance, and the characteristic function are equal to,*

$$\begin{aligned}
m_x &= E[x] = nr, \;\; q_x = E[(x-E[x])(x-E[x])^T] = nr(1-r), \\
c_x(w) &= E[\exp(iw\,x)] = (1-r+r\exp(iw))^n.
\end{aligned}$$

The special case of the Binomial pdf for $n = 1$ *is called the Bernoulli pdf. The expression for the variance* q_x *is easily derived by first calculating* $E[x(x-1)]$.

Proposition 2.6.5. *Consider a random variable* $x : \Omega \to \mathbb{N} = \{0,1,\ldots\}$ *having a Poisson pdf function with the parameter* $\lambda \in (0,\infty)$*, see Def. 2.1.5. Then the mean, the variance, and the characteristic function are equal to,*

$$\begin{aligned}
m_x &= E[x] = \lambda, \;\; q_x = E[(x-E[x])^2] = \lambda, \\
c_x(w) &= E[\exp(iw\,x)] = \exp(\lambda(\exp(iw)-1)).
\end{aligned}$$

The expression for the variance q_x *is easily derived by first calculating* $E[x(x-1)]$.

Proposition 2.6.6. *Consider a random variable $x : \Omega \to [0,1]$ having a Beta pdf function with the parameters $(\beta_1, \beta_2) \in (0,\infty)$, see Def. 2.1.7. Then the mean, the variance, and the characteristic function are equal to,*

$$\begin{aligned} m_x &= E[x] = \frac{\beta_1}{\beta_1+\beta_2}, \\ q_x &= E[(x-E[x])(x-E[x])^T] = \frac{\beta_1\beta_2}{(\beta_1+\beta_2)^2(\beta_1+\beta_2+1)}, \\ c_x(w_x) &= E[\exp(iw_x\, x)] = \frac{\Gamma(\beta_1+\beta_2)}{\Gamma(\beta_1)} \sum_{k=0}^{\infty} \frac{\Gamma(\beta_1+k)}{\Gamma(\beta_1+\beta_2+k)} \frac{(iw_x)^k}{k!}. \end{aligned}$$

Proposition 2.6.7. *Consider a random variable $x : \Omega \to \mathbb{R}_+$ having a Gamma pdf function with the parameters $(\gamma_1, \gamma_2) \in (0,\infty)$, see Def. 2.1.8. Then the mean, the variance, and the characteristic function are equal to,*

$$\begin{aligned} m_x &= E[x] = \gamma_1\ \gamma_2,\ \ q_x = E[(x-E[x])(x-E[x])^T] = \gamma_1\ \gamma_2^2, \\ c_x(w) &= E[\exp(iwx)] = (1-iw\ \gamma_2)^{-\gamma_1}. \end{aligned}$$

Proof. Note the calculations, using equation (2.5),

$$\begin{aligned} & E[\exp(iwx)] = \int_0^\infty \exp(iwv)\ v^{\gamma_1-1}\exp(-v/\gamma_2)dv\ \gamma_2^{-\gamma_1}/\Gamma(\gamma_1) \\ &= \int v^{\gamma_1-1}\exp(-v[-iw+\frac{1}{\gamma_2}])dv\ \gamma_2^{-\gamma_1}/\Gamma(\gamma_1) \\ &= (-iw+1/\gamma_2)^{-\gamma_1}\ \gamma_2^{-\gamma_1} = (1-iw\ \gamma_2)^{-\gamma_1}. \end{aligned}$$

Analogously,

$$\begin{aligned} E[x] &= \int_0^\infty v\ v^{\gamma_1-1}\exp(-v/\gamma_2)dv\gamma_2^{-\gamma_1}/\Gamma(\gamma_1) \\ &= \int_0^\infty v^{(\gamma_1+1)-1}\exp(-v/\gamma_2)dv\ \gamma_2^{-\gamma_1}/\Gamma(\gamma_1) = \frac{\Gamma(\gamma_1+1)}{\Gamma(\gamma_1)}\gamma_2 = \gamma_1\gamma_2; \\ E[x^2] &= \gamma_1(\gamma_1+1)\gamma_2^2,\ \ q_x = E[(x-E[x])^2] = E[x^2]-E[x]^2 = \gamma_1\gamma_2^2. \end{aligned}$$

□

Proposition 2.6.8. *Consider a random variable $y : \Omega \to \mathbb{R}_+$ with a Chi-Square probability distribution function with $\nu \in \mathbb{R}_{s+}$ degrees of freedom, see Def. 2.1.9. Then the mean, the variance, and the characteristic function are equal to,*

$$m_y = \nu,\ q_y = E[(y-m_y)^2] = 2\nu,\ \ c_y(w) = (1-i2w)^{-\nu/2}.$$

From the latter formula follows directly that if y_1 has a Chi-Square pdf with ν_1 degrees of freedom and y_2 has a Chi-Square pdf with ν_2 degrees of freedom then y_1+y_2 has a Chi-Square pdf with $\nu_1+\nu_2$ degrees of freedom.

Proposition 2.6.9. *Consider a random variable $x : \Omega \to \mathbb{R}^n$ having a Gaussian pdf function with the parameters $n \in \mathbb{Z}_+$ and $(\overline{m}_x, \overline{Q}_x) \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_x \times n_x}_{spds}$ see Def. 2.1.10. Then the mean, the variance, and the characteristic function are equal to,*

$$\begin{aligned} m_x &= E[x] = \overline{m}_x,\ \ Q_x = E[(x-E[x])(x-E[x])^T] = \overline{Q}_x, \\ c_x(w) &= E[\exp(iw^T\ x)] = \exp(iw^T\ \overline{m}_x - \frac{1}{2}w^T\overline{Q}_x w). \end{aligned}$$

2.7 Gaussian Random Variables

Random variables with Gaussian probability distribution functions are frequently used as mathematical models for variables with uncertainty. A justification for this modelling approximation is the theoretical result known as the central limit theorem. This result states, based on several conditions, that the normalized sum of independent random variables, none of which is necessarily Gaussian distributed, converges to a random variable with a Gaussian distribution. Thus if the variable to be modelled is the sum of many independent random variables, each of which is not necessarily Gaussian distributed, then the probability distribution of this sum may be approximated by a Gaussian probability distribution function.

Definition 2.7.1. A random variable $x : \Omega \to \mathbb{R}^n$ is said to be a *Gaussian random variable* with the parameters $(m_x, Q_x) \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_x \times n_x}_{pds}$ called the *mean* m_x and the *variance* Q_x, if its characteristic function has the form,

$$\begin{aligned} E[\exp(iw_x^T x)] &= \exp(iw_x^T m_x - \frac{1}{2} w_x^T Q_x w_x), \ \ \forall w_x \in \mathbb{R}^{n_x}, \\ \mathbb{R}^{n_x \times n_x}_{pds} &= \{Q \in \mathbb{R}^{n_x \times n_x} | 0 \preceq Q = Q^T\}. \end{aligned}$$

The notation $\text{pdf}(x) \in G(m_x, Q_x)$ and $x \in G(m_x, Q_x)$ will be used in this case. Note that this includes the case where $Q_x = 0$, in which case $x = m_x$ almost surely.

A *standard Gaussian random variable* taking values in $\mathbb{R}^{n_x}$ for an integer $n_x \in \mathbb{Z}_+$, is defined as a Gaussian random variable with $m_x = 0$ and $Q_x = I_{n_x}$.

Moreover, $\text{pdf}(x_1, ..., x_m) \in G(m, Q)$ denotes that, with $x^T = (x_1, ..., x_m)^T$, $x \in G(m, Q)$. In this case $(x_1, ..., x_m)$ will be said to be *jointly Gaussian* random variables.

One can prove that a Gaussian random variable with $Q > 0$ has a Gaussian distribution function that admits a density function as defined in Section 2.1, and, conversely, that a random variable with a Gaussian density function is a Gaussian random variable. The argument in a proof of this statement is based on the fact that there is a bijective correspondence between characteristic functions and probability distribution functions.

Properties of Gaussian random variables follow.

Proposition 2.7.2. *If $x : \Omega \to \mathbb{R}^{n_x}$, $x \in G(m_x, Q_x)$, $A \in \mathbb{R}^{n_y \times n_x}$, $b \in \mathbb{R}^{n_y}$, and $y = A\,x + b$ then $(Ax + b) \in G(Am_x + b, AQ_x A^T)$.*

Proof.

$$\begin{aligned} E[\exp(iw^T \ (Ax + b))] &= E[\exp(i(A^T w)^T x)] \exp(iw^T b) \\ &= \exp(iw^T (Am_x + b) - \frac{1}{2} w^T A Q_x A^T w), \ \forall\, w \in \mathbb{R}^{n_x}. \end{aligned}$$

□

Proposition 2.7.3. *Let $x : \Omega \rightarrow \mathbb{R}^{n_x}$ and $y : \Omega \rightarrow \mathbb{R}^{n_y}$ be jointly Gaussian random variables with,*

$$\text{pdf}(x,y) \in G\left(m, \begin{pmatrix} Q_x & Q_{xy} \\ Q_{xy}^T & Q_y \end{pmatrix}\right).$$

Then F^x, F^y are independent if and only if $Q_{xy} = 0$; or, in words, x, y are independent random variables if and only if they are uncorrelated.

Proof. Using properties of expectation it has been proven that two random variables are independent if and only if their joint characteristic function factorizes as the product of the individual characteristic function. Thus, $x, y \in G$ are independent if and only if,

$$\begin{aligned}
& E[\exp(iw_x^T\, x + iw_y^T\, y)] = E[\exp(iw_x^T\, x)]E[\exp(iw_y^T\, y)], \\
& \forall\ (w_x, w_y) \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_y}; \\
& E[\exp(iw_x^T\, x + iw_y^T\, y)] \\
& = \exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} m_x \\ m_y \end{pmatrix} - \frac{1}{2}\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} Q_x & Q_{xy} \\ Q_{xy}^T & Q_y \end{pmatrix}\begin{pmatrix} w_x \\ w_y \end{pmatrix}\right), \\
& = \exp\left(iw_x^T\, m_x - \frac{1}{2}w_x^T Q_x w_x + iw_y^T\, m_y - \frac{1}{2}w_y^T Q_y w_y\right) \times \exp(-w_x^T Q_{xy} w_y) \\
& = E[\exp(iw_x^T\, x)]E[\exp(iw_y^T\, y)] \times \exp(-w_x^T Q_{xy} w_y).
\end{aligned}$$

Thus the required equality holds iff $w_x^T Q_{xy} w_y = 0$ for all $(w_x, w_y) \in (\mathbb{R}^{n_x} \times \mathbb{R}^{n_y})$ if and only if $Q_{xy} = 0$. □

Proposition 2.7.4. *Let $x : \Omega \rightarrow \mathbb{R}^{n_x}$, $y : \Omega \rightarrow \mathbb{R}^{n_y}$ each be a Gaussian random variable. Assume that they are independent. Let $z : \Omega \rightarrow \mathbb{R}^{n_x+n_y}$, $z = (x^T, y^T)^T$. Then z is a Gaussian random variable.*

The elementary proof is omitted.

Proposition 2.7.5. *Consider a Gaussian random variable $x : \Omega \rightarrow \mathbb{R}^{n_x}$ with $n_x \in \mathbb{Z}_+$, $x \in G(0, Q_x)$, $Q_x \in \mathbb{R}_{pds}^{n \times n}$, $Q_x \neq 0$, $n_v = \text{rank}(Q_x)$, hence $1 \leq n_v \leq n_x$.*

There exists a standard Gaussian random variable $v : \Omega \rightarrow \mathbb{R}^{n_v}$ and a matrix $M \in \mathbb{R}^{n_x \times n_v}$ such that,

$$x = Mv,\ a.s.,\ \ v \in G(0, I_{n_v}),\ (x,\ v) \in G,\ \text{rank}(M) = n_v,\ Q_x = MM^T.$$

Because $v \in G(0, I)$, the components of v are independent random variables.

Proof. Because $Q_x = Q_x^T \succeq 0$, there exists a singular value decomposition of Q_x of the form, $Q_x = U\, D\, U^T$ in which $U \in \mathbb{R}^{n_x \times n_x}$ is an orthogonal matrix, thus satisfying $UU^T = I = U^T U$, and $D \in \mathbb{R}^{n_x \times n_x}$ is a diagonal matrix such that for all $i \in \mathbb{Z}_{n_x}$, $D_{i,i} \geq 0$. Because $\text{rank}(Q_x) = n_v \leq n_x$, there exists in general a decomposition of the form,

$$\begin{aligned}
D &= \begin{pmatrix} D_1 & 0 \\ 0 & 0 \end{pmatrix}, \ D_1 \in \mathbb{R}^{n_v \times n_v}, \ \forall\, i = 1,\ldots,n_v, \ D_{1,i,i} > 0, \\
U &= \begin{pmatrix} U_1 & U_2 \end{pmatrix}, \ U_1 \in \mathbb{R}^{n_x \times n_v}, \ U_2 \in \mathbb{R}^{n_x \times (n_x - n_v)}, \ U \text{ orthogonal}, \\
Q_x &= UDU^T = U_1 D_1 U_1^T, \\
M &= U_1 D_1^{1/2} \in \mathbb{R}^{n_x \times n_v}, \ \ MM^T = U_1 D_1 U_1^T = Q_x, \ \text{rank}(M) = n_v.
\end{aligned}$$

Define the random variable $v = Lx = D_1^{-1/2} U_1^T x$, $v : \Omega \to \mathbb{R}^{n_v}$. It follows from Proposition 2.7.2 that $(x,\ v)$ are jointly Gaussian random variables. Then note that,

$$\begin{aligned}
Q_v &= E[vv^T] = LQ_xL^T = D_1^{-1/2} U_1^T U_1 D_1 U_1^T U_1 D_1^{-1/2} = I, \ \ v \in G(0, I_{n_v}), \\
& E[(x - Mv)(x - Mv)^T] = Q_x - Q_{x,v} M^T - MQ_{x,v}^T + MQ_v M^T \\
&= Q_x - Q_x U_1 D_1^{-1/2} D_1^{1/2} U_1^T - Q_x + U_1 D_1 U_1^T = 0, \ \Rightarrow \\
x &= M\, v, \ a.s.
\end{aligned}$$

□

Proposition 2.7.6. *Let* $y : \Omega \to \mathbb{R}^{n_y}$*,* $\text{pdf}(y) \in G(m_y, Q_y)$*, and* $Q_c \in \mathbb{R}_{pds}^{n_y \times n_y}$*. Then* $E[y^T Q_c y] = \text{tr}(Q_c Q_y) + m_y^T Q_c m_y$*.*

Proof.

$$\begin{aligned}
E[y^T Q_c y] &= E[(y - m_y)^T Q_c (y - m_y)] + 2E[y^T Q_c m_y] - E[m_y^T Q_c m_y] \\
&= E[\sum_{i=1}^{n_y} \sum_{j=1}^{n_y} Q_{c,i,j} (y_i - m_{y,i})(y_j - m_{y,j})] + m_y^T Q_c m_y \\
&= \left(\sum_{j=1}^{n_y} \sum_{i=1}^{n_y} Q_{c,j,i} Q_{y,i,j} \right) + m_y^T Q_c m_y = \text{tr}(Q_c Q_y) + m_y^T Q_c m_y.
\end{aligned}$$

□

Proposition 2.7.7. *Consider two jointly Gaussian random variables* $x : \Omega \to \mathbb{R}^{n_x}$ *and* $y : \Omega \to \mathbb{R}^{n_y}$ *with* $(x, y) \in G$ *and* $\text{rank}(Q_x) = n_x$*. Assume that there is exists a matrix* $L \in \mathbb{R}^{n_y \times n_x}$ *such that* $y = Lx$*. Then* $F^y = F^x$ *if and only if* $\text{rank}(L) = n_x = \text{rank}(Q_x)$*. It follows from the assumption that* (x, y) *are jointly Gaussian random variables and from Theorem 2.8.3 that* $L = Q_{y,x} Q_x^{-1}$*.*

Proof. Note that $y = Lx$ hence y is a linear function of x. The linear map L is injective if and only if $n_x \leq n_y$ and $\text{rank}(L) = n_x$ by Proposition 17.4.2. The result then follows from Proposition 2.5.14. □

Example 2.7.8. Consider a probability space (Ω, F, P) and a Gaussian random variable $x : \Omega \to \mathbb{R}^n$ with $x \in G(0, Q_x)$.

If $Q_x \succ 0$ then $X_{supp} = X$ hence x has full support on the state set $X = \mathbb{R}^n$.

If $Q_x \succeq 0$ then it follows from Proposition 17.4.30 that there exists a linear transformation $L \in \mathbb{R}^{n \times n}$ such that,

$$LQ_xL^T = \begin{pmatrix} Q_1 & 0 \\ 0 & 0 \end{pmatrix},\; Q_1 \in \mathbb{R}^{n_1 \times n_1}_{spds},\; Q_1 \succ 0,\; n_1 \in \mathbb{N},\; n_1 < n.$$

Then $X_{supp} = \mathbb{R}^{n_1} \oplus I_{n-n_1} \subsetneq X = \mathbb{R}^n$ up to the linear transformation. Hence x does not have full support on the state set X.

2.8 Conditional Expectation

In this book a major operation is to take conditional expectation of a random variable conditioned on a σ-algebra. This concept is defined in this chapter because of its use in subsequent chapters. There follows the defintion of conditional expectation and the two foremost applications to Gaussian random variables and to finite-valued random variables.

Theorem 2.8.1. *Consider a measurable space (Ω, F) and a sub-σ-algebra $G \subseteq F$.*

*(a)*Conditional expectation of a positive-valued random variable. *Consider a positive random variable $x : \Omega \to \mathbb{R}_+$ with $E[x] < \infty$. Then there exists a positive random variable $E[x \mid G] : \Omega \to \mathbb{R}_+$ that is G measurable, such that,*

$$\forall\, A \in G,\; E[x\, I_A] = E[E[x|G]\, I_A]. \tag{2.6}$$

A random variable satisfying these conditions is unique up to an almost sure modification. Uniqueness up to an almost sure modification means that if $y : \Omega \to \mathbb{R}_+$ is also G measurable and such that $E[xI_A] = E[yI_A]$ for all $A \in G$, then $E[x|G] = y$ a.s.
Note that $E[E[x|G]] = E[x] < \infty$ by the definition of the conditional expectation with $A = \Omega$, hence $E[x|G] \in L_1$.
The random variable $E[x|G]$ will be called the conditional expectation *of the positive random variable x* given *G or* conditioned on *G.*

*(b)*Conditional expectation of an integrable random variable. *If $x : \Omega \to \mathbb{R}$ satisfies $E|x| < \infty$ then there exists a random variable $E[x|G] : \Omega \to \mathbb{R}_+$ that is G-measurable, such that,*

$$\begin{aligned} x = x^+ - x^-,&\;\; E[x^+] + E[x^-] = E[x^+ + x^-] = E|x| < \infty, \\ &\Rightarrow\;\; E[x^+] < \infty,\; E[x^-] < \infty \\ &\Rightarrow^{by\ (a)}\;\; E[x^+|G] < \infty,\; E[x^-|G] < \infty, \\ E[x|G] &= E[x^+|G] - E[x^-|G],\; \text{by (a) is well defined.} \\ &\text{The function } E[.\mid G] : \{x : \Omega \to \mathbb{R} | x \text{ a random variable},\; x \in L_1\} \\ &\mapsto \{y : \Omega \to \mathbb{R} | y \text{ is a } G\text{-measurable random variable}\}, \end{aligned}$$

will be called the conditional expectation operator *of conditioning on G.*
If $y : \Omega \to \mathbb{R}$ satisfies that (1) y is G-measurable and (2) for all $A \in G$, $E[I_A\, x] = E[I_A\, y]$, then $E[x|G] = y$ a.s.

For an integrable random variable $x : \Omega \rightarrow \mathbb{R}^{n_x}$ one defines its conditional expectation componentwise.

(c)For $x : \Omega \rightarrow \mathbb{R}^{n_x}$ and $G \subseteq F$ a sub-σ-algebra define the conditional characteristic function *of x conditioned on G as,*

$$c_x : \Omega \times \mathbb{R}^{n_x} \rightarrow \mathbb{C}, \;\; c_x(\omega, w|G) = E[\exp(iw^T\, x) \mid G];$$
$$\textit{notation, } c_x(w|G) = c_x(\omega, w|G).$$

Proof. Below use is made of the concept of absolute continuity of probability measures and the Radon-Nikodym derivative, see Section 19.9.

(a) Let $x : \Omega \rightarrow \mathbb{R}_+$. Define $P_x : F \rightarrow \mathbb{R}_+$, $P_x(A) = E[xI_A]$. Then P_x is a positive measure on (Ω, F). Let $P_x^G : G \rightarrow \mathbb{R}_+,\; P_x^G(A) = P_x(A)$ be the restriction of P_x from F to G. Then P_x^G is absolute continuous with respect to P_x and by the Radon-Nikodym Theorem 19.9.4, there exists a random variable $E[x \mid G] : \Omega \rightarrow \mathbb{R}_+$ that is G measurable and is such that for all $A \in G$,

$$E[xI_A] = P_x(A) = P_x^G(A) = E[E[x \mid G]I_A].$$

As indicated below, if $y : \Omega \rightarrow \mathbb{R}_+$ is another random variable that is G measurable and such that (2.6) holds, then from $E[E[x \mid G]I_A] = E[xI_A] = E[yI_A]$ follows that $y = E[x|G]$ a.s.

(b) The construction of the conditional expectation is described in the theorem statement and it directly obvious that the concept is well defined. Note that then,

$$E[I_A E[x|G]] = E[I_A x] = E[I_A y]$$
$$\Rightarrow E[I_A(y - E[x|G])] = 0, \; \forall \, A \in G \;\; \Rightarrow \;\; y - E[x|G] = 0 \; a.s..$$

(c) The expression is well defined because $|\exp(iwx)| \leq 1$ hence $E|\exp(iwx)| < 1$ and the result follows from (b). □

An alternative construction of the existence of the conditional expectation operator may be given starting with the geometric projection operation for a square-integrable random variable such that $E[x|G] \in L_2$ and then extending this operation by a limit argument from L_2 to L_1; see [37].

Properties of Conditional Expectation

Theorem 2.8.2. *Let $x, y : \Omega \rightarrow \mathbb{R}$ be integrable random variables and G, G_1, $G_2 \subseteq F$ be sub-σ-algebras of F. Then the conditional expectation operator has the following properties:*

*(a)*Linearity:

$$E[x + y|G] = E[x|G] + E[y|G], \;\; E[cx|G] = cE[x|G], \; \forall c \in \mathbb{R}.$$

*(b)*Order preservation: *if $x \leq y$ a.s. then $E[x|G] \leq E[y|G]$ a.s.*

*(c)*Measurability: *if y is G measurable, and if $xy \in L_1$ then $E[xy|G] = yE[x|G]$. In particular, $E[y|G] = y$.*

*(d)*Reconditioning: *if* $G_1 \subseteq G_2$ *then* $E[x|G_1] = E[E[x|G_2]|G_1]$. *In particular, with* $G_1 = \{\Omega, \emptyset\}$, $E[E[x|G]] = E[x]$.
*(e)*Independence: *if* F^x *and* G *are independent sub-*σ*-algebras with respect to* P *and if* $f : \mathbb{R} \to \mathbb{R}$ *is a Borel measurable function such that* $f(x) \in L_1$, *then* $E[f(x)|G] = E[f(x)]$.
(f) F^x, G *are independent if and only if, for all* $w \in \mathbb{R}$, $E[exp(iw\,x)|G] = E[exp(iw\,x)]$.

If $x, y : \Omega \to \mathbb{R}^n$ *then the results above hold for each component of* x *and* y.

Proof. (a) Let $A \in G$. Then

$$\begin{aligned} E|x+y| \le E|x| + E|y| &< \infty, \text{ because } x, y \text{ are integrable,} \\ E[I_A E[x+y|G]] &= E[I_A(x+y)] = E[I_A x] + E[I_A y] \\ &= E[I_A E[x|G] + E[I_A E[y|G]] \\ &= E[I_A(E[x|G] + E[y|G])], \text{ thus,} \\ E[x+y|G] &= E[x|G] + E[y|G], \text{ a.s.,} \end{aligned}$$

by the uniqueness of $E[x+y|G]$ up to a.s. modifications. The proof that $E[cx|G] = cE[x|G]$ for any $c \in \mathbb{R}$ is similar.
(b) Because x and y are integrable, the conditional expectations $E[x|G]$ and $E[y|G]$ are well defined. Suppose that there exists a set $A \in G$ with $P(A) > 0$ and $I_A(E[x|G] - E[y|G]) > 0$. Then,

$$\begin{aligned} 0 &< E[I_A(E[x|G] - E[y|G])] = E[I_A E[x|G]] - E[I_A E[y|G]] \\ &= E[I_A x] - E[I_A y] = E[I_A(x-y)] \le 0, \end{aligned}$$

which is a contradiction. Thus $E[x|G] \le E[y|G]$ a.s.
(c) Consider first the special case when $y = I_C$ for a set $C \in G$. Thus, for $A \in G$,

$$\begin{aligned} E[I_A E[xy|G]] &= E[I_A I_C x] = E[I_{A\cap C} E[x|G]], \text{ because } A \cap C \in G, \\ &= E[I_A y E[x|G]], \text{ hence,} \\ E[xI_C|G] &= I_C E[x|G]. \end{aligned}$$

From (a) follows then that for $\{C_i \in G, i \in \mathbb{Z}_n\} \subset G$ and $\{c_i \in \mathbb{R}, i \in \mathbb{Z}_n\}$,

$$E[x \sum_{i=1}^{n} c_i I_{C_i} |G] = \sum_{i=1}^{n} c_i I_{C_i} E[x|G].$$

An application of the monotone class Theorem 19.1.3 then yields the result.
(d) Let $A \in G_1 \subseteq G_2$. Then

$$\begin{aligned} & E[I_A E[E[x|G_2]|G_1] \\ &= E[I_A E[x|G_2]], \text{ because } A \in G_1, \\ &= E[I_A x], \text{ because } A \in G_1 \subseteq G_2, \\ &= E[I_A E[x|G_1]], \text{ and from uniqueness upto modification,} \\ E[x|G_1] &= E[E[x|G_2]|G_1]. \end{aligned}$$

(e) Let F^x and G be independent σ-algebras and $A \in G$. Then

$$\begin{aligned} E[I_A E[f(x)|G]] &= E[I_A f(x)] = E[I_A]E[f(x)], \text{ by independence of } F^x,\ G, \\ &= E[I_A E[f(x)]], \text{ hence} \\ E[f(x)|G] &= E[f(x)] \text{ a.s.} \end{aligned}$$

because $E[f(x)]$ is G measurable and by the uniqueness a.s.
(f) Omitted. □

Special Cases of Conditional Expectation

In several cases it is possible to present explicit formula's for the conditional expectation of random variables. See also Chapter 9 for additional examples.

Theorem 2.8.3. *Let* $x : \Omega \to \mathbb{R}^{n_x}, y : \Omega \to \mathbb{R}^{n_y}$,

$$(x,\ y) \in G\left(\begin{pmatrix} m_x \\ m_y \end{pmatrix}, \begin{pmatrix} Q_x & Q_{xy} \\ Q_{xy}^T & Q_y \end{pmatrix} \right).$$

Assume that $0 \prec Q_y$. *For the case* $0 \preceq Q_y$ *see [30]. Then,*

(a) $E[x \mid F^y] = m_x + Q_{xy} Q_y^{-1}(y - m_y)$.

(b) $E[(x - E[x \mid F^y])(x - E[x \mid F^y])^T \mid F^y]$
$= E[(x - E[x \mid F^y])(x - E[x \mid F^y])^T] = Q_x - Q_{xy} Q_y^{-1} Q_{xy}^T = Q_{x|y} \in \mathbb{R}^{n_x \times n_x}$.

(c) $E[\exp(iw^T x)|F^y] = \exp(iw^T E[x|F^y] - \frac{1}{2} w^T Q_{x|y} w)$, *for all* $w \in \mathbb{R}^{n_x}$.

(d) $E[\exp(iw^T E[x \mid F^y])]$
$= \exp(iw^T m_x - \frac{1}{2} w^T Q_{xy} Q_y^{-1} Q_{xy}^T w)$, *for all* $w \in \mathbb{R}^{n_x}$.

Proof. (a) Define the random variable,

$$r : \Omega \to \mathbb{R}^{n_x},\ r = m_x + Q_{xy} Q_y^{-1}(y - \mu_y).$$

Then,

$$\begin{pmatrix} x - r \\ y - m_y \end{pmatrix} = \begin{pmatrix} I\ -Q_{xy}Q_y^{-1} \\ 0\ I \end{pmatrix} \begin{pmatrix} x - m_x \\ y - m_y \end{pmatrix},$$

and the assumption that $x,\ y$ are jointly Gaussian random variables imply by Proposition 2.7.2 that $(x - r,\ y - m_y)$ are jointly Gaussian random variables. It will be proven that $E[x|F^y] = r$.

Because $x - r$ and $y - m_y$ are jointly Gaussian,

$$\begin{aligned} E[(x - r)(y - m_y)^T] &= E[(x - m_x - Q_{xy} Q_y^{-1}(y - m_y))(y - m_y)^T] \\ &= Q_{xy} - Q_{xy} Q_y^{-1} Q_y = 0. \end{aligned}$$

From Proposition 2.7.3 follows that $(x-r)$ and $(y-m_y)$ are independent random variables. From Theorem 2.8.2.(e) and (c) follows that

$$\begin{aligned} & E[x-r|F^y] = E[x-r] = E[(x-m_x) - Q_{xy}Q_y^{-1}(y-m_y)] = 0 \\ \Rightarrow\ & E[x|F^y] = E[r|F^y] = E[m_x + Q_{xy}Q_y^{-1}(y-m_y)|F^y] \\ =\ & m_x + Q_{xy}Q_y^{-1}(y-m_y) = r. \end{aligned}$$

(b) From Theorem 2.8.2 and from the property that $(x-r)$ is independent of F^y follows that,

$$\begin{aligned} & E[(x-E[x|F^y])(x-E[x|F^y])^T|F^y] = E[(x-r)(x-r)^T] \\ =\ & Q_{xx} - Q_{xy}Q_y^{-1}Q_{yx} = Q_{x|y}. \end{aligned}$$

(c)

$$\begin{aligned} & E[\exp(iw^T(x-r))|F^y] \\ =\ & E[\exp(iw^T(x-r))],\ \text{ because } (x-r) \text{ is independent of } y, \\ =\ & \exp(-\frac{1}{2}w^T Q_{x|y}w),\ \ \text{by (b) and by } \mathrm{pdf}(x-r) \in G(0, Q_{x|y}), \\ \Rightarrow\ & E[\exp(iw^T x)|F^y] = \exp(iw^T E[x|F^y] - \frac{1}{2}w^T Q_{x|y}w),\ \forall\, w \in \mathbb{R}^{n_x}. \end{aligned}$$

(d) This follows directly from (a) and (b). □

Proposition 2.8.4. *Let x be a positive integrable random variable and y be a simple random variable with the indicator representation, see Def. 2.5.9,*

$$\begin{aligned} & x : \Omega \to \mathbb{R}_+,\ x \in L_1, \\ & y_k = I_{A_k},\ \forall\, k \in \mathbb{R}_{n_y},\ y : \Omega \to \mathbb{R}^{n_y},\ \{A_k \in F,\ k \in \mathbb{Z}_n\} \textit{ a finite partition of } \Omega, \\ & p_y = E[y] \in \mathbb{R}_{s+} \ \Leftrightarrow\ \forall\, k \in \mathbb{Z}_n,\ E[I_{A_k}] = P(A_k) > 0. \end{aligned}$$

(a)Then

$$E[x|F^y] = d^T y = \sum_{k=1}^{n} d_k I_{A_k},\ \forall\, k \in \mathbb{R}_+,\ d_k = E[xI_{A_k}]/E[I_{A_k}];\ d \in \mathbb{R}_+^{n_y}.$$

(b)If, in addition, the random variable x is also finite valued in the indicator representation, hence $x : \Omega \to \mathbb{R}_+^{n_x}$, then,

$$E[x|\ F^y] = Q_{x,y}\mathrm{Diag}(p_y)^{-1}y,\ \textit{ where } Q_{x,y} = E[xy^T].$$

The expression of the conditional expectation in (a) has been described as that conditional expectation is averaging the random variable x over the atoms of F^y.

Proof. (a) Note that the σ-algebra F^y is generated by the finite partition of the representation. Then,

$$E[x|F^y] = \sum_{k=1}^{n} d_k I_{A_k} \;\Leftrightarrow\; \forall \; A \in F^y, \;\; E[xI_A] = E[\sum_{k=1}^{n} d_k I_{A_k} I_A],$$

$$\Leftrightarrow \forall \, m \in \mathbb{Z}_n, \text{ and } \forall \, A_m, \; E[xI_{A_m}] = E[\sum_{k=1}^{n} d_k I_{A_k} I_{A_m}],$$

$$\Leftrightarrow \forall \, m \in \mathbb{Z}_n, \;\; E[xI_{A_m}] = \sum_{k=1}^{n} d_k E[I_{A_k \cap A_m}] = d_m E[I_{A_m}],$$

$$\Leftrightarrow \forall \, m \in \mathbb{Z}_n, \;\; d_m = E[xI_{A_m}]/E[I_{A_m}].$$

which is true because $\{A_k, k \in \mathbb{Z}_n\}$ is a partition of Ω and because of the definition of d_m.
(b) Note that,

$$Q_{x,y;i,j} = E[x_i y_j] = E[I_{\{x=e_i\}} I_{\{y=e_j\}}] = P(\{x = e_i\} \cap \{y = e_j\});$$
$$p_y = E[y] = (E[1_{n_x}^T x y^T])^T = (1^T Q_{x,y})^T.$$

The result then follows from (a). □

Example 2.8.5. Consider a tuple of finite-valued random variables both in the indicator representation. Then the conditional expectation equals,

$$x, \; y : \Omega \to \mathbb{R}^2, \;\; Q_{x,y} = \begin{pmatrix} 0.4 \; 0.2 \\ 0.1 \; 0.3 \end{pmatrix}, \;\; p_y = (1^T Q_{x,y})^T = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix};$$
$$E[x| \; F^y] = Q_{x,y} \text{Diag}(p_y)^{-1} y = \begin{pmatrix} 0.8 \; 0.4 \\ 0.2 \; 0.6 \end{pmatrix} y.$$

The next result is closely related to the concept of a conditional probability, see Section 19.6. The result will be used in Chapter 9.

Theorem 2.8.6. *Consider two random variables for $n_x, \; n_v \in \mathbb{Z}_+$, $x : \Omega \to \mathbb{R}^{n_x}$ and $v : \Omega \to \mathbb{R}^{n_v}$ and denote their joint probability distribution function by $f_{x,v} : \mathbb{R}^{n_x} \times \mathbb{R}^{n_v} \to \mathbb{R}$. Consider a measurable function $h : \mathbb{R}^{n_x} \times \mathbb{R}^{n_v} \to \mathbb{R}$. Assume that:*

1. *The σ-algebras F^x and F^v are independent. It then follows from the definition of independence of σ-algebras that the joint probability distribution function $f_{x,v}$ factorizes as $f_{x,v} = f_x \times f_v$.*
2. $$\int_{\mathbb{R}^{n_v}} |h(w_x, w_v)| f_v(dw_v) < \infty, \; \forall \, w_x \in \mathbb{R}^{n_x}; \textit{ and,}$$
   $$\int_{\mathbb{R}^{n_x}} \int_{\mathbb{R}^{n_v}} |h(w_x, w_v)| f_x(dw_x) f_v(dw_v) < \infty.$$

(a)Then

$$E[h(x,v)|F^x] = \int_{\mathbb{R}^{n_v}} h(x, w_v) \; f_v(dw_v) \; a.s. \tag{2.7}$$

(b)If in addition to the above conditions, the probability distribution functions f_x and f_v each admit a probability density function denoted by $p_x : \mathbb{R}^{n_x} \rightarrow \mathbb{R}_+$ and $p_v : \mathbb{R}^{n_v} \rightarrow \mathbb{R}_+$ respectively then,

$$E[h(x,v)|F^x] = \int_{\mathbb{R}^{n_v}} h(x,w_v)\ p_v(w_v)dw_v\ a.s. \tag{2.8}$$

Proof. (a) By assumption (2),

$$\begin{aligned}
&\int_{\mathbb{R}^{n_v}} |h(w_x,w_v)|\ f_v(dw_v) < \infty,\ \forall\ w_x \in \mathbb{R}^{n_x}, \\
\Rightarrow &\int_{\mathbb{R}^{n_v}} |h(x,w_v)|\ f_v(dw_v) < \infty\ a.s., \\
&\int_{\mathbb{R}^{n_x}} \left[\int_{\mathbb{R}^{n_v}} |h(w_x,w_v)|\ f_v(dw_v) \right] f_x(dw_x) \\
= &\int_{\mathbb{R}^{n_x} \times \mathbb{R}^{n_v}} |h(w_x,w_v)|\ f_{x,v}(dw_x,dw_v) < \infty.
\end{aligned}$$

Thus the right-hand side of equation (2.7) is well defined. Note that $\forall\ A \in F^x$,

$$\begin{aligned}
&E\left[I_A \int_{\mathbb{R}^{n_v}} h(x,w_v)\ f_v(dw_v) \right] \\
= &\int_{\mathbb{R}^{n_x}} \int_{\mathbb{R}^{n_v}} I_A(w_x) h(w_x,w_v) f_x(dw_x) f_v(dw_v) \\
= &\int_{\mathbb{R}^{n_x} \times \mathbb{R}^{n_v}} I_A(w_x)\ h(w_x,w_v)\ f_{x,v}(dw_v,dw_v) = E[I_A\ h(x,v)],
\end{aligned}$$

and the statement (a) follows from Theorem 2.8.1.
(b) This follows directly from (a) using the facts that, in case the probability densities exist, then $f_x(dw_x) = p_x(w_x)dw_x$ and $f_v(dw_v) = p_v(w_v)dw_v$. □

2.9 Conditional Independence

In stochastic system theory, the main concept is conditional independence of σ-algebras. Below the concept is introduced and illustrated by the case of conditional independence of Gaussian random variables. In Section 19.8 the readers finds additional results for the conditional independence relation.

The reader is reminded that two σ-algebras F_1 and F_2 are independent if,

$$E[x_1 x_2] = E[x_1]\ E[x_2],$$

where the random variable $x_1 : \Omega \rightarrow \mathbb{R}_+$ is F_1-measurable which is denoted by $x_1 \in L((\Omega,F_1),(\mathbb{R}_+,B(\mathbb{R}_+)))$, and similarly $x_2 \in L((\Omega,F_2),(\mathbb{R}_+,B(\mathbb{R}_+)))$.

Definition 2.9.1. Let $F_{1,}$, F_2, and G be sub-σ-algebras of (Ω,F,P). One calls F_1 and F_2 *conditionally independent* given G or conditioned on G, or one says that G makes F_1 and F_2 *conditional independent*, if

$$\forall\, x_1 \in L((\Omega,F_1),(\mathbb{R}_+,B(\mathbb{R}_+))),\ \forall\, x_2 \in L((\Omega,F_2),(\mathbb{R}_+,B(\mathbb{R}_+)),$$
$$E[x_1x_2|G] = E[x_1|G]\ E[x_2|G]. \tag{2.9}$$

The notation $(F_1,F_2|G) \in CI$ will be used to denote that $F_1,\ F_2$ are conditionally independent given G. If the random variables x_1 and x_2 are both integrable and such that $x_1x_2 \in L_1$ then equation (2.9) can be extended from positive random variables to integrable random variables with the same formula.

Conditional independence is like independence but formulated in terms of conditional expectations. Below equivalent conditions for conditional independence are presented.

Proposition 2.9.2. *Given the sub-σ-algebras $F_{1,}\ F_2,\ G$ on Ω. The following statements are equivalent:*

(a)$(F_1,F_2|G) \in CI$;
(b)$(F_2,F_1|G) \in CI$;
(c)for all $x_1 \in L((\Omega,F_1),(\mathbb{R}_+,B(\mathbb{R}_+)))$, $E[x_1|F_2 \vee G] \ = \ E[x_1|G]$;
(d)for all $x_1 \in L((\Omega,F_1),(\mathbb{R}_+,B(\mathbb{R}_+)))$, $E[x_1|F_2 \vee G]$ is G-measurable.

Proof. (a) ⇔ (b) This equivalence follows directly from the symmetry of the definition of the conditional independence relation with respect to the subindices 1 and 2.
(a) ⇒ (c) Let $x_1 \in L(\Omega,\mathbb{R}_+),\ A_2 \in F_2$ and $A_3 \in G$. Then

$$\begin{aligned}
&E[I_{A_2}I_{A_3}E[x_1|F_2 \vee G]]\\
&= E[I_{A_2}I_{A_3}x_1] = E[I_{A_3}E[x_1I_{A_2}|G]],\\
&\quad \text{by reconditioning on } G \text{ and using that } A_3 \in G,\\
&= E[I_{A_3}E[x_1|G]E[I_{A_2}|G]], \text{by CI,,}\\
&= E[I_{A_2}I_{A_3}E[x_1|G]],\ \text{by conditional expectation of } E[I_{A_2}|G]\\
&\quad \text{using that } I_{A_3}E[x_1|G] \text{ is } G\text{-measureable.}
\end{aligned}$$

By using the monotone class theorem one then shows that for all
$x_3 \in L((\Omega,F_2 \vee G),(\mathbb{R}_+,B(\mathbb{R}_+)))$, $E[x_3E[x_1|F_2 \vee G]] = E[x_3E[x_1|G]]$ from which follows that $E[x_1|F_2 \vee G] \ = \ E[x_1|G]$.
(c) ⇒ (a) $E[x_1x_2|G] = E[x_2E[x_1|F_2 \vee G]|G] = E[x_2E[x_1|G]|G] = E[x_1|G]E[x_2|G]$.
(c) ⇒ (d) Let $x_1 \in L((\Omega,F_1),(\mathbb{R}_+,B(\mathbb{R}_+))),\ x_3 \in L((\Omega,G),(\mathbb{R}_+,B(\mathbb{R}_+)))$.
Then $E[x_1x_3|(F_2 \vee G) \vee G] = x_3E[x_1|F_2 \vee G] = x_3E[x_1|G] = E[x_1x_3|G]$, by (c). With the aid of the monotone class theorem one then shows that for any
$x_1 \in L((\Omega,F_1 \vee G),(\mathbb{R}_+,B(\mathbb{R}_+)))$, $E[x_1|(F_1 \vee G) \vee G] \ = \ E[x_1|G]$, hence the result.
(d) ⇒ (c) This follows directly from the definition of conditional independence by restriction of the σ-algebras according to, $E[x_1|F_2 \vee G] = E[E[x_1|F_2 \vee G]|G] = E[x_1|G]$, where for the first equality (d) is used and for the second equality reconditioning of conditional expectation is used. □

Proposition 2.9.3. Sufficient conditions for conditional independence. *Consider sub-σ-algebras F_1,F_2,G of F*

(a)If $F_1 \subseteq G$ or $F_2 \subseteq G$ then $(F_1, F_2|G) \in \mathrm{CI}$. In particular,
$(F_1, F_2|F_1) \in \mathrm{CI}$ and $(F_1, F_2|F_2) \in \mathrm{CI}$.
(b)If the σ-algebras F_1 and $F_2 \vee G$ are independent then $(F_1, F_2|G) \in CI$.
(c)Let $G_0 = \{\Omega, \emptyset\}$ up to null sets of F.
Then F_1 and F_2 are independent if and only if $(F_1, F_2|G_0) \in \mathrm{CI}$.

Proof. (a) If $F_1 \subset G$ then by Theorem 2.8.2.(c) and (b), and, for all $x_1 \in L_+(F_1)$, $E[x_1|F_2 \vee G] = x_1 = E[x_1|G]$. The conclusion follows from
Proposition 2.9.2 (c) $\Leftrightarrow$ (a).
(b) For all $x_1 \in L_+(F_1)$, $E[x_1|F_2 \vee G] = E[x_1] = E[x_1|G]$, because of Proposition 2.8.2 and the result follows from Proposition 2.9.2.
(c) ($\Rightarrow$) Note that G is the trivial algebra. Every random variable measure measurable with respect to G is almost surely equal to a constant. Then,

$$\begin{aligned} &\forall\, x_1 \in L_+(F_1),\; x_2 \in L_+(F_2), \\ &E[x_1x_2|G] = E[x_1x_2], \text{ by the trivial } \sigma\text{-algebra } G, \\ &= E[x_1]E[x_2], \text{ by independence}, \\ &= E[x_1|G]E[x_2|G]. \end{aligned}$$

($\Leftarrow$) As above, $\forall\, x_1 \in L(F_1),\; x_2 \in L(F_2)$,

$$E[x_1x_2] = E[x_1x_2|G_0] = E[x_1|G_0]E[x_2|G_0] = E[x_1]E[x_2],$$

hence F_1 and F_2 are independent. □

Proposition 2.9.4. *Consider the sub-σ-algebras F_1, F_2, G of F.*

(a)If $(F_1, F_2|G) \in \mathrm{CI}$ then $F_1 \cap F_2 \subseteq G$.
(b)Assume that $F_2 \subseteq F_1$. Then $(F_1, F_2|G) \in \mathrm{CI}$ if and only if $F_2 \subseteq G$. In particular, $(F_1, F_1|G) \in \mathrm{CI}$ if and only if $F_1 \subseteq G$.

Proof. (a) Let $A \in F_1 \cap F_2$. Then,

$$\begin{aligned} E[I_A|G] &= E[I_A|F_2 \vee G], \text{ because } (F_1, F_2|\, G) \in \mathrm{CI}, \\ &\quad \text{because of Proposition 2.9.2.(a-c), and because } A \in F_1 \cap F_2 \subseteq F_1, \\ &= I_A, \text{ because } A \in F_1 \cap F_2 \subseteq F_2;\; \Rightarrow\; A \text{ is } G\text{-measurable}. \end{aligned}$$

(b) ($\Rightarrow$) This follows directly from (a) and because $F_2 \subseteq F_1$ implies that $F_2 = F_1 \cap F_2 \subseteq G$.
($\Leftarrow$) Let $x_1 \in L_+(F_1)$. Then $F_2 \subseteq G$ implies that $F_2 \vee G \subseteq G$ hence $F_2 \vee G = G$. This implies that $E[x_1|G] = E[x_1|F_2 \vee G]$ and one concludes with Proposition 2.9.2.(a-c). □

In case the σ-algebras are generated by Gaussian random variabels one can obtain an explicit characterization of conditional independence in terms of the joint distribution of the variables.

Proposition 2.9.5. *Let $x : \Omega \to \mathbb{R}^{n_x}$, $y_1 : \Omega \to \mathbb{R}^{n_{y_1}}$, and $y_2 : \Omega \to \mathbb{R}^{n_{y_2}}$ be jointly Gaussian random variables with $0 \prec Q_{xx}$. Then $(F^{y_1}, F^{y_2}|F^x) \in CI$ if and only if $Q_{y_1 y_2} = Q_{y_1 x} Q_{xx}^{-1} Q_{x y_2}$.*

Proof. The characterization of the conditional independence relation holds if and only if

$$\begin{aligned} &\forall\, v \in \mathbb{R}^{n_{y_1}},\ w \in \mathbb{R}^{n_{y_2}}, \\ &E[\exp(iv^T y_1 + iw^T y_2)|F^x] = E[\exp(iv^T y_1)|F^x] E[\exp(iw^T y_2) \mid F^x]. \end{aligned}$$

By the formula for the conditional characteristic function of Gaussian random variables of Proposition 2.8.3.(c) and calculations, one obtains the relation of the proposition. □

Example 2.9.6. Let $x,\ y_1,\ y_2 : \Omega \to \mathbb{R}^2$, $(y_1,\ y_2) \in G(0,Q)$ with

$$y_1 = \begin{pmatrix} y_{11} \\ y_{12} \end{pmatrix},\ y_2 = \begin{pmatrix} y_{21} \\ y_{22} \end{pmatrix},\ x = \begin{pmatrix} y_{11} \\ y_{22} \end{pmatrix},\ \ Q_{(y_1,y_2)} = \begin{pmatrix} 1 & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 0 & \frac{1}{3} \\ \frac{1}{2} & 0 & 1 & 0 \\ 0 & \frac{1}{3} & 0 & 1 \end{pmatrix};$$

$$Q_{y_1,y_2} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1/3 \end{pmatrix} I \begin{pmatrix} 1/2 & 0 \\ 0 & 1 \end{pmatrix} = Q_{y_1,x} Q_x^{-1} Q_{y_2,x}^T.$$

Then $(F^{y_1}, F^{y_2}|F^x) \in$ CI.

2.10 Computations

The purpose of this section is to summarize several computations for results of this chapter. The focus is on numerical computations, not on analytic calculations. No explicit procedures are written because the formulas are often stated explicitly in the theorem statements or in proofs. Readers with an interest in computations find in this section an overview of those results for which computations are of interest to engineering.

1. Computation of the conditional mean and the conditional variance of one Gaussian random variable on another, see Theorem 2.8.3. This concerns the computations of the matrix $Q_{x,y} Q_y^{-1}$ and that of the conditional variance matrix $Q_{x|y}$.
2. Computation of the conditional mean of a random variable x on a finite-valued random variable y, see Proposition 2.8.4.
3. The representation of a Gaussian random variable as a function of a standard Gaussian random variable, see Proposition 2.7.5.

2.11 Exercises

Problem 2.11.1. Let $x : \Omega \to \mathbb{R}^{n_x}$ be a random variable. Prove that x is a Gaussian random variable if and only if for every $c \in \mathbb{R}^{n_x}$, $c^T x : \Omega \to \mathbb{R}$ is a real-valued Gaussian random variable.

Problem 2.11.2. Let $x : \Omega \to \mathbb{R}^{n_x}$, $x \in G(0, Q_x)$, $v : \Omega \to \mathbb{R}^{n_y}$, $v \in G(0, Q_v)$, $Q_v = Q_v^T \succ 0$, $C \in \mathbb{R}^{n_y \times n_x}$. Assume that $x,\ v$ are independent random variables. Define,

$$y = C\,x + v,\ y : \Omega \to \mathbb{R}^{n_y}.$$

Determine an expression for $E[x \mid F^y]$ and for $E[\exp(iw_x^T x) \mid F^y]$ with $w_x \in \mathbb{R}^{n_x}$.

Problem 2.11.3. Let $x : \Omega \to \mathbb{R}^{n_x}$, $x \in G(0, Q_x)$, $Q_x = Q_x^T \succ 0$, $y : \Omega \to \mathbb{R}^{n_y}$, $y \in G(0, Q_y)$. Assume that x, y are jointly Gaussian random variables and that,

$$E[\exp(iw^T\ y)|F^x] = \exp(iw^T\ Cx - \frac{1}{2} w^T Q_{y|x} w),$$

for a matrix $C \in \mathbb{R}^{n_y \times n_x}$ and a matrix $Q_{y|x} \in \mathbb{R}^{n_y \times n_y}$ satisfying $Q_{y|x} = Q_{y|x}^T \geq 0$.

Prove that there exists a random variable $v : \Omega \to \mathbb{R}^{n_y}$, such that $x,\ v$ are independent, $v \in G(0, Q_v)$, and $y = Cx + v$.

Problem 2.11.4. Let $x : \Omega \to \mathbb{R}$, $x \in L_1$, $y : \Omega \to \mathbb{R}$ be a bounded random variable, and let G be a sub-σ-algebra. Prove that

$$E[E[x|G]y] = E[xE[y|G]].$$

Problem 2.11.5. Let $x, y_1, y_2 : \Omega \to \mathbb{R}$ be jointly Gaussian random variables with zero mean values.

(a) Prove that if $F^x \subseteq F^{y_1} \vee F^{y_2}$ that then $x = c_1 y_1 + c_2 y_2 + c_3$ for some $c_1, c_2, c_3 \in \mathbb{R}$.
(b) Suppose that the variance matrix of (x, y_1, y_2) has unit diagonal and that $E[y_1 y_2] \neq \pm 1$. Prove that then $(F^{y_1}, F^{y_2} \mid F^x) \in CI$ and $F^x \subseteq (F^{y_1} \vee F^{y_2})$ if and only if either $x = y_1$ or $x = y_2$ almost surely.

Problem 2.11.6. Let $x, y : \Omega \to \mathbb{R}^2$ be random variables that are jointly Gaussian, $(x, y) \in G(0, Q)$, with $\lambda \in (0, 1)$ and

$$Q = \begin{pmatrix} 1 & 0 & \lambda & 0 \\ 0 & 1 & 0 & 0 \\ \lambda & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

One calls $z : \Omega \to \mathbb{R}^n$ a *sufficient statistic* for x given y if (1)$F^z \subseteq F^y$ and (2)$E[\exp(iw^T x) \mid F^y] = E[\exp(iw^T x) \mid F^z]$.

(a) Determine $E[\exp(iw^T x) \mid F^y]$.
(b) Determine a sufficient statistic z for x given y with $z \neq y$.
(c) Is the sufficient statistic minimal in a to-be-specified sense?

2.12 Further Reading

History Probability in elementary form has been used for centuries. It was used to answer problems of daily life and of gambling. The formulation of measure theory by E. Borel was a basis for probability theory as used in this book. Probability theory was formulated and practiced in The Netherlands, France, and the United Kingdom, several centuries ago. It may also have been practiced in other countries and at other continents. Major schools of probability theory of the last century are those of France, Russia, India, and the United States of America.

Measure theory was advanced by F.E.J.E. Borel (1871 – 1956), a French mathematician. The related theory for integration was formulated by H.L. Lebesgue (1875 – 1941), another French mathematician.

The measure theoretic formulation of stochastic processes is due to the Russian mathematician A. Kolmogorov, see [27] for the English translation. In the USA, D. Blackwell contributed significantly in the period 1940 - 1970, [6, 7, 9]. The India school of probability is famous and it includes C. Rao, [38], and others. In Europe P. Lévy and later J. Neveu [35, 36, 37] contributed to probability theory. In Russia probability theory was advanced by A. Kolmogorov, A. Shiryayev, and R.S. Liptser, [27, 42, 28].

Books on probability theory at an elementary level. For an elementary introduction to probability theory and stochastic processes the author recommends for engineers the books [21, 40], and the book [20] for mathematics students. Other books at an elementary level are [11, 14].

Books at the level of this course. A useful reference for probability theory at the level of this book is [23]. Additional books on probability theory at the level of this chapter include [2, 5, 13, 35, 42].

Books in probability at an advanced level include [10, 16, 19, 29].

Books on measure theory include [5, 22]. Papers on measure theory without probability include [39].

Probability distribution functions. Books with a description of many probability distributions are [24, 25, 26]. For the Beta pdf see [25, Ch. 24], [1, p. 133], and [20, pp. 168–172].

Probability spaces. For Borel sets and Borel spaces, the reader is referred to the books [12, Ch. 9] and [41, Ch. 3] for further details. The concept of a Borel space was introduced by D. Blackwell, [6, 8]. For a complete separable metric space see [17, Ch. 3]. The concept of a *universal measureable set* is referred to [18, Def. 13.9.2] and [4].

Conditional expectation This is covered in many of the books mentioned above. For example, [23]. Theorem 2.8.6 is a generalization of [10, Cor. 4.38].

Conditional independence. The concept of conditional independence is rather old. The elementary properties are described by P.A. Meyer, [31, 32]. Several properties used in this chapter and in Chapter 19 are from the report and papers, [15, 33, 34, 43, 44].

References

1. R.B. Ash. *Basic probability theory*. John Wiley & Sons, New York, 1970. 49
2. R.B. Ash. *Real analysis and probability*. Academic Press, New York, 1972. 12, 49, 741
3. R.R. Bahadur and E.L. Lehmann. Two comments on 'Sufficiency and statistical decision functions'. *Ann. Math. Statist.*, 26:139–142, 1955. 31
4. D.P. Bertsekas and S.E. Shreve. *Stochastic optimal control: The discrete time case*. Academic Press, New York, 1978. 49, 428, 431, 468, 575, 595
5. P. Billingsley. *Probability and measure (Third Edition)*. John Wiley & Sons, New York, 1995. 49, 73
6. D. Blackwell. On a class of probability spaces. In *Proc. Third Berkeley Symp. Math. Statist. Prob.*, volume 2, pages 1–6, Berkeley, CA, 1956. University of California. 49, 419
7. D. Blackwell. A Borel set not containing a graph. *Ann. Math. Statist.*, 39:1345–1347, 1968. 49
8. D. Blackwell. The stochastic processes of Borel gambling and dynamic programming. *Ann. Statist.*, 4:370–374, 1976. 49, 376, 419
9. D. Blackwell and C. Ryll-Nardzewski. Nonexistence of everywhere proper conditional distributions. *Ann. Math. Statist.*, 34:223–225, 1963. 49
10. L. Breiman. *Probability*. Addison-Wesley Publ. Co., Reading, MA, 1968. 49, 73, 741, 758
11. P. Brémaud. *An introduction to probabilistic modeling*. Springer-Verlag, Berlin, 1988. 49
12. A. Browder. *Mathematical analysis - An introduction*. Undergraduate texts in mathematics. Springer-Verlag, New York, 1996. 30, 49, 424, 426, 475, 526, 635, 636, 677, 815
13. K.L. Chung. *A course in probability theory*. Academic Press, New York, 1974. 12, 49, 741, 758
14. K.L. Chung. *Elementary probability with stochastic processes*. Springer-Verlag, Berlin, 1975. 49, 73
15. A.P. Dawid. Conditional independence for statistical operations. *Ann. Math. Statist.*, 8:598–617, 1980. 49, 276, 742
16. C. Dellacherie and P.A. Meyer. *Probabilités et potentiel, CH. I à IV*. Hermann, Paris, 1975. 49
17. J. Dieudonné. *Foundations of modern analysis*. Academic Press, New York, 1969. 49, 636
18. J. Dieudonné. *Treatise on analysis, volume II*. Academic Press, New York, 1970. 49
19. R.M. Dudley. *Real analysis and probability*. Wadsworth and Brooks/Cole, Pacific Grove, CA, 1989. 49
20. C.M. Grinstead and J.L. Snell. *Introduction to probability, 2nd revised edition*. American Mathematical Society, Boston, 1997. 49, 698
21. Bruce Hajek. *Random processes for engineers*. Cambridge University Press, Cambridge, 2015. 49, 73
22. P.R. Halmos. *Measure theory*. Van Nostrand, New York, 1950. 49
23. J. Jacod and Ph. Protter. *Probability essentials*. Universitext. Springer, Berlin, 1999. 49
24. N.L. Johnson and S. Kotz. *Distributions in statistics: Discrete distributions*. Houghton Mifflin, Boston, 1969. 49
25. N.L. Johnson and S. Kotz. *Distributions in statistics: Continuous univariate distributions – 2*. Houghton-Mifflin, Boston, 1970. 49
26. N.L. Johnson and S. Kotz. *Distributions in statistics: Continuous multivariate distributions*. Houghton-Mifflin, Boston, 1972. 49
27. A.N. Kolmogorov. *Foundations of probability (translation)*. Chelsea, New York, 1950. 49, 72, 73
28. R.S. Liptser and A.N. Shiryayev. *Statistics of random processes: I. General theory; II. Applications*. Springer-Verlag, Berlin, 1977,1978. 49, 742
29. M. Loève. *Probability theory, 3rd edition*. Van Nostrand Reinhold Co. Inc., New York, 1963. 49, 72, 742
30. G. Marsaglia. Conditional means and covariances of normal variables with singular covariance matrix. *Amer. Statist. Assoc. J.*, 59:1203–1204, 1964. 41

31. P.A. Meyer. *Probability and Potentials*. Blaisdell Publishing Company, Waltham, MA, 1966. 49, 336, 741, 758
32. P.A. Meyer. *Processus de Markov*, volume 26 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1967. 49, 73, 341
33. M. Mouchart and J.-M. Rolin. A note on conditional independence (with statistical applications). Report 129, Institut de Mathématique Pure et Appliquée, Université Catholique de Louvain, Louvain-la-Neuve, 1979. 49, 276, 723, 742
34. M. Mouchart and J.-M. Rolin. A note on conditional independence with statistical applications. *Statistica*, 44:557–584, 1984. 49, 276, 723, 742
35. J. Neveu. *Mathematical foundations of the calculus of probability*. Holden-Day Inc., San Francisco, 1965. 49, 741
36. J. Neveu. *Processus aléatoires Gaussiens*. Presses Universitaires Montréal, Montréal, 1968. 49, 73
37. J. Neveu. *Martingales à temps discrets*. Masson et Cie, Paris, 1972. 39, 49, 755, 758
38. C.R. Rao. *Linear statistical inference and its applications (Second Edition)*. John Wiley & Sons, New York, 1973. 49
39. V.A. Rohlin. On the fundamental ideas of measure theory. *Amer.Math.Soc. Transl.*, 1:1–54, 1949. 49
40. Sheldon Ross. *A first course in probability*. Prentice-Hall, Upper Saddle River, 1998. 49
41. H.L. Royden. *Real analysis, 2nd edition*. MacMillan Co., New York, 1968. 49, 526, 636
42. A.N. Shiryaev. *Probability*. Springer, New York, 1984. 49
43. C. van Putten and J.H. van Schuppen. The weak and strong Gaussian probabilistic realization problem. *J. Multivariate Anal.*, 13:118–137, 1983. 49, 237, 275, 276
44. C. van Putten and J.H. van Schuppen. Invariance properties of the conditional independence relation. *Ann. Probab.*, 13:934–945, 1985. 49, 276, 723, 741, 742

Chapter 3
Stochastic Processes

Abstract Elementary concepts and results of the theory of stochastic processes are summarized in this chapter. Concepts presented include: a stochastic process, equivalent processes, a Gaussian process, stationarity, time-reversibility, and a Markov process. It is shown how to go from the definition of a Gauss-Markov process to a representation of such a process in the form of a state equation of a stochastic system. Advanced topics of the theory of stochastic processes are provided in Chapter 20.

Key words: Stochastic process. Gaussian process. Markov process.

Examples of phenomena that may be modelled as a stochastic process include: (1) the position of a moored tanker at sea, (2) the deviations of a ship from its planned course, (3) electric energy demand per hour as measured by a power company, (4) the weight per product in quality control at a commercial company, (5) arrivals of call requests at a telephone switch, (6) the daily movement of the price of a stock on the market, etc. In Chapter 4 it will be described how to model a dynamic phenomenon with uncertainty as a stochastic system.

3.1 Concepts

The set T in the definition of a stochastic process will be called the *time index set* or the *time axis*. Examples of this time index set are: $T = \{0,1,2,\ldots,n\}$ for an integer $n \in \mathbb{Z}_+$, the set of the natural numbers $T = \mathbb{N} = \{0,1,2,\ldots\}$, the intervals of the real numbers $T = [0,1]$, $T = \mathbb{R}_+$ and $T = \mathbb{R}$. In all of these cases T is totally ordered. Depending on whether T is a discrete or a continuous subset of $\mathbb{R}$ one calls the stochastic process a *discrete-time process* or a *continuous-time process* respectively. In this book only discrete-time stochastic processes are treated. The time index set will thus primarily be either $T = \mathbb{N}_{t_1} = \{0,1,\ldots,t_1\} = T(0:t_1)$ for an integer $t_1 \in \mathbb{Z}_+$, or $T = \mathbb{N} = \{0,1,\ldots\}$, or $T = \mathbb{Z}$. In case $T = \mathbb{N}\times\mathbb{N}$ or $T = \mathbb{Z}\times\mathbb{Z}$ then one calls the associated process a *random field.*

If the process takes values in the real numbers then the associated σ-algebra is that of the Borel sets, unless mentioned otherwise. Thus if $X = \mathbb{R}^n$ then the associated σ-algebra is denoted by $B(\mathbb{R}^n)$.

Definition 3.1.1. Consider a probability space (Ω,F,P), an index set T, and a measurable space (X,G). A *stochastic process*, or a *random process*, or just a *process*, on these spaces is a function $x : \Omega \times T \to X$ such that for all $t \in T$, the map $x(.,t) : \Omega \to X$ is a measurable map or, in other words, $x(.,t)$ is a random variable. Notation: A stochastic process is denoted by $x = \{x_t, t \in T\}$, and for $x(\omega,t)$ one uses one of the following equivalent notations $x(t) = x_t = x_t(\omega) = x(\omega,t)$. For $\omega \in \Omega$ fixed, one calls $x(\omega,.) : T \to X$ a *sample path* of the process x.

A phenomenon may be modelled as a stochastic process. Of the phemenon, one obtains observations which are general vectors in the vector space of tuples of the real numbers. The observations correspond in the model of a stochastic process to a realization or a sample path of the stochastic process. In general, one can not observe what corresponds to all sample paths of a stochastic process.

Construction of a Stochastic Process

Of interest is the construction of a probability measure on a stochastic process.

Definition 3.1.2. *Construction of a stochastic process on a finite time index set.* Consider the finite time index set $T = T(0:t_1) = \{0,1,2,\ldots,t_1\}$ for a positive integer $t_1 \in \mathbb{Z}_+$. The stochastic process is to take values in the measurable space $(\mathbb{R}^n, B(\mathbb{R}^n))$.

Define the set of finite-length trajectories as,

$$\Omega = \{y : T \to \mathbb{R}^n |\ y \text{ is a function}\}.$$

Due to the fact that the time index set is a finite set, it is not necessary to take measureable functions in the set Ω.

Define a *cylinder set* of Ω by the formula,

$$\begin{aligned}
&\forall\, s \in T,\ A_s \in B(\mathbb{R}^n),\\
A &= \prod_{s=1}^{t_1} A_s = A_1 \times A_2 \times \ldots \times A_{t_1} \in B(\mathbb{R}^{t_1 n});\ \text{ for example},\\
A &= \prod_{s=1}^{t_1} \prod_{k=1}^{n} (-\infty, w_{s,k}] \in B(\mathbb{R}^{t_1 n}),\ \forall\, w_{s,k} \in \mathbb{R}.
\end{aligned}$$

Thus a cylinder set is a finite product of measureable sets; in the particular case, each measureable set is a left-closed interval of $\mathbb{R}$. Define the σ-algebra generated by the cylinder set and denote it by $F = \sigma(\text{all cylinder sets})$. Then (Ω,F) is a measurable space.

Because the space $\mathbb{R}^{t_1 n}$ is a space of a finite tuple of finite-dimensional real numbers, one can define a probability measure on this space by an associated probability distribution function. For example, by constructing a probability measure based on a Gaussian probability distribution function on the space $(\mathbb{R}^{t_1 n}, B(\mathbb{R}^{t_1 n}))$. Then (Ω, F, P) is a probability space called the *probability space of the set of finite sequences of* $\mathbb{R}^n$*-valued random variables*.

Definition 3.1.3. *Construction of a stochastic process on a half-infinite time index set*. Consider next the half-infinite time index set $T = \mathbb{N} = \{0, 1, 2, \ldots, \}$. Define the set of infinite-length sequences as,

$$\Omega = \{y : T \to \mathbb{R}^n |\ y \text{ is a function}\}.$$

Define a cylinder set of a finite-length sequence as in Def. 3.1.2. One then defines the σ-algebra F on the set Ω of infinite sequences by the formula,

$$F = \sigma(B(\mathbb{R}^{tn}),\ \forall\, t \in \mathbb{N}).$$

Then (Ω, F) is a measureable space.

The construction of a probability measure on the latter measureable space requires a new argument. One approach is to define a probability measure on the set of finite sequences on the time index set $T = \{0, 1, 2, \ldots, t\}$, for any time moment $t \in T$. Then this defines by Def. 3.1.2 a probability measure on any finite sequence. The alternative approach follows.

A.N. Kolmogorov has formulated a procedure on how to go from a family of finite-dimensional probability distributions to a probability measure on countable sequences and even on stochastic processes defined on the continuous time-index set $T = \mathbb{R}_+$. This procedure is described below.

Definition 3.1.4. Define a *family of finite-dimensional probability distributions* F_{fdpdf} with values in the measurable space $(\mathbb{R}^n,\ B(\mathbb{R}^n))$, for $n \in \mathbb{Z}_+$, a time index set $T = \mathbb{N}$, as a set,

$$F_{fdpdf} = \left\{ \begin{array}{l} \text{pdf}(.; (t_1, \ldots, t_m)) \in \text{pdf}(\mathbb{R}^{mn}, B(\mathbb{R}^{mn}))| \\ \forall\, m \in \mathbb{Z}_+,\ \forall\, t_1,\ t_2, \ldots,\ t_m \in T,\ \text{such that}\ t_1 < t_2 < \ldots < t_m, \end{array} \right\}.$$

The family is said to be *consistent* if for all $m \in \mathbb{Z}_+$ and for all $i \in \mathbb{Z}_m$, the probability distribution defined below, again belongs to the family; equivalently,

$$\begin{aligned} &\forall\, m \in \mathbb{Z}_+,\ \forall\, f_m \in F_{fdpdf}, \\ &f_{m-1}(.;\ w_1,\ w_2, \ldots,\ w_{i-1},\ w_{i+1}, \ldots, w_m) \\ =\ &f_m(.;\ w_1, \ldots, w_{i-1}, +\infty, w_{i+1}, \ldots, w_m) \in F_{fdpdf}. \end{aligned}$$

Thus each member of F_{fdpdf} is associated with a natural number $m \in \mathbb{N}$ and a finite time sequence $\{t_1, t_2, \ldots, t_m\}$. Then $f_m(.; \ldots) \in F_{fdpdf}$ is a probability distribution function on the space $(\mathbb{R}^{mn}, B(\mathbb{R}^{mn}))$. It then follows from Theorem 2.4.9 that there exists a probability measure P_{mn} on the space of a sequence with m members, of which each element is defined on $(\mathbb{R}^n, B(\mathbb{R}^n))$.

Theorem 3.1.5. *Consider a half-infinite time interval $T = \mathbb{N}$ and a family of finite-dimensional probability distribution functions F_{fdpdf} which is assumed to be consistent as defined in Def. 3.1.4.*

Then there exists a probability measure $P_{\mathbb{N}}$ on the space of half-infinite sequences $(\mathbb{R}^{\mathbb{N}}, B(\mathbb{R}^{\mathbb{N}}))$ such that, for any natural number $m \in \mathbb{N}$, the following condition holds,

$$\begin{aligned} & P_{\mathbb{N}}\left(\left\{\begin{array}{l} \omega \in \Omega |\, x(\omega,t_i) \leq w_i, \ \forall\, i \in \mathbb{Z}_m, \\ \forall\, t_1, t_2, \ldots, t_m \in T, \text{ such that } t_1 < t_2 < \ldots < t_m \end{array}\right\}\right) \\ & = f_m((w_1, w_2, \ldots, w_m); (t_1, t_2, \ldots, t_m)) \in F_{fdpdf}. \end{aligned}$$

Equivalent Processes

The construction of the probability measure on a stochastic process leaves unanswered the question whether the spaces on which a stochastic process has been defined are unique. In general there is no uniqueness. Hence the need for the concept of equivalent processes.

Definition 3.1.6. Consider a time index set T and a measurable space (X, G). Two stochastic processes x, y defined on these spaces and on possibly different probability spaces $(\Omega_x,\ F_x,\ P_x)$, $(\Omega_y,\ F_y,\ P_y)$ are called *equivalent* if they have the same family of finite-dimensional distribution. Thus, if

$$\begin{aligned} & \forall\, n \in \mathbb{Z}_+,\ \forall\, S_n = \{t_1, \ldots, t_n\} \subseteq T,\ \forall\, A_1, \ldots, A_n \in G, \\ & P_x(\{\omega \in \Omega_x |\ \forall\, k \in Z_n,\ x(t_k) \in A_k\}) = P_y(\{\omega \in \Omega_y |\ \forall\, k \in Z_n,\ y(t_k) \in A_k\}). \end{aligned}$$

It can be proven that the concept of equivalence of stochastic processes defined above induces an equivalence relation on the set of stochastic processes defined on the sets T and (X, G). If the stochastic processes x and y take values in the real numbers with $(X, G) = (\mathbb{R}, B(\mathbb{R}))$, then one can replace in the above definition the set A_k by $\{\omega \in \Omega |\ x(t_k) \leq u_k\}$ for $u_k \in \mathbb{R}$ for all $k \in \mathbb{Z}_n$.

Definition 3.1.7. Two stochastic processes $x,\ y : \Omega \times T \to X$ defined on the same probability space and with $(X, G) = (\mathbb{R}^n, B(\mathbb{R}^n))$ are called *modifications of each other* if for all $t \in T$, $x_t = y_t$ a.s., or, equivalently, if,

$$P(\{\omega \in \Omega \mid x(\omega,t) = y(\omega,t)\}) = 1,\ \forall\, t \in T.$$

Note that the set $\{\omega \in \Omega |\ x(\omega,t) \neq y(\omega,t)\}$ may depend on $t \in T$.

Definition 3.1.8. Two stochastic processes $x,\ y : \Omega \times T \to X$ are called *indistinguishable* if,

$$P(\{\omega \in \Omega \mid x(\omega,t) = y(\omega,t),\ \forall\, t \in T\}) = 1.$$

One also says then that y is *indistinguishable* from x or that x is *indistinguishable* from y.

The convention is adopted that a stochastic process has a certain property if one of its modifications has the property.

If the processes x and y are indistinguishable then they are modifications of each other. This follows from the inequality,

$$\begin{aligned} 1 &= P(\{\omega \in \Omega |\ x(\omega,s) = y(\omega,s),\ \forall\ s \in T\}) \\ &\leq P(\{\omega \in \Omega |\ x(\omega,t) = y(\omega,t)\}),\ \ \forall\ t \in T. \end{aligned}$$

The converse of the above statement is in general not true.

3.2 Special Subsets of Stochastic Processes

In the literature one finds formal definitions of stochastic processes in one of the following ways:

1. A specification of the entire family of finite-dimensional probability distributions. This is easily done for Gaussian processes as defined below
2. A specification that the process is a sequence of independent random variables while each random variable of the sequence has the same probability distribution function. Afterwards this stochastic process can be transformed by either an algebraic or an analytic operation.
3. A specification that the stochastic process is a Markov process with a specified transition function. This modeling approach is often followed in control engineering. It is described below for Gauss-Markov processes.

Definition 3.2.1. A stochastic process x is called a *Bernoulli process* with intensity function $q : T \rightarrow [0,1]$, if $x : \Omega \times T \rightarrow \{0,\ 1\}$ on $T = \mathbb{N}$ satisfies:
(1) $\{x(t),\ t \in T\}$ is a sequence of independent random variables;
(2) for all $t \in T$, the random variable $x(.,t) : \Omega \rightarrow \{0,\ 1\}$ has a Bernoulli probability distribution with parameter $q(t)$,

$$P(\{\omega \in \Omega | x(\omega,t) = 1\}) = q(t),\ \ P(\{\omega \in \Omega | x(\omega,t) = 0\}) = 1 - q(t).$$

A Bernoulli process may be used to model the flow of bits in a digital communication channel.

Definition 3.2.2. A stochastic process x is called a *discrete-time Poisson process* or just a *Poisson process* with rate function $\lambda : T \rightarrow \mathbb{R}_{s+} = (0,\infty)$, if $x : \Omega \times T \rightarrow \mathbb{N}$ on $T = \mathbb{N}$ satisfies:
(1) $\{x(t),\ t \in T\}$ is a sequence of independent random variables;
(2) for all $t \in T$, the random variable $x(.,t) : \Omega \rightarrow \mathbb{N}$ has a Poisson probability distribution with parameter $\lambda(t)$,

$$\forall\ k \in \mathbb{N},\ P(\{\omega \in \Omega | x(\omega,t) = k\}) = \lambda(t)^k\ \exp(-\lambda(t))/k!$$

A Poisson process may be used as a model for the number of call requests arriving in a sequence of minute intervals at a call service center where the rate parameter λ varies during the day.

Definition 3.2.3. A stochastic process x is called a *Gamma process* with rate functions $\gamma_1,\ \gamma_2 : T \rightarrow \mathbb{R}_{s+}$, if $x : \Omega \times T \rightarrow \mathbb{R}_+$ on $T = \mathbb{N}$ satisfies:
(1) $\{x(t),\ t \in T\}$ is a sequence of independent random variables;
(2) for all $t \in T$, the random variable $x(.,t) : \Omega \rightarrow \mathbb{R}_+$ has a Gamma probability distribution with parameters $(\gamma_1(t), \gamma_2(t))$,

$$P(\{\omega \in \Omega | x(\omega,t) \in A\}) = \int_A p_{x(t)}(v; \gamma_1(t), \gamma_2(t)) dv,\ \forall\, A \in B(\mathbb{R}_+),$$
$$p_x(v; \gamma_1(t), \gamma_2(t)) = v^{\gamma_1(t)-1}\ \exp(-v/\gamma_2(t)) \gamma_2(t)^{-\gamma_1(t)} / \Gamma(\gamma_1(t)).$$

A gamma process may be used to model the subsequent passsage times of cars at a fixed location of a motorway.

The most well known subset of stochastic processes is the set of Gaussian processes.

Definition 3.2.4. A stochastic process $x : \Omega \times T \rightarrow \mathbb{R}^n$ is called a *Gaussian (stochastic) process* if every finite-dimensional distribution is Gaussian. Equivalently, if

$$\forall\, m \in Z_+,\ \forall\, (t_1, ..., t_m) \in T,\ \ \mathrm{pdf}(x(t_1), \ldots, x(t_m)) \in G.$$

Definition 3.2.5. Two stochastic processes x and y are called *jointly Gaussian* if every joint finite-dimensional distribution is Gaussian. Equivalently, if

$$(\forall\, m,\ k \in Z_+,\ \forall\, (t_1, ..., t_m) \in T,\ \forall\, (s_1, ..., s_k) \in T),$$
$$\mathrm{pdf}(x(t_1), \ldots,\ x(t_m),\ y(s_1), \ldots,\ y(s_k)) \in G.$$

If x, y are jointly Gaussian processes, then each of them is Gaussian. If both x and y are Gaussian processes and if they are independent, then it follows from Proposition 2.7.3 that they are jointly Gaussian.

3.3 Properties of Stochastic Processes

Integrability of Stochastic Processes

Definition 3.3.1. Consider a stochastic process $x : \Omega \times T \rightarrow \mathbb{R}^{n_x}$ for $n_x \in \mathbb{Z}_+$.

(a) One says that this process is *integrable* or of *first order* if for all $t \in T$, and for all $i \in \mathbb{Z}_{n_x}$, $E|x_i(t)| < \infty$. If x is integrable then one defines the *mean-value function* of x as $m_x : T \rightarrow \mathbb{R}^{n_x}$, $m_x(t) = E[x(t)]$.

(b) One says also that this process is *square integrable* or of *second order* if for all $t \in T$, and for all $i \in \mathbb{Z}_{n_x}$, $E|x_i(t)|^2 < \infty$. If the process is square integrable then it follows from the Cauchy-Schwartz inequality that,

$$\forall\, i, j \in \mathbb{Z}_{n_x},\ \forall\, t \in T,\ E|x_i(t) x_j(t)| \leq (E|x_i(t)|^2)^{1/2} (E|x_j(t)|^2)^{1/2} < \infty.$$

If x is square integrable then one defines the *correlation function* and the *covariance function* respectively as

$$\begin{aligned} C_x(t,s) &= E[x(t)x(s)^T], \;\; C_x : T \times T \to \mathbb{R}^{n_x \times n_x}, \\ W_x(t,s) &= E[(x(t)-E[x(t)])(x(s)-E[x(s)])^T], \;\; W_x : T \times T \to \mathbb{R}^{n_x \times n_x}. \end{aligned}$$

Proposition 3.3.2. *Consider a square-integrable stochastic process with mean-value function $m : T \to \mathbb{R}^n$ and with covariance function $W : T \times T \to \mathbb{R}^{n \times n}$. Then:*

(a)for all $t \in T$, $W(t,t) \succeq 0$ or, equivalently, $W(t,t)$ is a positive-definite matrix;
(b)for all $t,s \in T$, $W(t,s) \; = \; W(s,t)^T$;
(c)for all $t,s \in T$, $W(t,s) \; = \; C(t,s) - m(t)m(s)^T$ where C is the correlation function and m the mean-value function of the process.

These properties follow directly from the definitions by algebraic operations.

Stationarity and Time-Reversibility

Modelling of phenomena by stochastic processes motivates the concept of stationarity. A *phenomenon* is considered *stationary* if it behaves more or less the same regardless of the time moment one considers it. An example of stationarity of a phenomenon is the arrival process of call requests at a telephone server. This phenomemon is in general not stationary and one may have to a use a modeling technique to be able to restrict attention to a stationary process. For example, it may be stationary between 14:00 and 15:00 hours on work days. But at night between 02:00 and 03:00 hours, the stochastic process will have completely different behavior, hence a shift by 12 hours changes the family of finite-dimensional probability distributions significantly. In practice, the modeler has to decide whether or not the phenomenon can be approximated by a stationary phenomenon. The concept of stationarity of a stochastic process is the mathematical equivalent of stationarity of a phenomenon.

Definition 3.3.3. A stochastic process $x : \Omega \times T \to X$ defined on T is called *stationary* if

$$\begin{aligned} &\forall\, m \in \mathbb{Z}_+, \; \forall\, s \in \mathbb{Z}, \; \forall\, t_1, \ldots t_m \in T, \; \text{such that}\; t_1 + s, \ldots t_m + s \in T, \\ &\text{pdf}(x(t_1), \ldots, \; x(t_m)) = \text{pdf}(x(t_1 + s), \ldots, \; x(t_m + s)). \end{aligned}$$

A property that certain stochastic processes have is time-reversibility. In physics the concept of time-reversibility has been formulated and noticed to hold for particular phenomena. This property is occasionally a useful technical tool such as for the performance analysis of communication networks.

Definition 3.3.4. A stochastic process $x : \Omega \times T \to X$ will be called *time-reversible* if

$$\begin{aligned} &(\forall\, m \in Z_+, \; \forall\, t_1, \ldots, t_m \in T, \; \forall\, t \in \mathbb{Z}, \; \text{such that}\; t - t_1, \ldots, t - t_m \in T), \\ &\text{pdf}(x(t_1), \ldots, x(t_m)) = \text{pdf}(x(t - t_1), \; \ldots, \; x(t - t_m)). \end{aligned}$$

Proposition 3.3.5. *If $x : \Omega \times T \to X$ is a time-reversible stochastic process then it is stationary.*

Proof. Let $m \in Z_+$ and $t, t_{1,} \ldots t_m \in T$. Then

$$\begin{aligned} &\text{distribution of } (x(t_1), \ldots x(t_m)) \\ &= \text{distribution of } (x(-t_1), ... x(-t_m)), \text{ by time-reversibility,} \\ &= \text{distribution of } (x(t+t_1), \ldots, x(t+t_m)), \text{ again by time-reversibility,} \end{aligned}$$

hence x is stationary. □

Time-reversibility of a Gaussian process is investigated in Section 3.4.

Markov Processes

A Markov process has been defined as the state of a stochastic system and it corresponds to the state process of a deterministic system. Markov processes are named after the Russian mathematician A.A. Markov (1856-1922). This is the father Markov to be distinguished from the son Markov who was also a mathematician.

In this section Markov processes are defined and elementary properties are characterized. This subsection makes use of the concept of conditional independence of σ-algebras which is described in Section 2.9.

Definition 3.3.6. A discrete-time *Markov process* is a stochastic process $x : \Omega \times T \to X$, with either $T = \mathbb{N}$ of $T = \{0, 1, 2, \ldots, t_1\}$, such that for all $t \in T$,

$$\begin{aligned} &\forall\, t \in T,\ (F_t^{x+}, F_t^x | F^{x(t)}) \in CI; \text{ where,} \\ F_t^{x+} &= \sigma(\{x(s),\ \forall\, s,\ t \in T,\ s \geq t\}), \\ F_t^x &= F_t^{x-} = \sigma(\{x(s),\ \forall\, s,\ t \in T,\ s \leq t\}). \end{aligned}$$

It follows from Proposition 19.8.2.(f) that, with $T = \mathbb{N}$,

$$\begin{aligned} &\forall\, t \in T,\ (F_t^{x+},\ F_t^{x-} |\ F^{x(t)}) \in \text{CI}, \\ \Leftrightarrow &\ \forall\, t \in T \backslash \{0\},\ (F_{t+1}^{x+},\ F_{t-1}^{x-} |\ F^{x(t)}) \in \text{CI}. \end{aligned}$$

The interpretation of the above definition is that the future and the past of a Markov process are conditionally independent given the current state, and that for any time moment.

Proposition 3.3.7. *Let $x : \Omega \times T \to \mathbb{R}^n$ be a stochastic process. The following statements are equivalent:*

(a) *x is a Markov process;*

(b) $(F^{x+}_{t+1},\, F^{x}_{t-1} |\, F^{x(t)}) \in CI,\ \forall\, t \in T;$

(c) $\forall\, s,\ t \in T,\ s < t,\ \forall\, w \in \mathbb{R}^n,\ E[\exp(iw^T x(t))|F^x_s] = E[\exp(iw^T x(t))|F^{x(s)}];$

(d) $\forall\ s,t \in T,\ s < t,\ \forall\, f : \mathbb{R}^n \to \mathbb{R},$ *with* $f(x(t)) \in L_1,$
$E[f(x(t))|F^x_s] = E[f(x(t))|F^{x(s)}];$

(e) $\forall\, m \in \mathbb{Z}_+,\ r_1,\ldots,r_m,s,t \in T,\ r_1 < \ldots \leq r_m \leq s \leq t,$
$f : \mathbb{R}^n \to \mathbb{R},\ f(x(t)) \in L_1,$
$E[f(x(t))|\sigma(\ x(r_1),\ \ldots,\ x(r_m),\ x(s)\ )] = E[f(x(t))|F^{x(s)}].$

The interpretation of the definition of a Markov process is rather direct in case of the asymmetric formulation of conditional independence as stated in (c) above: the future of the process conditioned on the past and the present of the process, depends only on the present of the Markov process.

Proof. Proof of Proposition 3.3.7. (a) ⇔ (b). See Proposition 19.8.2.(f).
(a) ⇒ (c). This follows from 2.9.2.(a) & (c).
(c) ⇔ (d). This result can be proven with techniques of analysis and of probability theory.
(c) ⇒ (a). It will be proven that for all $m \in \mathbb{Z}_+\ t_1,\ \ldots,\ t_m,\ s \in T,\ s < t_1 < \ldots < t_m,$ $\forall\, w_j \in \mathbb{R}^n$

$$E[\exp(i\sum_{j=1}^{m} w_j^T x(t_j))|F^x_s] = E[\exp(i\sum_{j=1}^{m} w_j^T x(t_j))|F^{x(s)}].$$

The conclusion then follows by applying the monotone class theorem. The above equality is proven for $m = 2$. Then,

$$\begin{aligned}
&E[\exp(i\sum_{j=1}^{2} w_j^T x_{t_j})|F^x_s] = E[E[\exp(iw_2^T x_{t_2})|F^x_{t_1}]\exp(iw_1^T x_{t_1})|F^x_s]\\
&\quad \text{by reconditioning,}\\
&= E[E[\exp(iw_2^T x_{t_2})|F^{x_{t_1}}]\exp(iw_1^T x_{t_1})|F^x_s],\ \text{ by (c),}\\
&= E[E[\exp(iw_2^T x_{t_2})|F^{x_{t_1}}]\exp(iw_1^T x_{t_1})|F^{x_s}],\ \text{ by (d),}\\
&= E[E[\exp(iw_2^T x_{t_2})|F^x_{t_1}]\exp(iw_1^T x_{t_1})|F^{x_s}],\ \text{ by (c),}\\
&= E[\exp(i\sum_{j=1}^{2} w_j^T x_{t_j})|F^{x_s}].
\end{aligned}$$

(d) ⇒ (e).

$$\begin{aligned}
&E[f(x_t)|\sigma(x_{r_1},\ldots,x_{r_m},x_s)] = E[E[f(x_t)|F^x_s]|\sigma(x_{r_1},\ldots,x_{r_m},x_s)]\\
&= E[f(x_t)|F^{x_s}].
\end{aligned}$$

(e) ⇒ (d). This follows with the use of the monotone class theorem, Theorem 19.1.4. □

A Markov process is thus characterized by one of the equivalent properties of Proposition 3.3.7. The main characteristic of a Markov process is that the conditional distribution of the future of the process given the past and the present of the process, depends only on its present. This property is the formalization for a stochastic process of the concept of state of a deterministic system.

3.4 Gaussian Processes

For many phenomena with random fluctuations, a Gaussian process is a realistic mathematical model. In such phenomena, the fluctuations are caused by very many independent contributions hence the probability distribution of a variable of this phenomenon is realistically modelled by a Gaussian probability distribution. Below the elementary properties of Gaussian processes are discussed.

Recall from Def. 3.2.4 that a stochastic process $x : \Omega \times T \to \mathbb{R}^n$ is called a Gaussian process if for any $m \in \mathbb{Z}_+$ and for any $t_1, \ldots, t_m \in T \subset \mathbb{Z}$, the finite-dimensional probability distribution function of $(x(t_1), \ldots, x(t_m))$ is Gaussian.

Definition 3.4.1. A *discrete-time Gaussian white noise process* with values in $\mathbb{R}^{n_v}$ and intensity $Q_v : T \to \mathbb{R}^{n_v \times n_v}_{pds}$, is a stochastic process $v : \Omega \times T \to \mathbb{R}^{n_v}$ such that:

(1)v is an independent sequence; or, equivalently, $\{v(t), t \in T\}$ is a collection of independent random variables;

(2)for all $t \in T$, $v_t \in G(0, Q_v(t))$; or, equivalently, $v(t)$ has a Gaussian probability distribution with the indicated parameters.

It is called a *stationary Gaussian white noise process* if the variance does not depend on time; equivalently, if for all $t \in T$, $Q_v(t) = Q_v(0)$.

It is called a *standard Gaussian white noise process* if, (1) it is a stationary Gaussian white noise process and (2) the variance equals the identity matrix (for all $t \in T$, $Q_v(t) = I_{n_v} \in \mathbb{R}^{n_v \times n_v}$).

Covariance Functions

A Gaussian process is determined by its family of finite-dimensional distributions. A family of Gaussian finite-dimensional distributions is in turn completely determined by the mean-value function and the covariance function of the process. The mean-value function is often assumed to be identically zero. This assumption is justified by the argument that the mean value function of a stochastic process is a deterministic function and can therefore be modelled separately from the stochastic process. Therefore a Gaussian process is completely determined by its covariance function, hence the interest in this class of functions.

Definition 3.4.2. (a)A function $W : T \times T \to \mathbb{R}^{n \times n}$ is called *positive definite* if

$$\forall\, m \in \mathbb{Z}_+,\ \forall\, t_1,\ldots,t_m \in T \subset \mathbb{R},\ \forall\, c_1,\ldots,c_m \in \mathbb{R}^n,$$
$$\sum_{i=1}^{m}\sum_{j=1}^{m} c_i^T W(t_i,t_j)c_j \geq 0.$$

(b)The function W is called *uniformly strictly positive-definite*, (in operator theory called a *coercive operator*), if

$$\exists\, r_W \in (0,\infty) \text{ such that } \forall\, e : T \to \mathbb{R}^n,$$
$$e \neq 0 \;\Rightarrow\; \sum_{s\in T} e(s)^T W(s,s)e(s) \geq r_W\ (\sum_{s\in T} e(s)^2) > 0.$$

This condition implies that, for all $s \in T$ and for a sequence e such that $e(s)=1$ and $\forall\, t \in T\backslash\{s\}$, $e(t)=0$, then, for all $s \in T$, $W(s,s) \succeq r_W I_n \succ 0$.

Proposition 3.4.3. *The function $W : T \times T \to \mathbb{R}^{n\times n}$ is a covariance function of a stochastic process if and only if (1) for all $t,s \in T$, $W(t,s) = W(s,t)^T$; and (2) W is positive definite.*

Proof. ($\Rightarrow$) Suppose that W is the covariance function of a process x that has zero mean function; if it does not have zero mean then consider the process $\bar{x} : \Omega \times T \to \mathbb{R}^n$, $\bar{x}(t) = x(t) - E[x(t)]$ that has zero mean. Then $W(t,s) = E[x(t)x(s)^T] = (E[x(s)x(t)^T])^T = W(s,t)^T$. Let $m \in \mathbb{Z}_+$, $t_1,\ldots,t_m \in T \subset \mathbb{R}$ and $c_1,\ldots,c_m \in \mathbb{R}^n$. Then

$$\sum_{i=1}^{m}\sum_{j=1}^{m} c_i^T W(t_i,t_j)c_j = \sum_{i=1}^{m}\sum_{j=1}^{m} E[c_i^T x_{t_i} x_{t_j}^T c_j] = E[(\sum_{i=1}^{m} c_i^T x_{t_i})^2] \geq 0.$$

($\Leftarrow$) Let $m \in \mathbb{Z}_+$, $t_1,\ldots,t_m \in T$. Define $Q \in \mathbb{R}^{nm\times nm}$ whose (i,j)-th block is equal to $W(t_i,t_j)$. From the assumptions follows that $Q = Q^T \succeq 0$. Hence $(0,Q)$ are the parameters of a jointly Gaussian distribution of $\mathbb{R}^{nm}$. This way one can construct a family of Gaussian distributions. Then there exists a stochastic process x that is Gaussian and that has W as its covariance function. □

Stationarity and Time-Reversibility of Gaussian Processes

In this subsection several properties of Gaussian processes are expressed in terms of its covariance function.

Proposition 3.4.4. *Let $x : \Omega \times T \to \mathbb{R}^{n_x}$ be a Gaussian process on $T \subset \mathbb{Z}$ with mean-value function m_x and covariance function W_x. Then x is stationary if and only if the following conditions both hold:*

(1)$m_x(t) \;=\; m_x(0)$, for all $t \in T$;
(2)$W_x(t,s) \;=\; W_x(t+r,s+r)$ for all $s,\ t \in T$ and $r \in \mathbb{Z}_+$ such that $s+r,\ t+r \in T$.

A consequence of Condition (2) above is that the variance function of the process is constant, for all $t \in T$,

$$Q_x(t) = E[(x(t)-m_x(t))(x(t)-m_x(t))^T] = W_x(t,t) = W_x(0,0) = Q_x(0).$$

Proof. Let

$$\begin{aligned} & m \in \mathbb{Z}_+,\ t_1,\dots,t_m \in T,\ r \in \mathbb{Z},\ \text{such that, } t_1+r,\dots,t_m+r \in T, \\ y_1 &= (x(t_1),\dots,x(t_m))^T,\ \ y_2 = (x(t_1+r),\dots,x(t_m+r))^T, \\ Q_1 &= E[(y_1 - E[y_1])(y_1 - E[y_1])^T], \\ Q_2 &= E[(y_2 - E[y_2])(y_2 - E[y_2])^T]. \end{aligned}$$

Then x is stationary if and only if y_1, y_2 have the same distribution if and only if $E[y_1] = E[y_2]$ and $Q_1 = Q_2$, because x is a Gaussian process; if and only if for all $i, j \in \mathbb{Z}_k$, $m_x(t_i) = m_x(t_i + r)$, $W_x(t_{i,}t_j) = W_x(t_i + r, t_j + r)$. □

Definition 3.4.5. Of a stationary Gaussian process $x : \Omega \times T \to \mathbb{R}^{n_x}$ define respectively its mean value, its covariance function, and its variance, as

$$\begin{aligned} m_x &= m_x(t) \in \mathbb{R}^{n_x},\ \forall\, t \in T;\ \ W_x : T \to \mathbb{R}^{n_x \times n_x}, \\ W_x(t) &= W(t,0) = W(t+s,s),\ \ \forall\, s \in T,\ \text{such that } t+s \in T; \\ Q_x &= W_x(0) = W_x(0,0) \in \mathbb{R}^{n_x \times n_x}_{pds};\ \ \text{then, } x(t) \in G(0,Q_x),\ \forall\, t \in T. \end{aligned}$$

It follows from Proposition 3.4.4.(2) that the above new covariance function is well defined. When considering the covariance function of a stationary Gaussian process then it will from now on depend on only one time argument. Because the process is stationary, for all time $t \in T$, $W_x(t,t) = W_x(0,0) = Q_x$ which does not depend on time.

Proposition 3.4.6. *The function $W : T \to \mathbb{R}^{n \times n}$ is the covariance function of a stationary Gaussian process if and only if:*

(1)W is para-symmetric*; or, equivalently, if for all $t \in T$, $W(t) = W(-t)^T$;*
(2)it is positive definite:

$$\forall\, m \in \mathbb{Z}_+,\ \forall\, t_1,\dots,t_m \in T,\ \forall\, c_1,\dots,c_m \in \mathbb{R}^n,\ \sum_{i=1}^{m}\sum_{j=1}^{m} c_i^T W(t_i - t_j) c_j \geq 0.$$

Proof. This follows directly from Proposition 3.4.3 and the definitions. □

Proposition 3.4.7. Representation of a stationary Gaussian white noise process. *Consider a stationary Gaussian white noise process $w : \Omega \times T \to \mathbb{R}^{n_w}$ with $n_w \in \mathbb{Z}_+$ with zero mean value. There exists a standard Gaussian white noise process $v : \Omega \times T \to \mathbb{R}^{n_v}$ with $n_v \in \mathbb{N}$ and a matrix $M \in \mathbb{R}^{n_w \times n_v}$ such that for all $t \in T$, $w(t) = Mv(t)$ a.s.*

Proof. Because the Gaussian white noise process w is stationary, it is true that $Q_w(t,t) = Q_w(t) = Q_w(0)$ for all $t \in T$. For $t = 0 \in T$, it follows from Proposition 2.7.5 that there exists a matrix $M \in \mathbb{R}^{n_w \times n_v}$ and an integer $n_v \in \mathbb{Z}_+$ such that $w(0) = Mv(0)$ for a Gaussian random variable $v(0) : \Omega \to \mathbb{R}^{n_v}$, $v(0) \in G(0,I)$ hence $Q_w(0) = MM^T$. Then for all $t \in T$, $Q_w(t,t) = Q_w(0) = MM^T$ and, similarly as above, $w(t) = Mv(t)$. Because w is a Gaussian white noise process, so is $v : \Omega \times T \to \mathbb{R}^{n_v}$ while, due to the construction, $v(t) \in G(0,I)$, hence v is a standard Gaussian white noise process. □

Proposition 3.4.8. *Consider a stationary Gaussian process taking values in $(\mathbb{R}^n, B(\mathbb{R}^n))$, with mean-value function equal to zero, and with covariance function $W : T \rightarrow \mathbb{R}^{n\times n}$.*

Then this process is time-reversible if and only if, for all $t \in T$, $W(t) = W(-t)$; or, equivalently, if for all $t \in T$, $W(t) = W(t)^T$. A scalar stationary Gaussian process, thus with $n = 1$, is always time-reversible.

Proof. The process is time-reversible if and only if for all $m \in \mathbb{Z}_+$, $t_1, \ldots, t_m, t \in T$

$$
\begin{aligned}
&E[\exp(i\sum_{j=1}^{m} w_j^T x(t_j))] = E[\exp(i\sum_{j=1}^{m} w_j^T x(t-t_j))]\\
&\Leftrightarrow \exp(-\frac{1}{2} \sum_{j=1}^{m}\sum_{k=1}^{m} w_j^T W(t_j - t_k) w_k)\\
&\quad = \exp(-\frac{1}{2}\sum_{j=1}^{m}\sum_{k=1}^{m} w_j^T W((t-t_j)-(t-t_k)) w_k),\ \forall\, w_1, \ldots, w_m \in \mathbb{R}^n;\\
&\Leftrightarrow W(t_j - t_k) = W(t_k - t_j).
\end{aligned}
$$

The last part of the proposition follows from the first part and Proposition 3.4.3. □

Gauss-Markov Processes

Definition 3.4.9. A stochastic process is called a *Gauss-Markov* process if it is both a Gaussian process and a Markov process.

Proposition 3.4.10. *Let $x : \Omega \times T \rightarrow \mathbb{R}^{n_x}$ be a Gaussian process with covariance function $W_x : T \times T \rightarrow \mathbb{R}^{n_x \times n_x}$. Assume that for all $t \in T$, $W_x(t,t) \succ 0$.*

Then the process x is a Markov process if and only if,

$$W_x(t,s) = W_x(t,u)W_x(u,u)^{-1}W_x(u,s),\ \ \forall\, t,\ s,\ u \in T,\ \textit{such that } s < u < t.$$

Proof. Without loss of generality one may suppose that $E[x(t)] = 0$ for all $t \in T$. $(\Rightarrow)$ Let $t, s, u \in T, s < u < t$. Then

$$
\begin{aligned}
W_x(t,s) &= E[x(t)x(s)^T] = E[E[x(t)x(s)^T|F_u^x]], \text{ because of Theorem 2.8.2.(e)},\\
&= E[E[x(t)|F_u^x]x(s)^T], \text{ because of Theorem 2.8.2.(d) and } s < u,\\
&= E[E[x(t)|F^{x(u)}]x(s)^T], \text{ because } x \text{ is Markov},\\
&= E[W_x(t,u)W_x(u,u)^{-1}x(u)x(s)^T], \text{ because } (x(t),x(s)) \text{ is Gaussian},\\
&= W_x(t,u)W_x(u,u)^{-1}W_x(u,s).
\end{aligned}
$$

$(\Leftarrow)$ Let $m \in \mathbb{Z}_+$, $s_1, \ldots, s_m, t \in T$, $s_1 < s_2 < \ldots < s_m < t$, and $w \in \mathbb{R}^{n_x}$. Let $F_m = \sigma(\{x(s_1), \ldots, x(s_m)\})$. Because x is a Gaussian process, $(x(s_1), \ldots, x(s_m), x_t)$ is jointly Gaussian. It follows then from Theorem 2.8.3 that,

$$E[\exp(iw^T x(t))|F_m] = \exp(iw^T E[x(t)|F_m] - \frac{1}{2} w^T Q w),$$

with $Q = E[(x(t) - E[x(t)|F_m])(x(t) - E[x(t)|F_m])^T]$. Then,

$$\begin{aligned}
& E[x_t|F_m] \\
&= (W_x(t,s_1)\ldots W_x(t,s_m)) \begin{pmatrix} W_x(s_1,s_1) & \ldots & W_x(s_1,s_m) \\ \vdots & \vdots & \vdots \\ W_x(s_m,s_1) & \ldots & W_x(s_m,s_m) \end{pmatrix}^{-1} \begin{pmatrix} x(s_1) \\ \vdots \\ x(s_m) \end{pmatrix} \\
&= W_x(t,s_m) W_x(s_m,s_m)^{-1} (W_x(s_m,s_1)\ldots W_x(s_m,s_m))(*)^{-1} \begin{pmatrix} x(s_1) \\ \vdots \\ x(s_m) \end{pmatrix}, \text{ by assumption,} \\
&= W_x(t,s_m) W_x(s_m,s_m)^{-1} x(s_m) = E[x(t)|F^{x(s_m)}].
\end{aligned}$$

In case the matrix whose inverse is used above does not have an inverse, the result still holds but the proof has to be modified. Then,

$$\begin{aligned}
& E[\exp(iw^T x(t))|F_m] \\
&= \exp(iw^T E[x(t)|F_m] - \frac{1}{2} w^T E[(x(t) - E[x(t)|F_m])(x(t) - E[x(t)|F_m])^T]w) \\
&= \exp(iw^T E[x(t)|F^{x(s_m)}]) \times \\
&\quad \times \exp(-\frac{1}{2} w^T E[(x(t) - E[x(t)|F^{x(s_m)}])(x(t) - E[x(t)|F^{x(s_m)}])^T]w) \\
&= E[\exp(iw^T x(t))|F^{x(s_m)}],
\end{aligned}$$

and the result follows from Proposition 3.3.7. □

Proposition 3.4.11. *Let $x : \Omega \times T \to \mathbb{R}^{n_x}$ be a Gaussian process with $T = N$, $x(t) \in G(0, Q_x(t))$ and covariance function $W_x : T \times T \to \mathbb{R}^{n_x \times n_x}$. Assume that for all $t \in T$, $Q_x(t) \succ 0$.*

The following statements are equivalent:

(a)The process x is a Markov process.
(b)$E[x(t)|F_s^x] = E[x(t)|F^{x(s)}]$, for all $s,\ t \in T$ with $s < t$.
(c)$E[x(t+1)|F_t^x] = E[x(t+1)|F^{x(t)}]$, for all $t \in T$.
(d)The covariance function W_x satisfies

$$W_x(t,s) = W_x(t,u) W_x(u,u)^{-1} W_x(u,s),\ \forall\ s,u,t \in T \text{ such that } s < u < t.$$

(e)The process x satisfies the recursion

$$x(t+1) = A(t)x(t) + M(t)v(t),\ x(0) = x_0,$$

where $x_0 : \Omega \to \mathbb{R}^{n_x}$, $x_0 \in G(0, Q_x(0))$, and $Q_x(0) \succ 0$, $v : \Omega \times T \to \mathbb{R}^{n_v}$ is a standard Gaussian white noise process, F^{x_0} and F_∞^v are independent, $A : T \to \mathbb{R}^{n_x \times n_x}$, and $M : T \to \mathbb{R}^{n_x \times n_v}$.

(f) If in addition the Gaussian process is stationary and the assumptions of Def. 3.4.5 hold, then the process satisfies the recursion,

$$x(t+1) = Ax(t) + Mv(t),\ x(0) = x_0,$$

where $x_0 : \Omega \rightarrow \mathbb{R}^{n_x}$, $x_0 \in G(0, Q_{x_0})$, $Q_{x_0} \succ 0$, $v : \Omega \times T \rightarrow \mathbb{R}^{n_v}$ for $n_v \in \mathbb{N}$, is a standard stationary Gaussian white noise process satisfying that F^{x_0} and F^v_∞ are independent, $v(t) \in G(0, I_{n_v})$, $A \in \mathbb{R}^{n_x \times n_x}$, and $M \in \mathbb{R}^{n_x \times n_v}$.

Proof. (a) ⇒ (b). See Proposition 3.3.7. (b) ⇒ (c). Take $s = t-1$ in (b).
(c) ⇒ (b). This follows by an induction argument.
(b) ⇒ (d) and (d) ⇒ (a). See Proposition 3.4.10.
(a) ⇒ (e). Let $t \in T$. Then

$$E[x(t+1)|F^x_t] = E[x(t+1)|F^{x(t)}],\ \text{by (c)},\ = A(t)x(t),$$

because x is a Gaussian process and by Theorem 2.8.3. Define the process $w : \Omega \times T \rightarrow \mathbb{R}^{n_x}$, $w(t) = x(t+1) - A(t)x(t)$. Then $(x(t+1), x(t), w(t)) \in G$. Let $F_t = F^w_{t-1} \vee F^{x_0}$. then $F_t \subseteq F^x_t$ for all $t \in T$, and

$$\begin{aligned}
&E[\exp(iu^T w(t))|F_t] = E[E[\exp(iu^T w(t))|F^x_t]|F_t] \\
&= E[E[\exp(iu^T w(t))|F^{x(t)}]|F_t],\ \text{by reconditioning and by (a)}, \\
&= \exp(-\frac{1}{2}u^T Q_w(t)u) \\
&\quad \text{by } (w(t+1), x(t)) \in G,\ E[w(t)|F^x_t] = E[x(t+1)|F^x_t] - A(t)x(t) = 0, \\
&= E[\exp(iu^T w(t))].
\end{aligned}$$

From Theorem 2.8.2 follows that $w(t)$ is independent of $F_t = F^{x_0} \vee F^w_{t-1}$. Hence w is a white noise process and by the above calculation it is Gaussian white noise. From Proposition 3.4.7 follows that there exists a standard stationary Gaussian white noise process $v : \Omega \rightarrow \mathbb{R}^{n_v}$ and a matrix $M : \rightarrow \mathbb{R}^{n_v}$ such that $w(t) = Mv(t)$ a.s. for all $t \in T$.

(e) ⇒ (c). Let $t \in T$ and $F_t = F^{x_0} \vee F^v_{t-1}$. By induction it may be proven that $F^x_t \subseteq F_t$. Because x_0 and v are independent, and, because v is Gaussian white noise, it follows that $v(t+1)$ is independent of F_t. By assumption, $E[v(t)] = 0$. Then,

$$\begin{aligned}
&E[x(t+1)|F^x_t] \\
&= E[A(t)x(t) + M(t)v(t)|F^x_t] \\
&= A(t)x(t) + M(t)E[E[v(t)|F_{t-1}]|F^x_t] = A(t)x(t) + 0 = E[A(t)x(t)|F^{x(t)}] \\
&= E[E[x(t+1)|F^x_t]|F^{x(t)}] = E[x(t+1)|F^{x(t)}].
\end{aligned}$$

(f) (⇒) (e) This follows directly from (e) and the property of stationarity.
(e) (⇒) (f) Note that,

$$\begin{aligned}
&E[x(t+1)|F^{x(t)}] = Ax(t), \\
w(t) &= x(t+1) - Ax(t),\ w(t) \in G(0, Q_w), \\
&\exists\ M \in \mathbb{R}^{n_x \times n_v},\ v : \Omega \times T \rightarrow \mathbb{R}^{n_v},\ v(t) \in G(0, I),\ Q_w = MM^T, \\
w(t) &= Mv(t), \\
x(t+1) &= Ax(t) + Mv(t).
\end{aligned}$$

□

Theorem 3.4.12. ***Characterization and classification of a time-reversible stationary Gauss-Markov process****. Consider a stochastic process $x : \Omega \times T \to \mathbb{R}^n$ on $T = \mathbb{Z}$ which is Gaussian, stationary, and has a mean value function which is identically zero.*

(a)If the covariance function admits the representation,

$$W(t) = (LD_{+-}L^{-1})^t W(0),\ \forall\, t \in T,\ \textit{where},$$
$$D_{+-} = \begin{pmatrix} I_{n_1} & 0 \\ 0 & -I_{n_2} \end{pmatrix},\ n_1,\ n_2 \in \mathbb{N},\ n_1 + n_2 = n,$$
$$W(0),\ L \in \mathbb{R}^{n \times n},\ \textit{both nonsingular matrices,}$$

then the process is both a Markov process and a time-reversible process.

(b)If the Gaussian process is Markov and time-reversible then the covariance function admits a representation as described in part (a).

In case the representation of the Gauss-Markov process is such that the matrix $L = I$, then the process x consists of two components, the first component x_{n_1} of dimension $n_1 \in \mathbb{N}_n$ and the second component x_{n_2} of dimension $n_2 \in \mathbb{N}_n$. The first component is such that $W_{x_{n_1}}(t) = W_{x_{n_1}}(0)$ for all $t \in T$ hence is called the even component *and the second component x_{n_2} is such that $W_{x_{n_2}}(t) = (-1)^t W_{x_{n_2}}(0)$ and is called the* odd component *of the process.*

The proof of Theorem 3.4.12 is simple and listed as an exercise.

3.5 Finite-Valued Stochastic Processes

In stochastic control theory, finite-valued processes are also used. The description of finite-valued processes is more complex than that of Gaussian processes. Limited use is made in this section of concepts of positive matrices, see Chapter 18.

Definition 3.5.1. A *finite-valued stochastic process* is a stochastic process as defined in Def. 3.1.1 with the sets and maps,

$$(\Omega, F, P),\ T = \{0, 1, \ldots, t_1\} = T(0 : t_1),\ \text{or}\ T = \mathbb{N} = \{0, 1, \ldots\},$$
$$x : \Omega \times T \to X = \mathbb{Z}_{n_x} = \{1, 2, \ldots, n_x\},\ n_x \in \mathbb{Z}_+.$$

If the process takes values in a different space than indicated above, then the range space can be redefined such that the above representation is obtained.

Rather than with the notation defined above, one prefers to work with the indicator representation of finite-valued random variables formulated in Def. 2.5.9. Thus,

$$
\begin{aligned}
x(t) = Cx_I(t),\ C = \begin{pmatrix} 1 \\ 2 \\ \vdots \\ n_x \end{pmatrix},\ X_e = \{e_1,\ e_2,\ \ldots,\ e_{n_x}\} \subset \mathbb{R}^{n_x},\\
x_I : \Omega \times T \to X_e,\ x_{I,i}(\omega) = I_{\{x(\omega,t)=i\}}(\omega),\ \forall\, i \in \mathbb{Z}_{n_x},\\
y = f(x(t)) = \sum_{i=1}^{n_x} f(i) I_{\{x(t)=i\}} = f_I^T x_I(t),\\
\forall\, f : X_e \to \mathbb{R},\ f_I = \big(f(1)\ f(2)\ \ldots\ f(n_x)\big)^T \in \mathbb{R}^{n_x}.
\end{aligned}
$$

Thus the process x_I takes values in the set X_e of unit vectors of $\mathbb{R}^{n_x}$, $X_e = \{e_1, e_2, \ldots, e_{n_x}\}$. From now onwards, the indicator representation will be used for a finite-valued stochastic process.

The probability measure of the random variable $x(t)$ for a time $t \in T$, is denoted by $p_{x(t)} \in \mathbb{R}_{st}^{n_x}$. The joint probability distribution of $x(t_1)$ and $x(t_2)$ is then denoted by,

$$p_{x(t_1),x(t_2)} \in \mathbb{R}_+^{n_x \times n_x},\ p_{x(t_1),x(t_2)}(i_1,\ i_2) = P(\{x(\omega,t_1) = i_1,\ x(\omega,t_2) = i_2\}),$$

Note that the vector $p_{x(t_1),x(t_2)}$ has to satisfy conditions not stated here. The joint distribution of the entire finite-valued process on the time axis $T(0:t_1)$ is thus a vector,

$$p_{x(0),\ \ldots,\ x(t_1)} \in \mathbb{R}_+^{n_x^{(t_1+1)}}.$$

Practically it is not simple to specify such a vector. Hence the interest to investigate particular finite-valued stochastic processes.

If the finite-valued stochastic processs x is such that $\{x(t) \in \mathbb{R}^{n_x},\ t \in T\}$ is a sequence of independent random variables then the probability distributions factorize according to,

$$
\begin{aligned}
p_{x(t_1),x(t_2)} &= p_{x(t_1)} p_{x(t_2)}^T,\\
p_{x(0),\ldots,x(t_1)}(i_0, i_1, \ldots i_{t_1}) &= p_{x(t_1)}(i_1) p_{x(t_2)}(i_2) \ldots p_{x(t_1)}(i_{t_1}) = \prod_{s=0}^{t_1} p_{x(s)}(i_s).
\end{aligned}
$$

If a finite-valued process is also a Markov process then a simple description of the process can be formulated.

Definition 3.5.2. Define a *finite-valued Markov process* by the probability distribution $p_{x_0} \in \mathbb{R}_{st}^{n_x}$ of the initial state x_0 and the transition probability denoted by,

$$
\begin{aligned}
p_{x(t+1)|x(t)}(i) &= E[x(t+1) = e_i | F_t^x] = E[x(t+1) = e_i | F^{x(t)}] = (A(t)x(t))_i,\\
p_{x(t+1)|x(t)} &= A(t)x(t),\ \ x : \Omega \times T \to X_e = \{e_1, e_2, \ldots, e_{n_x}\} \subset \mathbb{R}^{n_x},\\
& \quad A : T \to \mathbb{R}_{st}^{n_x \times n_x}, x_0 : \Omega \to X_e.
\end{aligned}
$$

Call the function A the *transition matrix function* of the Markov process. Call the transition matrix function *time invariant* if the matrix function A does not depend on time; equivalently, $A(t) = A(0)$ for all $t \in T$. Call then $A = A(0) \in \mathbb{R}_{st}^{n_x \times n_x}$ the *transition matrix* of the Markov process.

It follows directly from the relation $p_{x(t+1)|x(t)} = A(t)x(t)$ and from Theorem 3.3.7 that the process x of Def. 3.5.2 is a Markov process.

Example 3.5.3. *Binary communication channel.* A *communication channel* is used as a model of a hardwired channel in communication engineering and in information theory. A communication channel can be *memoryless* or have *memory* depending on whether it consists of a probabilistic map or of a stochastic control system as defined in Chapter 10. A channel is called *binary* if the input set is the binary set $X_2 = \{0,1\}$ and likewise the output set is the same binary set with a probabilistic map between these spaces as defined below.

Consider a binary communication channel with the representation,

$$\begin{aligned}
& U_c = \{0,1\},\ Y_c = \{0,1\}, \\
& f_c : U_c \to P(Y_c),\ p_{y|u} = f_c(u) = A_c u, \\
& p_{y|u=0} = A_{c,0,0} I_{\{u=0\}} + A_{c,0,1} I_{\{u=1\}},\ p_{y|u=1} = A_{c,1,0} I_{\{u=0\}} + A_{c,1,1} I_{\{u=1\}}, \\
A_c = & \begin{pmatrix} p_{0,0} & p_{0,1} \\ p_{1,0} & p_{1,1} \end{pmatrix} \in \mathbb{R}_{st}^{2\times 2}, \text{ hence } p_{0,0} + p_{1,0} = 1,\ p_{0,1} + p_{1,1} = 1,
\end{aligned}$$

The operation of the binary communication channel is thus that for $u = 0$ the channel produces a probability distribution on the output set Y_c of the form,

$$\begin{aligned}
p_{y|u} &= \begin{pmatrix} p_{y,0} \\ p_{y,1} \end{pmatrix} = A_c \begin{pmatrix} 1 \\ 0 \end{pmatrix};\ \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} I_{\{u=0\}} \\ I_{\{u=1\}} \end{pmatrix}, \\
p_y &= A_c p_u.
\end{aligned}$$

A *symmetric binary channel* is defined as a binary channel such that $p_{0,1} = p_{1,0}$. Then it follows from the above relations that $p_{0,0} = p_{1,1}$.

Example 3.5.4. *Model of a communication buffer.* At every node of a communication network there are buffers for packets or messages. For a communication channel from station A to station B there is a buffer of packets waiting to be sent. A communication channel may cary much traffic in the form of packets and the new packets have to be temporarily buffered before they can be sent via the communication channel.

Note that the arrivals of packets is irregular because the decision to send packets is made by humans or by machines in an uncoordinated way. One says that the *arrival rate* of packets is varying over time. However, the output of the buffer to the communication channel is usually at a fixed *output rate*. Because of the difference between the arrival rate and the departure rate, a buffer is needed. Such a model is also used in manufacturing systems, in ware houses for physical packets to be delivered, in shops for goods for customers, etc.

A model for a communication buffer can be formulated as a Markov process. Because any physical buffer has a finite number of waiting places, the size of the buffer is assumed to be finite. Below the buffer size is assumed to be equal to $n_x = 2$

to keep the example simple. The arrival process is assumed to be a sequence of independent random variables with a constant rate. The departure process is, due to the service of the communication line, assumed to have a constant deterministic rate if there is at least one packet present. The operations of the buffer are such that during a time step, first a packet is removed from the buffer to be sent, and then zero, or one, or two packets are put into the buffer depending on the arrivals.

A particular model of a communication buffer is then,

$$\begin{aligned} & T = \mathbb{N},\ X = \mathbb{N}_2 = \{0,1,2\},\ x : \Omega \times T \to \mathbb{R}^3, \\ p_{x_0} &= \begin{pmatrix} 0.7\ 0.2\ 0.1 \end{pmatrix}^T, \\ p_{x(t+1)|x(t)} &= E[x(t+1)|\ F^{x(t)}] = Ax(t),\ \ p_{x(t+1)} = Ap_{x(t)} = A^t p_{x_0}, \\ A &= \begin{pmatrix} 0.2\ 0.1\ 0.1 \\ 0.6\ 0.5\ 0.3 \\ 0.2\ 0.4\ 0.6 \end{pmatrix} \in \mathbb{R}_{st}^{3\times 3}, \end{aligned}$$

Then the state process x is a stationary Markov chain as proven below.

A model of a communication buffer can be used to analyse the required buffer size for a communication switch, or to evalutate the performance of a call center including the number of calls which could not be answered due to buffer overflow, etc.

Proposition 3.5.5. *Consider a finite-valued Markov process with a time-invariant transition function. Assume that the state transition matrix A is irreducible and non-periodic.*

(a)Then there exists a unique vector $p_s \in \mathbb{R}_{st}^{n_x}$ which is the solution of the steady state equation,

$$p_s = Ap_s,\ p_s \in \mathbb{R}_{st}^{n_x}.$$

(b)If the probability distribution of the initial state x_0 equals the vector p_s defined in (a) and if the transition matrix function is time-invariant then the probability measure p_s is the invariant measure of the Markov process; equivalently, for all $t \in T$, $p_{x(t)} = p_s$. In this case the process is a stationary Markov process.

Proof. (a) This follows from Theorem 18.8.7.(a).
(b) Note that by the result of (a), for any $t \in T$, $p_{x(t)} = A^{t+1} p_s = A^t(Ap_s) = A^t p_s = p_s$, hence p_s is the invariant measure of the Markov process.

The joint probability distribution function of any tuple $(x(s_1), \ldots x(s_k))$ for $s_1, \ldots s_k \in T$ with $s_1 < s_2 < \ldots < s_k$, can be written as a product of the elements of the vector p_{x_0} and of the elements of the matrix A in which only the differences of the type $s_{i+1} - s_i$ for $i = 1, \ldots, k-1$ occur. The result then follows from the definition of a stationary process. □

Proposition 3.5.6. *Consider a finite-valued stationary Markov process specified by p_{x_0} with the transition matrix A. Assume that the transition matrix A is irreducible.*

The process is time-reversible if and only if the transition matrix is symmetric upto diagonal scaling with the invariant probability distribution; equivalently,

$$A = D(p_x)A^T D(p_x)^{-1}.$$

3.6 Exercises

Problem 3.6.1. *Prediction of the state of a Gaussian state process.* Consider the system representation,

$$\begin{aligned}
x(t+1) = A(t)x(t) + M(t)v(t),\ & x(t_0) = x_0,\\
& T = \{t_0, t_0+1, \ldots,\} \subset \mathbb{Z}_+,\ x_0 : \Omega \times \mathbb{R}^{n_x},\ x_0 \in G(0, Q_0),\\
& v : \Omega \times T \to \mathbb{R}^{n_v},\ \text{is a standard Gaussian white noise process},\\
& v(t) \in G(0, I),\ \forall t \in T,\ F^{x_0},\ F^v_\infty \text{ are independent } \sigma\text{-algebras}\\
& A:\ T \to \mathbb{R}^{n_x \times n_x},\ M:\ T \to \mathbb{R}^{n_x \times n_v};\ \text{assume that } \forall\, t \in T,\\
& Q_x(t) = E[(x(t) - E[x(t)])(x(t) - E[x(t)])^T] \succ 0.
\end{aligned}$$

Derive a formula for

$$E[x(t+2)|F^x_t],\ \forall\, t \in \mathbb{Z}_+,\ \text{where } F^x_t = \sigma(\{x(s),\ s \leq t\}).$$

Problem 3.6.2. *Characterization of a time-reversible stationary Gauss-Markov process.* Write out the proof of Theorem 3.4.12.

3.7 Further Reading

History. Initially, what is now called a stochastic process on a totally-ordered time index set, was a sequence of random variables. The measure-theoretic treatment of stochastic processes is due to A.N. Kolmogorov with his book, see the English translation [17], where the theory stochastic processes on the time axis of the real numbers is treated.

Stochastic processes in engineering. The interest in stochastic processes in electrical engineering started early in the twentieth century with the operation of radio tubes, of communication equipment, and of radar signals. The problems concerned primarily filtering which is the estimation of a signal which is corrupted by a noise signal. Early books on stochastic processes for engineers include [12, 24]. N. Wiener and colleagues at MIT developed prediction etc. for second-order processes, [25, 26]. Later on the books of [11] and of E. Wong [27] were published for engineers. In Russia there were the books of A.M. Yaglom, [29, 30].

Stochastic processes in mathematics. In the USA, the basic textbook was that of J.L. Doob, [6]. Also, the book of M. Loève, [18] is known. In France, P. Lévy was a major author. Later on the developments of martingale theory and stochastic integrals received much interest.

Books at an elementary level. The author recommends the book of Bruce Hajek, [8]. See also [4, 13, 20].

Books at an advanced level. For references on the theory of stochastic processes at the level of this course see [27, 28]. Other references at an advanced level and with a broad emphasis are [3, 6].

For the construction of a probability measure on the range space of a sequence of random variables or on stochastic processes defined on the real numbers, see [17], [2, Ch. 20], and [3, Section 2.4].

The time-reversibility of a stochastic process and the characterization of time-reversibility of a state-finite Markov process are both due to A. Kolmogorov, [16]; the text of that paper is in German. Time-reversibility of a finite-valued stationary Markov process is an important tool in the analysis of the operation of communication networks, see [14].

Markov processes. For Markov processes see [7, 15, 19, 22].

Gaussian processes. An early reference on Gaussian processes is the paper [5]. Books on Gaussian processes are [10, 21]. References on the spectral approach to second-order stochastic processes are [1, 9, 23].

References

1. R.B. Ash and M.F. Gardner. *Topics in stochastic processes*. Academic Press, New York, 1975. 73
2. P. Billingsley. *Probability and measure (Third Edition)*. John Wiley & Sons, New York, 1995. 49, 73
3. L. Breiman. *Probability*. Addison-Wesley Publ. Co., Reading, MA, 1968. 49, 73, 741, 758
4. K.L. Chung. *Elementary probability with stochastic processes*. Springer-Verlag, Berlin, 1975. 49, 73
5. J.L. Doob. The elementary Gaussian processes. *Ann. Math. Statist.*, 15:229–282, 1944. 73
6. J.L. Doob. *Stochastic processes*. Wiley, New York, 1953. 72, 73, 721, 747, 754, 758
7. E.B. Dynkin. *Markov processes, volume 1, volume 2*. Academic Press Inc., Publishers, New York, 1965. 73, 574
8. Bruce Hajek. *Random processes for engineers*. Cambridge University Press, Cambridge, 2015. 49, 73
9. E.J. Hannan. *Multiple time series*. Wiley, New York, 1970. 73, 120
10. I.A. Ibragimov and Y.A. Rozanov. *Gaussian random processes*. Springer-Verlag, Berlin, 1978. 73
11. W.B. Davenport Jr. *Probability and random processes*. McGraw-Hill Book Co., New York, 1970. 72
12. W.B. Davenport Jr. and W.L. Root. *An introduction to the theory of random signals and noise*. McGraw-Hill Book Co., New York, 1958. 9, 72, 310
13. S. Karlin and H.M. Taylor. *A first course in stochastic processes, 2nd edition*. Academic Press, New York, 1975. 73
14. F.P. Kelly. *Reversibility and stochastic networks*. John Wiley & Sons, Chichester, 1979. 73, 169, 468
15. J.G. Kemeny, J.L. Snell, and A.W. Knapp. *Denumerable Markov chains*. Springer-Verlag, Berlin, 1976. 73, 697

16. A. Kolmogorov. Zur Theorie der Markoffschen Ketten. *Berliner Berichte*, 144:155–160, 1931. 73
17. A.N. Kolmogorov. *Foundations of probability (translation)*. Chelsea, New York, 1950. 49, 72, 73
18. M. Loève. *Probability theory, 3rd edition*. Van Nostrand Reinhold Co. Inc., New York, 1963. 49, 72, 742
19. P.A. Meyer. *Processus de Markov*, volume 26 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1967. 49, 73, 341
20. R.E. Mortensen. *Random signals and systems*. Wiley, New York, 1987. 73
21. J. Neveu. *Processus aléatoires Gaussiens*. Presses Universitaires Montréal, Montréal, 1968. 49, 73
22. D. Revuz. *Markov chains*. North-Holland Publ. Co., Amsterdam, 1975. 73
23. Yu.A. Rozanov. *Stationary random processes*. Holden-Day, San Francisco, 1967. 73
24. H.L. van Trees. *Detection, estimation, and modulation theory*. Wiley, 1968. 9, 72
25. N. Wiener. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. Technology Press of the MIT, Cambridge, MA, U.S.A., 1949. MIT Press, Cambridge, MA, U.S.A., 1966. 72, 310
26. N. Wiener and P. Masani. The prediction theory of multivariate stochastic processes – part ii. *Acta Math.*, 99:93–137, 1958. 72, 310
27. E. Wong. *Stochastic processes in information and dynamical systems*. McGraw-Hill Book Co., New York, 1971. 72, 73, 310, 352
28. E. Wong and B. Hajek. *Stochastic processes in engineering systems*. Springer-Verlag, Berlin, 1985. 73, 310, 352
29. A.M. Yaglom. *An introduction to the theory of stationary random functions*. Prentice-Hall, Englewood Cliffs, 1962. 72
30. A.M. Yaglom. *Correlation theory of stationary and related random functions I*. Springer-Verlag, Berlin, 1987. 72

Chapter 4
Gaussian Stochastic Systems

Abstract A stochastic system (without input) is a mathematical model of a dynamic phenomenon exhibiting uncertain signals. Such a system is mathematically characterized by the transition map from the current state to the joint probability distribution of the next state and the current output. In this chapter are formulated only Gaussian systems. Forward and backward Gaussian stochastic systems are defined and related. Stochastic observability and stochastic co-observability are defined and characterized.

Key words: Stochastic system. Gaussian system. Stochastic observability.

A reader not familiar with the subject of stochastic systems may in a first reading focus attention on Section 4.3.

4.1 Modeling of Phenomena as a Stochastic System

In this section it is described how to go from a concrete dynamic phenomenon to a mathematical structure in the form of a stochastic system representation. The stochastic system formulated should be a realistic model of the phenomenon. The modelling procedure will be illustrated by an example.

Example 4.1.1. *Control of a paper machine.*
An example of stochastic control is that of control of a paper machine. The project was carried out by K.J. Aström in the early 1960's. The aim of that investigation has been to show how computers can be used to control a paper machine and to establish a design procedure for control of industrial processes. This example is a prototype of a quality control problem as they appear in the process industry.

The paper machine under consideration belonged to the Billerud Kraft Paper Mill at Gruvön in Sweden. A simplified diagram of the paper machine is shown in Figure 4.1. Only those parts are displayed which are of interest to basis weight control.

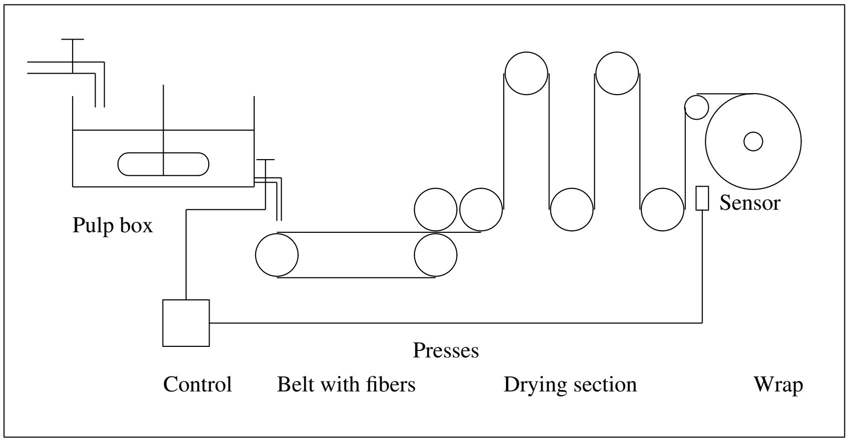


Fig. 4.1 Diagram of paper machine, Example 4.1.1.

The thick stock, a water fibre mixture with a fibre concentration of about 3% comes from the machine chest. The thick stock is diluted with white water so that the headstock concentration is reduced to between 0.2 and 0.5%. On a special moving underground, the fibres are separated from the water and a web is formed. Water is pressed out of the paper by presses and the paper is then dried on steam-heated cylinders in the dryer section.

In this plant it is possible to influence the basis weight by varying the thick stock flow and the thick stock consistency.

Measurements of the basis weight are taken by a beta-ray gauge set at a fixed position at the dry end. The output of this instrument will be proportional to the mass of fibres and water per unit area, which will be called the *wet basis weight.* In order to obtain the *dry basis weight,* that is the mass of fibres per unit area, the beta-ray gauge reading has to be compensated for the moisture in the paper sheet. Moisture is measured by a capacitance gauge. An estimate of the dry basis weight is provided by the formula $y(t) = WSP(t)(1 - MSP(t))$, where WSP is the calibrated beta-ray gauge signal and MSP is the signal from the moisture gauge.

The investigation resulted in control of basis weight by a control law based on feedback from the measurements at the dry end of the machine to thick stock flow. An increase of thick stock flow or thick stock consistency results in an increase in moisture content as well as basis weight. A change of steam pressure in the drying section will however influence the moisture content of the paper but not the dry basis weight.

Control has been applied to the mill during normal operations, of course with permission of the operators of the plant. Specifically, the control objective is to control the variance of the dry basis weight. With a reduced variance the set point of the control may be placed closer to the specified limit which reduces the need for raw material and power consumption and hence increases productivity.

Before computer control, the weight had a standard deviation of 1.3 g/m^2. The initial goal of the investigation was to reduce this to 0.7 g/m^2. With computer control, in continuous operation since the beginning of 1966, the standard deviations for wet basis weight and dry basis weight at the end of the investigation were respectively 0.5 g/m^2 and 0.3 g/m^2.

The phases of the investigation are briefly described. Let y represent the variable of dry basis weight, and u the thick stock flow. The dynamics of the process is not explicitly modelled. The physical dynamics consists primarily of that of the drying process and that of the delay caused by the transportation of the paper from the beginning of the belt with fibers to the sensor at the end of the drying section. The set of dynamical systems with which to model the plant is chosen to be an *autoregressive moving average representation with an external input* (ARMAX) of the form,

$$y(t) = \sum_{i=1}^{n} a_i y(t-i) + \sum_{i=0}^{n} b_i u(t-i-k) + \lambda \sum_{i=0}^{n} c_i v(t-i), \tag{4.1}$$

where v is a Gaussian white noise process, $c_0 = 1$, and k is a delay to account for the time lag between a control action at the thick stock control valve and its effect at the measurement station. The uncertainty in the paper production process is modelled by a Gaussian white noise process appearing in the third term in the right-hand side of equation (4.1).

Experiments with the paper mill had to take place during normal operation. A pseudo random binary signal was used to generate the input data. A maximum likelihood approach has been used to estimate the parameters $n, \{a_i, b_i, c_i, i \in \mathbb{Z}_n\}$, and λ of the stochastic system.

The control problem at the level of the physical model is then to determine a control program for a computer so that the variance of the dry basis weight is as small as possible. The control problem at the level of control theory is then to determine a control law so that the variance of a function of the state process is as small as possible. Such a control law has been determined using the theory of minimum variance control. This stochastic control problem will be discussed later in this course.

A brief summary of the system identification phase and the control phase of the investigation follows. Since the output of the plant was found to drift slowly the external variables were changed from y, u to $\Delta y(t) = y(t) - y(t-1)$, $\Delta u(t) = u(t) - u(t-1)$. The sampling time has been chosen to be 0.01 hour which equals 36 seconds. Based on the results of the maximum likelihood algorithm it has been decided to adopt a 4-th order ARMAX model for the model relating dry basis weight to thick stock flow. Several quantitative measures for the performance of the control algorithm indicate that it behaves reasonably well, see Fig. 7 of the paper.

In this example both the dynamics of the physical plant and that of the noise process are modelled by the ARMAX representation (4.1). This modelling procedure is called black-box modelling. The ARMAX system representation is a special case of a Gaussian stochastic control system which is described in this chapter for systems without input and in Chapter 10 for systems with input.

A List of Phenomena for Stochastic Systems

There follows a list of control and signal processing problems in which results from stochastic control and system theory have been or may be used.

Control engineering: Control of a paper machine [8], control of an ore crusher [13], and prediction of power demand [12].

Communication engineering: Echo cancellation and channel egalization [27, 94, 95], interpolation of signals in a compact disc recorder [36], overload control of telephone switches [25, 80, 81, 82, 38].

Electrical engineering: Control and filtering of power systems [57].

Civil engineering - Hydrology and hydraulics: Prediction of water levels [33].

Civil engineering - Spatial interpolation: Kriging [9, 21, 37, 45, 60].

Transportation engineering - Road traffic: Control of freeway traffic flow [84], control of urban road networks.

Transportation engineering - Shipping: Control of a single-point moored large tanker [59].

Aerospace: Stability of a helicopter [67].

Geological prospecting - Seismic exploration and production: Inverse scattering problems [16, 51, 65, 101], extraction of oil or gas from a reservoir.

Economics: Portfolio selection and consumption investment problems [58, 62, 78]; When to cut a tree? - An optimal stopping problem [64]; and production planning [24].

Other applications of control and filtering may be found in, for example, the journals *Automatica, IEEE Circuits and Systems Magazine, IEEE Control Systems Magazine, IEEE Transactions on Automatic Control, Journal of Economic Dynamics and Control.*

4.2 The Concept of a Stochastic System

Control of stochastic systems requires the concept of a stochastic system. The definition should be applicable to a control system with any type of probability distribution.

A concept of a stochastic system for only the state process of a stochastic system was mentioned by R.E. Kalman in the book [42, p. 5 footnote]. Another source, [83], credits C. Shannon for the concept of a stochastic system of the form, $(x(t), u(t)) \mapsto \text{cpdf}(x(t+1), y(t)) | F^{x(t), u(t)})$, thus the current state and the current input determine the probability distribution of the next state and the current output. But the author has not found this concept in any paper of C. Shannon. The model is also used in [10].

Below a general definition is proposed which is phrased in terms of the conditional independence relation. An informal discussion proceeds the formal definition.

A stochastic system with a state process x and an output process y is herewith defined with the probability distribution of the initial state and with the probabilistic transition function,

$$\begin{aligned}
&X = \mathbb{R}^{n_x},\ Y = \mathbb{R}^{n_y},\ T = \mathbb{N} = \{0,1,2,\ldots\},\\
&x(t) \mapsto \text{cpdf}\left(\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix} | F_t^{x-} \vee F_{t-1}^{y-} \right),\ \forall\, t \in T,\ \ \text{pdf}(x_0),\\
F_t^{x-} &= \sigma(\{x(s),\ \forall\, s \leq t\}),\ \forall\, t \in T;\\
F_{t-1}^{y-} &= \sigma(\{y(s),\ \forall\, s \leq t-1\}),\ \forall\, t \in T,\ t \geq 1.
\end{aligned}$$

In the above cpdf denotes a *conditional probability distribution function*. A stochastic system is defined by the conditional probabilistic transition function in combination with the probability distribution function $\text{pdf}(x_0)$ of the initial state x_0. The conditional probabilistic transition function maps, for any time $t \in T$, the current state $x(t)$ to the conditional probability distribution of the next state and the current output, $(x(t+1), y(t))$, conditioned on the indicated past states and past outputs $F_t^x \vee F_{t-1}^y$. From the probability distribution of the initial state and from the conditional probabilistic transition function one can construct the probability measure for the entire state and output processes on the considered horizon.

The above definition does not restrict the measurability of the state and the output processes. The reader may learn from the chapters on stochastic realization, Chapter 6 and Chapter 7, that there exists a set of stochastic system all having the same output process either in terms of finite-dimensional probability distributions of the output process or in terms of a sequence of random variables of the output process.

If one uses a forward representation of a stochastic system starting from an initial state x_0 and driven by a noise process v then it can be proven that that the state $x(t+1)$ and the output $y(t)$ are measurable with respect to the past of these processes, or,equivalently,

$$F^{x(t+1)} \vee F^{y(t)} \subseteq F^{x_0} \vee F_t^v,\ \forall\, t \in T.$$

For such a stochastic system representation one can prove that the condition of a stochastic system holds. See for this Example 4.2.1.

A direct consequence of the above definition is that,

$$\begin{aligned}
& \text{cpdf}((x(t+1), y(t)) | F_t^{x-} \vee F_{t-1}^{y-})\\
=\ & \text{cpdf}((x(t+1), y(t)) | F^{x(t)}),\ \forall\, t \in T;\\
\Leftrightarrow\ & (F^{x(t+1)} \vee F^{y(t)}, F_t^{x-} \vee F_{t-1}^{y-} | F^{x(t)}) \in \text{CI},\ \forall\, t \in T,\\
\Leftrightarrow\ & (F_t^{x+} \vee F_t^{y+}, F_t^{x-} \vee F_{t-1}^{y-} | F^{x(t)}) \in \text{CI},\ \forall\, t \in T; \qquad (4.2)\\
F_t^{x+}\ =\ & \sigma(\{x(s),\ \forall\, s \geq t\}),\ F_t^{y+} = \sigma(\{y(s),\ \forall\, s \geq t\}),\ \forall\, t \in T.
\end{aligned}$$

The characterization of equation (4.2) in terms of conditional independence will be taken as the general definition of a stochastic system. The equivalences above are formalized and proven in Section 5.10. There the asymmetry between the time arguments in the above formulas is discussed.

Note that the above definition does not impose any restriction on the measurability of the state process. The state at time $t \in T$ may be a function of the initial state and of a noise process, but it could also be a function of the past outputs. The reader will learn from stochastic realization theory that for any output process there exists a large set of state processes.

There follows an illustration of the concept of a stochastic system for what will be defined as a time-invariant Gaussian system representation below in this chapter.

Example 4.2.1. *A time-invariant Gaussian system.* Students in engineering most learn stochastic systems for the object defined below. Therefore this is a useful starting point to explain the concept of a stochastic system.

Consider the time-invariant Gaussian system representation,

$$\begin{aligned} x(t+1) &= Ax(t)+Mv(t),\ x_0, \\ y(t) &= Cx(t)+Nv(t), \end{aligned}$$

where $x_0 : \Omega \rightarrow \mathbb{R}^{n_x}, x_0 \in G(m_0,Q_0)$, $v : \Omega \times T \rightarrow \mathbb{R}^{n_v}$ is a Gaussian white noise process with $v(t) \in G(0,I_{n_v})$, F^{x_0}, F^v_∞ are independent σ-algebras, $A \in \mathbb{R}^{n_x \times n_x}$, $M \in \mathbb{R}^{n_x \times n_v}$, $C \in \mathbb{R}^{n_y \times n_x}$, $N \in \mathbb{R}^{n_y \times n_v}$, and $x : \Omega \times T \rightarrow \mathbb{R}^{n_x}$ and $y : \Omega \times T \rightarrow \mathbb{R}^{n_y}$ defined by the above equations. It will be argued that this Gaussian system representation satisfies the informal formulation of a stochastic system stated above. Note that,

$$\begin{aligned} & E\left[\exp\left(i\begin{pmatrix}w_x\\w_y\end{pmatrix}^T\begin{pmatrix}x(t+1)\\y(t)\end{pmatrix}\right)|F_t^{x-}\vee F_{t-1}^{y-}\right] \\ & \overset{(1)}{=} \exp\left(i\begin{pmatrix}w_x\\w_y\end{pmatrix}^T\begin{pmatrix}A\\C\end{pmatrix}x(t)\right)\times E\left[\exp\left(i\begin{pmatrix}w_x\\w_y\end{pmatrix}^T\begin{pmatrix}M\\N\end{pmatrix}v(t)\right)\right] \\ & \overset{(2)}{=} \exp\left(i\begin{pmatrix}w_x\\w_y\end{pmatrix}^T\begin{pmatrix}A\\C\end{pmatrix}x(t)-\frac{1}{2}\begin{pmatrix}w_x\\w_y\end{pmatrix}^T Q_r\begin{pmatrix}w_x\\w_y\end{pmatrix}\right), \\ & \qquad \forall\, t \in T,\ \forall\, (w_x,w_y) \in \mathbb{R}^{n_x}\times\mathbb{R}^{n_y},\ Q_r = \begin{pmatrix}M\\N\end{pmatrix}\begin{pmatrix}M\\N\end{pmatrix}^T \succeq 0,\ x_0 \in G(m_0,Q_0). \end{aligned}$$

(1) because $x(t)$ is measurable with respect to the σ-algebra $F_t^{x-}\vee F_{t-1}^{y-}$ and $v(t)$ is independent of that σ-algebra; and (2) because $\text{pdf}(v(t)) \in G(0,I)$.

Note that the probabilistic transition map yields a conditional Gaussian characteristic function of the form,

$$\begin{aligned} & x(t) \mapsto E\left[\exp\left(i\begin{pmatrix}w_x\\w_y\end{pmatrix}\begin{pmatrix}x(t+1)\\y(t))\end{pmatrix}\right)|F_t^{x-}\vee F_{t-1}^{y-}\right] \\ & = \exp\left(i\begin{pmatrix}w_x\\w_y\end{pmatrix}^T\begin{pmatrix}A\\C\end{pmatrix}x(t)-\frac{1}{2}\begin{pmatrix}w_x\\w_y\end{pmatrix}^T Q_r\begin{pmatrix}w_x\\w_y\end{pmatrix}\right),\ \forall\, t \in T. \end{aligned}$$

Note also that in the conditional characteristic function of $(x(t+1),y(t))$ given $(F_t^{x-}\vee F_{t-1}^{y-})$ the conditional mean value depends on the state $x(t)$ but the conditional variance does not, and neither does it depend on the other random variables $x(t-1),x(t-2),\ldots,x(0)$ nor on $y(t-1),y(t-2),\ldots,y(0)$. The function depends on the matrices A, C, and Q_r.

The reader is assumed to be familiar with the concept of conditional independence, see Section 2.9 and Section 19.8.

Definition 4.2.2. A discrete-time *stochastic (dynamic) system* in terms of a state process and an output process (without input) is a collection,

$$\{\Omega, F, P, T, Y, B_Y, X, B_X, y, x\},$$

where (Ω, F, P) is a complete probability space; $T \subseteq \mathbb{Z}$, to be called the *time index set;* (Y, B_Y) is a measurable space, to be called the *output space;* (X, B_X) is a measurable space, to be called the *state space;* $y : \Omega \times T \to Y$ is a stochastic process, to be called the *output process;* $x : \Omega \times T \to X$ is a stochastic process, to be called the *state process;* such that,

$$\forall\, t \in T, \;\; (F_t^{y+} \vee F_t^{x+}, F_{t-1}^{y-} \vee F_t^{x-} | F^{x(t)}) \in \text{CI}. \tag{4.3}$$

The definition does not restrict the measurability of the state process. The reader will learn in Chapter 6 of stochastic realization theory in which it is proven that a stochastic system has in general many different state processes for the same output process.

The convention is adopted that, for $t = 0$, the σ-algebra F_{-1}^{y-} is the trivial σ-algebra generated by the set Ω, the empty set, and the null sets of the complete probability space. In addition, in case of the finite interval $T(0 : t_1) = \{0, 1, \ldots, t_1\}$, that for $t = t_1$ the σ-algebra $F_{t_1}^{x+} = F^{x(t_1)}$ and $F_{t_1}^{y+} = F^{y(t_1)}$. The class of stochastic systems is denoted by StocS and a stochastic system is denoted by the tuple,

$$\{\Omega, F, P, T, Y, B_Y, X, B_X, y, x\} \in \text{StocS}.$$

Arguments to favor the above formulation of a stochastic system are that: (1) the use of the conditional independence relation; and (2) the formulation in terms of future and of past state and output processes. The distinction into future and past leads to a forward and a back representation of a stochastic system as described below in this chapter.

More details on the above definition of a stochastic system and properties may be found in Section 5.10.

Based on the historical practice of the formulation of the dynamics of a stochastic system, one uses either a forward or a backward representation of a stochastic system. The representation itself can be specified by (1) either by a tuple of a conditional probability distribution for the transition probability and the probability distribution of the initial state, (2) or by a tuple of a conditional characteristic function for the transition function and the characteristic function of the initial state.

Definition 4.2.3. Define a *forward conditional-probability-distribution representation of a stochastic system* by a tuple consisting of a conditional probability distribution of the transition, and the probability distribution of the initial state x_0, denoted by,

$$T = \mathbb{N}, \; n_y \in \mathbb{Z}_+, \; n_x \in \mathbb{N},$$
$$\{\Omega, F, P, T, \mathbb{R}^{n_y}, B(\mathbb{R}^{n_y}), \mathbb{R}^{n_x}, B(\mathbb{R}^{n_x}), \; y, x\}, \;\; \forall\, t \in T,$$
$$f(.;(x(t+1), y(t)) | F_t^{x-} \vee F_{t-1}^{y-}) = f_s(t, x(t)), \; \text{pdf}(x_0).$$

In the above equations, f denotes a conditional probability distribution for the transition function and f_s denotes a function.

Equivalently, define the *forward conditional-characteristic-function representation* of a stochastic system by a tuple consisting of a conditional characteristic function of the transition, and the conditional characteristic function of the initial state x_0, denoted by,

$$\begin{aligned}
&T = \mathbb{N},\ n_y \in \mathbb{Z}_+,\ n_x \in \mathbb{N},\\
&\{\Omega, F, P, T, \mathbb{R}^{n_y}, B(\mathbb{R}^{n_y}), \mathbb{R}^{n_x}, B(\mathbb{R}^{n_x}),\ y, x\},\ \ \forall\, t \in T,\\
&E\left[\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix}\right) \mid F_t^{x-} \vee F_{t-1}^{y-}\right] = f_{tr,ccf}(t,(w_x,w_y);x(t)),
\end{aligned}$$

and the characteristic function of x_0.

Proposition 4.2.4. A sufficient condition for the specification of a stochastic system. *Consider a conditional probability distribution representation of a stochastic system of Def. 4.2.3. Then this representation defines a stochastic system of Def. 4.2.2. The same conclusion holds for a conditional characteristic representation of a stochastic system.*

Proof. It is to be proven that (1) the specification of the representation of a stochastic system constructs the complete measure on the state and the output process; and (2) that the process satisfy the conditional independence condition of the definition of a stochastic process.

The probability distribution of the initial state is specified in the system representation. From the specification of the conditional probability distribution of the transition then follows that the measure on the random variables $(x(1), y(0))$ is specified. By induction it then follows that, for every time $t \in T$, the probability measure on $(x(t+1), y(t))$ is specified.

It follows from the condition of Def. 4.2.3, from $F^{x(t)} \subset F_t^{x-} \vee F_t^{y-}$, and from Proposition 2.9.2 that for all $t \in T$,

$$(F^{x(t+1)} \vee F^{y(t)},\ F_t^{x-} \vee F_{t-1}^{y-} |\ F^{x(t)}) \in \text{CI}.$$

From Proposition 5.10.1 then follows that for all $t \in T$,

$$(F_t^{x+} \vee F_t^{y+},\ F_t^{x-} \vee F_{t-1}^{y-} |\ F^{x(t)}) \in \text{CI}.$$

hence the system is a stochastic system as defined in Def. 4.2.3. □

The definition of a stochastic dynamic system implies by restriction of σ-algebras that the state process satisfies the condition,

$$(F_t^{x+}, F_t^{x-} | F^{x(t)}) \in CI,\ \forall\, t \in T = \mathbb{N}.$$

This statement equals the definition of a Markov process, see Def. 3.3.6. The state process of a stochastic system is therefore a Markov process.

Definition 4.2.5. Consider a stochastic dynamic system $\{\Omega, F, P, T, Y, B_Y, X, B_X, y, x\} \in \text{StocS}$. This system is called:

(a) A *stationary* stochastic system if (x, y) is a pair of jointly stationary stochastic processes;

(b) The forward representation of a stochastic system is callaed *time-invariant* if the conditional probability distribution of the transition does not depend on the time moment, and the probability distribution of the initial state equals the invariant distribution of the state, see Section 4.4;

(c) A *Gaussian stochastic system* if there exist $n_x \in \mathbb{N}$ and $n_y \in \mathbb{Z}_+$ such that

$$(Y, B_Y) = (\mathbb{R}^{n_y}, B(\mathbb{R}^{n_y})), \ \ (X, B_X) = (\mathbb{R}^{n_x}, B(\mathbb{R}^{n_x})),$$

and if (x, y) is a jointly Gaussian process; a *finite-dimensional Gaussian system* would be a more appropriate definition but that term is too cumbersome; the class of Gaussian systems is denoted by GStocS;

(d) A *finite stochastic system* if Y, X are finite sets and B_Y, B_X are the σ-algebras on Y, X generated by all subsets; and the class of such systems is denoted by FStocS.

4.3 Time-Varying Gaussian Systems

The reader finds in this section results for time-varying Gaussian systems, both for forward and for backward Gaussian system representations.

For forward Gaussian system representations, the time index set will be either an *infinite horizon* $T = \mathbb{N} = \{0, 1, 2, \ldots\}$ or a *finite horizon* $T_1 = \{0, 1, 2, \ldots, t_1\} \subset \mathbb{N}$ for a $t_1 \in \mathbb{Z}_+$. Because in these notes attention is restricted to discrete-time processes and systems, the adjective *discrete-time* will often be omitted, except in formal definitions.

Definition 4.3.1. A *forward Gaussian system representation* is a collection

$$\{\Omega, F, P, T, \mathbb{R}^{n_y}, B(\mathbb{R}^{n_y}), \mathbb{R}^{n_x}, B(\mathbb{R}^{n_x}), y, x, v, x_0, A, C, M, N\}$$

in which the processes x and y satisfy the equations,

$$x(t+1) = A(t)x(t) + M(t)v(t), \ x(t_0) = x_0, \tag{4.4}$$

$$y(t) = C(t)x(t) + N(t)v(t), \tag{4.5}$$

where (Ω, F, P) is a complete probability space; $T = \mathbb{N} = \{0, 1, 2, \ldots\}$, to be called the *time index set;* $x_0 : \Omega \rightarrow \mathbb{R}^{n_x}$ is a Gaussian random variable with $x_0 \in G(m_0, Q_0)$ called the *initial state*; $v : \Omega \times T \rightarrow \mathbb{R}^{n_v}$ a standard Gaussian white noise process with intensity I_{n_v}, hence $\forall\ t \in T$, $v(t) \in G(0, I_{n_v})$; it will be assumed that x_0, v are independent objects in the sense that F^{x_0}, F^v_∞ are independent σ-algebras; $A : T \rightarrow \mathbb{R}^{n_x \times n_x}$, $M : T \rightarrow \mathbb{R}^{n_x \times n_v}$, $C : T \rightarrow \mathbb{R}^{n_y \times n_x}$ $N : T \rightarrow \mathbb{R}^{n_y \times n_v}$; $x : \Omega \times T \rightarrow \mathbb{R}^{n_x}$, and $y : \Omega \times T \rightarrow \mathbb{R}^{n_y}$ are stochastic processes defined by the above recursions.

A Gaussian system representation will in the remainder of the book often be specified only by the equations (4.4,4.5) with the understanding that the above conditions hold.

A *time-invariant forward Gaussian system representation* is a forward Gaussian system representation in which the functions (A,C,M,N) do not explicitly depend on time, thus, for all $t \in T$, $A(t) = A(0) = A \in \mathbb{R}^{n_x \times n_x}$, etc. Then the parameters of the time-invariant system are denoted by $(n_y, n_x, n_v, A, C, M, N)$ of which the variance of the initial state is not listed. Note that by definition of a standard Gaussian white noise, for all $t \in T$, $v(t) \in G(0, I_{n_v})$ with $Q_v = I_{n_v}$.

Define the conditions,

$$(1)\ n_y \leq n_v,\ \text{rank}(N) = n_y,\ \text{rank}\begin{pmatrix} M \\ N \end{pmatrix} = n_v;\ \ (2)\ n_y > n_v.$$

When one of these conditions is imposed, this will be explicitly stated.

Gaussian system representations are frequently used stochastic models in stochastic system theory. They are useful because they model Gaussian processes and because Gaussian processes are appropriate models for signals of many engineering problems.

Control engineers refer to the Gaussian white noise process as *noise*. The interpretation of the output equation is that the output, $y(t)$, is a signal, $Cx(t)$, based on the state, plus noise $v(t)$. If it helps engineers to think of this noise interpretation then that is OK. However, the model does not define any physical property of the noise. In case of other stochastic systems with different probability distributions, the noise term will in general have a different representation and a different interpretation. The noise term may not determine the conditional distribution of $y(t)$ conditioned on $x(t)$.

In Section 4.2 the definition of a stochastic system has been introduced, as well as the concept of a Gaussian system. A Gaussian system representation as defined above will be argued to be a representation of such a Gaussian system. In Section 4.1 an example has been presented for which Gaussian system representations is a useful model.

In case of a time-invariant forward Gaussian system with (2) of Def. 4.3.1, thus if $n_y > n_v$, then there exists a transformation matrix such that,

$$L_y y(t) = L_y C x(t) + L_y N v(t) = L_y C x(t) + \begin{pmatrix} N_1 \\ 0 \end{pmatrix} v(t),$$
$$\text{where, } L_y \in \mathbb{R}^{n_y \times n_y},\ \text{rank}(L_y) = n_y.$$

Thus there are components of the transformed output process $L_y y(t)$ which are not affected by the noise process v. This case then requires a separate analysis for the two component processes: (a) those output components which are affected by the noise; and (b) those output components which are not affected by the noise. The latter case will not be treated in this book. The reader may find theory for that case in the theory of observers of a linear deterministic system. From now on the assumption $n_y \leq n_v$ will often be assumed.

The column rank condition on the column matrix consisting of M, N is to eliminate dependence of the noise components $v(t)$.

In the literature one may find a representation of the form

$$\begin{aligned} x(t+1) &= A(t)x(t)+M_1(t)r(t), \\ y(t) &= C(t)x(t)+N_2(t)w(t), \end{aligned}$$

in which $r : \Omega \times T \to \mathbb{R}^{n_r}$, $w : \Omega \times T \to \mathbb{R}^{n_w}$ are independent standard Gaussian white noise processes. Such a representation is a special case of the Gaussian system representation defined above as the following transformation shows.

$$v : \Omega \times T \to \mathbb{R}^{n_r+n_w}, \; v(t) = \begin{pmatrix} r(t) \\ w(t) \end{pmatrix}, v(t) \in G(0, I_{n_v}),$$

$$\begin{aligned} Q_v(t) &= \begin{pmatrix} I_{n_r} & 0 \\ 0 & I_{n_w} \end{pmatrix} = I_{n_v}, \\ M(t) &= \begin{pmatrix} M_1(t) & 0 \end{pmatrix}, \; N(t) = \begin{pmatrix} 0 & N_2(t) \end{pmatrix}; \text{ then,} \\ & M(t)v(t) = M_1(t)r(t), \;\; N(t)v(t) = N_2(t)w(t). \end{aligned}$$

Then one obtains the representation of equations (4.4,4.5). The case in which the processes r and w are not independent but jointly Gaussian and correlated, may be transformed in an analogous way, it requires a second transformation of the noise process to one with a variance matrix equal to the identity matrix.

In the literature researchers also use as a model of a stationary Gaussian process the system representations of either an *autoregressive* (AR) recursion or of an *autoregressive moving-average* (ARMA) recursion. Both of these recursions can be converted into a time-invariant Gaussian system. The transformation is difficult for most researchers. The user has to think of what the state of the system is and then write a Gaussian system for it. In general, that first system representation will not be a minimal weak Gaussian stochastic realization of the output process but from the realization one can construct a minimal realization using the procedures of Chapter 6. There are several other models of stationary Gaussian processes but all those can also be converted to a time-invariant Gaussian system. It does not seem useful to take space of this book for these transformations.

Definition 4.3.2. Consider a Gaussian stochastic system representation,

$$\begin{aligned} x(t+1) &= A(t)x(t)+M(t)v(t), \\ y(t) &= C(t)x(t)+N(t)v(t). \end{aligned}$$

The system is said to be *state-output conditionally independent* conditioned on past-states and past-outputs if,

$$(F^{x(t+1)}, F^{y(t)} | F_t^x \vee F_{t-1}^y) \in \text{CI}, \; \forall \, t \in T.$$

Proposition 4.3.3. *A Gaussian system representation is state-output conditionally independent conditioned on past states and on past outputs if and only if,*

$$0 = M(t)N(t)^T, \; \forall \, t \in T.$$

Proof. See the following calculations where $t \in T$ is fixed.

$$
\begin{aligned}
&(F^{x(t+1)}, F^{y(t)} | F_t^x \vee F_{t-1}^y) \in \text{CI}, \\
&\Leftrightarrow E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix}\right) | F_t^x \vee F_{t-1}^y\right] \\
&\quad = E\left[\exp\left(iw_x^T x(t+1)\right) | F_t^x \vee F_{t-1}^y\right] E\left[\exp\left(iw_y^T y(t)\right) | F_t^x \vee F_{t-1}^y\right], \\
&\Leftrightarrow E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} (A(t)x(t)+M(t)v(t)) \\ (C(t)x(t)+N(t)v(t)) \end{pmatrix}\right) | F_t^x \vee F_{t-1}^y\right] \\
&\quad = E\left[\exp\left(iw_x^T (A(t)x(t)+M(t)v(t))\right) | F_t^x \vee F_{t-1}^y\right] \times \\
&\qquad \times E\left[\exp\left(iw_y^T (C(t)x(t)+N(t)v(t))\right) | F_t^x \vee F_{t-1}^y\right], \\
&\Leftrightarrow E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} M(t)v(t) \\ N(t)v(t) \end{pmatrix}\right) | F_t^x \vee F_{t-1}^y\right] \\
&\quad = E\left[\exp\left(iw_x^T M(t)v(t)\right) | F_t^x \vee F_{t-1}^y\right] E\left[\exp\left(iw_y^T N(t)v(t)\right) | F_t^x \vee F_{t-1}^y\right], \\
&\quad = E\left[\exp\left(iw_x^T M(t)v(t)\right)\right] E\left[\exp\left(iw_y^T N(t)v(t)\right)\right], \\
&\Leftrightarrow \exp\left(-\frac{1}{2}\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} M(t)M(t)^T & M(t)N(t)^T \\ N(t)M(t)^T & N(t)N(t)^T \end{pmatrix} \begin{pmatrix} w_x \\ w_y \end{pmatrix}\right) \\
&\quad = \exp\left(-\frac{1}{2}w_x^T M(t)M(t)^T w_x - \frac{1}{2}w_y^T N(t)N(t)^T w_y\right), \\
&\Leftrightarrow \exp(-w_x^T M(t)N(t)^T w_y) = 1,\ \forall\ (w_x, w_y) \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_y}, \\
&\Leftrightarrow M(t)N(t)^T = 0.
\end{aligned}
$$

The third equivalence is due to multiplication of the factor missing from the equality, or, in the other direction, by the inverse of the factor. □

Questions about the properties of Gaussian system representations are: (1) What are the families of finite-dimensional distributions of the state process x and the output process y? (2) Is the state process x a Markov process? (3) What are the asymptotic properties of the state and output process? These questions are answered below.

Definition 4.3.4. The *state-transition function* associated with the state-transition matrix $A : T \to \mathbb{R}^{n_x \times n_x}$ is defined to be the function $\Phi : T \times T \to \mathbb{R}^{n_x \times n_x}$,

$$
\begin{aligned}
\Phi(t+1, s) &= \begin{cases} A(t)\Phi(t,s), & \text{if } s < t+1, \\ I, & \text{if } s = t+1, \\ 0, & \text{if } s > t+1; \end{cases} \\
\Phi(t,s) &= A(t-1)A(t-2)\ldots A(s),\ \text{ if } s \le t-1.
\end{aligned}
$$

Theorem 4.3.5. The probability distributions of a forward Gaussian system representation. *Consider the Gaussian system representation,*

$$
\begin{aligned}
x(t+1) &= A(t)x(t) + M(t)v(t),\ x(0) = x_0 \in G(m_{x_0}, Q_{x_0}), \\
y(t) &= C(t)x(t) + N(t)v(t), \\
&\quad T = \mathbb{N},\ \text{pdf}(x_0) \in G(m_{x_0}, Q_{x_0}),\ \ \text{pdf}(v(t)) \in G(0, I).
\end{aligned}
$$

Let $\Phi : T \times T \rightarrow \mathbb{R}^{n_x \times n_x}$ be the state transition function of $A : T \rightarrow \mathbb{R}^{n_x \times n_x}$.

(a)For all $t \in T$, the σ-algebras $F_t^{v+} = \sigma(\{v(s), \forall\, s \geq t\})$ and $(F_t^x \vee F_{t-1}^y)$ are independent.

(b)Explicit expressions for x, y are, with $s < t$,

$$\begin{aligned} x(t) &= \Phi(t,s)x(s) + \sum_{r=s}^{t-1} \Phi(t-1,r)M(r)v(r), \\ y(t) &= C(t)\Phi(t,s)x(s) + \left[\sum_{r=s}^{t-1} C(t)\Phi(t-1,r)M(r)v(r) \right] + N(t)v(t). \end{aligned}$$

The following two processes are adapted, Def. 20.1.3, to the indicated filtrations,

$$\{x(t),\ F_{t-1}^v \vee F^{x_0},\ t \in T\},\ \{y(t),\ F_t^v \vee F^{x_0},\ t \in T\}.$$

(c)The process (x,y) is a jointly Gaussian process.

(d)The Gaussian system representation defines a Gaussian system as defined in Def. 4.2.5.(b).

(e)The state process x is a Gauss-Markov process with mean value function m_x, variance function Q_x, *and* covariance function W_x,

$$\begin{aligned} & m_x : T \rightarrow \mathbb{R}^{n_x},\ Q_x : T \rightarrow \mathbb{R}^{n_x \times n_x},\ W_x : T \times T \rightarrow \mathbb{R}^{n_x \times n_x}, \\ & x(t) \in G(m_x(t), Q_x(t)), \\ m_x(t+1) &= A(t)m_x(t),\ m_x(0) = m_{x_0}, \\ Q_x(t+1) &= A(t)Q_x(t)A(t)^T + M(t)M(t)^T,\ Q_x(0) = Q_{x_0}, \\ W_x(t,s) &= \begin{cases} Q_x(t), & \text{if } t = s, \\ \Phi(t,s)Q_x(s), & \text{if } s < t, \\ Q_x(t)\Phi(s,t)^T, & \text{if } s > t. \end{cases} \end{aligned}$$

(f) The output process y is a Gaussian process with with mean value function m_y, variance function Q_y, *and* covariance function W_y,

$$\begin{aligned} & m_y : T \rightarrow \mathbb{R}^{n_y},\ Q_y : T \rightarrow \mathbb{R}^{n_y \times n_y},\ Q_{x^+,y} : T \rightarrow \mathbb{R}^{n_y \times n_y}, \\ & W_y : T \times T \rightarrow \mathbb{R}^{n_y \times n_y}, \\ & y(t) \in G(m_y(t), Q_y(t)), \\ m_y(t) &= C(t)m_x(t),\ Q_y(t) = C(t)Q_x(t)C(t)^T + N(t)N(t)^T, \\ W_y(t,s) &= \begin{cases} Q_y(t), & \text{if } t = s, \\ C(t)\Phi(t,s)Q_x(s)C(s)^T + C(t)\Phi(t-1,s)M(s)N(s)^T, & \text{if } s < t, \end{cases} \\ Q_{x^+,y}(t) &= E[(x(t+1) - m_x(t+1))(y(t) - m_y(t))^T] \\ &= A(t)Q_x(t)C(t)^T + M(t)N(t)^T. \end{aligned}$$

Proof. (a) The assumption that v is a Gaussian white noise process implies that for all $t \in T$, F_t^{v+}, F_{t-1}^{v-} are independent σ-algebras. Note that $F_{t-1}^{v-} = \sigma(\{v(s), \forall s \leq t-1\})$. Again, by assumption, F^{x_0}, F_∞^{v-} are independent σ-algebras. Hence F_t^{v+}, $F_{t-1}^{v-} \vee F^{x_0}$ are independent. From the system representation (4.4,4.5) follows that $(F_1^x \vee$

$F_0^y) \subset (F_0^v \vee F^{x_0})$ and, by induction, that for all $t \in T$, $(F_t^x \vee F_{t-1}^y) \subset (F_{t-1}^v \vee F^{x_0})$. Thus F_t^{v+}, $(F_t^x \vee F_{t-1}^y)$ are independent σ-algebras for all $t \in T$.
(b) This is a simple verification by induction.
(c) Because $x_0 \in G(m_0, Q_0)$, v is a Gaussian process, F^{x_0}, F_∞^v are independent. From this, the linear equations of (b), and, Proposition 2.7.3, follows that $(x(t), y(t))$ are jointly Gaussian distributed. Similarly one proves that, for a finite sequence of times, the processes (x, y) sampled at those times are linear functions of the x_0 and past of the process v, hence are jointly Gaussian distributed. Thus (x, y) are jointly Gaussian processes.
(d) As in Example 4.2.1, and because of the properties of the Gaussian white noise process v,

$$\begin{aligned}
&E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix}\right) \mid F_{t-1}^v \vee F^{x_0}\right] \\
&= \exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} A \\ C \end{pmatrix} x(t) - \frac{1}{2}\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T Q_v \begin{pmatrix} w_x \\ w_y \end{pmatrix}\right), \\
&E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix}\right) \mid F_t^x \vee F_{t-1}^y\right] \\
&= E\left[E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix}\right) \mid F_{t-1}^v \vee F^{x_0}\right] \mid F_t^x \vee F_{t-1}^y\right] \\
&= \exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} A \\ C \end{pmatrix} x(t) - \frac{1}{2}\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T Q_v \begin{pmatrix} w_x \\ w_y \end{pmatrix}\right) \\
&= E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix}\right) \mid F^{x(t)}\right],
\end{aligned}$$

hence the forward Gaussian system representation defines the characteristic function of the transition and from Proposition 4.2.4 follows that the system representation defines a stochastic system.
(e) Let $t \in T$. Then,

$$\begin{aligned}
m_x(t+1) &= E[x(t+1)] = E[A(t)x(t) + M(t)v(t)] = A(t)m_x(t), \\
Q_x(t+1) &= E[(x(t+1) - m_x(t+1))(x(t+1) - m_x(t+1))^T] \\
&= E[(A(t)(x(t) - m_x(t)) + M(t)v(t))(A(t)(x(t) - m_x(t)) + M(t)v(t))^T] \\
&= A(t)Q_x(t)A(t)^T \; + \; M(t)M(t)^T,
\end{aligned}$$

because by (a) $v(t)$ is independent of $x(t)$ and $E[x(t) - m_x(t)] = 0$. Let $s, t \in T$, $s < t$. Then,

$$\begin{aligned}
x(t) - m_x(t) &= \Phi(t,s)(x(s) - m_x(s)) + \sum_{r=s}^{t-1} \Phi(t-1,r)M(r)v(r), \\
W_x(t,s) &= E[(x(t) - m_x(t))(x(s) - m_x(s))^T] \\
&= \Phi(t,s)W_x(s,s) \; = \; \Phi(t,s)Q_x(s),
\end{aligned}$$

because by (a), $\{v(s),...,v(t-1)\}$ are independent of F_s^x. Let $t \in T$. Then

$$\begin{aligned}
&E[\exp(iw^T x(t+1))|F_t^x] \\
&= E[E[\exp(iw^T(A(t)x(t)+M(t)v(t))|F_t^x \vee F_{t-1}^v]|F_t^x] \\
&= \exp(iw^T A(t)x(t))E[\exp(iw^T M(t)v(t))] \\
&= \exp(iw^T A(t)x(t) \;-\; \frac{1}{2}w^T M(t)M(t)^T w),
\end{aligned}$$

by measurability of $x(t)$ on F_t^x, by (a), and by $v(t) \in G(0, I_{n_v})$,

$$\begin{aligned}
&E[\exp(iw^T x(t+1))|F^{x(t)}] \\
&= E[E[\exp(iw^T x(t+1))|F_t^x]|F^{x(t)}], \text{ by reconditioning,} \\
&= \exp(iw^T A(t)x(t) \;-\; \frac{1}{2}w^T M(t)M(t)^T w) \\
&= E[\exp(iw^T x(t+1))|F_t^x], \; \forall\, w \in \mathbb{R}^{n_x}, \; \forall\, t \in T,
\end{aligned}$$

twice by the above relation, hence x is a Markov process by Proposition 3.3.7.
(f) Let $t, s \in T, \; s < t$. Then

$$\begin{aligned}
m_y(t) &= E[y(t)] = E[C(t)x(t)+N(t)v(t)] = C(t)m_x(t), \\
Q_y(t) &= E[(y(t)-m_y(t))(y(t)-m_y(t))^T]] \\
&= E[(C(t)(x(t)-m_x(t))+N(t)v(t))(C(t)(x(t)-m_x(t))+N(t)v(t))^T] \\
&= C(t)Q_x(t)C(t)^T + N(t)N(t)^T,
\end{aligned}$$

because by (a), $x(t)$ and $v(t)$ are independent, and $E[v(t)] = 0$.

$$\begin{aligned}
y(t)-m_y(t) &= C(t)\Phi(t,s)(x(s)-m_x(s)) \\
&\quad + \sum_{r=s}^{t-1} C(t)\Phi(t-1,r)M(r)v(r) + N(t)v(t), \\
y(s)-m_y(s) &= C(s)(x(s)-m_x(s)) + N(s)v(s), \\
W_y(t,s) &= E[(y(t)-m_y(t))(y(s)-m_y(s))^T] \\
&= C(t)\Phi(t,s)Q_x(s)C(s)^T + C(t)\Phi(t-1,s)M(s)N(s)^T.
\end{aligned}$$

Further,

$$\begin{aligned}
Q_{x^+,y}(t) &= E[(x(t+1)-m_x(t+1))(y(t)-m_y(t))^T] \\
&= E[(A(t)(x(t)-m_x(t))+M(t)v(t))(C(t)(x(t)-m_x(t))+N(t)v(t))^T] \\
&= A(t)Q_x(t)C(t)^T + M(t)N(t)^T.
\end{aligned}$$

□

The reader finds below results for time-varying backward Gaussian systems. For the backward system the time index set used is that of the negative integers. Thus, $T_- = \{\ldots, \; -2, \; -1, \; 0\} \subset \mathbb{N}$ or $T(-t_1:0) = \{-t_1, -t_1+1, \ldots, -1, 0\}$ for $t_1 \in \mathbb{Z}_-$. The backward system starts at time 0 and then moves backward in the time index set till time $-t_1 \in T_-$.

Definition 4.3.6. A *backward Gaussian system representation* is a collection

$$\{\Omega, F, P, T, \mathbb{R}^{n_y}, B(\mathbb{R}^{n_y}), \mathbb{R}^{n_x}, B(\mathbb{R}^{n_x}), y, x, v, x_0, A, C, M, N\}$$

in which the processes x and y satisfy the equations,

$$x(t-1) = A_b(t)x(t) + M_b(t)v(t),\ x(0) = x_0, \tag{4.6}$$

$$y(t-1) = C_b(t)x(t) + N_b(t)v(t), \tag{4.7}$$

where (Ω, F, P) is a complete probability space; $T_-(t_1:0) = \{t_1, t_1+1, \ldots, -1, 0\}$, to be called the *backward time index set;* $x_0 : \Omega \rightarrow \mathbb{R}^{n_x}$ is a Gaussian random variable with $x_1 \in G(m_{x_1}, Q_{x_1})$ called the *terminal state*; $v_b : \Omega \times T \rightarrow \mathbb{R}^{n_{v_b}}$ a standard Gaussian white noise process with intensity $I_{n_{v_b}}$, hence $\forall\ t \in T$, $v_b(t) \in G(0, I_{n_{v_b}})$; it will be assumed that x_1, v_b are independent objects in the sense that F^{x_1}, $F^{v_b}_{t_1} = F(\{v(s),\ \forall\ s \in T_-\})$ are independent σ-algebras; $A_b : T \rightarrow \mathbb{R}^{n_x \times n_x}$, $M_b : T \rightarrow \mathbb{R}^{n_x \times n_{v_b}}$, $C_b : T \rightarrow \mathbb{R}^{n_y \times n_x}$ $N_b : T \rightarrow \mathbb{R}^{n_y \times n_{v_b}}$; $x : \Omega \times T \rightarrow \mathbb{R}^{n_x}$, and $y : \Omega \times T \rightarrow \mathbb{R}^{n_y}$ are stochastic processes defined by the above recursions.

For short, a backward Gaussian system representation is specified by the equations (4.6,4.7).

A *time-invariant backward Gaussian system representation* is a backward Gaussian system representation in which the functions (A_b, C_b, M_b, N_b) do not explicitly depend on time, thus, for all $t \in T_-$, $A_b(t) = A_b(0) = A_b \in \mathbb{R}^{n_x \times n_x}$, etc. Then the parameters of the time-invariant system are denoted by $(n_y, n_x, n_{v_b}, A_b, C_b, M_b, N_b)$ of which the variance of the initial state is not listed.

Define the conditions,

$$(1) n_y \leq n_{v_b},\ \text{rowrank}(N) = n_y,\ \text{colrank}\begin{pmatrix} M_b \\ N_b \end{pmatrix} = n_{v_b};\ `(2) n_y > n_{v_b}.$$

When one of these conditions is imposed, this will be explicitly stated.

The reader may have noticed that in the system equations (4.6, 4.7) on the left-hand side, the random variables $(x(t-1),\ y(t-1))$ have been used. It is a direct consequence from the choice to use for the forward system the variables $(x(t+1),\ y(t))$ that for the backward system the indicated variables are to be used. Thus $y(t)$ is specified by the forward representation and not by the backward system representation at that time. As for the forward system representation, the probability distributions of the state and the output process have to be determined in terms of the system matrices of the backward system.

Definition 4.3.7. The *backward state-transition function* associated with the transition matrix $A_b : T_- \rightarrow \mathbb{R}^{n_x \times n_x}$ is defined to be the function $\Phi_b : T_- \times T_- \rightarrow \mathbb{R}^{n_x \times n_x}$,

$$\Phi_b(t,s) = \begin{cases} A_b(t+1)\Phi_b(t+1,s), & \text{if } t < s, \\ I, & \text{if } t = s, \\ 0, & \text{if } t > s; \end{cases}$$

$$\Phi_b(s,s) = I,\ \Phi_b(s-1,s) = A_b(s),$$

$$\Phi_b(t,s) = A_b(t+1)A_b(t+2)\ldots A_b(s),\ \text{ if } t < s.$$

Theorem 4.3.8. *Consider the Gaussian system representation*

$$\begin{aligned}
x(t-1) &= A_b(t)x(t)+M_b(t)v_b(t),\ x(0)=x_0,\ \text{pdf}(x_0)\in G(m_{x_0},Q_{x_0}),\\
y(t-1) &= C_b(t)x(t)+N_b(t)v_b(t),\ T_-(t_1:0)=\{t_1,t_1+1,\ldots,-1,0\},\\
&\qquad \text{pdf}(x_0)\in G(m_{x_0},Q_{x_0}),\ \text{pdf}v_b(t)\in G(0,I).
\end{aligned}$$

Let $\Phi_b : T_- \times T_- \to \mathbb{R}^{n_x\times n_x}$ be the state transition function of the system matrix $A_b : T_- \to \mathbb{R}^{n_x\times n_x}$.

(a)For all $t\in T_-$, the σ-algebras $F_t^{v_b-}=\sigma(\{v_b(s),\forall\ s\le t\})$ and $(F_t^x\vee F_t^y)$ are independent.

(b)Explicit expressions for x,y are, with $t<s$,

$$\begin{aligned}
x(t) &= \Phi_b(t,s)x(s)+\sum_{r=t+1}^{s}\Phi_b(t+1,r)M_b(r)v_b(r),\\
y(t) &= C_b(t+1)\Phi_b(t+1,s)x(s)+\\
&\quad +\left[\sum_{r=t+2}^{s}C_b(t+1)\Phi_b(t+2,r)M_b(r)v_b(r)\right]+N_b(t+1)v_b(t+1).
\end{aligned}$$

(c)The process (x,y) is a jointly Gaussian process.

(d)The Gaussian system representation defines a Gaussian system as defined in Def. 4.2.5.(b).

(e)The state process x is a Gauss-Markov process with mean value function m_x, variance function Q_x, *and* covariance function W_x,

$$\begin{aligned}
&m_x : T\to\mathbb{R}^{n_x},\ Q_x : T\to\mathbb{R}^{n_x\times n_x},\ W_x : T\times T\to\mathbb{R}^{n_x\times n_x},\\
&x(t)\in G(m_x(t),Q_x(t)),\\
m_x(t-1) &= A_b(t)m_x(t),\ m_x(0)=m_{x_0},\\
Q_x(t-1) &= A_b(t)Q_x(t)A_b(t)^T+M_b(t)M_b(t)^T,\ Q_x(0)=Q_{x_0},\\
W_x(t,s) &= \begin{cases} Q_x(t), & \text{if } t=s,\\ \Phi_b(t,s)Q_x(s), & \text{if } t<s,\\ Q_x(t)\Phi_b(s,t)^T, & \text{if } t>s.\end{cases}
\end{aligned}$$

(f) The output process y is a Gaussian process with mean value function m_y, variance function Q_y, *and* covariance function W_y,

$$\begin{aligned}
&m_y : T\to\mathbb{R}^{n_y},\ Q_y : T\to\mathbb{R}^{n_y\times n_y},\ Q_{x,y} : T\to\mathbb{R}^{n_y\times n_y},\\
&W_y : T_-\times T_-\to\mathbb{R}^{n_y\times n_y},\\
&y(t)\in G(m_y(t),Q_y(t)),\\
m_y(t) &= C_b(t+1)m_x(t+1),\\
Q_y(t) &= C_b(t+1)Q_x(t+1)C_b(t+1)^T+N_b(t+1)N_b(t+1)^T,\\
W_y(t,s) &= \begin{cases} Q_y(t), & \text{if } t=s,\\ C_b(t+1)\Phi_b(t+1,s+1)Q_x(s+1)C_b(s+1)^T & \\ +C_b(t+1)\Phi_b(t+2,s+1)M_b(s+1)N_b(s+1)^T, & \text{if } t<s,\end{cases}\\
Q_{x,y}(t) &= E[(x(t)-m_x(t))(y(t)-m_y(t))^T]\\
&= A_b(t+1)Q_x(t+1)C_b(t+1)^T+M_b(t+1)N_b(t+1)^T.
\end{aligned}$$

The proof is analogous to that of Theorem 4.3.8 and therefore omitted.

4.4 Time-Invariant Gaussian Systems

A stochastic system may be time-invariant and such systems arise often as models in engineering and in other research domains. Such time-invariant stochastic systems may have an invariant measure on the product of the state set and the output set. Even if the measure of the initial state differs from the invariant measure for the state set then the measure of the state set may converge in distribution to the invariant measure for the state set.

In this section the existence of an invariant measure is discussed for both a time-invariant forward Gaussian system representation and a time-invariant backward Gaussian system representation.

The reader is reminded of a standard result of probability theory, how a random variable and a probability measure jointly induce a measure on the image space of the random variable.

Consider a stochastic system with state process $x : \Omega \times T \rightarrow X$ and output process $y : \Omega \times T \rightarrow Y$. Consider for any time $t \in T$, the vector of random variables,

$$\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix} : \Omega \rightarrow \mathbb{R}^{n_x+n_y}.$$

This random vector induces a probability measure on the image space which objects are denoted by,

$$\begin{aligned}
& X = \mathbb{R}^{n_x},\ Y = \mathbb{R}^{n_y}, \\
& (X \times Y, B(X \times Y)) = \left(\mathbb{R}^{(n_x+n_y)\times(n_x+n_y)}, B(\mathbb{R}^{(n_x+n_y)\times(n_x+n_y)}\right), \\
P_{(x^+,y),t}(A) = {} & P(\{\omega \in \Omega | (x(\omega,t+1), y(\omega,t)) \in A\}), \\
& P_{(x^+,y),t} : B(X \times Y) \rightarrow [0,1],\ P_{(x^+,y),t} \in P(B(X \times Y)); \\
P(B(X \times Y)) \quad & \text{denotes the set of probability measures on } B(X \times Y).
\end{aligned}$$

It follows from Theorem 2.5.4 that $P_{(x+,y),t}$ is a probability measure for all $t \in T$. The invariant measure defined below is always defined on the range space as described above. For a Gaussian system representation with $(x(t+1), y(t))$, the measure on the range space is a Gaussian probability measure on $(\mathbb{R}^{n_x} \times \mathbb{R}^{n_y}, B(\mathbb{R}^{n_x} \times \mathbb{R}^{n_y}))$.

Definition 4.4.1. Consider a stochastic system

$$\{\Omega, F, P, T, Y, B(Y), X, B(X), y, x\}.$$

A probability measure $P_{(x^+,y)} : B(X) \otimes B(Y) \rightarrow [0,1]$ is said to be an *invariant (probability) measure* of the stochastic system if for all $t \in T$ the probability measure $P_{(x^+,y),t}$, induced by the state and output process on $(X \times Y, B(X) \otimes B(Y))$ equals $P_{x+,y}$; thus if,

$$P_{(x^+,y)} = P_{(x^+,y),t}, \;\; \forall t \in T. \tag{4.8}$$

Call then $P_x = P_{(x^+,y)}|_{B(X)}$ the *invariant state (probability) measure* and $P_y = P_{(x^+,y)}|_{B(Y)}$ the *invariant output (probability) measure*.

In case the state and the output spaces are such that $(X,Y) = (\mathbb{R}^{n_x}, \mathbb{R}^{n_y})$ for $n_x, n_y \in \mathbb{Z}_+$ then one associates with an invariant measure an *invariant (probability) distribution function* by,

$$\begin{aligned} P_{(x^+,y)}(A) = &\int_A f_{(x^+,y)}(du,dv), \;\; \forall A \in B(X) \otimes B(Y), \\ &f_{(x^+,y)} : \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \to \mathbb{R}_+, \end{aligned} \tag{4.9}$$

and, correspondingly, to $P_x = P_{x^+,y}|_X$ the invariant pdf of the state f_x and to $P_y = P_{x^+,y}|_Y$ the invariant pdf of the output f_y, respectively.

Moreover, if the invariant distribution function admits a density with respect to Lebesgue measure then one defines the *invariant (probability) density function* as

$$p_{(x^+,y)} : \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \to \mathbb{R}_+, \;\; P_{(x^+,y)}(A) = \int_A p_{(x^+,y)}(u,v)dudv, \tag{4.10}$$

and, correspondingly, p_x for P_x and p_y for P_y.

Proposition 4.4.2. Sufficient condition for a stochastic system. *Consider a forward representation of a stochastic system consisting of:*

1. *a probability measure $P_{x_0} : B(X) \to [0,1]$ of the initial state $x(t_0)$;*
2. *a probabilistic forward-transition function,*
   $f : T \times X \to \text{cpdf}(X \times Y)$; in more detail, for all $t \in T$,

$$\begin{aligned} &f(t,.) : X \to \text{cpdf}(.;(x(t+1),y(t))|F_t^{x-} \vee F_{t-1}^{y-}) \\ &= \text{cpdf}(.;(x(t+1),y(t))|F^{x(t)}). \end{aligned}$$

If the measure $P_{(x^+,y)} : B(X) \otimes B(Y) \to [0,1]$ is such that
(1) the measure induced by x_0 on $B(X) = B(\mathbb{R}^{n_x})$ equals $P_{x_0} = P_{(x^+,y)}|_{B(\mathbb{R}^{n_x})}$ and
(2) for any $t \in T$, the unconditional measure of $(x(t+1),y(t))$ equals $P_{(x,y)}$; then $P_{(x,y)}$ is an invariant measure of the system. The unconditional measure $(x(t+1),y(t))$ is calculated by restriction of the unconditional measure on $(x(t),y(t-1))|x(t-1)$ to the component $x(t)$ and integrating that measure with respect to the transition measure $(x(t+1),y(t))|\, x(t)$.

Proof. This follows directly from the definitions. □

The Invariant Measure of a Forward Gaussian System Representation

Consider a time-invariant Gaussian system. It will be proven that for this system an invariant measure exists, with respect to conditions specified in the theorem below.

Of interest to system representation are the answers to the following questions: (1) What is the support set X_{supp} of the invariant measure? (2) Does the support set equal the state set X? Recall from Example 2.7.8 that the support set of a Gaussian probability measure equals the range of the variance matrix. Thus the support set equals the state set if and only if the variance matrix of the invariant distribution is strictly positive definite. If the support set is a strict subset of the state set than the state set can be reduced in size, which reduces the complexity of the system representation. A state set which equals the support set of the invariant measure is thus of interest. The following concept is therefore motivated.

Definition 4.4.3. Consider the time-invariant Gaussian system,

$$x(t+1) = Ax(t) + Mv(t),\ x(t_0) = x_0.$$

Call the matrix tuple $(A,M) \in \mathbb{R}^{n_x \times n_x} \times \mathbb{R}^{n_x \times n_v}$ a *supportable pair* respectively and a *supportable-stable pair* if it is a controllable pair, see Def. 21.2.6, respectively a stabilizable pair, see Def. 21.2.10. By a characterization of a controllable pair, it is a supportable pair if,

$$n_x = \text{rank}(\text{conmat}(A,M)) = \text{rank}\left(M\ AM\ \ldots\ A^{n_x-1}M\right). \tag{4.11}$$

Call the Gaussian system representation a *supportable representation* or a *supportable-stable representation* if (A,M) is a supportable pair or a supportable-stable pair respectively.

In the above definition, use is made of the assumption that in a Gaussian system representation the variance of the noise process is the identity matrix, $Q_v = I_{n_v}$.

Readers who are familiar with system theory for linear systems recognize equation (4.11) as a controllability condition of an associated linear deterministic system. For stochastic systems the expression *supportable pair* is used rather than *controllable pair* because stochastic controllability will be defined differently in a subsequent chapter of this book.

The usefulness of the concept of a supportable pair and of a supportable-stable is explained below by an example.

Example 4.4.4. *A forward Gaussian system representation which is not supportable but supportable-stable*. Consider the forward system representation,

$$\begin{aligned} x(t+1) = & \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} + \begin{pmatrix} M_1 \\ 0 \end{pmatrix} v(t),\ x(0) = x_0; \\ & x_1(t) \in \mathbb{R}^{n_1},\ x_2(t) \in \mathbb{R}^{n_2},\ n_1,\ n_2 \in \mathbb{N},\ n_1 + n_2 = n; \\ & (A_{11}, M_1) \text{ supportable pair, } \text{spec}(A_{22}) \subset \mathrm{D}_o \end{aligned}$$

Note that by Proposition 21.2.9 any matrix tuple (A,M) can be decomposed as above where a component may be missing if either the pair (A,M) is a supportable pair or if $M = 0$.

Note that then,

$$\begin{aligned} x_2(t+1) = & A_{22}x_2(t),\ x_2(0) = x_{2,0}, \\ & \text{spec}(A_{22}) \subset \mathrm{D}_o \ \Rightarrow\ \text{a.s.} - \lim_{t\to\infty} x_2(t) = 0. \end{aligned}$$

Thus the part of the state set $\mathbb{R}^{n_2}$ is not useful if one is interested in a model of a stochastic system on the infinite horizon $T = \mathbb{N}$.

The subspace $\mathbb{R}^{n_2}$ may in this case be deleted from the state space, hence only the set $\mathbb{R}^{n_1}$ remains as a useful state set. This example motivates the usefulness of the concept of a supportable pair of system matrices.

The concept of spectral index of a state transition matrix will be used below, see Def. 17.4.16. The spectral index of a square state transition matrix $A \in \mathbb{R}^{n_x \times n_x}$ consists of a triple of indices of the form $n_{si}(A) = (n_+, n_1, n_-)$ which represent respectively the number of eigenvalues of the matrix A with modules strictly larger than one, equal to one, and strictly less than one.

The existence of an invariant measure depends on the spectral index:

(a)$n_{si}(A) = (n, 0, 0)$ in which case the system is unstable hence no invariant measure exists;
(b)$n_{si}(A) = (0, n, 0)$ in which case the matrix A has only eigenvalues on the unit circle and the state trajectory will move on a circle in the state space of which the radius depends on the probability distribution of the initial state; and
(c)$n_{si}(A) = (0, 0, n)$ in which case the system matrix is exponentially stable. This case is treated below, an invariant measure exists, and a necessary and sufficient condition for the invariant measure to cover the entire state set is provided in terms of a supportable pair of matrices.

The case (a) and (b) are not discussed further. It is also possible to consider a system matrix in which the types (a), (b), and (c) are mixed but conclusions for such a matrix can be deduced from the three pure cases (a), (b), and (c).

Theorem 4.4.5. *Consider a time-invariant Gaussian system,*

$$\begin{aligned} x(t+1) &= Ax(t) + Mv(t), \;\; x(t_0) = x_0, \\ y(t) &= Cx(t) + Nv(t), \;\; v(t) \in G(0, Q_v). \end{aligned}$$

Assume that the system matrix $A \in \mathbb{R}^{n_x \times n_x}$ is an exponentially stable matrix, or, equivalently, that $\text{spec}(A) \subset \text{D}_o$*, or, equivalently, that the spectral index equals* $n_{si}(A) = (0, 0, n)$.

(a)There exists an invariant measure of the system on the product of the state space and the output space which is Gaussian, denoted by,

$$\text{pdf}(x(t+1), y(t)) \in G(0, Q_{(x^+,y)}), \; (\mathbb{R}^{n_x} \times \mathbb{R}^{n_y}, \, B(\mathbb{R}^{n_x}) \otimes B(\mathbb{R}^{n_y})), \quad (4.12)$$

where $Q_x \in \mathbb{R}^{n_x \times n_x}$ is the unique solution of the following discrete-time Lyapunov equation and the other matrices are provided by the relations,

$$Q_x = AQ_xA^T + MM^T, \; Q_y = CQ_xC^T + NN^T, \quad (4.13)$$

$$Q_{x^+,y} = AQ_xC^T + MN^T, \; Q_{(x^+,y)} = \begin{pmatrix} Q_x & Q_{x^+,y} \\ Q_{x^+,y}^T & Q_y \end{pmatrix}. \quad (4.14)$$

The reader has to distinguish the following expressions,

$$
\begin{aligned}
Q_{(x^+,y)} &= E\left[\begin{pmatrix} x(t+1)-m_x(t+1) \\ y(t)-m_y(t) \end{pmatrix}\begin{pmatrix} x(t+1)-m_x(t+1) \\ y(t)-m_y(t) \end{pmatrix}^T\right], \\
Q_{x^+,y} &= E[(x(t+1)-m_x(t+1))(y(t)-m_y(t))^T], \\
Q_{x,y} &= E[(x(t)-m_x(t))(y(t)-m_y(t))^T].
\end{aligned}
$$

(b)Assume that the probability distribution of the initial state equals the invariant state distribution, $x_0 \in G(0,Q_x)$. From the fact that there exists an invariant measure follows by definition that,

$$
\begin{aligned}
& x_0 \in G(0,Q_x) \;\Rightarrow\; \forall\, t \in T,\ \mathrm{pdf}(x(t+1),y(t)) \in G(0,Q_{(x^+,y)}), \\
& \textit{in particular, } x(t+1) \in G(0,Q_x).
\end{aligned}
$$

Then it follows also that the processes $(x,\ y)$ are jointly stationary, that the state process x is a stationary Gauss-Markov process, and that the coveriance function of these processes are,

$$
\begin{aligned}
W_x(t) &= A^t Q_x,\ \ t \geq 0, \\
W_y(t) &= \begin{cases} CA^{t-1}Q_{x^+y}, & \textit{if } t > 0, \\ Q_y, & \textit{if } t = 0, \end{cases} \\
W_{yx}(t) &= CA^t Q_x,\ \ t \geq 0; \\
& \quad NN^T \succ 0 \;\Rightarrow\; Q_y = W_y(0) = CQ_xC^T + NN^T \succeq NN^T \succ 0.
\end{aligned}
$$

Then also,

$$
\begin{aligned}
& x_0 \in G(0,Q_x),\ w(t) = \begin{pmatrix} M \\ N \end{pmatrix} v(t) \in G(0,Q_w), \\
& (F,\ G,\ H,\ J+J^T) = (A,\ Q_{x^+,y},\ C,\ Q_y), \\
W_y(t) &= \begin{cases} HF^{t-1}G, & t > 0, \\ Q_y = CQ_xC^T + NN^T = J + J^T, & t = 0, \end{cases} \\
& \begin{pmatrix} Q_x - FQ_xF^T & G - FQ_xH^T \\ G^T - HQ_xF^T & J + J^T - HQ_xH^T \end{pmatrix} \\
&= \begin{pmatrix} Q_x - AQ_xA^T & Q_{x^+,y} - AQ_xC^T \\ Q_{x^+,y}^T - CQ_xA^T & Q_y - CQ_xC^T \end{pmatrix} = \begin{pmatrix} M \\ N \end{pmatrix}\begin{pmatrix} M \\ N \end{pmatrix}^T = Q_w \succeq 0, \\
& \Rightarrow\ Q_x \in \mathbf{Q_{lsdp}},\ \textit{see Def. 24.1.1.}
\end{aligned}
$$

(c)Assume next that the probability distribution of the initial state does not equal to the invariant probability distribution of the state process, $x_0 \notin G(0,Q_x)$. Then the measure induced by $(x(t+1),y(t))$ on $(\mathbb{R}^{n_x} \times \mathbb{R}^{n_y},\ B(\mathbb{R}^{n_x} \times \mathbb{R}^{n_y})$ converges in distribution to the invariant measure on the state set and the output set, with,

$$
\mathrm{D}-\lim_{t\to\infty} G\left(\begin{pmatrix} m_x(t) \\ m_y(t) \end{pmatrix}, \begin{pmatrix} Q_{x^+}(t) & Q_{x^+,y}(t) \\ Q_{x^+,y}(t)^T & Q_y(t) \end{pmatrix}\right) = G(0,Q_{(x^+,y)}).
$$

(d)The following statements are equivalent:

(d.1)The support of the state process equals the state space $X = \mathbb{R}^{n_x}$.

(d.2)The invariant Gaussian measure of the state process $G(0,Q_x)$ *satisfies* $Q_x \succ 0$.
(d.3)The matrix tuple (A,M) *is a supportable pair.*

Proof. (a) From the system representation follows for all $t \in T$ that if $x(t) \in G(0,Q_x)$ then

$$\begin{aligned}
&E\left[\exp\left(i\begin{pmatrix}w_x\\w_y\end{pmatrix}^T\begin{pmatrix}x(t+1)\\y(t)\end{pmatrix}\right)|F_t^x\right]\\
&= \exp(i\begin{pmatrix}w_x\\w_y\end{pmatrix}^T\begin{pmatrix}Ax(t)\\Cx(t)\end{pmatrix}-\frac{1}{2}\begin{pmatrix}w_x\\w_y\end{pmatrix}^T\begin{pmatrix}M\\N\end{pmatrix}\begin{pmatrix}M\\N\end{pmatrix}^T\begin{pmatrix}w_x\\w_y\end{pmatrix}),\\
&\quad E\left[\exp\left(i\begin{pmatrix}w_x\\w_y\end{pmatrix}^T\begin{pmatrix}x(t+1)\\y(t)\end{pmatrix}\right)\right]\\
&= E\left[E\left[\exp\left(i\begin{pmatrix}w_x\\w_y\end{pmatrix}^T\begin{pmatrix}x(t+1)\\y(t)\end{pmatrix}\right)|F_t^x\right]\right]\\
&= \exp\left(-\frac{1}{2}\begin{pmatrix}w_x\\w_y\end{pmatrix}^T\begin{pmatrix}AQ_xA^T+MM^T & AQ_xC^T+MN^T\\(AQ_xC^T+MN^T)^T & CQ_xC^T+NN^T\end{pmatrix}\begin{pmatrix}w_x\\w_y\end{pmatrix}\right)\\
&= \exp\left(-\frac{1}{2}\begin{pmatrix}w_x\\w_y\end{pmatrix}^T\begin{pmatrix}Q_x & Q_{x^+,y}\\Q_{x^+,y}^T & Q_y\end{pmatrix}\begin{pmatrix}w_x\\w_y\end{pmatrix}\right),
\end{aligned}$$

where the latter equality follows from the equations (4.13, 4.14). Thus the measure (4.12) is invariant. In the above use has been made of the fact that the matrix Q_x is a solution of the discrete-time Lyapunov equation and of the definitions of $Q_{x^+,y}$ and Q_y. The result then follows from Proposition 4.4.2. The formulas for the covariance function follow directly from Theorem 4.3.8.(e). and (f). Because $W_x(t+s,s) = W_x(t,0)$ for all $t,s \in T$, $W_y(t+s,s) = W_y(t,0)$, and $W_{(x^+,y)}(t+s,s) = W_{(x^+,y)}(t,0)$ it follows from Proposition 3.4.4 that the process $(x,\ y)$ is jointly stationary.
(b) This is a direct calculation using the conditional probability distribution of the transition.
(c) The indicated convergence holds if and only if

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

$$\begin{aligned}
\lim_{s\to\infty} W_x(t+s,s) &= A^tQ_x,\\
\lim_{s\to\infty} W_y(t+s,s) &= \begin{cases} CA^{t-1}Q_{x^+y}, & t>0,\\ Q_y, & t=0,\end{cases}\\
\lim_{s\to\infty} W_{xy}(t+s,s) &= CA^tQ_{x^+,y}.
\end{aligned}$$

But from Theorem 4.3.8 follows that, because $\text{spec}(A) \subset \text{D}_o$,

$$\begin{aligned}
&\lim_{t\to\infty} m_x(t) = A^tm_x(0) = 0,\ \ \lim_{t\to\infty} m_y(t) = CA^tm_x(0) = 0,\\
&\lim_{t\to\infty}\ Q_x(t) = Q_x,\\
&\lim_{t\to\infty}\ Q_y(t) = \lim\ CQ_x(t)C^T + NN^T = CQ_xC^T + NN^T = Q_y,
\end{aligned}$$

$$\lim_{s\rightarrow\infty} W_x(t+s,s) = \lim_{s\rightarrow\infty} A^t Q_x(s) = A^t Q_x, \ \forall \ t \in T,$$
$$\lim_{s\rightarrow\infty} \ W_y(t+s,s) = \lim \ CA^{t-1}[AQ_x(s)C^T + MN^T]$$
$$= CA^{t-1}[AQ_xC^T + MN^T] = CA^{t-1}Q_{x^+,y}.$$

(d) This follows from Theorem 22.1.2.(d) for the matrix Lyapunov equation. □

The Invariant Measure of a Backward Gaussian System

Definition 4.4.6. Consider a time-invariant backward Gaussian system representation,

$$x(t-1) = A_b x(t) + M_b v_b(t), \ x(t_1) = x_1.$$

Call the matrix tuple $(A_b, M_b) \in \mathbb{R}^{n_x \times n_x} \times \mathbb{R}^{n_x \times n_{v_b}}$ a *backward supportable pair* respectively a *backward supportable-stable pair* if it is a controllable pair, see Def. 21.2.6, respectively a stabilizable pair, see Def. 21.2.10.

According to a characterization of a controllable pair, the tuple (A_b, M_b) is a supportable pair if and only if,

$$n_x = \text{rank}(\text{conmat}(A_b, M_b)) = \text{rank}\left(\left(\, M_b \ A_bM_b \ \ldots \ A_b^{n_x-1}M_b \,\right)\right).$$

Call the backward Gaussian system representation *backward supportable* if (A_b, M_b) is a backward supportable pair and *backward supportable-stable* if (A_b, M_b) is a backward supportable-stable pair.

In the above definition, use is made of the assumption that in a Gaussian system representation the variance of the noise process is the identity matrix, $Q_{v_b} = I_{n_{v_b}}$.

Theorem 4.4.7. *Consider a time-invariant backward Gaussian system,*

$$\begin{aligned} x(t-1) &= A_b x(t) + M_b v_b(t), \ \ x(0) = x_0, \\ y(t-1) &= C_b x(t) + N_b v_b(t), \ \ v_b(t) \in G(0,I), \\ T &= -\mathbb{N} = \{\ldots, \ -2, \ -1, \ 0\}. \end{aligned}$$

Assume that the system matrix $A_b \in \mathbb{R}^{n_x \times n_x}$ is an exponentially stable matrix, or, equivalently, that $\text{spec}(A_b) \subset \text{D}_o$.

(a)There exists an invariant measure of the system on the product of the state space and the output space which is Gaussian, denoted by,

$$\text{pdf}(x(t-1), y(t-1)) \in G(0, Q_{(x,y)}), \ (\mathbb{R}^{n_x} \times \mathbb{R}^{n_y}, \ B(\mathbb{R}^{n_x}) \otimes B(\mathbb{R}^{n_y})), \ (4.15)$$

where $Q_x \in \mathbb{R}^{n_x \times n_x}$ is the unique solution of the following discrete-time Lyapunov equation and the other matrices are provided by the relations,

$$Q_x = A_b Q_x A_b^T + M_b M_b^T, \ Q_y = C_b Q_x C_b^T + N_b N_b^T, \quad (4.16)$$

$$Q_{(x,y)} = \begin{pmatrix} Q_x & Q_{x,y} \\ Q_{x,y}^T & Q_y \end{pmatrix}, \ \ Q_{x,y} = A_b Q_x C_b^T + M_b N_b^T. \quad (4.17)$$

The reader has to distinguish,

$$\begin{aligned} Q_{x,y} &= E[(x(t-1)-E[x(t-1)])(y(t-1)-E[y(t-1)])^T], \textit{ from,} \\ Q_{(x,y)} &= E\left[\begin{pmatrix} x(t)-E[x(t)] \\ y(t)-E[y(t)] \end{pmatrix} \begin{pmatrix} x(t)-E[x(t)] \\ y(t)-E[y(t)] \end{pmatrix}^T \right]. \end{aligned}$$

(b)Assume that the probability distribution function of the initial state equals the invariant measure, $x_0 \in G(0,Q_x)$. *Then, by definition of the invariant measure of the time-invariant Gaussian system,*

$$\begin{aligned} &x_0 \in G(0,Q_x) \ \Rightarrow \ \forall\, t \in T, \ x(t-1), y(t-1) \in G(0,Q_{(x,y)}), \\ &\textit{in particular, } x(t) \in G(0,Q_x). \end{aligned}$$

Then the state process and the output process are jointly stationary Gaussian processes with covariance functions,

$$\begin{aligned} W_x(t) &= A_b^{t_1-t} Q_x, \ \ t<0, \\ W_y(t) &= \begin{cases} C_b A_b^{t_1-t} Q_{xy}, & \textit{if } t<0, \\ Q_y, & \textit{if } t=0, \end{cases} \\ W_{yx}(t) &= \begin{cases} Q_{x,y}^T, & t=0, \\ C_b A_b^{-t-1} Q_x, & t<0; \end{cases} \\ & N_b N_b^T \succ 0 \ \Rightarrow \ Q_y = W_y(0) = C_b Q_x C_b^T + N_b N_b^T \succeq N_b N_n^T \succ 0. \end{aligned}$$

Then also,

$$\begin{aligned} & x_0 \in G(0,Q_x), \ w_b(t) = \begin{pmatrix} M_b \\ N_b \end{pmatrix} v_b(t), \\ W_y(t) \ &= \ \begin{cases} C_b A_b^{-t-1} Q_{x,y} = G^T (F^T)^{-t-1} H^T, & \textit{if } t<0, \\ C_b Q_x C_b^T + N_b N_b^T = J + J^T, & \textit{if } t=0; \end{cases} \\ & (F^T,\, H^T,\, G^T,\, J+J^T) = (A_b,\, C_b,\, Q_{x,y},\, Q_y), \\ & \begin{pmatrix} Q_x - F^T Q_x F & Q_{x,y} - F^T Q_x G \\ Q_{x,y}^T - G^T Q_x F & Q_y - G^T Q_x G \end{pmatrix} \\ &= \begin{pmatrix} Q_x - A_b Q_x A_b^T & Q_{x,y} - A_b Q_x C_b^T \\ Q_{x,y}^T - C_b Q_x A_b^T & Q_y - C_b Q_x C_b^T \end{pmatrix} = \begin{pmatrix} M_b \\ N_b \end{pmatrix} = Q_{w_b} \succeq 0, \\ & \Rightarrow Q_x \in \mathbf{Q_{lsp}}, \ \textit{see Def. 24.1.1.} \end{aligned}$$

Note that the two sets $\mathbf{Q_{lsp}}$ *and* $\mathbf{Q_{lsdp}}$ *are different.*

(c)If the measure of the terminnal state x_1 *is not equal to the invariant measure then the measure induced by* $(x(t-1),y(t-1))$ *on* $B(\mathbb{R}^{n_x}) \otimes B(\mathbb{R}^{n_y})$ *converges in distribution to the invariant measure on the state set and the output set, with*

$$\mathrm{D}-\lim_{t\to\infty} G\left(\begin{pmatrix} m_x(t) \\ m_y(t) \end{pmatrix}, \begin{pmatrix} Q_x(t) & Q_{x,y}(t) \\ Q_{x,y}(t)^T & Q_y(t) \end{pmatrix} \right) = G(0,Q_{(x,y)}).$$

(d)The following statements are equivalent:

(d.1)The support of the state process equals the state space $X = \mathbb{R}^{n_x}$.

(d.2)The invariant Gaussian measure of the state process $G(0,Q_x)$ satisfies $Q_x \succ 0$.
(d.3)The matrix tuple (A_b,M_b) is a backward supportable pair.

The proof of the above theorem is analogous to that of Theorem 4.4.5 and therefore omitted.

Transformations of a Gaussian System

The reader finds in this section several transformations of a Gaussian system which are used subsequently in this book. These transformations are related to the inverse of a linear control system, see Section 21.6.

Definition 4.4.8. *Space transformations of a time-invariant Gaussian system*. Consider a time-invariant Gaussian system representation of the form,

$$\begin{aligned} x(t+1) &= Ax(t) + Mv(t),\ x(0) = x_0, \\ y(t) &= Cx(t) + Nv(t). \end{aligned}$$

Define state-space transformations of the state space $X = \mathbb{R}^{n_x}$, of the output space $Y = \mathbb{R}^{n_y}$, and of the noise space $V = \mathbb{R}^{n_v}$. Choose transformation matrices and define new stochastic processes and consequently obtain the new state-space representation,

$$\begin{aligned} & L_x \in \mathbb{R}^{n_x \times n_x}_{nsng},\ L_y \in \mathbb{R}^{n_y \times n_y}_{nsng},\ U_v \in \mathbb{R}^{n_v \times n_v}_{ortg}, \\ & \overline{x}(t) = L_x x(t),\ \overline{y}(t) = L_y y(t),\ \overline{v}(t) = Uv(t), \\ \overline{x}(t+1) &= L_x A L_x^{-1} \overline{x}(t) + L_x M U_v^T \overline{v}(t),\ \overline{x}(0) = L_x x_0, \\ \overline{y}(t) &= L_y C L_x^{-1} \overline{x}(t) + L_y N U^T \overline{v}(t), \\ & (A,C,M,N) \ \mapsto \ (L_x A L_x^{-1}, L_y C L_x^{-1}, L_x M U_v^T, L_y N U_v^T). \end{aligned}$$

Note that $E[\overline{v}(t)] = 0$ and $Q_{\overline{v}} = UQ_vU^T = UU^T = I$ because $U \in \mathbb{R}^{n_v \times n_v}_{ortg}$.

Definition 4.4.9. *Transformation of the inverse of a Gaussian system in a special case*. Consider a time-invariant Gaussian system with representation,

$$\begin{aligned} x(t+1) &= Ax(t) + Mv(t),\ x(0) = x_0, \\ y(t) &= Cx(t) + Nv(t),\ \text{assume } n_v = n_y,\ N \in \mathbb{R}^{n_v \times n_v},\ \text{rank}(N) = n_v. \end{aligned}$$

Define the inverse of this Gaussian system as the system with representation,

$$\begin{aligned} x(t+1) &= (A - MN^{-1}C)x(t) + MN^{-1}y(t),\ x(0) = x_0, \\ v(t) &= -N^{-1}Cx(t) + N^{-1}y(t), \\ & (A - MN^{-1}C, -N^{-1}C, MN^{-1}, N^{-1}). \end{aligned}$$

Note that the above system representation is not a Gaussian system representation because in general y is not a Gaussian white noise process.

Proposition 4.4.10. *Consider the transformation of the inverse of a Gaussian system in a special case of Def. 4.4.9. The representation of the inverse system is well defined, and both a left-inverse and a right-inverse of the original system.*

Proof. Because $n_v = n_y$, $N \in \mathbb{R}^{n_v \times n_v}$, and $\text{rank}(N) = n_v$, the inverse matrix $N^{-1} \in \mathbb{R}^{n_v \times n_v}$ exists. Then,

$$\begin{aligned}
N^{-1}y(t) &= N^{-1}Cx(t) + v(t),\ v(t) = N^{-1}y(t) - N^{-1}Cx(t),\\
x(t+1) &= Ax(t) + Mv(t) = Ax(t) + MN^{-1}y(t) - MN^{-1}Cx(t)\\
&= (A - MN^{-1}C)x(t) + MN^{-1}y(t).
\end{aligned}$$

That the second system is both a left-inverse and a right-inverse of the first system follows from the following calculations and induction,

$$\begin{aligned}
x(t+1) &= Ax(t) + Mv(t),\ x(0) = x_0,\\
y(t) &= Cx(t) + Nv(t),\\
\bar{x}(t+1) &= (A - MN^{-1}C)\bar{x}(t) + MN^{-1}\bar{y}(t),\ \bar{x}(0) = x_0,\\
\bar{v}(t) &= -N^{-1}C\bar{x}(t) + N^{-1}\bar{y}(t),\\
&\quad \bar{x}_0 = x_0 \ \Rightarrow\ \forall\, s \in T,\ \bar{x}(s) = x(s);\\
&\quad \text{to prove that the second system is a left-inverse of first system}\\
&\quad \text{assume that } y(t) = \bar{y}(t),\ \forall\, t \in T;\ \text{then,}\\
\bar{v}(t) &= -N^{-1}C\bar{x}(t) + N^{-1}\bar{y}(t) = -N^{-1}Cx(t) + N^{-1}y(t)\\
&= -N^{-1}Cx(t) + N^{-1}[Cx(t) + Nv(t)] = v(t);
\end{aligned}$$

$$\begin{aligned}
&\quad \text{to prove that the second system is a right-inverse of first system}\\
&\quad \text{assume that } v(t) = \bar{v}(t),\ \forall\, t \in T;\ \text{then,}\\
y(t) &= Cx(t) + Nv(t) = C\bar{x}(t) + N\bar{v}(t)\\
&= C\bar{x}(t) + N[-N^{-1}C\bar{x}(t) + N^{-1}\bar{y}(t)] = \bar{y}(t), \forall\, t \in T.
\end{aligned}$$

□

4.5 Relation of Forward and Backward Gaussian System Representations

In this subsection it is proven that a Gaussian system has both a forward and a backward representation and it is described how the system matrices of these two representations are related. That either a forward and a backward representation are possible follows from the abstract definition of a stochastic system, see Section 4.2.

Theorem 4.5.1. *Let*

$$\begin{aligned}
&\{\Omega, F, P, T, \mathbb{R}^{n_y}, B(\mathbb{R}^{n_y}), \mathbb{R}^{n_x}, B(\mathbb{R}^{n_x}), y, x\} \in \text{GStocS},\\
&T = T(0:t_1) = \{0, 1, \ldots, t_1\},
\end{aligned}$$

be a Gaussian system. Assume that for all $t \in T$, $E[x(t)] = 0$, $E[y(t)] = 0$, and that $Q_x : T \to \mathbb{R}^{n \times n}$ satisfies for all $t \in T$, $Q_x(t) = E[x(t)x(t)^T] \succ 0$.

(a)*There exists a* forward Gaussian system representation *of the form,*

$$\begin{aligned}
x(t+1) &= A_f(t)x(t) + M_f v_f(t), x_0,\\
y(t) &= C_f(t)x(t) + N_f v_f(t);\\
& \quad v_f : \Omega \times T \to \mathbb{R}^{n_x+n_y}.\\
A_f(t) &= E[x(t+1)x(t)^T]Q_x(t)^{-1},\ C_f(t) = E[y(t)x(t)^T]Q_x(t)^{-1},\\
v_f(t) &= \begin{pmatrix} x(t+1) - A_f(t)x(t) \\ y(t) - C_f(t)x(t) \end{pmatrix},\ v_f(t) \in G(0, Q_{v_f}(t)),\ \forall\, t \in T,\\
Q_{v_f}(t) &= \begin{pmatrix} Q_x(t+1) & E[x(t+1)y(t)^T] \\ E[y(t)x(t+1)^T] & Q_y(t) \end{pmatrix}\\
& \quad - \begin{pmatrix} A_f(t) \\ C_f(t) \end{pmatrix} Q_x(t)^{-1} \begin{pmatrix} A_f(t) \\ C_f(t) \end{pmatrix}^T,\\
M_f &= \begin{pmatrix} I_{n_x} & 0 \end{pmatrix} \in \mathbb{R}^{n_x \times (n_x+n_y)},\ N_f = \begin{pmatrix} 0 & I_{n_y} \end{pmatrix} \in \mathbb{R}^{n_y \times (n_x+n_y)},
\end{aligned}$$

Note that for all $t \in T$, $Q_{v_f}(t) \in \mathbb{R}_{pds}^{n_{v_f} \times n_{v_f}}$ and that F^{x_0} and $F_{t_1-1}^{v_f}$ are independent. Conversely, consider a forward Gaussian system representation with $A_f, C_f, M_f, N_f, Q_{v_f}$ functions and x, y defined by the above forward representation. Then the forward difference representation defines a Gaussian system.

(b)*There exists a* backward Gaussian system representation *of the form,*

$$\begin{aligned}
x(t-1) &= A_b(t)x(t) + M_b v_b(t), x(t_1) = x_1,\\
y(t-1) &= C_b(t)x(t) + N_b v_b(t);
\end{aligned}$$

where $v_b : \Omega \times T \to \mathbb{R}^{n_x+n_y}$ is a Gaussian white noise process with intensity Q_{v_b}, and $x_1 : \Omega \to \mathbb{R}^{n_x}$, $x_1 \in G(0, Q_{x_1})$, with F^{x_1} and $F_{t_1}^{v_b}$ independent; and

$$\begin{aligned}
A_b(t) &= E[x(t-1)x(t)^T]Q_x(t)^{-1},\ C_b(t) = E[y(t-1)x(t)^T]Q_x(t)^{-1},\\
v_b(t) &= \begin{pmatrix} x(t-1) - A_b(t)x(t) \\ y(t-1) - C_b(t)x(t) \end{pmatrix},\ v_b(t) \in G(0, Q_{v_b}(t)),\ \forall\, t \in T,\\
Q_{v_b}(t) &= \begin{pmatrix} Q_x(t-1) & E[x(t-1)y(t-1)^T] \\ E[y(t-1)x(t-1)^T] & E[y(t-1)y(t-1)^T] \end{pmatrix},\\
& \quad - \begin{pmatrix} A_b(t) \\ C_b(t) \end{pmatrix} Q_x(t)^{-1} \begin{pmatrix} A_b(t) \\ C_b(t) \end{pmatrix}^T,\\
M_b &= \begin{pmatrix} I_{n_x} & 0 \end{pmatrix} \in \mathbb{R}^{n_x \times (n_x+n_y)},\ N_b = \begin{pmatrix} 0 & I_{n_y} \end{pmatrix} \in \mathbb{R}^{n_y \times (n_x+n_y)}.
\end{aligned}$$

Conversely, consider a backward Gaussian system representation with $A_b, C_b, Q_{v_b}, M_b, N_b$ and x, y as defined by the above backward representation. Then this backward system representation defines a Gaussian system.

(c)*The relation between the forward and the backward representation of a Gaussian system is specified by the equations,*

$$\begin{aligned}
A_f(t)Q_x(t) &= Q_x(t+1)(A_b(t+1))^T, \\
C_b(t)Q_x(t) &= C_f(t-1)Q_x(t-1)(A_f(t-1))^T + N_b Q_{v_f}(t-1)M_b^T, \\
C_f(t)Q_x(t) &= C_b(t+1)Q_x(t+1)(A_b(t+1))^T + N Q_{v_b}(t+1)M_b^T.
\end{aligned}$$

Proof. (a) Consider the collection in GStocS and $t \in T$. Then

$$E\left[\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix} | F_t^{x-} \vee F_{t-1}^{y-}\right] = E\left[\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix} | F^{x(t)}\right] = \begin{pmatrix} A_f(t) \\ C_f(t) \end{pmatrix} x(t),$$

$$\begin{aligned}
&E\left[\exp\left(iw^T \begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix}\right) | F_t^{x-} \vee F_{t-1}^{y-}\right] \\
&= \exp\left(iw^T \begin{pmatrix} A_f(t) \\ C_f(t) \end{pmatrix} x(t) - \frac{1}{2} w^T Q_{v_f}(t) w\right),
\end{aligned}$$

because the collection is a Gaussian system, because (x,y) is jointly Gaussian, hence $(x(t+1),\ y(t),\ x(t))$ are jointly Gaussian random variables, and by Theorem 2.8.3. The matrix function $Q_{v_f} : T \rightarrow \mathbb{R}^{n_{v_f} \times n_{v_f}}$ follows from the calculation. Define the process $v_f : \Omega \times T \rightarrow \mathbb{R}^{n_x+n_y}$ as in the theorem statement. The forward representation then follows from the definition of M_f, N_f. Also

$$\begin{aligned}
&E[\exp(iw^T v_f(t)) | F_{t-1}^{v_f-}] \\
&= E[E[\exp(iw^T \begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix}) | F_t^{x-} \vee F_{t-1}^{y-}] \exp(iw^T \begin{pmatrix} -A_f(t)x(t) \\ -C_f(t)x(t) \end{pmatrix}) | F_{t-1}^{v_f-}] \\
&= \exp(-\frac{1}{2} w^T Q_{v_f}(t) w) = E[\exp(iw^T v_f(t))],
\end{aligned}$$

by Theorem 2.8.3. Thus $v_f(t)$ is independent of $F_{t-1}^{v_f}$, by Theorem 2.8.2.(f). This proves that v_f is a Gaussian white noise process with intensity $Q_{v_f}(t)$.

The converse is a direct calculation,

$$\begin{aligned}
&E[\exp\left(iw^T \begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix}\right) | F_t^{x-} \vee F_{t-1}^{y-}] \\
&= E[E[\exp(iw^T v_f(t)) | F_{t-1}^{v_f-}] | F_t^{x-} \vee F_{t-1}^{y-}] \exp(iw^T \begin{pmatrix} A_f(t)x(t) \\ C_f(t)x(t) \end{pmatrix})) \\
&= \exp(iw^T \begin{pmatrix} A_f(t)x(t) \\ C_f(t)x(t) \end{pmatrix} - \frac{1}{2} w^T Q_{v_f}(t) w)
\end{aligned}$$

and one concludes with Proposition 4.2.4. The formula's for A_f, C_f, Q_{v_f} follow from Proposition 2.8.3.

(b) This proof is similar to that of (a).

(c)

$$\begin{aligned}
&A_f(t)Q_x(t) = E[x(t+1)x(t)^T], \text{by (a)} \\
&= (E[x(t)x(t+1)^T])^T = (A_b(t+1)Q_x(t+1))^T, \text{ by (b)} \\
&= Q_x(t+1)(A_b(t+1))^T.
\end{aligned}$$

$$\begin{aligned}
C_b(t)Q_x(t) &= E[y(t-1)x(t)^T], \text{by (b)}, \\
&= E[y(t-1)x(t-1)^T](A_f(t-1))^T + \\
&\quad +E[(C_f(t-1)x(t-1)+N_b v_f(t-1))(v_f(t-1))^T M_f^T], \text{by (a)}, \\
&= C_f(t-1)Q_x(t-1)(A_f(t-1))^T + N_f Q_{v_f}(t-1)M_f^T. \\
C_f(t)Q_x(t) &= E[y(t)x(t)^T], \text{by (a)}, \\
&= E[y(t)x(t+1)^T(A_b(t+1))^T] + \\
&\quad +E[(C_b(t+1)x(t+1)+N_b v_b(t+1))(v_b(t+1))^T M_b^T], \text{by (b)} \\
&= C_b(t+1)Q_x(t+1)(A_b(t+1))^T + N_b Q_{v_b}(t+1)M_b^T.
\end{aligned}$$

□

Theorem 4.5.2. *Consider a time-invariant Gaussian system and the associated forward and backward system representation of Theorem 4.5.1 with the notation introduced there.*

Then $A_f, C_f, Q_{v_f}, A_b, C_b, Q_{v_b}$, do not depend explicitly on $t \in T$. Assume that the system matrix A_f is exponentially stable, $\text{spec}(A_f) \subset \mathrm{D}_o$. *Then there exists a unique solution $Q_x \in \mathbb{R}_{pds}^{n_x \times n_x}$ for the state variance matrix to the Lyapunov equation of Theorem 4.4.5.(a). Recall the assumption of Theorem 4.5.1 that* $0 \prec Q_x$.

(a)The relation between the forward and the backward Gaussian system representation is then given by,

$$\begin{aligned}
A_f &= Q_x(A_b)^T Q_x^{-1}, \\
C_f &= C_b Q_x (A_b)^T Q_x^{-1} + N_b Q_{v_b} M_b^T Q_x^{-1} = C_b A_f + N_b Q_{v_b} M_b^T Q_x^{-1}, \\
A_b &= Q_x(A_f)^T Q_x^{-1}, \\
C_b &= C_f Q_x (A_f)^T Q_x^{-1} + N_f Q_{v_f} M_f^T Q_x^{-1} = C_f A_b + N_f Q_{v_f} M_f^T Q_x^{-1}.
\end{aligned}$$

In the above system representations, the transformation to standard stationary Gaussian white noise is not described hence Q_{v_b} and Q_{v_f} need not be unit matrices.

From the matrices of a forward representation one can compute the system matrix and the output matrix of the backward representation by the formulas,

$$\begin{aligned}
A_b &= Q_x(A_f)^T Q_x^{-1}, \\
C_b &= (A_f Q_x C_f^T + M_f Q_{v_f} N_f^T)^T Q_x^{-1} = (Q_{x_f^+, y})^T Q_x^{-1}.
\end{aligned}$$

(b)Note that the spectra of the forward and backward system matrices are identical, $\text{spec}(A_f) = \text{spec}(A_b)$. *Hence, if* $\text{spec}(A_f) \subset \mathrm{D}_o$ *then* $\text{spec}(A_b) \subset \mathrm{D}_o$ *and conversely.*

(c)Equivalence holds of (1) $\text{spec}(A_f) \subset \mathrm{D}_o$ *and (A_f, M_f) a supportable pair; and (2)* $\text{spec}(A_b) \subset \mathrm{D}_o$ *and (A_b, M_b) a supportable pair.*

Proof. (a) The results follow directly from Theorem 4.5.1.(c).

(b)That the spectra of the forward and backward system matrices are identical follows because,

$$\det(sI - A_b) = \det(sI - Q_x(A_f)^T Q_x^{-1}) = \det(Q_x(sI - (A_f)^T)Q_x^{-1})$$
$$= \det(Q_x)\det(Q_x^{-1})\det(sI - (A_f)^T) = \det(sI - A_f).$$

(c) The conditions (1) and Theorem 22.1.2.(d) imply that the state variance matrix $Q_x \in \mathbb{R}^{n_x \times n_x}$ satisfies $0 \prec Q_x$, where $Q_x = A_f Q_x A_f^T + M_f M_f^T$. Then $0 \prec Q_x$, $\text{spec}(A_b) \subset \text{D}_o$, $Q_x = A_b Q_x A_b^T + M_b M_b^T$, and the same theorem, imply that (A_b, M_b) is a supportable pair. The proof of the converse direction is analogous. □

In the following sections the subindices f and b will be omitted when it is clear from the context which representation is referred to. Because forward difference representations of Gaussian systems are used most often in these notes, the adjective forward will be omitted when confusion cannot arise. The concept of a forward difference representation then coincides with that presented in Def. 4.2.2.

The combination of a forward and a backward Gaussian system has also been defined. One then obtains a concept related to a Hamiltonian linear system. Such systems are used in scattering theory, see [15, 34, 72] This concept is not further explored in this book due to space limitations.

4.6 Stochastic Observability and Stochastic Co-Observability

The concepts of stochastic observability and stochastic co-observability are basic concepts of control and system theory. They are used in Chapter 6 for stochastic realization of Gaussian systems, in Kalman filtering, and in stochastic control with partial observations. Below these two concepts are defined and subsequently characterized for a Gaussian system.

4.6.1 Observability and Co-Observability

As an introduction, the concept of observability is first introduced for a map and subsequently for a deterministic linear system. Consider first the simple setting of sets and maps.

Definition 4.6.1. Consider the sets and maps,

$$(Y,\, X,\, U,\, g,\, h),$$
$$Y,\, X,\, U \text{ are sets},\; g : U \to X,\; h : X \to Y \text{ are maps}.$$

Call the observation map h *observable* if the map h is injective; equivalently, if for all $x_a,\, x_b \in X$, $h(x_a) = h(x_b) \in Y$ implies that $x_a = x_b$. See Definition 17.1.11 for the formal definition of an injective and a surjective map.

A direct consequence of an injective map is that from the value $h(x_a) \in Y$ one can uniquely determine the corresponding argument $x_a \in X$.

Proposition 4.6.2. *Consider the linear function $h(x) = Cx$ for $X = \mathbb{R}^{n_x}$, $Y = \mathbb{R}^{n_y}$, and for $h : X \to Y$, with $C \in \mathbb{R}^{n_y \times n_x}$. This map is injective if and only if*

$$\ker(C) = \{x_a \in \mathbb{R}^{n_y} |\, Cx_a = 0\} = \{0\} \;\; \Leftrightarrow \; \text{rank}(C) = n_x.$$

Proof. Note that,

$$\begin{aligned}
& h(x) = Cx \text{ is injective} \\
\Leftrightarrow & \;\forall\, x_a,\, x_b \in X,\; Cx_a = Cx_b \;\Rightarrow x_a = x_b, \\
\Leftrightarrow & \;\forall\, x_a \in X,\; Cx_a = 0 \;\Rightarrow x_a = 0 \;\Leftrightarrow\; \ker(C) \subseteq \{0\}, \\
\Leftrightarrow & \;\ker(C) = \{0\}, \text{ the converse inclusion is always true}, \\
\Leftrightarrow & \;\text{rank}(C) = n_x; \text{ by linear algebra}, \\
& \text{rank}(C) + 0 = \dim(\text{Range}(C)) + \dim(\ker(C)) \\
= & \;\dim(\text{Domain}(C)) = n_x.
\end{aligned}$$

For the formula of linear algebra used in the proof, see Proposition **??**. □

In case injectivity holds then it is possible, for example by using the singular value decompostion, to compute a state x_a such that $y_a = h(x_a) = Cx_a$.

Next observability of a deterministic linear system is discussed.

Definition 4.6.3. Consider a deterministic time-varying linear system without input with both a forward and a backward representation

$$\begin{aligned}
x(t+1) &= A_f(t)x(t),\; x(0) = x_0, \\
y(t) &= C_f(t)x(t), \\
x(t-1) &= A_b(t)x(t),\; x(0) = x_0, \\
y(t-1) &= C_b(t)x(t),\;\; x : T \to \mathbb{R}^{n_x},\; y : T \to \mathbb{R}^{n_y}.
\end{aligned}$$

Call this linear system *observable from the future outputs on the interval* $\{t_0,\; t_0+1,\; \ldots,\; t_0+t_1-1\} \subseteq T$ if the following *state-to-future-output map of the interval* is injective,

$$x(t_0) \mapsto \{y(t_0),\, y(t_0+1),\; \ldots,\; y(t_0+t_1-1)\}\,.$$

It is to be understood that, because the map is injective, injectivity must hold for all initial states in $\mathbb{R}^{n_x}$.

Call this linear system *co-observable from the past outputs on the interval* $\{t_0-1,\; t_0-2,\; \ldots,\; t_0-t_1\} \subseteq T$ if the following *state-to-past-output map of the interval* is injective,

$$x(t_0) \mapsto \{y(t_0-1),\, y(t_0-2),\; \ldots,\; y(t_0-t_1)\}\,.$$

Define the *observability matrix* respectively the *co-observability matrix* on the intervals described above by the formulas

$$
\begin{aligned}
&O_f(A_f,\, C_f,\, t_0 : t_0+t_1-1) \\
&= \begin{pmatrix} C_f(t_0) \\ C_f(t_0+1)\Phi_f(t_0+1,t_0) \\ \vdots \\ C_f(t_0+t_1-1)\Phi_f(t_0+t_1-1,t_0) \end{pmatrix} \in \mathbb{R}^{t_1 n_y \times n_x}, \\
&O_b(A_b,\, C_b,\, t_0-1 : t_0-t_1) \\
&= \begin{pmatrix} C_b(t_0) \\ C_b(t_0-1)\Phi_b(t_0-1,t_0) \\ \vdots \\ C_b(t_0-t_1+1)\Phi_b(t_0-t_1+1,t_0) \end{pmatrix} \in \mathbb{R}^{t_1 n_y \times n_x},
\end{aligned}
$$

Denote the vectors of future outputs and of past outputs for the intervals by

$$
\begin{aligned}
&y(t_0 : t_0+t_1-1) \\
&= \begin{pmatrix} y(t_0) \\ y(t_0+1) \\ \vdots \\ y(t_0+t_1-1) \end{pmatrix}, \;\; y(t_0-1 : t_0-t_1) = \begin{pmatrix} y(t_0-1) \\ y(t_0-2) \\ \vdots \\ y(t_0-t_1) \end{pmatrix}.
\end{aligned}
$$

Proposition 4.6.4. *Consider the linear system of Definition 4.6.3 with the indicated forward and backward representations.*

(a)The linear system is observable from the future outputs on the interval $\{t_0, t_0+1, \ldots, t_0+t_1-1\}$ *if and only if the observability matrix* O_f *satisfies*

$$n_x = \text{rank}(O_f(A_f, C_f, t_0 : t_0+t_1-1)).$$

(b)The linear system is co-observable from the past outputs on the interval $\{t_0-1, t_0-2, \ldots, t_0-t_1\}$ *if and only if the observability matrix* O_b *satisfies*

$$n_x = \text{rank}(O_b(A_b, C_b, t_0-1 : t_0-t_1)).$$

Proof. Note the linear observation map from the current state $x(t_0)$ to the future outputs and the linear observation map from the current state to the past outputs,

$$
\begin{aligned}
&y(t_0 : t_0+t_1-1) = O_f(A_f,\, C_f,\, t_0 : t_0+t_1-1)\, x(t_0), \\
&y(t_0-1 : t_0-t_1) = O_b(A_b,\, C_b,\, t_0 : t_0+t_1-1)\, x(t_0).
\end{aligned}
$$

The rank condition then follows from Proposition 4.6.2. □

The consequence of a nonobservable linear system are illustrated by the following example.

Example 4.6.5. Consider a time-invariant linear system. Suppose that the system is not observable, hence the rank of the observability matrix is strictly less than the dimension of the system. It can then be proven, see Proposition 21.3.6, that there exists a linear transformation of the considered system such that the transformed system has the following system representation.

$$\begin{aligned} x(t+1) &= \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix},\ x(0) = \begin{pmatrix} x_{1,0} \\ x_{2,0} \end{pmatrix}, \\ y(t) &= \begin{pmatrix} C_1 & 0 \end{pmatrix}, \\ & x_1(t) \in \mathbb{R}^{n_{x1}},\ x_2(t) \in \mathbb{R}^{n_{x2}},\ n_{x_2} \in \mathbb{Z}_+, \\ & \text{hence } 0 < n_{x_2},\ n_{x_1} + n_{x_2} = n_x, \\ & (A_{11},\ C_1) \text{ is an observable pair.} \end{aligned}$$

Note that the second component of the state, x_2, does not directly affect the output y. Note also that the second component of the state, x_2, does not affect the first component of the state, x_1. Therefore, if one is interested in a model for the output, only the first state component is of interest and the second state component can be discarded. Thus the subsystem represented by the second state component of the considered linear system is not relevant for a model of the output.

This is the main consequence of nonobservability of a linear system. A corresponding conclusion holds if the linear system is not co-observable.

Note that the interconnection of several observable linear systems may be nonobservable. See Exercise 4.11.4.

4.6.2 Stochastic Observability and Stochastic Co-Observability

The concepts of observability and of co-observability are extended to stochastic systems.

Definition 4.6.6. Consider a stochastic system,

$$\{\Omega, F, P, T, \mathbb{R}^{n_y}, B(\mathbb{R}^{n_y}), \mathbb{R}^{n_x}, B(\mathbb{R}^{n_x}), y, x\} \in \text{StocS}.$$

(a) This stochastic system is called *stochastically observable on the interval* $\{t_0, t_0+1, \ldots, t_0+t_1-1\} \subseteq T$ if the following *stochastic state-to-future-output map* is injective on the support of $x(t_0)$,

$$\begin{aligned} & x(t_0) \mapsto \text{cpdf}\left(\begin{pmatrix} y(t_0) & y(t_0+1) & \ldots & y(t_0+t_1-1) \end{pmatrix} \mid F^{x(t_0)} \right) \\ \Leftrightarrow\ & \sigma(F^{y(t_0:t_0+t_1-1)} \mid F^{x(t_0)}) = F^{x(t_0)}. \end{aligned}$$

The equivalence of the two above conditions follows from Proposition 2.5.14.

(b) Assume that the stochastic system is time-invariant and that the state and output process are jointly stationary. Then the system is called *stochastically observable* if there exists a $t_0 \in T$ and there exists a $t_1 \in \mathbb{Z}_+$ such that $\{t_0,\ t_0+1,\ \ldots,\ t_0+t_1-1\} \subseteq T$ and such that this system is stochastically observable on the indicated interval. By stationarity this then holds for any initial time $t_0 \in T$ such that the corresponding interval is contained in the set T.

(c) This stochastic system is called *stochastically co-observable on the interval* $\{t_0-1, t_0-2, \ldots, t_0-t_1\} \subseteq T$ if the following *stochastic state-to-past-output map* is injective on the support of $x(t_0)$

$$\begin{aligned}&x(t_0) \mapsto \text{cpdf}\Big(\big(y(t_0-1)\ y(t_0-2)\ \ldots\ y(t_0-t_1)\big) \mid F^{x(t_0)}\big)\Big)\\ &\Leftrightarrow \sigma(F^{y(t_0-1:t_0-t_1)} \mid F^{x(t_0)}) = F^{x(t_0)}.\end{aligned}$$

(d) Assume that the stochastic system is time-invariant and that the state and output process are jointly stationary. Then the system is called *stochastically co-observable* if there exists a $t_0 \in T$ and there exists a $t_1 \in \mathbb{Z}_+$ such that $\{t_0-1,\ t_0-2,\ \ldots,\ t_0-t_1\} \subseteq T$ and such that this system is stochastically co-observable on the indicated interval. By stationarity this then holds for any initial time $t_0 \in T$ such that the interval is contained in the set T.

The interpretation of a stochastic system which is stochastically observable on an interval is that, if one knows the conditional distribution of the future observations conditioned on the initial state of the interval, then one can uniquely determine the value of the initial state.

Note that the conditional distribution of the future observations conditioned on the initial state can in principle be determined from measurements.

The stochastic state-to-future-output map is a map from the random variable of the initial state to a conditional probability measure. A corresponding statement holds for the stochastic state-to-past-output map.

The difference between stochastic observability and stochastic co-observability is that the first concept refers to future outputs and the second concept refers to past outputs.

4.6.3 Stochastic Observability and Stochastic Co-Observability of Gaussian Systems

A characterization of stochastic observability and stochastic co-observability of a Gaussian system is formulated below. The concept of an observability matrix is useful.

Definition 4.6.7. Consider a time-varying forward Gaussian system representation. Define the *observability matrix of the indicated interval* of this system representation by the formula

$$\begin{aligned}&\{t_0,\ t_0+1,\ \ldots,\ t_0+t_1-1\} \subseteq T,\\ &O_f(A_f,\ C_f,\ t_0 : t_0+t_1-1)\\ &= \begin{pmatrix} C_f(t_0) \\ C_f(t_0+1)\Phi_f(t_0+1,t_0) \\ \vdots \\ C_f(t_0+t_1-1)\Phi_f(t_0+t_1-1,t_0) \end{pmatrix} \in \mathbb{R}^{(t_1 n_y)\times n_x}.\end{aligned}$$

Here Φ_f denotes the state transition function associated with the system matrix A_f of the forward representation, see Def. 4.3.4.

Consider a time-varying backward Gaussian system representation. Correspondingly, define the *co-observability matrix of the indicated interval* of this backward system representation by the formula

$$\begin{aligned}
&\{t_0-1,\ t_0-2,\ \ldots,\ t_0-t_1\} \subseteq T,\\
&O_b(A_b,\ C_b,\ t_0-1:t_0-t_1)\\
&= \begin{pmatrix} C_b(t_0) \\ C_b(t_0+1)\Phi_b(t_0+1,t_0) \\ \vdots \\ C_b(t_0+t_1-1)\Phi_b(t_0+t_1-1,t_0) \end{pmatrix} \in \mathbb{R}^{(t_1 n_y)\times n_x}.
\end{aligned}$$

Consider next a time-invariant forward Gaussian system representation. Define the *observability matrix* of this system representation by the formula

$$O_f(A_f,C_f) = \begin{pmatrix} C_f \\ C_f A_f \\ C_f A_f^2 \\ \vdots \\ C_f A_f^{n_x-1} \end{pmatrix} \in \mathbb{R}^{n_x n_y \times n_x}.$$

Call the matrix tuple $(A_f,\ C_f) \in \mathbb{R}^{n_x\times n_x} \times \mathbb{R}^{n_y\times n_x}$ an *observable pair* if $n_x = \operatorname{rank}(O_f(A_f,\ C_f))$.

Consider finally a time-invariant backward Gaussian system representation. Define the *co-observability matrix* of this system representation by the formula

$$O_b(A_b,C_b) = \begin{pmatrix} C_b \\ C_b A_b \\ C_b A_b^2 \\ \vdots \\ C_b A_b^{n_x-1} \end{pmatrix} \in \mathbb{R}^{n_x n_y \times n_x}.$$

Call the matrix tuple $(A_b,\ C_b) \in \mathbb{R}^{n_x\times n_x} \times \mathbb{R}^{n_y\times n_x}$ a *co-observable pair* if $n_x = \operatorname{rank}(O_b(A_b,\ C_b))$.

For a time-invariant forward Gaussian system representation the following assumptions will be used in the theorem stated directly below. Denote the system representation by

$$x(t+1) = A_f x(t) + M_f(t) v_f(t),\ x(0) = x_0, \tag{4.18}$$

$$y(t) = C_f x(t) + N_f(t) v_f(t),\ v_f(t) \in G(0,I), \tag{4.19}$$

$$\operatorname{spec}(A_f) \subseteq \mathrm{D}_o,\ T = \mathbb{N}. \tag{4.20}$$

Assume that the initial state has the invariant state measure $x_0 \in G(0,Q_x)$. Then the state and the output process are jointly stationary Gaussian processes.

Theorem 4.6.8. *Consider the following Gaussian system representation,*

$$\begin{aligned}
&\{\Omega,F,P,T(0:t_1),\mathbb{R}^{n_y},B(\mathbb{R}^{n_y}),\mathbb{R}^{n_x},B(\mathbb{R}^{n_x}),y,x\} \in \text{StocS},\\
x(t+1) &= A_f(t)x(t)+M_f(t)v_f(t),\\
y(t) &= C_f(t)x(t)+N_f(t)v_f(t),\ v_f(t) \in G(0,I),\\
&x(t) \in G(0,Q_x(t)).
\end{aligned}$$

(a)The following statements are equivalent.

(a.a)The system is stochastically observable on the interval $\{t_0,t_0+1,\ldots,t_0+t_1-1\} \subseteq T$.
(a.b) $\ker(O_f(t_0:t_0+t_1)Q_x(t_0)) = \ker(Q_x(t_0))$.
(a.c) $\sigma(F^{y(t_0:t_0+t_1-1)} | \, F^{x(t_0)}) = F^{x(t_0)}$.

If, in addition, for all $t \in T$, the state variance matrix $Q_x(t_0)$ is strictly-positive definite, then this time-varying Gaussian system is stochastically observable on the considered time interval if and only if $n_x = \text{rank}(O_f(A_f,C_f,\ t_0,t_0+t_1-1))$.

(b)Consider next a time-invariant forward Gaussian system representation such that the state and the output process are jointly Gaussian processes. The following statements are equivalent:

(b.a)This time-invariant Gaussian system is stochastically observable.
(b.b)

$$\ker(O_f(A_f,C_f)Q_x) = \ker(Q_x). \tag{4.21}$$

(b.c)$\exists\, t_0 \in T,\ \exists\, t_1 \in \mathbb{Z}_+$ such that $\{t_0,t_0+1,\ldots,t_0+t_1-1\} \subseteq T$ and $\sigma(F^{y(t_0:t_0+t_1-1)} | \, F^{x(t_0)}) = F^{x(t_0)}$.

(c)Consider the assumptions of (b). Assume that $(A_f,\ M_f)$ is a supportable pair hence $0 \prec Q_x$. Then the system is stochastically observable if and only if (A_f,C_f) is an observable pair.

Proof. (a) Let $t_0,\ t_1 \in T$, $t_1 > 0$, and,

$$\bar{y} = \begin{pmatrix} y(t_0) \\ y(t_0+1) \\ \vdots \\ y(t_0+t_1-1) \end{pmatrix} \in \mathbb{R}^{t_1 n_y},\ \ \bar{v}_f = \begin{pmatrix} v_f(t_0) \\ v_f(t_0+1) \\ \vdots \\ v_f(t_0+t_1-1) \end{pmatrix} \in \mathbb{R}^{t_1 n_{v_f}},$$

$$w = \begin{pmatrix} w(t_0) \\ w(t_0+1) \\ \vdots \\ w(t_0+t_1-1) \end{pmatrix} \in \mathbb{R}^{t_1 n_y}.$$

Then,

$$y(t_0+s) = C_f(t_0+s)\Phi_f(t_0+s,t_0)x(t_0)+$$
$$+\sum_{\tau=t_0}^{t_0+s-1}[C_f(t_0+s-1)\Phi_f(t_0+s-1,\tau)M_f(\tau)v_f(\tau)]+N_f(t_0+s)v_f(t_0+s),$$
$$\overline{y} = O_f(t_0)x(t_0)+\overline{M}_f(t_0)\overline{v}_f,$$
$$x(t_0)\mapsto E[\exp(iw^T\overline{y})|F^{x(t_0)}] = \exp(iw^T O_f(t_0)x(t_0)-\frac{1}{2}w^T Qw),$$

for a deterministic matrix Q. This map is injective on the support of x if and only if the map $x(t_0)\mapsto O^f(t)x(t_0)$ is injective on the support of $x(t_0)$. The support of $x(t_0)$ is $\text{Range}(Q_x(t_0))$. Thus the map $x(t_0)\mapsto O^f(t_0)x(t_0)$ is injective on the support of $x(t_0)$ if and only if for all $w\in\mathbb{R}^{n_x}$, $O_f(t_0)Q_x(t_0)w=0$ implies that $Q_x(t_0)w=0$, which is true if and only if $\ker(O_f(t_0)Q_x(t_0))\subseteq\ker(Q_x(t_0))$ if and only if $\ker(O_f(t_0)Q_x(t_0))=\ker(Q_x(t_0))$, since $\ker(Q_x(t_0))\subseteq\ker(O_f(t_0)Q_x(t_0))$ always holds. The characterization in terms of the equality of the two σ-algebras follows from Proposition 2.5.14. Note that,

$$\sigma(F^{y(t_0:t_0+t_1)}|\,F^{x(t_0)}) = \sigma(F^{\Phi_f(t_0:t_0+t_1)x(t_0)}|\,F^{x(t_0)}),$$
$$= F^{x(t_0)},\text{ by the next-to-last characterization.}$$

(b) ($\Rightarrow$) If σ is stochastically observable then there exists a $t_1>0$ such that for any $t_0\in T$ it is stochastically observable on the interval $\{t_0,t_0+1,\ldots,t_0+t_1-1\}$. From (a) then follows that, with the definition of the observability matrix, $\ker(O_f(A_f,C_f)Q_x)=\ker(Q_x)$.

If $t_1<n_x$ then (4.21) holds, else the same conclusion is reached by application of the Cayley-Hamilton theorem.

($\Leftarrow$) Conversely, the Condition (4.21) and (a) imply that the system is stochastically observable on the interval $\{t_0,t_0+1,\ldots,t_0+n_x-1\}$. Because Condition (4.21) does not depend on $t_0\in T$, it is equivalent to the existence of a $s\in T$ such that $F^{x(s)}=F^{O_f(A_f,C_f)x(s)}$. By stationarity this then holds for all $s\in T$.

(c) It follows from Theorem 22.1.2.(d) and the assumption that (A_f,M_f) is a supportable pair, that the solution $Q_x\in\mathbb{R}^{n_x\times n_x}$ of the Lyapunov equation is such that $Q_x\succ 0$. Hence the condition $\ker(O_f(A_f,C_f)Q_x)=\ker(Q_x)$ is equivalent to $\ker(O_f(A_f,C_f))=\{0\}$, to $\text{rank}(O_f(A_f,C_f))=n_x$, and to (A_f,C_f) an observable pair by Theorem 21.3.4. □

Theorem 4.6.9. *Consider the backward Gaussian system representation*

$$\{\Omega,F,P,T,\mathbb{R}^{n_y},B(\mathbb{R}^{n_y}),\mathbb{R}^{n_x},B(\mathbb{R}^{n_x}),y,x\}\in\text{StocS},$$
$$x(t-1) = A_b(t)x(t)+M_b(t)v_b(t),\ x(t_1)=x_1,$$
$$y(t-1) = C_b(t)x(t)+N_b(t)v_b(t),\ v_b(t)\in G(0,I),$$
$$x(t)\in G(0,Q_x(t)).$$

(a)The following statements are equivalent.

(a.i)The system is stochastically co-observable on the interval $\{t_0-1,t_0-2,\ldots,t_0-t_1\}\subseteq T$.

(a.b) $\ker(O_b(t_0 : t_0 - t_1)Q_x(t_0)) = \ker(Q_x(t_0))$.
(a.c) $\sigma(F^{y(t_0-1:t_0-t_1)}|F^{x(t_0)}) = F^{x(t_0)}$.

(b) Consider next a time-invariant backward Gaussian system representation such that the state and the output process are jointly stationary. The following statements are equivalent.

(b.a) The system is stochastically co-observable.
(b.b)

$$\ker(O_b(A_b, C_b)Q_x) = \ker(Q_x). \tag{4.22}$$

(b.c) $\exists\, t_0 \in T,\ \exists\, t_1 \in \mathbb{Z}_+$ *such that* $\{t_0 - 1, \ldots, t_0 - t_1\} \subseteq T$ *and* $\sigma(F^{y(t_0-1:t_0-t_1)}|\, F^{x(t_0)}) = F^{x(t_0)}$.

(c) Consider the assumptions of (b). Assume that (A_b, M_b) *is a supportable pair or, equivalently, that* $0 \prec Q_x$*. Then the system is stochastically co-observable if and only if* (A_b, C_b) *is an observable pair.*

The proof of this result is analogous to that of Theorem 4.6.8 and therefore omitted.

Note that the condition (4.21) is expressed in terms of the matrices (A_f, C_f) of the forward representation of the Gaussian system and the condition (4.22) is expressed in terms of the matrices (A_b, C_b) of the backward representation. See Section 4.5 for the way the matrices of the forward and backward representation are related.

Example 4.6.10. *A forward Gaussian system representation which is not stochastically observable.* Consider a time-invariant forward Gaussian system representation.

Assume that the system representation is not stochastically observable. Then it follows from the Kalman observable form, see Proposition 21.3.6, that the system representation can be transformed to the following representation,

$$\begin{aligned}
x(t+1) &= \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} + \begin{pmatrix} M_1 \\ M_2 \end{pmatrix} v(t),\ x(0) = x_0, \\
y(t) &= \begin{pmatrix} C_1 & 0 \end{pmatrix} x(t) + Nv(t),\ v(t) \in G(0, I_{n_v}), \\
&\quad \text{spec}(A) \subset \mathrm{D}_o,\ n_{x_1} \in \mathbb{N},\ n_{x_2} \in \mathbb{Z}_+,\ n_x = n_{x,1} + n_{x,2}, \\
&\quad (A, M) \text{ a supportable pair, } (A_{11}, C_1) \text{ an observable pair.}
\end{aligned}$$

From the above representation follows that the output proces y is not at all influenced by the second component x_2 of the state x. If one is interested in this Gaussian system representation as a model for the output process y then the second component x_2 of the state can be deleted from the system representation. If so deleted, then the remaining forward Gaussian system representation is,

$$\begin{aligned}
x_1(t+1) &= A_{11}x_1(t) + M_1 v(t),\ x_1(0) = x_{0,1}, \\
y(t) &= C_1 x_1(t) + Nv(t).
\end{aligned}$$

This example illustrates a consequence of the concept of stochastic observability of a Gaussian system representation. This example is developed further in Chapter 6 with stochastic realization of Gaussian stochastic systems.

4.7 Interconnections of Gaussian Systems

In engineering a particular modeling technique is used which is useful for obtaining realistic models for phenomena exhibiting random fluctuations. The approach is called the *noise-shaping approach* to modeling of Gaussian systems. A special case follows.

Example 4.7.1. *Noise-shaping modeling of a Gaussian system.* Consider a linear control system driven by a stochastic process w which is not Gaussian white noise, with the representation,

$$\begin{aligned} x(t+1) &= Ax(t) + Bw(t),\ x(0) = x_0, \\ y(t) &= Cx(t) + Dw(t), \\ & T = \mathbb{N},\ n_x,\ n_w,\ n_y \in \mathbb{Z}_+,\ X = \mathbb{R}^{n_x},\ W = \mathbb{R}^{n_w},\ Y = \mathbb{R}^{n_y}. \end{aligned}$$

It is then assumed that the stochastic process w can be approximated as the output of a time-invariant Gaussian system with the following represenation,

$$\begin{aligned} x_w(t+1) &= A_w x_w(t) + B_w v(t),\ x_w(0) = x_{w,0}, \\ w(t) &= C_w x_w(t) + D_w v(t), \\ & T = \mathbb{N},\ n_{x_w},\ n_v \in \mathbb{Z}_+,\ X_w = \mathbb{R}^{n_{xw}},\ V = \mathbb{R}^{n_v}, \end{aligned}$$

where v is standard Gaussian white noise. The combined system is then a time-invariant Gaussian system with representation,

$$\begin{aligned} \begin{pmatrix} x(t+1) \\ x_w(t+1) \end{pmatrix} &= \begin{pmatrix} A & BC_w \\ 0 & A_w \end{pmatrix} \begin{pmatrix} x(t) \\ x_w(t) \end{pmatrix} + \begin{pmatrix} BN_w \\ M_w \end{pmatrix} v(t), \\ y(t) &= \begin{pmatrix} C & DC_w \end{pmatrix} \begin{pmatrix} x(t) \\ x_w(t) \end{pmatrix} + DN_w v(t). \end{aligned}$$

Such a combined system is then often a more realistic model of the considered phenomenon than that by assuming that w is a Gaussian white noise process. Depending on the values of the matrices, stochastic observability of the Gaussian system representation may or may not hold. The extra effort to solve a filtering problem and of solving a stochastic control problem is often worth the effort.

This modeling technique has been used to model the disturbances of a fully-developed sea in a model for the behavior of a mooring tanker, see Example 1.1.1.

Noise shaping systems for other types of stochastic systems can also be defined. Interconnections of stochastic control system are treated in Chapter 10.

4.8 Stochastic Stability

Readers who are familiar with stability of deterministic systems have an understanding of stability of the state process of such a system. Stochastic stability is the subject which describes the stability-like properties of stochastic systems of which the state is a stochastic process. However, the concepts of stability of a stochastic system are different in character than those of a deterministic system.

In stability of a deterministic system, stability refers either to (1) boundedness of the state trajectory or to (2) convergence of the state trajectory to a steady state and the system is often transformed such that the steady state is the zero state.

In stability of a stochastic system there are two main concepts: (1) convergence of the state trajectory to a deterministic steady state; and (2) existence of a nontrivial invariant measure or of a nontrivial invariant probability distribution of the state trajectory, such that the state trajectory fluctuates with that invariant measure. The second concept was discussed for Gaussian systems in Section 4.4. The existence of an invariant measures of a state-finite stochastic system, defined in the next chapter, is extensively discussed in Section 18.8. The second concept has no clear equivalence in stability theory of deterministic systems.

Below a result is stated on convergence of the state trajectory of a stochastic system to a deterministic steady state, the zero state. Such a result is useful in particular cases.

Proposition 4.8.1. *Consider the special time-varying Gaussian system,*

$$\begin{aligned} x(t+1) &= Ax(t)+M(t)v(t),\ x(t_0)=x_0, \\ & \qquad v(t) \in G(0,I),\ \text{spec}(A) \subset \text{D}_o,\ \lim_{t\to\infty} M(t) = 0. \end{aligned}$$

Then $L_2 - \lim_{t\to\infty}\ x(t) = 0$.

Proof. From Theorem 4.3.5 follows that, for all $t \in T$, $x(t) \in G(0, Q_x(t))$,

$$Q_x(t+1) = AQ_x(t)A^T + M(t)M(t)^T,\ Q_x(0) = Q_{x_0}.$$

Then the assumptions $\lim_{t\to\infty}\ M(t) = 0$ and $\text{spec}(A) \subset \text{D}_o$ imply that,

$$\begin{aligned} \overline{Q}_x &= \lim_{t\to\infty}\ Q_x(t+1) = \lim\ AQ_x(t)A^T + \lim_{t\to\infty}\ M(t)M(t)^T \\ &= \lim\ AQ_x(t)A^T = A\overline{Q}_xA^T \in \mathbb{R}^{n_x\times n_x}. \end{aligned}$$

From Theorem 22.1.2 and $\text{spec}(A) \subset \text{D}_o$ follows that the equation $\overline{Q}_x = A\overline{Q}_xA^T$ has the unique solution $\overline{Q}_x = 0$. Thus,

$$\begin{aligned} 0 &= \lim_{t\to\infty} E[(x(t)-E[x(t)])(x(t)-E[x(t)])^T] = \lim\ Q_x(t) = \overline{Q}_x = 0,\ \Rightarrow \\ 0 &= L_2 - \lim_{t\to\infty} x(t). \end{aligned}$$

□

4.9 Gaussian Factor Models and Gaussian Factor Systems

The factor model was proposed early in the twentieth century. Factor analysis is used as a quantitative model in sociology and psychology. R. Frisch has suggested to use the factor model to determine relations among random variables [26]. He was awarded the Nobel Prize in Economics also for this research. R.E. Kalman has emphasized this model and formulated the associated stochastic realization problem [39, 40, 41]. Since then several researchers have considered the stochastic realization problem for this model class. This problem is still unsolved.

From economic data that exhibit variability one may estimate a covariance. Suppose that this data vector may be modelled by a Gaussian random variable. Effectively one is thus given a Gaussian measure, say on $\mathbb{R}^{n_y}$. The initial problem may then be stated as: how to represent this measure such that the dependencies between the components of the vector are exhibited? The factor model will be used to describe these dependencies.

The next definition is as used in the literature. It is proven below that this definition is characterized by the concept of conditional independence.

Definition 4.9.1. A *Gaussian factor model* or a *Gaussian factor model* of a Gaussian measure on $\mathbb{R}^k$ for an integer $k \in \mathbb{Z}_+$ is defined by the relation,

$$y = Hx + w, \tag{4.23}$$

$$y_i = H_i x + w_i, \;\; i = 1, \ldots, k, \tag{4.24}$$

where $x : \Omega \to \mathbb{R}^{n_x}$, $x \in G(0, Q_x)$ is called the *factor*, $w : \Omega \to \mathbb{R}^{n_w}$, $w \in G(0, Q_w)$ is called the *noise,* $y : \Omega \to \mathbb{R}^{n_y}$, $y \in G(0, Q_y)$, is called the *observation vector,* $H \in \mathbb{R}^{n_y \times n_x}$ is called the *matrix of factor loadings,* Q_w is a diagonal matrix, and (x, w) are independent random variables. The main characteristic of the model is that the matrix Q_w is diagonal.

The interpretation of the Gaussian factor model (4.24) is that each component of the observation vector consists of a systematic part $H_i x$ and a noise part w_i. Observe that the condition that Q_w is diagonal is equivalent to the condition that $(w_1, \ldots, w_{n_w})$ are independent random variables. A generalization of the above definition may be given to the case in which Q_w is block diagonal. The Gaussian factor model in rudimentary form goes back to [86]. The Gaussian factor model is equivalent to the *confluence analysis model* introduced by R. Frisch [26]. In this model the representation of the observation vector is specified by,

$$y = u + w, \; Au = 0,$$

in which $A \in \mathbb{R}^{(n_y - n_x) \times n_u}$, (u, w) are independent Gaussian random variables, and Q_w is a diagonal matrix. For other references on confluence analysis see the publications of O. Reiersol [73, 74].

The Gaussian factor model has been suggested as an alternative to regression analysis. Strong arguments for this approach are the introduction of the book by R. Frisch [26], and the papers of R.E. Kalman [39, 40, 41]. Within the economic and statistics literature the questions regarding regression and factor models have been recognized, see for example [6, 79, 87, 98].

What is the main characteristic of the Gaussian factor model? To answer this question one has to learn about the concept of conditional independence, see section 2.9. The main definition is repeated here to allow a smooth continuation of the discussion. The σ-algebra's $F_1, F_2, ..., F_m$ are called *multiply conditionally independent* conditioned on the σ-algebra G if

$$E[z_1 \dots z_m|G] = E[z_1|G] \dots E[z_m|G], \ \forall \, z_i \in L^+(F_i);$$
$$\text{the notation } (F_1, F_2, \dots, F_m|G) \in CI,$$

will be used to denote that $F_1, \dots, F_m$ are conditionally independent given G. The following elementary result then establishes the relation between the Gaussian factor model and the conditional independence relation.

Proposition 4.9.2. *Let $y_i : \Omega \to R, \ i = 1, 2, ..., k, \ x : \Omega \to \mathbb{R}^n$. The following statements are equivalent:*

(a)The random variables $(y_{1,}.., y_k, x)$ are jointly Gaussian with zero mean and satisfy the conditional independence relation,

$$(F^{y_1}, .., F^{y_k}|F^x) \in CI.$$

(b)The random variables y, x satisfy the conditions of the Gaussian factor model of Def. 4.9.1 with the representation $y = Hx + w$ for a matrix $H \in \mathbb{R}^{n_y \times n_x}$.

The conditional independence property of a Gaussian factor model is now seen to be its main characteristic. It will be called the *factor property* of a Gaussian factor model. It allows extensions to non-Gaussian random variables. Such extensions have been considered in the literature, see for references [91].

The factor property is a generalization of the concept of state for a stochastic system. In such a system the future of the state and output process on one hand, and the past of the state and output process on the other hand are conditionally independent given the present state. The analogy is such that the state corresponds to the factor and the output process to the observation vector of the factor model. The factor property or the conditional independence property occurs in many mathematical models in widely different application areas.

4.10 Computations

Computations for time-invariant Gaussian systems include the following operations severally of which are available in Matlab programs.

1. Computation of the solution of the discrete-time Lyapunov equation in case exponential stability holds, see Theorem 4.4.5.
2. Computation of the covariance function of the state and of the output of a time-invariant Gaussian system, see Theorem 4.4.5.
3. Computation of the backward Gaussian system representation from a forward time-invariant Gaussian system representation, and conversely. See Theorem 4.5.2.
4. Computations for several transformations of a time-invariant Gaussian system, see Section 4.4.

5. Check on the stochastic observability and the stochastic co-observability of a time-invarant Gaussian system. See Theorem 4.6.8 and Theorem 4.6.9.

For the benefit of a subgroup of readers there follows the algorithm for the computation of the covariance function on a finite time set of the combined stochastic processes (x^+, y) which formulas are stated in Theorem 4.4.5.

Procedure 4.10.1 *Computation of the covariance function of the state and the output process of a time-invariant forward Gaussian system representation.*
Data: Let $\{n_y, n_x, n_v, A, C, M, N\} \in \text{GStocSP}$ *be the parameters of a time-invariant forward Gaussian system representation and* $t_1 \in \mathbb{Z}_+$.

1. *Check whether the eigenvalues of the matrix A are in* $\text{D}_o = \{c \in \mathbb{C} | \, |c| < 1\}$. *Continue if this is the case, stop otherwise.*
2. *Solve the Lyapunov equation* $Q_x = AQ_xA^T + MM^T$ *for the matrix* $Q_x \in \mathbb{R}^{n_x \times n_x}$.
3. *Compute*

$$\begin{aligned} W_x(0) &= Q_x, \; W_y(0) = CQ_xC^T + NN^T, \\ Q_{x^+,y} &= AQ_xC^T + MN^T, \; W_{yx}(0) = CQ_x, \\ &\quad \textit{and for } t \in \mathbb{Z}_{t_1} = \{1, 2, \ldots, t_1\}, \textit{ recursively,} \\ W_x(t) &= A^tQ_x, \; W_y(t) = CA^{t-1}Q_{x^+,y}, \; W_{yx}(t) = CA^tQ_x. \end{aligned}$$

4.11 Exercises

Problem 4.11.1. *Time-varying covariance functions of a Gaussian system*. Consider a scalar time-invariant Gaussian system, thus with $n_x = 1$, $n_y = 1$, but $n_v > 1$, and with representation,

$$\begin{aligned} x(t+1) &= ax(t) + Mv(t), \; x(0) = x_0, \\ y(t) &= cx(t) + Nv(t), \;\; v(t) \in G(0, I). \end{aligned}$$

Assume that $0 < M^TM \in \mathbb{R}$. Describe the behavior of the time-varying variance function of the state process x and of the time-varying covariance function of y for the two cases of the variable a: either (1) $|a| < 1$ or (2) $|a| = 1$.

Problem 4.11.2. *Computation of the covariance functions of a forward Gaussian system representation*. Consider the following time-invariant forward Gaussian system representation,

$$\begin{aligned} x(t+1) &= \begin{pmatrix} 0 & 1 \\ 0.16 & 0.60 \end{pmatrix} x(t) + \begin{pmatrix} 0.6 & 0.4 \\ 1 & 0 \end{pmatrix} v(t), \\ y(t) &= \begin{pmatrix} 1 & 0 \end{pmatrix} x(t) + \begin{pmatrix} 0 & 1 \end{pmatrix} v(t), \; v(t) \in G(0, I). \end{aligned}$$

Compute with the aid of a computer program, for example with MATLAB, the following quantities.

(a) The eigenvalues of the system matrix A.
(b) The variance of the invariant probability distribution function of the state.
(c) The covariance function of the output process for several time steps, say $W_y(0)$, $W_y(1)$, and $W_y(2)$.

Problem 4.11.3. *Conditional distribution of the future outputs conditioned on past state and past outputs.* Consider a time-invariant forward Gaussian system representation with the equations,

$$\begin{aligned} x(t+1) &= Ax(t)+Mv(t),\ x(0)=x_0 \in G(0,Q_0), \\ y(t) &= Cx(t)+Nv(t),\ v(t) \in G(0,I). \end{aligned}$$

(a) Calculate the conditional characteristic function of the vector,

$$\bar{y} = \begin{pmatrix} y(t) \\ y(t+1) \\ \vdots \\ y(t+n_x-1) \end{pmatrix},\ \ \bar{y} : \Omega \rightarrow \mathbb{R}^{n_x n_y},$$
$$E[\exp(iw^T\bar{y})|F_t^x \vee F_{t-1}^y],\ \forall\, w \in \mathbb{R}^{n_x n_y},\ \forall\, t \in T.$$

It is not necessary to calculate a formula for the conditional variance.
(b) Which condition implies that from the conditional characteristic function one can uniquely determine the value of the randon variable $x(t)$?

Problem 4.11.4. *Observability of a series connection of two Gaussian systems.* Consider a series connection of two time-invariant Gaussian systems,

$$\begin{aligned} x_1(t+1) &= A_1x_1(t)+M_1v_1(t),\ x_1(t_0)=x_{1,0}, \\ &\quad x_1 : \Omega \times T \rightarrow \mathbb{R}^{n_{x_1}}, \\ y_1(t) &= C_1x_1(t)+N_1v_1(t),\ y_1 : \Omega \times T \rightarrow \mathbb{R}^{n_{y_1}}, \\ x_2(t+1) &= A_2x_2(t)+B_2y_1(t)+M_2v_2(t),\ x_2(t_0)=x_{2,0}, \\ &\quad x_2 : \Omega \times T \rightarrow \mathbb{R}^{n_{x_2}}, \\ y_2(t) &= C_2x_2(t)+D_2y_1(t)+N_2v_2(t),\ y_2 : \Omega \times T \rightarrow \mathbb{R}^{n_{y_2}}. \end{aligned}$$

Assume that $F^{x_{1,0}}$, $F^{x_{2,0}}$, $F_\infty^{v_1}$, $F_\infty^{v_2}$, are independent σ-algebras.

(a) Derive the forward difference representation of the series connection of the two forward difference representations of the above defined stochastic systems with state process (x_1,x_2) and output process y_2.
(b) Assume that the systems are scalar, thus that $1=n_{x_1}=n_{x_2}=n_{y_1}=n_{y_2}$ and that $a_1,a_2 \in (0,1)$, $c_1 \neq 0$, and $c_2 \neq 0$. Derive conditions on the system matrices such that the series connection has an observability matrix, Def. 21.3.3, which satisfies $\text{rank}(\text{obsm}(A,C))=1$, where A,C are the matrices of the forward difference representations of the series connection in the notation used. Thus, a series connection of two stochastically observable Gaussian systems is not necessarily stochastically observable. (The concept of stochastic observability is defined in Chapter 6 and this concept is not needed for this exercise.)

4.12 Further Reading

Books. General books on stochastic systems are [7, 17, 19, 30, 49, 61]. These references do not use the concept of a stochastic dynamic system as defined in this chapter. A book in the style of these lecture notes is [54] though it is restricted to stationary Gaussian processes. A reference on time series analysis is [29].

Modeling of phenomena by stochastic systems. The example of control of a paper machine is adjusted from a paper of K.J. Aström, [8]. Books on stochastic models of economic behavior are [18, 43]. Books from the area of signal processing are [66, 71].

Definition of a stochastic system. An early reference on the definition of a stochastic dynamic system is [42, p.5, footnote]. The definition presented in this chapter is inspired by [68]. Early versions of the definition are in [88, 90, 92]. The definition of a stochastic dynamic system in terms of the conditional independence relation is based on related concepts given in [55, 69, 89, 90]. A definition of a Gaussian stochastic dynamic system phrased in terms of Hilbert spaces is proposed in [55]. A brief survey on stochastic systems and stochastic realization problems is [92]. J.C. Willems has proposed open stochastic systems in [96].

Gaussian system representations. The contents of this section is standard, [49, Ch. 6] though the presentation of this chapter is a generalization. For additional information on the different models of a stationary Gaussian process see [17, 30]. Forward and backward representations of Gaussian systems are discussed in [5, 56, 93]. Reciprocal Gaussian processes are discussed in [35, 50, 52, 53]. Backward representations of Gaussian systems are discussed in [93].

The concept of stochastic observability and of stochastic co-observability was first defined by G. Ruckebusch in [75, 76, 77]. It is now a standard in the Hilbert-space formulation of stationary Gaussian systems, [54]. The measure theoretic formulation of stochastic observability and of stochastic co-observability used in this chapter was stated in [92].

Dynamic factor systems. The factor model was proposed early in the 20th century. For references on the factor model see [6, 91]. Factor analysis is used as a quantitative model in sociology and psychology. R. Frisch has suggested to use the factor model to determine relations among random variables [26]. R.E. Kalman has emphasized this model and formulated the associated stochastic realization problem [40, 39, 41]. Since then several researchers have considered the stochastic realization problem for this model class. This problem is still unsolved. For publications on this problem see the special issue of *J. of Econometrics* that is opened by the paper [1].

Concepts similar to that of a Gaussian factor system have been introduced in the literature. An elementary version of a Gaussian factor system with H a constant matrix is introduced in [73]. In [28] a Gaussian factor system is defined without the rationality and causality conditions. In [23] one can find a definition of a Gaussian factor system. In [70] and in [54, Subsection 2.6.3] a generalization a dynamic factor system is presented in which the spectral density of the process w is not diagonal but block-diagonal and in which the transfer function H not be causal. The term *dynamic errors-in-variables systems* is used instead of Gaussian factor system in the publications [3, 2, 4, 20].

The formulation of a factor system is stated in [92]. See also the book [54].

Stochastic stability. A dated survey on stochastic stability is [44]. Books with discussion of stochastic stability are [31, 47, 48]. Papers that discuss stochastic stability of Gaussian stochastic control systems are [32, 85, 99, 100, 102]. Papers that discuss stochastic stability of nonlinear stochastic control systems are [11, 46, 97]. Papers that discuss stochastic stability of Markov chains are [14, 22, 63].

References

1. D.J. Aigner and M. Deistler. Latent variables models - Editor's introduction. *J. Econometrics*, 41:1–3, 1989. 120
2. B.D.O. Anderson. Identification of scalar errors-in-variables models with dynamics. *Automatica J.-IFAC*, 21:709–716, 1985. 121
3. B.D.O. Anderson and M. Deistler. Identifiability in dynamic errors-in-variables models. *J. Time Series Anal.*, 5:1–13, 1984. 121
4. B.D.O. Anderson and M. Deistler. Dynamic errors-in-variables systems with three variables. *Automatica J.-IFAC*, 23:611–616, 1987. 121
5. B.D.O. Anderson and T. Kailath. Forwards, backwards, and dynamically reversible Markovian models of second-order processes. *IEEE Trans. Circuits and Systems*, 26:956–965, 1979. 120, 217
6. T.W. Anderson and H. Rubin. Statistical inference in factor analysis. In J. Neyman, editor, *Proc. Third Berkeley Symposium on Mathematical Statistics and Probability, Volume V*, pages 111–150. University of California Press, Berkeley, 1956. 116, 120
7. M. Aoki. *State space modeling of time series*. Springer-Verlag, Berlin, 1987. 120, 410
8. K.J. Aström. Computer control of a paper machine - An application of linear stochastic control theory. *IBM J. Res. & Developm.*, 11:389–405, 1967. 9, 78, 120, 522, 575, 596
9. G. Bastin and M. Gevers. Identification and optimal estimation of random fields from scattered point-wise data. *Automatica*, 21:139–155, 1985. 78, 170
10. D. Blackwell, L. Breiman, and A.J. Thomasian. Proof of Shannon's transmission theorem finite-state indecomposible channels. *Ann. Math. Statist.*, 29:1209–1220, 1958. 78
11. J. Blankenship. Frequency domain stability criteria for stochastic nonlinear feedback systems. *SIAM J. Control & Optim.*, 14:1107–1123, 1976. 121
12. T. Bohlin. Four cases of identification of changing systems. In R.K. Mehra and D.G. Lainiotis, editors, *System identification - Advances and case studies*, pages 441–518. Academic Press, New York, 1976. 78, 282
13. U. Borisson and R. Syding. Self-tuning control of an ore crusher. *Automatica J. IFAC*, 12:1–7, 1976. 78
14. A.M. Bruckstein. On the invariant measures of some discrete-time Markov processes. *IEEE Trans. Inform. Theory*, 30:125–126, 1984. 121
15. A.M. Bruckstein and T. Kailath. Inverse scattering for discrete transmission-line models. *SIAM Review*, 29:359–389, 1987. 105
16. A.M. Bruckstein, B.C. Levy, and T. Kailath. Differential methods in inverse scattering. *SIAM J. Appl. Math.*, 45:312–335, 1985. 78
17. P.E. Caines. *Linear stochastic systems*. John Wiley & Sons, New York, 1988. 120, 276, 302, 310, 575
18. G. Chow. *Econometric analysis by control methods*. John Wiley, New York, 1981. 120, 410

19. M.H.A. Davis and R.B. Vinter. *Stochastic modelling and control*. Chapman and Hall, London, 1985. 120, 376, 410, 468, 575, 595
20. M. Deistler. Linear errors-in-variables models. In S. Bittanti, editor, *Time series and linear systems*, Lecture Notes in Control and Information Sciences, pages 37–67. Springer-Verlag, Berlin, 1986. 121
21. J.P. Delhomme. Kriging in the hydrosciences. *Adv. Water Resources*, 1:251–266, 1978. 78, 170
22. P. Echeverria. A criterion for invariant measures of Markov processes. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 61:1–16, 1982. 121
23. R. Engle and M. Watson. A one-factor multivariate time series model of metropolitan wage rates. *J. Amer. Statist. Assoc.*, 76:774–781, 1981. 120
24. V.H. Fleming, S.P. Sethi, and H.M. Soner. Turnpike sets in optimal stochastic production planning problems. Report 86-2, Lefschetz Center for Dynamical Systems, Providence, 1986. 78
25. L.J. Forys. Performance analysis of a new overload strategy. In *10th International Teletraffic Congres*, 1983. 78, 379
26. R. Frisch. *Statistical confluence analysis by means of complete regression systems*. Publ. no. 5. University of Oslo Economic Institute, Oslo, 1934. 116, 120
27. P.J. Van Gerwen, W.A.M. Snijders, and N.A.M. Verhoeckx. An integrated echo canceller for baseband data transmission. *Philips tech. Rev.*, 39:102–117, 1980. 78
28. J.F. Geweke and K.J. Singleton. Maximum likelihood ‘confirmatory’ factor analysis of econometric time series. *Int. Economic Rev.*, 22:37–54, 1981. 120
29. E.J. Hannan. *Multiple time series*. Wiley, New York, 1970. 73, 120
30. E.J. Hannan and M. Deistler. *The statistical theory of linear systems*. John Wiley & Sons, New York, 1988. 120, 217
31. R.Z. Hasminski. *Stochastic stability of differential equations*. Sijthoff & Noordhoff, Alphen aan de Rijn, 1980. 121, 466
32. U.G. Hausmann. On the existence of moments of stationary linear systems with multiplicative noise. *SIAM J. Control*, 12:99–105, 1974. 121
33. A.W. Heemink. *Storm surge prediction using Kalman filtering*. Thesis, Twente University, Enschede, 1986. 78, 282
34. J. William Helton. Discrete time systems, operator models, and scattering theory. *J. Functional Analysis*, 16:15–38, 1974. 105
35. B. Jamison. Reciprocal processes: The stationary gaussian case. *Ann. Math. Statist.*, 41:1624–1630, 1970. 120
36. A.J.E.M. Janssen, R.N.J. Veldhuis, and L.B. Vries. Adaptive interpolation of discrete-time signals that can be modelled as autoregressive processes. *IEEE Trans. Acoustics, Speech & Signal Processing*, 34:317–330, 1986. 78, 303, 310
37. A. Journel and C. Huijbregts. *Kriging geostatistics*. Academic Press, New York, 1978. 78, 170
38. S. Stidham Jr. Optimal control of admission to a queueing system. *IEEE Trans. Automatic Control*, 30:705–713, 1985. 78, 379
39. R.E. Kalman. Identification from real data. In M. Hazewinkel and A.H.G. Rinnooy Kan, editors, *Current developments in the interface: Economics, Econometrics, Mathematics*, pages 161–196. D. Reidel Publishing Company, Dordrecht, 1982. 116, 120
40. R.E. Kalman. System identification from noisy data. In A.R. Bednarek and L. Cesari, editors, *Dynamical Systems II*, pages 135–164. Academic Press, New York, 1982. 116, 120, 217
41. R.E. Kalman. Identifiability and modeling in econometrics. In P.R. Krishnaiah, editor, *Developments in Statistics*, volume 4, pages 97–136. Academic Press, New York, 1983. 116, 120
42. R.E. Kalman, P.L. Falb, and M.A. Arbib. *Topics in mathematical systems theory*. McGraw-Hill Book Co., New York, 1969. 78, 120, 807
43. D. Kendrick. *Stochastic control for economic models*. McGraw-Hill Book Co., New York, 1981. 120, 410, 575

44. F. Kozin. A survey of stability of stochastic systems. *Automatica J.-IFAC*, 5:95–112, 1969. 121

45. D.G. Krige. Two dimensional weighted moving average trend surfaces for ore evaluation. *Journal of the South African Institute of Mining and*, pages 13–38, 1966. 78, 170

46. H. Kunita. Supports of diffusion processes and controllability problems. In *Proceedings International Symposium on Stochastic Differential Equations, Kyoto 1976*, pages 163–185, 1976. 121

47. H.J. Kushner. *Stochastic stability and control*. Academic Press, New York, 1967. 121, 376, 410, 467

48. H.J. Kushner. *Introduction to stochastic control*. Holt, Rinehart and Winston Inc., New York, 1971. 121, 376, 410, 467, 525

49. H. Kwakernaak and R. Sivan. *Linear optimal control systems*. Wiley-Interscience, New York, 1972. 120, 376, 410, 467, 489, 593, 822, 823

50. B. Levy. Regular and reciprocal multivariate stationary gaussian processes over z are necessarily Markov. In C. Commault et al., editor, *Proceeding First European Control Conference*, pages 602–607, Paris, 1991. Hermès. 120

51. B.C. Levy. Layer by layer reconstruction methods for the earth resistivity from direct current measurements. *IEEE Trans. Geosc. Rem. Sensing*, 23:841–850, 1985. 78

52. B.C. Levy. Regular and reciprocal multivariate stationary gaussian processes over z are necessarily Markov. Report, Department of Electrical Engineering and Computer Science, University of California, Davis, 1991. 120

53. B.C. Levy, R. Frezza, and A.J. Krener. Modeling and estimation of discrete-time gaussian reciprocal processes. *IEEE Trans. Automatic Control*, 35:1013–1023, 1990. 120

54. A. Lindquist and G. Picci. *Stochastic realization of Gaussian processes – A geometric approach to modeling, estimation and identification*. Springer, Heidelberg, 2015. 120, 121, 175, 176, 180, 217, 246, 253, 254, 275, 276

55. A. Lindquist, G. Picci, and G. Ruckebusch. On minimal splitting subspaces and Markovian representations. *Math. Systems Th.*, 12:271–279, 1979. 120, 175, 217, 275

56. L. Ljung and T. Kailath. Backwards Markovian models for second-order stochastic processes. *IEEE Trans. Information Theory*, 22:488–491, 1976. 120

57. D. Mallieu, P. Rousseaux, Th. Van Cutsem, and M. Ribbens-Pavella. Dynamic multilevel filtering for real-time estimation of electric power systems. *Control - Theory and Advanced Technology*, 2:255–272, 1986. 78

58. H.M. Markowitz. *Portfolio selection: Efficient diversification of investments*. Yale University Press, New Haven, 1959. 78

59. G.B. Di Masi, L. Finesso, and G. Picci. Design of LQG controller for single point moored large tankers. *Automatica J.-IFAC*, 22:155–169, 1986. 8, 78, 575

60. G. Matheron. The intrinsic random functions and their applications. *Adv. Appl. Probab.*, 5:439–468, 1973. 78, 170

61. P.S. Maybeck. *Stochastic models, estimation and control: Volume 1,2 and 3*. Academic Press, New York, 1979. 120, 376, 410

62. R.C. Merton. Lifetime portfolio selection under uncertainty: The continuous-time case. *Rev. Econ. Statist.*, 51:247–257, 1969. 78

63. S.P. Meyn and R.L. Tweedie. *Markov chains and stochastic stability*. Springer, New York, 1993. 121

64. R.A. Miller and K. Voltaire. A stochastic analysis of the tree paradigm. *J. Econ. Dyn. & Control*, 6:371–386, 1983. 78

65. R.G. Newton. Inversion of reflection data for layered media: A review of exact methods. *Geophys. J.R. Astr. Soc.*, 65:191–215, 1981. 78

66. A.V. Oppenheim and R.W. Schafer. *Digital signal processing*. Prentice-Hall Inc., Englewood Cliffs, 1975. 120

67. E. Pardoux and M. Pignol. Etude de la stabilité de la solution d'une eds bilinéaire à coefficients périodiques - application au movement des pales d'helicoptère. In J.L. Lions A. Bensoussan, editor, *Analysis and Optimization of Systems, Part 2*, volume 63 of *Lecture Notes in Control and Information Sciences*, pages 92–103. Springer-Verlag, Berlin, 1984. 78

68. G. Picci. Stochastic realization of gaussian processes. *Proc. IEEE*, 64:112–122, 1976. 120, 175, 217, 275
69. G. Picci. On the internal structure of finite-state stochastic processes. In *Proc. of a U.S.-Italy Seminar*, volume 162 of *Lecture Notes in Economics and Mathematical Systems*, pages 288–304. Springer-Verlag, Berlin, 1978. 120, 150, 277
70. G. Picci and S. Pinzoni. Dynamic factor analysis models for stationary processes. *IMA J. Math. Control and Information*, 3:185 – 210, 1986. 120
71. L.R. Rabiner and B. Gold. *Theory and application of digital signal processing*. Prentice-Hall Inc., Englewood Cliffs, NJ, 1975. 120
72. R. Redheffer. On the relation of transmission-line theory to scattering and transfer. *J. Math. Phys.*, 41:1–41, 1962. 105
73. O. Reiersol. Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica*, 9:1–24, 1941. 116, 120
74. O. Reiersol. Confluence analysis by means of instrumental sets of variables. *Arkiv für Mathematik, Astronomi och Fysik*, 32A, 1945. 116
75. G. Ruckebusch. Représentations markoviennes de processus gaussiens stationnaires. *C. R. Acad. Sc. Paris, Série A*, 282:649–651, 1976. 120, 175, 217, 248, 253, 275
76. G. Ruckebusch. Représentations markoviennes de processus gaussiens stationnaires et applications statistiques. Rapport interne 18, Ecole Polytechnique, Centre de Mathématiques Appliquées, 1977. 120, 175, 248, 253, 275, 291, 310
77. G. Ruckebusch. Théorie géométrique de la représentation markovienne. *Ann. Inst. Henri Poincaré*, 16:225–297, 1980. 120, 175, 217, 248, 253, 275
78. P.A. Samuelson. Lifetime portfolio selection by dynamic stochastic programming. *Rev. Econ. Statist.*, 51:239–246, 1969. 78
79. P.A. Samuelson. A note on alternative regressions. In J.E. Stiglitz, editor, *The collected scientific papers of Paul A. Samuelson, Fifth Printing*, pages 694–697. M.I.T. Press, Cambridge, 1979. 116
80. F.C. Schoute. Optimal control and call acceptance in a SPC exchange. In *9th International Teletraffic Congres*, 1981. 78, 379
81. F.C. Schoute. Adaptive overload control of an SPC exchange. In *10th International Teletraffic Congres*, 1983. 78, 379
82. F.C. Schoute. Overload control in spc processors. *Philips Telecommunication Review*, 41:300–310, 1983. 78, 379
83. C.E. Shannon. A mathematical theory of communication. *Bell System Techn. Journal*, 27:379–423, 623–656, 1948. 78
84. S.A. Smulders. Control of freeway traffic flow. Report OS-R8817, Centrum voor Wiskunde en Informatica, Amsterdam, 1988. 9, 78, 169
85. J. Snyders. Stationary probability distributions for linear time-invariant systems. *SIAM J. Control & Optim.*, 15:428–437, 1977. 121
86. C.A. Spearman. General intelligence, objectively determined and measured. *Amer. J. Psych.*, 15:201–293, 1904. 116
87. J.H. Steiger. Factor indeterminacy in the 1930's and the 1970's some interesting parallels. *Psychometrika*, 44:157–167, 1979. 116
88. C. van Putten and J.H. van Schuppen. On stochastic dynamical systems. In *Proceedings Fourth International Symposium on the Mathematical Theory of Networks and Systems (MTNS79)*, pages Volume 3, 350–355, North Hollywood, CA, 1979. Western Periodical. 120, 276
89. J.H. van Schuppen. Stochastic filtering theory: A discussion of concepts, methods and results. In W. Vogel M. Kohlmann, editor, *Stochastic control theory and stochastic differential systems*, volume 16 of *Lecture Notes in Control and Information Sciences*, pages 209–226, Berlin, 1979. Springer-Verlag. 120, 352
90. J.H. van Schuppen. The strong finite stochastic realization problem - Preliminary results. In A. Bensoussan and J.L. Lions, editors, *Analysis and optimization of systems*, volume 44 of *Lecture Notes in Control and Information Sciences*, pages 179–190, Berlin, 1982. Springer-Verlag. 120, 276, 742

91. J.H. van Schuppen. Stochastic realization problems motivated by econometric modeling. In C.I. Byrnes and A. Lindquist, editors, *Modelling, Identification and Robust Control*, pages 259–275. Elsevier Science Publishers B.V. (North-Holland), Amsterdam, 1986. 117, 120, 217

92. J.H. van Schuppen. Stochastic realization problems. In J.M. Schumacher H. Nijmeijer, editor, *Three decades of mathematical system theory*, volume 135 of *Lecture Notes in Control and Information Sciences*, pages 480–523. Springer-Verlag, Berlin, 1989. 120, 121

93. G. Verghese and T. Kailath. A further note on backwards Markovian models. *IEEE Trans. Information Th.*, 25:121–124, 1979. 120

94. N.A.M. Verhoeckx, H.C. Van den Elzen, F.A.M. Snijders, and P.J. Van Gerwen. Digital echo cancellation for baseband data transmission. *IEEE Trans. Acoustics, Speech & Signal Proc.*, 27:768–781, 1979. 78

95. B. Widrow. Adaptive noise cancelling - Principles and applications. *Proc. IEEE*, 63:1692–1716, 1975. 78

96. Jan C. Willems. Open stochastic systems. *IEEE Trans. Automatic Control*, 58:406–421, 2013. 120

97. J.C. Willems and G.L. Blankenship. Frequency domain stability criteria for stochastic systems. *IEEE Trans. Automatic Control*, 16:292–299, 1971. 121

98. E.B. Wilson. Review of ‘Crossroads in the mind of man: A study of differentiable mental abilities’ by t.l. Kelley. *J. Gen. Psychology*, 2:153–169, 1929. 116

99. W.M. Wonham. Liapunov criteria for weak stochastic stability. *J. Diff. Eqns.*, 2:195–207, 1966. 121

100. W.M. Wonham. Liapunov method for the estimation of statistical averages. *J. Diff. Eqns.*, 2:365–377, 1966. 121

101. A.E. Yagle and B.C. Levy. A layer-stripping solution of the inverse problem for a one-dimensional elastic medium. *Geophysics*, 50:425–433, 1985. 78

102. M. Zakai. A Liapunov criterion for the existence of stationary probability distributions for systems perturbed by noise. *SIAM J. Control*, 7:390–397, 1969. 121

Chapter 5
Stochastic Systems

Abstract Several sets of stochastic systems are defined in this chapter. The sets are selected based on the sets in which the outputs take values. Conditions are provided for the selection of the output-state conditional distribution function and for the selection of the conditional distribution function of the next-state on the current state. Useful are a Poisson-Gamma stochastic system, a Beta-Gamma stochastic system, an output-finite-state-polytopic stochastic system, a sigma-algebraic system, and a multiple conditional independent relation.

Key words: Stochastic systems. Probability distributions.

Several specific subsets of stochastic systems are defined which are not Gaussian systems. Successively are treated a Bernoulli-Beta system, a Poisson-Gamma system, a Gamma-Gamma system, a output-finite-state-polytopic stochastic system, a sigma-algebraic system, and a multiple conditional independent relation. The choice for these particular subclasses is based on the concept of estimation-conjugate distributions and control-conjugate functions as explained below. The motivation to discuss these sets of stochastic systems is the search for subsets of stochastic systems for which the filtering problem and the stochastic control problem admit finite-dimensional solutions as will become clear in subsequent chapters of this book.

5.1 Stochastic Systems and Probability Distributions

Which stochastic systems (1) are motivated by control engineering problems and (2) yield a substantial control theory for the filtering problem and for the stochastic control problem? The motivation of a stochastic system is problem dependent for which examples have been described in previous chapters. Below the focus is on stochastic systems for which a substantial control theory can be developed.

In principle one can define a stochastic system as defined in Def. 4.2.2 with any probability distributions, according to the formulation,

$$x(t) \mapsto \text{cpdf}(x(t+1), y(t)|F_t^{x-} \vee F_{t-1}^{y-}) = \text{cpdf}(x(t+1), y(t)|F^{x(t)}),$$

which maps, for any time $t \in T$, the current state $x(t)$ into the conditional probability distribution function of the combined variables $(x(t+1),\ y(t))$ of the next state $x(t+1)$ and of the current output $y(t)$ conditioned on the past of the state and the past of the output process. In addition, the stochastic system is specified by the probability distribution of the initial state. The probability distributions of the vector of the state process and of the output process of such a system, can in principle be calculated as for a Gaussian stochastic system, see Chapter 4.

However, if one wants to obtain explicit solutions for a particular filter problem or for a stochastic control problem, then the defined stochastic system may have a solution which is difficult to calculate or with a different analytic expression for every time. It will be useful if control theory can provide several sets of stochastic systems for which a finite-dimensional filter exists or for which an optimal stochastic control problem has an analytic formula for the optimal control law at every time.

For the requested properties of a stochastic system, it will be useful if the conditional distribution of the state has the same analytic form for every time but with possible different parameters. To this one may refer by the term *invariance of a subset of probability distributions* of the state of a stochastic system. Analogously, it will be useful if the value function of an optimal stochastic control problem has the same analytic form for every time. To this one may refer by the term *invariance of a subset of value functions*. Needed are thus sufficient conditions on the probability distribution functions of the system which guarantee these invariance properties.

Sufficient conditions imposed on a stochastic system which imply invariance conditions of probability distributions are a combination of:

1. a tuple of *filter-conjugate* probability distributions for the conditional probability distribution of the output conditioned on the state, in the literature also known as *conjugate* or *estimation-conjugate* pdfs; and
2. a tuple of *control-conjugate* functions for the conditional probability distribution of the next state conditioned on the current-state and the current input.

The remainder of this section describes the above formulated sufficient conditions. Additional theory may be found in the chapters: for filter-conjugate probability distributions in Chapter 9 and for control-conjugate functions in Chapter 12.

Output-State Conditional Probability Distribution

The observed process of a stochastic system takes by definition values in the set Y and this set can take one of the following values: $Y = \mathbb{Z}_{n_y}$ or $Y = \mathbb{N}_{n_y}$ for a positive integer $n_y \in \mathbb{Z}_+$ which is a finite set; $Y = \mathbb{N} = \{0, 1, \ldots\}$ the natural numbers; $Y = [0, 1]$ a bounded interval of the real numbers; $Y = \mathbb{R}_+$ the positive numbers; and $Y = \mathbb{R}$ or $Y = \mathbb{R}^{n_y}$, the real numbers and the vector space of tuples of the real numbers. There exist other subsets of the real numbers not listed here. For each of the subsets one may choose probability distributions, formulate a stochastic system, and solve a filter problem and a stochastic control problem.

For estimation and filtering, the tuple of a (conditional Poisson distribution, Gamma distribution) for respectively the output conditioned on the state and the state, is called a estimation-conjugate tuple and is such that the conditional distribution of the state conditioned on the output, is of gamma type. See Table 5.1 for tuples of filter-conjugate probability distributions used in this book. For control, the tuple of a (quadratic cost rate, a Gaussian probability distribution) for respectively the cost rate, and the conditional probability distribution of the next state conditioned on the state and on the input is called a control-conjugate tuple and is such that the value function has the same analytic form as the cost rate.

Y	$f(.;y\|x)$	State	X	$f(.;x)$	$f(.;x\|y)$
	$E[\exp(iw_y^T y)\|F^x]$			$E[\exp(iw_x^T x)]$	$E[\exp(iw_x^T x)\|F^y]$
$\mathbb{N}_m$	Bernouli (q)	q	$(0,1)$	Beta	Beta
$\mathbb{N}$	Poisson (λ)	λ	$\mathbb{R}_+$	Gamma	Gamma
$\mathbb{R}_+$	Gamma (γ_1,γ_2)	γ_1	$\mathbb{R}_+$	Gamma	Gamma
$\mathbb{R}$	Gaussian (m,q)	m	$\mathbb{R}$	Gaussian	Gaussian
$\mathbb{R}$	Gaussian (m,q)	q	$\mathbb{R}_+$	Gamma	Gamma

Table 5.1 Table of pairs of estimation-conjugate probability distributions and their associated conditional distribution.

Next-State-Current-State Conditional Probability Distributions

Below subsets of conditional probability distribution functions are formulated which are invariant with respect to composition of such functions.

Notation for the probability distribution of the state process of a stochastic system is introduced. Denote the unconditional probability distribution of $x(t)$ by $f(.;\ x(t))$ and the set of such unconditional probability distribtutions by $F(X)$. This set of distributions is assumed to be the same for all time. Denote for any time $t \in T$, the conditional probability distribution function of $x(t+1)$ conditioned on $x(t)$ by $\text{cpdf}(x(t+1)|\ x(t)) = f(.|.;x(t+1)|x(t))$. The variables of the distribution will often be denoted by $f(v_{x(t+1)}|\ v_{x(t)};x(t+1)|x(t))$. Denote by the $F(X|\ X))$ the set of all such conditional probability distributions on the state set X. Special subsets of distributions are denoted by $F_s(X) \subseteq F(X)$ and $F_s(X|\ X) \subseteq F(X|\ X)$ where the subindex s denotes special.

Example 5.1.1. Define the set of Gaussian probability density functions on $(\mathbb{R}, B(\mathbb{R}))$ by the formula,

$$\begin{aligned} p(w;m,q) &= \exp(-(w-m)^2/2q)\ (2\pi q)^{-1/2}, \\ & \quad p : \mathbb{R} \to \mathbb{R}_+,\ m \in \mathbb{R},\ q \in \mathbb{R}_{s+} = (0,\infty); \\ F_s(X) &= \{p(.;\ m,q)|\ m \in \mathbb{R},\ q \in \mathbb{R}_{s+}\}. \end{aligned}$$

Example 5.1.2. Define on the finite set $X = \mathbb{Z}_n = \{1,\ 2,\ \ldots,\ n\}$ for an integer $n \in \mathbb{Z}_+$ a probability measure by the vector $p \in \mathbb{R}_{st}^n$, which is thus such that $p \in \mathbb{R}_+$ and $1 = 1_n^T p = \sum_{i=1}^n\ p_i$. Then $F_s(X) = \mathbb{R}_{st}^n = F(X)$.

Example 5.1.3. Define the set of Gamma probability density functions on $(\mathbb{R}_+, B(\mathbb{R}_+))$ by the formula,

$$\begin{aligned} p(w;\gamma_1,\gamma_2) &= w^{\gamma_1 - 1}\ \exp(-w/\gamma_2)\ \gamma_2^{-\gamma_1}\ /\Gamma(\gamma_1), \\ &\qquad p : \mathbb{R}_+ \to \mathbb{R}_+,\ \gamma_1,\ \gamma_2 \in \mathbb{R}_{s+} = (0,\infty); \\ F_s(X) &= \{p(.;\ \gamma_1,\ \gamma_2) |\ \gamma_1,\ \gamma_2 \in \mathbb{R}_{s+}\}. \end{aligned}$$

Formulas are needed for the composition of two probability distribtution functions or of two conditional probability distribution functions.

The relation from the unconditional probability distribution function $f(.;\ x(t))$ to the unconditional probability distribution function $f(.;\ x(t+1))$ is provided by the formula,

$$\begin{aligned} & f(w_{x(t+1)};\ x(t+1)) = f(.|.;\ x(t+1)|x(t)) \circ f(.;\ x(t)) \\ &= \int_X f(w_{x(t+1)} |\ w_{x(t)};\ x(t+1)|x(t))\ f(dw_{x(t)};\ x(t)). \end{aligned}$$

The formula is due to integrating out over the probability distribution function $f(.;\ x(t))$ of $x(t)$. In case that for both probability distribution functions there exist probability density functions then one obtains the relation,

$$\begin{aligned} & p(v_{x(t+1)} |\ v_{x(t)};\ x(t+1)|x(t)) \\ &= \int_X p(v_{x(t+1)} |\ w_{x(t)};\ x(t+1)|x(t))\ p(w_{x(t)};\ x(t))\ dw_{x(t)}. \end{aligned}$$

Notation is also needed for the composition of two conditional probability distribution functions or for the conditional probability density functions. A conditional probability distribution function depends on parameters while these parameters are functions of the random variables on which is conditioned. Denote the composition of two conditional probability distribution functions by the formula,

$$\begin{aligned} & f(w_{x(t+2)} |\ w_{x(t)};\ x(t+2)|x(t)) \\ &= f(.|.;\ x(t+2)|x(t+1)) \circ f(.|.;\ x(t+1)|x(t)) \\ &= \int_X f(w_{x(t+2)} |\ w_{x(t+1)};\ x(t+2)|x(t+1))\ f(dw_{x(t+1)} |\ w_{x(t)}; x(t+1)|x(t)). \end{aligned}$$

In case there exists conditional density functions then the formula becomes,

$$\begin{aligned} & p(w_{x(t+2)} |\ w_{x(t)};\ x(t+2)|x(t)) \\ &= \int_X p(w_{x(t+2)} |\ w_{x(t+1)};\ x(t+2)|x(t+1))\ \times \\ &\qquad p(w_{x(t+1)} |\ w_{x(t)};\ x(t+1)|x(t))\ dw_{x(t+1)}. \end{aligned}$$

Definition 5.1.4. Consider a state set X, a set of special unconditional probability distribution functions $F_s(X) \subseteq F(X)$ and a set of special conditional probability distribution functions $F_s(X|X) \subseteq F(X|X)$. Consider a stochastic system with the conditional probability distribution $f(.|.;\, x(t+1)|x(t)) \in F_s(X|\, X)$.

(a) Call the set of special conditional probability distribution functions $F_s(X|X)$ *invariant with respect to composition* if, for all $t \in T$, $f(.;x(t+2)|x(t+1)) \in F_s(X|X)$ and $f(.;x(t+1)|x(t)) \in F_s(X|X)$ imply that the composition $f(.;x(t+2)|x(t) = f(.|.;\, x(t+2)|x(t_1)) \circ f(.|.;\, x(t+1)|x(t)) \in F_s(X|X)$,

(b) Call the tuple of subsets $(F_s(X|X),\, F_s(X))$ of a conditional and an unconditional probability distribution functions *invariant with respect to composition* if, for all time $t \in T$, $(f(.|.;\, x(t+1)|x(t)),\, f(.;\, x(t)) \in (F_s(X|X) \times F_s(X))$ imply that $f(.;\, x(t+1)) \in F_s(X)$.

Problem 5.1.5. Consider a stochastic system with state set denoted by X. Consider a subset of special unconditional probability distributions $F_s(X) \subseteq F(X)$ on the state set and a subset of special unconditional probability distributions on the state set denoted by $F_s(X|\, X) \subseteq F(X|\, X)$.

(a) Which subset $F_s(X|\, X)$ of special conditional probability distribution functions is such that it is invariant with respect to composition?

(b) Which tuple of subsets $(F_s(X|X),\, F_s(X))$ is such that it is invariant with respect to composition?

The following answers to the above formulated problem have been obtained as will be argued below.

1. The subset of special Gaussian conditional probability distribution functions on the state set defined below, is invariant with respect to composition.
2. The subset of conditional probability measures on the finite set $\mathbb{Z}_n$ for an integer $n \in \mathbb{Z}_+$ is invariant with respect to composition.

Below the case of a time-invariant Gaussian system is described. The result stated below also follows from the results of Chapter 4. The conditional distribution function of a time-invariant Gaussian system is,

$$\begin{aligned}
x(t+1) &= Ax(t) + Mv(t),\; x(0) = x_0,\\
&\quad p(w_{x(t+1)}|w_{x(t)};\, x(t+1)|x(t))\\
&= \exp(-(w_{x(t+1)} - Aw_{x(t)})^T Q_{1|0}(w_{x(t+1)} - Aw_{x(t)})/2) \times\\
&\quad \times\; (2\pi \det(Q_{1|0}))^{-1/2};\\
&\quad E[\exp(iw^T x(t+1))|\, F^{x(t)}]\\
&= \exp(iw^T\, Ax(t) - w^T Q_{1|0} w/2),\; \forall\, w \in \mathbb{R}^{n_x},\; Q_{1|0} = MM^T \succ 0.
\end{aligned}$$

Denote the set of conditional Gaussian probability distribution functions on $(X, B(X)) = (\mathbb{R}^{n_x}, B(\mathbb{R}^{n_x}))$ for this time-invariant Gaussian system by its corresponding probability density function according to,

$$F_s(X|X) = \left\{ \begin{array}{l} p(w_1|w_0;\, x(1)|\, x(0)) \\ = \exp(-(w_1 - Aw_0)^T Q_{1|0}^{-1}(w_1 - Aw_0)/2)\times \\ \times(2\pi \det(Q_{1|0}))^{-1/2},\ w_1,\ w_0 \in \mathbb{R}^{n_x},\ Q_{1|0} \in \mathbb{R}^{n_x\times n_x}_{spds} \end{array} \right\};$$

$$F_s(X) = \left\{ \begin{array}{l} p(w.;\, x) = \exp(-(w - Am_x)^T Q_x^{-1}(w - Aw_x)/2)\times \\ \times(2\pi \det(Q_x))^{-1/2},\ w,\ m_x \in \mathbb{R}^{n_x},\ Q_x \in \mathbb{R}^{n_x\times n_x}_{spds} \end{array} \right\}.$$

Proposition 5.1.6. *Consider a time-invariant Gaussian system with the above notation. The set of conditional Gaussian probability distributions $F_s(X|\ X)$ defined above, is invariant with respect to composition.*

Similarly, the tuple of subsets of Gaussian probability distribution functions $(F_s(X|X),\ F_s(X))$ is invariant with respect to composition.

Proof. The requested invariance with respect to composition of $F_s(X|X)$ holds if and only if, in terms of the characteristic functions,

$$E[E[\exp(iw^T\ x(t+2))|\ F^{x(t+1)}]|\ F^{x(t)}],$$

is again a characteristic function of the type stated in the definition of $F_s(X|X)$.

The proof of invariance is written out in terms of the characteristic function, see the calculations,

$$\begin{aligned} & f(.|.;\, x(t+2)|x(t+1)),\ f(.|.;\, x(t+1)|x(t)) \in F_s(X|X) \\ \Rightarrow\ & f(.|.;\, x(t+2)|x(t+1)) \circ f(.|.;\, x(t+1)|x(t)) \in F_s(X|X), \text{ because,} \\ =\ & E[E[\exp(iw^T\ x(t+2))|\ F^{x(t+1)}]|\ F^{x(t)}] \\ =\ & E[\exp(iw^T\ Ax(t+1) - w^T Q_{2|1} w)|\ F^{x(t)}] = \exp(iw^T A^2 x(t) - w^T Q_{2|0} w/2), \\ & \text{which is a Gaussian characteristic function,} \\ \Rightarrow\ & f(.|.;\, x(t+2)|x(t)) \in F_s(X|X). \end{aligned}$$

In the above A and $Q_{2|0}$ are deterministic matrices.

The proof of the second claim is similar. □

Proposition 5.1.7. *The set of conditional probability distributions on a finite set $X = \mathbb{Z}_{n_x}$ is invariant with respect to composition of the state dynamics of a state-finite stochastic system. Assume that the unconditional probability measure of the state is a strictly positive vector, in $\mathbb{R}^{n_x}_{s+}$.*

Proof. This follows directly from Proposition 2.8.4.(b) along the lines of the previous proposition.

$$\begin{aligned} & E[x(t+2)|\ F^{x(t)}] = E[E[x(t+2)|\ F^{x(t+1),x(t)}]|\ F^{x(t)}] \\ =\ & E[A(t+1)x(t+1)|\ F^{x(t)}] = A(t+1)A(t)x(t). \end{aligned}$$

□

In the remainder of this chapter several stochastic systems are defined for which the tuples of probability distribution functions $(F_s(X|X),\ F_s(X))$ are invariant with respect to composition, see Proposition 5.3.4.(d) and Proposition 5.4.2. There are also negative conclusions not stated here.

5.2 Output in Binary Set

The focus in this section is on modeling of stochastic systems of which the output process takes values in the binary set $\mathbb{N}_1 = \{0,1\}$ or in a finite set with the binomial probability distribution.

Engineering phenomema with such an output process are common in information theory, in control theory for particular systems, and, in many research areas, for a stochastic system obtained after approximation of an arbitrary system by one with a binary output.

There are in the literature two subsets of models for stochastic processes with binary outputs:

1. an output-finite-state-polytopic stochastic system as formulated in Section 5.7;
2. and a Bernoulli-Beta stochastic system defined below in this section.

Which of these models is appropriate for a particular engineering phenomenon has to be investigated in each case.

The reader is reminded of the probability distribution functions of Bernoulli and of Beta type, which may be found in Def. 2.1.3 and Def. 2.1.7.

Example 5.2.1. *Phenomena with a binary output.* Chemical tests are performed to check for the presence of particular chemicals or of bacteria with harmful effects. The outcome of any test is either zero, denoting no presence, or one, denoting presence of the chemical or bacterium. The state model then models the dynamics of the bacterium whether it grows over time or declines. In either case, one obtains as model a stochastic system with a binary output process.

Another case is in modeling of manufacturing systems where the output represents with one the good operating state and with zero the down state when the machine does not operate. The state equation then models the probability that the machine is in the operating state.

In information theory, a binary output process can model the output of a communication source where for example a set of messages generated by a Markov process is encoded. The output of the source is then sent to a communication channel to be communicated to a receiver. The state process is of course related to the state of the communication source and may have a high complexity or state-space dimension.

Definition 5.2.2. Define the *Bernoulli–Beta stochastic system* by the objects and relations,

$$
\begin{aligned}
& y : \Omega \times T \to \mathbb{N}_1 = \{0,1\},\ x : \Omega \times T \to (0,1), \\
& E[I_{\{y(t)=k\}} | F_t^x \vee F_{t-1}^y] \\
& = \binom{1}{k}\ x(t)^k (1-x(t))^{1-k} = x(t)^k (1-x(t))^{(1-k)},\ \ \forall\, k \in \mathbb{N}_1, \\
& = \begin{cases} 1-x(t), & \text{if } k=0, \\ x(t), & \text{if } k=1; \end{cases} \\
& E[\exp(iw\, x(t+1))|\ F_t^x \vee F_{t-1}^y] \\
& \text{is a Beta conditional characteristic function} \\
& \text{with parameters } (\beta_1(t+1), \beta_2(t+1)) \in \mathbb{R}_{s+}^2, \\
& \text{and with density function } p_{x(t+1)} : (0,1) \to \mathbb{R}_+, \\
& p_{x(t+1)|\ F_t^x \vee F_{t-1}^y}(v) = v^{\beta_1(t+1)-1}(1-v)^{\beta_2(t+1)-1} / B(\beta_1(t+1), \beta_2(t+1)) \\
& \beta_1(t+1) = a_1 x(t) + b_1,\ \ \beta_1 : \Omega \times T \to \mathbb{R}_{s+} = (0,\infty),\ a_1,\ b_1 \in (0,\infty), \\
& \beta_2(t+1) = a_2 x(t) + b_2,\ \ \beta_2 : \Omega \times T \to \mathbb{R}_{s+} = (0,\infty),\ a_2,\ b_2 \in (0,\infty); \\
& x_0 \text{ has a Beta pdf with parameters } (\beta_{x_0,1}, \beta_{x_0,2}) \in (0,\infty)^2; \\
& E[I_{\{y(t)=k\}} \exp(iw\, x(t+1)) | F_{t-1}^x \vee F_{t-1}^y] \\
& = E[I_{\{y(t)=k\}} | F_{t-1}^x \vee F_{t-1}^y]\ E[\exp(iw\, x(t+1)) | F_{t-1}^x \vee F_{t-1}^y], \\
& \forall\, k \in N_1,\ w \in \mathbb{R}.
\end{aligned}
$$

Because of the last equation, there holds a condition of state-output conditional independence similar to that in the Gaussian case, see Def. 4.3.2.

Another way to model the recursion of the state process is to assume that,

$$
x(t+1) = \frac{x(t)^2}{x(t)^2 + v(t)^2},\ x(0) = x_0,\ x : \Omega \times T \to (0,1),\ x_0 : \Omega \to (0,1),
$$

where the process $v : \Omega \times T \to (0,\infty)$ is a sequence of independent random variables. The above fractional model is reminiscent of the model leading to the Beta pdf from the Chi-Square pdf, see below Def. 2.1.7.

5.3 Output in the Natural Numbers

The reader finds in this section several stochastic systems of which the output process is integer valued, without any upper bound on the values taken. In addition, it is illustrated for which engineering problems such systems are useful mathematical models.

Example 5.3.1. Consider for example the number of call requests arriving at a telephone exchange within a certain period, the number of parts arriving at a machine in a production process, and the number of pulses arriving at a nerve cell in a nervous system.

A stochastic model of these phenomena is a counting process. A Poisson process is a well known continuous-time counting process. In discrete-time one may define an analogous process. The value of a discrete-time counting process represents for each time the number of pulses received in a time interval of which the duration is specified and fixed. At each discrete-time moment this process takes values in the set of the natural numbers $N = \{0,1,2,\ldots\}$ and a model is to assume that the distribution of this variable is a Poisson distribution. A generalization of this model is to let the parameters of this Poisson distribution also be a stochastic process, say a Markov process. This structure will be called a Poisson-Markov-process system.

Definition 5.3.2. A *Poisson-state-finite-Markov-process system* is a collection, $\{\Omega, F, P, T, \mathbb{N}, X, C, p\}$, specified by the forward difference representation: the distribution of $x(0)$ and the transition map,

$$\begin{aligned} &E[I_{\{x(t+1)=i,y(t)=k\}}|F_t^{x-} \vee F_{t-1}^{y-}] \\ &= (Cx(t))^k \exp(-Cx(t))p(i,x(t))/k!,\ \forall\, k \in \mathbb{N},\ \forall\, i \in \mathbb{Z}_{n_x}, \end{aligned} \tag{5.1}$$

where (Ω, F, P) is a probability space, $T \subseteq \mathbb{N}$ is a time index set, $\mathbb{N} = \{0,1,2,\ldots\}$ is the set of natural numbers, X a finite set with n elements, say $X = \{a_1, \ldots, a_{n_x}\}$, $C \in \mathbb{R}^{1\times n_x}$, $p : \mathbb{Z}_{n_x} \times \mathbb{Z}_{n_x} \rightarrow \mathbb{R}_+$ satisfying $\sum_{i=1}^{n} p(i,j) = 1$ for all $j \in \mathbb{Z}_{n_x}$, $x : \Omega \times T \rightarrow X$, $y : \Omega \times T \rightarrow \mathbb{N}$ are stochastic processes whose probability measures are recursively defined by equation (5.1). As in Proposition 4.2.4 it may be proven that the above stochastic system representation is a stochastic system as defined in Definition 4.2.2.

Definition 5.3.3. Define the *Poisson-Gamma stochastic system* as a stochastic system with the following objects and relations,

$$(\Omega, F, P),\ T = \mathbb{N} = \{0,1,2,\ldots\},\ x : \Omega \times T \rightarrow \mathbb{R}_+,\ y : \Omega \times T \rightarrow \mathbb{N},$$

$$E[I_{\{y(t)=k\}}|F_t^x \vee F_{t-1}^y] = \frac{x(t)^k}{k!}\exp(-x(t)),\ \forall\, k \in \mathbb{N}, \tag{5.2}$$

$$\begin{aligned} &E[\exp(iw\, x(t+1))|\, F_t^x \vee F_{t-1}^y] \\ &= \exp(iw\, ax(t))\,(1 - iw\, \gamma_{v,2}(t))^{-\gamma_{v,1}(t)},\ \forall\, w \in \mathbb{R}, \end{aligned} \tag{5.3}$$

$$x_0 \text{ Gamma pdf with parameters } (\gamma_{x_0,1}(0), \gamma_{x_0,2}(0)) \in (0,\infty)^2,\ a \in (0,\infty),$$

$$\begin{aligned} &E_1[I_{\{y(t)=k\}}\exp(iw\, x(t+1))|F_t^x \vee F_{t-1}^y] \\ &= E_1[I_{\{y(t)=k\}}|F_t^x \vee F_{t-1}^y] \times E[\exp(iw\, x(t+1))|\, F_t^x \vee F_{t-1}^y], \\ &\forall\, t \in T,\ k \in \mathbb{N},\ \forall\, w \in \mathbb{R}. \end{aligned} \tag{5.4}$$

Define a *Gamma process* as a stochastic process $z : \Omega \times T \rightarrow \mathbb{R}_+$ such that, for all $t \in T$, $z(t)$ has a Gamma pdf. No restrictions are imposed on the relation of $z(t)$ and $z(s)$ for $s, t \in T$ with $s \neq t$.

The interpretation of this stochastic system is that the output $y(t)$ conditioned on the past states and the past outputs has a conditional Poisson distribution with as rate the state process, while the state process is defined as a conditional Gamma process as proven below.

Proposition 5.3.4. *Consider a Poisson-Gamma stochastic system of Def. 5.3.3.*

(a)The Poisson-Gamma system of Def.5.3.3 is a stochastic system as defined in Def. 4.2.2.
(b)Define the stochastic process $v : \Omega \times T \to \mathbb{R}$ $v(t) = x(t+1) - ax(t)$. Then v is a Gamma process, v is a sequence of independent random variables, and v and x_0 are independent.
(c)The system has the system representation,

$$\begin{aligned} x(t+1) &= a\,x(t) + v(t+1),\ x(0) = x_0, \qquad (5.5)\\ y(t) &= x(t) + v_o(t), \qquad (5.6)\\ &\{v(t),\ t \in T\},\ v : \Omega \times T \to \mathbb{R}_+,\ \textit{an independent sequence,}\\ &\textit{pdf of } v(t) \textit{ is a Gamma pdf}\\ &\textit{with parameters } (\gamma_{v,1}(t), \gamma_{v,2}(t)) \in (0,\infty)^2,\\ &v_o : \Omega \times T \to \mathbb{R},\ E[v_o(t)|\, F_t^x \vee F_{t-1}^y] = 0,\ \forall\, t \in T,\\ &F^{x_0},\ F_\infty^v,\ F_\infty^{v_0},\ \textit{independent.} \end{aligned}$$

(d)Assume that for all $t \in T$, $\gamma_{v,2}(t) = a\gamma_{x,2}(t)$. Then the state process is a Gamma process with the parameters,

$$\begin{aligned} E[\exp(iwx(t+1))] &= (1 - iw\gamma_{x,2}(t+1))^{-\gamma_{x,1}(t+1)},\ \forall\, w \in \mathbb{R},\\ \gamma_{x,1}(t+1) &= \gamma_{x,1}(t) + \gamma_{v,1}(t),\ \gamma_{x,1}(0) = \gamma_{x_0,1},\\ \gamma_{x,2}(t+1) &= a\gamma_{x,2}(t) = \gamma_{v,2}(t),\ \gamma_{x,2}(0) = \gamma_{x_0,2},\\ &\gamma_{x,1} : T \to (0,\infty),\ \gamma_{x,2} : T \to (0,\infty). \end{aligned}$$

Proof. (a) This follows directly from the formulas of the stochastic system.
(b) Define the stochastic process $v : \Omega \times T \to \mathbb{R}$ $v(t) = x(t+1) - ax(t)$. Then one calculates,

$$\begin{aligned} &\forall\, t \in T,\ \forall\, w \in \mathbb{R},\ E[\exp(iw\, v(t))|\, F_t^x \vee F_{t-1}^y]\\ &= E[\exp(iw\, x(t+1))|\, F_t^x \vee F_{t-1}^y]\exp(-iw\, ax(t))]\\ &= (1 - iw\, \gamma_{v,2}(t))^{-\gamma_{v,1}(t)} = E[\exp(iw\, v(t))],\\ &\quad \Rightarrow\ F^{v(t)},\ F_t^x \vee F_{t-1}^y \text{ are independent } \Rightarrow\ F^{v(t)},\ F_{t-1}^v \vee F^{x_0} \text{ are independent.} \end{aligned}$$

That the above conditional characteristic function of $v(t)$ equals the characteristic function is due to the sequences $(\gamma_{v,1},\ \gamma_{v,2})$ being deterministic. From the equality of these two characteristic functions then follows with Theorem 2.8.2.(f) that $v(t)$ is independent of the σ-algebra appearing in the conditional characteristic function.

Hence v is a Gamma process, v is a sequence of independent random variables, and v and x_0 are independent.
(c) Note that,

$$\begin{aligned}
&E[y(t)|\, F_t^x \vee F_{t-1}^y] = E[\sum_{k=0}^{\infty} k\, I_{\{y(t)=k\}}|\, F_t^x \vee F_{t-1}^y] \\
&= \sum k\, x(t)^k \exp(-x(t))/k! = \sum_{k=1}^{\infty} x(t)^k \exp(-x(t))/(k-1)! \\
&= x(t) \exp(-x(t)) \sum_{m=0}^{\infty} x(t)^m/m! = x(t),
\end{aligned}$$

$$v_o(t) = y(t) - x(t),\ E[v_o(t)|F_t^x \vee F_{t-1}^y] = 0, \text{ by the above calculation.}$$

(d) Note that because x_0 has a Gamma probability distribution that $E[\exp(iw\,x(0))] = (1-iw\gamma_{x_0,2})^{-\gamma_{x_0,1}}$. Suppose that for $s = 1,2,\ldots,t \in T$ the probability distribution of $x(s)$ is Gamma with the parameters $(\gamma_{x,1}(s), \gamma_{x,2}(s)) \in (0,\infty)^2$. It will be proven that the corresponding formula holds for $s = t+1$.

Note that,

$$\begin{aligned}
&E[\exp(iwx(t+1))] = E[\exp(iwx(t+1))|F_t^x \vee F_{t-1}^y]] \\
&= E[\exp(iw\, ax(t))]\ (1-iw\gamma v, 2(t))^{-\gamma_{v,1}(t)}, \text{ by Def. 5.3.3,} \\
&= (1-iwa\gamma_{x,2}(t))^{-\gamma_{x,1}(t)}(1-iw\gamma_{v,2}(t))^{-\gamma_{v,1}(t)}, \text{ by the induction hypothesis,} \\
&= (1-iw\gamma_{x,2}(t+1))^{-\gamma_{x,1}(t+1)},
\end{aligned}$$

because of the formulas for the recursion of $(\gamma_{x,1}(t+1), \gamma_{x,2}(t+1))$.

□

5.4 Output in a Bounded Interval

Stochastic systems are considered for which the output process takes values in a bounded interval of the real numbers. Attention is therefore restricted to the interval $(0,1)$ which particular interval can be arranged by a transformation.

Examples of a stochastic system of which the output takes values in the interval $(0,1)$ are: (1) the forking percentage at a splitt of a road into two or more roads, or at an off-ramp of a motorway network; (2) the forking percentage at a node of a communication network where packets or messages are directed to two or more different lines; and (3) the reactions of humans to advertisement campaigns of commercial companies for a particular product. In each of these cases it is of interest to determine an estimate of the state of the output and to predict the state for future times. These predictions may then be used for advertisement campaigns.

Possible stochastic systems for a phenomenon with an output taking values in the interval $(0,1)$ are:

1. A Beta-Gamma stochastic system defined below;
2. An output-Beta-state-polytopic stochastic system in which the state process takes values in a polytope within the probability simplex; such a system is related to a output-finite-state-polytopic system, see Def. 5.7.1, but different.

The reader has to choose which of the above two possible stochastic systems fits best the considered phenomenon.

Definition 5.4.1. *Beta-Gamma stochastic system*. Define the *Beta-Gamma stochastic system* by the sets and the relations,

$$\begin{aligned}
&\text{cpdf}(y(t)|\, F_t^x \vee F_{t-1}^y) \text{ is of Beta type with parameters } (x_1(t),\, x_2(t)),\\
&\text{cpdf}(x_i(t+1)|\, F_t^x \vee F_{t-1}^y),\ i=1,\ 2,\\
&\quad \text{is of Gamma type with parameters } (\gamma_{i,1}(t+1),\ \gamma_{i,2}(t+1)),\\
\gamma_{i,1}(t+1) &= a_{i,1}x_i(t)+b_{i,1},\ \gamma_{i,1}(0)=x_{i,1,0},\\
\gamma_{i,2}(t+1) &= a_{i,2}x_i(t)+b_{i,2},\ \gamma_{i,2}(0)=x_{i,2,0},\ \forall\, i=1,\ 2;\\
&x_1,\ x_2 : \Omega\times T \to (0,\infty),\ \gamma_{i,1},\ \gamma_{i,2} : \Omega\times T\to(0,\infty),\\
&a_{i,1},\ a_{i,2},\ b_{i,1},\ b_{i,2}\ \in(0,\infty),\ \forall\, i=1,\ 2;\\
&x_{i,1,0},\ x_{i,2,0}:\Omega\to(0,\infty),\text{ have a Gamma pdf } \Gamma(\gamma_{1,0},\gamma_{2,0}).
\end{aligned}$$

Proposition 5.4.2. The probability distributions of the state and of the output process of a Beta-Gamma system. *Consider the Beta-Gamma stochastic system of Def. 5.4.1. The probability distributions of the state of this system are specified by their characteristic functions according to,*

$$\begin{aligned}
E[\exp(iw_1\ x_1(t+1))] &= \Big(1-iw_1\ (a_{1,1}^{t+1}\gamma_{1,2,0})\Big)^{-\gamma_{1,1,0}}\ \exp(iw_1r_1(t)),\\
E[\exp(iw_2\ x_2(t+1))] &= \Big(1-iw_2\ (a_{2,1}^{t+1}\gamma_{2,2,0})\Big)^{-\gamma_{2,1,0}}\ \exp(iw_1r_1(t)),\\
&\forall\, t\in T,\ w_1,\ w_2\in\mathbb{R};\\
&r_1(t)=b_{i,1}+a_{i,1}b_{i,1}+\ldots+a_{i,1}^t b_{i,1}.
\end{aligned}$$

If in addition, $b_{i,1}=0$ and $b_{i,2}=0$ for $i=1,\ 2$ then $r_i(t)=0$ for all $t\in T$. Hence the probability distributions of $x_1(t),\ x_2(t)$ are Gamma.

Proof. Note the calculation,

$$\begin{aligned}
&E[\exp(iw_1x_1(t+1))] = E[E[\exp(iw_1x_1(t+1))|F_t^x\vee F_{t-1}^y]]\\
&= E[\exp(iw_1[a_{1,1}x_1(t)+b_1])] = E[\exp(i(w_1a_{1,1})x_1(t))]\exp(iw_1b_1) = \ldots\\
&= E[\exp(iw_1a_{1,1}^{t+1}\ x(0))]\exp(iw_1r_1(t))\\
&= \Big(1-iw_1a_{i,1}^{t+1}\ \gamma_{2,0}\Big)^{-\gamma_{1,0}}\exp(iw_1\ r_1(t)).
\end{aligned}$$

□

5.5 Output in the Positive Real Numbers

Engineering phenomena with positive outputs arise in chemical engineering where the outputs are concentrations of chemicals, in civil engineering as rates in transport.

In economics there are also phenomena where the outputs are positive. Hence there is a need for a stochastic system with a positive output which will usually have a positive state process.

Below is defined a Gamma-Gamma stochastic system of which the output is a positive stochastic process.

Definition 5.5.1. *The Gamma-Gamma stochastic system.*
Define the *Gamma-Gamma stochastic system* by the sets and maps according to,

$$\begin{aligned}
& E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix}\right) \mid F_t^x \vee F_{t-1}^y\right] \\
&= (1 - iw_x\gamma_{x,2}(t+1))^{-\gamma_{x,1}(t+1)}\,(1 - iw_y\gamma_{y,2}(t+1))^{-\gamma_{y,1}(t+1)}, \\
&\quad \forall\, w_x,\ w_y \in \mathbb{R},\ \forall\, t \in T, \\
\gamma_{x,1}(t+1) &= a_1x(t) + b_1,\ \gamma_{x,2}(t+1) = a_2x(t) + b_2, \\
\gamma_{y,1}(t) &= x(t),\ \gamma_{y,2}(t) = \gamma_{y,2}(0) \in (0,\infty),\ a_1,\ a_2,\ b_1,\ b_2(0,\infty); \\
& x_0 : \Omega \to \mathbb{R}_+ \text{ has a Gamma pdf with parameters } (\gamma_{x_0,1}, \gamma_{x_0,2}).
\end{aligned}$$

Proposition 5.5.2. *Consider the Gamma-Gamma stochastic system of Def. 5.5.1. A Gamma-Gamma stochastic system has the property of being state-output conditionally independent, see Def. 4.3.2 for the corresponding definition of a Gaussian stochastic system.*

Proof. This follows directly from the definition of the system,

$$\begin{aligned}
& E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix}\right) \mid F_t^x \vee F_{t-1}^y\right] \\
&= (1 - iw_x\gamma_{x,2}(t+1))^{-\gamma_{x,1}(t+1)}\,(1 - iw_y\gamma_{y,2}(t+1))^{-\gamma_{y,1}(t+1)}, \\
&= E\left[\exp\left(iw_xx(t+1)\right) \mid F_t^x \vee F_{t-1}^y\right] \times E\left[\exp\left(iw_yy(t)\right) \mid F_t^x \vee F_{t-1}^y\right], \\
&\quad \forall\, w_x,\ w_y \in \mathbb{R},\ \forall\, t \in T.
\end{aligned}$$

□

5.6 Output in the Real Numbers

In the literature mathematical models have been introduced that are called *nonlinear stochastic systems.* Such an object is often formed by first considering a deterministic nonlinear system and then adding a noise process in the form of a Gaussian white noise process or a Brownian motion process. The nonlinear system may be derived from mechanical laws or from engineering modeling procedures. The addition of a noise process or disturbance with a Gaussian distribution is not easy to justify.

What is needed is a theory which, starting from time series or stochastic process, derives a stochastic system. Such a theory does not yet exist. To motivate and structure such a theory examples are needed. References to such examples are provided in the Section *Further Reading* of this chapter.

Bilinear Gaussian Stochastic Systems

The reader may wonder why it is useful to consider this class of models. Bilinearity is a particular form of nonlinearity. At first sight it is not clear what the system specifies about the distributions of the process. A bilinear Gaussian stochastic system is defined to that it can be analysed. Examples have to show whether this stochastic system is a useful model for engineering or for the sciences.

Definition 5.6.1. Define a *bilinear Gaussian stochastic system* by the sets and relations,

$$x(t+1) = Ax(t) + \sum_{j=1}^{n_v} B_j x(t) v_j(t), x(0) = x_0, \tag{5.7}$$

$$y(t) = Cx(t) + Dv_o(t), \tag{5.8}$$

where $T = \mathbb{N}$, $A \in \mathbb{R}^{n_x \times n_x}$, $B_1, \ldots, B_{n_v} \in \mathbb{R}^{n_x \times n_x}$, $C \in \mathbb{R}^{n_y \times n_x}$, $D \in \mathbb{R}^{n_y \times n_{v_o}}$, $x_0 : \Omega \to \mathbb{R}^{n_x}$, $x_0 \in G(0, Q_0)$, $v : \Omega \times T \to \mathbb{R}^{n_v}$ is a Gaussian white noise process hence with independent components, $v(t) \in G(0, I_{n_v})$, $v_o : \Omega \times T \to \mathbb{R}^{n_{v_o}}$ is a Gaussian white noise process with $v_o(t) \in G(0, I_{n_{v_o}})$, and $x : \Omega \times T \to \mathbb{R}^{n_x}$ and $y : \Omega \times T \to \mathbb{R}^{n_y}$ are stochastic processes defined by the above equations. Assume that F^{x_0}, F^{v}_{∞}, and $F^{v_o}_{\infty}$ are independent objects.

It may be proven for a bilinear Gaussian stochastic system that,

$$\begin{aligned} &E[\exp(iw^T\, x(t+1))|F^x_t \vee F^y_{t-1}] \\ &= \exp(iw^T\, Ax(t) - \frac{1}{2} w^T [\sum_{j=1}^{m} B_j x(t) x(t)^T B_j^T] w),\ \forall\, w \in \mathbb{R}^{n_x}. \end{aligned} \tag{5.9}$$

This formula shows that a bilinear Gaussian stochastic system is such that the state transition map is Gaussian with as a mean a linear map in the state $x(t)$ and as variance a function which is quadratic in the state. Note that both the mean and the variance of this transition function are dependent on the state $x(t)$ and there is a relation between these dependences. It remains to be seen for which phenomena this stochastic system is a useful mathematical model.

5.7 Output-Finite-State-Polytopic Stochastic Systems

5.7.1 Time-Varying Stochastic Systems

A *finite stochastic system* is a stochastic system in which the state and the output process take values in finite sets, Def. 4.2.5. From the viewpoint of stochastic system theory, the set of finite stochastic systems is too narrow. The set has to be enlarged as described below. This viewpoint creates a difference of this book with much of the literature on this set of stochastic systems.

That the set of finite stochastic systems has to be enlarged is motivated by the set of Gaussian systems. Recall that the set of Gaussian stochastic realizations contains both an arbitrary time-invariant system and its associated Kalman realization, related to the Kalman filter. The Kalman realization is a weak Gaussian stochastic realization of the output process and hence also a Gaussian system. Thus the set of Gaussian systems contains both the original system and its associated filter system.

Next consider a finite stochastic system. Such a system in the indicator representation is such that the state vector is a unit vector of the positive orthant, hence, in any state vector, there exists one component equal to the value one while all other components have the value zero. The filter of a finite stochastic system is not a finite stochastic system because the conditional estimate of the state conditioned on past outputs is a vector in the tuples of the positive real numbers with almost-always two or more elements in the interval $(0,1)$; with respect to a condition, all components have values in the interval $(0,1)$. If one wants to obtain a useful stochastic system theory then the set of finite stochastic systems has to be enlarged so that the state set admits vectors in a subset $X \subset \mathbb{R}_{st}^{n_x}$ which also includes elements not equal to the unit vectors of the positive orthant. This enlargement will be formulated below. After further definitions, the choice of the subset of finite stochastic systems will be discussed in more detail.

The reader is expected to have read the introduction of Chapter 18 with notation on positive vectors. Recall from there that a *stochastic vector*, denoted by $p \in \mathbb{R}_{st}^{n}$, is defined to be a vector $p \in \mathbb{R}_{+}^{n}$ such that $1_n^T p = \sum_{i=1}^{n} p_i = 1$. The *probability simplex* is defined according to,

$$\mathbb{R}_{st}^{n} = \{p \in \mathbb{R}_{+}^{n} | 1_n^T p = 1\}.$$

The reader is expected to be familiar with the indicator representation of a finite-valued random variable, Def. 2.5.9, and of the associated representation of a finite-valued stochastic process, Def. 3.5.1. The reader is also expected to be familiar with the concept of a polytope, Def. 17.6.5.

Definition 5.7.1. Define a *forward representation of an output-finite-state-finite stochastic system* as a stochastic system such that,

$$\begin{aligned}
&\{\Omega, F, P, T, Y_e, X_e, y, x, \text{cpdf}, p_{x_0}\} \in \text{OFSFstocsys}, \\
X_e &= \{e_1, \ldots, e_{n_x} \in \mathbb{R}_{st}^{n_x}\},\ Y_e = \{e_1, \ldots, e_{n_y} \in \mathbb{R}_{st}^{n_y}\},\ n_y,\ n_x \in \mathbb{Z}_+, \\
&\text{cpdf}(x(t+1), y(t)) | F_t^x \vee F_{t-1}^y) = \text{cpdf}(x(t+1), y(t)) | F^{x(t)}), \qquad (5.10) \\
&E[x(t+1) y(t)^T |\ F_t^x \vee F_{t-1}^y] = f(t, x(t)), \qquad (5.11) \\
&\text{the support of the cpdf is a subset of } X_e \times Y_e,\ p_{x_0} \in \mathbb{R}_{st}^{n_x};
\end{aligned}$$

where $\{\Omega, F, P\}$ is a complete probability space, either the *time index set* is $T = \{0, 1, 2, \ldots, t_1\} \subset \mathbb{N}$ or $T = \mathbb{N}$, n_x, $n_y \in \mathbb{Z}_+$, X_e, Y_e are finite sets, called respectively the *state set*, the *output set,* $x : \Omega \times T \to X_e$, and $y : \Omega \times T \to Y_e$ are stochastic processes. The understanding is that the state $x(t)$ determines the conditional probability distribution on the joint atoms of the next state $x(t+1)$ and the output $y(t)$ and that the randomization mechanism selects the random variables $x(t+1), y(t))$ according to that distribution.

It is called a *time-invariant forward representation* of a output-finite-state-finite stochastic system if the transition map $x(t) \mapsto \text{cpdf}(x(t+1), y(t))$ does not depend explicitly on the time variable t.

Define a *forward representation of a time-invariant output-finite-state-polytopic stochastic system* (OFSPstocsys) as a stochastic system such that,

$$\{\Omega, F, P, T, Y_e, X_p, B(X_p), y, x, \text{cpdf}, X_{p0}\} \in \text{OFSPstocsys},$$

$$X_p \subseteq \mathbb{R}_{st}^{n_x}, \text{ a polytope,} \tag{5.12}$$

$$X_{p_0} \subseteq X_p, \text{ representing the set of initial states,} \tag{5.13}$$

$$Y_e = \{e_1, \ldots, e_{n_y} \in \mathbb{R}_{st}^{n_y}\},\ n_y,\ n_x \in \mathbb{Z}_+,$$

$$\text{cpdf}(x(t+1), y(t)) | F_t^x \vee F_{t-1}^y) = \text{cpdf}(x(t+1), y(t)) | F^{x(t)}), \tag{5.14}$$

$$\text{such that if } x_0 \in X_{p_0} \text{ and if, } \forall\, t \in T,\ x(t) \in X_p \text{ then } x(t+1) \in X_p; \tag{5.15}$$

$$\text{the support of the cpdf is a subset of } X_p \times Y_e; \tag{5.16}$$

where $\{\Omega, F, P\}$ is a complete probability space, either the *time index set* is $T = \{0, 1, 2, \ldots, t_1\} \subset \mathbb{N}$ or $T = \mathbb{N}$, n_x, $n_y \in \mathbb{Z}_+$, $X_p \subset \mathbb{R}_{st}^{n_x}$ is a polytope within the probability simplex, Y_e is a finite set, called respectively the *state set*, the *output set,* $x : \Omega \times T \to X_p$, and $y : \Omega \times T \to Y_e$ are stochastic processes. The main restriction of an output-finite-state-polytopic stochastic system are equation (5.14) and equation (5.15) on the invariance of the state polytope X_p.

It is called a *time-invariant forward representation* of a output-finite-state-polytopic stochastic system if the transition map $x(t) \mapsto \text{cpdf}(x(t+1), y(t))$ does not depend explicitly on the time variable t.

The term *output-finite-state-polytopic stochastic system* is complex. But the term best describes the characteristics of this set of stochastic systems. To assist the reader with reminding the term, keep in mind that the set of states is a polytope X_p within the probability simplex $\mathbb{R}_{st}^{n_x}$. The output is always finite-valued.

It follows directly from Def. 5.7.1 that an output-finite-state-polytopic stochastic system is a stochastic system as defined in Definition 4.2.2. For the finite state set X_e, a σ-algebra is not relevant but it is relevant for the state set $X_p \mathbb{R}_{st}^{n_x}$ in the form of a polytope within the probability simple.

The operation of an output-finite-state-finite stochastic system is then such that, at the initial time, $x_0 \in X_e$ while its probability measure is p_{x_0}; and, at every time $t \in T$, the current state $x(t)$ determines the conditional probability measure on the next state and the current output, thus on $(x(t+1), y(t))$. The underlying randomization mechanism then determines a realization of the random variables $(x(t+1), y(t)) \in X_e \times Y_e$. In the case of an output-finite-state-polytopic stochastic system, $(x(t+1), y(t))$ takes values in $X_p \times Y_e$.

The difference between (1) an output-finite-state-polytopic stochastic system and (2) an output-finite-state-finite stochastic system is that in (1) the state set $X_{pst} \in \mathbb{R}_{st}^{n_x}$ is a polytope while in (2) the state set is X_e, the set of unit vectors, while in general $X_e \subsetneq X_{pst}$.

For a more detailed discussion on the output-finite-state-polytopic stochastic systems see Subsection 5.7.2.

In the literature the name of a *hidden Markov model* is used for a output-finite stochastic system. That name will not be used further in this book. That term does not describe the properties of the state process in sufficient detail. Also a Gaussian system with an output process is then a hidden Markov model because the state process, which is a Markov process, is not observed.

Example 5.7.2. Consider the particular output-finite-state-finite stochastic system with the representation,

$$\begin{aligned}
& E\left[\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix} \mid F_t^x \vee F_{t-1}^y\right] = \begin{pmatrix} A \\ C \end{pmatrix} x(t),\ x(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \\
& X_e = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix},\ \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\},\ Y_e = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix},\ \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\},\ n_y = 2,\ n_x = 2, \\
& A = \begin{pmatrix} 0.8\ 0.4 \\ 0.2\ 0.6 \end{pmatrix},\ C = \begin{pmatrix} 0.9\ 0.3 \\ 0.1\ 0.7 \end{pmatrix}.
\end{aligned}$$

The understanding of this system is that the initial state x_0 and the matrices A and C determine the probability measure on the vector $(x(1), y(0))$ by the above equation and that the randomization mechanism then generates random variables $(x(1), y(0))$ taking values in $X_e \times Y_e$, etc.

Example 5.7.3. An example follows which is a output-finite-state-polytopic stochastic system. In fact, it is the filter system of the previous example as will be proven in Section 9.7. This system is defined by the representation,

$$\begin{aligned}
\hat{x}_{un}(t+1) &= A\ \mathrm{Diag}(C^T y(t))\ \hat{x}_{un}(t),\ \hat{x}_{un}(0) = E[x_0] = p_{x_0}, \\
\hat{x}(t+1) &= \frac{\hat{x}_{un}(t+1)}{1_{n_x}^T\ \hat{x}_{un}(t+1)} \in X_p \subseteq \mathbb{R}_{st}^{n_x}, \\
E[y(t) |\ F_t^{\hat{x}} \vee F_{t-1}^y] &= C\hat{x}(t),\ n_y = 2,\ n_{\hat{x}} = n_{\hat{x}_{un}} = 2, \\
& \qquad \hat{x}(t) \in X_p \subsetneq \mathbb{R}_{st}^{n_x},\ X_{un} = \mathbb{R}_+^{n_x},\ y(t) \in Y = Y_e.
\end{aligned}$$

The representation of a finite stochastic system is simplified if a condition of conditional independence holds.

Definition 5.7.4. Consider an output-finite-state-finite stochastic system. The system is called *state-output conditionally independent* conditioned on past states and past outputs if,

$$\begin{aligned}
& E[I_{\{x(t+1)=w_x, y(t)=w_y\}} | F_t^{x-} \vee F_{t-1}^{y-}] \\
&= E[I_{\{x(t+1)=w_x\}} | F_t^{x-} \vee F_{t-1}^{y-}]\ E[I_{\{y(t)=w_y\}} | F_t^{x-} \vee F_{t-1}^{y-}],\ \forall\ (w_x, w_y) \in X_e \times Y_e; \\
&\Leftrightarrow (F^{x(t+1)}, F^{y(t)} | F_t^{x-} \vee F_{t-1}^{y-}) \in \mathrm{CI},\ \forall\ t \in T.
\end{aligned} \tag{5.17}$$

It depends on the specification of the finite stochastic system considered whether or not the conditional independence property holds. Next a new representation is defined.

Definition 5.7.5. Consider a output-finite-state-finite stochastic system in the indicator representation, $\{\Omega, F, P, T, Y, X, y, x, \text{cpdf}, p_{x_0}\} \in$ OFSFstocsys. Assume that for all $t \in T$, $E[x(t)] > 0$ which by convention means that for all $i \in \mathbb{Z}_{n_x}$, $p_{x(t),i} = E[x_i(t)] > 0$.

Define respectively the *state-transition matrix function* and the *output-transition matrix function* as,

$$\begin{aligned}
&A : T \to \mathbb{R}_{st}^{n_x \times n_x},\ C : T \to \mathbb{R}_{st}^{n_y \times n_x},\ n_x,\ n_y \in \mathbb{Z}_+,\\
&A_{i,j}(t) = E[x_i(t+1)x_j(t)](E[x_j(t)])^{-1},\ \forall\, i, j \in \mathbb{Z}_{n_x},\\
&C_{i,j}(t) = E[y_i(t)x_j(t)](E[x_j(t)])^{-1},\ \forall\, i \in \mathbb{Z}_{n_y},\ j \in \mathbb{Z}_{n_x}.
\end{aligned}$$

Note that then,

$$\begin{aligned}
&1_{n_x}^T x(t) = \sum_{j=1}^{n_x} x_j(t) = 1,\ 1_{n_y}^T y(t) = 1,\ \forall\, t \in T,\\
&1_{n_x}^T A(t) = (E[1_{n_x}^T x_i(t+1)x_j(t)]\ (E[x_j(t)])^{-1})_{j \in \mathbb{Z}_{n_x}} = 1_{n_x}^T,\\
&\Rightarrow A(t) \in \mathbb{R}_{st}^{n_x \times n_x};\ C(t) \in \mathbb{R}_{st}^{n_y \times n_x},\ \text{similarly;}\\
&E[x(t+1)|F_t^{x-} \vee F_{t-1}^{y-}] = E[x(t+1)|F^{x(t)}] = A(t)x(t),\\
&\quad \text{where the first equality holds because it is a stochastic system,}\\
&E[y(t)|F_t^{x-} \vee F_{t-1}^{y-}] = E[y(t)|F^{x(t)}] = C(t)x(t),\ \forall\, t \in T.
\end{aligned}$$

The last two formulas follow from Proposition 2.8.4.

The system representation is then,

$$E\left[\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix} |F_t^x \vee F_{t-1}^y\right] = \begin{pmatrix} A(t) \\ C(t) \end{pmatrix} x(t),\ \forall\, t \in T.$$

Proposition 5.7.6. *Consider an output-finite-state-finite stochastic system in the indicator representation.*

(a)In general, the transition function of the finite stochastic system in the indicator represenation of Def. 5.7.1, cannot be determined from the matrix functions $A : T \to \mathbb{R}_+^{n \times n}$ and $C : T \to \mathbb{R}_+^{p \times n}$.

(b)Assume that the system is state-output conditionally independent given the past of the state and of the output process as defined in Def. 5.7.4. Then the matrix functions A and C with the current state $x(t)$ determine the probabilistic transition function of the finite stochastic system, in fact,

$$E[x(t+1)y(t)^T \mid F_t^x \vee F_{t-1}^y] = A(t)\text{Diag}(x(t))C(t)^T,\ \forall\, t \in T.$$

Proof. (a) The following expression cannot be determined from the matrix functions A and C, $E[x(t+1)y(t)^T|F_t^{x-} \vee F_{t-1}^{y-}]$.

(b) Note that, because of the state-output conditional independence of Def. 5.7.4,

$$\begin{aligned}
&\forall\, t \in T,\ (i,j) \in \mathbb{Z}_{n_x} \times \mathbb{Z}_{n_y}, \\
&E[x_i(t+1) y_j(t) | F_t^{x-} \vee F_{t-1}^{y-}] \\
= {} &E[x_i(t+1) | F_t^{x-} \vee F_{t-1}^{y-}] \times E[y_j(t) | F_t^{x-} \vee F_{t-1}^{y-}] \\
&\text{by the assumed state-output conditional independence,} \\
= {} &E[x_i(t+1) | F^{x(t)}] E[y_j(t) | F^{x(t)}],\ \text{by definition of a OFSFstocsys,} \\
= {} &(\sum_{k \in \mathbb{Z}_{n_x}} A_{ik}(t) x_k(t))(\sum_{m \in \mathbb{Z}_{n_x}} C_{jm}(t) x_m(t)) = \sum_{k \in \mathbb{Z}_{n_x}} A_{ik}(t) C_{jk}(t)\, x_k(t); \\
&\text{because } k \neq m \ \Rightarrow\ x_k(t) x_m(t) = 0,\ \ k = m \ \Rightarrow\ x_k(t) x_m(t) = x_k(t).
\end{aligned}$$

□

The reader should be careful with the literature on output-finite-state-finite stochastic systems. Many papers use as model that of Def. 5.7.5 and do not realize that this model is a strict subset of the general case of Def. 5.7.1.

Proposition 5.7.7. *Consider a time-varying output-finite-state-finite stochastic system in the indicator representation. Assume that the condition of state-output conditional independence holds, see Def. 5.7.4.*

(a)Then the probability measures of the state process x and of the output process y satisfy,

$$\begin{aligned}
p_x : T \to \mathbb{R}_{st}^{n_x},\ p_x(t) = E[x(t)], \\
p_y : T \to \mathbb{R}_{st}^{n_y},\ p_y(t) = E[y(t)], \\
\begin{pmatrix} p_x(t+1) \\ p_y(t) \end{pmatrix} = \begin{pmatrix} A(t) \\ C(t) \end{pmatrix} p_x(t),\ p_x(0) = p_{x_0} \in \mathbb{R}_{st}^{n_x}.
\end{aligned}$$

(b)For all $t \in T$,

$$\begin{aligned}
E[x(t) x(t)^T] &= \mathrm{Diag}(p_x(t)), \\
E[\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix} x(t)^T] &= \begin{pmatrix} A(t) \\ C(t) \end{pmatrix} \mathrm{Diag}(p_x(t)), \\
E[y(t+1) x(t)^T] &= C(t) A(t) \mathrm{Diag}(p_x(t)), \\
E[y(t+1) y(t)^T] &= C(t) A(t) \mathrm{Diag}(p_x(t)) C(t)^T; \\
E[x(t) x(s)^T] &= \left(\prod_{r=s}^{t-1} A(r) \right) \mathrm{Diag}(p_x(s)),\ \forall\, t,\ s \in T,\ s < t, \\
E[y(t) y(s)^T] &= C(t) \left(\prod_{r=s}^{t-1} A(r) \right) \mathrm{Diag}(p_x(s)) C(s)^T, \\
&\forall\, t,\ s \in T,\ s < t.
\end{aligned}$$

Proof. (a) One calculates using conditional independence,

$$\begin{aligned}
p_x(t+1) &= E[x(t+1)] = E[E[x(t+1) | F_t^x \vee F_{t-1}^y]] = E[A(t) x(t)] = A(t) p_x(t), \\
p_y(t) &= E[y(t)] = E[E[y(t) | F_t^x \vee F_{t-1}^y]] = E[C(t) x(t)] = C(t) p_x(t).
\end{aligned}$$

(b)

$$
\begin{aligned}
&\forall\, i,\ j \in \mathbb{Z}_{n_x},\ i \neq j,\ \ E[x_i(t)x_j(t)] = E[I_{\overline{x}(t)=\overline{x}_i} I_{\overline{x}(t)=\overline{x}_j}] = 0,\\
&E[x_i(t)x_i(t)] = E[x_i(t)] = p_{x,i}(t)\ \Rightarrow\ E[x(t)x(t)^T] = \text{Diag}(p_x(t));\\
&E[y(t)x(t)^T] = E[E[y(t)|\, F_t^x \vee F_{t-1}^y]x(t)^T]\\
=\ &E[C(t)x(t)x(t)^T] = C(t)\text{Diag}(p_x(t));\\
&E[y(t+1)x(t)^T] = E[E[y(t+1)x(t)^T|F_{t+1}^x \vee F_t^y]]\\
=\ &E[E[y(t+1)|F_{t+1}^x \vee F_t^y]x(t)^T] = E[C(t+1)x(t+1)x(t)^T]\\
=\ &C(t+1)E[E[x(t+1)x(t)^T|F_t^x \vee F_{t-1}^y]]\\
=\ &C(t+1)E[E[x(t+1)|F_t^x \vee F_{t-1}^y]x(t)^T]\\
=\ &C(t+1)A(t)E[x(t)x(t)^T] = C(t+1)A(t)\text{Diag}(p_x(t)),\ \text{etc.}
\end{aligned}
$$

□

Of interest is the formula of the probability distributions of a finite set of states and of a finite set of outputs as in,

$$
\begin{aligned}
&P(\{x(t_1),\ x(t_2),\ \ldots,\ x(t_k)\} \in H_k),\ \ P(\{y(t_1),\ y(t_2),\ \ldots,\ y(t_k)\} \in G_k),\\
&t_1,\ t_2,\ \ldots,\ t_k \in T,\ t_1 < t_2 < \ldots < t_k.
\end{aligned}
$$

The formulas for these finite-dimensional probability distributies can be written out as functions of the matrix functions A and C but the resulting algebraic expressions do not have the nice structure as those of Gaussian processes. In the above proposition are presented only the finite-dimensional probability distributions in case of tuples of states and of tuples of outputs.

Output-finite-state-polytopic stochastic systems are useful mathematical models of, for example, communication systems where the variables often take values in finite sets. They may also be useful as mathematical models for other areas of engineering in combination with approximation techniques. In such an approximation technique, the continuous state and output spaces are discretized into finite sets, for example $\mathbb{Z}_k$ for an integer $k \in \mathbb{Z}_+$, and the state and output process are redefined accordingly. This approximation technique is used for example in control of stochastic systems.

5.7.2 Time-Invariant Stochastic Systems

Example 5.7.8. A *communication system* was defined by C. Shannon of Bell Laboratories in a paper published in 1948. The engineering model consists of: a source, a communication channel, and a receiver. The reader may think for a source of a Markov process which generates a sequence of states and the state process is a Markov process. The states are sent via a communication channel to a receiver. The main contribution of Shannon to science is the formulation of abstract problems of information, the formulation of the communication system and the main concepts of information theory, and for solving problems of information theory for special cases using the formulated concepts.

Consider the specific model in which the channel has no memory. Assume that the indicator representation of these processes is used and that the system is time-invariant. Moreover, the state-output conditional independence relation will be assumed. Then one may model the communication system by the formulas,

$$\begin{aligned}
X_e &= \{e_1,\, e_2 \in \mathbb{R}^2_+\},\ Y_e = \{e_1,\, e_2 \in \mathbb{R}^2_+\},\\
x_1(t) &= I_{\{\bar{x}(t)=0\}},\ x_2(t) = I_{\{\bar{x}(t)=1\}},\\
y_1(t) &= I_{\{\bar{y}(t)=0\}},\ y_2(t) = I_{\{\bar{y}(t)=1\}},\\
E\left[\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix} | F^x_t \vee F^y_{t-1}\right] &= \begin{pmatrix} A \\ C \end{pmatrix} x(t), \qquad (5.18)\\
A = \begin{pmatrix} 1-p_1 & p_2 \\ p_1 & 1-p_2 \end{pmatrix}&,\ C = \begin{pmatrix} 1-p_3 & p_4 \\ p_3 & 1-p_4 \end{pmatrix}. \qquad (5.19)
\end{aligned}$$

Realistic values for the parameters of a simple engineering model are $p_1 = 0.45$, $p_2 = 0.53$, $p_3 = 10^{-4}$, and $p_4 = 10^{-4}$.

An engineer may be interested in the formulas for the following expressions,

$$E[y(t)|F^x_t \vee F^y_{t-1}],\ \ E[y(t+1)|F^x_t \vee F^y_{t-1}],\ \ E[y(t)|F^y_{t-1}].$$

The formulas for those expressions are developed in Chapter 9 on filtering of stochastic systems.

Definition 5.7.9. Consider a *time-invariant output-finite-state-finite stochastic system* in the indicator represenation, assuming that state-output conditional independence holds. This system has the representation,

$$\begin{aligned}
&E\left[\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix} | F^x_t \vee F^y_{t-1}\right] = \begin{pmatrix} A \\ C \end{pmatrix} x(t),\ x(0) = x_0,\\
&X_e = \{e_1,\, e_2, \ldots,\, e_{n_x} \in \mathbb{R}^{n_x}_{st}\},\ Y_e = \{e_1,\, e_2, \ldots,\, e_{n_y} \in \mathbb{R}^{n_y}_{st}\},\ n_x,\ n_y \in \mathbb{Z}_+,\\
&x : \Omega \times T \to X_e,\ y : \Omega \times T \to Y_e,\ A \in \mathbb{R}^{n_x \times n_x}_{st},\ C \in \mathbb{R}^{n_y \times n_x}_{st}.
\end{aligned}$$

Corollary 5.7.10. *Consider the time-invariant finite stochastic system of Def. 5.7.9. From Proposition 5.7.7 follows that,*

$$\begin{aligned}
p_x(t) &= E[x(t)] = A^t p_x(0),\ \ p_y(t) = E[y(t)] = CA^t p_x(0),\\
&\quad p_x : T \to \mathbb{R}^{n_x}_{st},\ p_y : T \to \mathbb{R}^{n_y}_{st},\\
E[\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix} x(t)^T] &= \begin{pmatrix} A \\ C \end{pmatrix} \text{Diag}(p_x(t)),\\
E[x(t)x(s)^T] &= A^{t-s}\text{Diag}(p_x(s)),\ \forall\, t,\ s \in T,\ s < t,\\
E[y(t)y(s)^T] &= CA^{t-s}\text{Diag}(p_x(s))C^T,\ \forall\, t,\ s \in T,\ s < t.
\end{aligned}$$

5.7.3 State Set a Polytope

Below is explored the concept of an output-finite-state-polytopic stochastic system and its representation.

Recall from Def. 5.7.1 that an output-finite-state-polytopic stochastic system has as state set a polytope $X_p \subseteq \mathbb{R}_{st}^{n_x}$ within the probability simplex. In that definition, the condition is imposed that the state set is invariant with respect to the state-transition dynamics. Consider a time-invariant output-finite-state-polytopic stochastic system in the representation of Def. 5.7.9 which is in the indicator representation and satisfies the state-output conditional independence assumption.

To satisfy for the selected representation the condition of invariance of the state polytope with respect to the state dynamics of the state-transition matrix, there are two approaches possible:

1. One fixes the state transition matrix $A \in \mathbb{R}_{st}^{n_x \times n_x}$. Then one determines the state polytope X_p such that it is invariant with respect to the state transition matrix A. This approach is followed below.
2. One fixes the state polytope X_p. Then one chooses the state transition matrix A such that this matrix leaves the state polytope invariant. This approach is less natural than the previous one. This approach is not detailed further in this book.

In either case, the condition of the invariance of the state polytope with respect to the invariance by the state dynamics is achieved.

The above distinction for the construction of a time-invariant output-finite-state-polytopic stochastic system has a direct effect on the concept of the state-space dimension of such a system. It will hopefully become clear that the minimal state-space dimension equals the minimal number of positively-independent vectors which span the state polytope. This number will in general be different from what that of the system representation of Def. 5.7.9 if that representation did not yet satisfy the invariance of the state polytope by the state dynamics. This issue may become clear after further discussion and examples.

The author has chosen to write in this section the first approach formulated above. The second approach is an alternative. It remains to be seen which approach is most useful in the long run.

Below the first approach mentioned above is adopted. Consider thus a system representation of a time-invariant output-finite-state-polytopic stochastic system of Def. 5.7.9. The question is whether the state polytope is invariant with respect to the state dynamics.

The next result is based on joint research of Y. Zeinaly and B. De Schutter with the author.

Theorem 5.7.11. *Consider a time-invariant output-finite-state-polytopic stochastic system, Def. 5.7.1, and assume that state-output conditional independence holds. The system representation in the indicator representation is,*

$$E\left[\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix} \mid F_t^x \vee F_{t-1}^y\right] = \begin{pmatrix} A \\ C \end{pmatrix} x(t),\ x(0) = x_0 \in X_{p,0},$$
$$X_p \subseteq \mathbb{R}_{st}^{n_x},\ \textit{a polytope},\ X_{p,0} \subseteq X_p,\ Y_e = \{e_1,\ e_2,\ e_{n_y} \in \mathbb{R}_{st}^{n_y}\},$$
$$A \in \mathbb{R}_{st}^{n_x \times n_x},\ C \in \mathbb{R}_{st}^{n_y \times n_x}.$$

The state set X_p is a polytope and is invariant with respect to the state-transition matrix A if and only if,

$$\exists\ n_r \in \mathbb{Z}_+,\ \exists\ a_0,\ a_1,\ \ldots,\ a_{n_r-1} \in \mathbb{R}_+,\ \textit{such that}$$
$$A^{n_r} = \sum_{i=0}^{n_r-1} a_i A^i \ \textit{and} \sum_{i=0}^{n_r-1} a_i = 1.$$

Definition 5.7.12. Consider Theorem 5.7.11. Define the following expression as a *positive recursion* of the system matrix A,

$$\exists\ n_r \in \mathbb{Z}_+,\ \exists\ a_0,\ a_1,\ \ldots,\ a_{n_r-1} \in \mathbb{R}_+,\ \text{such that}\ A^{n_r} = \sum_{i=0}^{n_r-1} a_i A^i. \tag{5.20}$$

Call the smallest integer $n_r \in \mathbb{Z}_+$ for which a positive state-transition recursion holds, the *positive-recursion number* of the state-transition matrix A. If a positive recursion exists with the coefficients not all zero, then division of all coefficients by the sum $(\sum_{i=0}^{n_r-1} a_i)$ yields a convex recursion of A as needed in Theorem 5.7.11.

In general, the problem to determine for a stochastic matrix A whether a finite positive-recursion number exists, is an undecidable problem as understood in computer science. This undecidability is similar to the existence of a finite-rank of the Hankel matrix in realization of a linear system, see Theorem 21.8.9. This analogy is quite appropriate for the positive-recursion number.

Assume that a positive-recursion number n_r exists. Define the polytope, $X(A) \subseteq \mathbb{R}_{st}^{n_x}$ as the smallest polytope containing all columns of the matrices $I,\ A,\ A^2,\ \ldots,\ A^{n_r-1}$. Choose a generator for the polytope $X(A)$ by selection of a matrix $S \in \mathbb{R}_{st}^{n_x \times n_p}$ such that $X(A) = \text{cone}(S) \cap \mathbb{R}_{st}^{n_x} = \text{Polytope}(S)$. This is called the implicit representation in Def. 18.4.1. The columns of the matrix S have thus to be positively independent as defined in Def. 18.4.4. Call the columns of the matrix S a *polytopic basis-representation* of the polytope $X(A)$. The computations are best done first for the polyhedral cone $\text{cone}(X(A))$ after which a restriction is imposed to the probability simplex by intersection with $\mathbb{R}_{st}^{n_x}$. There exist software packages for polyhedral cones.

In terms of formulas,

$$A^{n_r} = \sum_{i=0}^{n_r-1} a_i A^i,\ \sum_{j=0}^{n_r-1} a_j = 1;\ n_r,\ n_p \in \mathbb{Z}_+,$$
$$\text{cone}(A) \subseteq \mathbb{R}_{st}^{n_x},\ \text{the smallest cone such that,}$$
$$\forall\ i \in \mathbb{N}_{n_r-1} = \{0, 1,\ \ldots, n_r-1\},\ \text{cone}(A^i) \subseteq \text{cone}(A);$$
$$X(A) = \text{cone}(A) \cap \mathbb{R}_{st}^{n_x},\ \text{a polytope;}$$
$$S \in \mathbb{R}_{st}^{n_x \times n_p},\ \ n_p \in \mathbb{Z}_+,\ \text{columns of } S \text{ positively independent such that,}$$
$$\text{Polytope}(S) = X(A).$$

Thus the columns of the stochastic matrix S form a positively-independent basis of the polytope $X(A)$. In general $n_p \neq n_r$.

An output-finite-state-polytopic stochastic system is thus described by the tuple, $(n_y, \, n_x, \, n_r, \, n_p, \, A, \, C, \, S)$.

A long explanation of the above theorem and definition follows.

It is known from realization theory of output-finite-state-finite stochastic systems, see [28], that the set of conditional probability measures of the output process, is a polytope which is invariant by application of the system matrix. The conditions that the state set is a polytopen and that the state set is invariant with respect to the state-transition dynamics implies the existence of a positive-recursion of the state-transition matrix. Negation of the positive-recursion results in a state set which is neither a polytope nor a polyhedral set. The assumption of a polytopic state set implies the condition that,

$$\exists \; n_r \in \mathbb{Z}_+, \; \exists \; a_0, \; a_1, \ldots, a_{n_r} \in \mathbb{R}_+, \text{such that, } A^{n_r} = \sum_{i=0}^{n_r - 1} a_i A^i.$$

Note that if $n_r \in \mathbb{Z}_+$ exists, then for all $k \geq n_r$, A^k satisfies a similar positive recursion. This is directly proven by recursive substition of powers of A using the above formula. The positive-recursion number n_r is defined to be the smallest integer for which a positive recursion exists.

Geometrically, this means that $X(A) \subseteq \mathbb{R}_{st}^{n_x}$ is a polytope which is invariant with respect to applications with the matrix A and it is the smallest polytope satisfying this invariance condition.

Geometrically there are several related objects, see Section 18.4. The first type of concept is that of a cone and a polyhedral cone. The second type of concept is that of a polytope inside the probability simplex. Both objects are related. A polyhedral cone determines a polytope by intersection with the probability simplex as in $X(A) = \text{cone}(A) \cap \mathbb{R}_{st}^{n_x}$. Conversely, a polytope generates a polyhedral cone by $\text{cone}(A) = \text{cone}(\text{Polytope}(A))$.

The approach to determine the minimal positive-recursion number of the state transition matrix is related to the concept of an extremal polyhedral cone for a considered polyhedral cone. Algebraically it is related to the concept of a prime in the positive matrices. See Section 18.9.6.

If the condition on the existence of n_r was not imposed then there exists an example where $X(A) \subset \mathbb{R}_{st}^{n_x}$ is a subset which is not a polytope, more specifically, it is not a polyhedral set. If the subset $X(A)$ is considered as generating a cone in $\mathbb{R}_+^{n_x}$ then that cone can in a particular cases be a nonpolyhedral cone. If one accepts the condition on the existence of a stochastic realization with a polytopic state set then the subset X_p is a polytope and consequently there exists a positive-recursion integer $n_r \in \mathbb{Z}_+$.

A research issue of output-finite-state-polytopic stochastic systems is the dimension of the state vector. Stochastic realization theory of this set of stochastic systems, Section 7.5 yields a necessary and sufficient condition for the existence of a stochastic realization as an output-finite-state-polytopic stochastic system. The object of interest is a polytope of conditional measures in the probability simplex. Because it is a polytope, it has a finite number of spanning vectors. In addition, the polytope is invariant with respect to shifts of the outputs which is equivalent to invariance by the state dynamics.

From the above result of weak stochastic realization of an output-finite stochastic process one concludes that the number of states $n_p \in \mathbb{Z}_+$ is best chosen as the number of spanning vectors of the polytope as described above. The system representation is then such that the state-transition matrix is a stochastic matrix $A \in \mathbb{R}_{st}^{n_x \times n_x}$ which leaves the state polytope invariant by its application. Hence,
$X_p = \text{Polytope}(A) = \text{cone}(A) \cap \mathbb{R}_{st}^{n_x}$ represents the state polytope with n_p spanning vectors. Then the effective dimension of the state set is the integer $n_p \in \mathbb{Z}_+$. In general this is different from the integer n_x with which one started the system representation.

An output-finite-state-finite stochastic system has as a characteristic thus four integers $(n_y,\ n_x,\ n_r,\ n_p) \in \mathbb{Z}_+^4$. In addition to the matrices $A,\ C$ the system has also a polytopic basis formed by the columns of the matrix $S \in \mathbb{R}_{st}^{n_x \times n_p}$ with $X(A) = \text{Polytope}(S)$. An output-finite-state-finite stochastic system is thus described by the tuple, $(n_y,\ n_x,\ n_r,\ n_p,\ A,\ C,\ S)$. The reader may notice that the linear rank of the matrix S and hence the linear rank of the representation of the state polytope, does not play any role in the characterization of this stochastic system.

Note that in general, the state polytope with n_p spanning vectors constructed in stochastic realization need not be such that the set of all measures of this stochastic realization equals $\mathbb{R}_{st}^{n_x}$. It can be strictly smaller, $X_p \subsetneq \mathbb{R}_{st}^{n_x}$.

There follow several examples which illustrate the definition of an output-finite-state-polytopic stochastic system.

Example 5.7.13. There exists a positive matrix $A \in \mathbb{R}_+^{n_x \times n_x}$ for which there does not exist a finite integer $n_p \in \mathbb{Z}_+$ satisfying a positive recursion of the state-transition matrix A. The example does this for the controllable subset of a positive linear system, its controllable subset keeps growing forever and the limit of the controllable subset is not a polyhedral set.

Example 5.7.14. There exists a positive matrix $A \in \mathbb{R}_+^{n_x \times n_x}$ of a discrete-time positive linear system where the controllable subset keeps growing forever, yet the limit of the controllable subset is a polyhedral set,

$$\begin{aligned} & x(t+1) = Ax(t) + Bu(t),\ x(0) = x_0,\ x(t) \in X = \mathbb{R}_+^3; \\ & \qquad \text{then } \forall\ k \in \mathbb{Z}_+,\ A^k \notin \text{cone}(\{I,\ A,\ A^2, \ldots, A^{k-1}\}). \end{aligned}$$

However, the closure of the infinite-time controllable subset is a polyhedral set.

Example 5.7.15. Consider a time-invariant output-finite-state-polytopic stochastic system as defined in Def. 5.7.1. Then the positive recursion number n_r, if it exists, is strictly larger than n_x.

$$A = \begin{pmatrix} 1\ 0\ 2 \\ 2\ 1\ 0 \\ 0\ 2\ 1 \end{pmatrix} \times (1/3) \in \mathbb{R}_{st}^{3\times 3},\ n_x = 3,$$

$$\Rightarrow \text{cone}(A^3) \not\subseteq \text{cone}(I,\ A,\ A^2),\ \text{cone}(A^4) \not\subseteq \text{cone}(I,\ A,\ A^2,\ A^3).$$

The claims are simple computations. If $\text{cone}(A^3) \subseteq \text{cone}(I,\ A,\ A^2)$ hence there exist $a_0,\ a_1,\ a_2 \in \mathbb{R}_+$ such that $A^3 = \sum_{i=0}^{2} a_i A^i$, then one solves the above equation for the coefficients $a_0,\ a_1,\ a_2$. It is then quickly discovered that such coefficients cannot exist in $\mathbb{R}_+$.

The conclusion is that in general the positive recursion number $n_r \in \mathbb{Z}_+$ can be strictly larger than n_x. Nothing is said about the existence of an $n_r \in \mathbb{Z}_+$ for this matrix.

5.7.4 Decompositions of the State Set

The reader finds below a classification of output-finite-state-finite stochastic systems in terms of a decomposition into subsystems. This decomposition is fully determined by the classification of the state transition matrix A, which is a stochastic matrix, and is described in Section 18.8. The topic of this subsection is quite well known in the literature.

The reader of this section is expected to have knowledge of stochastic matrices, in particular on the existence of a steady state and of convergence to a steady state, as provided in Section 18.8.

To assist the reader, please find below an overview of the decomposition of an output-finite-state-finite stochastic system as clarified below:

- irreducible subsystems:
  - irreducible and nonperiodic subsystem,
  - irreducible and partly-periodic subsystem,
  - irreducible and pure-periodic subsystem;
- fully-reduced subsystem.

The set of state transformations for a finite-state Markov process is the set of the permutation matrices, denoted by $\mathbb{R}_{perm}^{n_x \times n_x}$. The reader is expected to have read the definition of a permutation matrix, Def. 18.3.1, and the property that the inverse of a permutation matrix is its transpose, $Q \in \mathbb{R}_{perm}^{n_x \times n_x}$ implies that $Q^{-1} = Q^T$, see Theorem 18.5.3. Note that for any permutation matrix $Q \in \mathbb{R}_{perm}^{n_x \times n_x}$ and, for any $i \in \mathbb{Z}_{n_x}$, there exists a $j \in \mathbb{Z}_{n_x}$ such that the vector e_i is mapped by Q to the vector e_j, $Qe_i = e_j$.

Of interest is the transformation of the state set of a finite-state Markov process to a particular form. In the literature such a form has been defined and investigated. The reader could check the decompostion of a matrix based on permutation similarity, see Def. 18.6.2.

The decomposition of a finite-state Markov process is motivated by the problem: Whether there exists an invariant measure of the state process? What is the invariant measure if it exists? Does convergence to an invariant measure take place? The reader should keep this problem in mind when reading the following examples.

The general definition is preceded by several examples. The first example introduces the notation and the terminology.

Example 5.7.16. Consider a finite-state Markov process in the indicator representation,

$$E[x(t+1)|\, F_t^x] = Ax(t),\ x(0) = x_0,$$

$$A = \left(\begin{array}{cc|cc|cc|cc} 0 & 1 & 0 & 1/3 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 2/3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 & 0 & 3/4 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{array} \right) \in \mathbb{R}_{st}^{n_x \times n_x}.$$

Define several subsets of the state set based on the structure of the above defined transition matrix A as,

$$X = \mathbb{Z}_8 = \{1,2,\ldots,7,8\},$$
$$X_1 = \{1,2\},\ X_2 = \{3,4\},\ X_3 = \{5,6\},\ X_4 = \{7,8\},\ \cup_{i=1}^4\ X_i = X;$$

the state transitions of the subsets of the state set are,

$$X_4\ \rightarrow\ X_3 \cup X_4,\ X_3\ \rightarrow\ X_3,\ X_2\ \rightarrow\ X_1 \cup X_2,\ X_1\ \rightarrow\ X_1.$$

These transitions can be read from the above displayed matrix: For the transitions from X_4, read the fourth block-column of the matrix A. If there are nonzero entries then read off the subsystems according to the block-rows, thus $X_4 \rightarrow X_3 \cup X_4$. In the third column there is only a transition to the third block-row then the transition is $X_3 \rightarrow X_3$. The set $\{X_1,\ X_2,\ X_3,\ X_4\}$ is called a partition of the state set X.

Consider an initial state of the form,

$$x_0 = \begin{pmatrix} 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0 & 0 \end{pmatrix}^T \in \mathbb{R}_{st}^8,\ \ n_{s+}(x_0) = 2,\ i(x_0) = \{3,4\} = X_2.$$

The initial state is said to have *support* in the state subset X_2. The support of the initial state of a finite stochastic system is defined to be those states of the state set for which the probability distribution of the initial state is strictly positive. Thus the support of the initial state above is $i(x_0) = \{3,4\} = X_2$.

It is then clear from the state transition matrix that the future states of the system, $x(1)$, $x(2)$, etc. have support only in the state subsets $X_1 \cup X_2$. This is denoted above by the relations, $X_2\ \rightarrow\ X_1 \cup X_2$ and $X_1 \rightarrow X_1$.

Consider the case in which the probability distribution of the initial state x_0 is chosed such that $i(x_0) = \{3,4\}$. It is then clear that the probability distribution of the state, $p_x(t) = E[x(t)] \in \mathbb{R}_{st}^8$, when times goes to infinity, is such that the probability mass on the state subset X_2 goes to zero while eventually the support of the state will be only in the state subset X_1. Call the state at time infinity the *terminal state*, even though the terminal state is reached only asymptotically, when time goes to infinity. Thus, eventually, the support of the probability measure will be $X_1 = \{1,2\}$, which is denoted by $n_{s+}(p_x(\infty)) = 2$ and $i(p_x(\infty)) = \{j \in \mathbb{Z}_{n_x} |\ p_{x,j}(0) > 0\} = \{1,2\} = X_1$ using notation of Chapter 18.

Correspondingly, if the initial state has as support only the state subset X_4 then the transition matrix implies the transformations $X_4 \rightarrow X_3 \cup X_4$ and $X_3 \rightarrow X_3$. Eventually, the probability distribution of the state has as support only the state subset X_3.

Call the state subsets X_4 and X_2 the *initial state subsets* and the state subsets X_3 and X_1 the *terminal state subsets*.

The conclusions of this example are that for a finite stochastic system: (1) the dynamic behavior of the probability distribution of the state depends on the probability distribution of the initial state; (2) the support of the probability distribution of the terminal state may be different from the support of the probability distribution of the initial state.

Example 5.7.17. Consider a finite-state Markov process with representation,

$$
\begin{aligned}
& E[x(t+1)|\ F_t^x] = Ax(t),\ x(0) = x_0, \\
A = & \left(\begin{array}{cc|cc|cc}
2/3 & 1/3 & 0 & 0 & 0 & 0 \\
1/3 & 2/3 & 0 & 0 & 1/3 & 0 \\
\hline
0 & 0 & 1/2 & 1/2 & 0 & 1/3 \\
0 & 0 & 1/2 & 1/2 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 2/3 \\
0 & 0 & 0 & 0 & 2/3 & 0
\end{array}\right) \in \mathbb{R}_{st}^{6\times 6}; \\
X = & \{1,2,\ldots,5,6\},\ X_1 = \{1,2\},\ X_2 = \{3,4\},\ x_3 = \{5,6\}; \\
& X_3 \rightarrow X_1 \cup X_2 \cup X_3,\ X_2 \rightarrow X_2,\ X_1 \rightarrow X_1.
\end{aligned}
$$

If the initial state has a probability distribution with support on the state subset X_3 then the terminal state will have a probability distribution on the state set $X_1 \cup X_2 \cup X_3$ even though that subset consists of two parts with separate dynamics, $X_1 \rightarrow X_1$ and $X_2 \rightarrow X_2$. In this example X_3 is the initial state subset while both X_1 and X_3 are terminal state subsets.

The conclusion of this example of a finite stochastic system is that: (1) the terminal state may have a probability distribution with support in the union of two or more state subsets; and (2) the state transition function determines how much mass each of the subsets receives in the probability distribution of the terminal state.

Example 5.7.18. Consider a finite-state Markov process with representation,

$$\begin{aligned}
& E[x(t+1)|\, F_t^x] = Ax(t),\ x(0) = x_0, \\
A = & \left(\begin{array}{cc|cc|cc}
1/9 & 7/8 & 0 & 1/5 & 0 & 0 \\
8/9 & 1/8 & 1/6 & 0 & 0 & 0 \\ \hline
0 & 0 & 0 & 4/5 & 0 & 1/3 \\
0 & 0 & 5/6 & 0 & 1/4 & 0 \\ \hline
0 & 0 & 0 & 0 & 0 & 2/3 \\
0 & 0 & 0 & 0 & 3/4 & 0
\end{array}\right) \in \mathbb{R}_{st}^{6\times 6}; \\
X = & \{1,2,\ldots,5,6\},\ X_1 = \{1,2\},\ X_2 = \{3,4\},\ x_3 = \{5,6\}; \\
& X_3 \rightarrow X_2 \cup X_3,\ X_2 \rightarrow X_1 \cup X_2,\ X_1 \rightarrow X_1.
\end{aligned}$$

If the initial state has a probability distribution with support on the state subset X_3 then the probability distribution moves to the state subset X_2 and from X_2 it moves to the state subset X_1 eventually. The support of the terminal state is only the subset X_1 and neither X_3 nor X_2. Call the state subset X_2 the *transient state subset* because the probability distribution passes through it but does not remain there, hence is transient. Transient means that it passes by in time but will not remain.

The conclusions of this finite stochastic system is: (1) there exists a subset of the state set, X_2, which has neither support in the initial state nor in the terminal state but has support during intermediate times; (2) there is terminal subsystem which in the limit, when time goes to infinity, has all the support of the probability measure.

Example 5.7.19. Consider a finite-state Markov process with representation,

$$\begin{aligned}
& E[x(t+1)|\, F_t^x] = Ax(t),\ x(0) = x_0, \\
A = & \left(\begin{array}{cc|cc}
1/2 & 0 & 0 & 1/5 \\
1/2 & 2/3 & 0 & 0 \\ \hline
0 & 1/3 & 3/4 & 0 \\
0 & 0 & 1/4 & 4/5
\end{array}\right) \in \mathbb{R}_{st}^{4\times 4};\ X = \{1,2,3,4\},\ X \rightarrow X.
\end{aligned}$$

If the initial state has a probability distribution with support on the state subset X then the probability distribution at all times has support in the full state set X and so has the terminal state. The state transition matrix is an irreducible matrix as defined in Def. 18.6.2.

The conclusion of this example of a finite stochastic system is that: (1) there exists a system such that the support of the initial state, of the terminal state, and of all states equal the entire state set; and (2) in case the support of the initial state is strictly smaller than the full state set then the support of the terminal state equals the entire state set. The claim of the support of the state set follows from Theorem 5.7.22 stated below.

Definition 5.7.20. Consider a output-finite-state-finite stochastic system with only the state process $\bar{x}$. Based on permutation similarity, there exists a state set transformation in the form of a permutation matrix $Q \in \mathbb{R}_{perm}^{n_x \times n_x}$ such that the transformed system with state process $x(t) = Q\bar{x}(t)$ has a state transition matrix which is fully reduced, see the definition of the Frobenius canonical form, Def. 18.8.11, and as displayed in the equations below.

$$A = \left(\begin{array}{cccc|cccc}
A_{11} & 0 & 0 & 0 & A_{1,n_1+1} & A_{1,n_1+2} & \dots A_{1,n_2-1} & A_{1,n_2} \\
0 & A_{22} & \dots & 0 & A_{2,n_1+1} & A_{2,n_1+2} & \dots A_{2,n_2-1} & A_{2,n_2} \\
0 & 0 & \ddots & 0 & \vdots & \vdots & \vdots \;\; \vdots & \vdots \\
0 & 0 & \dots & A_{n_1,n_1} & A_{n_1,n_1+1} & A_{n_1,n_1+1} & \dots A_{n_1,n_2-1} & A_{n_1,n_2} \\
\hline
0 & 0 & \dots & 0 & A_{n_1+1,n_1+1} & A_{n_1+1,n_1+2} & \dots A_{n_1+1,n_2-1} & A_{n_1+1,n_2} \\
0 & 0 & \dots & 0 & 0 & A_{n_1+2,n_1+2} & \dots A_{n_1+2,n_2-1} & A_{n_1+2,n_2} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \vdots & \vdots \\
0 & 0 & \dots & 0 & 0 & 0 & \dots A_{n_2-1,n_2-1} & A_{n_2-1,n_2} \\
0 & 0 & \dots & 0 & 0 & 0 & \dots 0 & A_{n_2,n_2}
\end{array}\right)$$

$$\begin{aligned}
&\in \mathbb{R}_{st}^{n \times n}, \;\; \exists\, m_1,\, m_2 \in \mathbb{Z}_+, \\
&\exists\, n_1,\, n_2,\, \dots,\, n_{m_1} \in \mathbb{Z}_+,\, n_{m_1+1},\, n_{m_1+2}, \dots,\, n_{m_1+m_2} \in \mathbb{Z}_+, \\
&n_1 + \ldots + n_{m_1} + n_{m_1+1} + \ldots + n_{m_2} = n, \\
&\forall\, i \in \mathbb{Z}_{n_1},\, A_{i,i} \in \mathbb{R}_{st}^{n_i \times n_i} \text{ is an irreducible stochastic matrix}, \\
&\forall\, i \in \mathbb{Z}_{n_1},\, \forall\, j = n_1+1, \ldots, n_2,\, A_{i,j} \in \mathbb{R}_+^{n_i \times n_j}, \\
&\forall\, k_2 = n_1+1, \ldots, n_2,\, \exists\, k_1 \in \{1, \ldots, k_2-1\},\, A_{(i,j),(k_1,k_2)} \neq 0; \\
&\forall\, i = n_1+1, \ldots, n_2,\, A_{i,i} \in \mathbb{R}_+^{n_i \times n_i} \text{ is a substochastic matrix.}
\end{aligned}$$

Call the submatrices $A_{i,i}$ for all $i \in \mathbb{Z}_{n_1}$ the *terminal submatrices* and the submatrices $A_{i,i}$ for $i = n_1+1, \ldots, n_2$ the *transient submatrices*.

Call the subset $X_i \subset X$ for $i \in \mathbb{Z}_{m_1+i}$ an *initial subset* if either $i = 1$ or, for $i = 2, \ldots m_2$, for all $j \in \mathbb{Z}_{m_2}$ with $i < j$, $A_{i,j} = 0$. Call the subset $X_i \subset X$ for $i \in \mathbb{Z}_{m_1}$ a *terminal subset* if either $i = 1$ or for all $k \in \mathbb{Z}_m$ with $k < i$, $A_{k,i} = 0$. All subsets which are neither initial subsets nor terminal subsets are called *transient subsets* regardless of the initial state.

One may call the object described by $(X_i, A_{i,i})$ for all $i \in \mathbb{Z}_m$, the i-th subsystem of the finite stochastic system though this ignores the transitions to subsystem i from other subsystems, and it ignores the transitions from substem i to other subsystems.

The state transition matrix A then describes the dynamics between the state subsets according to the equations, $\forall\, i \in \mathbb{Z}_{n_x}$,

initial subsets of states $X_i \to \cup_{j \in \mathbb{Z}_m, j \leq i} X_j$;
transient subsets of states $X_i \to \cup_{j \in \mathbb{Z}_m, j \leq i} X_j$;
terminal subsets of states $X_i \to X_i$.

The examples above the definition describe several special cases of finite stochastic systems.

The complexity of a finite stochastic systems can be large due to the presence of many state subsets each with its own dynamics. However, asymptotically the support of the state probability distribution will be in a subset of the state set corresponding to an irreducible submatrix of the state transition matrix. It is to be recalled that the support of the probability distribution of the state of the system may be in two or more subsets each associated with an irreducible state transition matrix.

The standard assumption made in the literature for the asymptotic behavior of a finite state stochastic systems is to assume that the state transition matrix A is an irreducible matrix. This way one ignores the behavior of the system state on initial state subsets and on transient state subsets. The assumption therefore is to focus on the long term behavior of the state of the stochastic system. Even if the state has support in a nonterminal subsystem then it will eventually reach an irreducible terminal subsystem. This standard assumption is adopted in this book.

Definition 5.7.21. *An irreducible and nonperiodic output-finite-state-finite stochastic system.* Consider a time-invariant output-finite-state-finite stochastic system in the indicator representation and assume that state-output conditional independence holds. This stochastic system will be denoted by the forward representation,

$$E\left[\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix} | F_t^x \vee F_{t-1}^y\right] = \begin{pmatrix} A \\ C \end{pmatrix} x(t),\ x(0) = x_0,\ A \in \mathbb{R}_{st}^{n_x \times n_x},\ C \in \mathbb{R}_{st}^{n_y \times n_x}.$$

Call this finite stochastic system an *irreducible and nonperiodic output-finite-state-finite stochastic system* if the system matrix A is an irreducible and nonperiodic stochastic matrix.

Theorem 5.7.22. *Theorem 18.8.2 is copied below for the convenience of the reader.*

Consider an irreducible and nonperiodic output-finite-state-finite stochastic system of Def. 5.7.9 with as transition matrix the irreducible and nonperiodic stochastic matrix $A \in \mathbb{R}_{st}^{n_x \times n_x}$. Hence with the spectral index $n_{s+}(A) = (1, n-1)$, see Def. 18.8.3 and Def. 18.7.2.

(a)*There exists a unique vector $p_{x_s} \in \mathbb{R}_{st}^{n_x}$ which is a solution of the* steady state equations,

$$p_{x_s} = A p_{x_s},\ 1_{n_x}^T p_{x_s} = 1. \tag{5.21}$$

Call the stochastic vector $p_x \in \mathbb{R}_{st}^{n_x}$ the invariant state measure *or the* invariant measure *of this stochastic system. Define then $p_y = C p_x \in \mathbb{R}_{st}^{n_y}$. Call then $p_{y_s} = C p_{x_s} \in \mathbb{R}_{st}^{n_y}$ the* invariant output measure.

(b)*Moreover, the invariant measure has strictly positive support on the state set; or, equivalently, for all $i \in \mathbb{Z}_{n_x}$, $p_x(i) > 0$; consequently $p_x \in \mathbb{R}_{s+,st}^{n_x}$, $n_{s+}(p_x) = n_x$, $i(p_x) = \mathbb{Z}_{n_x}$, and $X_{supp}(p_x) = int(\mathbb{R}_{st}^{n_x})$.*

In this case, the invariant output probability measure is strictly positive, in terms of notation, $p_y = C p_x \in \mathbb{R}_{s+}^{n_y}$, if and only if the matrix C has no zero rows; equivalently, if for all $i \in \mathbb{Z}_{n_y}$ there exists a $j \in \mathbb{Z}_{n_x}$ such that $C_{i,j} > 0$.

(c)*Define the sequences of stochastic vectors,*

$$p_x(t+1) = A p_x(t),\ p_x(0) = p_{x,0},\ \ p_y(t) = C p_x(t).$$

If $p_{x,0} = p_{x_s} \in \mathbb{R}_{s+,st}^{n_x}$ then,

$$\forall\, t \in T,\ \ p_x(t) = p_{x_s},\ p_y(t) = C p_{x_s} = p_{y_s} \in \mathbb{R}_{st}^{n_y}.$$

Hence $p_x \in \mathbb{R}_{st}^{n_x}$ is the invariant state probability measure and p_y is the invariant-output probability measure of this output-finite-state-finite stochastic system.

(d)For any $p_0 \in \mathbb{R}_{st}^{n_x}$, if the sequence $p_x(.;0,p_0) : T \to \mathbb{R}_{st}^{n}$ is defined as in Corollary 5.7.10, then convergence to the invariant measure holds,

$$\lim_{t\to\infty} p_x(t;0,p_0) = p_{x,s}, \ \lim_{t\to\infty} p_y(t;0,p_0) = p_{y,s}.$$

The case of a state-finite stochastic system with a fully-reduced state transition matrix is such that there is in general no unique steady-state stochastic vector or steady state probability distribution.

5.7.5 Forward and Backward System Representations

It is proven that for every output-finite-state-finite stochastic system satisfying an assumption there exists both a forward representation and a backward representation. Moreover, the two representations are related.

Theorem 5.7.23. *Consider a output-finite-state-finite stochastic system,*

$$(\Omega, F, P, T, Y, X, y, x),$$

Assume that the condition of state-output conditional independence holds, Def. 5.7.4. Assume that for all times $t \in T$, the stochastic vector of the state satisfies $p_x(t) = E[x(t)] \in \mathbb{R}_{s+,st}^{n_x}$ meaning that every component of this vector is strictly positive.

(a)There exists a forward representation *of this stochastic system of the form,*

$$E\left[\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix} \mid F_t^x \vee F_{t-1}^y\right] = \begin{pmatrix} A_f(t) \\ C_f(t) \end{pmatrix} x(t), \ \forall\, t \in T,$$
$$A_f(t) = E[x(t+1)x(t)^T] \ \text{Diag}(p_x(t))^{-1} \in \mathbb{R}_{st}^{n_x \times n_x},$$
$$C_f(t) = E[y(t)x(t)^T] \ \text{Diag}(p_x(t))^{-1} \in \mathbb{R}_{st}^{n_y \times n_x}.$$

(b)There exists a backward representation *of this stochastic system of the form,*

$$E\left[\begin{pmatrix} x(t-1) \\ y(t-) \end{pmatrix} \mid F_t^{x+} \vee F_t^{y+}\right] = \begin{pmatrix} A_b(t) \\ C_b(t) \end{pmatrix} x(t), \ \forall\, t \in T,$$
$$A_b(t) = E[x(t-1)x(t)^T] \ \text{Diag}(p_x(t))^{-1} \in \mathbb{R}_{st}^{n_x \times n_x},$$
$$C_b(t) = E[y(t-1)x(t)^T] \ \text{Diag}(p_x(t))^{-1} \in \mathbb{R}_{st}^{n_y \times n_x}.$$

(c)The system matrices of the forward and backward representation are related by the equations,

$$A_b(t) = \text{Diag}(p_x(t-1)) \, A_f(t-1)^T \ \text{Diag}(p_x(t))^{-1},$$
$$C_b(t) = C_f(t-1) \, A_f(t-1)^T \ \text{Diag}(p_x(t))^{-1};$$
$$A_f(t) = \text{Diag}(p_x(t+1)) \, A_b(t+1)^T \ \text{Diag}(p_x(t))^{-1},$$
$$C_f(t) = C_b(t+1) \, A_b(t+1)^T \ \text{Diag}(p_x(t))^{-1}, \ \forall\, t \in T.$$

(d)Assume that the system is time-invariant, irreducible, and nonperiodic. From Theorem 18.8.7 follows that there exists a unique invariant state stochastic vector $p_{x_s} \in \mathbb{R}^{n_x}_{s+,st}$. Then the relations between the system matrices of the forward and backward representation are,

$$\begin{aligned} A_b &= \text{Diag}(p_{x_s})\, A_f^T \, \text{Diag}(p_{x_s})^{-1}, \ \ C_b = C_f \, A_f^T \, \text{Diag}(p_{x_s})^{-1}, \\ A_f &= \text{Diag}(p_{x_s})\, A_b^T \, \text{Diag}(p_{x_s})^{-1}, \ \ C_f = C_b \, A_b^T \, \text{Diag}(p_{x_s})^{-1}. \end{aligned}$$

Proof. (a & b) Note that from Proposition 2.8.4.(b) follows that for finite-valued random variables,

$$E[y(t)|F^{x(t)}] = E[y(t)x(t)^T] \, \text{Diag}(p_x(t))^{-1}.$$

The results follow then immediately from the definition of a stochastic system according to,

$$E\left[\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix} \mid F_t^x \vee F_{t-1}^y\right] = E\left[\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix} \mid F^{x(t)}\right].$$

and with Theorem 2.8.4. The formulas for the backward representation are derived correspondingly.

(c) The formulas follow directly from the formulas of the system matrices as stated in (a) and in (b) with the observation that,

$$\begin{aligned} C_b(t) &= E[y(t-1)x(t)^T] \, \text{Diag}(p_x(t))^{-1} \\ &= E[y(t-1)(E[x(t)|F_{t-1}^x \vee F_{t-1}^y)^T] \, \text{Diag}(p_x(t))^{-1} \\ &= E[y(t-1)(A_f(t-1)\text{Diag}(p_x(t-1))^{-1}x(t-1))^T] \, \text{Diag}(p_x(t))^{-1} \\ &= C_f(t-1)A_f(t-1)^T\text{Diag}(p_x(t))^{-1}; \\ A_b(t) &= E[x(t-1)x(t)^T] \, \text{Diag}(p_x(t))^{-1} \\ &= E[x(t-1)E[x(t)| \, F_{t-1}^x \vee F_{t-2}^y]] \, \text{Diag}(p_x(t))^{-1} \\ &= E[x(t-1)(A_f(t-1)x(t-1))^T] \, \text{Diag}(p_x(t))^{-1} \\ &= \text{Diag}(p_x(t-1)) \, A_f(t-1)^T \, \text{Diag}(p_x(t))^{-1}. \end{aligned}$$

The expression of $C_f(t)$ in terms of $C_b(t+1)$ follows similarly. □

Theorem 5.7.24. *Consider a time-invariant state-finite stochastic system in the indicator representation,*

$$\begin{aligned} E[x(t+1)| \, F_t^x] &= Ax(t), \ x(0) = x_0, \\ X &= X_e \subset \mathbb{R}_+^{n_x}, \ n_x \in \mathbb{Z}_+. \end{aligned}$$

Assume that the state transition matrix $A \in \mathbb{R}^{n_x \times n_x}_{st}$ is irreducible and nonperiodic. Then there exists a unique steady state vector which is strictly positive, $p_{x_s} \in \mathbb{R}^{n_x}_{s+,st}$. Assume that the initial state has the steady state measure, $p_{x_0} = p_{x_s}$. Then the state process x is stationary.

The stationary state process x is time-reversible if and only if the state transition matrix satisfies the equation,

$$A = \text{Diag}(p_{x_s})A^T\text{Diag}(p_{x_s})^{-1}; \ \ \Leftrightarrow \ p_{x_s,i}A_{i,j} = p_{x_s,j}A_{j,i}, \ \forall \ i, \ j \in \mathbb{R}_{n_x}.$$

The characterization of a time-reversible state-finite Markov process is related to the system matrix A being a doubly-stochastic matrix. This relation requires further investigation.

It is of interest to formulate a theorem on the time reversibility of a stationary output process of an output-finite-state-finite stochastic system.

5.7.6 Stochastic Observability

Stochastic observability was defined in general terms in Def. 4.6.6. In this section stochastic observability is characterized for output-finite-state-polytopic stochastic systems.

Definition 5.7.25. Consider an output-finite-state-polytopic stochastic system in the indicator representation, Def. 5.7.1. Assume that state-output conditional independence holds.

(a) This system is called *stochastically observable on the time interval* $\{t : t+t_1-1\} = \{t,\ t+1,\ \ldots,\ t+t_1-1\}$ if the map,

$$x(t) \mapsto E\left[\left(\begin{array}{l} y(t) \\ y(t+1) \\ \vdots \\ y(t+t_1-1) \end{array}\right) |F^{x(t)}\right],$$

is injective on the support of the measure of $x(t)$.

(b) Assume that the system is time-invariant. Then it is called *stochastic observable* if there exist $t,\ t_1 \in T$ such that it is stochastically observable on the interval $\{t : t+t_1-1\}$ as defined in (a). Because the system is time-invariant this then holds for all t.

(c) This system is called *stochastically co-observable on the time interval* $\{0 : -t_1+1\} = \{0,\ -1,\ \ldots,\ -t_1+1\}$ with $t_1 \in \mathbb{Z}_+$, where one counts backward on the subset T, if the map,

$$x(t) \mapsto E\left[\left(\begin{array}{l} y(0) \\ y(-1) \\ \vdots \\ y(-t_1+1) \end{array}\right) |F^{x(t)}\right],$$

is injective on the support of the measure of $x(t)$.

(d) Assume that the system is time-invariant. Then it is called *stochastic observable* if there exist $t,\ t_1 \in T$ such that it is stochastically observable on the interval $\{0 : -t_1+1\}$ as defined in (c). Because the system is time-invariant this then holds for all t.

The characterization of stochastic observability of a finite system is first treated not for a finite stochastic system but for finite-valued random variables.

Proposition 5.7.26. Stochastic observability of a tuple of finite-valued random variables. *Consider a finite-valued state vector x, a finite-valued output vector y both in the indicator representation, and their relation according to,*

$$\begin{aligned} & x : \Omega \rightarrow \mathbb{R}^{n_x},\ y : \Omega \rightarrow \mathbb{R}^{n_y},\ C \in \mathbb{R}_{st}^{n_y \times n_x}, \\ & n_x,\ n_y \in \mathbb{Z}_+,\ p_x \in \mathbb{R}_{st}^{n_x},\ n_{s+}(p_x) = n_x (\Leftrightarrow\ \forall\, i \in \mathbb{Z}_{n_x},\ p_{x,i} > 0), \\ E[y|F^x] &= Cx. \end{aligned}$$

(a)Assume that the state set equals $X = \mathbb{R}_{st}^{n_x}$. Then the relation is stochastically observable if and only if $\text{rank}(C) = n_x$.

(b)Assume that the state set X_p equals a polytope generated by the nonsquare stochastic matrix $S \in \mathbb{R}_{st}^{n_x \times n_p}$ with $n_p \in \mathbb{Z}_+$, $\text{Polytope}(S) = X_p \subset X = \mathbb{R}_{st}^{n_x}$. *Assume that the column vectors of S are positively independent. Thus only states in the subset* $\text{Polytope}(S)$ *are considered.*
Then the relation $E[y|\ F^x] = CSx$ is stochastically observable if and only if $\text{rank}(CS) = n_p$*, where $CS \in \mathbb{R}_{st}^{n_y \times n_p}$.*

Proof. (a) ($\Leftarrow$) If the map is not stochastically observable then there exist two vectors $p_1,\ p_2 \in \mathbb{R}_{st}^{n_x}$ such that $p_1 \neq p_2$ and $Cp_1 = Cp_2$. Then $C(p_1 - p_2) = 0$ and $p_1 \neq p_2$, hence $p_1 - p_2 \neq 0$, imply that $\text{rank}(C) < n_x$ which is a contradiction.
($\Rightarrow$) By contradiction, suppose that $\text{rank}(C) < n_x$. Then there exists a vector $x \in \mathbb{R}^{n_x}$ such that $x \neq 0$ and $Cx = 0$. Define $x_+ = \max\{x,\ 0\}$ and $x_- = -\max\{-x,\ 0\}$. Then $x_+,\ x_- \in \mathbb{R}_{s+}^{n_x}$, $x = x_+ - x_-$, and $0 = Cx$ implies that $Cx_+ = Cx_-$. Note that $0 = 1^T Cx = 1^T C(x_+ - x_-) = 1^T x_+ - 1^T x_-$, and $1^T x_+ \neq 0$ and $1^T x_- \neq 0$. Define then the vectors $z_+ = x_+/[1^T x_+]$ and $z_- = x_-/[1^T x_-]$. By definition of these vectors, $1^T z_+ = 1$ and $1^T z_- = 1$ hence $z_+,\ z_- \in \mathbb{R}_{st}^{n_x}$. Then $Cz_+ - Cz_- = [Cx_+ - Cx_-]/[1^T x_+] = 0$ and $z_+ - z_- = [x_+ - x_-]/[1^T x_+] = x/[1^T x] \neq 0$. Thus $Cz_+ = Cz_-$, $z_+,\ z_- \in \mathbb{R}_{st}^{n_x}$, and $z_+ \neq z_-$ imply that the map is not stochastically observable.
(b) This follows from (a) if one replaces the matrix C by the matrix CS. Note that then the state set is the polytope X_p which is, by the transformation S, transformed to $\mathbb{R}_{st}^{n_p}$. □

Theorem 5.7.27. *Consider a time-invariant output-finite-state-polytopic stochastic system in the indicator representation. Assume that state-output conditional independence holds.*

$$\begin{aligned} E\left[\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix} |F^{x(t)} \right] &= \begin{pmatrix} A \\ C \end{pmatrix} x(t),\ A \in \mathbb{R}_{st}^{n_x \times n_x},\ C \in \mathbb{R}_{st}^{n_y \times n_x}; \\ X_p = X(A) = \mathbb{R}_{st}^{n_x} \cap \text{cone}(S),&\ S \in \mathbb{R}_{st}^{n_x \times n_p}. \end{aligned}$$

(a)Assume that the state set is the full set $X_p = \mathbb{R}_{st}^{n_x \times n_x}$. The output-finite-state-polytopic stochastic system is stochastically observable on the interval $\{t : t + t_1\} = \{t, t+1, \ldots, t+t_1\}$ if and only if the following condition holds,

$F^{x(t)} = F^{O_s x(t)}$, *where,*

$$\text{obsmat}(A, C, t_1) = \begin{pmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{t_1 - 1} \end{pmatrix} \in \mathbb{R}_+^{t_1 n_y \times n_x},$$

if and only if $n_x = \text{rank}(\text{obsmat}(A, C, t_1)$.

(b)Assume that the stochastic system is such that the state equals the polytope $X_p \subset \mathbb{R}_{st}^{n_x}$*; equivalently, that* $x_0 \in X_p$ *and, for all* $t \in T$, $x(t) \in X_p$ *implies that* $x(t+1) \in X_p$*. Represent the state polytope by the stochastic matrix* $S \in \mathbb{R}_{st}^{n_x \times n_p}$ *according to* $X_p = \text{Polytope}(S)$ *and recall that* $n_r \in \mathbb{Z}_+$ *denotes the positive-recursion number, Def. 5.7.12.*

Then the system is stochastically observable if and only if,

$$n_p = \text{rank}(\text{obsmat}(A, C, n_r)),$$

$$\text{obsmat}(A, C, n_r) = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n_r - 1} \end{pmatrix} S \in \mathbb{R}^{n_y n_r \times n_p}.$$

Note that only n_r *block rows are needed in the observability matrix due to the assumed existence of a minimal positive recursion and the representation of the state polytope* $X(A) = \text{Polytope}(S)$.

Proof. (a) This follows from Proposition 5.7.26.(a).
(b) This follows from Proposition 5.7.26.(b). Note that the set of state probability measures is the polytope $X_p = X(A) \subseteq \mathbb{R}_{st}^{n_x}$. But this polytope is described as generated by the columns of the matrix S. Then the representation follows. No higher power than $n_r - 1$ is needed due to the fact that n_r is the smallest integer for which a positive recursion exists. Also $X(A)$ is represented by the matrix S, $X(A) = \text{Polytope}(S)$. □

A result like Theorem 5.7.27 holds for a positive linear system.

5.8 Sigma-Algebraic Stochastic System

Definition 5.8.1. A *σ-algebraic stochastic system*.
The σ-algebra families $\{F_t,\ G_t \subseteq F,\ \forall\ t \in T = \mathbb{N}\}$ are said to form a *σ-algebraic system* if,

$$\forall\ t \in T,\ (F_t^+ \vee G_t^+,\ F_{t-1}^- \vee G_t^- \mid G_t) \in \text{CI};$$

where $F_t^+ = \vee_{s \geq t} F_s$, $F_{t-1}^- = \vee_{s \leq t-1} F_s$, $G_t^+ = \vee_{s \geq t} G_s$, $G_t^- = \vee_{s \leq t} G_s$.

In a σ-algebraic stochastic system, the main objects are σ-algebras consistent with the geometric approach to systems. The reader may think of the σ-algebra F_t as representing the space of the output of a system at time $t \in T$ and of G_t as representing the space of the state of a σ-algebra system at the same time. The definition imposes the restriction that, at any time $t \in T$, the current state σ-algebra G_t makes the combined future of the output process and of the state process conditionally independent from the combined past of the output process and of the state process. The above definition is consistent with the general definition of a stochastic systems, see Def. 4.2.2.

Note that $\{F_t,\ t \in T\}$ is not a filtration because in general $F_t \not\subseteq F_{t+1}$, for $t \in T$. However, $\{F_t^-,\ \forall\, t \in T\}$ is a filtration because, by definition of F_t^-, $F_t^- \subseteq F_{t+1}^-$.

A σ-algebraic stochastic system is used in Section 7.4 for strong stochastic realization of a σ-algebra family.

5.9 The Multiple Conditional-Independence Relation

The reader finds in this section the concept of the multiple conditional independence relation and several of its properties.

The motivation for this concept is the use of databases in computer science, data science, and mathematics. A major problem of the area of databases is to determine relations between variables. Commerical companies can then exploit those relations. The database entries can be regarded as deterministic variables or as random variables. For the description of relations between deterministic variables the concept of a normal form has been introduced. In case the model is that database entries are random variables then useful relations between these variables are the conditional independence relations as described in this section.

The reader should recall that a stochastic system, without input, is defined by a the tuple of a state and output process such that the current state makes conditionally independent the past and the future of the output and the state processes. The conditional independence of the state applies only to the tuple of the past and the future processes.

Next a generalization of the concept of state is introduced. The state of a multiple conditional independence relation makes conditionally independent not two but $n \in \mathbb{Z}_+$ σ-algebras. This concept is useful for modeling relations of databases. More useful is the set of state σ-algebras in which each state makes $k \in \mathbb{Z}_n$ a subset of the database σ-algebras conditionally independent. There then appears a combinatorial structure of state σ-algebras. Of interest to database research is: Which of these possible state σ-algebras is most useful in answering a set of questions for the database? It is clear that statistics and analysis can answer this question but the theoretical framework is best based on the concept of the multiple conditional independence relation.

The concepts and results of this section are a direct extension of the concept of a stochastic system.

Definition 5.9.1. *The multiple conditional independence relation*. Consider a probability space (Ω, F, P) and a family of sub-σ-algebras $F_1,\ F_2,\ \ldots,\ F_n,\ G \subseteq F$ for an integer $n \in \mathbb{Z}_+$, $n \geq 3$. One says that the σ-algebras $F_1, \ldots, F_n$ are *multiply conditionally independent* conditioned on or given the state σ-algebra G, if,

$$\begin{aligned} & E[x_1 x_2 \ldots x_n | G] = E[x_1|G]E[x_2|G] \ldots E[x_n|G], \\ & \forall\ i \in \mathbb{Z}_n,\ \forall\ x_i \in L_+((\Omega, F_i, P), \mathbb{R}_+). \end{aligned}$$

One says also that the *state σ-algebra* σ-algebra G makes the family $F_1, \ldots, F_n$ *multiply conditionally independent*. The term for the relation is the *multiple conditional independence relation*. Denote this form of multiple conditional independence by the notation,

$$(F_1, F_2, \ldots, F_n|\ G) \in \text{CI}.$$

From examples one learns that one needs a concept of a set of state σ-algebras each of which makes a set of the variable σ-algebras conditionally independent. Of interest then is the set of relations between the state σ-algebras. For an application, one can partition the set of attributes of a human being into: (1) her or his physical properties like length, weight, hair color, eye color, etc.; (2) marital state if any; (3) her or his educational level; (4) her or his work experience; etc. Another set of attributes can be formulated for companies. or for government organizations. From the data of the database it can then be determined which sets of attributes are conditionally independent. Once the conditional independence is determined, the search engine can carry out its task more efficiently than without this knowledge.

Definition 5.9.2. Consider the setting of Def. 5.9.1. The notation $|I| \in \mathbb{N}$ for a finite set $I \subseteq \mathbb{N}$, denotes the number of distinct elements belonging to I.

Define the *family of state σ-algebras* of the multiple conditional independence relation if the following conditions all hold,

$$\begin{aligned} G_s = & \left\{ G(I) \subseteq F |\ \forall\ I \subset \mathbb{Z}_n \backslash \{1\} \right\}, \\ & \text{where } I = \{i_1,\ i_2, \ldots, i_{n_k}\} \subseteq \mathbb{Z} \backslash \{1\},\ |I| = k \geq 2, \\ & (F_{i_1},\ F_{i_2},\ \ldots,\ F_{i_{n_k}} | G(I)) \in \text{CI}. \end{aligned}$$

Example 5.9.3. The special case of multiple conditional independence for $n = 3$ is specified in detail below. The notation reads that,

$$\begin{aligned} & n = 3,\ k = 2,\ k = 3; \\ & k = 2:\ \ I = \{1,2\},\ \{1,3\},\ \{2,3\} \subseteq \mathbb{Z}_3;\ \ k = 3,\ I = \{1,2,3\} = \mathbb{Z}_3; \\ & G(I) = G(\{1,2\}) = G(1,2),\ \text{note abuse of notation}; \\ & k = 2,\ (F_1,\ F_2|G(2,1)) \in \text{CI},\ (F_1,\ F_3|G(2,2)) \in \text{CI}, \\ & \quad (F_2,\ F_3|G(2,3)) \in \text{CI}, \\ & k = 3,\ (F_1, F_2, F_3|G(1,2,3)) \in \text{CI}; \\ G_s = & \{G(1,2),\ G(1,3),\ G(2,3),\ G(1,2,3) \in F\}. \end{aligned}$$

Research problems for the multiple conditional independence relation are:

1. What is the relation between the elements of the set of state σ-algebras?
2. How to construct and to classify the set of minimal state σ-algebras?
3. How to construct subsets of the set of variables for which multiple conditional independence holds?

Proposition 5.9.4. *Consider the multiple conditional independence relation of Def. 5.9.1. If* $(F_1, F_2, \ldots, F_n|\ G) \in \mathrm{CI}$ *then* $F_1 \cap F_2 \cap \ldots \cap F_n \subseteq G$.

Proof. Consider a subset $A \in (F_1 \cap F_2 \cap \ldots \cap F_n)$. Then,

$$\begin{aligned} x &= E[I_A|G] = E[I_A I_A \ldots I_A|G] = \prod_{i=1}^{n} E[I_A|G], \text{ by conditional independence,} \\ &= x^n \;\Rightarrow\; x(x^{n-1}-1) = 0, \text{ because } x = E[I_A|G] \geq 0, \\ &\quad \text{either } x(\omega) = 0 \text{ or } x(\omega) = 1 \ a.s., \\ &\Rightarrow x = E[I_A|G] = I_A \;\Rightarrow\; A \in G; \text{ or } x = 0 = \emptyset \ a.s.\ \emptyset \in G. \end{aligned}$$

□

A special case of the above proposition is that,

$$(F_1, F_2, \ldots, F_n|\ G) \in \mathrm{CI},\ F_n \subseteq F_{n-1} \subseteq \ldots F_2 \subseteq F_1 \quad \Rightarrow \quad F_n \subseteq G.$$

5.10 Technicalities

A more detailed discussion follows of the concept of a stochastic dynamic system without inputs. Recall Def. 4.2.2 of a stochastic system.

Condition (4.3) of the definition of a stochastic system is asymmetric with respect to the output process. This is a convention. A priori there are four possible conditions for a stochastic dynamic system which are listed below:

$$(F_t^{y+} \vee F_t^{x+}, F_t^{y-} \vee F_t^{x-}|F^{x(t)}) \in CI,\ \ \forall\, t \in T; \tag{5.22}$$

$$(F_{t+1}^{y+} \vee F_t^{x+}, F_{t-1}^{y-} \vee F_t^{x-}|F^{x(t)}) \in CI,\ \ \forall\, t \in T; \tag{5.23}$$

$$(F_t^{y+} \vee F_t^{x+}, F_{t-1}^{y-} \vee F_t^{x-}|F^{x(t)}) \in CI,\ \ \forall\, t \in T; \tag{5.24}$$

$$(F_{t+1}^{y+} \vee F_t^{x+}, F_t^{y-} \vee F_t^{x-}|F^{x(t)}) \in CI,\ \ \forall\, t \in T. \tag{5.25}$$

Condition (5.22) and a property of conditional expectation, see Proposition 2.9.4.(a), imply that

$$F^{y(t)} \subseteq (F_t^{y+} \cap F_t^{y-}) \subseteq F^{x(t)},$$

which fact is not compatible with the intuitive concept of state because the output is in general not part of the state. Condition (5.23) is not suitable because it would allow examples that are counter-intuitive to the concept of state, see Example 5.10.5. It seems logical to require that the past and the future of the σ-algebras of the output process add up to F_∞^y as in $F_{t+1}^{y+} \vee F_t^{y-} = F_\infty^y$. This condition excludes the case (5.23).

The conditions (5.24) and (5.25) thus remain, of which Condition 3 has been chosen. This is a convention. Condition (5.25) results in the following representation of a Gaussian stochastic system,

$$x(t+1) = Ax(t) + Mv(t), \tag{5.26}$$

$$y(t+1) = Cx(t) + Nv(t), \tag{5.27}$$

or that the system is specified by the map

$$x(t) \mapsto \text{cpdf}(x(t+1), y(t+1)|F_t^{x-} \vee F_t^{y-}),$$

which form is inconsistent with the system theoretic convention of (4.4) & (4.5). The system representation (5.26,5.27) is occasionally used in the literature. The results are comparable to the Gaussian system representation used in these notes though the formulas are slightly different.

The definition of a stochastic system is formulated in terms of σ-algebras rather than in terms of stochastic processes. This is a geometric formulation in which emphasis is put on spaces and subspaces rather than on the variables or processes that generate those spaces.

Proposition 5.10.1. *Consider a collection*

$$\{\Omega, F, P, T, Y, B_Y, X, B_X, y, x\}$$

as defined in Definition 4.2.2 but without condition (4.3). The following statements are equivalent:

(a)for all $t \in T$, $(F_t^{y+} \vee F_t^{x+}, F_{t-1}^{y-} \vee F_t^{x-}|F^{x(t)}) \in CI$;
(b)for all $t \in T$, $(F^{y(t)} \vee F^{x(t+1)}, F_{t-1}^{y-} \vee F_t^{x-}|F^{x(t)}) \in CI$;
(c)for all $t \in T$, $(F_t^{y+} \vee F_t^{x+}, F^{y(t-1)} \vee F^{x(t-1)}|F^{x(t)}) \in CI$.

Proof. (a) ⇒ (b). This is immediate by restriction.
(b) ⇒ (a). Let $t \in T$, $n \in \mathbb{Z}_+$ and, for $k \in \mathbb{Z}_n$, $z_{t+k} \in L_b(F^{x(t+k)} \vee F^{y(t+k-1)})$ is a bounded random variable. Then,

$$\begin{aligned}
&E[z_{t+1} \ldots z_{t+n}|F_t^{x-} \vee F_{t-1}^{y-}] \\
&= E[z_{t+1} \ldots z_{t+n-1} E[z_{t+n}|F_{t+n-1}^{x-} \vee F_{t+n-2}^{y-}]|F_t^{x-} \vee F_{t-1}^{y-}] \\
&= E[z_{t+1} \ldots z_{t+n-1} E[z_{t+n}|F^{x(t+n-1)}]|F_t^{x-} \vee F_{t-1}^{y-}], \text{ by assumption (b),} \\
&= E[z_{t+1} \ldots z_{t+n-2} \tilde{z}_{t+n-1}|F_t^{x-} \vee F_{t-1}^{y-}], \text{ where,}
\end{aligned}$$

$$\tilde{z}_{t+n-1} = z_{t+n-1} E[z_{t+n}|F^{x(t+n-1)}] \in L_b(F^{x(t+n-1)} \vee F^{y(t+n-2)}), \text{ thus,}$$

$$\begin{aligned}
&E[z_{t+1} ... z_{t+n}|F_t^{x-} \vee F_{t-1}^{y-}] \\
&= E[z_{t+1} \ldots z_{t+n-2} \tilde{z}_{t+n-1}|F_t^{x-} \vee F_{t-1}^{y-}] = \ldots = E[\tilde{z}_{t+1}|F_t^{x-} \vee F_{t-1}^{y-}] \\
&= E[\tilde{z}_{t+1}|F^{x(t)}] = E[z_{t+1} ... z_{t+n}|F^{x(t)}].
\end{aligned}$$

where the latter equality follows from a property of conditional independence. This and a monotone class argument establishes that

$$(\vee_{k=1}^{n} (F^{y(t+k-1)} \vee F^{x(t+k)}), F_t^{x-} \vee F_{t-1}^{y-} | F^{x(t)}) \in CI.$$

Yet another monotone class argument yields

$$(F_t^{y+} \vee F_{t+1}^{x+}, F_t^{x-} \vee F_{t-1}^{y-} | F^{x(t)}) \in CI.$$

From Proposition 2.9.2 then follows that

$$(F_t^{y+} \vee F_t^{x+}, F_{t-1}^{y-} \vee F_t^{x-} | F^{x(t)}) \in CI.$$

(a) ⇔ (c) This follows along the lines of (a) ⇔ (b). □

The following result is a useful way to specify a stochastic dynamic system.

Definition 5.10.2. A *forward difference representation of a stochastic dynamic system* in terms of a state and output process is a collection,

$$\{\Omega, F, P, T, Y, B_Y, X, B_X, y, x, p_{x_0}, f\}, \tag{5.28}$$

where $\Omega, F, P, T, Y, B_y, X, B_X, y, x$ are as defined in Definition 4.2.2 and

- $T = \mathbb{N}$ or $T = \{0, 1, \ldots, t_1\} = T(0 : t_1)$,
- p_{x_0} is a probability measure on (X, B_X) which represents the probability distribution of the initial state $x_0 = x(t_0)$;
- the probabilistic forward transition function is specified by f according to

$$\begin{aligned} & f : T \times X \to \mathbf{P}(X \times Y), \\ & P(\{(x(t+1), y(t)) \in A | F_t^{x-} \vee F_{t-1}^{y-}\}) = f(A, t, x(t)),\ \forall t \in T \backslash \{t_1\}. \end{aligned} \tag{5.29}$$

Note that Equation (5.29) implies that

$$\text{cpdf}((x(t+1), y(t)) | F_t^{x-} \vee F_{t-1}^{y-}) = \text{cpdf}((x(t+1), y(t)) | F^{x(t)}),\ \forall t \in T,$$

or, equivalently, that,

$$(F^{x(t+1)} \vee F^{y(t)}, F_t^{x-} \vee F_{t-1}^{y-} | F^{x(t)}) \in CI,\ \forall t \in T.$$

Proposition 5.10.3. *A forward difference representation of a stochastic dynamic system,*

$$\{\Omega, F, P, T, Y, B_Y, X, B_X, y, x, p_{x_0}, f\},$$

is such that the collection $\{\Omega, F, P, T, Y, B_Y, X, B_X, y, x, \}$, *is a stochastic dynamic system.*

Proof. The result follows from Proposition 5.10.1 (a) ⇔ (b). □

The following result is a useful sufficient condition for a stochastic dynamic system.

Proposition 5.10.4. *Consider the collection*

$$\{\Omega, F, P, T, Y, B_Y, X, B_X, y, x\}$$

as defined in Definition 4.2.2 but without condition (4.3). If for all $t \in T$,

(1) $(F_t^{y+}, F_\infty^{x-} \vee F_{t-1}^{y-} | F^{x(t)}) \in CI \; \forall t \in T$; *and,*
(2) $(F_t^{x+}, F_t^{x-} \vee F_{t-1}^{y-} | F^{x(t)}) \in CI \; \forall t \in T$;

then this collection is a stochastic dynamic system.

Proof. Let $t \in T$, $A_1 \in F_t^{x+}$, $A_2 \in F_t^{y+}$. Then

$$\begin{aligned} & E[I_{A_1} I_{A_2} | F_t^{x-} \vee F_{t-1}^{y-}] \\ & = E[I_{A_1} E[I_{A_2} | F_\infty^{x-} \vee F_{t-1}^{y-}] | F_t^{x-} \vee F_{t-1}^{y-}] \text{ because } A_1 \in F_t^{x+} \subseteq F_\infty^{x-}, \\ & = E[I_{A_1} | F_t^{x-} \vee F_{t-1}^{y-}] E[I_{A_2} | F^{x(t)}], \text{ by } (1), \\ & = E[I_{A_1} | F^{x(t)}] E[I_{A_2} | F^{x(t)}], \text{ by } (2), \end{aligned}$$

is $F^{x(t)}$ measurable and, as in the proof of Proposition 5.10.1, an application of the monotone class theorem yields the result. □

Below an example of a stochastic dynamic system is presented.

Example 5.10.5. *Gaussian system with state noise.* Let $v : \Omega \times T \to \mathbb{R}$ be a standard Gaussian white noise process. Define $y : \Omega \times T \to \mathbb{R}$, $x : \Omega \times T \to \mathbb{R}$ by,

$$\begin{aligned} & x(t) = v(t-1), \\ & y(t) = x(t) + v(t) = v(t-1) + v(t). \end{aligned}$$

Then the following hold.

(a) For all $t \in T$, $(F_{t+1}^{y+}, F_{t-1}^{y-} | G_0) \in CI$, where $G_0 \subset F$ is the trivial σ-algebra. Thus the process y is the output process of a stochastic dynamic system according to the condition (5.23) with a trivial state space.
(b) For all $t \in T$, $E[\exp(iuy(t)) | F_{t-1}^{y+}]$ is nondeterministic, indicating that the process y has a kind of memory.
(c) $(F_t^{y+} \vee F_t^{x+}, F_{t-1}^{y-} \vee F_t^{x-} | F^{x(t)}) \in$ CI for all $t \in T$, hence $\{\Omega, F, P, T, Y, B, X, B, y, x\} \in$ GStocS.

Proof. (a) For $t \in T$ and for all $u \in \mathbb{R}$,

$$\begin{aligned} & E[\exp(iuy(t+1)) | F_{t-1}^{y-}] \\ & = E[E[\exp(iuv(t) + iuv(t+1)) | F_{t-1}^{v-}] | F_{t-1}^{y-}] = E[\exp(iuv(t) + iuv(t+1))] \end{aligned}$$

because v is a Gaussian white noise process. Hence $(F^{y_{t+1}}, F_{t-1}^{y-} | G_0) \in CI$ and the result follows from Proposition 5.10.1.(b).
(b) Apply Theorem 8.3.2 for the Kalman filter to obtain,

$$\begin{aligned}
x(t+1) &= v(t), \;\; y(t) = x(t) + v(t), \\
E[\exp(iuy(t))|F_{t-1}^{y-}] &= \exp(iuE[y(t)|F_{t-1}^{y-}] - \frac{1}{2}u^2\bar{q}), \\
\hat{x}(t+1|t) &= k(y(t) - \hat{x}(t|t-1)), \; k = (q+1)^{-1}, \; q = 1-(q+1)^{-1},
\end{aligned}$$

from which follows that $q = 0$ and $k = 1$,

$$\begin{aligned}
\hat{y}(t+1|t) &= E[y(t+1)|F_t^{y-}] = \hat{x}(t+1|t) = y(t) - \hat{x}(t|t-1) = y(t) - \hat{y}(t|t-1), \\
\bar{q} &= E[(y(t) - E[y(t)|F_{t-1}^{y-}])^2] = q + 1 = 1, \\
E[\exp(iuy(t))|F_{t-1}^{y-}] &= \exp(iu\hat{y}(t|t-1) - \frac{1}{2}u^2).
\end{aligned}$$

(c) With $x(t+1) = v(t)$, $y(t) = x(t) + v(t)$, and v a Gaussian white noise process, the result follows from Example 4.2. □

5.11 Further Reading

Examples of stochastic systems. Examples of nonlinear stochastic systems are provided in the following references. Electrical noise in nonlinear RLC networks [2, 39]. A model for freeway traffic flow [36, 37]. Modeling of stochastic systems for physical structures is presented in the book [27].

Output-finite-state-polytopic stochastic systems. A early book on finite stochastic systems is [26] which presents stochastic realization theory of these systems. A more recent book is [21]. For system identification of finite stochastic systems see [6]. An entry into the use of finite stochastic systems in communication is [33].

The state process of a finite stochastic system is also called a Markov chain for which references include [19]. The asymptotic analysis of Markov chains is described in [8, 22].

The characterization of a stochastic matrix which leaves a polytope invariant is due to Y. Zeinaly, the author, and B. De Schutter, [44, Thm. 4.5 and 5.2]. Examples 5.7.13 and 5.7.14 are due to [44, Ex. 4.8, Ex. 4.7, p. 293].

Stochastic observability of an output-finite-state-polytopic stochastic system
Proposition 5.7.26 is a generalization of a result of [15].

Social networks. The modeling of belief in networks leads to problems of analysis of Markov chains. Tutorial papers on this research issue are [29, 30] and related books include [16, 11].

Stochastic systems with outputs in the natural numbers. Most books on communication networks present stochastic systems with as output a counting process. Examples are [18, 42].

Stochastic systems with outputs in a finite interval of the real numbers. Systems with beta probability distributions were considered in [3, 10, 25].

Stochastic systems with outputs in the positive real numbers. Positive stochastic systems are discussed in [31]. Deterministic positive systems are treated in [5, 13].

Stochastic systems with outputs in the real numbers. References on bilinear Gaussian stochastic systems are [1, 12, 14, 32, 40].

Random fields. This topic is not treated in this book. Therefore a few references follow. Random fields are treated in [9, 35, 34, 41, 43]. References on homogeneous random fields and spatial interpolation are [20], the studies of G. Matheron [23, 24], the book [17], and for applications see [4, 7, 38].

References

1. S.I. Akamanam, M. Bhaskara Rao, and K. Subramanyam. On the ergodicity of bilinear time series models. *J. Time Series Anal.*, 7:157–163, 1986. 170
2. B.D.O. Anderson. Nonlinear networks and Onsager-Casimir reversibility. *IEEE Transactions on Circuits and Systems*, 27:1051–1058, 1980. 169
3. A. Azzilini. A Markov process with beta marginal distribution. *Statistica*, 44:241–243, 1984. 169
4. G. Bastin and M. Gevers. Identification and optimal estimation of random fields from scattered point-wise data. *Automatica*, 21:139–155, 1985. 78, 170
5. A. Berman, M. Neumann, and R.J. Stern. *Nonnegative matrices in dynamic systems*. John Wiley & Sons, New York, 1989. 169, 697
6. O. Cappé, E. Moulines, and T. Rydén. *Inference in hidden Markov models*. Springer, Berlin, 2005. 169, 277, 353
7. J.P. Delhomme. Kriging in the hydrosciences. *Adv. Water Resources*, 1:251–266, 1978. 78, 170
8. P. Diaconis and D. Stroock. Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Probab.*, 1:36–61, 1991. 169, 698
9. M. Dozzi. *Stochastic processes with a multidimensional parameter*. Longman Scientific & Technical, Harlow, 1989. 170
10. S.L.P. Ferrari and F. Cribari-Neto. Beta regression for modeling rates and proportions. *J. Applied Statist.*, 31:799–815, 2004. 169
11. N.E. Friedkin and E.C. Johnson. *Social influence network theory: A social examination of small group dynamics*. Cambridge University Press, Cambridge, 2011. 169
12. C.W.J. Granger and A.P. Andersen. *An introduction to bilinear time series models*. Vandehoeck & Ruprecht, Göttingen, 1978. 170
13. Wassim M. Haddad, VijaySekhar Chellaboina, and Qing Hui. *Nonnegative and Compartmental Dynamical Systems*. Princeton University Press, Princeton, 2010. 169, 697
14. E.J. Hannan. A note on bilinear time series models. *Stoc. Processes Appl.*, 12:221–224, 1982. 170
15. Hanna M. Härdin and Jan H. van Schuppen. Observers for linear positive systems. *Linear Algebra & its Applications*, 425:571–607, 2007. 169
16. M.O. Jackson. *Social and economic networks*. Princeton University Press, Princton, 2010. 169
17. A. Journel and C. Huijbregts. *Kriging geostatistics*. Academic Press, New York, 1978. 78, 170
18. F.P. Kelly. *Reversibility and stochastic networks*. John Wiley & Sons, Chichester, 1979. 73, 169, 468
19. J.G. Kemeny and J.L. Snell. *Finite Markov chains*. Springer, New York, 1983. 169
20. D.G. Krige. Two dimensional weighted moving average trend surfaces for ore evaluation. *Journal of the South African Institute of Mining and*, pages 13–38, 1966. 78, 170
21. V. Krishnamurthy. *Partially observed Markov decision processes*. Cambridge University Press, Cambridge, 2016. 169, 277, 353, 575

22. D.A. Levin, Y. Peres, and E.L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, Providence, 2009. 169
23. G. Matheron. *Les variables regionalisees et leur estimation*. Masson, Paris, 1965. 170
24. G. Matheron. The intrinsic random functions and their applications. *Adv. Appl. Probab.*, 5:439–468, 1973. 78, 170
25. E. McKenzie. An autoregressive process for beta random variables. *Manag. Sci.*, 31:988–997, 1985. 169
26. A. Paz. *Introduction to probabilistic automata*. Academic Press, New York, 1971. 169, 277
27. Y. Peng and J. Li. *Stochastic optimal control of structures*. Springer, Berlin, 2019. 169
28. G. Picci. On the internal structure of finite-state stochastic processes. In *Proc. of a U.S.-Italy Seminar*, volume 162 of *Lecture Notes in Economics and Mathematical Systems*, pages 288–304. Springer-Verlag, Berlin, 1978. 120, 150, 277
29. A.V. Proskurnikov and R. Tempo. A tutorial on modeling and analysis of dynamic social networks – part i. *Annual Review Control*, 43:65–79, 2017. 169
30. A.V. Proskurnikov and R. Tempo. A tutorial on modeling and analysis of dynamic social networks – part ii. *Annual Review Control*, 44:0–0, 2018. 169
31. P. Purdue. Stochastic compartmental models: A review of the mathematical theory with ecological applications. In J.H. Matis, B.C. Patten, and G.C. White, editors, *Compartmental analysis of ecosystem models*, pages 223–260. International Co-operative Publishing House, Fairland, MD, 1979. 169
32. B.G. Quinn. Stationarity and invertibility of simple bilinear models. *Stoc. Processes Appl.*, 12:225–230, 1982. 170
33. L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77:257–286, 1989. 169, 353
34. M. Rosenblatt. *Stationary sequences and random fields*. Birkäuser Verlag, Basel, 1985. 170
35. Yu.A. Rozanov. *Markov random fields*. Springer-Verlag, New York, 1982. 170
36. S.A. Smulders. Control of freeway traffic flow. Report OS-R8817, Centrum voor Wiskunde en Informatica, Amsterdam, 1988. 9, 78, 169
37. S.A. Smulders. *Control of freeway traffic flow*. PhD thesis, University of Twente, Enschede, 1989. 9, 169
38. A. Stein, W. van Dooremolen, J. Bouma, and A.K. Bregt. Cokriging point data on moisture deficit. *Soil science Society of America Journal*, 52:1418–1423, 1988. 170
39. H.-N. Tan and J.L. Wyatt Jr. Thermodynamics of electrical noise in a class of nonlinear rlc networks. *IEEE Trans. Circuits & Systems*, 32:540–558, 1985. 169
40. H. Tong. A note on a Markov bilinear stochastic process in discrete time. *J. Time Series Anal.*, 2:279–284, 1981. 170
41. E. Vanmarcke. *Random fields: analysis and synthesis*. The MIT Press, Cambridge, MA, 1983. 170
42. J. Walrand and P.P. Varaiya. *High-performance communication networks (2nd ed.)*. Morgan Kaufmann, San Francisco, 2000. 169, 605
43. J.W. Woods. Two-dimensional discrete Markovian fields. *IEEE Trans. Information Theory*, 18:232–240, 1972. 170
44. Yashar Zeinaly, Jan H. van Schuppen, and Bart de Schutter. Linear positive systems may have a reachable subset from the origin that is either polyhedral or nonpolyhedral. *SIAM J. Math. Anal.*, 41:279–307, 2020. 169

Chapter 6
Stochastic Realization of Gaussian Systems

Abstract The weak stochastic realization problem is to determine all stochastic systems whose output equals a considered output process in terms of its finite-dimensional distributions. Such a system is then said to be a *stochastic realization* of the considered output process. The problem encompasses: (1) an equivalent condition for the existence of a realization, (2) characterizing when such a system is a minimal stochastic realization, and (3) classifying all minimal stochastic realizations. The concepts of stochastic realization theory are the basis of filter theory and of control theory. In this chapter the stochastic realization is presented for stationary Gaussian stochastic processes. Particular concepts discussed include: stochastic observability, stochastic co-observability, covariance realization, minimality of a stochastic realization, a classification map, and a canonical form for a stochastic system.

Key words: Stochastic realization. Gaussian systems.

The reader who first learns about the stochastic realization problem is advised to concentrate attention on the Sections 6.2 through 6.5. The study of this chapter may be complemented with the reading of the appendices 23 and 24 but this can be postponed till a study of the proof of Theorem 6.4.3 presented in Section 6.6.

The problem formulation of stochastic realization is due to R.E. Kalman. The theory of weak Gaussian stochastic realization is primarily due to P. Faurre and his co-workers. The theory of strong Gaussian stochastic realization is due to A. Lindquist, G. Picci, and G. Ruckebusch with his advisor M. Metivier.

6.1 Introduction to Realization Theory

Realization theory is a major component of control and system theory used in this book. In control and system theory, realization theory mostly refers to realization of deterministic systems. The term *stochastic realization* is used for realization of stochastic systems. There follow historical comments on realization theory. The reader is also referred to the section *Further Reading* of this chapter in which many references are mentioned.

The term *realization* originates in circuit theory, also called electric network theory. Consider engineers who have formulated mathematically an impedance matrix of an electric network, either with two or more entry points called poles. Their problem is then: Characterize those impedance matrices for which there exists a finite passive electric network whose open-circuit impedance matrix equals the considered impedance matrix. This realization problem was considered already in the 1920's with major contributions dating from 1920 to 1960. The research area in which this problem was solved is no longer studied that intensively as before. See the two papers of B. McMillan of A.T.T. Bell Laboratories, [68, 69], for an entry in the early literature. The main result of realization theory of electric circuit is the characterization of those impedance matrices which admit a realization of a prespecified form. For recent references on realization of electric circuits see [50, 49, 90]. Currently there is a research interest to relate synthesis of mechanical systems to realization of electric circuits.

There is also an analogy of realization theory with the theory of regular languages in computer science. The paper of S.C. Kleene from 1936, [59], is often quoted in this regard. Later significant papers include those of A. Nerode, [70], of M.O. Rabin with D. Scott, [77], and those of J.A. Brzozowski [19, 20]. The concept of a regular language and the result that a regular language can be characterized as the output of finite-state automaton or generator, is now included in many text books for computer scientists. It is not clear to the author whether there is a relation between the realization theory of circuit theory and that of computer science.

It is well known that there is a generalization of realization theory which contains both realization of automata and realization of specific subsets of finite-dimensional linear systems, see [31] and [13]. Realization of bilinear systems can also be derived using realizaton of automata by using a realization over a Boolean semiring, [60, 84].

R.E. Kalman, [54], formulated and solved the realization problem for finite-dimensional linear systems. Products of this investigation were the concepts of: *controllability* and *observability* of a finite-dimensional linear system, *minimality of a realization*, and the *classication of all minimal realizations*. These concepts are by now used to characterize the existence of control laws and the existence of observers. Later on *canonical forms* were formulated for the considered set of systems, motivated by the research area of system identification.

Realization theory was extended to particular classes of nonlinear deterministic systems, notably bilinear systems, polynomial systems, rational systems, and nonlinear systems on differential-geometric manifolds.

It was noticed by several researchers, including H.H. Rosenbrock in the United Kingdom, [80], that in electric circuit theory there are difficulties in distinguishing inputs and outputs. The voltage can be an input and the current an output of a circuit, or conversely. J.C. Willems has formulated the concept of a *behavior* of a deterministic system, [93, 94, 95]. The reader may think of a behavior as the set of trajectories of the vector of inputs and outputs of a deterministic system in which the observation vector is no longer distinguished into inputs and outputs. If one wants to distinguish an observed variable into inputs and outputs, there are in particular cases several choices possible. The researcher may then to make a choice. Behaviors are also useful models for several other domains of engineering and the sciences. A book on the behavioral approach is [76].

R.E. Kalman in a conference paper [55], when reviewing the usefullness of the Kalman filter, formulated indirectly the realization problem for stochastic systems. The Frenchman P. Faurre worked with R.E. Kalman at Stanford University on what is now called the weak-Gaussian stochastic realization problem. The problem is: How to realize a stationary Gaussian process as the output of a time-invariant finite-dimensional Gaussian system? Call such a stochastic system then a *stochastic realization* of the considered stochastic process. Realization theory should be understood as solving the questions: Does a realization exist? When is a realization minimal in the sense of having a minimal state set or space? How to classify all minimal stochastic realizations? P. Faurre solved the problem for stationary Gaussian processes, see the book [34] which is written in the French language, and the report, [32], also written in the French language.

A generalization was proposed by G. Picci, [73], to the strong stochastic realization problem for stationary Gaussian processes. In this formulation the probability space is already specified and the stochastic realization has to be constructed on the considered probability space. The framework for the theory is Hilbert spaces and a key concept is conditional independence in Hilbert spaces. The resulting theory is fully described in the book of A. Lindquist and G. Picci, [65]. The initial work on this approach is due to G. Ruckebusch and his research advisor M. Metivier. Later G. Ruckebusch cooperated with A. Lindquist and G. Picci, [81, 82, 83, 66].

Stochastic realization theory for finite-valued stochastic systems is motivated by a system identification problem for such system, formulated by D. Blackwell and L. Koopmans, [16]. This problem is partly solved but not completely. For references see the section *Further Reading*.

Stochastic realization problems where there are not only outputs but also inputs, is not satisfactorily solved. Chapter 10 provides an introduction to the problem. However, there is a publication of a Stanford researcher who refers to R.E. Kalman as the originator of the problem, see [24].

The generalization of stochastic realization problems to other subsets of stochastic systems, other than Gaussian systems and other than finite systems, was developed by the author of this book. The main concept is that of the conditional independence relation of σ-algebras and that of a σ-algebraic system. This leads to problems of realization of σ-algebras and of σ-algebraic families indexed by the time index set. These problems are treated in Chapter 7.

A further generalization leads to the multiple conditional independence relation for a set of σ-algebras, see Section 5.9. This model is useful for the analysis of databases but the theory is currently underdeveloped.

Once the reader has grasped the concept of a system with outputs and states, the concept of a stochastic control system, and understood the realization problem, then the realization problem can be extended to other sets of stochastic systems. In the long run the most useful products of the theory are the concepts of observability, controllability, canonical forms of sets of stochastic realizations, filters, and optimal control laws characterized as extremal realizations.

The reader finds in this chapter an exposition of the Gaussian stochastic realization problem, primarily about the weak problem formulated in terms of probability distributions. Novel for most readers is that the characterization of minimality of such a realization is that it is stochastically observable and stochastically co-observable. The latter concept is defined for the backward representation of a Gaussian system. The strong Gaussian stochastic realization problem is well described in the book [65].

A very useful product of the Gaussian stochastic realization problem is the subspace identification algorithm of Gaussian systems, see [65, 87]

Chapter 7 treats primarily the stochastic realization of a tuple of σ-algebras and of a σ-algebra family. In addition, it provides an introduction to the stochastic realization problem for output-finite stochastic systems.

Section 21.8 provides a description of the realization of a time-invariant finite-dimensional linear system as developed by R.E. Kalman.

A note of warning is issued to computer scientists. The realization problem is, by its formulation, an undecidable problem. That realization theory is nevertheless useful is due to the conditions of controllability and of observability of a control system, which do admit a finite computational complexity.

6.2 Motivation

Practical problems of control and communication may be formulated as mathematical problems of stochastic control and filtering. Solution of the latter problems requires a model in the form of a stochastic system. The question is thus motivated: How to go from a time series to a stochastic system representation with its parameter values? This question leads to two major problems: (1) the *stochastic realization problem*: how to determine a stochastic system of which the family of finite-dimensional distributions *equals* the finite-dimensional distributions of the observed time series, formulated in this chapter as a realization problem of stochastic system theory; and (2) the *system identification problem*: how to determine a stochastic system of which the family of finite-dimensional distributions *approximates* the family of finite-dimensional distributions of the observed time series.

Depending on the type of problem, the adjectives weak, or strong, will be added to the expression of *stochastic realization*. In this chapter primary attention will be given to the weak and the strong realization problem for stationary Gaussian processes. Stochastic realization problems for non-Gaussian processes are discussed in Chapter 7.

Example 6.2.1. *Modelling of a paper machine* Consider Example 4.1.1 of control of a paper machine. An engineering model consists of the paper production process, and the input signals and output signals of the model. There are two system identification situations: one situation with constant input and another situation with a varying input. For the purpose of this chapter, only the case of a constant input is treated. Suppose then that the input signals are held constant during the data collection experiment. System identification with inputs requires additional concepts it is the topic of Chapter 10. The output signal of wet basis weight will fluctuate. Suppose that data are available on the fluctuation of wet basis weight, say $\{z(t), t \in T_1\}$, $T_1 = \{1,2,\ldots,t_1\}$. How to construct a mathematical model of this process? Suppose that the data are stationary. Suppose that the class of time-invariant Gaussian system representations is a suitable set of models. How to select a Gaussian stochastic system representation from the considered set which is a realistic model of the observations?

One approach to this problem is to fit a Gaussian system representation to an estimate of the covariance function. Another approach is to apply the maximum likelihood procedure.

From the data one may compute a function that can be considered as an estimate of the covariance function of the observations. First the data are inspected whether they do not have any deterministic trend, for example a periodic signal. If such a trend is present it must first be removed. Then one computes the average of the data according to,

$$z_a = \frac{1}{t_1} \sum_{s=1}^{t_1} z(s).$$

Next compute an estimate of the covariance function of the observations for the time indices $t = 0,1,\ldots,t_2 < t_1$,

$$\hat{W}(t) = \frac{1}{t_1 - t} \sum_{s=1}^{t_1 - t} (z(t+s) - z_a)(z(s) - z_a)^T.$$

Other formulas for a covariance function estimate have also been proposed but are not discussed in this chapter. Because $\hat{W}$ is to be a covariance function one may set for $t = -1,\ldots,-t_2$, $\hat{W}(t) = \hat{W}(-t)^T$, because a covariance function must satisfy that property, see Def. 3.4.6. For $\hat{W}$ to be a covariance function on the interval $\{-t_2,\ldots,-1,0,1,\ldots,t_2\}$ it is necessary and sufficient that the following block-Toeplitz matrix is positive-definite symmetric,

$$\begin{pmatrix} \hat{W}(0) & \hat{W}(1)^T & \ldots & \hat{W}(t_2)^T \\ \hat{W}(1) & \ddots & & \hat{W}(t_2-1)^T \\ \vdots & & \ddots & \vdots \\ \hat{W}(t_2) & \hat{W}(t_2-1) & \ldots & \hat{W}(0) \end{pmatrix} \succeq 0.$$

In working with actual data this condition may not be satisfied due to numerical approximations in which case one should modify the estimate. Such techniques are not in the scope of this book.

The problem is to determine a time-invariant Gaussian system representation, say

$$\begin{aligned} gsp &= \{n_y, n_x, n_v, A, C, M, N\} \in \text{GStocSP}, \\ x(t+1) &= Ax(t) + Mv(t), \\ y(t) &= Cx(t) + Nv(t),\ v(t) \in G(0,I), \end{aligned}$$

such that it is a model of the data. The covariance function of the above Gaussian system representation is according to Theorem 4.4.5 specified by,

$$W_y(t) = \begin{cases} CA^{t-1}Q_{(x^+,y)}, & t > 0, \\ Q_y, & t = 0. \end{cases}$$

As a criterion of fit consider the difference between the covariance function of the output of the system and that of the estimated covariance function: $\Delta W(t) = \hat{W}(t) - W_y(t),\ t = 0, 1, \ldots, t_2$. How to choose the parameters of the Gaussian system representation, the state-space dimension $n_x \in \mathbb{N}$, the space dimension $n_v \in \mathbb{N}$, and the matrices (A, C, M, N), such that the defined difference is small in terms of a specified distance function?

Several problems may now be formulated. Assume that the time series is a stationary Gaussian process characterized by its finite-dimensional probability distributions; by restriction it is characterized by its covariance function if the mean value function is assumed to be zero.

There is first the *weak Gaussian stochastic realization problem* in which one assumes that a function $\hat{W}$ is available for all $t \in T = \mathbb{Z}$, and in which one demands that $\hat{W}(t) = W(t)$ for all $t \in T = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. Next there is the *partial weak Gaussian stochastic realization problem* in which one demands that $\hat{W}(t) = W(t)$ for all $t = 0, 1, \ldots, t_2$, for a $t_2 \in \mathbb{Z}_+$, but in which there is neither a constraint on $W(t)$ for $t > t_2$ nor for $t < -t_2$. Finally there is the *system identification problem* in which one demands that the difference,

$$\Delta W(t) = \hat{W}(t) - W(t)$$

is small in terms of a distance, either for all $t \in T$ or for $t = 0, 1, \ldots, t_2$. The example is herewith concluded.

The general motivations for the stochastic realization problem include:

- the system identification problem discussed in the above example;
- the identifiability problem of the parametrization of a stochastic system which leads to the concept of a canonical form for such a system;
- existence of a filter system such that the error system is exponentially stable; and
- existence of a control law for a stochastic control system such that the closed-loop system meets control objectives of stability and of performance minimization.

6.3 Weak Gaussian Stochastic Realization Problem

Below the weak stochastic realization problem for Gaussian processes is formulated. Recall that a stationary Gaussian process, characterized by its family of finite-dimensional probability distributions, is completely characterized by its mean value function and its covariance function.

Problem 6.3.1. The *weak Gaussian stochastic realization problem* for a stationary Gaussian process is, considering a stationary Gaussian process on $T = \mathbb{Z}$ taking values in $(\mathbb{R}^{n_y}, B(\mathbb{R}^{n_y}))$ having mean value function zero and covariance function $W : T \to \mathbb{R}^{n_y \times n_y}$, to solve the following subproblems.

(a) Does there exist a time-invariant Gaussian system

$$\{\Omega, F, P, T, \mathbb{R}^{n_y}, B(\mathbb{R}^{n_x}), \mathbb{R}^{n_x}, B(\mathbb{R}^{n_x}), y, x\} \in \text{GStocS}$$

such that the output process y of this system equals the considered process in terms of their families of finite-dimensional distributions? Effectively this means that the covariance function of the output process must be equal to the considered covariance function W because both processes are Gaussian and have mean value function zero.

If such a system exists, then one calls this system a *weak Gaussian stochastic realization* of the considered process, or, if the context is clear, a *stochastic realization.*

(b) Classify all minimal stochastic realizations of the considered process. A weak Gaussian stochastic realization is called *minimal* if the dimension of the state space is minimal within the class of all stochastic realizations. The following points must be addressed:

(1) characterize those stochastic realizations that are minimal;
(2) obtain the classification as such;
(3) indicate the relation between two minimal stochastic realizations;
(4) formulate a procedure which constructs all minimal weak Gaussian stochastic realizations of the considered process.

In Problem 6.3.1 one considers a stationary Gaussian process with zero mean value function. Such a process is thus completely characterized by its covariance function. In part (a) of this problem the question is whether for the considered process, there exists a Gaussian system which equals the considered process in distribution. Because by definition such a Gaussian system has a finite-dimensional state space, not all stationary Gaussian processes can be the output process of a Gaussian system. The question should therefore be interpreted as to determine a necessary and sufficient condition on the considered stochastic process, or on its covariance function, such that it can be the output process of a finite-dimensional ($n_x < \infty$) Gaussian system.

In part (b) of Problem 6.3.1 a classification is asked for. This question arises because a stochastic realization, if it exists, is in general nonunique as will be proven later in this chapter. The dimensions of the state space of two stochastic realizations may also be different in general. For system theoretic reasons, such as identifiability, one should restrict attention to those stochastic realizations for which the dimension of the state space is minimal. Such a realization is called a *minimal realization.* In general minimal stochastic realizations are also nonunique. A classification of all minimal stochastic realizations is then useful for the identifiability question.

Publications have appeared on the partial version of the weak stochastic realization problem. In this case one is presented with a covariance function on a finite horizon, rather than a covariance function on the infinite horizon $T = \mathbb{Z}$ as in Problem 6.3.1. Thus let $t_1 \in \mathbb{Z}_+$,

$$T(-t_1 : +t_1) = T_1 = \{-t_1, -t_1+1, ..., -1, 0, +1, ..., t_1\},$$

and $W_1 : T_1 \to \mathbb{R}^{n_y \times n_y}$ be a covariance function. The *partial weak Gaussian stochastic realization problem* is then to classify all Gaussian system representations such that the covariance function of the output process equals the considered covariance function W_1 on the finite horizon T_1.

The motivation of this problem is the following. Given a finite time series $T_2 = \{1, 2, \ldots, t_2\}$, $t_2 \in \mathbb{Z}_+$, and $y : T_2 \to \mathbb{R}^{n_y}$. An estimate of the covariance function has been presented in Example 6.2.1. It provides an estimate $W_1 : T_1 \to \mathbb{R}^{n_y \times n_y}$ of the covariance function on the finite horizon $T_1 = \{-t_1, \ldots, -1, 0, +1, \ldots, t_1\}$, $t_1 \in \mathbb{Z}_+$. For certain procedures in system identification a covariance function on an infinite-time horizon, or at least a rather large time horizon, is needed. In several papers one defines the function $W_3 : T \to \mathbb{R}^{p \times p}$ on $T = Z$ by

$$W_3 : T \to \mathbb{R}^{n_y \times n_y},\ T = \mathbb{Z},\ W_3(t) = \begin{cases} W_1(t), & \text{for } t \in T_1, \\ 0, & \text{for } t \in T \cap T_1^c \end{cases}$$

Thus W_3 is chosen to be zero on the set $T \cap T(-t_1 : +t_1)^c$. This somewhat arbitrary choice is not easy to justify mathematically. From a theoretical viewpoint it seems useful to investigate which covariance functions like W_1 defined on a finite horizon are equal to the restriction to the same finite horizon of a covariance function defined on the infinite interval. See for this issue the book [65].

6.4 The Theorem

The reader finds in this section the solution of the weak Gaussian stochastic realization Problem 6.3.1 for a stationary Gaussian process.

The theory of the weak Gaussian stochastic realization problem as described in this chapter was developed by P. Faurre and colleagues, [32, 33, 34].

In this section use is made of realization theory of finite-dimensional linear systems as stated in Chapter 21, in particular, the Sections 21.8 – 21.8.

The reader is reminded of the concept of an infinite Hankel matrix and its rank.

Definition 6.4.1. Associate with a covariance function $W : T \to \mathbb{R}^{n_y \times n_y}$ on $T = \mathbb{N}$ of a stationary stochastic process, the *finite Hankel matrix* of block sizes $k,\ m \in \mathbb{Z}_+$ by the formula,

$$H_W(k,m) = \begin{pmatrix} W(1) & W(2) & W(3) & \ldots & W(m-1) & W(m) \\ W(2) & W(3) & W(4) & \ldots & W(m) & W(m+1) \\ W(3) & W(4) & W(5) & \ldots & W(m+1) & W(m+2) \\ \vdots & & & \vdots & & \vdots \\ W(k) & W(k+1) & W(k+2) & \ldots & W(k+m-2) & W(k+m-1) \end{pmatrix}$$
$$\in \mathbb{R}^{kn_y \times mn_y}.$$

Define the *infinite Hankel matrix* $H_W \in \mathbb{R}^{\infty \times \infty}$ as an infinite by infinite matrix such that, for any $k,\ m \in \mathbb{Z}_+$, the upper-left block of the infinite Hankel matrix equals the finite Hankel matrix displayed above. Define the rank of the infinite Hankel matrix,

$$\text{rank}(H_W) = \sup_{k,\ m \in \mathbb{Z}_+} \text{rank}(H_W(k,m)) \in \mathbb{Z}_+ \cup \{+\infty\}.$$

One says that infinite Hankel matrix has *finite rank* if $\text{rank}(H_W) < \infty$.

In the formulation of Theorem 6.4.3 use is made of the set $\mathbf{Q_{lsdp}}$ presented in Definition 24.1.1 of which the definition is repeated here for the convenience of the reader.

Definition 6.4.2. Define the set of state variance matrices associated with a set of matrices describing a covariance realization according to,

$$\begin{aligned} & (n_y, n_x, n_y, F, G, H, J) \in \text{LSP}, \\ \mathbf{Q_{lsdp}}(F,G,H,J) &= \{Q \in \mathbb{R}^{n_x \times n_x}_{pds} \mid Q_{v,lsdp}(Q) \succeq 0\}, \\ Q_{v,lsdp}(Q) &= \left(\begin{array}{c|c} Q - FQF^T & G - FQH^T \\ \hline G^T - HQF^T & J + J^T - HQH^T \end{array} \right). \end{aligned}$$

Theorem 6.4.3. Weak Gaussian stochastic realization of a stationary Gaussian process.

Consider the weak Gaussian Stochastic Realization Problem 6.3.1 for a stationary Gaussian process. Assume that (1) the time index set is $T = \mathbb{Z}$, (2) that the considered process has zero mean value function, (3) that $\lim_{t \to \infty} W(t) = 0$*, and (4) that* $W(0) \succ 0$.

(a)There exists a time-invariant Gaussian system with the parameters,

$$\begin{aligned} & (n_y, n_x, n_v, A, C, M, N), \\ x(t+1) &= Ax(t) + Mv(t), \qquad (6.1) \\ y(t) &= Cx(t) + Nv(t), \ \ v(t) \in G(0,I), \qquad (6.2) \end{aligned}$$

such that the family of finite-dimensional probability distribution functions of the output process equals the family of finite-dimensional probability distribution *functions of the considered process, if and only if the Hankel matrix H_W associated with the covariance function W has finite rank; equivalently, if it satisfies* $\text{rank}(H_W) < \infty$.

If the above condition holds then there exists a tuple,

$$\{n_y, n_x, n_y, F, G, H, J\} \in \text{LSP},$$

such that the covariance function W satisfies,

$$W(t) = \begin{cases} HF^{t-1}G, & \text{if } t > 0, \\ J + J^T, & \text{if } t = 0, \\ G^T(F^T)^{-t-1}H^T, & \text{if } t < 0. \end{cases} \quad (6.3)$$

A linear system specified by the parameter values $\{n_y, n_x, n_y, F, G, H, J\} \in \text{LSP}$ having the form of equation (6.3) will be called a covariance realization *of the covariance function W.*

The time-invariant Gaussian system described by the equations (6.1,6.2) is then called a weak-Gaussian stochastic realization *of the considered stationary Gaussian process.*

(b)*Consider a time-invariant Gaussian system. For any stochastic realization there exists a state variance matrix $Q_x \in \mathbb{R}^{n_x \times n_x} \cap \mathbf{Q_{lsdp}}$.*

Conversely, for any covariance realization $(F, G, H, J) \in \text{LSP}$, there exists exists a matrix $Q \in \mathbf{Q_{lsdp}}(F, G, H, J)$ and a time-invariant Gaussian system such that the invariant distribution of this system has the same state variance, thus $Q = Q_x$.

There exists a minimal and maximal element of the set of state-variance matrices $\mathbf{Q_{lsdp}}$ denoted by,

$$\exists\, Q^-,\ Q^+ \in \mathbf{Q_{lsdp}}(F, G, H, J),\ \textit{such that,}$$
$$\forall Q_x \in \mathbf{Q_{lsdp}}(F, G, H, J),\ Q^- \preceq Q_x \preceq Q^+. \quad (6.4)$$

(c)*Consider a time-invariant Gaussian system described by the equations (6.1,6.2) and assume that its system matrix is such that* $\text{spec}(A) \subset \text{D}_o$. *If the system is started at time $t = 0$ in its invariant distribution then the output process is a stationary Gaussian process.*

This Gaussian system is a minimal weak Gaussian stochastic realization of its output process if and only if the following conditions all hold:

1. *the support of the state process is all of the vector space $\mathbb{R}^{n_x}$; equivalently, if $0 \prec Q_x$; or, equivalently, if the matrix tuple (A, M) is a supportable pair;*
2. *the system representation is stochastically observable; and*
3. *the system representation is stochastically co-observable.*

(d)*A minimal weak Gaussian stochastic realization is nonunique in two ways.*

1. *If $gsp_1 = \{n_y, n_x, n_v, A, C, M, N\} \in$* GStocSP *are the parameters of a forward representation of a minimal stochastic realization, and if $L_x \in \mathbb{R}^{n_x \times n_x}$ is a nonsingular matrix and $U_v \in \mathbb{R}^{n_v \times n_v}$ is an orthogonal matrix, then*

$$gsp_2 = \{n_y, n_x, n_v, L_x A L_x^{-1}, C L_x^{-1}, L_x M U_v, N U_v\} \in \text{GStocSP},$$

are also the parameters of a forward representation of a minimal stochastic realization.

2. *If*

$$gsp_1 = \{n_y, n_x, n_v, A_1, C_1, M_1, N_1\} \in \text{GStocSP},$$
$$gsp_2 = \{n_y, n_x, n_v, A_2, C_2, M_2, N_2\} \in \text{GStocSP},$$

are the parameters of two minimal weak stochastic realizations of the output process then there exists matrices $L_x \in \mathbb{R}^{n_x \times n_x}$ nonsingular and $U_v \in \mathbb{R}^{n_v \times n_v}$ orthogonal, such that,

$$A_1 = L_x A_2 L_x^{-1},\ M_1 = L_x M_2 U_v,\ C_1 = C_2 L_x^{-1},\ N_1 = N_2 U_v.$$

3. *Fix the parameters of a minimal covariance realization as given in (a) above, $lsp = \{n_y, n_x, n_y, F, G, H, J\} \in \text{LSP}_{min}$. Denote the parameters of a forward representation of a minimal Gaussian stochastic realization by $\{n_y, n_x, n_v, A, C, M, N\}$ and the set of such parameters by $WGSRP_{min}$. Define the classification map,*

$$\begin{aligned} c_{lsp}(Q) &= \{n_y, n_x, A, C, M, N\},\ c_{lsp} : \mathbf{Q_{lsdp}} \to WGSRP_{min}, \\ lsp &= \{n_y, n_x, n_y, F, G, H, J\} \in \text{LSP}_{min}, \\ A &= F,\ C = H,\ n_x \text{ is provided by } lsp, \\ n_v &= n_x + n_y,\ M \in \mathbb{R}^{n_x \times n_v},\ N \in \mathbb{R}^{n_y \times n_v}, \\ \begin{pmatrix} M \\ N \end{pmatrix} \begin{pmatrix} M \\ N \end{pmatrix}^T &= Q_{v,d}(Q) = \left(\begin{array}{c|c} Q - FQF^T & G - FQH^T \\ \hline G^T - HQF^T & J + J^T - HQH^T \end{array} \right). \end{aligned}$$

Then, for fixed $lsp \in \text{LSP}_{min}$, c_{lsp} is a bijection. Thus, for fixed $lsp \in \text{LSP}_{min}$, all minimal weak Gaussian stochastic realization are parametrized by the elements of the set of state-variance matrices $\mathbf{Q_{lsdp}}$.

(e) The stochastic realization Procedure 6.4.4 is well defined and constructs all minimal weak Gaussian stochastic realizations.

Procedure 6.4.4 The stochastic realization procedure for
a weak Gaussian stochastic realization of a stationary Gaussian process.
Data: *Consider a stationary Gaussian processes with zero mean-value function and covariance function $W : T \to \mathbb{R}^{n_y \times n_y}$. Assume that the conditions of Theorem 6.4.3.(a) hold.*

1. *Determine a minimal covariance realization of W via the realization procedure for a time-invariant finite-dimensional linear system, or $lsp = \{n_y, n_x, n_y, F, G, H, J\} \in \text{LS}_{min}$, such that,*

$$W(t) = \begin{cases} HF^{t-1}G, & \text{if } t > 0, \\ J + J^T, & \text{if } t = 0, \\ G^T (F^T)^{-t-1} H^T, & \text{if } t < 0. \end{cases}$$

The tuple (F, G, H, J) is called a covariance realization *of the covariance function W. The procedure to determine this matrix tuple is described in the proof of the* *theorem of realization of a time-invariant linear system, Theorem 21.8.9. For procedures for this step see also references on linear system theory. This is in principle a procedure which takes an infinite number of steps.*

2. *Determine a matrix* $Q_x \in \mathbf{Q_{lsdp}}$*, or, equivalently, a* $Q_x \in \mathbb{R}_{pds}^{n_x \times n_x}$ *satisfying,*

$$\begin{pmatrix} Q_x - FQ_xF^T & G - FQ_xH^T \\ G^T - HQ_xF^T & J + J^T - HQ_xH^T \end{pmatrix} \succeq 0, \ Q_x = Q_x^T \succeq 0.$$

For procedures, see Section 22.3 and Chapter 24.

3. *Define the integer and matrices,*

$$\begin{aligned} n_v &= n_x + n_y, \ A = F, \ C = H, \ M \in \mathbb{R}^{n_x \times n_v}, \ N \in \mathbb{R}^{n_y \times n_v}, \\ & \begin{pmatrix} M \\ N \end{pmatrix} \begin{pmatrix} M \\ N \end{pmatrix}^T = Q_{v,d}(Q_x) \\ &= \begin{pmatrix} Q_x - FQ_xF^T & G - FQ_xH^T \\ G^T - HQ_xF^T & J + J^T - HQ_xH^T \end{pmatrix} \in \mathbb{R}^{(n_x+n_y)\times(n_x+n_y)}, \end{aligned}$$

construct a probability space by

$$\begin{aligned} \Omega &= \mathbb{R}^{n_x} \times (\mathbb{R}^{n_v})^T, \ \omega = (\omega_1, \omega_2), \ (\mathbb{R}^{n_v})^T = \{h : T \to \mathbb{R}^{n_v}\}, \\ F &= \sigma\text{-}algebra \text{ } generated \text{ } by \text{ } the \text{ } subsets \text{ } of \, B(\mathbb{R}^{n_x}) \otimes B(cylinder \text{ } sets), \\ & x_0 : \Omega \to \mathbb{R}^n, \ x_0(\omega) = \omega_1, \\ & v : \Omega \times T \to \mathbb{R}^{(n+p)}, v(\omega, t) = \omega_2(t), \ P : F \to [0,1], \end{aligned}$$

a probability measure such that v is a standard Gaussian white noise process with intensity I_{n_v}*,* $(F^{x_0}, \ F^v_\infty)$ *are independent* σ*-algebras, and* $x : \Omega \times T \to \mathbb{R}^{n_x}$ *and* $y : \Omega \times T \to \mathbb{R}^{n_y}$ *are defined by*

$$x(t+1) = Ax(t) + Mv(t), \ x_0 \in G(0, Q_x), \quad (6.5)$$

$$y(t) = Cx(t) + Nv(t). \quad (6.6)$$

Then

$$\{\Omega, F, P, T, \mathbb{R}^{n_y}, B(\mathbb{R}^{n_y}), \mathbb{R}^{n_x}, B(\mathbb{R}^{n_x}), y, x\} \in \text{GStocS}$$

is a minimal weak Gaussian stochastic realization of the considered process, meaning that the output process y is a Gaussian process with covariance function equal to the considered covariance function W.

Theorem 6.4.5. *Consider Problem 6.3.1. The following statements are equivalent:*

(a)The covariance function W is uniformly strictly positive definite, see Def. 3.4.2.

(b)Denote the minimal and maximal elements of the set of state-variance matrices by Q^-*,* $Q^+ \in \mathbf{Q_{lsdp}}$*; then* $Q^+ - Q^- > 0$*;*

(c)There exists a covariance realization with the matrices (n_x, F, G, H, J) *which is regular (meaning that* $n_y = \text{rank}(J+J^T)$*) and the following four spectral conditions all hold:*

$$\begin{aligned}
&\text{spec}(F - G(J+J^T)^{-1}(H - G^T Q^-)) \subset \text{D}_o, \\
&\text{spec}(F - G(J+J^T)^{-1}(H - G^T Q^+)) \subset \text{D}_o, \\
&\text{spec}(-(F - (G - \tilde{Q}^- H^T)(J+J^T)^{-1}H)) \subset \text{D}_o, \\
&\text{spec}(-(F - (G - \tilde{Q}^+ H^T)(J+J^T)^{-1}H)) \subset \text{D}_o.
\end{aligned}$$

(d)The interior int$(\mathbf{Q_{lsdp}}) \neq \emptyset$ *if and only if,*

$$\begin{aligned}
&\exists\, Q_x \in \partial\mathbf{Q_{lsdp}} \textit{ such that, } Q_x = Q_x^T \succ 0 \textit{ and} \\
&\begin{pmatrix} Q_x - FQ_xF^T & G - FQ_xH^T \\ G^T - HQ_xF^T & J + J^T - HQ_xH^T \end{pmatrix} \succ 0.
\end{aligned}$$

Definition 6.4.6. (a)A *(forward) Kalman realization* is a weak Gaussian stochastic realization with the following representation and satisfying the following conditions,

$$\begin{aligned}
x(t+1) &= Ax(t) + Mv(t),\ x(0) = x_0, \\
y(t) &= Cx(t) + Nv(t),\ v(t) \in G(0, I_{n_v}), \\
&\quad \text{spec}(A) \subset \text{D}_o,\ n_v = n_y,\ \text{rank}(N) = n_y, \\
Q_x &= AQ_xA^T + MM^T \ \Rightarrow\ Q_x = Q^- \in \partial\mathbf{Q_{lsdp,r,s}}.
\end{aligned}$$

It is called a *minimal (forward) Kalman realization* if in addition (1) (A, M) is a supportable pair, (2) (A, C) is an observable pair, and (3) (A_b, C_b) is an observable pair.

(b)It is called a *backward Kalman realization* if it has the following representation satisfying also the variance relation,

$$\begin{aligned}
x(t-1) &= A_bx(t) + M_bv(t),\ x(0) = x_0, \\
y(t-1) &= C_bx(t) + N_bv(t),\ v(t) \in G(0, I_{n_v}), \\
&\quad \text{spec}(A) \subset \text{D}_o,\ n_v = n_y,\ \text{rank}(N) = n_y, \\
Q_x &= AQ_xA^T + MM^T \ \Rightarrow\ Q_x = Q^-.
\end{aligned}$$

Proposition 6.4.7. *Consider the weak Gaussian stochastic realization problem with the conditions of Theorem 6.4.3. Then there exist minimal and maximal state variance* Q^-, Q^+. *The following statements are equivalent:*

(a)$Q \in \mathbf{Q_{lsdp}}$ *satisfies* $Q = Q^-$*; and*
(b)$Q \in \mathbf{Q_{lsdp}}$ *satisfies,*

$$\begin{aligned}
&0 \prec J + J^T - GQG^T, \quad \text{spec}(A - K(Q)C) \subset \text{D}_o; \textit{ where,} \\
K(Q) &= [AQC^T + MN^T][CQC^T + NN^T]^{-1}.
\end{aligned}$$

Proof. The equivalence follows from Proposition 24.6.2 and from Proposition 24.6.3. The condition for the matrix $(A - K(Q)C)$ then follows from the theory of the realization Riccati equation, see Section 22.3. □

6.5 Explanation

The text of this section is best considered as an explanation of Theorem 6.4.3. It will be argued that the conditions of Theorem 6.4.3.(b) are necessary for minimality of a weak Gaussian stochastic realization.

The eigenvalues of the system matrix of a time-invariant Gaussian system can be either in the interior of the unit disc, on the unit circle, or in the interior of the outside of the unit disc, or in various combinations of these subsets of the complex plane. The case of eigenvalues on the unit disc has to be treated separately.

If the system matrix A is not exponentially stable, say it has an eigenvalue outside the unit disc, but the other three conditions of Theorem 6.4.3.(c) hold, then the output process has a covariance function which goes off to infinity. That contradicts the assumption that $\lim_{t\to\infty} W(t) = 0$. Hence the system matrix cannot have strictly unstable eigenvalues, those with $|\lambda| > 1$.

There follow several examples which could also be considered as special cases of the theorem.

Example 6.5.1. Suppose that the tuple of matrices (A,M) is not a supportable pair while the other conditions of Theorem 6.4.3.(c) hold. Then it follows from the Kalman controllable form of a linear system that there exists a state space transformation matrix $L_x \in \mathbb{R}^{n_x \times n_x}$ nonsingular, such that, if one defines $\bar{x}(t) = L_x x(t)$, then, with respect to the new basis, the system has the representation,

$$\bar{x}(t+1) = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} \bar{x}(t) + \begin{pmatrix} M_1 \\ 0 \end{pmatrix} v(t), \;\; \bar{x}(t) = \begin{pmatrix} \bar{x}_1(t) \\ \bar{x}_2(t) \end{pmatrix}.$$

By the assumption on the system matrix A being exponentially stable, it follows that $\lim_{t\to\infty} Q_{\bar{x},2,2}(t) = 0$. The process $\bar{x}_2$ is therefore not useful in the long run of the stationary process and is therefore best deleted from the minimal weak Gaussian stochastic realization. Thus the matrix tuple (A,M) is best taken to be a supportable pair.

Example 6.5.2. Suppose that the system is not stochastically observable while the other conditions of Theorem 6.4.3.(b) hold. It then follows from the characterization of stochastic observability of Theorem 4.6.8 and from the Kalman observable form of a linear system, Proposition 21.2.9, that there exists a state-space transformation to a Gaussian system of the form,

$$\begin{aligned} x(t+1) &= \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix} x(t) + Mv(t), \\ y(t) &= \begin{pmatrix} C_1 & 0 \end{pmatrix} x(t) + Nv(t). \end{aligned}$$

It is then directly obvious from these system matrices that the second component of the state x never influences the output process y. Note that the weak Gaussian stochastic realization problem asks for a representation of the output process only. That second component can therefore be safely deleted from the system. With the second component present, the system is not a minimal realization.

Example 6.5.3. Next suppose that the Gaussian system is not stochastically co-observable. From the characterization of the stochastic co-observability of Theorem 4.6.9 and a transformation of linear systems see Proposition 21.3.6, follows that the stochastic realization then has the form,

$$\begin{aligned} x(t-1) &= \begin{pmatrix} A_{b,11} & 0 \\ A_{b,21} & A_{b,22} \end{pmatrix} x(t) + M_b v_b(t), \\ y(t-1) &= \begin{pmatrix} C_{b,1} & 0 \end{pmatrix} x(t) + N_b v_b(t). \end{aligned}$$

As in the previous case, the second component of this state process does not influence the output process at all and hence can be deleted from the system representation.

It is a fact that stochastic observability of a Gaussian system is not equivalent with stochastic co-observability of this system. See Exercise 6.12.2.

A mistake that is sometimes made is the following. Consider the following forward representation of a Gaussian system

$$\begin{aligned} x(t+1) &= Ax(t) + Mv(t), \\ y(t) &= Cx(t) + Nv(t), \end{aligned}$$

with $v(t) \in G(0,Q_v)$. A statement is that if the pair of matrices (A,M) is a controllable pair and if (A,C) is an observable pair, that then the stochastic realization described by the above system representation is a minimal realization of the output process. This statement is false as the following example shows.

Example 6.5.4. Consider the Gaussian system with forward representation,

$$\begin{aligned} & \{\Omega, F, P, T, \mathbb{R}, B(\mathbb{R}), \mathbb{R}, B(\mathbb{R}, y, x\} \in \text{GStocS}, \\ x(t+1) &= ax(t) + mv(t), \\ y(t) &= x(t) + v(t),\ v(t) \in G(0,1),\ a \in (-1,+1), a \neq 0,\ m = (a^2-1)/a. \end{aligned}$$

(a)Then (a,m) is a controllable pair and $(a,1)$ is an observable pair.
(b)The system is a nonminimal realization of its output process.
(c)The Gaussian system reprensentation is not stochastically co-observable.

Proof. (a) The conditions $a \in (-1,+1)$, $a \neq 0$, imply that $m \neq 0$, hence the result. (b) From Theorem 4.4.5 follows that the covariance function of the output process is specified by

$$\begin{aligned} q &= a^2 q + m^2, \\ E[y(t)y(0)] &= \begin{cases} q+1, & \text{if } t = 0, \\ a^{t-1}[aq+m], & \text{if } t > 0. \end{cases} \quad \text{Then} \\ q &= (1-a^2)/a^2,\ aq+m = 0,\ \text{hence,} \\ E[y(t)y(0)] &= \begin{cases} a^{-2}, & \text{if } t = 0, \\ 0, & \text{if } t \neq 0, \end{cases} \end{aligned}$$

and y is a Gaussian white noise process. For a Gaussian white noise process there exists a weak Gaussian stochastic realization with state-space dimension $n_x = 0$ and with $y(t) = v(t)$, while the stochastic realization above has state-space dimension $n_x = 1$. Thus the realization above is not minimal. □

An implication of the weak Gaussian stochastic realization problem for the identifiability question is illustrated by the following example.

Example 6.5.5. Consider the time-invariant Gaussian system with forward representation,

$$\{\Omega, F, P, T, \mathbb{R}, B(\mathbb{R}), \mathbb{R}, B(\mathbb{R}), y, x\} \in \text{GStocS},$$
$$x(t+1) = ax(t) + \begin{pmatrix} 1 & 0 \end{pmatrix} v(t), \tag{6.7}$$
$$y(t) = cx(t) + \begin{pmatrix} 0 & 1 \end{pmatrix} v(t), \quad v(t) \in G(0, Q_v), \tag{6.8}$$
$$Q_v = \begin{pmatrix} q_{11} & 0 \\ 0 & q_{22} \end{pmatrix}. \tag{6.9}$$

Consider the asymptotic Kalman filter for the Gaussian system (6.7) and (6.8)

$$\hat{x}(t+1) = a\hat{x}(t) + k[y(t) - c\hat{x}(t)],$$
$$\overline{v}(t) = y(t) - c\hat{x}(t),$$

in which $\overline{v} : \Omega \times T \mapsto R$ is a Gaussian white noise process with $\overline{v}(t) \in G(0, q_{\overline{v}})$. This asymptotic Kalman filter may be rewritten as

$$\hat{x}(t+1) = a\hat{x}(t) + k\overline{v}(t), \tag{6.10}$$
$$y(t) = c\hat{x}(t) + \overline{v}(t), \tag{6.11}$$

in which $\overline{v} : \Omega \times T \mapsto \mathbb{R}$ is a Gaussian white noise process with $\overline{v}(t) \in G(0, Q_{\overline{v}})$. Then $(k\overline{v}(t),\ \overline{v}(t))^T$ has the variance matrix,

$$Q_{(k\overline{v},\overline{v})} = \begin{pmatrix} k \\ 1 \end{pmatrix} Q_{\overline{v}} \begin{pmatrix} k & 1 \end{pmatrix}. \tag{6.12}$$

From these forward representations one deduces that (6.7, 6.8), and (6.10, 6.11) are both weak Gaussian stochastic realizations of the output process y. This may be verified by computing the covariance function of the output process. This example shows that one may not be able to uniquely determine the parameters of the noise process of a Gaussian system, here (6.9) and (6.12), from the covariance function of the output process. For system theoretic reasons one chooses the stochastic realization (6.10, 6.11) as the representative of all stochastic realizations. This point is discussed in more detail in Section 6.9.

6.6 The Proof

Proof. Of Theorem 6.4.3. (1) Assume that there exists a weak-Gaussian stochastic realization of the considered process with the forward representation,

$$\begin{aligned}
&\{\Omega, F, P, T, \mathbb{R}^{n_y}, B(\mathbb{R}^{n_y}), \mathbb{R}^{n_x}, B(\mathbb{R}^{n_x}), y, x\} \in \text{GStocS}, \\
x(t+1) &= Ax(t) + Mv(t), \\
y(t) &= Cx(t) + Nv(t), \; v(t) \in G(0, I).
\end{aligned}$$

One may suppose that $\text{spec}(A) \subset \text{D}_o$; for if not then stationarity of the considered output process, the assumed convergence of the covariance function $\lim_{t\to\infty} W(t) = 0$, and that (A, C) is an observable pair imply that the dimension of the state space can be reduced such that for the reduced system $\text{spec}(A) \subset \text{D}_o$. It follows from Theorem 4.4.5 that there exists a $Q_x \in \mathbb{R}_{pds}^{n_x \times n_x}$ such that,

$$Q_x = AQ_xA^T + MM^T, \; Q_x = Q_x^T \succeq 0,$$

and, for $x_0 \in G(0, Q_x)$, it follows from Theorem 4.4.5.(b) that y is a stationary Gaussian process with,

$$\begin{aligned}
&E[y(t)] = 0, \;\; \forall\, t \in T, \\
&E[y(t)y(s)^T] = \begin{cases} CQ_xC^T + NN^T, & \text{if } t = s, \\ CA^{t-s-1}[AQ_xC^T + MN^T] = CA^{t-s-1}Q_{(x^+,y)}, & \text{if } s < t. \end{cases}
\end{aligned}$$

Let $F = A$, $H = C$, $G = Q_{(x^+,y)} = AQ_xC^T + MN^T$, $J + J^T = CQ_xC^T + NN^T$, and $CA^tQ_{(x^+,y)} = HF^tG$. Then, because the system is a realization,

$$\begin{aligned}
W(t) &= E[y(t)y(0)^T] = \begin{cases} HF^{t-1}G, & \text{if } t > 0, \\ J + J^T, & \text{if } t = 0, \\ G^T(F^T)^{-t-1}H^T, & \text{if } t < 0; \end{cases} \\
H_W(k, m) &= \begin{pmatrix} H \\ HF \\ HF^2 \\ \vdots \\ HF^{k-1} \end{pmatrix} \begin{pmatrix} G \; FG \; F^2G \ldots F^{m-1}G \end{pmatrix}, \\
&\Rightarrow \text{rank}(H_W(k, m)) \le n_x \;\forall\, k, m \in \mathbb{Z}_+, \text{ and } \text{rank}(H_W) \le n_x.
\end{aligned}$$

Thus W is a covariance function that has a finite rank of a Hankel matrix and admits a finite-dimensional realization. Then $Q_x = Q_x^T \succeq 0$ and,

$$\begin{pmatrix} Q_x - FQ_xF^T & G - FQ_xH^T \\ G^T - HQ_xF^T & J + J^T - HQ_xH^T \end{pmatrix} = \begin{pmatrix} M \\ N \end{pmatrix} \begin{pmatrix} M \\ N \end{pmatrix}^T \succeq 0 \;\Rightarrow\; Q_x \in \mathbf{Q_{lsdp}}.$$

(2) If the rank of the infinite Hankel matrix is finite it follows from Theorem 21.8.9 that there exists a covariance realization of the covariance function W denoted as in Equation (6.3). Next the steps of Procedure 6.4.4 will be checked.

Because W admits a representation as in equation (6.3) there exists a $\text{lsp} = \{n_y, n_x, n_y, F, G, H, J\} \in \text{LSP}_{min}$ such that

$$W(t) = \begin{cases} HF^{t-1}G, & \text{if } t > 0, \\ J + J^T, & \text{if } t = 0. \end{cases}$$

This representation is called a *covariance realization.* Note that because $W(t) = W(-t)^T$ for all $t \in T$, hence $W(0) = W(0)^T$, one can set $J = W(0)/2 = J^T$. Hence one needs only to determine n_x, F, G, H such that

$$W(t) = HF^{t-1}G, \text{ for } t > 0.$$

It follows from Theorem 21.8.9.(b) of the realization theory of a finite-dimensional linear system that minimality of the covariance realization is equivalent with (F,G) a controllable pair and (F,H) an observable pair. The conditions that $\lim_{t\to\infty} W(t) = 0$ and that $\text{lsp} \in \text{LSP}_{min}$ imply then that $\text{spec}(F) \subset \text{D}_o$.

The covariance function W is positive definite, see the remark following 23.1.5. This, the expression for W obtained as above, Theorem 23.4.2, and Proposition 23.4.6 imply that there exists a matrix $Q_x \in \mathbf{Q_{lsdp}}$.

The third step of the procedure is a construction.

(3) It will be shown that the stochastic system that is constructed by Procedure 6.4.4 is indeed a weak Gaussian stochastic realization. For this system with the forward representation (6.5) and (6.6) it follows from (1) above that the output process is a stationary Gaussian process with zero mean-value function and

$$E[y(t)y(0)^T] = \begin{cases} CQ_1C^T + NN^T, & \text{for } t = 0, \\ CA^{t-1}[AQ_1C^T + MN^T], & \text{for } t > 0, \end{cases} \quad \text{where,}$$
$$Q_1 = AQ_1A^T + MM^T,\ Q_1 = Q_1^T \succeq 0.$$

From the definition of $Q_v = Q_{v,d}(Q)$ in Step (2) of Procedure 6.4.4 follows that $Q_x = AQ_xA^T + MM^T$, where Q_x satisfies the same equation as that for Q_1. From $\text{spec}(A) \subset \text{D}_o$ and Theorem 22.1.2.(b) follows that $Q_1 = Q_x$. Then, by definition of Q_v and by the expression for W,

$$E[y(t)y(0)^T] = \begin{cases} CQ_xC^T + NN^T = J + J^T, & \text{for } t = 0, \\ CA^{t-1}[AQ_xC^T + MN^T] = HF^{t-1}G, & \text{for } t > 0, \end{cases}$$
$$= W(t),$$

and the system is a weak Gaussian stochastic realization of the considered process. The result Theorem 6.4.3.(a) then follows from (1), (2), and (3) above.

It is proven in Step (1) of the proof that the existence of a stochastic realization imples that there exists a state-variance matrix $Q_x \in \mathbf{Q_{lsdp}}$. In Step (2) it is proven that if there exists a matrix $Q_x \in \mathbf{Q_{lsdp}}$ then there exists a stochastic realization. It follows from Theorem 23.4.2.(b) that there exist matrices Q^-, Q^+ which are respectively the minimal and the maximal element of the set $\mathbf{Q_{lsdp}}$.

(4) It is proven in Step (1) of this proof that the existence of a weak Gaussian stochastic realization implies that there exists a matrix $Q_x \in \mathbf{Q_{lsdp}}$. In Step (2) it is proven that if there exists a matrix $Q_x \in \mathbf{Q_{lsdp}}$ then there exists a weak Gaussian stochastic realization.

The existence of the minimal and maximal state variance matrices Q^-, $Q^+ \in \mathbf{Q_{lsdp}}$ follows from Theorem 23.4.2 and from the fact that W is a covariance function.

(5) It will be argued that $\sigma \in$ GStocS is a minimal weak Gaussian realization if and only if the covariance realization (6.3) is a minimal covariance realization. Assume that the stochastic realization is minimal. If the covariance function is not minimal then there exists a covariance realization of dimension $n_1 < n_x$. From the steps (2) and (3) of the proof above then follows that there exists a weak stochastic realization with dimension of the state space equal to $n_1 < n_x$, which contradicts that the system is a minimal stochastic realization. Assume that the covariance realization is minimal. If the stochastic realization is not minimal, then there exists a stochastic realization with state space dimension $n_1 < n_x$. Then from (1) above follows that W is a covariance function with a realization of dimension $n_1 < n_x$, contradicting that this realization is minimal.
(6) Let a stochastic realization have the following forward and backward representations

$$\begin{aligned} &x(t+1) = A_f x(t) + N_f v_f(t),\ y(t) = C_f x(t) + N_f v_f(t), \\ &x(t-1) = A_b x(t) + M_b v_b(t),\ y(t-1) = C_b x(t) + N_b v_b(t). \end{aligned}$$

Without loss of generality this Gaussian system representation may be reduced so that (A_f, M_f) and (A_b, M_b) are supportable pairs while y stays a weak stochastic realization of the considered process. See Example 6.5.1 for a detailed explanation. Then $x(t) \in G(0, Q_x)$ with $Q_x = Q_x^T \succ 0$. Then, by Theorem 4.4.5 and Theorem 4.4.7, for $t > 0$,

$$\begin{aligned} W(t) &= C_f (A_f)^{t-1} [A_f Q_x (C_f)^T + M_f Q_{f,v} N_f^T] \\ &= [A_b Q_x (C_b)^T + M_b Q_{b,v} N_b^T]^T ((A_b)^T)^{t-1} (C_b)^T. \end{aligned}$$

From Theorem 4.5.2 follows that,

$$F = A_f = Q_x (A_b)^T Q_x^{-1},\ G = A_f Q_x (C_f)^T + M Q_{f,v} N_f^T = Q_x (C_b)^T,\ H = C_f.$$

Then (F, G) is a controllable pair
if and only if $(A_f, A_f Q_x (C_f)^T + M_f Q_{f,v} N_f^T)$ is a controllable pair,
if and only if $(Q_x (A_b)^T Q_x^{-1}, Q_x (C_b)^T)$ is a controllable pair,
if and only if $((A_b)^T, (C_b)^T)$ is a controllable pair,
if and only if (A_b, C_b) is an observable pair.
Thus $\sigma \in$ GStocS is a minimal weak Gaussian stochastic realization,
if and only if W has the minimal covariance realization specified in (6.3) by (5) above,
if and only if (F, G) is a controllable pair and (F, H) is an observable pair by linear system theory,
if and only if (A_f, C_f) is an observable pair and (A_b, C_b) is an observable pair, by the above argument,
if and only if σ is stochastically observable and stochastically co-observable, by Propositions 4.6.8 and 4.6.9. Hence 6.4.3.(c) is proven.

(7) Statement (d.1) is verified by an elementary calculation. Consider Statement (d.2) and the map $c_{lsp} : \mathbf{Q_{lsdp}} \mapsto WGSRP_{min}$
with lsp $= \{n_y, n_x, n_y, F, G, H, J\} \in \text{LSP}_{min}$ fixed. It follows from the Steps (2), (3),

and (6) above that this map is well defined. To prove that c_{lsp} is surjective let gsp $= \{n_y, n_x, A, C, Q_v\} \in WGSRP_{min}$. It follows from (1) above that there exists a $Q_x \in \mathbb{R}_{pds}^{n_x \times n_x}$. As in Step (3) of the proof, one establishes that Q_x corresponds to a gsp via c_{lsp}, or that gsp $= c_{lsp}(Q_x)$. Thus c_{lsp} is surjective. To prove that c_{lsp} is injective let,

$$c_{lsp}(Q_1) = c_{lsp}(Q_2) = \{n_y, n_x, A, C, M, N\} \in WGSRP_{min}.$$

Then

$$\begin{aligned}
\begin{pmatrix} M \\ N \end{pmatrix} \begin{pmatrix} M \\ N \end{pmatrix}^T &= Q_{v,d}(Q_1) = Q_{v,d}(Q_2) \\
&= \begin{pmatrix} Q_2 - FQ_2F^T & G - FQ_2H^T \\ G^T - HQ_2F^T & J + J^T - HQ_2H^T \end{pmatrix}, \Rightarrow \\
Q_1 - FQ_1F^T &= Q_2 - FQ_2F^T \ \Rightarrow \ (Q_1 - Q_2) = F(Q_1 - Q_2)F^T.
\end{aligned}$$

This, $\text{spec}(F) = \text{spec}(A) \subset \mathrm{D}_o$, and Theorem 22.1.2 imply that $Q_1 - Q_2 = 0$.
(8) Finally (d) follows from the Steps (2) and (3) above. □

6.7 Realization Procedures

The purpose of this subsection is to present several procedures for the weak stochastic realization problem for stationary Gaussian processes.

The motivation for the stochastic realization problem is the modelling of data by stochastic systems. This modelling problem is also considered to be part of the topic of system identification. System identification procedures have been developed primarily for single-output processes. The system identification problem for multi-output processes is still far from being completely solved. There is an effort to exploit stochastic realization theory for the system identification problem. Hence the motivation for the discussion of the stochastic realization procedure.

The stochastic realization procedure for a weak Gaussian stochastic realization is presented in Procedure 6.4.4. The essential steps of the stochastic realization procedures are the covariance realization of Step (1) and the variance selection of Step (2). The covariance realization of Step (1) may be executed by applying a realization procedure for deterministic finite-dimensional linear systems. Such procedures may be found in books treating linear system theory. The variance selection of Step (2) is particular for the stochastic realization problem. In the following discussion attention is therefore restricted to this variance selection.

Results on the structure of the set $\mathbf{Q_{lsdp}}$ may be found in Chapter 24. There it is proven that there exist matrices $Q^-, Q^+ \in \mathbf{Q_{lsdp}}$ such that for all $Q \in \mathbf{Q_{lsdp}}$, $Q^- \preceq Q \preceq Q^+$. Furthermore all elements of the set $\mathbf{Q_{lsdp}}$ can be constructed via the procedure that is indicated in Section 24.7. Below attention is concentrated on procedures for the computation of Q_x^-. The stochastic realization corresponding to Q_x^- will be argued to be related to the Kalman filter.

If one has determined the matrix $Q_x^- \in \mathbb{R}_{pd}^{n_x \times n_x}$ then the parameters of the associated stochastic realization may be simplified as follows. Because
$Q_x^- \in \partial \mathbf{Q}_{\mathbf{lsdp},\mathbf{r},\mathbf{s}}$, $\text{rank}(Q_{v,d}(Q_x^-)) = n_y$, hence one may write,

$$Q_{v,d}(Q_x^-) = \begin{pmatrix} KK^T & KN^T \\ NK^T & NN^T \end{pmatrix} = \begin{pmatrix} K \\ N \end{pmatrix} \begin{pmatrix} K^T & N \end{pmatrix} = \begin{pmatrix} M^- \\ N^- \end{pmatrix} \begin{pmatrix} M^- \\ N^- \end{pmatrix}^T,$$
$$\text{rank}(N^-(N^-)^T) = n_y.$$

The forward representation of the stochastic realization associated with Q^- can then be written as

$$\hat{x}(t+1) = A\hat{x}(t) + M^- \bar{v}(t), \tag{6.13}$$
$$y(t) = C\hat{x}(t) + N^- \bar{v}(t), \; \bar{v}(t) \in G(0,I). \tag{6.14}$$

To obtain the matrix Q_x^- one can use Procedure 6.7.1.

Procedure 6.7.1 . *The procedure for the computation of the minimal variance matrix $Q_x^- \in \mathbb{R}_{pd}^{n_x \times n_x}$.*
Declarations: n_x, $n_y \in \mathbb{Z}_+$, F, Q_x, $Q_x^- \in \mathbb{R}^{n_x \times n_x}$, $G \in \mathbb{R}^{n_x \times n_y}$, $H \in \mathbb{R}^{n_y \times n_x}$, $J \in \mathbb{R}^{n_y \times n_y}$ with $J = J^T > 0$.
Data: n_x, n_y, F, G, H, J.

1. *Set $Q_0 = 0$.*
2. *For $k = 1$ step 1 to ∞ do*

$$Q(k+1) = FQ(k)F^T + [G - FQ(k)H^T][J + J^T - HQ(k)H^T]^{-1}[G - FQ(k)H^T]^T. \tag{6.15}$$

3. *Set $Q_x^- = Q(\infty)$.*

In practice one stops the steps of the procedure if the sequence has converged upto the preset accuracy.

The procedure above may be deduced from Proposition 24.6.2 by observing that $Q^- \in \mathbf{Q}_{\mathbf{lsdp}}$ implies that $(Q^-)^{-1} = Q_d^+ \in \mathbf{Q}_{\mathbf{lsp}}$.

One may pose the following questions for the above procedure:

- What are the stationary points of the recursion (6.15)?;
- What is the domain of attraction of Q^-?;
- What is the convergence speed of the procedure?

There are reports in the literature [30, 89] that this iterative procedure converges slowly with particular data. To overcome the difficulties with the iterative procedure for $Q^- \in \mathbf{Q}_{\mathbf{lsdp}}$, a noniterative procedure has been developed by P. Van Dooren [30].

Procedure 6.7.2 *A noniterative procedure for the computation of $Q_x^- \in \mathbf{Q}_{\mathbf{lsdp}}$.*
Declarations: n_x, $n_y \in \mathbb{Z}_+$, F, $Q_x^- \in \mathbb{R}^{n_x \times n_x}$, $G \in \mathbb{R}^{n_x \times n_y}$, $H \in \mathbb{R}^{n_y \times n_x}$, $J \in \mathbb{R}^{n_y \times n_y}$, $\Lambda, X_1, X_2 \in C^{n_x \times n_x}$.

1. *Check if $J = J^T > 0$.*

2. *Determine* $X_1, X_2, \Lambda \in \mathbb{C}^{n_x \times n_x}$ *such that* $\operatorname{spec}(\Lambda) \subset \mathrm{D}_o$,

$$\begin{pmatrix} (F - G(J+J^T)^{-1}H)^T & 0 \\ -G(J+J^T)^{-1}G^T & I \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} I & -H^T(J+J^T)^{-1}H \\ 0 & F - G(J+J^T)^{-1}H \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \Lambda, \tag{6.16}$$

   in which Λ *is in Jordan canonical form with only those generalized eigenvalues associated with (6.16) that are in* D_o.
3. *Set* $Q_x^- = X_2 X_1^{-1}$.

That Procedure 6.7.2 indeed produces $Q^- \in \mathbf{Q}_{\mathbf{lsdp}}$ is proven in Theorem 22.3.11. There the procedure is proven for the dual case, for the set $\mathbf{Q}_{\mathbf{lsp}}$, and the reader obtains the above procedure by applying the transformation lsp ↦ lsdp. A numerical stable procedure for step 2 of the above procedure, in fact for a modification of it, may be found in [30].

6.8 State-Space Reduction of a Gaussian System

Of interest to modeling by Gaussian system representations is the problem of how to reduce a Gaussian system representation to one of a lower dimension which preferably is stochastically observable, stochastically co-observable, and supportable, hence a minimal stochastic realization. The next procedure describes how to obtain such a representation.

Procedure 6.8.1 *Consider a time-invariant Gaussian system representation with the parameter value* $gsp = (n_y, n_x, n_v, A, C, M, N)$. *Assume that the transition matrix is exponentially stable. Suppose that the Gaussian system is* not *a minimal realization of its output process.*

Construct a minimal stochastic realization of the output process of this systems by executing the following steps.

1. *Reduce the Gaussian system in state-space dimension from* n_x *to* $n_1 \leq n_x$ *such that the reduced Gaussian system is such that the corresponding system matrices* (A_1, M_1) *are a supportable pair. This can be done by the decomposition of the system matrices to the Kalman control-canonical form and a reduction step, see Proposition 21.2.9. Denote the parameters of the new forward Gaussian system representation by* $(n_y, n_1, n_v, A_1, C_1, M_1, N_1)$.
2. *Reduce the Gaussian system in state-space dimension from* n_1 *to* $n_2 \leq n_1$ *such that the reduced Gaussian system is such that the corresponding system matrices* (A, C) *form an observable pair. This can be done by a decomposition of the system matrices to the Kalman observable-canonical form and a reduction step. See, Proposition 21.3.6. Denote the parameters of the new forward Gaussian system representation by* $(n_y, n_2, n_v, A_2, C_2, M_2, N_2)$.

3. *Transform the system matrices from the forward representation* $(n_y, n_2, n_v, A_2, C_2, M_2, N_2)$*. to those of the associated backward representation,* (A_b, C_b, M_b, N_b)*, see Theorem 4.5.2 for the transformation.*

   $$\begin{aligned} Q_2 &= A_2 Q_2 A_2^T + M_2 M_2^T, \;\; Q_2 \in \mathbb{R}_{spds}^{n_2 \times n_2}, \\ A_b &= Q_2 A_2^T Q_2^{-1}, \;\; C_b = C_{2,f} Q_2 A_2^T Q_2^{-1} + N_{2,b} M_{2,b}^T Q_2^{-1}. \end{aligned}$$

   If possible, reduce the backward Gaussian system representation in state-space dimension from n_2 *to* $n_3 \leq n_2$ *such that the system matrices of the reduced system* $(n_y, n_3, n_v, A_{b,r}, C_{b,r}, M_{b,r}, N_{b,r})$ *are such that* $(A_{b,r}, C_{b,r})$ *is an observable pair. Transform the system matrices from the backward representation to the forward representation according to the formulas of Theorem 4.5.2.*

   $$\begin{aligned} Q_3 &= A_{b,r} Q_3 A_{b,r}^T + M_{b,r} M_{b,r}^T, \; Q_3 \in \mathbb{R}_{spds}^{n_3 \times n_3}, \\ A_3 &= Q_3 A_{b,r}^T Q_3^{-1}, \;\; C_3 = C_{b,r} Q_3 A_{b,r}^T Q_3^{-1} + N_r Q_{b,v} M_{b,r}^T Q_3^{-1}. \end{aligned}$$

   Denote the parameters of the new forward Gaussian system representation by $(n_y, n_3, n_v, A_3, C_3, M_3, N_3)$*.*
4. *Output the system matrices of the resulting forward representation of the Gaussian system* $(n_y, n_{3,x}, n_v, A_3, C_3, M_3, N_3)$*. That system is a minimal weak-Gaussian stochastic realization of the output of the considered Gaussian system.*

Proposition 6.8.2. *Consider a time-invariant Gaussian system with an exponentially stable system matrix. Procedure 6.8.1 is correct and computes a minimal weak Gaussian stochastic realization of the output of the considered Gaussian system. Thus the covariance function of the reduced Gaussian system equals the covariance function of the output of the original Gaussian system.*

Proof. It follows along the lines of Theorem 21.8.9 that the system matrices of the reduced system, $(n_{3,x}, n_y, n_v, A_3, C_3, M_3, N_3)$ are such that (A_3, C_3) is an observable pair hence that the system is stochastically observable, and $(Q_{3,x} A_3 Q_{1,x}^{-1}), C_3 Q_{3,x}^{-1}))$ is an observable pair hence that the reduced system is stochastically co-observable. Because by Step 3 of the procedure $Q_3 \in \mathbb{R}_{spds}^{n_3 \times n_3}$, and by the steps of the procedure that $\text{spec}(A_3) \subset \mathrm{D}_o$, it follows from Theorem 22.1.2.(d) that (A_3, M_3) is a controllable pair and thus a supportable pair.

That the reduced system then is a minimal weak Gaussian system realization follows from Theorem 6.4.3.(c). □

6.9 Special Stochastic Realizations-1

The purpose of this section is to present properties of specific Gaussian system representations as stochastic realizations of their output process. The set of weak Gaussian stochastic realizations has been described in Theorem 6.4.3. Elements of this set have received names. These names are defined below. If the observed process that is to be realized has a particular property then the associated Gaussian system, which is a stochastic realization of the observed process, will satisfy related particular conditions. The problem to be investigated in this section is to characterize these other particular conditions of a Gaussian system representation for several properties.

The set of weak Gaussian stochastic realizations has been described in Theorem 6.4.3. Elements within this set have received names. This terminology is introduced below. Recall from Def. 3.1.6 that two stochastic processes y_1, y_2 defined on $T = \mathbb{Z}$ taking values in $\mathbb{R}^{n_y}$ are said to be *equivalent* if their families of finite-dimensional distributions are identical. Also two stochastic processes $y_1, y_2 : \Omega \times T \to \mathbb{R}^{n_y}$ are said to be *modifications* of each other if,

$$P(\{\omega \in \Omega | y_1(\omega,t) = y_2(\omega,t)\}) = 1, \ \forall\, t \in T.$$

Definition 6.9.1. Consider two time-invariant Gaussian system representations,

$$\begin{aligned} gsp_1 &= \{n_y, n_{x_1}, n_{v_1}, A_1, C_1, M_1, N_1\} \in \text{GStocSP}, \\ x_1(t+1) &= A_1 x_1(t) + M_1 v_1(t), \\ y_1(t) &= C_1 x_1(t) + N_1 v_1(t), \ v_1(t) \in G(0, I_{n_{v_1}}), \\ gsp_2 &= \{n_y, n_{x_2}, n_{v_2}, A_2, C_2, M_2, N_2\} \in \text{GStocSP}, \\ x_2(t+1) &= A_2 x_2(t) + M_2 v_2(t), \\ y_2(t) &= C_2 x_2(t) + N_2 v_2(t), \ v_2(t) \in G(0, I_{n_{v_2}}), \end{aligned}$$

possibly defined on different probability spaces. These Gaussian system representations are said to be:

(a) *Weakly-equivalent stochastic realizations* if the output processes y_1, y_2 are equivalent processes. In this case both systems are stochastic realizations of the same output process in terms of distributions. The qualifier *weakly* refers to the equality of y_1, y_2 in distributions.

(b) *Strongly equivalent stochastic realizations* if the systems are defined on the same probability space, and y_1, y_2 are modifications of each other; equivalently, if $y_1(t) = y_2(t)$ *a.s.* for all $t \in T$.

(c) *Strongly equivalent systems* if they are defined on the same probability space, $n_{x_1} = n_{x_2}$, y_1, y_2 are modifications of each other, and x_1, x_2 are modifications of each other.

(d) *Strongly equivalent representations* if $n_{v_1} = n_{v_2}$, v_1, v_2 are modifications of each other, and y_1, y_2 are modifications of each other.

Definition 6.9.2. Consider a time-invariant Gaussian system representation

$$\begin{aligned} gsp &= \{n_y, n_x, n_v, A, C, M, N\} \in \text{GStocSP}, \\ x(t+1) &= Ax(t) + Mv(t), \ \text{spec}(A) \subset \mathrm{D}_o, \\ y(t) &= Cx(t) + Nv(t), \ \ v(t) \in G(0, I). \end{aligned}$$

This Gaussian system representation is said to be:

(a) *regular* if $rank(NN^T) = n_y$;

(b) *regular and square* if it is regular and $n_v = n_y$, and *regular and non-square* if it is regular and $n_v > n_y$.

(c) Assume that $\text{spec}(A) \subset \text{D}_o$. The system is called *supported* on the full state space if $x(t) \in G(0, Q_x)$, where $Q_x \in \mathbb{R}_{pds}^{n_x \times n_x}$ is the unique solution of the Lyapunov equation $Q_x = AQ_xA^T + MM^T$, then $Q_x \in \mathbb{R}_{spds}^{n_x \times n_x}$, hence $0 \prec Q_x$.

Theorem 6.4.3 implies that as a stochastic realization of its output process, the Gaussian system of the above definition satisfies,

$$W(0) = J + J^T = CQ_xC^T + NN^T; \;\Rightarrow\; \text{rank}(NN^T) = \text{rank}(J + J^T - CQ_xC^T).$$

The condition of regularity of a Gaussian system is therefore equivalent with the condition of regularity of the set $\mathbf{Q_{lsdp}}$, see 24.4.1.

Proposition 6.9.3. *Consider a time-invariant Gaussian system representation*

$$\begin{aligned} gsp &= \{n_y, n_x, n_v, A, C, M, N\} \in \text{GStocSP}, \\ x(t+1) &= Ax(t) + Mv(t), && (6.17)\\ y(t) &= Cx(t) + Nv(t), \;\; v(t) \in G(0, I). && (6.18) \end{aligned}$$

(a) This system is regular if and only if it is strongly equivalent with the Gaussian system representation,

$$\begin{aligned} x(t+1) &= Ax(t) + M_1w_1(t) + M_2w_2(t), \\ y(t) &= Cx(t) + N_2w_2(t), \; F^{x_0}, \; F_\infty^w, \; \textit{independent}, \\ w(t) &= \begin{pmatrix} w_1(t) \\ w_2(t) \end{pmatrix} \in G(0, I), \; w : \Omega \times T \to \mathbb{R}^{n_w}, \; n_w \geq n_y, \\ & M_1 \in \mathbb{R}^{n_x \times (n_w - n_y)}, \; M_2 \in \mathbb{R}^{n_x \times n_y}, \; N_2 \in \mathbb{R}^{n_y \times n_w}, \; \text{rank}(N_2) = n_y. \end{aligned}$$

w is a standard Gaussian white noise process.

(b) The system is regular and square if and only if it is a strongly equivalent system with representation,

$$\begin{aligned} gsp_1 &= \{n_y, n_x, n_y, A, C, M_2, I, Q_{w_2}\} \in \text{GStocSP}, \\ x(t+1) &= Ax(t) + M_2w_2(t), && (6.19)\\ y(t) &= Cx(t) + N_2w_2(t), \;\; w_2(t) \in G(0, I), \; w_2 : \Omega \times T \to \mathbb{R}^{n_y}. && (6.20) \end{aligned}$$

where w_2 is a standard Gaussian white noise process.

(c) Assume that $\text{spec}(A) \subset \text{D}_o$. *Then this system representation is supported on the full state space if and only if $(A, \; M)$ is a supportable pair.*

(d) If the time-invariant Gaussian system representation is not supported on the full state space, then there exists a time-invariant Gaussian system representation that is supported on the full state space and that is strongly equivalent with considered representation. It may be taken as $\{n_y, n_{x_1}, n_v, A_1, C_1, M_1, N_1\} \in \text{GStocSP}$,

$$\begin{aligned} x_1(t+1) &= A_1x_1(t) + M_1v(t), \; \text{spec}(A) \subset \text{D}_o, \\ y(t) &= C_1x_1(t) + N_1v(t), \; v(t) \in G(0, I), \\ & (A_1, M_1) \textit{ a supportable pair}. \end{aligned}$$

The pair (A_1, C_1) in (d) above may or may not be an observable pair.

Proof. (a) By definition of regularity, $\text{rank}(NN^T) = n_y$. Hence $n_v \geq n_y$. Let $L \in \mathbb{R}^{(n_v - n_y) \times n_v}$ be such that,

$$S = \begin{pmatrix} L \\ N \end{pmatrix} \in \mathbb{R}^{n_v \times n_v}, \; LN^T = 0, \; \text{rank}(S) = n_v.$$

Let $V \in \mathbb{R}^{n_v \times n_v}$ be such that $VSS^TV^T = I_{n_v}$. Define the stochastic process $w : \Omega \times T \to \mathbb{R}^{n_v}$, $w(t) = VSv(t)$. Then w is a Gaussian white noise with $w(t) \in G(0, Q_w)$, $Q_w = VSS^TV^T = I_{n_v} \succ 0$. Define compatible with S,

$$\begin{pmatrix} w_1(t) \\ w_2(t) \end{pmatrix} = w(t) = Sv(t) = \begin{pmatrix} L \\ N \end{pmatrix} v(t).$$

Then,

$$\begin{aligned} Nv(t) &= \begin{pmatrix} 0 \; I_{n_y} \end{pmatrix} Sv(t) = \begin{pmatrix} 0 \; I_{n_y} \end{pmatrix} w(t) = w_2(t), \\ Mv(t) &= MS^{-1}Sv(t) = MS^{-1} \begin{pmatrix} w_1(t) \\ w_2(t) \end{pmatrix} = \begin{pmatrix} M_1 \; M_2 \end{pmatrix} \begin{pmatrix} w_1(t) \\ w_2(t) \end{pmatrix}, \\ Q_w &= \begin{pmatrix} LL^T & 0 \\ 0 & NN^T \end{pmatrix} = \begin{pmatrix} Q_{w_1} & 0 \\ 0 & Q_{w_2} \end{pmatrix}, \text{ because } LN^T = 0. \end{aligned}$$

Thus the two components $w_1(t)$ and $w_2(t)$ are independent for every time $t \in T$, and, because they are Gaussian white noise processes, they are independent processes. One obtains the system representation,

$$\begin{aligned} x(t+1) &= Ax(t) + M_1 w_1(t) + M_2 w_2(t), \\ y(t) &= Cx(t) + w_2(t). \end{aligned}$$

Finally one transforms each of the two Gaussian white noise processes w_1 and w_2 to standard Gaussian white noise processes while transforming the matrices M_1, M_2, and N_2 correspondingly. In general this yields a nontrivial matrix N_2.
(b) This follows from (a) and the definition of a square Gaussian system ($n_v = n_y$).
(c) By the result for the invariant distribution of a Gaussian system $x(t) \in G(0, Q_x)$ with $Q_x = AQ_xA^T + MQ_vM^T$. It follows from Theorem 22.1.2.(d) and the assumption $\text{spec}(A) \subset \mathrm{D}_o$, that $Q_x \succ 0$ if and only if (A, M) is a controllable pair.
(d) According to the Kalman controllable form, Proposition 21.2.9, that there exists a transformation matrix $S \in \mathbb{R}^{n_1 \times n_1}$ such that,

$$SA_1S^{-1} = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \; SM_1 = \begin{pmatrix} M_2 \\ 0 \end{pmatrix}, \; C_1S^{-1} = \begin{pmatrix} C_2 \; C_3 \end{pmatrix}, \tag{6.21}$$

with $n_2 \in \mathbb{N}$, $A_{11} \in \mathbb{R}^{n_2 \times n_2}$, $A_{12} \in \mathbb{R}^{n_2 \times (n_1 - n_2)}$, $A_{22} \in \mathbb{R}^{(n_1 - n_2) \times (n_1 - n_2)}$, $M_2 \in \mathbb{R}^{n_2 \times m_2}$, and (A_{11}, M_1) a controllable pair. Let $A_2 = A_{11}$, $M_2 = M_1$. Then also (A_2, M_2) is a controllable pair. $\text{spec}(A_1) \subset \mathrm{D}_o$ implies that $\text{spec}(A_2) \subset \mathrm{D}_o$. Let Q_2 be the solution of

$$Q_2 = A_{22}Q_2A_{22}^T + M_2M_2^T,$$

which solution exists by $\text{spec}(A_{22}) \subset \text{D}_o$ and Theorem 22.1.2. Then (A_{22}, M_2) a controllable pair implies that $Q_2 = Q_2^T \succ 0$. Define,

$$\begin{aligned} x_2(t) &= \begin{pmatrix} I_{n_2} & 0 \end{pmatrix} Sx_1(t),\ \ x_2 : \Omega \times T \to \mathbb{R}^{n_2}; \text{ it follows from (6.21) that,} \\ Sx_1(t) &= \begin{pmatrix} x_2(t) \\ 0 \end{pmatrix} \quad \text{a.s. Then,} \\ x_2(t+1) &= \begin{pmatrix} I & 0 \end{pmatrix} Sx_1(t+1) = \begin{pmatrix} I & 0 \end{pmatrix} SA_1S^{-1}Sx_1(t) + \begin{pmatrix} I & 0 \end{pmatrix} SM_1v_2(t) \\ &= A_2x_2(t) + M_2v_2(t). \end{aligned}$$

□

Definition 6.9.4. Consider a time-invariant Gaussian system representation

$$\begin{aligned} gsp &= \{n_y, n_x, n_v, A, C, M, N\} \in \text{GStocSP}, \\ x(t+1) &= Ax(t) + Mv(t), && (6.22) \\ y(t) &= Cx(t) + Nv(t),\ \ v(t) \in G(0, I). && (6.23) \end{aligned}$$

This Gaussian system is said to be:

(a) An *output-based* stochastic realization of its output process if $n_v = \text{rank}(NN^T)$. An output-based stochastic realization is in the literature also called an internal stochastic realization.

(b) A *Kalman realization* of the associated output process if the system is regular, output-based, and such that,

$$\text{spec}(A) \subset \text{D}_o \text{ and } \text{spec}(A - MN^{-1}C) \subset \text{D}_o.$$

Proposition 6.9.5. *Consider a time-invariant Gaussian system representation*

$$\begin{aligned} gsp &= \{n_y, n_x, n_v, A, C, M, N\} \in \text{GStocSP}, \\ x(t+1) &= Ax(t) + Mv(t), \\ y(t) &= Cx(t) + Nv(t),\ \ v(t) \in G(0, I). \end{aligned}$$

(a) Assume that the system is regular. Then it is an output-based stochastic realization if and only if it is a strongly equivalent system with the Gaussian system representation,

$$\begin{aligned} gsp &= \{n_y, n_x, n_v, A, C, K, N\} \in \text{GStocSP},\ \ n_v = n_y, \\ x(t+1) &= Ax(t) + Kw(t), && (6.24) \\ y(t) &= Cx(t) + Nw(t),\ \ w(t) \in G(0, I),\ \text{rank}(N) = n_y = n_v. && (6.25) \end{aligned}$$

(b) Assume the conditions of (a). The Gaussian system representation may then be written as the linear system

$$\begin{aligned} x(t+1) &= (A - KN^{-1}C)x(t) + KN^{-1}y(t),\ x(t_0) = x_0, && (6.26) \\ w(t) &= -N^{-1}Cx(t) + N^{-1}y(t). && (6.27) \end{aligned}$$

This will be called a noise generating system of the output process y. From this representation follows that,

$$F^{x(t)} \subset (F^y_{t-1} \vee F^{x_0}),$$

for all $t \in T$, hence the state $x(t)$ is a function of the past output process and the initial state. This relation explains the term output-based stochastic realization.

(c)*Assume that the system is regular. All output based stochastic realizations of the output process y are classified in the sense of Theorem 6.4.3.(d.3) by the boundary set $\partial \mathbf{Q}_{\mathbf{lsdp},\mathbf{r},\mathbf{s}}$, which are the elements of the singular boundary points of the regular part of $\partial \mathbf{Q}_{\mathbf{lsdp},\mathbf{r}}$,*

$$\begin{aligned}\partial \mathbf{Q}_{\mathbf{lsdp},\mathbf{r},\mathbf{s}} &= \{Q \in \mathbb{R}^{n_x \times n_x}_{pds} | \ J + J^T - HQH^T \succ 0, \ D_d(Q) = 0\}, \\ D_d(Q) &= Q - FQF^T - [G - FQH^T][J + J^T - HQH^T]^{-1}[G - FQH^T]^T.\end{aligned}$$

(d)*The system representation (6.22,6.23) is a Kalman realization if and only if it is strongly equivalent to the Gaussian system representation,*

$$\begin{aligned}gsp &= \{n_y, n_x, n_y, A, C, K, N\} \in \text{GStocSP}, \\ x(t+1) &= Ax(t) + Kw(t), \\ y(t) &= Cx(t) + Nw(t), \ \ w(t) \in G(0,I), \\ &\quad \text{spec}(A) \subset \text{D}_o \ \textit{and} \ \text{spec}(A - KN^{-1}C) \subset \text{D}_o.\end{aligned}$$

All Kalman realizations of the output process y are classified in the sense of Theorem 6.4.3 by $Q^- \in \mathbf{Q}_{\mathbf{lsdp}}$. Note that for a Kalman realization, the noise generating system of (b), (6.26,6.27), is exponentially stable.

Proof. (a) Because the system is assumed to be regular, $\text{rank}(NN^T) = n_y$. If the system is output-based then $n_v = \text{rank}(NN^T) = n_y$, hence the system is square. The result then follows from 6.9.3.(b). Conversely, the Gaussian system representation with the equations (6.24) and (6.25) implies that $\text{rank}(NN^T) = n_y = n_v$, hence the stochastic realization is output-based.

(b) The Gaussian system representation follows directly from (a) using the invertibility of the matrix N and algebraic operations. By induction the σ-algebra inclusion follows.

(c) An output-based stochastic realization is such that $n_v = n_y$, $K \in \mathbb{R}^{n_x \times n_v}$, and $N \in \mathbb{R}^{n_y \times n_v}$, thus

$$\begin{pmatrix} K \\ N \end{pmatrix} \in \mathbb{R}^{(n_x+n_y)\times n_v}, \ \text{rank} \begin{pmatrix} K \\ N \end{pmatrix} = n_v = n_y, \ \text{rank} \left(\begin{pmatrix} K \\ N \end{pmatrix} \begin{pmatrix} K \\ N \end{pmatrix}^T \right) = n_v = n_y.$$

Because the system is a weak Gaussian stochastic realization of its output process,

$$Q_{v,d}(Q_x) = \begin{pmatrix} Q - FQF^T & G - FQH^T \\ G^T - HQF^T & J + J^T - HQH^T \end{pmatrix} = \begin{pmatrix} K \\ N \end{pmatrix} \begin{pmatrix} K \\ N \end{pmatrix}^T,$$

implies that $\text{rank}(Q_{v,d}(Q)) = n_y$. Hence $Q \in \partial \mathbf{Q}_{\mathbf{lsdp},\mathbf{r},\mathbf{s}}$.

(d) The strong equivalence follows from (a) and the definition of a Kalman realization. □

An overview of the different Gaussian system representations as weak Gaussian stochastic realization is provided by Table 6.1.

Term	*Subset of* $\mathbf{Q_{lsdp}}$
Regular	$\mathbf{Q_{lsdp,r}}$
Regular and output-based	$\partial \mathbf{Q_{lsdp,r,s}} \subset \mathbf{Q_{lsdp,r}}$
Kalman realization	$\{ Q^- \} \subset \partial \mathbf{Q_{lsdp,r,s}}$

Table 6.1 Table relating particular realizations to their subsets of state variance matrices.

6.10 Special Stochastic Realizations-2

The question of characterizing Gaussian system representations for particular output processes, will be answered for the properties of the output process being a Gaussian white noise process, and that of the output process being time-reversible. Stochastic realizations in balanced form will also be mentioned.

Definition 6.10.1. Consider the square time-invariant Gaussian system representation,

$$\begin{aligned} gsp &= \{n_y, n_x, n_y, A, C, M, N\} \in \text{GStocSP}, \\ x(t+1) &= Ax(t) + Mv(t), \\ y(t) &= Cx(t) + Nv(t), \ v(t) \in G(0,I), \ n_y = n_v. \end{aligned}$$

This system is called a *unitary Gaussian stochastic realization* if the output process y is a Gaussian white noise process with $y(t) \in G(0,I)$. Impose the convention that the system, considered with v as input and y as output, is a minimal linear system realization of its transfer function, or, equivalently, that (A,M) is a controllable pair and (A,C) is an observable pair. Note that this definition is different from the system being a weak Gaussian stochastic realization of the output process.

The square transfer function of a unitary Gaussian stochastic realization from v to y, $T(z) = C(zI - A)^{-1}M + N$, is called an *all pass* function in circuit theory and an *inner function* in operator theory.

Proposition 6.10.2. *Consider a time-invariant Gaussian system representation*

$$\begin{aligned} gsp &= \{n_y, n_x, n_y, A, C, M, N\} \in \text{GStocSP}, \\ x(t+1) &= Ax(t) + Mv(t), \\ y(t) &= Cx(t) + Nv(t), \ v(t) \in G(0,I), \ n_y = n_v. \end{aligned}$$

Assume that (A,M) is a controllable pair, (A,C) an observable pair, that the system representation is regular and output based, and that it satisfies $\text{spec}(A) \subset \text{D}_o$.

(a)Then the output of this system is a unitary Gaussian stochastic realization if and only if,

$$M = -AQC^T N^{-T},\ I = CQC^T + NN^T,\ \textit{where } Q = AQA^T + MM^T.$$

(b)A Gaussian system representation is a unitary Gaussian stochastic realization if and only if it is strongly equivalent with the Gaussian system representation,

$$\begin{aligned} &\{n_y, n_x, n_y, A, C, K, N\} \in \text{GStocSP}, \\ x(t+1) &= Ax(t) + Kw(t), \\ y(t) &= Cx(t) + Nw(t),\ w(t) \in G(0,I),\ N \in \mathbb{R}^{n_y \times n_y},\ \text{rank}(N) = n_y, \\ K &= -AQC^T N^{-T},\ I = CQC^T + NN^T,\ Q = AQA^T + KK^T. \end{aligned}$$

Proof. (a) As in the proof of Theorem 6.4.3 the covariance function of the output process of the Gaussian system is,

$$\begin{aligned} W(t) &= \begin{cases} CA^{t-1}G, & t > 0, \\ I, & t = 0, \end{cases} \\ G &= AQC^T + MN^T,\ J + J^T = I = CQC^T + NN^T,\ Q = AQA^T + MM^T. \end{aligned}$$

The discrete-time Lyapunov equation has a unique solution Q because $\text{spec}(A) \subset \text{D}_o$. The output process is a unitary Gaussian stochastic realization if and only if $W(t) = CA^{t-1}G = 0$, for all $t > 0$ and $W(0) = I$. Because (A,C) is an observable pair and $\text{spec}(A) \subset \text{D}_o$, these relations are equivalent with

$$0 = G = AQC^T + MN^T, \quad I = CQC^T + NN^T. \tag{6.28}$$

Because the stochastic realization is square, regular, and output-based, it follows that $n_v = n_y$, $\text{rank}(N) = n_y$. Hence N is invertible. From Equation (6.28) follows that $M = -AQC^T N^{-T}$. Conversely, if this relation holds and $I = CQC^T + NN^T$, then $G = 0$ and the output process is standard Gaussian white noise.
(b) This follows from Proposition 6.9.5.(a) and from (a). □

Realization of a Time-Reversible Output Process

The problem to be considered in this subsection is to derive the structure of a time-invariant Gaussian system that is a weak Gaussian stochastic realization of a time-reversible stationary Gaussian process.

Consider a stationary Gaussian process $y : \Omega \times T \to \mathbb{R}^{n_y}$ on a probability space (Ω, F, P) and the time index set $T = \mathbb{N}$. Assume that the mean value function of the process is zero. Recall from Definition 3.3.4 that a stochastic process y is called time reserversible if for all $m \in \mathbb{Z}_+$, t, $t_1, \ldots, t_m \in T$ the joint distribution of $(y(t_1), \ldots, y(t_m))$ equals the joint distribution of $(y(t - t_1), \ldots, y(t - t_m))$. Moreover, recall from Proposition 3.4.8 that a stationary Gaussian process with zero mean value function and covariance function $W : T \to \mathbb{R}^{n_y \times n_y}$ is time-reversible if and only if for all $t \in T$ $W(t) = W(-t)$, if and only if $W(t) = W(t)^T$ for all $t \in T$. A scalar ($n_y = 1$) stationary Gaussian process is always time-reversible. The question to be pursued is whether a weak Gaussian stochastic realization of which the output process is time-reversible, has a special structure.

Theorem 6.10.3. *(a)Consider a time-invariant Gaussian system representation,*

$$\begin{aligned} x(t+1) &= Ax(t) + Mv(t),\ x_0 \in G(0,Q), \\ y(t) &= Cx(t) + Nv(t),\ v(t) \in G(0,I), \\ Q &= AQA^T + MQ_vM^T, \end{aligned}$$

that is a minimal weak-Gaussian stochastic realization. This system has a time-reversible output process if and only if there exists a nonsingular matrix,

$$L \in \mathbb{R}^{n_x \times n_x},\ LA = A^T L,\ C = [AQC^T + MN^T]^T L.$$

If such an L exists then it satisfies $L = L^T$.

(b)Consider a stationary Gaussian process with a zero mean value function and a covariance function that admits a finite-dimensional covariance realization as defined in Theorem 6.4.3.(a), and moreover satisfies $\lim_{t\to\infty} W(t) = 0$ *and* $W(0) \succ 0$. *If this process is time-reversible then there exists a weak Gaussian stochastic realization of the form,*

$$\begin{pmatrix} x_+(t+1) \\ x_-(t+1) \end{pmatrix} = \begin{pmatrix} D_3 & A_{12} \\ -A_{12}^T & D_4 \end{pmatrix} \begin{pmatrix} x_+(t) \\ x_-(t) \end{pmatrix} + Mv(t), \tag{6.29}$$

$$y(t) = \begin{pmatrix} C_1 & C_2 \end{pmatrix} \begin{pmatrix} x_+(t) \\ x_-(t) \end{pmatrix} + Nv(t),\ v(t) \in G(0,I). \tag{6.30}$$

where n_{x_+}, $n_{x_-} \in \mathbb{N}$, $n_{x_+} + n_{x_-} = n_x$, $x_+ : \Omega \times T \to R^{n_{x+}}$, $x_- : \Omega \times T \to R^{n_{x-}}$, the matrices D_3, D_4, A_{12}, C_1, C_2, M, N are compatible with this decomposition, such that D_2 is a signature matrix,

$$\begin{aligned} D_2 &= \begin{pmatrix} I_{n_{x_+}} & 0 \\ 0 & -I_{n_{x_-}} \end{pmatrix} \in \mathbb{R}^{n_x \times n_x},\ D_3 \in \mathbb{R}_{diag}^{n_{x_+} \times n_{x_+}},\ D_4 \in \mathbb{R}_{diag}^{n_{x_-} \times n_{x_-}}, \\ A &= \begin{pmatrix} D_3 & A_{12} \\ -A_{12}^T & D_4 \end{pmatrix},\ \text{spec}(A) \subset \mathrm{D}_o, \\ &\begin{pmatrix} M \\ N \end{pmatrix} \begin{pmatrix} M \\ N \end{pmatrix}^T = \begin{pmatrix} I - AA^T & (D_2 - A)C^T \\ C(D_2 - A^T) & J + J^T - CC^T \end{pmatrix}. \end{aligned}$$

Conversely, a Gaussian system of the form (6.29) and (6.30) with the conditions specified has a time reversible output process.

(c)Procedure 6.10.4 constructs from a covariance function satisfying the conditions of (b) the stochastic realization (6.29) and (6.30).

An interpretation of Theorem 6.10.3 follows. Recall from analysis that there are even and odd functions. An example of an even function is $f(t) = \cos(t) = \cos(-t) = f(-t)$, and one of an odd function $g(t) = \sin(t) = -\sin(-t) = -g(-t)$. Next consider the linear system,

$$\sigma = \left\{ \begin{array}{l} w : T \rightarrow R^{n_u+n_y} | w = \begin{pmatrix} u \\ y \end{pmatrix}, \\ \exists\, x : T \rightarrow \mathbb{R}^n, \;\; (6.32)\ \&\ (6.33) \text{ hold}, \end{array} \right\} \tag{6.31}$$

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{6.32}$$

$$y(t) = Cx(t) + Du(t), \tag{6.33}$$

and the operator $R : \Sigma \rightarrow \Sigma$, $(Rw)(t) = w(-t)$. Let $\Sigma_R = \{w \in W^T | Rw \in \Sigma\}$. The system Σ is called *time reversible* if $\Sigma = \Sigma_R$. If the system is a minimal realization and time reversible then with respect to some basis it may be represented by,

$$\frac{d}{dt}\begin{pmatrix} x_+(t) \\ x_-(t) \end{pmatrix} = \begin{pmatrix} 0 & A_+ \\ A_- & 0 \end{pmatrix} \begin{pmatrix} x_+(t) \\ x_-(t) \end{pmatrix} + \begin{pmatrix} 0 \\ B_+ \end{pmatrix} u(t), \tag{6.34}$$

$$y(t) = \begin{pmatrix} C_+ & 0 \end{pmatrix} \begin{pmatrix} x_+(t) \\ x_-(t) \end{pmatrix} + Du(t). \tag{6.35}$$

If the matrix S is a signature matrix compatible with A and if,

$$S = \begin{pmatrix} I_{n_x^+} & 0 \\ 0 & -I_{n_x^-} \end{pmatrix},$$

$$\overline{x}(t) = \begin{pmatrix} \overline{x}_+(t) \\ \overline{x}_-(t) \end{pmatrix} = Sx(-t) = S\begin{pmatrix} x_+(-t) \\ x_-(-t) \end{pmatrix} = \begin{pmatrix} x_+(-t) \\ -x_-(-t) \end{pmatrix},$$

then the system $\overline{\Sigma}$ with $\overline{x}$, $\overline{y}(t) = y(-t)$, $\overline{u}(t) = u(-t)$ satisfies also (6.34) and (6.35). One may call x_+ the *even state components* and x_- the *odd state components*.

Consider next the Gaussian system (6.29) and (6.30). Let

$$\bar{x}(t) = \begin{pmatrix} \overline{x}_+(t) \\ \overline{x}_-(t) \end{pmatrix} = D_2 \begin{pmatrix} x_+(-t) \\ x_-(-t) \end{pmatrix} = \begin{pmatrix} x_+(-t) \\ -x_-(-t) \end{pmatrix},$$

$$\overline{y}(t) = y(-t).$$

Using the relation $D_2A = A^TD_2$ from the proof presented below, the process $\overline{y}$ has the realization,

$$\bar{x}(t+1) = A^T\bar{x}(t) + D_2M\bar{v}(t), \tag{6.36}$$

$$\bar{y}(t) = CD_2\bar{x}(t) + N\bar{v}(t), \tag{6.37}$$

with $\bar{v}(t) = v(-t)$. This is not the backwards representation as defined in Section 4.5. Because the output process y is time-reversible the pair (6.36) and (6.37) is also a stochastic realization of y.

Procedure 6.10.4 Computation of the time-reversible weak Gaussian stochastic realization of Theorem 6.10.3 from a covariance function
Data: $n_y \in Z_+$, $W : Z \rightarrow \mathbb{R}^{n_y \times n_y}$.

1. *Construct a covariance realization* $lsp = \{n_y,\ n_x,\ n_y,\ F,\ G,\ H,\ J\} \in \text{LSP}_{min}$ *such that,*

   $$W(0) = J,\ W(t) = HF^{t-1}G,\ \ t \geq 1.$$

2. *Determine $S \in \mathbb{R}^{n_x \times n_x}_{nsng}$ nonsingular such that $SF = F^T S$ and $H = G^T S$. See realization theory for linear time-invariant systems for a procedure. Then also $S = S^T$.*
3. *Determine $Q \in \mathbb{R}^{n_x \times n_x}$ such that $Q^{-1} = SQS$. See Section 17.4 in the Subsection Contra Gradient Transform, for information on a procedure.*
4. *Determine the decompositions*

$$\begin{aligned}
Q &= U_1 D_1 U_1^T,\ U_1 \in \mathbb{R}^{n_x \times n_x}_{ortg}\ D_1 \in \mathbb{R}^{n_x \times n_x}_{diag,+}, \\
D_1^{\frac{1}{2}} U_1^T S U_1 D_1^{\frac{1}{2}} &= U_2 D_2 U_2^T,\ U_2 \in \mathbb{R}^{n_x \times n_x}_{ortg} \\
D_2 &= \text{Block} - \text{diag}\left(I_{n_1}\ -I_{n_2} \right) \in \mathbb{R}^{n_x \times n_x}\ n_1,\ n_2 \in \mathbb{N},\ n_1 + n_2 = n_x; \\
T &= U_1 D_1^{\frac{1}{2}} U_2 \ \Rightarrow\ Q = TT^T,\ S = T^{-T} D_2 T^{-1}.
\end{aligned}$$

5. *Define $F_2 = T^{-1} F_1 T$, $G_2 = T^{-1} G_1$, $H_2 = H_1 T$.*
6. *Decompose compatible with D_2*

$$F_2 = \begin{pmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{pmatrix},\ F_{11} = U_3 D_3 U_3^T,\ F_{22} = U_4 D_4 U_4^T.$$

7. *Define,*

$$\begin{aligned}
T_1 &= \text{Block} - \text{diag}\left(U_3^T\ U_4^T \right), \\
A &= T_1 F_2 T_1^{-1} = \begin{pmatrix} D_3 & A_{12} \\ -A_{12}^T & D_4 \end{pmatrix},\ C = H_2 T_1^{-1}, \\
Q_{v,11} &= I - AA^T, Q_{v,12} = (D_2 - A)C^T,\ Q_{v,22} = J - CC^T, \\
\begin{pmatrix} M \\ N \end{pmatrix} \begin{pmatrix} M \\ N \end{pmatrix}^T &= Q_v = \begin{pmatrix} Q_{v,11} & Q_{v,12} \\ Q_{v,12}^T & Q_{v,22} \end{pmatrix},\ M \in \mathbb{R}^{n_x \times (n_x \times n_y)},\ N \in \mathbb{R}^{n_y \times (n_x \times n_y)}.
\end{aligned}$$

Then $\{n_y,\ n_x,\ n_x + n_y,\ A,\ C,\ M,\ N\}$ are the parameters of a weak Gaussian stochastic realization with a time-reversible output process.

The proof of Theorem 6.10.3 is based on the following intermediate result.

Proposition 6.10.5. *Let $y : \Omega \times T \to \mathbb{R}^{n_y}$ be a stationary Gaussian process with zero mean value function and covariance function $W : T \to \mathbb{R}^{n_y \times n_y}$. Assume that W admits a minimal covariance realization of the form,*

$$\begin{aligned}
lsp &= \{n_y,\ n_x,\ n_y,\ F,\ G,\ H,\ J\} \in \text{LSP}_{min}, \\
W(t) &= \begin{cases} HF^{t-1}G, & t > 0, \\ J + J^T, & t = 0, \\ G^T (F^T)^{-t-1} H^T, & t < 0. \end{cases}
\end{aligned}$$

Then the process y is time-reversible if and only if there exists a non-singular matrix $S \in \mathbb{R}^{n_x \times n_x}$ such that $F = S^{-1} F^T S$ and $H = G^T S$. Then the relation $S = S^T$ holds.

Proof. By the results recalled above, the process y is time-reversible if and only if for all $t \in T$ $W(t) = W(t)^T$, if and only if, $HF^{t-1}G = G^T (F^T)^{t-1} H^T$, for all $t \in T_+ = T \cap \mathbb{Z}_+$, if and only if there exists an unique non-singular matrix $S \in \mathbb{R}^{n_x \times n_x}$ such that $F = S^{-1}F^TS$, $H = G^TS$, $G = S^{-1}H^T$. The latter characterization, in particular the existence of a unique S, follows from realization theory for linear systems and the assumption that $\{n_y,\, n_x,\, n_y,\, F,\, G,\, H,\, J\}$ represents a minimal covariance realization. If in the above formulas, the matrix S is replaced by the matrix S^T then

$$HF^{t-1}G = G^TS^T(S^{-T}F^TS^T)^{t-1}S^{-T}H^T = G^T(F^T)^{t-1}H^T.$$

Hence the matrix S^T achieves the same property and by uniqueness $S = S^T$. Then the condition $H = G^TS = G^TS^T$ is identical to $G = S^{-1}H^TS$ and the latter condition may therefore be omitted. □

Proof. Of Theorem 6.10.3. (a) Let $G = AQC^T + MN^T$. By Theorem 4.4.5 the covariance function of the system is specified by,

$$W(t) = \begin{cases} CA^{t-1}G, & t > 0, \\ CQC^T + NN^T, & t = 0. \end{cases}$$

The assumptions that $SA = A^TS$ and $C = G^TS$, and Proposition 3.4.8, then imply that the output process is time reversible. Conversely, if the output process is time reversible then by Proposition 3.4.8 there exists a nonsingular $S \in \mathbb{R}^{n_x \times n_x}$ such that $SA = A^TS$ and $C = G^TS$.

(b) & (c). (1) It will be shown that Procedure 6.10.4 constructs the stochastic realization (6.29) and (6.30).

(2) Let $W : T \to \mathbb{R}^{n_y \times n_y}$ be the covariance function of a time reversible stationary Gaussian process. Because W admits by assumption of minimal covariance realization, there exists $p_1 = \{n_y,\, n_x,\, n_y,\, F,\, G,\, H,\, J\} \in L\Sigma P_{min}$ such that,

$$W(t) = \begin{cases} HF^{t-1}G, & \text{for } t \geq 0, \\ J + J^T, & \text{for } t = 0. \end{cases}$$

(3) By Proposition 6.10.5 there exists a nonsingular $S \in \mathbb{R}^{n_x \times n_x}$ such that $SF = F^TS$ and $H = G^TS$. Moreover, $S = S^T$.

(4) Let

$$p_1 = \{n_y, n_x, n_y,\, F,\, G,\, H,\, J\}, p_2 = \{n_y,\, n_x, n_y,\, F^T,\, H^T,\, G^T,\, J\}.$$

Then $p_2 = \bar{p}_1$ and $p_1 = \bar{p}_2$. Let $Q_1 \in \mathbf{Q}_{dp_1}$, $Q_1 \succ 0$. By Proposition 24.2.4, $Q_1^{-1} \in \mathbf{Q}_{p_1}$. Note that,

$$F_1^T = SF_1S^{-1},\; G_1^T = H_1S^{-1},\; H_1^T = SG_1.$$

Then by Proposition 24.2.3 $\mathbf{Q}_{p_1} = S\mathbf{Q}_{p_2}S$. Note that then $Q_1^{-1} \in \mathbf{Q}_{p_1} = S\mathbf{Q}_{p_2}S = S\mathbf{Q}_{dp_1}S$, or $S^{-1}Q_1^{-1}S^{-1} \in \mathbf{Q}_{dp_1}$. Consider now the map $f : Q_{\bar{p}_1} \to Q_{\bar{p}_1}$ $f(Q_1) = S^{-1}Q_1^{-1}S^{-1}$. Then the properties that $Q_{\bar{p}_1}$ is compact and convex, that the map f is continuous, and the fixed point theorem, imply that there exists a $Q \in Q_{\bar{p}_1}$ such that $Q = S^{-1}Q^{-1}S^{-1}$, or $Q^{-1} = SQS$.

(5) As in step 4 of the procedure, let,

$$Q = U_1 D_1 U_1^T, L = D_1^{\frac{1}{2}} U_1^T S U_1 D_1^{\frac{1}{2}}.$$

Note that then $L = L^T$ and that $Q^{-1} = SQS$ or, equivalently, $U_1 D_1^{-1} U_1^T = S U_1 D_1 U_1^T S$ implies that $I = L^2$. Let $L = U_2 D_2 U_2^T$ with U_2 orthogonal and D_2 diagonal. Then $L^2 = I$ and $L = L^T$ imply that $D_2 \in \mathbb{R}^{n_x \times n_x}$ is a signature matrix. By performing a permutation one concludes that D_2 may be decomposed as,

$$\begin{aligned} D_2 &= \text{Block} - \text{diag}\left(I_{n_1} \middle| -I_{n_2} \right) \;\; n_1,\, n_2 \in N,\; n_1 + n_2 = n_x;\; T = U_1 D_1^{\frac{1}{2}} U_2; \\ TT^T &= U_1 D_1 U_1^T = Q,\; T^{-T} D_2 T^{-1} = U_1 D_1^{-\frac{1}{2}} U_2 D_2 U_2^T D_1^{-\frac{1}{2}} U_1^T = S. \end{aligned}$$

(6) Perform a state space transformation by T. Then

$$\begin{aligned} F_2 &= T^{-1} F_1 T, G_2 = T^{-1} G_1, \\ H_2 &= H_1 T,\; Q_2 = T^{-1} Q_1 T^{-T} = T^{-1} T T^T T^{-T} = I. \text{ Note that} \\ SF_1 &= F_1^T S = T^{-T} D_2 T^{-1} F_1 = F_1^T T^{-T} D_2 T^{-1} \;\Rightarrow\; D_2 F_2 = F_2^T D_2, \\ H_2 &= H_1 T = G_1^T S T = G_1^T T^{-T} D_2 T^{-1} T = G_2^T D_2. \end{aligned}$$

(7) Decompose,

$$F_2 = \begin{pmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{pmatrix}$$

compatible with the decomposition of D_2. The relation $D_2 F_2 = F_2^T D_2$ then implies that $F_{11} = F_{11}^T$, $F_{22} = F_{22}^T$, and $F_{21} = -F_{12}^T$. Decompose, $F_{11} = U_3 D_3 U_3^T$ and $F_{22} = U_4 D_4 U_4^T$, with $U_3,\ U_4$ orthogonal and $D_3,\ D_4$ diagonal.
(8) Define,

$$T_1 = \text{Block} - \text{diag}\left(U_3^T \;\; U_4^T \right);\; T_1 D_2 T_1^T = D_2,\; A = T_1 F_2 T_1^{-1} = \begin{pmatrix} D_3 & A_{12} \\ -A_{12}^T & D_4 \end{pmatrix},$$

for a matrix $A_{12} \in R^{n_1 \times n_2}$, while $C := H_2 T_1^{-1}$ satisfies

$$\begin{aligned} C &= H_2 T_1^{-1} = G_2^T D_2 T_1^{-1} = G_2^T T_1^{-1} D_2 = G_3^T D_2, \\ Q_3 &= T_1 Q_2 T_1^T = T_1 T_1^T = I. \end{aligned}$$

(9) Let $Q_{v,11}, Q_{v,12}, Q_{v,22}, Q_v, M, N$ be as defined in 6.10.3. Then

$$\begin{aligned} M Q_v M^T &= Q_3 - A Q_3 A^T = I - A A^T = Q_{v,11}, \\ N Q_v N^T &= J - C Q_3 C^T = J - C C^T = Q_{v,22}, \\ M Q_v N^T &= G_3 - A Q_3 C^T = (D_2 - A) C^T = Q_{v,12}, \end{aligned}$$

because $C = G_3^T D_2 \Leftrightarrow G_3 = D_2 C^T$.
(10) Consider the Gaussian system (6.29) and (6.30) with the conditions provided there. The condition of (a) is checked for time-reversibility. Let $S = D_2 = \text{Block} - \text{diag}(I_{n_1}, -I_{n_2})$. It is then easily verified that $D_2 A = A^T D_2$. Then,

$$\begin{aligned} Q_5 &= A Q_5 A^T + M Q_v M^T = A Q_5 A^T + I - A A^T \\ &\Leftrightarrow (Q_5 - I) = A (Q_5 - I) A^T, \end{aligned}$$

and $\text{spec}(A) \subset \text{D}_o$ imply by Theorem 22.1.2 that $Q_5 - I = 0$ is the unique solution of this equation. Then

$$G = AQ_5C^T + MQ_vN^T = AC^T + (D_2 - A)C^T = D_2C^T \;\Leftrightarrow\; C = G^TD_2.$$

□

Realizations in Balanced Form

Consider a stationary Gaussian process with covariance function $W : T \to \mathbb{R}^{n_y \times n_y}$ admitting a finite-dimensional covariance realization of the form,

$$W(t) = \begin{cases} HF^{t-1}G, & \text{for} t > 0, \\ J + J^T, & \text{for } t = 0, \end{cases},$$

and the stochastic realization

$$\begin{aligned} x(t+1) &= Ax(t) + Mv(t), \\ y(t) &= Cx(t) + Nv(t), v(t) \in G(0, Q_v), \end{aligned}$$

with state covariance matrix $Q \in \mathbf{Q}_{\mathbf{lsdp}}$, $lsdp = \{n_y,\, n_x,\, n_y,\, F,\, G,\, H,\, J\} \in L\Sigma P_{min}$. Assume that $Q > 0$. By Proposition 3.4.8 the time-reversed process has covariance function $\overline{W} : T \to \mathbb{R}^{n_y \times n_y}$ $\overline{W}(t) := W(-t) = W(t)^T$. This process has also a stochastic realization for which the covariance realization is described by
lsp $= \{n_y,\, n_x,\, n_y,\, F^T,\, H^T,\, G^T,\, J\} \in L\Sigma P$ and $\overline{Q} \in \mathbf{Q}_{\mathbf{lsp}}$. By F.2.3 $Q_d \in \mathbf{Q}_{\mathbf{lsp}}$ if and only if $Q_d^{-1} \in \mathbf{Q}_{\mathbf{lsdp}}$. Recall that a weak Gaussian stochastic realization is nonunique in two ways. The first type of non-uniqueness is a state space transformation. This corresponds also to the non-uniqueness of the covariance realization,

$$W(t) = HF^{t-1}G, \text{ for } t > 1.$$

Proposition 6.10.6. *Let* $p_1 = \{n_y,\, n_x,\, n_y,\, F_1,\, G_1,\, H_1,\, J\} \in L\Sigma P$ *and* $p_2 = \{n_y,\, n_x,\, n_y,\, F_2,\, G_2,\, H_2,\, J\} \in L\Sigma P$ *be two equivalent covariance realizations that are related by* $T \in Gl_n(\mathbb{R})$ *according to*

$$F_2 = TF_1T^{-1},\; H_2 = H_1T^{-1},\; G_2 = TG_1.$$

Let $Q_{11} \in \mathbf{Q}_{dp_1}$, $Q_{12} \in \mathbf{Q}_{p_1}$, *be state covariance matrices for* p_1 *and its dual.*

(a)Then the corresponding covariance matrices for p_2 *are:*

$$Q_{21} = TQ_{11}T^T \in \mathbf{Q}_{dp_2},\; Q_{22} = T^{-T}Q_{12}T^{-1} \in \mathbf{Q}_{p_2}.$$

(b)$Q_{21}Q_{22} = TQ_{11}Q_{12}T^{-1}$.

Proof. (a) This follows from 24.2.3. (b) This follows directly from (a). □

A consequence of Proposition 6.10.6.(b) is that the eigenvalues of $Q_{11}Q_{12}$ are invariant with respect to a state space transformation. The question may therefore be posed: Is there a canonical form with respect to this invariant?

In the literature the result of Proposition 6.10.6 is usually stated for the case $Q_{11} = Q^- \in \mathbf{Q}_{\bar{p}_1}$ and $Q_{12} = Q_d^+ \in \mathbf{Q}_{p_1}$. That these conditions are not necessary has been pointed out in [42].

The statement of the canonical form requires a result from linear algebra, see Appendix 17.

Definition 6.10.7. A time-invariant Gaussian system is said to be a *balanced stochastic realization* of a stationary Gaussian process, or the stochastic realization is said to be in *balanced form,* if the state variance matrix Q of the forward representation satisfies $Q = Q^-$ and if Q^- and Q_d^+ satisfy,

$$Q^- = Q_d^+ = D = \text{Diag}(d_1, \ldots, d_{n_x}) \in \mathbb{R}_{diag}^{n_x \times n_x},\ d_1 \geq d_2 \geq \ldots \geq d_{n_x} > 0.$$

Proposition 6.10.8. *Consider a stationary Gaussian process with zero mean value function and a covariance function that admits a finite-dimensional covariance realization, say*

$$W(t) = \begin{cases} HF^{t-1}G, & for\, t > 0, \\ J + J^T, & for\, t = 0. \end{cases}$$

$$lsp = \{n_y,\ n_x,\ n_y,\ F,\ G,\ H,\ J\} \in L\Sigma P_{min},\ J = J^T.$$

Then:

(a)The process has a forward Kalman realization of the form

$$\begin{aligned} x^f(t+1) &= Ax^f(t) + K^f v^f(t), \\ y(t) &= Hx^f(t) + v^f(t), v^f(t) \in G(0, Q_{v^f}), \\ K^f &= [G - ADH^T][J + J^T - HDH^T]^{-1}, \\ Q_{v^f} &= J + J^T - HDH^T. \end{aligned}$$

(b)The process has a backward Kalman realization of the form

$$\begin{aligned} x^b(t-1) &= A^T x^b(t) + K^b v^b(t), \\ y(t) &= G^T x^b(t) + v^b(t), v^b(t) \in G(0, Q_{v^b}), \\ K^b &= [H^T - A^T DG][J + J^T - G^T DG]^{-1}, \\ Q_{v^b} &= J + J^T - G^T DG. \end{aligned}$$

(c)The state covariance matrices of the realizations of (a) and (b) are solutions to the following algebraic Riccati equations of stochastic realization:

$$\begin{aligned} D &= ADA^T + [G - ADH^T][J + J^T - HDH^T]^{-1}[G - ADH^T]^T, \\ D &= A^T DA + [H^T - A^T DG][J + J^T - G^T DG]^{-1}[H^T - A^T DG]^T. \end{aligned}$$

(d)Any two (forward) Kalman realizations are related by a transformation $T \in Gl_n(\mathbb{R})$ satisfying

$$
\begin{aligned}
&T \in Gl_n(\mathbb{R}),\ TD^2 = D^2T,\ T^TT = I,\\
A_2 &= TA_1T^{-1},\ \ H_2 = H_1T^{-1},\\
(A_2DH_2^T + K_2Q_{v,2}) &= T[A_1DH_1^T + K_1Q_{v,1}].
\end{aligned}
$$

Proof. This follows directly from Proposition 6.10.6 and Proposition 24.6.3. □

Procedure 6.10.9 Computation of a time-invariant Gaussian system representation in balanced form.
Data: $\{n_y,\ n_x,\ n_y,\ F,\ G,\ H,\ J\} \in \text{LSP}_{\min}$, $J = J^T$ *be the parameters of a covariance realization.*

1. *Compute* $Q^- \in \mathbb{R}^{n_x \times n_x}$ *as the minimal solution of the algebraic Riccati equation of stochastic realization for* $Q \in \mathbb{R}^{n_x \times n_x}$ *with two side conditions,*

$$
\begin{aligned}
Q &= AQA^T + [G - AQC^T][J + J^T - CQC^T]^{-1}[G - AQC^T]^T,\\
Q &= Q^T \succeq 0,\\
\text{spec}(A(Q)) &\subset \mathbb{C}_{outsideclosed disc} = \{c \in \mathbb{C} | |c| > 1\},
\end{aligned}
$$

and $Q_d^+ \in \mathbb{R}^{n_x \times n_x}$ *as the maximal solution of the algebraic Riccati equation of stochastic realization for* $Q \in \mathbb{R}^{n_x \times n_x}$ *with two side conditions,*

$$
\begin{aligned}
&Q = A^TQA + [C^T - A^TQG][J + J^T - G^TQG]^{-1}[C^T - A^TQG]^T,\\
&Q = Q^T \succeq 0,\\
&\quad \text{spec}(A(Q)) \subset \mathbb{C}_{outsideclosedisc}.
\end{aligned}
$$

2. *Compute a contra gradient transform* $T \in \mathbb{R}^{n_x \times n_x}$ *nonsingular such that* $TQ^-T^T = D = T^{-T}Q_d^+T^{-1}$. *See Procedure 17.4.36.*
3. *Define*

$$
\begin{aligned}
&A_1 = TAT^{-1},\ C_1 = CT^{-1},\ G_1 = TG,\\
&K_1 = [G_1 - A_1DC_1^T][J + J^T - C_1DC_1^T]^{-1},\ Q_{v,1} = [J + J^T - C_1DC_1^T].
\end{aligned}
$$

Then $\{n_y,\ n_x, n_y,\ A_1,\ C_1,\ K_1,\ I,\ Q_{v,1}\} \in G S\Sigma P$ *are the parameters of the Kalman realization in balanced form.*

6.11 A Canonical Form

The set of minimal weak Gaussian stochastic realizations consists not of a unique element but in general of a set of realizations. Each of these realizations is equivalent to all other minimal stochastic realizations where the equivalence relation is defined in Theorem 6.4.3.(d). It is therefore of interest to have a canonical form of this set of realizations where an element of the canonical form represents a subset of equivalent realizations. The concept of a canonical form is used for this purpose. The reader is expected to have read the definition of a canonical form as formulated in Def. 17.1.4.

The reader finds in this section the definition of a canonical form for the set of minimal weak Gaussian stochastic realizations but only for the case of a single output process. The multivariate case requires a description with a higher complexity.

The canonical form of a minimal weak Gaussian stochastic realization is used in system identification and in observer design.

Consider next a stationary Gaussian process. If it admits a finite-dimensional weak Gaussian stochastic realization then the set of minimal weak Gaussian stochastic realizations is characterized in Theorem 6.4.3.(d).

Definition 6.11.1. The *single-output observable canonical form* of a minimal weak Gaussian stochastic realization of the single-output process of a Gaussian system.

Consider a covariance function of a single-output stationary Gaussian process satisfying the conditions of Theorem 6.4.3. Consider the set of minimal weak Gaussian stochastic realizations of the considered stationary Gaussian process as described in Theorem 6.4.3.(d). That theorem defines an equivalence relation on the set of minimal weak Gaussian stochastic realizations.

Define the *single-output observable canonical form* of a minimal weak Gaussian stochastic realization of the considered stationary Gaussian process by the following representations and relations. The canonical form is defined as a subset of all minimal weak Gaussian stochastic realizations but with particular properties. For the system matrices of this canonical form the subindex $ocf1$ is used throughout to indicate that it concerns the *observable canonical form of a single-output Gaussian system*.

$$\begin{aligned}
x(t+1) &= A_{ocf1}x(t) + K_{ocf1}\overline{v}(t),\ x(0) = x_0,\\
y(t) &= C_{ocf1}x(t) + N_{ocf1}\overline{v}(t),\\
A_{ocf1} &= \begin{pmatrix}
0 & 1 & 0 & 0 & 0\ldots & 0 & 0\\
0 & 0 & 1 & 0 & 0\ldots & 0 & 0\\
0 & 0 & 0 & 1 & 0\ldots & 0 & 0\\
0 & 0 & 0 & 0 & 1\ldots & 0 & 0\\
\vdots & & & & \ddots & & \vdots\\
0 & 0 & 0 & 0 & 0\ldots & 1 & 0\\
0 & 0 & 0 & 0 & 0\ldots & 0 & 1\\
-a_0 & -a_1 & -a_2 & -a_3 & -a_4\ \ldots & -a_{n_x-2} & -a_{n_x-1}
\end{pmatrix},\ K = \begin{pmatrix} k_1\\ k_2\\ k_3\\ k_4\\ \vdots\\ k_{n_x-1}\\ k_{n_x}\end{pmatrix},\\
C_{ocf1} &= \begin{pmatrix}1\ 0\ 0\ \ldots\ 0\end{pmatrix} \in \mathbb{R}^{1\times n_x},\ N_{ocf1} \in \mathbb{R}_+,
\end{aligned}$$

$$\det(sI - A_{ocf1}) = s^{n_x} + \sum_{i=0}^{n_x-1} a_i\, s^i,$$

$\operatorname{spec}(A_{ocf1}) \subset \mathrm{D}_o,\ \operatorname{spec}(A_{ocf1} - K_{ocf1}C_{ocf1}) \subset \mathrm{D}_o,$

(A_{ocf1}, K_{ocf1}) is a supportable pair,

$\left((A_{ocf1} - K_{ocf1}C_{ocf1}), C_{ocf1}\right)$, is an observable pair,

$Q_x \in \mathbb{R}^{n_x\times n_x}_{pds}$ is the unique solution of the Lyapunov equation,

$$Q_x = A_{ocf1}Q_xA_{ocf1}^T + K_{ocf1}K_{ocf1}^T,\ Q_x \succ 0,\ \ x_0 \in G(0, Q_x),$$

$$A_{b,ocf1} = Q_x A_{ocf1}^T Q_x^{-1},\ C_{b,ocf1} = C_{ocf1} Q_x A_{ocf1}^T Q_x^{-1} + K_{ocf1}^T Q_x^{-1},$$
$(A_{b,ocf1}, C_{b,ocf1})$ is an observable pair,
$$(n_x, n_y, n_y, A_{ocf1}, C_{ocf1}, M_{ocf1}, N_{ocf1}) \in WGSR_{ocf1}.$$

That (A_{ocf1}, C_{ocf1}) is an observable pair follows directly from the above chosen representations of the system matrices. It is a result that the assumptions imply that $Q_x \succ 0$ as will be proven below.

To prove that the above defined form is actually a canonical form of the set of considered stochastic realizations, the following statements have to be proven:

- the stochastic system defined in the observable canonical form is a minimal weak Gaussian stochastic realization of the considered single-output process;
- any minimal stochastic realization can be transformed to a system in the observable canonical form; and
- two system presentations in the same form of the considered process which are equivalent, are identical.

These items will be proven below. First a procedure is defined which transforms any minimal weak Gaussian stochastic realization to a system in the single-output observable canonical form.

Procedure 6.11.2 Transformation of a single-output minimal weak Gaussian stochastic realization to the single-output observable canonincal form.
Input: the indices and system matrices, $(n_x, n_y = 1, n_v, A, M, C, N, Q_v)$ *where by definition the variance of the noise process is equal to* $Q_v = I_{n_y}$*. By assumption, these system matrices represent a minimal weak Gaussian stochastic realization of its output process.*
Output: the indices and system matrices, $(n_x, n_y, A_{ocf1}, K_{ocf1}, C_{ocf1}, N_{ocf1})$ *where the system matrices have the form specified in Def. 6.11.1.*

1. Check on input data. *Calculate the following matrices and check the stated conditions. If these conditions are all satisfied then the time-invariant Gaussian system is a minimal weak Gaussian stochastic realization of the output process of the Gaussian system specified by the input.*

$$Q_v = I,\ \text{rank}\left(M\ AM\ A^2M\ \ldots\ A^{n_x-1}M\right) = n_x,\ \text{rank}(NN^T) = n_y;$$
$$\text{spec}(A) \subset \text{D}_o,\ \text{spec}(A - M(NN^T)^{-1}C) \subset \text{D}_o,$$
$$A_b = Q_x A^T Q_x^{-1},\ C_b = CQ_x A^T Q_x^{-1} + K^T Q_x^{-1},$$
$$\text{rank}\begin{pmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n_x-1} \end{pmatrix} = n_x,\ \text{rank}\begin{pmatrix} C_b \\ C_bA_b \\ C_bA_b^2 \\ \vdots \\ C_bA_b^{n_x-1} \end{pmatrix} = n_x.$$

2. Solve the algebraic Riccati equation of Kalman filtering by determining a matrix Q_f *satisfying the following algebraic Riccati equation with side conditions,*

$$\begin{aligned}
& Q_f \in \mathbb{R}_{pds}^{n_x \times n_x}, \\
Q_f &= A Q_f A^T + M M^T + \\
& \quad -[A Q_f C^T + M N^T][C Q_f C^T + N N^T]^{-1}[A Q_f C^T + M N^T]^T, \\
& \quad \mathrm{spec}(A + K(Q_f)C) \subset \mathrm{D}_o; \\
K(Q_f) &= [A Q_f C^T + M N^T][C Q_f C^T + N N^T]^{-1}; \\
N_{\bar{v}} N_{\bar{v}}^T &= C Q_f C^T + N N^T \in \mathbb{R}_+, \; N_{\bar{v}} \in \mathbb{R}_{s+}, \; \mathrm{rank}(N_{\bar{v}}) = 1; \\
K_{\bar{v}} K_{\bar{v}}^T &= K(Q_f)[C Q_f C^T + N N^T] K(Q_f)^T, \; K_{\bar{v}} \in \mathbb{R}^{n_x \times n_y}, \\
\hat{x}(t+1) &= A \hat{x}(t) + K_{\bar{v}} \bar{v}(t), \; \hat{x}(0) = m_{x_0} = 0, \\
y(t) &= C \hat{x}(t) + N_{\bar{v}} \bar{v}(t); \\
& \quad \mathrm{spec}(A) \subset \mathrm{D}_o, \; \mathrm{spec}(A - K_{\bar{v}} C) \subset \mathrm{D}_o, \\
Q_{\hat{x}} &= A Q_{\hat{x}} A^T + K_{\bar{v}} K_{\bar{v}}^T, \; Q_{\hat{x}} \succ 0.
\end{aligned}$$

3. *Define the state-space transformation to the observable canonical form according to the computations,*

$$L = \begin{pmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n_x - 1} \end{pmatrix}, \; LAL^{-1} = \begin{pmatrix} 0 & 1 & 0 & \ldots 0 & 0 \\ 0 & 0 & 1 & \ldots 0 & 0 \\ 0 & 0 & 0 & \ldots 0 & 0 \\ \vdots & & & \ddots 0 & 0 \\ 0 & 0 & 0 & \ldots 1 & 0 \\ 0 & 0 & 0 & \ldots 0 & 1 \\ -a_0 & -a_1 & -a_2 & \ldots -a_{n_x-2} & -a_{n_x-1} \end{pmatrix},$$

$$\begin{aligned}
A_{ocf1} &= LAL^{-1} \in \mathbb{R}^{n_x \times n_x}, \textit{ where,} \\
0 &= A^{n_x} + \sum_{i=0}^{n_x - 1} a_i \, A^i, \; \forall \, i \in \mathbb{N}_{n_x - 1}, \; a_i \in \mathbb{R}, \\
C_{ocf1} &= CL^{-1} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \end{pmatrix} \in \mathbb{R}^{1 \times n_x}, \\
K_{ocf1} &= LK_{\bar{v}} = \begin{pmatrix} k_1 \\ k_2 \\ k_3 \\ \vdots \\ k_{n_x - 1} \\ k_{n_x} \end{pmatrix} \in \mathbb{R}^{n_x \times n_y}, \; \forall \, i \in \mathbb{Z}_{n_x}, \; k_i \in \mathbb{R}, \\
N_{ocf1} &= N_{\bar{v}} \in \mathbb{R}_{s+}.
\end{aligned}$$

The computation of LAL^{-1} will yield the matrix A_{ocf1} as displayed above and hence provides the real numbers $\{a_i \in \mathbb{R}, \; i \in \mathbb{N}_{n_x - 1}\}$ which satisfy the characteristic polynomial of A_{ocf1} as defined above. The matrix C_{ocf1} has the described form. These statements are proven below.

4. *Output the indices and matrices $(n_x, n_y = 1, n_v = 1, A_{ocf1}, K_{ocf1}, C_{ocf1}, N_{ocf1})$.*

Theorem 6.11.3. *Consider Def. 6.11.1 of the single-output observable canonical form of a minimal weak Gaussian stochastic realization and Procedure 6.11.2.*

(a)Any element of the set of observable canonical forms is a minimal weak Gaussian stochastic realization.

(b)Procedure 6.11.2 is well defined and transforms any minimal weak Gaussian stochastic realization to an element of the observable canonical form to which it is equivalent.

(c)Any two elements of the observable canonical form which are equivalent, are actually equal.

Proof. (a) Consider an element of the observable canonical form. By Def. 6.11.1 then (A_{ocf1}, K_{ocf1}) is a supportable pair, (A_b, C_b) is an observable pair, $\text{spec}(A_{ocf1}) \subset \text{D}_o$, and $\text{spec}(A_{ocf1} - K_{ocf1} C_{ocf1}) \subset \text{D}_o$. It is a simple calculation that the observability matrix associated with the pair of system matrices (A_{ocf1}, C_{ocf1}) is the identity matrix hence nonsingular and thus (A_{ocf1}, C_{ocf1}) is an observable pair.

It then follows from those conditions and from Theorem 6.4.3.(c) that the time-invariant Gaussian system of the single-output observable canonical form is a minimal weak Gaussian stochastic realization of the output process of the considered Gaussian system.

(b) Consider any minimal weak Gaussian stochastic realization of the considered output process. It will be argued that Procedure 6.11.2 is well defined and that it produces an element of the single-output observable canonical form. Step 1 of the procedure is a simple check. It follows from Theorem 6.4.3.(c) that if the conditions of this step are satisfied that then the Gaussian system is a minimal weak Gaussian stochastic realization of its output function.

Step 2 specifies the computation of the solution of the algebraic Riccati equation associated with the Gaussian system. It follows from Theorem 22.2.2.(f) that the algebraic Riccati equation as specified has a unique solution $Q_f \in \mathbb{R}_{spds}^{n_x \times n_x}$ hence satisfying $0 \prec Q_f$. Define then the Kalman gain matrix $K(Q_f)$ as described in the procedure.

It follows from Theorem 4.4.5.(a) that the variance matrix of the transformed system satisfies,

$$Q_{\hat{x}} = A Q_{\hat{x}} A^T + K(Q_f) Q_{\overline{v}} K(Q_f)^T .$$

Because the original Gaussian system is a minimal weak Gaussian stochastic realization with the conditions specified, it follows that $0 \prec Q_{\hat{x}}$. Because $\text{spec}(A) \subset \text{D}_o$ and $Q_{\hat{x}} \succ 0$ it follows from Theorem 22.1.2.(d) that $(A, K_{\overline{v}Q_{\overline{v}}^{-1}})$ is a controllable/supportable pair. That (A, C) is an observable pair follows from Step 1. It has to be proven that (A_b, C_b) is an observable pair. These matrices have changed due to the construction of the Kalman filter.

Thus the Gaussian system of Step (2) is a minimal weak Gaussian stochastic realization of its output process.

Step 3 concerns a state-space transformation. Because by Step (2), (A, C) is an observable pair, the rank of the observability matrix L satisfies $\text{rank}(L) = n_x$, hence is invertible. Then carry out the state transformation $\overline{x}(t) = Lx(t)$ which yields the new system matrices $A_{ocf1} = LAL^{-1}$, $K_{ocf1} = LK(Q_f)$, $C_{ocf1} = CL^{-1}$, and $N_{ocf1} = N_{\overline{v}}$. The forms of those matrices can be obtained from the following considerations,

$$\begin{aligned}
A_{ocf1} &= LAL^{-1} \ \Leftrightarrow \ A_{ocf1}L = LA = \begin{pmatrix} CA \\ CA^2 \\ \vdots \\ CA^{n_x} \end{pmatrix}; \\
C_{ocf1} &= CL^{-1} \ \Leftrightarrow \ C_{ocf1}L = C; \\
K_{ocf1} &= LK(Q_f),
\end{aligned}$$

where the particular forms of A_{ocf1} and C_{ocf1} follow from the equality of the above equations when considered for each of the elements of those matrices.

Finally one obtains the indices and matrices of the form, $(n_x, n_y, n_y, A_{ocf1}, C_{ocf1}, K_{ocf1}, N_{obs1})$. Because the transformation was by the nonsingular matrix L, and by writing out the explicit conditions such as for a supportable pair or an observable pair, one obtains that (A_{ocf1}, K_{ocf1}) is a supportable pair, (A_{ocf1}, C_{ocf1}) is an observable pair, $(A_{b,ocf1}, C_{b,ocf1})$ is an observable pair, $\text{spec}(A_{ocf1}) = \text{spec}(LAL^{-1}) = \text{spec}(A) \subset \text{D}_o$ and

$$\begin{aligned}
&\text{spec}(A_{ocf1} - K_{ocf1}C_{ocf1}) = \text{spec}(L(A - K(Q_f)C)L^{-1}) = \\
&= \text{spec}(A - K(Q_f)C) \subset \text{D}_o.
\end{aligned}$$

Hence the tuple describes an element of the observable canonical form.

(c) Consider two elements of the observable canonical form

$$(n_x, n_y, n_y, A_1, C_1, M_1, N_1), \ (n_x, n_y, n_y, A_2, C_2, M_2, N_2) \in \text{WGSR}_{ocf1},$$

which are equivalent as minimal weak Gaussian stochastic realizations. Because both elements are members of the observable canonical form, $\text{spec}(A_1) \subset \text{D}_o$, $\text{spec}(A_2) \subset \text{D}_o$, $\text{spec}(A_1 - K_1C_1) \subset \text{D}_o$, $\text{spec}(A_2 - K_2C_2) \subset \text{D}_o$. By the equivalence of the system matrices by a nonsingular transformation matrix $L_1 \in \mathbb{R}^{n_x \times n_x}$ it follows from the equalities that $C_{1,ocf} = C_{2,ocf}L_1^{-1}$ and $A_{1,ocf}L_1 = L_1A_{2,ocf}$ and consideration of the element wise equalities of these matrices, that the matrix L_1 is the identity matrix and that $A_{1,ocf} = A_{2,ocf}$, $K_{1,ocf} = K_{2,ocf}$, $C_{1,ocf} = C_{2,ocf}$, and $N_{1,ocf} = N_{2,ocf}$. From the equalities $A_{1,ocf} = A_{2,ocf}$ and $K_{1,ocf} = K_{2,ocf}$ then follows that the parameters of these matrices are equal, thus for all $i \in \mathbb{N}_{n_x-1}$, $a_{1,i} = a_{2,i}$ and $k_{1,i+1} = k_{2,i+1}$. Thus these two elements of the observable canonical form are equal. □

6.12 Exercises

Problem 6.12.1. Analyse the set $\mathbf{Q_{lsp}}$ in case $n_y = n_x = 1$. Hence, consider

$$\begin{aligned}
lsp &= \{1, 1, 1, f, g, h, j\} \in \text{LSP}_{min}, \\
\mathbf{Q_{lsp}} &= \left\{ q \in \mathbb{R}_+ | \ v(q) = \begin{pmatrix} q - f^2q & h - fgq \\ h - fgq & 2j - g^2q \end{pmatrix} \succeq 0 \right\}.
\end{aligned}$$

Assume that lsp $\in \text{LSP}_{min}$ is such that $\mathbf{Q_{lsp}}$ is regular (see definition F.4.1).

(a) Determine an equivalent condition for $\mathbf{Q_{lsp}} \neq \emptyset$ in terms of f, g, h, j.

(b) Describe the set $\mathbf{Q_{lsp}}$, for example its geometric structure, the boundary points, whether it has an interior etc.

Problem 6.12.2. Prove by the following steps that a time-invariant Gaussian system may at the same time be stochastically observable and not be stochastically co-observable.

(a) Consider the time-invariant backward Gaussian system representation,

$$\begin{aligned}
\sigma_1 &= \{\Omega, F, P, T, \mathbb{R}, B, \mathbb{R}, B, y, x\} \in \text{GStocS},\ T = \mathbb{Z},\\
x(t-1) &= a_1 x(t) + \begin{pmatrix} 1 & 0 \end{pmatrix} w(t),\\
y(t-1) &= c_1 x(t) + \begin{pmatrix} 0 & 1 \end{pmatrix} w(t),\\
&\quad w(t) \in G(0, Q_w),\ Q_w = Q_w^T \succ 0,\ Q_{w,2,1} \neq 0,\\
&\quad a_1 \in (-1, +1),\ c_1 = 0.
\end{aligned}$$

Provide arguments why this system is not stochastic co-observable. Note that because $c_1 = 0$ the output process y is Gaussian white noise.

(b) Construct the forward representation associated with the Gaussian system of (a) above, say

$$\begin{aligned}
x(t+1) &= a_2 x(t) + \begin{pmatrix} 1 & 0 \end{pmatrix} v(t),\\
y(t) &= c_2 x(t) + \begin{pmatrix} 0 & 1 \end{pmatrix} v(t),\ v(t) \in G(0, Q_v).
\end{aligned}$$

Determine expressions for a_2 and c_2 in terms of a_1 and Q_w.

(c) Provide arguments why the forward representation is stochastic observable.

Problem 6.12.3. Let $x : \Omega \times T \to \mathbb{R}^n$ be a stationary Gauss-Markov process with a zero mean-value function, variance $Q = E[x(t)x(t)^T] \succ 0$, (A, M) a supportable pair, and representation,

$$x(t+1) = Ax(t) + Mv(t), v(t) \in G(0, Q_v).$$

Prove that x is a time-reversible process if and only if there exists a diagonal matrix $\Lambda \in \mathbb{R}^{n \times n}$ with

$$\Lambda = \text{Diag}(\lambda_1, \ldots, \lambda_n),\ \forall\, i \in \mathbb{Z}_n,\ |\lambda_i| < 1,$$

a matrix $S \in \mathbb{R}^{n \times n}_{nsng}$ nonsingular, and a Gaussian white noise process $w : \Omega \times T \to \mathbb{R}^n$ with $w(t) \in G(0, I - \Lambda^2)$, such that, with $x(t) = Sz(t)$,

$$z(t+1) = \Lambda z(t) + w(t).$$

Explain why then the components of z are independent processes.

6.13 Further Reading

Books on stochastic realization of time-invariant Gaussian systems. A reference in book form on the stochastic realization problem at the level of this chapter on the weak Gaussian stochastic realization problem is [34]. See also the report [32]. Both of these references are in the French language. A chapter in the English language by P. Faurre is [33].

The weak Gaussian stochastic realization problem has been posed by R.E. Kalman, [56], though it is formulated indirectly and not very explicit.

The weak Gaussian stochastic realization problem has been solved by P. Faurre [32, 33]. The main results are described in the book [34]. A reference of a survey character is [63]. The relation between the different stochastic realizations is explored in [75].

References on the partial weak-Gaussian stochastic realization problem are [57, 58, 79].

Square stochastic realizations are treated in [35]. For procedures for the construction of a covariance realization, [18, 21, 85]. Theorem 6.4.5 is due to [34, p. 88, Th. 4.17]. The realization procedures mentioned are due to [34, pp. 232-235]. There exists a procedure which, from a covariance function, directly produces the system matrices of the Kalman realization. This procedure is due to G. Ruckebusch and my be found in [34, Annexe 8.A]. The result on realization of a time-reversible stationary Gaussian process is due to J.C. Willems, [92].

References on the weak Gaussian stochastic realization problem for nonstationary covariance functions are [52, 61, 97].

The stochastic realization problem for stochastic dynamic factor systems is treated in [26, 36, 39, 74, 88].

An application of Gaussian stochastic realization to the problem of modeling economic data by a stochastic system, may be found in [11, 12].

Continuous-time weak stochastic realizations that are time-reversible have been analysed in [7, 92]. Continuous-time weak stochastic realizations that are in balanced form have been treated in [43, 42]. The presentation of discrete-time weak stochastic realizations in balanced form is a generalization of unpublished notes of P.H.M. Janssen for which use the author is very grateful. Stochastic realizations of systems with a feedback structure are described in [38].

Parametrization of Gaussian stochastic systems as weak Gaussian stochastic realizations is discussed in [22, 67]. The case of multi-output systems requires concepts from differential geometry and topology. It has been explored by B. Hanzon [45, 46]. For the concept of canonical form and parametrization of finite-dimensional linear systems see [47, 48].

There is an extensive literature on other representations of Gaussian systems. Of interest are the transformations from one representation to another. Below references on this are mentioned. The transformation from a covariance function to an autoregressive (AR) representation is treated in [23, 91]. References on spectral factorization of spectral density functions or spectral density matrices are [4, 10, 8, 9, 5, 6, 14, 15, 28, 27, 35, 71, 86, 91, 96]. References on Toeplitz operators in stochastic realization are [1, 17, 25, 27, 53, 78]. See also the book [44].

For the strong Gaussian stochastic realization problem see the book [65] and the papers [2, 3, 37, 63, 66, 62, 64, 73, 81, 83].

Model reduction of Gaussian system representations is treated in [29, 40, 51]. This is a form of approximate weak Gaussian stochastic realization. Additional research of a broader character on approximate realizations include [41, 72].

References

1. H. Akaike. Block toeplitz matrix inversion. *SIAM J. Appl. Math.*, 24:234–241, 1973. 217
2. H. Akaike. Stochastic theory of minimal realization. *IEEE Trans. Automatic Control*, 19:667–674, 1974. 217, 275
3. H. Akaike. Markovian representation of stochastic processes by canonical variables. *SIAM J. Control*, 13:162–173, 1975. 217, 275
4. B.D.O. Anderson. A system theory criterion for positive real matrices. *SIAM J. Control*, 5:171–182, 1967. 217
5. B.D.O. Anderson. Algebraic properties of minimal degree spectral factors. *Automatica*, 9:491–500; Correction, 11(1975), 321–322, 1973. 217
6. B.D.O. Anderson, K.L. Hitz, and N.D. Diem. Recursive algorithm for spectral factorization. *IEEE Trans. Circuits & Systems*, 21:742–750, 1974. 217
7. B.D.O. Anderson and T. Kailath. Forwards, backwards, and dynamically reversible Markovian models of second-order processes. *IEEE Trans. Circuits and Systems*, 26:956–965, 1979. 120, 217
8. B.D.O. Anderson and J.B. Moore. Algebraic structure of generalized positive real matrices. *SIAM J. Control*, 6:615–624, 1968. 217
9. B.D.O. Anderson and S. Vongpanitlerd. *Network analysis and synthesis*. Prentice Hall, Englewood Cliffs, NJ, 1973. 217
10. Brian D. Anderson. Dual form of a positive real lemma. *Proceedings of the IEEE*, X:1749–1750, 1967. 217
11. M. Aoki. On alternative state space representations of time series models. *J. Econ. Dyn. Control*, 12:595–607, 1988. 217
12. M. Aoki and A. Havenner. Approximate state space models of some vector-valued macroeconomic time series for cross-country comparisons. *J. Econ. Dyn. & Control*, 10:149–155, 1986. 217
13. M.A. Arbib and E.G. Manes. Generalized Hankel matrices and system realization. *SIAM J. Math. Anal.*, 11:405–424, 1980. 174
14. H. Bart, I. Gohberg, M.A. Kaashoek, and P. Van Dooren. Factorizations of transfer functions. *SIAM J. Control & Optim.*, 18:675–696, 1980. 217
15. F.L. Bauer. Ein direktes iterationsverfahren zur Hurwitz-zerlegung eines polynoms. *Archiv. der elektrischen Übertragung*, 9:285–290, 1955. 217
16. D. Blackwell and L. Koopmans. On the identifiability problem for functions of finite Markov chains. *Ann. Math. Statist.*, 28:1011–1015, 1957. 175, 277
17. A. Böttcher and B. Silbermann. *Analysis of Toeplitz operators*. Springer-Verlag, Berlin, 1990. 217
18. R.W. Brockett. *Finite dimensional linear systems*. Wiley, New York, 1970. 217, 438
19. J.A. Brzozowski. A survey of regular expressions and their applications. *IEEE Trans. Electronic Computers*, 11:324–335, 1962. 174, 807
20. J.A. Brzozowski. Derivatives of regular expressions. *J. ACM*, 11:481–494, 1964. 174, 807
21. F.M. Callier and C.A. Desoer. *Linear system theory*. Springer-Verlag, New York, 1991. 217, 781, 808
22. J.V. Candy, T.E. Bullock, and M.E. Warren. Invariant system description of the stochastic realization. *Automatica J.-IFAC*, 15:493–495, 1979. 217
23. B.S. Choi. An algorithm for solving the extended yule-walker equations of an autoregressive moving-average time series. *IEEE Trans. Inform. Theory*, 32:417–419, 1986. 217

24. M.M. Connors. Controllability of discrete, linear, random dynamical systems. *SIAM J. Control*, 5:183–209, 1967. 175, 376

25. G. Cybenko. The numerical stability of the Levinson-Durbin algorithm for Toeplitz systems of equations. *SIAM J. Sci. Stat. Comput.*, 1:303–319, 1980. 217

26. M. Deistler and B.D.O. Anderson. Identification of dynamic systems from noisy data: The case $m* = 1$. Report, Institute of Econometrics, Operations Reseach, and System Theory, University of Technology Vienna, Vienna, 1990. 217

27. P. Delsarte, Y. Genin, and Y. Kamp. Schur parametrization of positive definite block-Toeplitz systems. *SIAM J. Appl. Math.*, 36:34–46, 1979. 217

28. P. Delsarte, Y.V. Genin, and Y.G. Kamp. Orthogonal polynomial matrices on the unit circle. *IEEE Trans. Circuits & Systems*, 25:149–160, 1978. 217

29. U.B. Desai, D. Pal, and R.D. Kirkpatrick. A realization approach to stochastic model reduction. *Int. J. Control*, 42:821–838, 1985. 218

30. Paul M. Van Dooren. The generalized eigenstructure problem in linear system theory. *IEEE Trans. Automatic Control*, 26:111–129, 1981. 193, 194, 808

31. H. Ehrig, K.-D. Kiermeier, H.-J. Kreowski, and W. Künel. *Universal theory of automata*. Teubner Studienbücher - Informatik. B.G. Teubner, Stuttgart, 1974. 174

32. P. Faurre. Réalization markoviennes de processus stationnaires. Rapport de recherche 13, IRIA, Rocquencourt, 1973. 175, 180, 217, 275

33. P. Faurre. Stochastic realization algorithms. In R.K. Mehra and D.G. Lainiotis, editors, *System Identification - Advances and Case Studies*, pages 1–25. Academic Press, New York, 1976. 180, 217, 275

34. P. Faurre, M. Clerget, and F. Germain. *Opérateurs rationnels positifs*. Dunod, Paris, 1979. 175, 180, 217, 275, 292, 310, 850, 865, 867, 877, 885

35. L. Finesso and G. Picci. A characterization of minimal square spectral factors. *IEEE Trans. Automatic Control*, 27:122–127, 1982. 217

36. L. Finesso and G. Picci. Linear statistical models and stochastic realization theory. In A. Bensoussan and J.L. Lions, editors, *Analysis and optimization of sysems, Part 1*, volume 62 of *Lecture Notes in Control and Information Sciences*, pages 445–470. Springer-Verlag, Berlin, 1984. 217

37. A.E. Frazho. On minimal splitting subspaces and stochastic realizations. *SIAM J. Control Optim.*, 20:553–562, 1982. 217

38. M.R. Gevers and B.D.O. Anderson. Representations of jointly stationary stochastic feedback processes. *Int. J. Control*, 33:777–809, 1981. 217

39. J. Geweke. The dynamic factor analysis of economic time-series models. In D. Aigner and A. Goldberger, editors, *Latent variables in socioeconomic models*, pages 365–383. North-Holland, Amsterdam, 1977. 217

40. K. Glover and E. Jonckheere. A comparison of two Hankel-norm methods for approximating spectra. In C.I. Byrnes and A. Lindquist, editors, *Modelling, Identification and Robust Control*, pages 297–306. Elsevier Science Publishers B.V. (North-Holland), Amsterdam, 1986. 218

41. A. Gombani. Consistent approximations of linear stochastic models. *SIAM J. Control & Opt.*, 27:83–107, 1989. 218

42. M. Green. Balanced stochastic realization. *J. Linear Algebra & Its Appl.*, 98:211–247, 1988. 209, 217

43. M. Green. A relative error bound for balanced stochastic trunction. *IEEE Trans. Automatic Control*, 33:961–965, 1988. 217

44. E.J. Hannan and M. Deistler. *The statistical theory of linear systems*. John Wiley & Sons, New York, 1988. 120, 217

45. B. Hanzon. *Identifiability, recursive identification, and spaces of linear dynamical systems*. Number 63 and 64 in CWI Tracts. Centre for Mathematics and Computer Science, Amsterdam, 1989. 217

46. B. Hanzon. On the differentiable manifold of fixed order stable linear systems. *Systems & Control Lett.*, 13:345–352, 1989. 217

47. M. Hazewinkel. Moduli and canonical forms for linear dynamical systems iii: The algebraic-geometric case. In C. Martin and R. Hermann, editors, *Proc. 1976 Ames Research Center (NASA) Conference on Geometric Control Theory*, pages 291–336. Math. Sci Press, Brookline, MA, 1977. 217
48. M. Hazewinkel and R.E. Kalman. On invariants, canonical forms and moduli for linear constant, finite dimensional dynamical systems. In *Proceedings CNR-CISM Symposium on Algebraic System Theory, Udine, 1975*, volume 131 of *Lecture Notes in Economics and Mathematical Systems*, pages 48–60. Springer Verlag, Berlin, 1976. 217
49. T.H. Hughes. Why RLC realizations of certain impedances need many more energy storage elements than expected. *IEEE Trans. Automatic Control*, 62:4333–4346, 2017. 174
50. T.H. Hughes, A. Morelli, and M.C. Smith. Electrical network synthesis: A survey of recent work. In R. Tempo, S. Yurkovich, and P. Misra, editors, *Emerging applications of control and systems theory*, volume 00 of *LNCIS*, pages 0–0. Springer, 2018. 174
51. E.A. Jonckheere and J.W. Helton. Power spectrum reduction by optimal Hankel norm approximation of the phase of the outer spectral factor. *IEEE Trans. Automatic Control*, 30:1192–1201, 1985. 218
52. T. Kailath and L. Ljung. Explicit strict sense state-space realizations of non-stationary processes. *Int. J. Control*, 42:971–988, 1985. 217
53. T. Kailath, A. Vieira, and M. Morf. Inverses of toeplitz operators, innovations, and orthogonal polynomials. *SIAM Review*, 20:106–119, 1978. 217
54. R.E. Kalman. Mathematical description of linear dynamical systems. *SIAM J. Control*, 1:152–192, 1963. 174, 761, 807, 808, 809
55. R.E. Kalman. New methods in Wiener filtering theory. In J.L. Bogdanoff and F. Kozin, editors, *Proceedings 1st Symposium Engineering Applications of Random Function Theory and Probability*, pages 270–388, New York, 1963. Wiley. 175, 275, 310
56. R.E. Kalman. Linear stochastic filtering - Reappraisal and outlook. In J. Fox, editor, *Proceedings Symposium on System Theory*, pages 197–205, New York, 1965. Polytechnic Press. 217, 310
57. R.E. Kalman. System identification from noisy data. In A.R. Bednarek and L. Cesari, editors, *Dynamical Systems II*, pages 135–164. Academic Press, New York, 1982. 116, 120, 217
58. H. Kimura. Positive partial realization of covariance sequences. In C.I. Byrnes and A. Lindquist, editors, *Modelling, Identification and Robust Control*, pages 499–513. Elsevier Science Publishers B.V. (North-Holland), Amsterdam, 1986. 217
59. S.C. Kleene. General recursive functions of natural numbers. *Mathematische Annalen*, 112:727–742, 1936. 174, 807
60. W. Kuich and A. Salomaa. *Semirings, automata, languages*. EATCS Monographs on Theoretical Computer Science. Springer-Verlag, 1986. 174
61. H. Lev-Ari and T. Kailath. Lattice filter parametrization and modeling of nonstationary processes. *IEEE Trans. Inform. Theory*, 30:2–16, 1984. 217
62. A. Lindquist, M. Pavon, and G. Picci. Recent trends in stochastic realization theory. In H. Saheli V. Mandrekar, editor, *Prediction theory and harmonic analysis*, pages 201–224. North-Holland Publ. Co., Amsterdam, 1983. 217
63. A. Lindquist and G. Picci. On the stochastic realization problem. *SIAM J. Control & Opt.*, 17:365–389, 1979. 217, 253
64. A. Lindquist and G. Picci. Realization theory for multivariate stationary gaussian processes. *SIAM J. Control & Opt.*, 23:809–857, 1985. 217
65. A. Lindquist and G. Picci. *Stochastic realization of Gaussian processes – A geometric approach to modeling, estimation and identification*. Springer, Heidelberg, 2015. 120, 121, 175, 176, 180, 217, 246, 253, 254, 275, 276
66. A. Lindquist, G. Picci, and G. Ruckebusch. On minimal splitting subspaces and Markovian representations. *Math. Systems Th.*, 12:271–279, 1979. 120, 175, 217, 275
67. B.P. McGinnie, R.J. Ober, and J.M. Maciejowski. Balanced parametrizations in time-series parametrization. In *Proceedings of the 29th Conference on Decision and Control*, pages 3202–3203, New York, 1990. IEEE Press. 217

68. B. McMillan. Introduction to formal realizability theory I. *Bell System Techn. J.*, 31:217–279, 1952. 174, 807, 865
69. B. McMillan. Introduction to formal realizability theory II. *Bell System Techn. J.*, 31:541–600, 1952. 174, 807, 865
70. A. Nerode. Linear automaton transformations. *Proc. Amer. Math. Soc.*, 9:541–544, 1958. 174, 807
71. M. Pagano. On the linear convergence of a covariance factorization algorithm. *J. A.C.M.*, 23:310–316, 1976. 217
72. V.V. Pellar and S.V. Khrushchev. Hankel operators, best approximations, and stationary gaussian processes. *Russian Math. Surveys*, 37:61–144, 1982. 218
73. G. Picci. Stochastic realization of gaussian processes. *Proc. IEEE*, 64:112–122, 1976. 120, 175, 217, 275
74. G. Picci and S. Pinzoni. Factor analysis models for stationary stochastic processes. In J.L. Lions A. Bensoussan, editor, *Analysis and Optimization of Systems*, volume 83 of *Lecture Notes in Control and Information Sciences*, pages 412–424, Berlin, 1986. Springer-Verlag. 217
75. G. Picci and S. Pinzoni. Acausal models of stationary processes. In C. Commault et al., editor, *Proceeding First European Control Conference*, pages 613–616, Paris, 1991. Hermès. 217
76. J.W. Polderman and J.C. Willems. *Introduction to mathematical system theory - A behavioral approach*. Number 26 in Texts in Applied Mathematics. Springer, New York, 1997. 175
77. M.O. Rabin and D. Scott. Finite automata and their decision problems. *IBM J.*, x:114–125, 1959. 174, 807
78. J. Rissanen. Algorithms for triangular decomposition of block Hankel and Toeplitz matrices with application to factoring positive matrix polynamial. *Math. of Computation*, 27:147–154, 1973. 217
79. J. Rissanen and T. Kailath. Partial realization of random systems. *Automatica J.-IFAC*, 8:389–396, 1972. 217
80. H.H. Rosenbrock. *State space and multivariable theory*. Wiley, New York, 1970. 174, 780
81. G. Ruckebusch. Représentations markoviennes de processus gaussiens stationnaires. *C. R. Acad. Sc. Paris, Série A*, 282:649–651, 1976. 120, 175, 217, 248, 253, 275
82. G. Ruckebusch. Représentations markoviennes de processus gaussiens stationnaires et applications statistiques. Rapport interne 18, Ecole Polytechnique, Centre de Mathématiques Appliquées, 1977. 120, 175, 248, 253, 275, 291, 310
83. G. Ruckebusch. Théorie géométrique de la représentation markovienne. *Ann. Inst. Henri Poincaré*, 16:225–297, 1980. 120, 175, 217, 248, 253, 275
84. A. Salomaa and M. Soittola. *Automata-theoretic aspects of formal power series*. Monographs in Computer Science. Springer, New York, 2012. 174
85. E.D. Sontag. *Mathematical control theory: Deterministic finite dimensional systems (2nd. Ed.)*. Number 6 in Graduate Text in Applied Mathematics. Springer, New York, 1998. 217, 277, 808
86. W.G. Tuel. Computer algorithm for spectral factorization of rational matrices. *IBM J. Research and Development*, pages 163–170, 1968. 217
87. P. van Overschee and B.L.R. De Moor. *Subspace identification for linear systems*. Kluwer Academic Publishers, Dordrecht, 1996. 176
88. J.H. van Schuppen. Stochastic realization problems motivated by econometric modeling. In C.I. Byrnes and A. Lindquist, editors, *Modelling, Identification and Robust Control*, pages 259–275. Elsevier Science Publishers B.V. (North-Holland), Amsterdam, 1986. 117, 120, 217
89. G.A. van Zee. *System identification for multivariable control*. Ph.d. thesis, Delft University of Technology, Delft, 1981. 193
90. K. Wang, M.Z.Q. Chen, C. Li, and G. Chen. Passive controller realization of a biquadratic impedance with double poles and zeros as a seven-element series–parallel network for effective mechanical control. *IEEE Trans. Automatic Control*, 63:3010–3015, 2018. 174

91. P. Whittle. On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix. *Biometrika*, 50:129–134, 1963. 217
92. J.C. Willems. Time reversibility in deterministic and stochastic dynamical systems. In A. Ruberti R.R. Mohler, editor, *Recent developments in variable structure systems, economics and biology*, volume 162 of *Lecture Notes in Economics and Mathematical Systems*, pages 318–326. Springer-Verlag, Berlin, 1978. 217
93. J.C. Willems. From time series to linear systems - Part I. Finite dimensional linear time invariant systems. *Automatica J. IFAC*, 22:561–580, 1986. 174, 276
94. J.C. Willems. From time series to linear systems - Part II. Exact modelling. *Automatica J. IFAC*, 22:675–694, 1986. 174, 276
95. J.C. Willems. From time series to linear systems - Part III. Approximate modelling. *Automatica J. IFAC*, 23:87–115, 1987. 174, 276
96. D.C. Youla. On the factorization of rational matrices. *IEEE Trans. Information Theory*, 7:172–189, 1961. 217, 865
97. J. Zachrisson. On the minimal state space dimension of non-stationary gaussian processes in discrete time. In N.S. Tzannes D.G. Lainiotis, editor, *Advances in Control*, volume 2, pages 130–138. D. Reidel Publ. Co., Dordrecht, 1980. 217

Chapter 7
Stochastic Realization

Abstract Stochastic realization problems are presented for a tuple of Gaussian random variables, for a tuple of σ-algebras, for a σ-algebra family, and for a finite stochastic system. The solution of the weak and of the strong stochastic realization of a tuple of Gaussian random variables is provided. The main theoretical contribution is the description of the strong stochastic realization of a tuple of σ-algebras. This is followed by stochastic realization of a family of σ-algebras. Finally the stochastic realization problem for finite-valued output processes is discussed.

Key words: Stochastic realization. σ-algebraic system.

The framework of stochastic realization is introduced in Section 7.1. The subsequent sections deal with several special cases which can be read independently though there are of course similarities between the subsequent sections.

In this chapter the concepts of system theory are formulated according to the geometric approach. The *geometric approach* to stochastic system theory is based on the view point of considering spaces of variables rather than the representation of these variables. For example, one does not consider the Gaussian random variables y_1, y_2, and x but the σ-algebras which they generate, F^{y_1}, F^{y_2}, and F^x. Therefore the formulations will be in terms of σ-algebras. Only for proofs and for illustration purposes are representations in terms of random variables used.

The reader of this chapter is expected to know the concept of a conditional independence relation of a triple of σ-algebras and its properties, as defined in Section 2.9 with additional results stated in Section 19.8.

The reader of this chapter is expected to have read the general introduction to realization theory of Section 6.1.

7.1 The Conceptual Framework of Stochastic Realization

The reader finds in this section an introduction to the conceptual framework of stochastic realization. This introduction is useful for the understanding of the subsequent sections of this chapter. Proofs are not provided in this section, they may be found in the remainder of this chapter.

For a deterministic control system, the realization problem starts with an *input-output map*, from an input to an output. A *control system* is then a mathematical structure with an input, a state function, and an output. The control system is called a *realization* of the considered input-output map if the input-output map of the control system equals the considered input-output map.

If one takes away the time axis from a control system then there is only an *input-output map* from an input variable u to an output variable y, $f : U \to Y$, $y = f(u)$, see Fig. 7.1. A *system* for such a map is then a tuple (U,X,Y,g,h) consisting, in addition to the input set U and the output set Y, of the state set X, of the input map $g : U \to X$, and of the output map $h : X \to Y$. The system is called a *realization* or a *factorization* of the input-output map f if for all $u \in U$, $f(u) = h(g(u))$. One calls such a factorization *controllable* if the input map g is surjective ($\Leftrightarrow$ $\forall\, x \in X$, $\exists\, u \in U$ such that $x = g(u)$). One calls such a factorization *observable* if the output map h is injective ($\Leftrightarrow h(x_1) = h(x_2)$ implies that $x_1 = x_2$). It can then be proven that this factorization is a *minimal factorization* in the sense that any other factorization receiving the same input-output map either has a larger state set or the state sets are isomorphic by a bijective map. This factorization is the basis of realization theory and the reader is referred to Section 21.7 for a more detailed exposition.

$$\begin{tikzcd}
\boxed{U} \arrow[rr, "f"] \arrow[dr, "g"'] & & \boxed{Y} \\
& \boxed{X} \arrow[ur, "h"'] &
\end{tikzcd}$$

Fig. 7.1 Canonical factorization of a deterministic map.

One can describe a probabilistic input-output map as a map from an input to a probability distribution on the output set. A stochastic system is defined as a tuple similar to a deterministic system with the input map g from an input to a probability measure on the state set and the output map h from a state to a probability measure on the output set. A stochastic system is a stochastic realization if its associated input-output map equals the considered input-output map. The factorization is displayed in Fig. 7.2.

In stochastic realization one does not describe an input-output map but a relation. The input is replaced by the past of the observed process, the output is replaced by

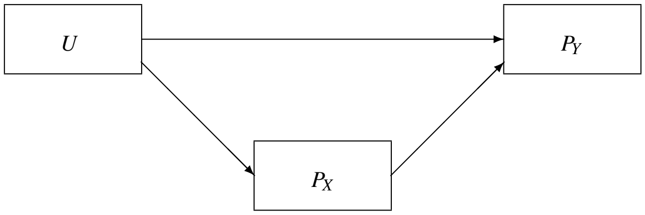


Fig. 7.2 Canonical factorization of a probabilistic map.

the future of the observed process, and the state is a variable which models the relation between the past and the future outputs. To be more specific, denote the past observation by the finite-dimensional random variable $y_- : \Omega \to \mathbb{R}^{n-}$, the future observation by the random variable $y_+ : \Omega \to \mathbb{R}^{n+}$, and the state by the random variable $x : \Omega \to \mathbb{R}^{n_x}$. These variables then form a system if the following conditional independence relation holds, $(F^{y+}, F^{y-} | \, F^x) \in \mathrm{CI}$. It is best to abstract even further and to denote the past, the future, and the state σ-algebras by F^-, F^+, G respectively. The system property is then denoted by the relation $(F^+, F^- | \, G) \in \mathrm{CI}$.

The weak stochastic realization problem is discussed first. Recall from Proposition 2.9.5 that for jointly Gaussian random variables $(y_+, y_-, x) \in G(0, Q)$ with $0 \prec Q_x$, conditional independence $(F^{y+}, F^{y-} | \, F^x) \in \mathrm{CI}$ holds if and only if the following matrix factorization holds $Q_{y_+,y_-} = Q_{y_+,x} Q_x^{-1} Q_{y_-,x}^T$. In the weak stochastic realization problem one is provided the matrix Q_{y_+,y_-}, which in general is a non-square matrix, and one has to construct the matrices in the right-hand side of the above equation. Using the singular value decomposition, Theorem 17.4.4, one obtains that $Q_{y_+,y_-} = USV^T$ where U and V are orthogonal matrices and S is a matrix with the singular values of the form,

$$S = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{n_{y+} \times n_{y-}}, \; D \in \mathbb{R}^{n_x \times n_x}_{s+,diag} (\Leftrightarrow \; 0 \prec D).$$

Because the orthogonal transformations leave the corresponding spaces invariant, these transformations are discarded from consideration in the following. A minimal factorization of Q_{y_+,y_-} is then directly obvious,

$$Q_{y_+,y_-} = S = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} D \\ 0 \end{pmatrix} I_{n_x} \begin{pmatrix} I & 0 \end{pmatrix} = Q_{y_+,x} Q_x^{-1} Q_{y_-,x}^T.$$

One may characterize this minimal factorization by the relations,

$$\begin{aligned} &\mathrm{Range}(Q_{y_+,y_-}) = \mathrm{Range}(Q_{y_+,x}), \; \ker(Q_{y_-,x}) = \{0\}, \\ &\mathrm{Range}(Q_{y_-,y_+}) = \mathrm{Range}(Q_{y_-,x}), \; \ker(Q_{y_+,x}) = \{0\}. \end{aligned}$$

The reader will learn in a formal definition later that the above conditions are respectively characterizations of stochastic co-controllability, stochastic co-observabilty, stochastic controllability, and stochastic observability of the above tuple of random variables. Another minimal factorization is,

$$Q_{y_+,y_-} = S = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} D \\ 0 \end{pmatrix} D^{-1} \begin{pmatrix} D & 0 \end{pmatrix} = Q_{y_+,x} Q_x^{-1} Q_{y_-,x}^T.$$

In case the factorization is not minimal it may take the form of,

$$Q_{y_+,y_-} = S = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} D & Q_2 \\ 0 & Q_3 \end{pmatrix} I_{n_x} \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} = Q_{y_+,x} Q_x^{-1} Q_{y_-,x}^T.$$

In this case $\text{Range}(Q_{y_+,y_-}) \subseteq \text{Range}(Q_{y_+,x})$ and $\ker(Q_{y_+,y_-}) \supseteq \ker(Q_{y_-,x}^T)$, etc.

Next consider the strong stochastic realization problem for a tuple of σ-algebras. First consider the past and the future σ-algebras F^-, F^+. Their relation can be defined to consist of the sub-σ-algebras $\sigma(F^- | \, F^+) \subseteq F^+$ and $\sigma(F^+ | \, F^-) \subseteq F^-$. The reader may think of the sub-σ-algebra $\sigma(F^- | \, F^+)$ as the effect of the past on the future and of $\sigma(F^+ | \, F^-)$ as the effect of the future on the past. Define for later use the *frame σ-algebra* as $F_f = \sigma(F^- | \, F^+) \vee \sigma(F^+ | \, F^-)$.

Next consider the system relation of a triple of σ-algebras according to $(F^+, \, F^- | \, G) \in \text{CI}$. This system relation is displayed in Fig. 7.3.

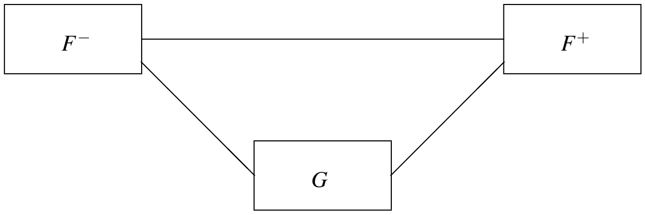


Fig. 7.3 Diagram presenting the conditional independence relation of a triple of σ-algebras representing, from left to right, the past F^-, the state G, and the future F^+ σ-algebras.

The *stochastic realization problem* is then to determine from a tuple $(F^+, \, F^-)$ of past and future σ-algebras, a state σ-algebra G such that $(F^+, \, F^- | \, G) \in \text{CI}$ and $G \subseteq F_f$. The stochastic realization problem has three subproblems.

The first subproblem is whether there exists a σ-algebra G such that $(F^+, \, F^- | \, G) \in \text{CI}$ and $G \subseteq F$. Existence follows directly from elementary properties of the conditional independence relation, for example $G_1 = F^+$, $G_2 = F^-$, $G_3 = \sigma(F^- | \, F^+)$, and $G_4 = \sigma(F^+ | \, F^-)$ all satisfy the conditional independence relation. Of interest are those state σ-algebras G which are contained in the frame σ-algebra F_f, thus if $G \subseteq F_f$. This means that the state σ-algebra is constructed only from the interaction of the past and the future σ-algebras.

The second subproblem is a characterization of minimal state sub-σ-algebras. A state σ-algebra G is called *minimal* if $(F^+, \, F^- | \, G) \in \text{CI}$, $G_1 \subseteq G$, and $(F^+, \, F^- | \, G_1) \in \text{CI}$ imply that $G = G_1$. This definition is identical to the concept of a minimal element of a partially-ordered set, Def. 17.1.6. Denote a minimal system by the relation $(F^+, \, F^- | \, G) \in \text{CI}_{\text{min}}$.

The characterization of a minimal state σ-algebra requires the following concepts. Consider a system triple satisfying the system relation $(F^+,\ F^-|\ G) \in \text{CI}$. Define this system to be: (1) *controllable* if $\sigma(G|\ F^-) = \sigma(F^+|\ F^-)$; (2) *co-controllable* if $\sigma(G|\ F^+) = \sigma(F^-|\ F^+)$; (3) *observable* if $\sigma(F^+|\ G) = G$; and (4) *co-observable* if $\sigma(F^-|\ G) = G$. It can then be proven that, if a state σ-algebra G satisfying $G \subseteq F_f$ is a minimal state σ-algebra, then it is controllable, co-controllable, observable, and co-observable. The converse is more involved and is stated in Section 7.3.

An example of a minimal state σ-algebra is $G = \sigma(F^+|\ F^-)$ which corresponds to the Kalman realization of the weak Gaussian stochastic realization problem.

A minimal state σ-algebra is not unique. There may exist two, or more, or an uncountable subset of such σ-algebras. The third subproblem of the stochastic realization problem is to relate two minimal state-σ-algebras and to describe the set of all minimal state σ-algebras for fixed $(F^+,\ F^-)$. A tuple of minimal state σ-algebras may be related by the projection relations: call two state σ-algebras G_1 and G_2 satisfying $(F^+,\ F^-|\ G_1) \in \text{CI}_{\text{min}}$, $(F^+,\ F^-|\ G_2) \in \text{CI}_{\text{min}}$, and $G_1,\ G_2 \subseteq F_f$, *isomorphic* if $\sigma(G_1|\ G_2) = G_2$ and $\sigma(G_2|\ G_1) = G_1$. However, not all such minimal state σ-algebras are isomorphic in this way. The results are described in Section 7.3.

Next consider the case of a stochastic system for an output process over time as defined in Def. 4.2.2. Let $T = \mathbb{N}$, $y : \Omega \times T \to Y \subseteq \mathbb{R}^{n_y}$ denote the output process, and let $x : \Omega \times T \to Y \subseteq \mathbb{R}^{n_x}$ denote the state process. A stochastic system is defined by the condition that,

$$(F_t^x \vee F_t^y,\ F_t^x \vee F_{t-1}^y|\ F^{x(t)}) \in \text{CI},\ \forall\, t \in T.$$

In words, the current state $x(t)$ makes the future of the state and the future of the output process conditionally independent from the past of the state and the past of the output process. The above definition of a stochastic system is identical to that of Def. 4.2.2.

The strong stochastic realization problem for a stochastic process starts with an output process, $y : \Omega \times T \to Y$. For any time $t \in T$, one constructs the past and the future of the output process at that time according to,

$$F_t^{y+} = \sigma(\{y(s),\ \forall\, s \geq t\}),\ F_{t-1}^{y-} = \sigma(\{y(s),\ \forall\, s \leq t-1\}).$$

Using the stochastic realization framework for a tuple of σ-algebras one can then construct a state σ-algebra $G_t \subseteq F_\infty^y$ such that $(F_t^{y+},\ F_{t-1}^{y-}|\ G_t) \in \text{CI}_{\text{min}}$. It requires in general a condition according to which there exists a finite-dimensional random variable $x(.,t) : \Omega \to X \subseteq \mathbb{R}^{n_x(t)}$ such that $G_t = F^{x(t)}$. Thus one obtains for all $t \in T$, $(F_t^{y+},\ F_{t-1}^{y-}|\ F^{x(t)}) \in \text{CI}_{\text{min}}$. See Fig. 7.4.

The above construction does not yet achieve a stochastic system. The construction of the state variable $x(.,t)$ for all times $t \in T$ has to be consistent for all $t \in T$. A condition is known, the σ-algebraic family $\{F_t,\ G_t,\ t \in T\}$ has to satisfy the condition that it is a *current-state σ-algebraic family*,

$$F_t^{y+} \vee F_t^{x+} = F_t^{y+} \vee F^{x(t)},\ \ F_{t-1}^{y-} \vee F_t^{x-} = F_{t-1}^{y-} \vee F^{x(t)},\ \ \forall\, t \in T.$$

Note that if the above condition holds then, for all $t \in T$,

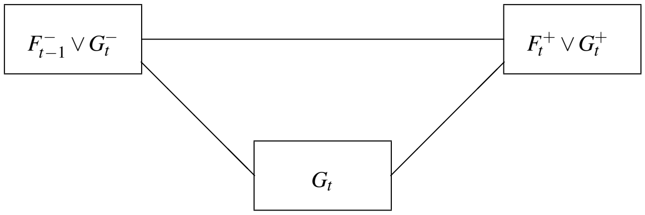


Fig. 7.4 Diagram presenting the conditional independence relation of a σ-algebraic system at time $t \in T$.

$$
\begin{aligned}
&(F_t^{y+},\, F_{t-1}^{y-} |\, F^{x(t)}) \in \mathrm{CI}_{\min} \\
&\Leftrightarrow (F_t^{y+} \vee F^{x(t)},\, F_{t-1}^{y-} \vee F^{x(t)} |\, F^{x(t)}) \in \mathrm{CI}_{\min},\ \text{by Proposition 19.8.2.(f)}, \\
&\Leftrightarrow (F_t^{y+} \vee F_t^{x+},\, F_{t-1}^{y-} \vee F_t^{x-} \,|\, F^{x(t)}) \in \mathrm{CI}_{\min}\ \text{ by the above condition.}
\end{aligned}
$$

Hence the processes (x, y) form a stochastic system which is then called a *strong stochastic realization* of the output process y.

Minimality of a stochastic realization can then be characterized directly by the concepts of stochastic controllability, stochastic co-controllability, stochastic observability, and stochastic co-observability,

$$
\begin{aligned}
&\forall\, t \in T,\ (F_t^+ \vee G_t^+,\, F_{t-1}^- \vee G_t^- \,|\, G_t) \in \mathrm{CI}_{\min},\ G_t \subseteq F_t^+ \vee F_{t-1}^- = F_\infty, \\
&\Leftrightarrow (F_t^+ \vee G_t,\, F_{t-1}^- \vee G_t |\, G_t) \in \mathrm{CI}_{\min},\ \text{if current-state } \sigma\text{-algebra family holds}, \\
&\Rightarrow (F_t^+,\, F_{t-1}^- |\, G_t) \in \mathrm{CI}_{\min}, \\
&\Rightarrow \sigma(F_t^+ | F_{t-1}^-) = \sigma(G_t |\, F_{t-1}^-),\ \sigma(F_{t-1}^- | F_t^+) = \sigma(G_t |\, F_t^+),\ \text{and} \\
&\quad \sigma(F_t^+ |\, G_t) = G_t,\ \sigma(F_{t-1}^- |\, G_t) = G_t.
\end{aligned}
$$

The above framework of stochastic realization applies regardless of which probability distributions are considered. From the past and the future σ-algebras one has to construct a minimal state σ-algebra which makes the past and the future minimally conditional independent. Then one can construct the stochastic system which is then a stochastic realization.

The filtering problem can now be regarded as a stochastic realization problem. One considers a stochastic realization in the form of a stochastic system, $(F_t^+ \vee G_t^+,\, F_{t-1}^- \vee G_t^- \,|\, G_t) \in \mathrm{CI}_{\min}$. The associated filter system is then the stochastic realization described by the formulas,

$$
\forall\, t \in T,\ \ G_{f,t} = \sigma(G_t |\, F_{t-1}^-),\ (F_t^+ \vee G_{f,t}^+,\, F_{t-1}^- \vee G_{f,t}^- |\, G_{f,t}) \in \mathrm{CI}.
$$

The filter system is such that at any time the state is measurable with respect to the past of the output process.

Correspondingly there is an interpretation of optimal stochastic control with complete observations. Consider a stochastic control system with complete observations and with a controlled-output process. The associated closed-loop system is another stochastic realization of the system. The closed-loop system associated with the optimal control law is then stochastic realization in which the state is measurable with respect to the past of the controlled output.

System identification can also be regarded from the view point of stochastic realization. This was first developed for system identification of Gaussian systems where it is called the *subspace-identification algorithm* or procedure. Suppose one has observed an output process y on a finite horizon $T = \{0,\ 1,\ 2,\ \ldots,\ t_1\}$. Take a time $t \in T$ near the middle of the time index set. Construct then the past and the future of the output process at time $t \in T$, $F^{y(0:t-1)} = \sigma(\{y(s),\ s = 0,\ldots,t-1\})$ and $F^{y(t:t_1)} = \sigma(\{y(s),\ s = t,\ t+1,\ \ldots,t_1\})$. Determine a σ-algebra G_t such that $(F^{y(t:t_1)}, F^{y(0:t-1)} |\ G_t) \in \mathrm{CI}$ and $G_t \subseteq F^{y(0:t-1)}$. For system identification one usually takes $G_t = \sigma(F^{y(t:t_1)} |\ F^{y(0:t-1)})$. The construction requires one to work with the conditioning displayed in the above projection operator. In the case of a stationary Gaussian process, the projection of the σ-algebras equals the σ-algebra generated by the associated conditional expectation $E[y(t:t_1) |\ F^{y(0:t-1)}]$. In case of other stochastic processes, the projection of the σ-algebras may lead to different formulas, for example a polynomial or a rational map. In practice, one determines a finite-dimensional random variable $x(.,t) : \Omega \to \mathbb{R}^{n_x}$ such that $G_t = F^{x(t)}$. The above formula then also leads to the relation of the output $y(t)$ and the state $x(t)$. Another projection operation, $\sigma(F^{x(t+1)} |\ F^{x(t)})$, allows one to construct the formula of the state recursion. This construction of a stochastic system for system identification is again universal for all stochastic systems, meaning with regard to the probability measures involved.

How is stochastic realization theory used in control and system theory? The condition of stochastic observability and of stochastic co-observability are used in filtering theory. The condition of stochastic controllability and of stochastic co-controllability are used in stochastic control of stochastic systems. In system identification the concept of a minimal realization has to be used, for which stochastic controllability, stochastic observability, etc. are the equivalent conditions. Moreover, the subspace identification procedure for Gaussian stochastic systems and its generalizations is based on stochastic realization theory. For system identification of other sets of systems, a corresponding approach based on the construction of the state space as used in stochastic realization is best explored.

The treatment of the stochastic realization problem for a finite-valued output process is still not satisfactory solved. Open are the stochastic realization problems for a counting process, for an output process in an interval of the real numbers, and for an output process in the positive real numbers. There are of course additional systems for which the stochastic realization problem can be formulated. The stochastic realization problem for the multiple conditional independence relation has not been investigated yet.

7.2 Stochastic Realization of a Tuple of Gaussian Random Variables

The reader may learn from this section the stochastic realization problem of a tuple of finite-dimensional Gaussian random variables and the structure of its solution. This will be useful when considering stochastic realization problems for other probability distributions and stochastic processes. The results of this section are also useful for information theory.

Concepts

For the benefit of the reader, the following definition is restated. The reader may recall the concept of conditional independence from Section 2.9 and from Section 19.8.

Definition 7.2.1. Consider a tuple of finite-dimensional Gaussian random variables,

$$y_1 : \Omega \rightarrow \mathbb{R}^{n_{y_1}},\ y_2 : \Omega \rightarrow \mathbb{R}^{n_{y_2}},\ (y_1,\ y_2) \in G(0, Q_{(y_1,y_2)}).$$

Define the *Gaussian conditional independence relation* of this tuple by the notation and conditions,

$$x : \Omega \rightarrow \mathbb{R}^{n_x}, (y_1,\ y_2,\ x) \in G,\ (F^{y_1},\ F^{y_2} |\ F^x) \in \text{CI};$$
denote the above conditions by $(F^{y_1},\ F^{y_2} |\ F^x) \in \text{CIG}$.

One also refers to the triple $(F^{y_1},\ F^{y_2} |\ F^x) \in \text{CIG}$ as a *Gaussian conditionally-independent triple*. Call the triple $(F^{y_1},\ F^{y_2} |\ F^x) \in \text{CIG}$ a *minimal Gaussian conditional-independent triple* if,

$$\begin{aligned} &(1) \quad (F^{y_1},\ F^{y_2} | F^x) \in \text{CIG}, \\ &(2) \quad z : \Omega \rightarrow \mathbb{R}^{n_z},\ F^z \subseteq F^x,\ (y_1, y_2, x, z) \in G,\ (F^{y_1},\ F^{y_2} |\ F^z) \in \text{CIG} \\ &\quad \Rightarrow F^z = F^x; \quad \text{denote these conditions by } (F^{y_1},\ F^{y_2} |\ F^x) \in \text{CIG}_{\text{min}}. \end{aligned}$$

The following result is recalled from Proposition 2.9.5. Consider a triple of Gaussian random variables $(y_1,\ y_2,\ x) \in G(0, Q)$ satisfying that $Q_x \in \mathbb{R}^{n_x \times n_x}_{spds}$. Then $(F^{y_1},\ F^{y_2} |\ F^x) \in \text{CIG}$ if and only if $Q_{y_1,y_2} = Q_{y_1,x} Q_x^{-1} Q_{y_2,x}^T$.

The canonical variable decomposition of a tuple of Gaussian random variables is used in this section, see Def. 19.4.5 of which the definition is recalled.

Definition 7.2.2. The tuple of jointly Gaussian random variables (y_1, y_2) with $y_1 : \Omega \rightarrow \mathbb{R}^{n_{y_1}}$ $y_2 : \Omega \rightarrow \mathbb{R}^{n_{y_2}}$ is said to be in the *canonical variable form* if the associated variance matrix has the form,

$$\begin{aligned}
Q_{(y_1,y_2)} &= \begin{pmatrix} I & Q_{12} \\ Q_{12}^T & I \end{pmatrix} \in \mathbb{R}^{(n_{y_1}+n_{y_2})\times(n_{y_1}+n_{y_2})}, \\
Q_{12} &= \begin{pmatrix} I_{n_{y_{11}}} & 0 & 0 \\ 0 & D_1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{n_{y_1}\times n_{y_2}}, \\
& n_{y_{11}},\ n_{y_{12}},\ n_{y_{13}},\ n_{y_{21}},\ n_{y_{22}},\ n_{y_{23}} \in \mathbb{N},\ n_{y_{11}} = n_{y_{21}},\ n_{y_{12}} = n_{y_{22}}, \\
& n_{y_1} = n_{y_{11}} + n_{y_{12}} + n_{y_{13}},\ n_{y_2} = n_{y_{21}} + n_{y_{22}} + n_{y_{23}},\ I \in \mathbb{R}^{n_{y_{11}}\times n_{y_{11}}}, \\
D_1 &= \text{Diag}(d_1,\ldots,d_{n_{y_{12}}}) \in \mathbb{R}^{n_{y_{12}}\times n_{y_{12}}},\ 1 > d_1 \geq \ldots \geq d_{n_{y_{12}}} > 0; \\
y_1 &= (y_{11},y_{12},y_{13})^T,\ y_2 = (y_{21},y_{22},y_{23})^T, \\
& \forall\ j = 1,2,3,\ y_{1j} : \Omega \to \mathbb{R}^{n_{1j}},\ y_{2j} : \Omega \to \mathbb{R}^{n_{2j}}.
\end{aligned}$$

Below the following special case will be used often.

$$\begin{aligned}
& (y_{12},y_{22}) \in G(0,Q_{y_{12},y_{22}}),\ n_{y_{12}} = n_{y_{22}},\ y_{12},\ y_{22} : \Omega \to \mathbb{R}^{n_{y_{12}}}, \\
Q_{y_1,y_2} &= \begin{pmatrix} I & D_1 \\ D_1 & I \end{pmatrix},\ D_2 = D_1^{-1} - D_1 \in \mathbb{R}^{n_x\times n_x}_{diag},\ D_2 \succ 0; \\
& D_1 D_2 = D_2 D_1,\ \text{equivalently, } D_1 \text{ and } D_2 \text{ commute.}
\end{aligned}$$

The reader is reminded of the convention to distinguish the following notations.

$$\begin{aligned}
Q_{y_1,y_2} &= E[y_1 y_2^T], \\
Q_{(y_1,y_2)} &= E\left[\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}^T\right] = \begin{pmatrix} Q_{y_1,y_1} & Q_{y_1,y_2} \\ Q_{y_1,y_2}^T & Q_{y_2,y_2} \end{pmatrix}.
\end{aligned}$$

The Problem

The expression *weak problem* signifies that stochastic realization is considered for a probability measure.

Problem 7.2.3. *Weak Gaussian stochastic realization problem.* Consider a Gaussian measure on a finite-dimensional probability space for a tuple of Gaussian random variables (y_1,y_2), rather than (y_+,y_-) as used above, specified by,

$$\left(\mathbb{R}^{n_{y_1}+n_{y_2}}, B(\mathbb{R}^{n_{y_1}+n_{y_2}}), G(0,Q_{(y_1,y_2)})\right).$$

(a) Construct all Gaussian measures on the finite-dimensional space,

$$\left(\mathbb{R}^{n_{y_1}+n_{y_2}+n_x}, B(\mathbb{R}^{n_{y_1}+n_{y_2}+n_x}), G(0,Q_{(y_1,y_2,x)})\right), n_x \in \mathbb{N},$$

such that, after defining the following random variables for the specified spaces, the following conditions hold,

$$\begin{aligned}
& (\Omega,F,P) = (\mathbb{R}^{n_{y_1}+n_{y_2}+n_x}, B(\mathbb{R}^{n_{y_1}+n_{y_2}+n_x}), G(0,Q_{(y_1,y_2,x)})), \\
& y_1 : \Omega \to \mathbb{R}^{n_{y_1}},\ y_1((\omega_1,\omega_2,\omega_3)) = \omega_1, \\
& y_2 : \Omega \to \mathbb{R}^{n_{y_2}},\ y_2((\omega_1,\omega_2,\omega_3)) = \omega_2,\ x : \Omega \to \mathbb{R}^{n_x},\ x((\omega_1,\omega_2,\omega_3)) = \omega_3; \\
& (F^{y_1},F^{y_2} \mid F^x) \in \text{CIG},\ G(0,Q_{(y_1,y_2,x)})|_{\mathbb{R}^{n_{y_1}+n_{y_2}}} = G(0,Q_{(y_1,y_2)}).
\end{aligned}$$

Call any such measure a *weak Gaussian stochastic realization* of the considered Gaussian measure.

(b)Formulate necessary and sufficient conditions such that, in addition to the conditions of (a), minimality holds,

$$(F^{y_1}, F^{y_2} \mid F^x) \in \text{CIG}_{\text{min}};$$

call any such measure a *minimal weak Gaussian stochastic realization* of the considered Gaussian measure $G(0, Q_{y_1,y_2})$.

(c)Classify or describe all measures as requested in (a) which are a minimal weak Gaussian stochastic realization.

The existence problem of a random variable x such that $(F^{y_1}, F^{y_2} \mid F^x) \in \text{CIG}$ is trivial, either $x = y_1$ or $x = y_2$ will do, see Proposition 2.9.3.(a). Below are discussed the characterization of minimality and the classification of minimal state variables.

Characterization of Minimality

Needed is a characterization of when a Gaussian conditional independent triple is a minimal such triple. For this the following concepts are useful.

Definition 7.2.4. *Stochastic observability and stochastic co-observability of Gaussian random variables*. Consider a Gaussian conditionally-independent triple of Gaussian random variables $(F^{y_1},\ F^{y_2} \mid F^x) \in \text{CIG}$.

Call the triple *stochastically observable* if the map,

$$x \mapsto E[\exp(iw^T y_1) | F^x],$$

is injective on the support of the random variable x.

Call the triple *stochastically co-observable* if the map,

$$x \mapsto E[\exp(iw^T y_2) | F^x],$$

is injective on the support of the random variable x.

Stochastic observability of a triple $(y_1,\ y_2,\ x) \in G$ is equivalent to the property that from the conditional characteristic function of y_1 conditioned on x, one can uniquely recover the value of the random variable x. The conditional characteristic function or the conditional distribution function can in principle be obtained or approximated by sampling many values of the random variable y_1. From either the estimated conditional distribution or from the conditional characteristic function, stochastic observability allows one to uniquely determine the value of the particular state related to that distribution or function. The concept of stochastic co-observability of y_2 conditioned on x has an analogous interpretation.

A characterization of stochastic observability and of stochastic co-observability is needed.

Proposition 7.2.5. Characterization of stochastic observability and stochastic co--observability. *Consider a conditionally-independent triple of Gaussian random variables* $(F^{y_1},\ F^{y_2}|\ F^x) \in \text{CIG}$. *Assume that a basis has been chosen such that* $(y_1,\ y_2,\ x) \in G(0,Q)$ *with* $x \in G(0,Q_x)$ *and* $Q_x \in \mathbb{R}_{spds}^{n_x \times n_x}$ *hence* $Q_x \succ 0$.

(a)The triple is stochastically observable if and only if $\text{rank}(Q_{y_1,x}) = n_x = \text{rank}(Q_x)$.
(b)The triple is stochastically co-observable if and only if $\text{rank}(Q_{y_2,x}) = n_x$.

Proof. (a) Consider the map from the state variable x to the conditional characteristic function of the output y_1 conditioned on the state variable,

$$x \mapsto E[\exp(iw^T y_1)|F^x] = \exp(iw^T E[y_1|F^x] - \frac{1}{2} w^T Q_{y_1|x} w),\ \forall\ w \in \mathbb{R}^{n_{y_1}},$$
$$E[y_1|F^x] = Q_{y_1,x} Q_x^{-1} x,\ \ Q_{y_1|x} = Q_{y_1} - Q_{y_1,x} Q_x^{-1} Q_{y_1,x}^T.$$

The latter formulas follow from Proposition 2.8.3. The map is injective on the support of x if and only if $\text{rank}(Q_{y_1,x} Q_x^{-1}) = \text{rank}(Q_x)$ if and only if $\text{rank}(Q_{y_1,x}) = n_x$, because, by the assumption that $Q_x \succ 0$, $\text{rank}(Q_x) = n_x$. Proof of (b) is analogous. □

Proposition 7.2.6. Dimension reduction *of the state variable of a weak Gaussian stochastic realization. Consider triple of random variables* $y_1,\ y_2,\ x$ *as in the previous proposition. If* $(F^{y_1}, F^{y_2}|\ F^x) \in \text{CIG}$, *and if one defines the random variables,*

$$x_1 = E[y_1|F^x],\ x_2 = E[y_2|F^{x_1}],\quad x_1 : \Omega \to \mathbb{R}^{n_{y_1}},\ x_2 : \Omega \to \mathbb{R}^{n_{y_2}},$$
$$\textit{then } (F^{y_1}, F^{y_2}|\ F^{x_2}) \in \text{CIG},\ \textit{and}$$
$$\text{rank}(Q_{y_1,y_2}) = \text{rank}(Q_{y_1,x_2}) = \text{rank}(Q_{x_2}) = \text{rank}(Q_{y_2,x_2}).$$

Proof. Because, by definition of x_1, $F^{x_1} \subseteq F^x$,

$$\begin{aligned} & E[y_1|F^{x_1} \vee F^{y_2}] = E[E[y_1|F^x \vee F^{y_2}]|F^{x_1} \vee F^{y_2}],\ \text{by Theorem 2.8.2.(d)},\\ &= E[E[y_1|F^x]|F^{x_1} \vee F^{y_2}],\ \text{by } (F_1, F_2|\ F^x) \in \text{CI},\\ &= E[y_1|F^x],\ \text{by definition of } x_1,\ E[y_1|F^x] \text{ is } F^{x_1}\text{-measurable},\\ &= E[E[y_1|F^x]|F^{x_1}],\ \text{by the same argument as above},\\ &= E[y_1|F^{x_1}], \text{by Theorem 2.8.2.(d).} \end{aligned}$$

and from Proposition 2.9.5 follows that $(F^{y_1}, F^{y_2}|\ F^{x_1}) \in \text{CIG}$. Analogously one proves that, $(F^{y_1}, F^{y_2}|\ F^{x_2}) \in \text{CIG}$. From Proposition 2.9.5 follows that $Q_{y_1,y_2} = Q_{y_1,x_2} Q_{x_2}^{-1} Q_{y_2,x_2}^T$. Note that by Proposition 2.8.3,

$$\begin{aligned}
x_1 \; &= \; E[y_1|F^x] = Q_{y_1,x}Q_x^{-1}x \;\;\Rightarrow\;\; Q_{x_1} = Q_{y_1,x}Q_x^{-1}Q_{y_1,x}^T \text{ and} \\
&\quad Q_{y_1,x_1} = E[y_1x_1^T] = Q_{y_1,x}Q_x^{-1}Q_{y_1,x}^T = Q_{x_1} \;\;\Rightarrow\;\; \text{rank}(Q_{y_1,x_1}) = \text{rank}(Q_{x_1}), \\
&\quad Q_{y_2,x_1} = E[y_2x_1^T] = Q_{y_2,x}Q_x^{-1}Q_{y_1,x}^T = Q_{y_2,y_1}, \text{ by conditional independence,} \\
&\Rightarrow \text{rank}(Q_{y_2,x_1}) = \text{rank}(Q_{y_2,y_1}) = \text{rank}(Q_{y_1,y_2}); \\
x_2 \; &= \; E[y_2|F^{x_1}] = Q_{y_2,x_1}Q_{x_1}^{-1}x_1 \;\;\Rightarrow\;\; Q_{x_2} = Q_{y_2,x_1}Q_{x_1}^{-1}Q_{y_2,x_1}^T \text{ and} \\
&\quad Q_{y_2,x_2} = E[y_2x_2^T] = Q_{y_2,x_1}Q_{x_1}^{-1}Q_{y_2,x_1}^T = Q_{x_2}, \; \text{rank}(Q_{x_2}) = \text{rank}(Q_{y_2,x_2}), \\
&\quad \text{rank}(Q_{x_2}) \leq \text{rank}(Q_{y_2,x_1}), \text{ by } Q_{x_2} = Q_{y_2,x_1}Q_{x_1}^{-1}Q_{y_2,x_1}^T, \\
&= \text{rank}(Q_{y_1,y_2}), \text{ by an above relation,} \\
&\leq \text{rank}(Q_{x_2}), \text{ by } (F^{y_1},F^{y_2}|\,F^{x_2}) \in \text{CI} \;\;\Rightarrow\;\; Q_{y_1,y_2} = Q_{y_1,x_2}Q_{x_2}^{-1}Q_{y_2,x_2}^T, \\
&\Rightarrow \text{rank}(Q_{y_1,y_2}) = \text{rank}(Q_{x_2}).
\end{aligned}$$

From the conditional independence $(F^{y_1},F^{y_2}|\,F^{x_2}) \in \text{CIG}$ one concludes that $Q_{y_1,y_2} = Q_{y_1,x_2}Q_{x_2}^{-1}Q_{y_2,x_2}^T$ is a minimal factorization. □

Theorem 7.2.7. Characterization of minimality *of a weak Gaussian stochastic realization. Consider a triple of Gaussian random variables satisfying,*

$$(F^{y_1},\; F^{y_2}|\,F^x) \in \text{CIG},\; y_1 : \Omega \to \mathbb{R}^{n_{y_1}},\; y_2 : \Omega \to \mathbb{R}^{n_{y_2}},\; x : \Omega \to \mathbb{R}^{n_x}.$$

Assume that a basis has been chosen such that and $Q_x \in \mathbb{R}^{n_x \times n_x}_{spds}$, hence $Q_x \succ 0$.
The following statements are equivalent:

(a)$(F^{y_1},\; F^{y_2}|F^x) \in \text{CIG}_{\text{min}}$;
(b)the external characterization*: $\text{rank}(Q_{y_1,y_2}) = \text{rank}(Q_x) = n_x$;*
(c)the intrinsic characterization*: the triple is both stochastically observable and stochastically co-observable.*

Proof. (b ⇔ c). From the assumption that $(F^{y_1},F^{y_2}|\; F^x) \in \text{CIG}$ and Proposition 2.9.5 follows that $Q_{y_1,y_2} = Q_{y_1,x}Q_x^{-1}Q_{y_2,x}^T$. It follows from consideration of the matrix product and linear algebra, that $\text{rank}(Q_{y_1,y_2}) = \text{rank}(Q_x)$ if and only if $\text{rank}(Q_{y_1,x}) = \text{rank}(Q_x^{-1}) = \text{rank}(Q_{y_2,x}^T)$. The latter condition is by Proposition 7.2.5 equivalent to (c).
(a ⇒ b) It follows from Proposition 2.9.5.(b) and $Q_{y_1,y_2} = Q_{y_1,x}Q_x^{-1}Q_{y_2,x}^T$, that $\text{rank}(Q_{y_1,y_2}) \leq \text{rank}(Q_x)$. Suppose that $\text{rank}(Q_{y_1,y_2}) < \text{rank}(Q_x)$. As in Proposition 7.2.6, define the random variables $x_1 = E[y_1|F^x]$ and $x_2 = E[y_2|F^{x_1}]$. From Proposition 7.2.6 follows that $(F^{y_1},F^{y_2}|\,F^{x_2}) \in \text{CIG}$ and $\text{rank}(Q_{x_2}) = \text{rank}(Q_{y_1,y_2}) < \text{rank}(Q_x)$. This and $F^{x_2} \subseteq F^x$ contradict the minimality of x. Thus $\text{rank}(Q_{y_1,y_2}) = \text{rank}(Q_x)$.
(b ⇒ a) Consider a random variable $x_1 : \Omega \to \mathbb{R}^{n_1}$ such that $F^{x_1} \subseteq F^x$, $(F^{y_1},F^{y_2}|F^{x_1}) \in \text{CIG}$, and $(y_1,y_2,x,x_1) \in G$. Suppose that a basis for x_1 has been chosen such that $n_{x_1} = \text{rank}(Q_{x_1})$. From the conditional independence and Proposition 2.9.5 follows that $Q_{y_1,y_2} = Q_{y_1,x}Q_x^{-1}Q_{y_2,x}^T$. From $(y_1,y_2,x,x_1) \in G$ and $F^{x_1} \subseteq F^x$ follows that $n_{x_1} \leq n_x$. Proposition 2.9.5 and the conditional independence with x_1 imply that $n_x = \text{rank}(Q_x) = \text{rank}(Q_{y_1,y_2}) \leq \text{rank}(Q_{x_1}) = n_{x_1}$ hence $n_x = n_{x_1}$. Thus the state variable x is of minimal dimension. □

Using the canonical variable decomposition of the observation vectors, the characterization of a minimal Gaussian conditional independence relation can now be formulated.

Proposition 7.2.8. *Consider a triple of Gaussian random variables $(y_1,\ y_2,\ x) \in G(0,Q)$ as formulated in Def. 7.2.1. Assume that the tuple $(y_1,\ y_2) \in G(0,Q_{(y_1,y_2)})$ is in the canonical variable form with $y_1 = (y_{11},\ y_{12},\ y_{13})$ and $y_2 = (y_{21},\ y_{22},\ y_{23})$ and such that $Q_x \in \mathbb{R}^{n_x \times n_x}_{spds}$. The following statements are equivalent:*

(a)$(F^{y_1},\ F^{y_2}|F^x) \in \mathrm{CIG}_{\min}$;
(b)there exists a basis for the random variable x such that,

$$\begin{aligned} x = &\begin{pmatrix} x_1 \\ x_2 \end{pmatrix},\ x_1 : \Omega \to \mathbb{R}^{n_{x_1}},\ x_2 : \Omega \to \mathbb{R}^{n_{x_2}}, \\ &n_{x_1} = n_{y_{11}} = n_{y_{21}},\ n_{x_2} = n_{y_{12}} = n_{y_{22}},\ n_x = n_{x_1} + n_{x_2}, \\ &(F^{y_{11}},\ F^{y_{21}}|\ F^{x_1}) \in \mathrm{CIG}_{\min},\ (F^{y_{12}},\ F^{y_{22}}|\ F^{x_2}) \in \mathrm{CIG}_{\min}. \end{aligned}$$

Proof. (a $\Rightarrow$ b) From the condition that $(F^{y_1}, F^{y_2}|\ F^x) \in \mathrm{CIG}$ and the use of the variables of the canonical variable decomposition follows that,

$$(F^{y_{11}} \vee F^{y_{12}}, F^{y_{21}} \vee F^{y_{22}}|\ F^x) \in \mathrm{CIG}.$$

Define the random variables, $x_1 : \Omega \to \mathbb{R}^{n_{x_1}}$ as $x_1 = E[x|F^{y_{11}}]$, $x_2 : \Omega \to \mathbb{R}^{n_{x_2}}$ as $x_2 = x - x_1$. Then $Q_{x_2,y_{11}} = E[x_2 y_{11}^T] = E[(x - E[x|F^{y_{11}})y_{11}^T] = 0$, hence x_2 and y_{11} are uncorrelated, and F^{x_2} and $F^{y_{11}}$ are independent. Because $y_{11} = y_{21}$ it follows that,

$$\begin{aligned} &(F^{y_{11}}, F^{y_{11}}|\ F^x) = (F^{y_{11}}, F^{y_{21}}|F^x) \in \mathrm{CIG}, \\ \Rightarrow\ &F^{y_{11}} \subseteq F^x, \text{ by Proposition 2.9.4, } \Rightarrow\ F^{x_1} \subseteq F^{y_{11}} \subseteq F^x, \end{aligned}$$

$x_2 = x - x_1$ implies that $F^{x_2} \subseteq (F^x \vee F^{x_1}) \subseteq F^x$. From $x = x_1 + x_2$ follows the converse inclusion, hence $F^{x_1} \vee F^{x_2} = F^x$. Next $F^{x_1} \subseteq F^{y_{11}} = F^{y_{21}}$ while by the above F^{x_2} is independent of $F^{y_{11}}$. Thus the following sub-σ-algebras $F^{y_{11}} \vee F^{y_{21}} \vee F^{x_1}$ and $F^{y_{12}} \vee F^{y_{22}} \vee F^{x_2}$ are independent. Then,

$$\begin{aligned} &F^{x_2},\ F^{y_{11}} \text{ independent, } F^{x_1} \subseteq F^{y_{11}}, \text{ and Proposition 19.8.5 imply that,} \\ &(F^{y_{11}} \vee F^{y_{12}}, F^{y_{21}} \vee F^{y_{22}}|F^{x_1} \vee F^{x_2}) \in \mathrm{CIG} \\ \Rightarrow\ &(F^{y_{11}}, F^{y_{21}}|F^{x_1}) \in \mathrm{CIG} \text{ and } (F^{y_{12}}, F^{y_{22}}|F^{x_2}) \in \mathrm{CIG}. \end{aligned}$$

From the minimality assumed in (a) and from Proposition 2.9.5 follows that $n_x = n_{y_{11}} + n_{y_{12}}$. Then,

$$\begin{aligned} n_x &= \mathrm{rank}(Q_x) = \mathrm{rank}(Q_{x_1}) + \mathrm{rank}(Q_{x_2}), \text{ by independence of } x_1 \text{ and } x_2, \\ &\geq \mathrm{rank}(Q_{y_{11},y_{21}}) + \mathrm{rank}(Q_{y_{12},y_{22}}), \\ &\quad \text{by the conditional independence established above,} \\ &= n_{y_{11}} + n_{y_{22}}, \text{ by the canonical variable decomposition,} \\ &= n_x, \text{ by the above relation.} \end{aligned}$$

Hence equality holds throughout, by dimensions rank$(Q_{x_1}) = n_{y_{11}}$ and rank$(Q_{x_2}) =$ rank$(Q_{y_{12},y_{22}}) = n_{y_{12}} = n_{y_{22}}$. Thus the minimality of the two conditional independence relations hold.
(b $\Rightarrow$ a). This is a simple verification of the equivalent condition specified in Theorem 7.2.7. □

Classification

Below the classification is described in the form of a set which parametrizes all measures describing measures of the form $G(0, Q_{y_1,y_2,x})$ which have the required conditional independent property.

Definition 7.2.9. The *set of weak Gaussian stochastic realizations*. Consider a Gaussian measure $G(0, Q_{y_1,y_2})$ on the space $(\mathbb{R}^{k_1+k_2}, B(\mathbb{R}^{k_1+k_2}))$. Define the set of *weak Gaussian stochastic realizations* for that space as,

$$\begin{aligned}
&\text{WGSR}\left(\mathbb{R}^{n_{y_1}+n_{y_2}},\ B(\mathbb{R}^{n_{y_1}+n_{y_2}}),\ G(0, Q_{(y_1,y_2)})\right) \\
&= \left\{ \begin{array}{l} (\mathbb{R}^{n_{y_1}+n_{y_2}+n_x},\ B(\mathbb{R}^{n_{y_1}+n_{y_2}+n_x}),\ G(0, Q_{(y_1,y_2,x)}))| \\ y_1 : \Omega \to \mathbb{R}^{n_{y_1}} = \omega_1,\ y_2 : \Omega \to \mathbb{R}^{n_{y_2}} = \omega_2,\ x : \Omega \to \mathbb{R}^{n_x} = \omega_3, \\ (F^{y_1}, F^{y_2} |\ F^x) \in \text{CIG},\ (y_1,\ y_2) \in G(0, Q_{(y_1,y_2)}) \end{array} \right\}.
\end{aligned}$$

The elements of WGSR,

$$(\mathbb{R}^{n_x}, B(\mathbb{R}^{n_x}, G(0, Q_{(y_1,y_2,x)}))),\ (\mathbb{R}^{n_{\bar{x}}}, B(\mathbb{R}^{n_{\bar{x}}}, G(0, Q_{(y_1,y_2,\bar{x})})) \in \text{WGSR},$$

are said to be *equivalent* if, up to a basis transformation of the underlying probability space, the state space dimensions are equal, $n_x = n_{\bar{x}}$ and the associated variance matrices are equal, $Q_{(y_1,y_2,x)} = Q_{(y_1,y_2,\bar{x})}$. In the following, the set WGSR will be identified with the set of equivalence classes obtained by dividing out the above equivalence relation.

The special case of dependent but not identical variables is considered first. The parametrization set is defined.

Definition 7.2.10. The *parameter set* WGSRP. Consider a tuple of random variables $(y_1, y_2) \in G(0, Q_{y_1,y_2})$ in the canonical variable form of Def. 7.2.2. Restrict attention in that definition to the correlated and nonidentical components,

$$\begin{aligned}
&(y_{12}, y_{22}) \in G(0, Q_{(y_{12},y_{22})}), \\
&y_{12} : \Omega \to \mathbb{R}^{n_{y_{12}}},\ y_{22} : \Omega \to \mathbb{R}^{n_{y_{22}}},\ n_{y_{12}} = n_{y_{22}}, \\
Q_{y_1,y_2} = &\begin{pmatrix} I & D_1 \\ D_1 & I \end{pmatrix},\ D_1 \in \mathbb{R}^{n_{y_{12}} \times n_{y_{12}}}_{s+,diag},\ D_2 = D_1^{-1} - D_1 \in \mathbb{R}^{n_{y_{12}} \times n_{y_{12}}}_{s+,diag}.
\end{aligned}$$

Define the integer $n_x = n_{y_{12}} = n_{y_{22}}$. Define the *weak Gaussian stochastic realization parametrization* for this particular tuple of random variables as the set WGSRP and define several other sets as,

$$\begin{aligned}
&\text{WGSRP}(\mathbb{R}^{n_{y_1}+n_{y_2}}, B(\mathbb{R}^{n_{y_1}+n_{y_2}}), G(0, Q_{(y_1,y_2)}))\\
&= \left\{ \begin{array}{l} Q \in \mathbb{R}^{n_x \times n_x}_{spds} | \\ Q - QD_2^{-1}Q - D_2^{-1} + QD_1D_2^{-1} + D_2^{-1}D_1Q \succeq 0 \end{array} \right\},\\
Q_{var} &= \left\{ Q \in \mathbb{R}^{n_x \times n_x}_{spds} | \begin{pmatrix} I & D_1 & D_1^{1/2} \\ D_1 & I & D_1^{1/2}Q \\ D_1^{1/2} & QD^{1/2} & Q \end{pmatrix} \succeq 0 \right\},\\
Q_{coneint}(n_x, D_1) &= \left\{ Q \in \mathbb{R}^{n_x \times n_x}_{spds} | D_1^{-1} \succeq Q \succeq D_1 \right\},\\
Q_{cone+}(n_x, D_1) &= \left\{ Q \in \mathbb{R}^{n_x \times n_x}_{spds} | Q \succeq D_1 \right\},\\
Q_{cone-}(n_x, D_1) &= \left\{ Q \in \mathbb{R}^{n_x \times n_x}_{spds} | D_1^{-1} \succeq Q \right\};\\
Q_{coneint}(n_x, D_1) &= Q_{cone+}(n_x, D_1) \cap Q_{cone-}(n_x, D_1).
\end{aligned}$$

Call Q_{cone+} the *forward cone* and Q_{cone-} the *backward cone* associated with $Q_{coneint}(n_x, D_1)$.

Both the forward cone $Q_{coneint}(n_x, D_1)$ and the backward cone $Q_{cone-}(n_x, D_1)$ of state variance matrices are convex sets hence is their intersection. The details of the structure of this convex body are of interest.

Example 7.2.11. Consider the special case of the set WGSRP_1 with $d_1 \in (0,1)$. Then,

$$\text{WGSRP} = \left\{ (d_1,\ d_1^{-1}) \subset \mathbb{R}_+ | \ \exists\ d_1 \in (0,1),\ 0 < d_1 < 1 < d_1^{-1} \right\}.$$

Example 7.2.12. Consider the special case of the set $\text{WGSRP}_1(2, D_1)$ with $D_1 = \text{Diag}(d_{1,1}, d_{1,2})$. Then,

$$\text{WGSRP}_1 = \{ Q \in \mathbb{R}^{2 \times 2}_{spd} | D_1 \leq Q \leq D_1^{-1} \},\ Q = \begin{pmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{pmatrix}.$$

The designer then selects $(q_{11}, q_{22}) \in \mathbb{R}_{s+}$ such that the conditions hold and subsequently chooses $q_{12} \in \mathbb{R}$ such that the conditions hold.

Lemma 7.2.13. *Consider a tuple of random variables*
$(y_1, y_2) \in G(0, Q_{y_1,y_2})$ *as defined in specified in Def. 7.2.10. Let* $n_x = n_{y_{12}} = n_{y_{22}}$.
The following statements are equivalent:
(a) $Q \in$ WGSRP *of Def. 7.2.10; (b)* $Q \in Q_{var}$*; (c)* $Q \in Q_{coneint}$.

Proof. Define for $Q \in \mathbb{R}^{n_x \times n_x}_{spds}$ the matrices,

$$Q_{(y_1,y_2)} = \begin{pmatrix} I & D_1 \\ D_1 & I \end{pmatrix} \succ 0,\ Q_{(y_1,y_2,x)} = \begin{pmatrix} I & D_1 & D_1^{1/2} \\ D_1 & I & D_1^{1/2}Q \\ D_1^{1/2} & QD_1^{1/2} & Q \end{pmatrix} \succeq 0.$$

An aside, the formula of $Q_{(y_1,y_2)}$ of [57, p. 123] is incorrect due to a typesetting error at the publisher, the above formula is the correct one.

Due to the fact that $D_1 = \text{Diag}(d_1, d_2, \ldots, d_{n_{y_{12}}})$ with $1 > d_1 \geq d_{n_{y_{12}}} > 0$, it follows that $I \succ D_1 \succ 0$, $D_2 = D_1^{-1} - D_1 \succ 0$, and that $Q_{(y_1,y_2)} \succ 0$ with the inverse matrix,

$$Q_{(y_1,y_2)}^{-1} = \begin{pmatrix} (I - D_1^2)^{-1} & -D_2^{-1} \\ -D_2^{-1} & (I - D_1^2)^{-1} \end{pmatrix} = \begin{pmatrix} D_1^{-1} D_2^{-1} & -D_2^{-1} \\ -D_2^{-1} & D_1^{-1} D_2^{-1} \end{pmatrix}.$$

Below Schur complements are used which are formulated in Proposition 17.4.33. The blocks considered differ by equivalence.

$$\begin{aligned}
L_1 &= \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ L_{31} & L_{32} & I \end{pmatrix}, \ \left(L_{31} \ L_{32} \right) = - \left(D_1^{1/2} \quad Q D_1^{1/2} \right) Q_{(y_1,y_2)}^{-1}, \\
& \quad L_1 Q_{(y_1,y_2,x)} L_1^T = \begin{pmatrix} I & D_1 & 0 \\ D_1 & I & 0 \\ 0 & 0 & Q_3 \end{pmatrix}, \\
Q_3 &= Q - \left(D_1^{1/2} \ Q D_1^{1/2} \right) Q_{y_1,y_2}^{-1} \begin{pmatrix} D_1^{1/2} \\ D_1^{1/2} Q \end{pmatrix} \\
&= Q - Q D_2^{-1} Q - D_2^{-1} + Q D_1 D_2^{-1} + D_1 D^{-2} Q; \\
& \quad Q_{(y_1,y_2,x)} \succeq 0 \ \Leftrightarrow \ Q_3 \succeq 0 \\
& \Leftrightarrow Q - Q D_2^{-1} Q - D_2^{-1} + Q D_1 D_2^{-1} + D_1 D^{-2} Q \succeq 0; \\
& \Leftrightarrow \begin{cases} (1) \begin{pmatrix} I & D_1^{1/2} \\ D_1^{1/2} & Q \end{pmatrix} \succeq 0 \ \Leftrightarrow \ I - D_1^{1/2} Q^{-1} D_1^{1/2} \succeq 0 \\ \quad \Leftrightarrow \ D_1^{-1} \succeq Q^{-1} \ \Leftrightarrow \ Q \succeq D_1, \\ (2) \begin{pmatrix} I & D_1^{1/2} Q \\ Q D_1^{1/2} & Q \end{pmatrix} \succeq 0 \ \Leftrightarrow \ I - D_1^{1/2} Q D_1^{1/2} \succeq 0 \ \Leftrightarrow \ D_1^{-1} \succeq Q, \\ (3) \ Q_{(y_1,y_2)} \succeq 0, \text{ which holds by assumption.} \end{cases}
\end{aligned}$$

The equivalence leading to the conditions (1), (2), and (3) above is due to the positive-definiteness of $Q_{(y_1,y_2,x)} \succeq 0$ being equivalent to the positive-definiteness of the three principle submatrices of the latter matrix. The equivalence of $Q \succeq D_1$ and $D^{-1} \succeq Q^{-1}$ follows from Lemma 17.4.34. □

Next a special case of the classification is proven. In this special case, the tuple of random variables (y_1, y_2) has a particular form as specified by the canonical variable decomposition. Afterwards the general classification result is stated.

Theorem 7.2.14. Classification in a special case. *Consider a tuple of random variables $(y_1, y_2) \in G(0, Q_{y_1,y_2})$ as defined in Problem 7.2.3. Consider the canonical variable decomposition of such a tuple and restrict attention to only the components (y_{12}, y_{22}) as defined in the canonical variable representation, with the probability measure,*

$$(y_{12}, y_{22}) \in G(0, Q_{(y_{12},y_{22})}), \ Q_{(y_{12},y_{22})} = \begin{pmatrix} I & D_1 \\ D_1 & I \end{pmatrix}, \ D_1 \in \mathbb{R}_{diag,s+}^{n_{y_{12}} \times n_{y_{12}}}.$$

Define the integer $n_x = n_{y_{12}} = n_{y_{22}}$. Recall the parametrization set of Def. 7.2.10.

Define the realization map r_1 by the formulas,

$$\begin{aligned} n_x &= n_{y_{12}} = n_{y_{22}},\ 3n_x = n_{y_{12}} + n_{y_{22}} + n_x, \\ r_1(Q) &= (\mathbb{R}^{3n_x}, B(\mathbb{R}^{3n_x}), G(0, Q_{(y_{12},y_{22},x)})),\ r_1 : \text{WGSRP} \to \text{WGSR}_{\min}, \\ Q_{(y_{12},y_{22},x)} &= \left\{ \begin{array}{cc|c} I & D_1 & D_1^{1/2} \\ D_1 & I & D_1^{1/2}Q \\ \hline D_1^{1/2} & QD_1^{1/2} & Q \end{array} \right) \in \mathbb{R}^{3n_x \times 3n_x}. \end{aligned}$$

Then the map r_1 is well defined and a bijection. This provides the requested classification of measures of Problem 7.2.3 for the special case defined above.

Proof. (1) It is proven that the map r_1 is well defined. Set $n_x = n_{y_{12}} = n_{y_{22}}$. If $Q_x \in \text{WGSRP}$ then it follows from Def. 7.2.10 and from Lemma 7.2.13 that $Q_x \succ 0$ and, with $3n_x = n_{y_{12}} + n_{y_{21}} + n_x$, $Q_{(y_1,y_2,x)} \in \mathbb{R}_{pds}^{3n_x \times 3n_x}$. Then $G(0, Q_{(y_1,y_2,x)})$ is a well defined Gaussian measure on the space $(\mathbb{R}^{3n_x}, B(\mathbb{R}^{3n_x}))$. Denote the canonical variables of this space by (y_{12}, y_{22}) which is correct because the measure $G(0, Q_{(y_1,y_2,x)})$ when restricted to the first two components equals the measure $G(0, Q_{(y_1,y_2)})$. Furthermore, from the formula $Q_{(y_1,y_2,x)}$ follows that,

$$\begin{aligned} &Q_{y_{12},x} Q_{x_2}^{-1} Q_{x,y_{22}} = D_1^{1/2} Q_x^{-1} Q_x D_1^{1/2} = D_1 = Q_{y_{12},y_{22}}, \\ &n_x = \text{rank}(Q_x) = \text{rank}(D_1) = \text{rank}(Q_{y_1,y_2}),\ \text{and} \\ &n_x = \text{rank}(Q_{y_1,x}) = \text{rank}(D_1) = \text{rank}(Q_x) = \text{rank}(D_1^{1/2} Q_x) = \text{rank}(Q_{y_{22},x_2}). \end{aligned}$$

With Proposition 7.2.7 one obtains that $(F^{y_{12}}, F^{y_{22}} | F^x) \in \text{CIG}_{\min}$.
(2) Surjectivenes of r_1. Because of the minimality in $(F^{y_{12}}, F^{y_{22}} |\ F^x) \in \text{CIG}_{\min}$ and of Proposition 7.2.7, one may choose a minimal basis for the state variable x of size $n_x \in \mathbb{Z}_+$; then $Q_x \succ 0$. From $(F^{y_{11}}, F^{y_{21}} |\ F^x) \in \text{CIG}$ follows that $D_1 = Q_{y_{12},y_{22}} = Q_{y_{12},x} Q_x^{-1} Q_{y_{22},x}^T$. Considerations of dimensions and the minimality imply by Proposition 7.2.7 that $Q_{y_1,x} \in \mathbb{R}^{n_{y_{11}} \times n_x} = \mathbb{R}^{n_x \times n_x}$ is nonsingular. Define $x_1 = D_1^{1/2} Q_{y_1,x}^{-T} x$. Then $Q_{y_1,x_1} = D_1^{1/2}$. Because $F^{x_1} = F^x$ it is true that $(F^{y_{12}}, F^{y_{22}} |\ F^{x_1}) \in \text{CIG}_{\min}$ and thus $D_1 = Q_{y_1,y_2} = Q_{y_1,x_1} Q_{x_1}^{-1} Q_{y_2,x_1}^T$. Consequently, $Q_{x_1} Q_{y_1,x_1}^{-1} D_1 = Q_{y_2,x_1}^T$ and $Q_{y_2,x_1} = D_1 Q_{y_1,x_1}^{-T} Q_{x_1} = D_1^{1/2} Q_{x_1}$. Then, $F^x = F^{x_1}$, $Q_{x_1} \in \mathbb{R}_{spds}^{n_x \times n_x}$, and $(y_1, y_2, x_1) \in G(0, Q_{(y_1,y_2,x_1)})$ with $Q_{(y_1,y_2,x_1)}$ as specified in the theorem statement. Because $Q_{(y_1,y_2,x_1)} \succeq 0$, it follows from Lemma 7.2.13 that $Q_{x_1} \in \text{WGSRP}$.
(3) Injectiveness. As indicated in Step (2) above, one may choose a basis for the probability space, such that with respect to this basis the corresponding covariance matrix has the form of $Q_{(y_1,y_2,x)}$ as displayed in the theorem statement. Consider two measures which are distinguished by the random variables x and z which produce the same realization, hence,

$$((n_{y_1}, n_{y_2}, n_x), G(0, Q_{(y_1,y_2,x)})) = ((n_{y_1}, n_{y_2}, n_x), G(0, Q_{(y_1,y_2,z)})) \in \text{WGSR}_{\min},$$

It is assumed that in both cases one has chosen a basis such that (y_1, y_2) are in the canonical variable form and that the two measures have the representation as in the theorem statement. Then it follows that the two state variance matrices $Q_x = Q_z \in \text{WGSR}_1$ of the domain of r_1 are equal. □

Theorem 7.2.15. Classification of $WGSR$ in the general case. *Consider a tuple of random variables $(y_1, y_2) \in G(0, Q_{(y_1,y_2)})$ as defined in Problem 7.2.3. Consider the canonical variable decomposition of such a tuple as formulated in Def. 7.2.2 and assume that a basis for these variables has been chosen such that the matrix $Q_{(y_1,y_2)}$ is in the canonical variable decomposition of Def. 7.2.2 with the diagonal matrix D_1, which notation is used below.*

Recall the parametrization set and define the realization map r by the formulas,

$$\begin{aligned}
&\text{WGSRP}(\mathbb{R}^{n_{y_1}+n_{y_2}}, B(\mathbb{R}^{n_{y_1}+n_{y_2}}), G(0, Q_{(y_1,y_2)})),\\
&r : \text{WGSRP} \to \text{WGSR}_{\text{min}},\\
r(Q) = {} &(\mathbb{R}^{n_{y_1}+n_{y_2}+n_x},\ B(\mathbb{R}^{n_{y_1}+n_{y_2}+n_x}), G(0, Q_{(y_1,y_2,x)})),\\
&n_{x_1} = n_{y_{11}} = n_{y_{21}},\ n_{x_2} = n_{y_{12}} = n_{y_{22}},\ n_x = n_{x_1} + n_{x_2};
\end{aligned}$$

$$Q_{(y_1,y_2,x)} = \left(\begin{array}{ccc|ccc|cc}
I & 0 & 0 & I & 0 & 0 & I & 0 \\
0 & I & 0 & 0 & D_1 & 0 & 0 & D_1^{1/2} \\
0 & 0 & I & 0 & 0 & 0 & 0 & 0 \\
\hline
I & 0 & 0 & I & 0 & 0 & I & 0 \\
0 & D_1 & 0 & 0 & I & 0 & 0 & D_1^{1/2}Q \\
0 & 0 & 0 & 0 & 0 & I & 0 & 0 \\
\hline
I & 0 & 0 & I & 0 & 0 & I & 0 \\
0 & D_1 & 0 & 0 & QD_1 & 0 & 0 & Q
\end{array}\right) \in \mathbb{R}_{pds}^{(n_{y_1}+n_{y_2}+n_x)\times(n_{y_1}+n_{y_2}+n_x)}.$$

Then the map r is a bijection and thus classifies all weak Gaussian stochastic realizations of the considered tuple of random variables.

Proof. (1) It will be shown that the map r is well defined. Consider a matrix $Q \in \text{WGSRP}$. Using Lemma 7.2.13.(a & b), one obtains that $Q_{(y_1,y_2,x)} = Q_{(y_1,y_2,x)}^T \succeq 0$. The Gaussian measure $G(0, Q_{(y_1,y_2,x)})$ is thus well defined. Note that,

$$\text{rank}(Q_x) = \text{rank}\begin{pmatrix} I & 0 \\ 0 & Q \end{pmatrix} = \text{rank}(I) + \text{rank}(Q) = n_{x_1} + n_{x_2} = n_x.$$

The following calculation, based on the structure of the matrix $Q_{(y_1,y_2,x)}$, establishes that,

$$\begin{aligned}
Q_{(y_1,x)} Q_x^{-1} Q_{(y_2,x)} &= \begin{pmatrix} I & 0 \\ 0 & D_1^{1/2} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & Q \end{pmatrix}^{-1} \begin{pmatrix} I & 0 & 0 \\ 0 & QD_1^{1/2} & 0 \end{pmatrix} \\
= \begin{pmatrix} I & 0 & 0 \\ 0 & D_1 & 0 \\ 0 & 0 & 0 \end{pmatrix} &= Q_{(y_1,y_2)},
\end{aligned}$$

$$\text{rank}(Q_{(y_1,x)}) = \text{rank}(Q_x) = \text{rank}(Q_{(y_2,x)})) = \text{rank}(Q_{(y_1,y_2)}) = n_x = n_{x_1} + n_{x_2}.$$

Using the above expression and Theorem 7.2.7 one concludes that $(F^{y_1}, F^{y_2} | F^x) \in \text{CIG}_{\text{min}}$, and $\left(\mathbb{R}^{n_{y_1}+n_{y_2}+n_x},\ B(\mathbb{R}^{n_{y_1}+n_{y_2}+n_x}),\ G(0, Q_{y_1,y_2,x})\right) \in WGSR_{min}$.

(2) Surjectiveness. Because of the minimality in $(F^{y_1}, F^{y_2} |\ F^x) \in \mathrm{CIG}_{\min}$ and because of Proposition 7.2.8 one may take a basis for the state random variable x of dimension $n_x = n_{y_{11}} + n_{y_{12}}$. From Proposition 7.2.8 follows that a basis for x_1 can be chosen such that $x_1 = y_{11} = y_{21}$ *a.s.*, $F^x = F^{x_1} \vee F^{x_2}$, $(F^{y_{11}}, F^{y_{21}} |\ F^{x_1}) \in \mathrm{CIG}_{\min}$, and $(F^{y_{12}}, F^{y_{22}} |\ F^{x_2}) \in \mathrm{CIG}_{\min}$. It follows from $(F^{y_{12}}, F^{y_{22}} |\ F^{x_2}) \in \mathrm{CIG}_{\min}$ and Theorem 7.2.14 that there exists a matrix $Q_{x_2} \in \mathrm{WGSRP}$ such that $r_1(Q_{x_2}) = (\mathbb{R}^{n_{y_{12},y_{22}}},\ B(\mathbb{R}^{n_{y_{12},y_{22}}}),\ G(0, Q_{(y_{12},y_{22},x_2)}))$. Thus $Q_{x_2} \in \mathrm{WGSRP}$ and $x = (x_1\ x_2)^T$ imply that $r(Q) = G(0, Q_{(y_1,y_2,x)})$.

(3) Injectiveness. Consider two domain values with equal range as in $(\mathbb{R}^{n_1}, B(\mathbb{R}^{n_1}), P_1) = (\mathbb{R}^{n_2}, B(\mathbb{R}^{n_2}), P_2) \in \mathrm{WGSRP}_{min}$. Because of the surjectiveness of the map r proven in Step (2) above, there exist domain values $Q_1,\ Q_2 \in \mathrm{WGSRP}$ such that $r(Q_1) = r(Q_2)$. Then $n_1 = n_2$ and, upto a selected basis chosen,

$$\begin{aligned} & G(0, Q_{(y_1,y_2,x)}) = G(0, Q_{(y_1,y_2,\bar{x})}) \\ \Rightarrow\ & Q_{(y_1,y_2,x)} = Q_{(y_1,y_2,\bar{x})}\ \Rightarrow\ Q_{x_1} = Q_{\bar{x}_1} = I,\ Q_{x_2} = Q_{\bar{x}_2}. \end{aligned}$$

□

Strong Stochastic Realization of a Tuple of Gaussian Random Variables

Problem 7.2.16. The *strong Gaussian stochastic realization of a tuple of Gaussian random variables* (SGSR problem). Consider a Gaussian measure on a finite-dimensional probability space and the representation of these variables according to,

$$\big(\Omega, F, G(0, Q_{(y_1,y_2)})\big),\ y_1 : \Omega \to \mathbb{R}^{n_{y_1}},\ y_2 : \Omega \to \mathbb{R}^{n_{y_2}},\ (y_1, y_2) \in G(0, Q_{(y_1,y_2)}).$$

(a) *Existence*. Construct a random variable x on the same probability space, $x : \Omega \to \mathbb{R}^{n_x}$ and the σ-algebra F^x, such that the following conditions all hold,

$$(F^{y_1}, F^{y_2} |\ F^x) \in \mathrm{CIG},\ (y_1, y_2, x) \in G(0, Q_{(y_1,y_2,x)}),\ F^x \subseteq F^{y_1} \vee F^{y_2}.$$

Call any triple (y_1, y_2, x) a *strong Gaussian stochastic realization* of the considered tuple of random variables $(y_1,\ y_2)$ and call the random variable x the *state* of the triple.

(b) *Minimality*. Formulate necessary and sufficient conditions such that, in addition to the conditions of (a), minimality holds,

$$(F^{y_1}, F^{y_2} |\ F^x) \in \mathrm{CIG}_{\min};$$

call any such triple (y_1, y_2, x) a *minimal strong Gaussian stochastic realization* of the considered Gaussian triple (y_1, y_2);

(c) *Classification*. Classify or describe all state variables x requested in (b) each of which is thus a minimal strong Gaussian stochastic realization.

Denote the set of minimal strong Gaussian stochastic realizations,

$$SGSR_{min} = \left\{ \begin{array}{l} (n_x, x) |\ n_x \in \mathbb{N},\ x : \Omega \to \mathbb{R}^{n_x}\ (y_1, y_2, x) \in G(0, Q_{(y_1,y_2,x)}), \\ F^x \subseteq (F^{y_1} \vee F^{y_2}),\ (F^{y_1}, F^{y_2} |\ F^x) \in \mathrm{CIG}_{\min} \end{array} \right\}.$$

States in the Frame σ-algebra

Theorem 7.2.17. *Consider a triple of Gaussian random variables $y+,\ y_-,\ x \in G$ as defined in Problem 7.2.16. Assume that $(F^{y_1}, F^{y_2} |\ F^x) \in \text{CIG}$.*

Then,

$$
\begin{aligned}
& G \subseteq F_f = \sigma(F^{y_-} |\ F^{y_+}) \vee \sigma(F^{y_+} |\ F^{y_-}) \\
\Leftrightarrow\ & \sigma(F^{y_+} |\ F^{y_-}) = \sigma(F^x |\ F^{y_-}),\ \sigma(F^{y_-} |\ F^{y_+}) = \sigma(F^x |\ F^{y_+}).
\end{aligned}
$$

Proof. This follows from Theorem 7.3.8. □

Theorem 7.2.18. *Consider a triple of Gaussian random variables $y+,\ y_-,\ x \in G$ as defined in Problem 7.2.16. Assume that $(F^{y_1}, F^{y_2} |\ F^x) \in \text{CIG}$ and that $0 \prec Q_{y_+}$, $0 \prec Q_{y_-}$, and $0 \prec Q_x$.*

(a)$\sigma(F^{y_+} |\ F^{y_-}) = \sigma(F^x |\ F^{y_-})$ *if and only if* $\text{Range}(Q_{y_+,x}) = \text{Range}(Q_{y_+,y_-})$ *and* $\text{rank}(Q_{y_+,x}) = n_x$.

(b)$\sigma(F^{y_-} |\ F^{y_+}) = \sigma(F^x |\ F^{y_+})$ *if and only if* $\text{Range}(Q_{y_-,x}) = \text{Range}(Q_{y_-,y_+})$ *and* $\text{rank}(Q_{y_-,x}) = n_x$.

Proof. (a) The joint-Gaussianness of the random variables $(y_+,\ y_-,\ x)$ and Theorem 2.8.3 imply that,

$$
\begin{aligned}
& E\left[\exp(iw^T\ y_+) |\ F^{y_-})\right] = \exp(iw^T\ E[y_+ |\ F^{y_-}] - w^T Q_{y_+|y_-} w/2), \\
& \forall\, w \in \mathbb{R}^{n_{y_+}};\ \ \Rightarrow\ \sigma(F^{y_+} |\ F^{y_-}) = \sigma(\{E[y_+ |\ F^{y_-}]\}).
\end{aligned}
$$

Similarly, $\sigma(F^x |\ F^{y_-}) = \sigma(E[x |\ F^{y_-}])$.

The conditional independence $(F^{y_1}, F^{y_2} |\ F^x) \in \text{CIG}$ and Theorem 2.8.3 imply,

$$
\begin{aligned}
& E[y_+ |\ F^{y_-}] = E[E[y_+ |\ F^{y_-} \vee F^x]\ |\ F^{y_-}] \\
& = E[E[y_+ |\ F^x] |\ F^{y_-}],\ \text{by conditional independence}, \\
& = E[Q_{y_+,x} Q_x^{-1} x |\ F^{y^-}] = Q_{y_+,x} Q_x^{-1}\ E[x |\ F^{y_-}] = LE[x |\ F^{y_-}],\ \ L = Q_{y_+,x} Q_x^{-1}.
\end{aligned}
$$

($\Leftarrow$) From the equality $E[y_+ |\ F^{y_-}] = LE[x |\ F^{y_-}]$ follows that $E[y_+ |\ F^{y_-}]$ is a function of $E[x |\ F^{y_-}]$. Hence,

$$
\sigma(F^{y_+} |\ F^{y_-}) = \sigma(E[y_+ |\ F^{y^-}]) \supseteq \sigma(E[x |\ F^{y_-}]) = \sigma(F^x |\ F^{y_-}).
$$

From the assumption $\text{rank}(Q_{y_+,x}) = n_x$ follows that the map L is injective hence there exists a left-inverse matrix $L_l^{-1} \in \mathbb{R}^{n_x \times n_{y_+}}$. Then $L_l^{-1} E[y_+ |\ F^{y_-}] = E[x |\ F^{y_-}]$ and $E[x |\ F^{y_-}]$ is a function of $E[y_+ |\ F^{y_-}]$. The same argument as used above then implies that,

$$
\begin{aligned}
& \sigma(F^{y_+} |\ F^{y_-}) = \sigma(E[y_+ |\ F^{y^-}]) \subseteq \sigma(E[x |\ F^{y_-}]) = \sigma(F^x |\ F^{y_-}); \\
& \Rightarrow\ \sigma(F^{y_+} |\ F^{y_-}) = \sigma(F^x |\ F^{y_-}).
\end{aligned}
$$

($\Rightarrow$) From the above follows that,

$$
\sigma(F^{y_+} |\ F^{y_-}) = \sigma(\{E[y_+ |\ F^{y_-}]\}),\ \ \sigma(F^x |\ F^{y_-}) = \sigma(E[x |\ F^{y_-}]).
$$

As argued above, the conditional independence and the joint-Gaussianness of the random variables imply that $E[y_+|\ F^{y_-}] = LE[x|\ F^{y_-}]$. It follows from Proposition 2.5.14 that $\sigma(\{E[y_+|\ F^{y_-}]\}) = \sigma(E[x|\ F^{y_-}])$ if and only if the function $E[y_+|\ F^{y_-}] = LE[x|\ F^{y_-}]$ is injective if and only if $n_x = \text{rank}(L) = \text{rank}(Q_{y_+,x})$ by Proposition 17.4.2.

(b) This result follows from (a) by symmetry for the variables y_+ and y_-. □

Classification

Theorem 7.2.19. *A* special case of the SGSR problem. *Consider the special case of Problem 7.2.16 in which the tuple of random variables (y_1,y_2) is in the canonical variable decomposition of Def. 7.2.2,*

$$(y_{12},y_{22}) \in G(0,Q_{(y_{12},y_{22})}),\ Q_{(y_{12},y_{22})} = \begin{pmatrix} I & D_1 \\ D_1 & I \end{pmatrix},\ D_1 \in \mathbb{R}^{n_{y_1}\times n_{y_1}}_{s+,diag},$$
$$D_2 = D_1^{-1} - D_1 \in \mathbb{R}^{n_{y12}\times n_{y12}}_{s+,diag}.$$

Define the integer $n_x = n_{y_{12}} = n_{y_{22}}$ and the parameter set,

$$\begin{aligned}&\text{SGSRP}(n_{y_{12}},n_{y_{22}},y_{12},y_{22},G(0,Q_{(y_{12},y_{22})}))\\ &= \left\{ Q \in \mathbb{R}^{n_x\times n_x}_{spds} | Q = QD_2^{-1}Q - D_2^{-1} + QD_2^{-1}D_1 + D_1D_2^{-1}Q \right\}.\end{aligned}$$

Note that the equation for the matrix Q inside the definition of SGSRP *is an algebraic Riccati equation for which the reader is referred to Chapter 22. Define the realization map,*

$$\begin{aligned}r_1(Q) &= (n_x,x),\ r_1 : \text{SGSRP} \to SGSR_{min} \text{ where},\\ &Q \in \text{SGSRP},\ n_x = n_{y_{12}} = n_{y_{22}},\ x : \Omega \to \mathbb{R}^{n_x},\\ P_1 &= (D_1^{-1} - Q)D_1^{1/2}D_2^{-1},\ P_2 = (Q - D_1)D_1^{-1/2}D_2^{-1} \in \mathbb{R}^{n_x\times n_x},\\ x &= P_1y_{12} + P_2y_{22}.\end{aligned}$$

Then the realization map r_1 is a bijection with respect to the chosen basis for the random variables (y_1,y_2).

The proof is based on the following lemma.

Lemma 7.2.20. *Consider the strong Gaussian stochastic realization problem of Problem 7.2.16. Consider the notation of Theorem 7.2.19. Define the variance matrix $Q_{(y_1,y_2,x)}$ and the set of variance matrices,*

$$Q_{(y_1,y_2,x)}(Q) = \left(\begin{array}{cc|c} I & D_1 & D_1^{1/2} \\ D_1 & I & D_1^{1/2}Q \\ \hline D_1^{1/2} & QD_1^{1/2} & Q \end{array}\right);$$

$$Q_{varrank} = \left\{ \begin{array}{l} Q \in \mathbb{R}^{n_x\times n_x}_{spds} |\ Q_{(y_{12},y_{22},x)}(Q) \succeq 0, \\ \text{rank}(Q_{(y_{12},y_{22},x)}(Q)) = n_{y_{12}} + n_{y_{22}} = 2n_x \end{array} \right\}.$$

Then $Q \in \text{SGSRP}$ *if and only if* $Q \in Q_{varrank}$.

Proof. Recall that $Q \in \text{SGSRP}$ implies that $Q \succ 0$, and that by the conditions on the diagonal elements of the matrix D_1, $\text{rank}(Q_{(y_1,y_2)}) = 2n_x$. It follows by taking a Schur complement, Proposition 17.4.33, that there exists a matrix L_1 such that,

$$
\begin{aligned}
L_1 Q_{(y_1,y_2,x)}(Q) L_1^T &= L_1 \left(\begin{array}{cc|c} I & D_1 & D_1^{1/2} \\ D_1 & I & D_1^{1/2}Q \\ \hline D_1^{1/2} & QD_1^{1/2} & Q \end{array}\right) L_1^T = \left(\begin{array}{cc|c} I & D_1 & 0 \\ D_1 & I & 0 \\ \hline 0 & 0 & Q_3 \end{array}\right), \\
\begin{pmatrix} I & D_1 \\ D_1 & I \end{pmatrix}^{-1} &= \begin{pmatrix} D_2^{-1}D_1^{-1} & -D_2^{-1} \\ -D_2^{-1} & D_2^{-1}D_1^{-1} \end{pmatrix}, \\
Q_3 &= Q - \begin{pmatrix} D_1^{1/2} & QD_1^{1/2} \end{pmatrix} \begin{pmatrix} I & D_1 \\ D_1 & I \end{pmatrix}^{-1} \begin{pmatrix} D_1^{1/2} \\ D_1^{1/2}Q \end{pmatrix} \\
&= Q - QD_2^{-1}Q - D_2^{-1} + QD_2^{-1}D_1 + D_1D_2^{-1}Q.
\end{aligned}
$$

Then it follows that,

$$
Q_{(y_1,y_2,x)} \succeq 0 \;\Leftrightarrow\; Q_3 \succeq 0; \;\; \text{rank}(Q_{(y_1,y_2,x)}) = 2n_x \;\Leftrightarrow\; Q_3 = 0 \;\Leftrightarrow\; Q \in Q_{varrank}.
$$

□

Proof. Proof of Theorem 7.2.19. (1) It is proven that the map r_1 is well defined. Because $Q \in \text{SGSRP}$, $Q \in \mathbb{R}_{spds}^{n_x \times n_x}$. Thus Q is a solution of the algebraic Riccati equation of the definition of SGSRP. It follows from Lemma 7.2.20 that the matrix $Q_{(y_1,y_2,x)} \succeq 0$. Define the matrices P_1 and P_2 as in the theorem statement and $x = P_1 y_1 + P_2 y_2$. One then calculates,

$$
\begin{aligned}
Q_{y_1,x} &= E[y_1 x^T] = P_1^T + D_1 P_2^T = D_1^{1/2}, \\
Q_{y_2,x} &= E[y_2 x^T] = D_1 P_1^T + P_2^T = D_1^{1/2} Q, \quad Q_x = E[xx^T] = Q, \\
& Q_{y_1,x} Q_x^{-1} Q_{y_2,x}^T = D_1^{1/2} Q_x^{-1} Q_x D_1^{1/2} = D_1 = Q_{y_1,y_2}, \\
& \text{rank}(Q_{y_1,x}) = n_x = \text{rank}(Q_x) = \text{rank}(Q_{y_2,x}), \\
& \Rightarrow (F^{y_1}, F^{y_2} \,|\, F^x) \in \text{CIG}_{\min}, \text{ by Theorem 7.2.7, } F^x \subseteq (F^{y12} \vee F^{y22}).
\end{aligned}
$$

Thus the map r_1 is well defined.

(2) Surjectiveness. Let $(n_x, x) \in \text{SGSRP}_1$. Then $(F^{y_1}, F^{y_2} \,|\, F^x) \in \text{CIG}_{\min}$ and $(y_1, y_2, x) \in G$. As in the proof of Theorem 7.2.15, one may choose a basis for x such that $Q_{y_1,x} = D_1^{1/2}$ and $Q_{y_2,x} = D_1^{1/2} Q$. It has to be proven that $Q_x \in \text{SGSRP}$ which is the case if Q_x satisfies the algebraic Riccati equation. Because $(y_1, y_2, x) \in G$ it follows that,

$$
\begin{aligned}
Q_{(y_1,y_2,x)} &= \left(\begin{array}{cc|c} I & D_1 & D_1^{1/2} \\ D_1 & I & D_1^{1/2}Q \\ \hline D_1^{1/2} & QD_1^{1/2} & Q \end{array}\right) \succeq 0; \;\; F^x \subseteq F^{y_1} \vee F^{y_2} \;\Rightarrow \\
& x = E[x | F^{y_1} \vee F^{y_2}] = P_1 y_1 + P_2 y_2 \;\Rightarrow\; -P_1 y_1 - P_2 y_2 + x = 0 \;\Rightarrow \\
& \quad \text{rank}(Q_{(y_1,y_2,x)}) = 2n_x,
\end{aligned}
$$

where use was made of the fact that $\text{rank}(Q_{y_1,y_2}) = 2n_x$. From Lemma 7.2.20 then follows that $Q_x \in Q_{varrank}$ is a solution of the algebraic Riccati equation of SGSRP hence $Q_x \in$ SGSRP.

(3) Injectiveness. Consider $Q_1,\ Q_2 \in$ SGSRP such that, $r_1(Q_1) = (n_x, x) = (n_{\bar{x}}, \bar{x}) = r_1(Q_2)$. It will be proven that $Q_1 = Q_2$. Choose for x a basis such that the joint measure $G(0, Q_{(y_1,y_2,x)})$ has the variance matrix,

$$Q_{(y_1,y_2,x)} = \begin{pmatrix} I & D_1 & D_1^{1/2} \\ D_1 & I & D_1^{1/2}Q_x \\ D_1^{1/2} & QD_1^{1/2} & Q_x \end{pmatrix}.$$

Because by the realization map $x = \bar{x}$, $(y_1,y_2,x) \in G$ and $(y_1,y_2,\bar{x}) \in G$ and $F^x \subseteq F^{y_1} \vee F^{y_2}$ and $F^{\bar{x}} \subseteq F^{y_1} \vee F^{y_2}$, the random variables x and $\bar{x}$ are jointly Gaussian and hence there exists a nonsingular transformation matrix $L \in \mathbb{R}^{n_x \times n_x}$ such that $\bar{x} = Lx$.

One denotes the following transformation matrices and calculates,

$$\begin{aligned}
& P_{11}y_1 + P_{12}y_2 = x = \bar{x} = L(P_{21}y_1 + P_{22}y_2), \\
\Leftrightarrow\ & (P_{11} - LP_{21})y_1 = (LP_{22} - P_{12})y_2, \\
\Rightarrow\ & (P_{11} - LP_{21}) = (LP_{22} - P_{12})D_1, \\
& \text{by right multiplication of } y_1^T \text{ and taking expectation,} \\
& (P_{11} - LP_{21})D_1 = (LP_{22} - P_{12}) \\
& \text{by right multiplication of } y_2^T \text{ and expectation;}
\end{aligned}$$

$$\begin{aligned}
P_{11} &= (D_1^{-1} - Q_1)D_1^{1/2}D_2^{-1} = (I - Q_1D_1)D_1^{-1/2}D_2^{-1}, \\
& P_{11} - LP_{21} = [(I - Q_1D_1) - L(Q - D_1)]D_1^{-1/2}D_2^{-1} \\
&= [(LQ_2 - Q_1) + (I - L)]D_1^{-1/2}D_2^{-1} \\
& \text{because } D_1(I - D_1^2)^{-1} = (D_1^{-1} - D_1)^{-1} = D_2^{-1}, \\
& LP_{22} - P_{12} = [L(Q_2 - D_1) - (Q_1 - D_1)]D_1^{-1/2}D_2^{-1} \\
&= [(LQ_2 - Q_1) + (I - L)D_1]D_1^{-1/2}D_2^{-1} \\
0 &= (P_{11} - LP_{21}) - (LP_{22} - P_{12})D_1 \\
&= [(LQ_2 - D_1)D_1 + (I - L) - (LQ_2 - Q_1)D_1 - (I - L)D_1^2]D_1^{-1/2}D_2^{-1} \\
&= (I - L)D_1^{-1/2}D_2^{-1} \ \Rightarrow\ L = I;
\end{aligned}$$

$$\begin{aligned}
0 &= (P_{11} - LP_{21})D_1 - (LP_{22} - P_{12}) \\
&= [(LQ_2 - D_1)D_1^2 + (I - L)D_1 - (LQ_2 - Q_1) - (I - L)D_1]D_1^{-1/2}D_2^{-1} \\
&= -(LQ_2 - Q_1)(I - D_1^2)D_1^{-1/2}D_2^{-1} \ \Rightarrow\ Q_1 = Q_2.
\end{aligned}$$

□

Theorem 7.2.21. *The* general case of SGSR of a tuple of Gaussian random variables. *Consider Problem 7.2.16. Assume that the tuple of random variables* (y_1, y_2) *is in canonical variable form.*

Recall from Theorem 7.2.19 the parametrization set,

$$\text{SGSRP} = \left\{ Q \in \mathbb{R}^{n_x \times n_x}_{spds} | Q = QD_2^{-1}Q - D_2^{-1} + QD_2^{-1}D_1 + D_1D_2^{-1}Q \right\}.$$

Define the realization map according to the formulas,

$$\begin{aligned} r(Q) &= (n_{x_1}, n_{x_2}, x),\ r : \text{SGSRP} \to \text{SGSR}_{min}, \\ n_{x_1} &= n_{y_{11}} = n_{y_{21}},\ n_{x_2} = n_{y_{12}} = n_{y_{22}}, n_x = n_{x_1} + n_{x_2}, \\ P_1 &= (D_1^{-1} - Q)D_1^{1/2}D_2^{-1} \in \mathbb{R}^{n_{x_1} \times n_{y_{12}}},\ P_2 = (Q - D_1)D_1^{-1/2}D_2^{-1} \in \mathbb{R}^{n_{x_2} \times n_{y_{22}}}, \\ x &= \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} y_{11} \\ P_1y_{12} + P_2y_{22} \end{pmatrix},\ x : \Omega \to \mathbb{R}^{n_x}. \end{aligned}$$

Then r is a well defined function and a bijection.

Proof. (1) It will be proven that the function r is well defined. Because $Q \in$ SGSRP, it follows from Lemma 7.2.20 that $Q \succ 0$. Note that $x_2 = P_1y_{12} + P_2y_{22}$. It follows from Theorem 7.2.19 that $(F^{y_{12}}, F^{y_{22}}|F^{x_2}) \in \text{CIG}_{\text{min}}$. From $x_1 = y_{11} = y_{21}$ it follows that $(F^{y_{11}}, F^{y_{21}}|\ F^{x_1}) \in \text{CIG}_{\text{min}}$. It then follows from Theorem 7.2.8 that $(F^{y_1}, F^{y_2}|F^{x}) \in \text{CIG}_{\text{min}}$. Hence $r(Q) = (n_{x_1}, n_{x_2}, Q) \in SGSR_{min}$.
(2) Surjectiveness. The tuple (y_1, y_2) is assumed to be already in the canonical variable form. Choose a basis for the random variable x such that $x = (x_1,\ x_2)$ with $F^{x_1} \subseteq F^{y_{11}} \vee F^{y_{21}}$, and $F^{x_2} \subseteq F^{y_{12}} \vee F^{y_{22}}$. Because x is of minimal dimension, $n_{x_1} = n_{y_{11}} = n_{y_{21}}$ and $n_{x_2} = n_{y_{12}} = n_{y_{22}}$. Then $(F^{y_1}, F^{y_2}|F^{x}) \in \text{CIG}_{\text{min}}$ and Proposition 7.2.8 imply that $(F^{y_{11}}, F^{y_{21}}|\ F^{x_1}) \in \text{CIG}_{\text{min}}$, and $(F^{y_{12}}, F^{y_{22}}|F^{x_2}) \in \text{CIG}_{\text{min}}$. From Theorem 7.2.19 then follows that $Q \in$ SGSRP, and $r_1(Q) = (n_{x_2}, x_2)$. Thus $r(Q) = (n_{x_1}, n_{x_2}, Q)$.
(3) Injectiveness. Consider $Q,\ \overline{Q} \in$ SGSRP such that
$r(Q) = (n_{x_1}, n_{x_2}, x) = (n_{x_1}, n_{x_2}, \overline{x}) = r(\overline{Q})$. By the realization map $x = (x_1, x_2) = \overline{x} = (\overline{x}_1, \overline{x}_2)$, hence $x_2 = \overline{x}_2$. It follows from $x_2 = r_1(Q) = r_1(\overline{Q}) = \overline{x}_2$ and the injectiveness of Theorem 7.2.19 that $Q = \overline{Q}$. □

7.3 Stochastic Realization of a Tuple of Sigma-Algebras

In the above title the expression *sigma-algebra* is used because the latex program signals an error if in a title of a section a mathematical symbol is used.

In this section the reader finds how to construct for a tuple of σ-algebras, a state σ-algebra which makes the considered triple of σ-algebras minimally conditionally independent. The construction and the properties are the basis of strong stochastic realization theory. In this section stochastic realization theory is formulated in terms of σ-algebras. In the next section it is extended to a family of σ-algebras indexed by time.

The reader is alerted of the fact that the theorems of this section are different from those of the Hilbert space framework described in [41].

7.3.1 Problem of Stochastic Realization

For a complete probability space (Ω,F,P) define the set of all sub-σ-algebras,

$$\mathbf{F} = \left\{ F_1 \subseteq F |\ F_1 \text{ a } \sigma\text{-algebra containing all null set of } F \right\}. \tag{7.1}$$

The above definition makes the concept dependent on the probability measure used. In this section all σ-algebras are members of the set $\mathbf{F}$. For a sub-σ-algebra $F_1 \in \mathbf{F}$ define the set of all positive random variables $x_1 : \Omega \to \mathbb{R}_+$ which are measurable with respect to F_1 and denote it by,

$$L_+(\Omega,\mathbb{R}_+;F_1) = L_+(F_1) = \{x_1 : \Omega \to \mathbb{R}_+|\ x_1 \text{ is } F_1 \text{ measurable}\}.$$

Below the definition and the notation of conditional independence of a triple of σ-algebras is repeated for the convenience of the reader and it is extended with the concept of minimality. See Def. 2.9.1 and Section 2.9.

Definition 7.3.1. Consider a complete probability space (Ω,F,P) and a tuple of sub-σ-algebras $(F_1,F_2) \in \mathbf{F}$. Call (F_1,F_2) *conditionally independent* conditioned on the sub-σ-algebra $G \in \mathbf{F}$, or *given* G, if the following factorization property holds,

$$E[x_1x_2|G] = E[x_1|G]E[x_2|G], \quad \forall\, x_1 \in L_+(F_1),\ x_2 \in L_+(F_2). \tag{7.2}$$

Notation to be used in this case is $(F_1,F_2|G) \in \text{CI}$.

Call G *minimally conditionally independent* for (F_1,F_2) if the following conditions both hold: (1) conditional independence, $(F_1,F_2|G) \in \text{CI}$; and (2) minimality: if $G_1 \in \mathbf{F}$, if $G_1 \subseteq G$, and if $(F_1,F_2|G_1) \in \text{CI}$ then $G_1 = G$ *a.s.* The notation to be used in this case is $(F_1,F_2|\ G) \in \text{CI}_{\text{min}}$.

In the above definition the concept of minimality is used which is related to the definition of minimality of a partially-ordered set, Def. 17.1.9.

Problem 7.3.2. *Strong stochastic realization for a tuple of σ-algebras.* Consider a complete probability space (Ω,F,P) and two σ-algebras: the *future σ-algebra* denoted by $F^+ \subseteq F$ and the *past σ-algebra* denoted by $F^- \subseteq F$.

If $(F^+,F^-|\ G) \in \text{CI}$ with $G \subseteq F^+ \vee F^-$ then call G a *state σ-algebra* for the tuple $F^+,\ F^-$. If in addition $(F^+,F^-|\ G) \in \text{CI}_{\text{min}}$ then call G a *minimal state σ-algebra* for the tuple.

The problem is: (1) to show existence of a σ-algebra $G \subseteq F^+ \vee F^-$ which make the future and the past σ-algebras conditionally independent; (2) to characterize those state σ-algebras which make the future and the past σ-algebra minimally conditional independent; and (3) to relate tuples of minimally conditionally independent state σ-algebras, and to classify all such σ-algebras.

The following notation is useful.

$$\mathbf{G}(F^+,F^-) = \left\{G \in \mathbf{F}|\ G \subseteq F^+ \vee F^-,\ (F^+,F^-|G) \in \text{CI}\right\},$$
$$\mathbf{G_{min}}(F^+,F^-) = \left\{G \in \mathbf{F}|\ G \subseteq F^+ \vee F^-,\ (F^+,F^-|G) \in \text{CI}_{\text{min}}\right\}.$$

Below attention is mainly focused on those state σ-algebras satisfying the condition that $G \subseteq (F^+ \vee F^-)$. There exists a corresponding theory for the case in which the state σ-algebra may be contained in a larger σ-algebra, $G \subseteq (F^+ \vee F^- \vee F_v)$ where the σ-algebra F_v can be specified by the user.

7.3.2 Concepts

The definition of the projection operator of σ-algebras is recalled for the benefit of the reader, see Def. 19.6.1. Consider a complete probability space (Ω, F, P) and two σ-algebras defined on it, $F_1,\ F_2 \in \mathbf{F}$. Define the projection of the σ-algebra F_1 on F_2 as the σ-algebra,

$$\sigma(F_1|\ F_2) = \sigma\left(\{E[x_1|F_2] : \Omega \to \mathbb{R}_+ \mid \forall\, x_1 : \Omega \to \mathbb{R}_+,\ x_1 \in L_+(F_1)\}\right).$$

Thus the σ-algebra $\sigma(F_1|\ F_2)$ is generated by the collection of random variables displayed above. It is assumed that all null sets of F, see Section 19.2, are included in the projection σ-algebra. In words, the projection is the smallest σ-algebra with respect to which the indicated conditional expectation is measureable of all positive random variables measurable with respect to the first σ-algebra and conditioned on the second σ-algebra. Below the qualification *a.s.* will not always be written when relating two σ-algebras even when it is useful to do so.

Definition 7.3.3. Consider Problem 7.3.2. Define the *interaction* σ-algebras of $F^+,\ F^-$ by the formulas $\sigma(F^-|\ F^+) \subseteq F^+$ and $\sigma(F^+|\ F^-) \subseteq F^-$.

Definition 7.3.4. *Stochastic controllability, co-controllability, observability, and co-observability*. Consider a triple of σ-algebras $F^+,\ F^-, G \subseteq F$ as in Problem 7.3.2, thus satisfying $(F^+, F^-|\ G) \in \mathrm{CI}$.

Call this triple *stochastically controllable* if $\sigma(F^{-+}|\ F^{+-}) = \sigma(G|\ F^{+-})$ and call it *stochastically co-controllable* if $\sigma(F^{+-}|\ F^{-+}) = \sigma(G|\ F^{-+})$.

Call this triple *stochastically observable* if $\sigma(F^+|G) = G$ and call it *stochastically co-observable* if $\sigma(F^-|G) = G$.

The above definitions correspond to observability and to co-observability of a linear system, see Section 21.3, and to stochastic observability and stochastic co-observability of a Gaussian system, see Section 4.6. The corresponding definition for Hilbert spaces was defined by G. Ruckebusch, [50, 51, 52].

Definition 7.3.5. Consider Problem 7.3.2. Define the *frame σ-algebra* by the expression,

$$F_f = F^{-+} \vee F^{+-} = \sigma(F^-|F^+) \vee \sigma(F^+|F^-).$$

7.3.3 Existence of Minimal State Sigma-Algebras

The first subproblem of the stochastic realization Problem 7.3.2 is to show existence of a minimal state σ-algebra.

Proposition 7.3.6. Existence of minimal state σ-algebras. *Consider a probability space (Ω, F, P) and two σ-algebras defined on it, $F^+,\ F^- \subseteq F$. Recall the notation, $F^{-+} = \sigma(F^-|\ F^+)$ and $F^{+-} = \sigma(F^+|\ F^-)$.*

(a)Then $(F^+, F^- | F^{+-}) \in \mathrm{CI}_{\min}$ *and* $(F^+, F^- | F^{-+}) \in \mathrm{CI}_{\min}$.
(b)If $(F^+, F^- | G) \in \mathrm{CI}$ *and* $G \subseteq F^-$ *then* $F^{+-} \subseteq G$.
If $(F^+, F^- | G) \in \mathrm{CI}$ *and* $G \subseteq F^+$ *then* $F^{-+} \subseteq G$.
(c)$(F^+, F^- | G) \in \mathrm{CI}_{\min}$ *and* $G \subseteq F^{+-}$ *if and only if* $F^{+-} = G$.
$(F^+, F^- | G) \in \mathrm{CI}_{\min}$ *and* $G \subseteq F^{-+}$ *if and only if* $F^{-+} = G$.

Proof. (a) It follows from Proposition 19.8.4.(d) that $(F^+, F^- | \sigma(F^+ | F^-)) \in$ CI. Consider a σ-algebra $G \subseteq \sigma(F^+|F^-) \subseteq F^-$ such that $(F^+, F^- | \; G) \in$ CI. It follows from Proposition 19.8.4.(f) that then $\sigma(F^+|F^-) \subseteq \sigma(G|F^-) = G$ hence $\sigma(F^+|F^-) = G$ and $(F^+, F^- | \; \sigma(F^+|F^-)) \in \mathrm{CI}_{\min}$. The version with $\sigma(F^-|F^+)$ is proven in a symmetric way.
(b) The conditional independence and $G \subseteq F^-$ imply by Proposition 19.8.2.(c) that $\sigma(F^+|F^-) \subseteq G$. The second statement follows by symmetry.
(c) ($\Leftarrow$) It follows from Proposition 19.8.4.(d) that $(F^+, F^- | G) \in$ CI where $G = \sigma(F^+| \; F^-)$. From (a) follows that $(F^+, F^- | \; G) \in \mathrm{CI}_{\min}$.
($\Rightarrow$) From (b) follows that $\sigma(F^+| \; F^-) \subseteq G$. Combined with the assumption this yields $\sigma(F^{-+}| \; F^{+-}) = G$. The second statement follows by symmetry. □

Below an example is provided of several σ-algebras defined on a finite set. The reader is refered to Example 19.8.8 for the procedure to calculate the conditional expectations used and the projection of two σ-algebras.

Example 7.3.7. *Existence of two minimal state σ-algebras with different numbers of atoms.* Consider the finite probability space and the sub-σ-algebras,

$$\begin{aligned} & (\Omega, F, P), \; \Omega = \mathbb{Z}_8, \; P \text{ a uniform measure;} \\ F^+ &= \sigma(\{1\},\{2,3\},\{4,5,6\},\{7,8\}), \; F^- = \sigma(\{1,2,3\},\{4,5,7\},\{6,8\}), \\ G_1 &= \sigma(\{1,2,4\},\{3\},\{5,6,7,8\}), \; G_2 = \sigma(\{1\},\{2,3,4,5\},\{6\},\{7,8\}). \end{aligned}$$

For a diagram of the probability space and the σ-algebras see Fig. 7.5. Then,

(a)$(F^+, F^- | G_1) \in \mathrm{CI}_{\min}$;
(b)$(F^+, F^- | G_2) \in \mathrm{CI}_{\min}$;
(c)State σ-algebra G_1 has three atoms and G_2 has four atoms.

This example shows that there is a conflict between minimality as in $\mathrm{CI}_{\min}$ based on the inclusion relation of state σ-algebras and the number of atoms of the state σ-algebra. The elementary calculations are omitted due to space constraints.

7.3.4 State Sigma-Algebras in the Frame Sigma-Algebra

Needed is a characterization when the state σ-algebra is a sub-σ-algebra of the frame σ-algebra; in terms of notation, when is $G \subseteq F_f$? Recall from Section 7.1 that the frame σ-algebra, $F_f = \sigma(F^-| \; F^+) \vee \sigma(F^+| \; F^-)$, describes the interaction spaces of the future F^+ and the past F^- σ-algebras.

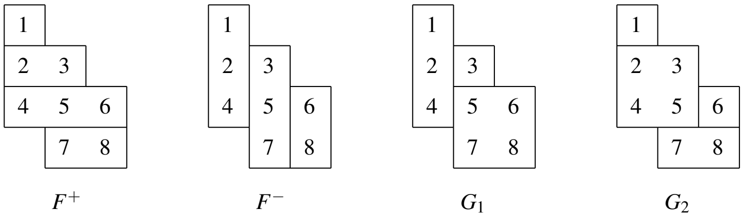


Fig. 7.5 Diagram of Example 7.3.7.

Theorem 7.3.8. Necessary and sufficient condition for a state σ-algebra to belong to the frame σ-algebra. *Consider the setting of Problem 7.3.2 for F^+, F^-, hence with $G \subseteq F^+ \vee F^-$ and $(F^+, F^-|G) \in \text{CI}$.*

Then $G \subseteq F_f = F^{-+} \vee F^{+-}$
if and only if $\sigma(G|F^+) = F^{-+}$ and $\sigma(G|F^-) = F^{+-}$
if and only if the triple is stochastically controllable and stochastically co-controllable.

Proof. Note the equivalences,

$$\begin{aligned}
& \sigma(G|F^+ \vee F^-) = G \subseteq \ F^{-+} \vee F^{+-} \ \subseteq \ F^+ \vee F^{+-}, \text{ by } G \subseteq F^{-+} \vee F^{+-}, \\
\Leftrightarrow\ & (F^+ \vee F^-, G|\ F^+ \vee F^{+-}) \in \text{CI}, \\
& (\Rightarrow \text{ by Proposition 2.9.3.(a)}, \ \Leftarrow \text{ by Proposition 2.9.4.(a),}) \\
\Leftrightarrow\ & (F^-, G|\ F^+ \vee F^{+-}) \in \text{CI}, \\
& (\Rightarrow \text{ by restriction}, \ \Leftarrow \text{ Proposition 19.8.2.(a),}) \\
\Leftrightarrow\ & (F^-, F^+ \vee G|\ F^{+-}) \in \text{CI}, \text{ by Proposition 19.8.2.(a)}, \\
& \text{using that } (F^+, F^-|\ F^{+-}) \in \text{CI}, \\
\Leftrightarrow\ & (F^-, G|\ F^{+-}) \in \text{CI}, \text{ and } (F^-, F^+|F^{+-} \vee G) \in \text{CI}, \\
& \text{by Proposition 19.8.2.(a)}, \\
\Leftrightarrow\ & \sigma(G|F^-) \subseteq F^{+-}, \text{ by Proposition 19.8.2.(c), and using } F^{+-} \subseteq F^-, \\
\Leftrightarrow\ & \sigma(G|F^-) = F^{+-}, \text{ because } (F^+, F^-|\ F^{+-}) \in \text{CI}_{\text{min}} \text{ by Prop. 7.3.6.}
\end{aligned}$$

Similary one shows that

$$\sigma(G|F^+ \vee F^{+-}) = G \subseteq F^{-+} \vee F^{+-} \ \Leftrightarrow \ \sigma(G|F^+) = F^{-+}.$$

□

Example 7.3.9. *Existence of a minimal state σ-algebra which is not included in the frame σ-algebra.*

Consider the finite probability space and the sub-σ-algebras,

$$\begin{aligned}
& (\Omega, F, P), \ \Omega = \mathbb{Z}_{10}, \ P \text{ a uniform probability measure;} \\
F^+ &= \sigma(\{1,4\}, \{2,5,8\}, \{3,6,9\}, \{7,10\}), \\
F^- &= \sigma(\{1,2,3\}, \{4,5,6,7\}, \{8,9,10\}), \\
G &= \sigma(\{1,2,4,5\}, \{3\}, \{6,7,9,10\}, \{8\}).
\end{aligned}$$

In Fig. 7.6 the atoms of these σ-algebras are displayed. Then,

(a) $(F^+, F^- | G) \in \mathrm{CI}_{\min}$;
(b) $\sigma(G|F^-) = F^{+-}$,
$\sigma(G|F^+) = F^+ \supsetneq F^{-+} = \sigma(\{1,4\},\{2,5,8,3,6,9\},\{7,10\})$, and
$G \not\subseteq F_f = F^{-+} \vee F^{+-}$.

See for the calculations, Example 7.3.16. The minimality of G is proven by showing that any strict sub-σ-algebra G_1 of G does not satisfy the condition of conditional independence. Any strict sub-σ-algebra of G_1 is obtained by either splitting the atom $\{1,2,4,5\}$ or by splitting the atom $\{6,7,9,10\}$. Note that the atoms $\{3\}$ and $\{8\}$ cannot be split. The σ-algebra G_1 has to satisfy the conditional independence relation hence has to have atoms which are rectangles and the atoms have to cover the set Ω.

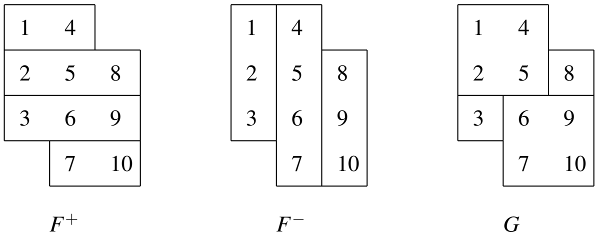


Fig. 7.6 Diagram of several σ-algebras of Example 7.3.9.

Below attention is restricted to those state σ-algebras G within the frame σ-algebra F_f, hence satisfying $G \subseteq F_f$.

Proposition 7.3.10. State σ-algebras within the frame σ-algebra. *Consider the setting of Problem 7.3.2. $F^+,\ F^- \in \mathbf{F}$. Assume that $G \subseteq F_f$.*

(a)$(F^+, F^- | G) \in \mathrm{CI}$ if and only if $(F^{-+}, F^{+-} | G) \in \mathrm{CI}$.
(b)$(F^+, F^- |\ G) \in \mathrm{CI}_{\min}$ if and only if $(F^{-+}, F^{+-} | G) \in \mathrm{CI}_{\min}$.

Because of the above results, attention below is restricted to the conditional independence relation $(F^{-+}, F^{+-} | G) \in \mathrm{CI}_{\min}$.

Proof. (a) ($\Rightarrow$) This follows by the restrictions of the σ-algebras $F^+ \supseteq F^{-+}$ and $F^- \supseteq F^{+-}$.
($\Leftarrow$) Note that,

$$\begin{aligned}
&\text{because } G \subseteq F^{-+} \vee F^{+-} \;\Rightarrow\; F^- \vee G \subseteq F^- \vee F^{-+},\\
\Rightarrow\; &\sigma(F^{-+}|F^- \vee G) \subseteq F^{+-} \vee G,\\
\Rightarrow\; &(F^{-+},F^-|G) \in \text{CI}, \text{ by Proposition 19.8.1 and}(F^{-+},F^{+-}\mid G) \in \text{CI};\\
&\text{because } G \subseteq F^{-+} \vee F^- \;\Rightarrow\; F^+ \vee G \subseteq F^+ \vee F^{+-},\\
&\sigma(F^+|F^- \vee G) \subseteq F^- \vee G,\\
\Rightarrow\; &(F^+,F^-|G) \in \text{CI}, \text{ by Proposition 19.8.1, and by}(F^{-+},F^-\mid G) \in \text{CI}.
\end{aligned}$$

(b) ($\Rightarrow$) From (a) follows that $(F^{-+},F^{+-}\mid G) \in \text{CI}$. Consider a σ-algebra $G_1 \subseteq G$ such that $(F^{-+},F^{+-}\mid G_1) \in \text{CI}$. It is to be shown that $G_1 = G$. Because $G_1 \subseteq G \subseteq F^{-+} \vee F^{+-}$ it follows from (a) that $(F^+,F^-\mid G_1) \in \text{CI}$. This, $G_1 \subseteq G$, and the minimality of $(F^+,F^-\mid G) \in \text{CI}_{\text{min}}$ imply that $G_1 = G$. Hence $(F^{-+},F^{+-}|G) \in \text{CI}_{\text{min}}$.
($\Leftarrow$) From (a) follows that $(F^+,F^-\mid G) \in \text{CI}$. Consider a σ-algebra $G_1 \subseteq G$ such that $(F^+,F^-\mid G_1) \in \text{CI}$. It is to be shown that $G_1 = G$. Because $G_1 \subseteq G \subseteq F^{-+} \vee F^{+-}$ it follows from (a) that $(F^{-+},F^{+-}\mid G_1) \in \text{CI}$. This, $G_1 \subseteq G$, and the minimality of $(F^{-+},F^{+-}\mid G) \in \text{CI}_{\text{min}}$ imply that $G_1 = G$. Hence $(F^+,F^-|G) \in \text{CI}_{\text{min}}$. □

Example 7.3.11. The structure of the set of state σ-algebras can be illustrated by the case in which the future and the past σ-algebras are generated by finite-dimensional Gaussian random variables. This case is solved in Section 7.2.

Consider a tuple of Gaussian random variables $(y^+,y^-) \in G(0,\; Q_{cvf})$ in the canonical variable form, see Def. 7.2.2, hence having the decomposition,

$$\begin{aligned}
&y^+ = (y_{11},y_{12},y_{13}),\; y^- = (y_{21},y_{22},y_{23}),\; y_{11} = y_{21},\; (y_{12},y_{22}) \in G(0,D_1),\\
&\qquad \{y_{11} = y_{21},(y_{12},y_{22}),y_{13},\; y_{23}\} \text{ are independent.}
\end{aligned}$$

Then,

$$\begin{aligned}
&E[y^-|F^{y^+}] = \begin{pmatrix} y_{11} \\ D_1 y_{12} \\ 0 \end{pmatrix}, E[y^+|F^{y^-}] = \begin{pmatrix} y_{21} \\ D_1 y_{22} \\ 0 \end{pmatrix},\\
&\text{rank}(D_1) = n_{y_{12}} = n_{y_{22}},\\
&F^{y^-|y^+} = F^{y_{11},y_{12}},\; F^{y^+|y^-} = F^{y_{21},y_{22}},\; F^{y^+} \cap F^{y^-} = F^{y_{11}} = F^{y_{21}} \subsetneq F^{y^-|y^+};\\
&\text{if } G \subseteq (F^{y^-|y^+} \vee F^{y^+|y^-}) = F^{y_{11},y_{12},y_{21},y_{22}},\text{ then } (F^{y^+},F^{y^-}\mid G) \in \text{CI}\\
\Leftrightarrow\; &(F^{y^-|y^+},F^{y^+|y^-}\mid G) = (F^{y_{11},y_{12}},F^{y_{21},y_{22}}|G) \in \text{CI}, \text{ by Proposition 7.3.10.}
\end{aligned}$$

7.3.5 Introduction to Characterization

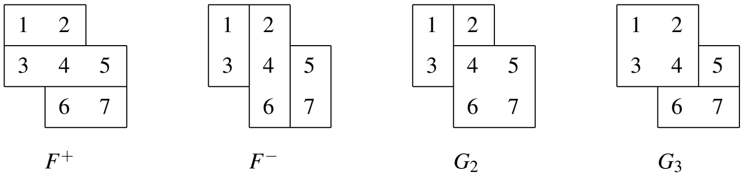
The characterization of minimal state σ-algebras is in terms of related σ-algebras. Thus for any σ algebra G such that $(F^+,\; F^-\mid G) \in \text{CI}_{\text{min}}$, the classification uses the σ-algebras $F^+ \vee G$ and $F^- \vee G$. Needed is therefore an understanding of the set theoretic structure of the sets of such σ-algebras.

The concepts formulated for the characterization of minimal state σ-algebras are inspired by the relation of conditional independence in Hilbert spaces formulated by G. Ruckebusch, [50, 51, 52], and by A. Lindquist and G. Picci, [40, 41]. The report [46] has a related framework with different conditions, see also the book [26].

There are essential differences between conditional independence in Hilbert spaces and conditional independence of σ-algebras. These differences result in different characterizations of minimal elements. For example, a theorem in the Hilbert space case has no equivalence for σ-algebras, see [41, Thm. 7.2.6]. There seems also a difference related to a partially ordered set of spaces. To be more explicit, if in the context of Hilbert space the two triples $(H^+, H^- \mid G_a)$ and $(H^+, H^- \mid G_b)$ are both minimal triples then it seems that (a) either $H^+ + G_a \subseteq H^+ + G_b$ or $H^+ + G_a \supseteq H^+ + G_b$; and (a) either $H^- + G_a \subseteq H^- + G_b$ or $H^- + G_a \supseteq H^- + G_b$. Thus it seems that on the sets of these Hilbert spaces, there is an order relation that is a total order relation.

For σ-algebras, the set consisting of $F^+ \vee G$ for minimal state σ-algebras G, is in general only partially ordered and not totally ordered. This conclusion leads to a realization theory of σ-algebras which is quite different than that for Hilbert spaces. This issue is explained with the following example and a definition.

Example 7.3.12. *Description of all minimal state σ-algebras.*



F^+ F^- G_2 G_3

Fig. 7.7 Diagram of several σ-algebras of Example 7.3.12.

Consider a finite probability space and several σ-algebras according to,

$$\begin{aligned}
&(\Omega, F, P),\ \Omega = \mathbb{Z}_7,\ P \text{ a uniform probability measure},\\
F^+ &= \sigma(\{1,2\},\{3,4,5\},\{6,7\}),\ \ F^- = \sigma(\{1,3\},\{2,4,6\},\{5,7\}),\\
G_1 &= \sigma(\{1,2\},\{3\},\{4,5,6,7\}),\ G_2 = \sigma(\{1,3\},\{2\},\{4,5,6,7\}),\\
G_3 &= \sigma(\{1,2,3,4\},\{5\},\{6,7\}),\ \ G_4 = \sigma(\{1,2,3,4\},\{5,7\},\{6\}),\\
G_5 &= \sigma(\{1\},\{2,6\},\{3,4,5\},\{7\}),\ \ G_6 = \sigma(\{1,3\},\{3,6\},\{4,5\},\{7\}),\\
G_7 &= \sigma(\{1\},\{2,6\},\{3,4\},\{5,7\}),\ \ G_8 = \sigma(\{1\},\{2,4,6\},\{3,5\},\{7\}),\\
G_9 &= \sigma(\{1\},\{2,4\},\{3,5\},\{6,7\}),\ \ G_{10} = \sigma(\{1,2\},\{3,5\},\{4,6\},\{7\}).
\end{aligned}$$

Several of these σ-algebras are displayed in Fig. 7.7.

It is straightforward to prove that,

$$\sigma(F^-|F^+) = F^+ \text{ and } \sigma(F^+|F^-) = F^- \ \Rightarrow\ F_f = F^+ \vee F^-.$$

According to Theorem 7.3.17, all minimal state σ-algebras G in the frame σ-algebra, thus such that $(F^+, F^- | G) \in \mathrm{CI}_{\mathrm{min}}$ and $G \subseteq F_f$, are: F^+, F^-, and G_1 through G_{10}. The calculations are a laborious exercise.

Definition 7.3.13. Consider a tuple of σ-algebras $F^+,\ F^- \subseteq F$ as specified in Problem 7.3.2. Consider the subset of all minimal triples $(F^+,\ F^- |\ G) \in \mathrm{CI}_{\mathrm{min}}$ with $G \subseteq F_f$. Define the sets,

$$\begin{aligned} FG^+_{min}(F^{-+}, F_f) &= \left\{ F^{-+} \vee G \subseteq F_f |\ (F^+,\ F^- |\ G) \in \mathrm{CI}_{\mathrm{min}} \right\}, \\ FG^-_{min}(F^{+-}, F_f) &= \left\{ F^{+-} \vee G \subseteq F_f |\ (F^+,\ F^- |\ G) \in \mathrm{CI}_{\mathrm{min}} \right\}. \end{aligned}$$

Define on each of these sets respectively a partial order relation by inclusion according to

$$\begin{aligned} F^{-+} \vee G_a \ &\subseteq\ F^{-+} \vee G_b,\ \ F^{-+} \vee G_a,\ F^{-+} \vee G_b \in FG^+_{min}(F^{-+}, F_f); \\ F^{+-} \vee G_c \ &\subseteq\ F^{+-} \vee G_d,\ \ F^{+-} \vee G_c,\ F^{+-} \vee G_d \in FG^-_{min}(F^{+-}, F_f). \end{aligned}$$

It is elementary to prove that each of the above relations is a partial order relation.

It will now be shown that the two partial order relations are in general not totally-ordered subsets. This is in distinction with the Hilbert space case, which is described in [41, Chapter 7].

Example 7.3.14. Consider Example 7.3.12 of a finite probability space with sub-σ-algebras $F^+,\ F^- \subseteq F$. Note that for this example $F^{-+} = F^+$ and $F^{+-} = F^-$. The quoted example specifies all σ-algebras $G \subseteq F_f$ such that $(F^+,\ F^- |\ G) \in \mathrm{CI}_{\mathrm{min}}$.

For future reference, the elements of the subset $FG^+_{min}(F^+, F_f)$ are all listed for this example.

$$\begin{aligned} F^+ &= \sigma(\{1,2\},\ \{3,4,5\},\ \{6,7\}), \\ F^+ \vee G_1 &= \sigma(\{1,2\},\ \{3\},\ \{4,5\},\ \{6,7\}), \\ F^+ \vee G_2 &= \sigma(\{1\},\ \{2\},\ \{3\},\ \{4,5\},\ \{6,7\}), \\ F^+ \vee G_3 &= \sigma(\{1,2\},\ \{3,4\},\ \{5\},\ \{6,7\}), \\ F^+ \vee G_4 &= \sigma(\{1,2\},\ \{3,4\},\ \{5\},\ \{6\},\{7\}), \\ F^+ \vee G_5 &= \sigma(\{1\},\ \{2\},\ \{3,4,5\},\ \{6\},\{7\}), \\ F^+ \vee G_6 &= \sigma(\{1\},\ \{2\},\ \{3\},\ \{4,5\},\ \{6\},\{7\}), \\ F^+ \vee G_7 &= \sigma(\{1\},\ \{2\},\ \{3,4\},\ \{5\},\ \{6\},\{7\}), \\ F^+ \vee G_8 &= \sigma(\{1\},\ \{2\},\ \{3,5\},\ \{4\},\ \{6\},\{7\}), \\ F^+ \vee G_9 &= \sigma(\{1\},\ \{2\},\ \{3,5\},\ \{4\},\ \{6,7\}), \\ F^+ \vee G_{10} &= \sigma(\{1,2\},\ \{3,5\},\ \{4\},\ \{6\},\{7\}). \end{aligned}$$

The set $FG^+_{min}(F^+, F_f)$ is a partially ordered subset which is *not* totally ordered. In fact, the set $FG^+_{min}(F^+, F_f)$ has several branches of elements from the root F^+ which are totally ordered while different elements of any two nonequal branches are not totally ordered according to the inclusion relations

$$
\begin{array}{ll}
F^+ \subseteq F^+ \vee G_2 \subseteq F^+ \vee G_1, & F^+ \subseteq F^+ \vee G_2 \subseteq F^+ \vee G_6, \\
F^+ \subseteq F^+ \vee G_5 \subseteq F^+ \vee G_6, & F^+ \subseteq F^+ \vee G_5 \subseteq F^+ \vee G_8, \\
F^+ \subseteq F^+ \vee G_5 \subseteq F^+ \vee G_7, & \\
F^+ \subseteq F^+ \vee G_9 \subseteq F^+ \vee G_8, & F^+ \subseteq F^+ \vee G_{10} \subseteq F^+ \vee G_8, \\
F^+ \subseteq F^+ \vee G_3 \subseteq F^+ \vee G_4 \subseteq F^+ \vee G_7. &
\end{array}
$$

For example,

$$
\begin{array}{l}
\{3,\ 4\} \not\in F^+ \vee G_2,\ \{3,\ 4\} \in F^+ \vee G_3, \\
\{4,\ 5\} \in F^+ \vee G_2,\ \{4,\ 5\} \not\in F^+ \vee G_3.
\end{array}
$$

The conclusion is that for this example, the set $FG^+_{min}(F^{-+}, F_f)$ is a partially ordered set which is not a totally ordered set. Moreover, the partially ordered set has several branches of elements where each branch has totally ordered elements. For the set FG^-_{min} analogous conclusions hold.

The consequences of the above example have an effect on the theory of the remaining parts of this section and on the next section.

7.3.6 Characterization of Minimal State Sigma-Algebras-1

The second subproblem of the stochastic realization Problem 7.3.2 is to characterize those triples of σ-algebras which are minimal stochastic realizations. The reader is asked to recall Proposition 7.3.6 on the existence of a minimal state σ-algebra.

Theorem 7.3.15. Minimality of a state σ-algebra implies stochastic observability and stochastic co-observability. *Consider Problem 7.3.2.*

(a)If $(F^+, F^- \mid G) \in \mathrm{CI}_{\mathrm{min}}$ and if $G \subseteq F^+ \vee F^-$ then

$$\sigma(F^+|G) = G = \sigma(F^-|G), \tag{7.3}$$

hence stochastic observability and stochastic co-observability both hold.

(b)If $(F^+, F^- \mid G) \in \mathrm{CI}_{\mathrm{min}}$ and if $G \subseteq F_f = F^{-+} \vee F^{+-}$ then

$$\sigma(F^{-+}|G) = G = \sigma(F^{+-}|G), \tag{7.4}$$

hence stochastic observability and stochastic co-observability both hold with respect to the interaction σ-algebras F^{-+} and F^{+-}.

Proof. (a) Note that $(F^+, F^-|G) \in \mathrm{CI}_{\mathrm{min}}$ implies by Proposition 19.8.4 that $(F^+, F^-|\sigma(F^+|G)) \in \mathrm{CI}$. The latter result, $\sigma(F^+|G) \subseteq G$, and the assumed minimality of the σ-algebra G imply that $\sigma(F^+|G) = G$. By symmetry, $\sigma(F^-|G) = G$.
(b) By Proposition 7.3.10.(b), $(F^{-+}, F^{+-}|G) \in \mathrm{CI}_{\mathrm{min}}$. By (a) the result follows. □

Example 7.3.16. *Existence of an example of a state σ-algebra which is stochastically observable and stochastically co-observable yet not minimal.*

Consider the finite probability space with a uniform measure, specified by $\Omega = \mathbb{Z}_9$. See Fig. 7.8 for a diagram of the σ-algebras. Define the following sub-σ-algebras of F by their atoms according to,

$$F^+ = \sigma(\{1,2,3\},\{4,5,6\},\{7,8,9\}),\ \ F^- = \sigma(\{1,4,7\},\{2,5,8\},\{3,6,9\}),$$
$$G_1 = \{\Omega,\emptyset\},\ \ G_2 = \sigma(\{2,3\},\{6,9\},\{7,8\},\{1,4\},\{5\}).$$

The example of G_2 is due to J.C. Willems in a personal communication with the author.

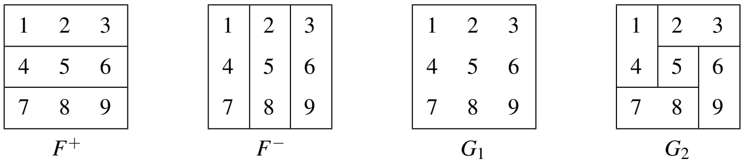


Fig. 7.8 Diagram of the finite probability space and sub-σ-algebras of Example 7.3.16. Each smallest rectangle of a σ-algebra is an atom of the corresponding σ-algebra. From the orientation of the atoms of F^+ and F^- and their overlap, it is directly clear that F^+ and F^- are independent σ-algebras.

(a) $F^+ \vee F^-$ is a finite σ-algebra generated of which the atoms are those with the numbers 1 through 9. Further, $F^{-+} = \sigma(F^-|F^+) = G_1$, $F^{+-} = \sigma(F^+|F^-) = G_1$, $F_f = F^{-+} \vee F^{+-} = G_1$.
(b) $(F^+,F^-|G_1) \in \mathrm{CI}_{\min}$.
(c) $G_2 \not\subseteq F_f = \sigma(F^-|\ F^+) \vee \sigma(F^+|\ F^-)$.
(d) $(F^+,F^-|G_2) \in \mathrm{CI}$ and $\sigma(F^+|G_2) = G_2 = \sigma(F^-|G_2)$. Hence stochastic observability and stochastic co-observability hold for $(F^+,F^-|G_2) \in \mathrm{CI}$.
(e) But $G_1 \subsetneq G_2$ hence $(F^+,F^-|G_2) \notin \mathrm{CI}_{\min}$; thus G_2 is not minimal for F^+ and F^-.

The proofs of the above statements may be found in Example 19.8.8.

There follows a first characterization of minimal state σ-algebras. The discussion of the partially ordered subset $FG^+_{min}(F^+,F_f)$ of Subsection 7.3.5 shows that the formulation of the next theorem is satisfactory. This first characterization is reformulated in Theorem 7.3.24.

Theorem 7.3.17. Minimal state σ-algebras – A first characterization.
Consider Problem 7.3.2. Consider a sub-σ-algebra $G \subseteq F_f = F^{-+} \vee F^{+-}$.

(a) If $(F^+,F^-|G) \in \mathrm{CI}_{\min}$ then
(1) $(F^{-+},F^{+-}|G) \in \mathrm{CI}$, (2) $\sigma(F^{-+}|G) = G = \sigma(F^{+-}|G)$, and
(3) for all sub-σ-algebras $G_1 \subseteq G$ such that $(F^{-+},F^{+-}|G_1) \in \mathrm{CI}$,
it is true that $G \subseteq F^{-+} \vee G_1$ and $G \subseteq F^{+-} \vee G_1$.

(b)If (1) $(F^{-+},F^{+-}|G) \in \mathrm{CI}$*, (2)* $\sigma(F^{-+}|G) = G = \sigma(F^{+-}|G)$*, and*
(3) for all sub-σ*-algebras* $G_1 \subseteq G$ *such that* $(F^{-+},F^{+-}|G_1) \in \mathrm{CI}$*,*
either $G \subseteq F^{-+} \vee G_1$ *or* $G \subseteq F^{+-} \vee G_1$*;*
then $(F^{-+},F^{+-}|G) \in \mathrm{CI}_{\min}$ *and* $(F^{+},F^{-}|G) \in \mathrm{CI}_{\min}$.

Proof. (a) Condition (1) follows directly from the assumption by restriction of the σ-algebras F^{+} to F^{-+} and of F^{-} to F^{+-}. (2) This follows from Theorem 7.3.15.(b). (3) Let G_1 be a σ-algebra such that $G_1 \subseteq G$ and $(F^{-+},F^{+-}|\, G_1) \in \mathrm{CI}$. It follows from the minimality of G, thus of $(F^{-+},F^{+-}|\, G) \in \mathrm{CI}_{\min}$, that $G_1 = G$. Hence $G \subseteq F^{-+} \vee G = F^{-+} \vee G_1$ and $G \subseteq F^{+-} \vee G = F^{+-} \vee G_1$.
(b) Consider a sub-σ-algebra $G_1 \subseteq G$ such that $(F^{-+},F^{+-}|\, G_1) \in \mathrm{CI}$. It will be proven that $G_1 = G$ from which follows that G is minimal. Note that,

$$\begin{aligned}
G &= \sigma(F^{-+}|\, G), \text{ by (2)},\\
&= \sigma(F^{-+}|\, F^{+-} \vee G), \text{ by (1)},\\
&\subseteq F^{+-} \vee G_1, \text{ by (3) and by selection},\\
&\Rightarrow F^{-+} \vee G \subseteq F^{-+} \vee G_1;\\
&\quad G_1 \subseteq G \;\Rightarrow F^{+-} \vee G_1 \subseteq F^{+-} \vee G \;\Rightarrow\; F^{+-} \vee G_1 = F^{+-} \vee G;\\
G &= \sigma(F^{-+}|\, F^{+-} \vee G), \text{ by an above relation},\\
&= \sigma(F^{-+}|\, F^{+-} \vee G_1), \text{ by the above conclusion},\\
&= \sigma(F^{-+}|\, G_1), \text{ because}(F^{-+},F^{+-}|\, G_1) \in \mathrm{CI},\\
&\subseteq G_1 \;\Rightarrow\; G_1 = G \;\Rightarrow\; (F^{-+},F^{+-}|\, G) \in \mathrm{CI}_{\min}\\
&\Rightarrow (F^{+},F^{-}|\, G) \in \mathrm{CI}_{\min}, \text{ by Proposition 7.3.10.(b)}.
\end{aligned}$$

□

Proposition 7.3.18. Minimal past and minimal future σ-algebras. *Consider Problem 7.3.2. Then* $(F^{+},F^{-}|\, F^{-+}) \in \mathrm{CI}_{\min}$*,* $(F^{-+},F^{+-}|\, F^{-+}) \in \mathrm{CI}_{\min}$*,*
$(F^{+},F^{-}|\, F^{+-}) \in \mathrm{CI}_{\min}$*, and* $(F^{-+},F^{+-}|\, F^{+-}) \in \mathrm{CI}_{\min}$.
Call $G^{-} = \sigma(F^{-+}|\, F^{+-})$ *and* $G^{+} = \sigma(F^{+-}|\, F^{-+})$ *respectively the* minimal past σ-algebra *and the* minimal future σ-algebra *of the tuple* $(F^{+},\, F^{-})$.

This result is identical to that of Proposition 7.3.6 but with a proof based on Theorem 7.3.17.

Proof. The conditions of Theorem 7.3.17.(b) are checked. (1) It follows from Proposition 2.9.3 that $(F^{-+},F^{+-}|\, F^{-+}) \in \mathrm{CI}$. (2) It follows from
Proposition 19.8.6.(h) that $\sigma(F^{-+}|\, F^{+-}) = F^{+-}$ and from Proposition 19.8.6.(b) that $\sigma(F^{+-}|\, F^{+-}) = F^{+-}$. (3) If $G_1 \subseteq F^{-+}$ is such that $(F^{-+},F^{+-}|\, G_1) \in \mathrm{CI}$ then $F^{-+} \subseteq F^{-+} \vee G_1$. From the theorem then follows that $(F^{-+},F^{+-}|\, F^{-+}) \in \mathrm{CI}_{\min}$. By symmetry of $+$ with respect to $-$, the conclusion for F^{+-} follows from the first part of the proof. □

Next the construction of a minimal state σ-algebra is addressed.

Procedure 7.3.19 Procedure for construction of a state σ-algebra which is stochastically observable and stochastically co-observable. *Consider a triple of* σ*-algebras* $(F^{+},F^{-}|\, G_0) \in \mathrm{CI}$ *with* $G_0 \subseteq F^{+} \vee F^{-}$*. Construct successively:*

1. *$G_1 = \sigma(F^+|G_0)$ and $G_2 = \sigma(F^-|\, G_1)$.*
2. *For $k = 1,2,\ldots,\infty$ do, if it is not true that $\sigma(F^+|G_{2k}) = G_{2k} = \sigma(F^-|\, G_{2k})$ then define $G_{2k+1} = \sigma(F^+|G_{2k})$ and $G_{2k+2} = \sigma(F^-|\, G_{2k+1})$.*
3. *If the procedure stops after a finite number of steps with $G_s \in \mathbf{F}$ then,*

   $$(F^+,F^-|\, G_s) \in \mathrm{CI},\ \sigma(F^+|\, G_s) = G_s = \sigma(F^-|\, G_s).$$

   Output G_s.

In general, the above procedure will not stop after two steps hence iteration is required.

Proposition 7.3.20. Construction of a minimal state σ-algebra. *Consider a triple $(F^+,F^-|\, G_0) \in \mathrm{CI}$ with $G_0 \subseteq F^+ \vee F^-$. Consider Procedure 7.3.19*

$$G_1 = \sigma(F^+|\, G),\ \ G_2 = \sigma(F^-|\, G_1).$$

(a)Then $(F^+,F^-|\, G_1) \in \mathrm{CI}$, $G_1 \subseteq F^+ \vee F^-$, and $G_1 = \sigma(F^+|G_1)$.
(b)Then $(F^+,F^-|\, G_2) \in \mathrm{CI}$, $G_2 \subseteq F^+ \vee F^-$, and $G_2 = \sigma(F^+|G_2)$.
(c)Results like (a) and (b) hold for all G_{k+1} and G_{k+2}. If the procedure converges in a finite number of steps then there exists a σ-algebra $G_s \in \mathbf{F}$ such that,

$$(F^+,F^-|\, G_s) \in \mathrm{CI},\ G_s \in \mathbf{F},\ \sigma(F^+|\, G_s) = G_s = \sigma(F^-|\, G_s).$$

(d)If the probability space has only a finite number of atoms then Procedure 7.3.19 stops in a finite number of steps.

Proof. (a) This follows from Proposition 19.8.4.(c) and the expression of G_1. $G_1 = \sigma(F^+|G) \subseteq G \subseteq F^+ \vee F^-$. $\sigma(F^+|G_1) = \sigma(F^+|\sigma(F^+|G_0)) = \sigma(F^+|G_0) = G_1$, where the second equality follows from Proposition 19.8.6.(d). (b) This follows from (a) by symmetry.
(c) The first part is obvious. As long as $\sigma(F^+|\, G) \subsetneq G$ then the first σ-algebra has a strictly smaller number of atoms than G.
(d) Because the probability space has only a finite number of atoms, the procedure stops after a finite number of steps. □

7.3.7 Characterization of Minimal State Sigma-Algebras-2

In this subsection it has to be kept in mind that the sets of σ-algebras $FG^+_{min}(F^{-+},F_f)$ and $FG^-_{min}(F^{+-},F_f)$ are partially ordered sets which are not totally ordered. See for this Subsection 7.3.5. A special case is stated first which provides the motivation for the general case.

Proposition 7.3.21. *Consider the σ-algebras F^+, $F^- \subseteq F$. The following statements are equivalent:*

(a)$(F^+,F^-|\, F^+ \cap F^-) \in \mathrm{CI}$;
(b)$F^{+-} \subseteq F^+ \cap F^-$ and $F^{-+} \subseteq F^+ \cap F^-$;

(c)$F^{+-} = F^{+} \cap F^{-} = F^{-+}$*;*
(d)$(F^{+}, F^{-} | \, F^{+} \cap F^{-}) \in \text{CI}_{\text{min}}$.

Proof. Recall that from Proposition 7.3.6.(a) follows that $(F^{+}, F^{-} | \, F^{-+}) \in \text{CI}_{\text{min}}$ and $(F^{+}, F^{-} | \, F^{+-}) \in \text{CI}_{\text{min}}$.
(a ⇒ b) From (a) follows that $(F^{+}, F^{-} | \, F^{+} \cap F^{-}) \in \text{CI}$. From Proposition 7.3.6.(b) follows that if $(F^{+}, F^{-} | \, G) \in \text{CI}$ and $G \subseteq F^{-}$ then $\sigma(F^{+}|F^{-}) \subseteq G$. Thus $F^{+-} = \sigma(F^{+}|F^{-}) \subseteq F^{+} \cap F^{-}$. By symmetry, $F^{-+} \subseteq F^{+} \cap F^{-}$.
(b ⇒ c) From $(F^{+}, F^{-} | \, F^{+-}) \in \text{CI}$ and from Proposition 2.9.4.(a) follows that $F^{+} \cap F^{-} \subseteq F^{+-}$. With (b) this yields $F^{+-} = F^{+} \cap F^{-}$. By symmetry the other equality follows.
(c ⇒ a) $(F^{+}, F^{-} | \, F^{+} \cap F^{-}) = (F^{+}, F^{-} | \, F^{+-}) \in \text{CI}$ where for the conditional independence use is made of Propositions 19.8.2.(c).
(d ⇒ a) This is obvious.
(c ⇒ d) $(F^{+}, F^{-} | \, F^{+} \cap F^{-}) = (F^{+}, F^{-} | F^{+-}) \in \text{CI}_{\text{min}}$ by Proposition 7.3.6. □

Proposition 7.3.22. *Consider* F^{+}, F^{-}, $G \subseteq F$. *Assume that* $(F^{+}, F^{-} | \, G) \in \text{CI}$. *Then the following statements all hold:*

(a)$(F^{+} \vee G, F^{-} \vee G | \, G) \in \text{CI}$*;*
(b)$G = (F^{+} \vee G) \cap (F^{-} \vee G)$*;*
(c)$\sigma(F^{+} \vee G | F^{-} \vee G) = (F^{+} \vee G) \cap (F^{-} \vee G) = \sigma(F^{-} \vee G | F^{+} \vee G)$*;*
(d)For any sub-σ*-algebra* $G_m \subseteq G$ *such that* $(F^{+} \vee G, F^{-} \vee G | \, G_m) \in \text{CI}$ *it is true that* $G_m = G$.

Proof. (a) Note that $(F^{+}, F^{-} | \, G) \in \text{CI}$ and Proposition 19.8.2.(f) imply that (a) holds.
(b) Then $G \subseteq (F^{+} \vee G) \cap (F^{-} \vee G) \subseteq G$ where the first inclusion holds because $G \subseteq F^{+} \vee G$ and $G \subseteq F^{-} \vee G$ and the second inclusion follows from (a) and Proposition 2.9.4, hence equality holds.
(c) From (a) and from Proposition 7.3.6.(a) follows that,

$$
\begin{aligned}
& (F^{+} \vee G, F^{-} \vee G | \sigma(F^{+} \vee G | F^{-} \vee G)) \in \text{CI}, \\
\Rightarrow \; & G \subseteq (F^{+} \vee G) \cap (F^{-} \vee G) \subseteq \sigma(F^{+} \vee G | F^{-} \vee G) \subseteq G, \\
& \text{where the second inclusion follows from Proposition 2.9.4.(a) and} \\
& \text{the third inclusion from (b),} \\
\Rightarrow \; & G = \sigma(F^{+} \vee G | F^{-} \vee G) = (F^{+} \vee G) \cap (F^{-} \vee G).
\end{aligned}
$$

By symmetry, $G = \sigma(F^{-} \vee G | F^{+} \vee G)$. From (b) the result follows.
(d) Note that by the assumed conditional independence and by Proposition 2.9.4.(a), $G \subseteq (F^{-+} \vee G) \cap (F^{+-} \vee G) \subseteq G_m$ hence $G_m = G$. □

Definition 7.3.23. Consider a probability space (Ω, F, P) and a tuple of sub-σ-algebras (F^{+}, F^{-}). Define the *set of tuples of Hamiltonian sub-*σ*-algebras* as,

$$
\mathbf{H}(F^{-+}, F^{+-}) = \left\{ \begin{array}{l} (H^{+}, H^{-}) \in \mathbf{F} \times \mathbf{F} | \\ F^{-+} \subseteq H^{+} \subseteq F_f, \; F^{+-} \subseteq H^{-} \subseteq F_f, \\ H^{+} \cap H^{-} = \sigma(H^{+}|H^{-}) = \sigma(H^{-}|H^{+}) \end{array} \right\}.
$$

Call any $(H^+,H^-) \in \mathbf{H}$ a *tuple of Hamiltonian sub-σ-algebras* of (F^{-+},F^{+-}).

Define on the set $\mathbf{H}$ the partial-order relation $\leq$,

$$(H_1^+,H_1^-) \leq (H_2^+,H_2^-) \text{ if } \left\{H_1^+ \subseteq H_2^+ \text{ and } H_1^- \subseteq H_2^-\right\}.$$

It follows from the discussion of Subsection 7.3.5 that the set $\mathbf{H}(F^{-+},F^{+-})$ is not a totally ordered subset.

An element $(H^+,H^-) \in \mathbf{H}(F^+,F^-)$ is said to be *minimal tuple of Hamiltonian sub-σ-algebras* if it is a minimal element with respect to the defined partial-order relation. (See Def. 17.1.9 for the concept of a minimal element of a partially ordered set.) Equivalently, $(H^+,H^-) \in \mathbf{H}_{min}(F^{-+},F^{+-})$ if for any $(H_1,H_2) \in \mathbf{H}$, with $(H_1,H_2) \leq (H^+,H^-)$, then $(H_1,H_2) = (H^+,H^-)$. Denote the subset of such minimal elements,

$$\mathbf{H}_{min}(F^{-+},F^{+-}) = \left\{ \begin{array}{l} (H^+,H^-) \in \mathbf{H}(F^{-+},F^{+-})| \\ (H^+,H^-) \text{ is a minimal element} \end{array} \right\}.$$

Define the *stochastic realization maps*,

$$\begin{aligned} f_{sr}(H^+,H^-) &= H^+ \cap H^-,\ f_{sr} : \mathbf{H}(F^{-+},F^{+-}) \rightarrow \mathbf{G}(F^+,F^-), \\ f_{sr,min}(H^+,H^-) &= H^+ \cap H^-,\ f_{sr,min} : \mathbf{H}_{min}(F^{-+},F^{+-}) \rightarrow \mathbf{G}_{min}(F^+,F^-). \end{aligned}$$

The sets $\mathbf{G}(F^+,F^-)$ and $\mathbf{G}_{min}(F^+,F^-)$ are defined in Def. 7.3.1.

Theorem 7.3.24. Minimal state σ-algebras – A second characterization. *Consider the setting of Def. 7.3.23. Recall from Problem 7.3.2 the notation of* $\mathbf{G}(F^+,F^-)$.

(a)The function f_{sr} is well defined.
(b)The function f_{sr} is a bijection.
(c)The inverse function of f_{sr} is specified by

$$f_{sr}^{-1}(G) = (F^{-+} \vee G, F^{+-} \vee G),\ f_{sr}^{-1} : \mathbf{G}(F^+,F^-) \rightarrow \mathbf{H}(F^{-+},F^{+-}).$$

(d)Consider $G_1 = f_{sr}(H_{11},H_{12})$ and $G_2 = f_{sr}(H_{21},H_{22})$. Then $G_1 \subseteq G_2$ if and only if $(H_{11},H_{12}) \leq (H_{21},H_{22})$. In words, both f_{sr} and f_{sr}^{-1} are monotone maps.
(e)The map $f_{sr,min} : \mathbf{H}_{min} \rightarrow \mathbf{G}_{min}$ is well defined and a bijection.
(f) If (H_1^+,H_1^-), $(H_2^+,H_2^-) \in \mathbf{H}_{min}$ then they are incomparable with respect to the partial order relation. Equivalently, neither $(H_1^+,H_1^-) \leq (H_2^+,H_2^-)$ nor $(H_2^+,H_2^-) \leq (H_1^+,H_1^-)$ holds.

The proof of the theorem uses the following lemma.

Lemma 7.3.25. *Consider a tuple of σ-algebras F^+, $F^- \subseteq F$ and a tuple $(H^+,H^-) \in \mathbf{H}(F^{-+},F^{+-})$.*

Then $H^+ = F^{-+} \vee (H^+ \cap H^-)$ and $H^- = F^{+-} \vee (H^+ \cap H^-)$.

Proof. From the definition of $(H^+,H^-) \in \mathbf{H}(F^{-+},F^{+-})$ follows that,

$$\begin{aligned}
& F^{-+} \subseteq H^{+} \subseteq F^{-+} \vee F^{+-},\ F^{+-} \subseteq H^{-} \subseteq F^{-+} \vee F^{+-} = F_f, \\
H^{+} \cap H^{-} \ &= \ \sigma(H^{+}|H^{-}) = \sigma(H^{-}|H^{+});\ \Rightarrow \\
(\supseteq) \quad & F^{-+} \subseteq H^{+} \ \Rightarrow\ F^{-+} \vee (H^{+} \cap H^{-}) \subseteq H^{+}; \\
(\subseteq) \quad & (H^{+}, H^{-} | H^{+} \cap H^{-}) = (H^{+}, H^{-} | \sigma(H^{-}|H^{+})) \in \mathrm{CI}_{\min}, \\
& \text{by Proposition 7.3.6.(a),} \\
& \Rightarrow F^{+-} \subseteq H^{-} \ \Rightarrow\ (H^{+}, F^{+-} | H^{+} \cap H^{-}) \in \mathrm{CI}, \text{ by restriction;} \\
& \Rightarrow \sigma(F^{+-} | H^{+} \cap H^{-}) \subseteq F^{-+} \vee (H^{+} \cap H^{-}) \subseteq H^{+} \vee (H^{+} \cap H^{-}), \\
& \Rightarrow (H^{+}, F^{+-} | F^{-+} \vee (H^{+} \cap H^{-})) \in \mathrm{CI}, \text{ by Proposition 19.8.4.(e),} \\
& \Leftrightarrow \sigma(H^{+} | F^{+-} \vee F^{-+} \vee (H^{+} \cap H^{-})) \subseteq F^{-+} \vee (H^{+} \cap H^{-}), \\
& \text{by Proposition 19.8.4.(e),} \\
& \Leftrightarrow H^{+} \subseteq F^{-+} \vee (H^{+} \cap H^{-}), \\
& \text{because } H^{+} \subseteq F^{-+} \vee F^{+-} \subseteq F^{-+} \vee F^{+-} \vee (H^{+} \cap H^{-}), \\
\Rightarrow \quad & H^{+} = F^{-+} \vee (H^{+} \cap H^{-}).
\end{aligned}$$

By symmetry the second statement follows. □

Proof. Proof of Theorem 7.3.24. (a) To prove that f_{sr} is well defined, let $(H^{+}, H^{-}) \in \mathbf{H}(F^{-+}, F^{+-})$. Define $G = H^{+} \cap H^{-}$. From $(H^{+}, H^{-}) \in \mathbf{H}$ follows that $G = f_{sr}(H^{+}, H^{-}) = H^{+} \cap H^{-} \subseteq H^{+} \subseteq F^{-+} \vee F^{+-}$ and similarly that $G \subseteq H^{-} \subseteq F^{-+} \vee F^{+-}$. From Proposition 7.3.6.(a) follows that $(H^{+}, H^{-} | \sigma(H^{+}|H^{-})) \in \mathrm{CI}$. Because by $(H^{+}, H^{-}) \in \mathbf{H}$ and $G = H^{+} \cap H^{-} = \sigma(H^{+}|H^{-})$, it follows that $(H^{+}, H^{-} | G) = (H^{+}, H^{-} | \sigma(H^{+}|H^{-})) \in \mathrm{CI}$. By the inclusions $F^{-+} \subseteq H^{+}$ and $F^{+-} \subseteq H^{-}$ and by the property that the conditional independence relations is closed with respect to restrictions, it follows that $(F^{-+}, F^{+-} | G) \in \mathrm{CI}$. Finally the above arguments, $G \subseteq F^{-+} \vee F^{+-}$, and Theorem 7.3.10.(b) imply that $(F^{+}, F^{-} | G) \in \mathrm{CI}$. Thus f_{sr} is well defined.

(b) It is proven that f_{sr} is surjective. Consider $G \subseteq F^{-+} \vee F^{+-}$ such that $(F^{+}, F^{-} |\ G) \in \mathrm{CI}$. By restriction it follows that $(F^{-+}, F^{+-} |\ G) \in \mathrm{CI}$. Define $H^{+} = F^{-+} \vee G$ and $H^{-} = F^{+-} \vee G$. Then $F^{-+} \subseteq H^{+} = F^{-+} \vee G \subseteq F^{-+} \vee F^{+-}$ and $F^{+-} \subseteq H^{-} = F^{+-} \vee G \subseteq F^{-+} \vee F^{+-}$. From $(F^{-+}, F^{+-} | G) \in \mathrm{CI}$ and Proposition 7.3.22.(b & c) follows that,

$$G = (F^{-+} \vee G) \cap (F^{+-} \vee G) = H^{+} \cap H^{-}.$$

Then,

$$\begin{aligned}
& (H^{+}, H^{-} |\ G) = (F^{-+} \vee G, F^{+-} \vee G |\ G) \in \mathrm{CI}, \text{ by Proposition 7.3.22,} \\
& \Rightarrow \text{ (by Proposition 7.3.21),} \\
G \ &= \ H^{+} \cap H^{-} = \sigma(H^{-}|H^{+}) = \sigma(H^{-}|H^{+}); \\
& \Rightarrow (H^{+}, H^{-}) \in \mathbf{H}(F^{-+}, F^{+-}),\ G = H^{+} \cap H^{-} = f_{sr}(H^{+}, H^{-}).
\end{aligned}$$

Next it is proven that f_{sr} is injective. Consider $G = f_{sr}(H_{11}, H_{12}) = f_{sr}(H_{21}, H_{22})$. From Lemma 7.3.25 follows that,

$$H_{11} = F^{-+} \vee (H_{11} \cap H_{12}) = F^{-+} \vee G = F^{-+} \vee (H_{21} \cap H_{22}) = H_{21};$$
$$H_{12} = H_{22}, \text{ similarly.}$$

Thus f_{sr} is injective and hence bijective.
(c) This follows from the proof of (b) for injectiveness.
(d)

$$\begin{aligned} (\Leftarrow)\ & G_1 = f_{sr}(H_{11}, H_{12}) = (H_{11} \cap H_{12}) \subseteq (H_{21} \cap H_{22}) = G_2; \\ (\Rightarrow)\ & H_{11} = F^{-+} \vee G_1 \subseteq F^{-+} \vee G_2 = H_{21}, \\ & H_{12} = F^{+-} \vee G_1 \subseteq F^{+-} \vee G_2 = H_{22}, \text{ by Lemma 7.3.25,} \\ & \Rightarrow\ (H_{11}, H_{12}) \leq (H_{21}, H_{22}). \end{aligned}$$

(e) It is first proven that the function $f_{sr,min}$ is well defined. Because $\mathbf{H}_{min}(F^{-+}, F^{+-}) \subseteq \mathbf{H}(F^{-+}, F^{+-})$ the function $f_{sr,min}$ maps to $\mathbf{G}(F^{-+}, F^{+-})$. It is to be proven that it maps to $\mathbf{G}_{min}$.

Consider $(H^+, H^-) \in \mathbf{H}_{min}(F^{-+}, F^{+-})$ and $G = f_{sr}(H^+, H^-)$. To prove minimality of G, consider a σ-algebra $G_1 \subseteq G$ such that $(F^{-+}, F^{+-} | \ G_1) \in \text{CI}$. From (c) follows that $(H_{11}, H_{12}) = f_{sr}^{-1}(G_1) \in \mathbf{H}$.
From $f_{sr}(H_{11}, H_{12}) = G_1 \subseteq G = f_{sr}(H^+, H^-)$ and from (d) follows that $(H_{11}, H_{12}) \leq (H^+, H^-)$ and the assumed minimality of $(H^+, H^-) \in \mathbf{H}_{min}(F^+, F^-)$ implies that $(H_{11}, H_{12}) = (H^+, H^-)$. Hence $G_1 = H_{11} \cap H_{12} = H^+ \cap H^- = G$. From this follows that $(F^{-+}, F^{+-} | G) \in \text{CI}_{\text{min}}$.

Next it is proven that $f_{sr,min}$ is surjective. Take $G \in \mathbf{G}_{min}(F^{-+}, F^{+-})$ hence $(F^{-+}, F^{+-} | \ G) \in \text{CI}_{\text{min}}$. Because of (a), $(H^+, H^-) = f_{sr}^{-1}(G) \in \mathbf{H}(F^{-+}, F^{+-})$. It is to be proven that $(H^+, H^-) \in \mathbf{H}_{min}$. Let $(H_1, H_2) \in \mathbf{H}$ be such that $(H_1, H_2) \leq (H^+, H^-)$. From (a) and (d) then follows that $G_1 = f_{sr}(H_1, H_2)$ satisfies $(F^{-+}, F^{+-} | G_1) \in \text{CI}$ and $G_1 = H_1 \cap H_2 \subset H^+ \cap H^- = G$. Because of the minimality of G, $(F^{-+}, F^{+-} | \ G) \in \text{CI}_{\text{min}}$, it follows that $G_1 = G$. Hence $(H_1, H_2) = f_{sr}^{-1}(G_1) = f_{sr}^{-1}(G) = (H^+, H^-)$ by (d). Thus $(H^+, H^-) \in \mathbf{H}_{min}$.

It is to be proven that $f_{sr,min}$ is injective. This follows directly from (b).
(f) Suppose that $(H_1^+, H_1^-) \leq (H_2^+, H_2^-)$ but not equal. Define $G_1 = H_1^+ \cap H_1^-$ and $G = H_2^+ \cap H_2^-$. From (e) follows that $G_1 = f_{sr,min}(H_1^+, H_1^-) \in \mathbf{G}_{min}$ and $G_2 = f_{sr,min}(H_2^+, H_2^-) \in \mathbf{G}_{min}$. Thus $(F^+, F^- | \ G_1) \in \text{CI}_{\text{min}}$ and $(F^+, F^- | \ G_2) \in \text{CI}_{\text{min}}$. From (d) follows that $G_1 = f_{sr,min}(H_1^+, H_1^-) = H_1^+ \cap H_1^- \subsetneq H_2^+ \cap H_2^- = G_2$. But then $(F^+, F^- | \ G_2) \notin \text{CI}_{\text{min}}$. This is a contradiction of the supposition. The alternative case is similar. □

Example 7.3.26. There follows a special case of Theorem 7.3.24. Recall from Proposition 7.3.18 the notation $G^- = \sigma(F^+ | \ F^-)$ and that then $(F^+, F^- | \ G^-) \in \text{CI}_{\text{min}}$ and $G^- \subseteq F^f$. The associated tuple of Hamiltonian σ-algebras is then $H^+ = \sigma(F^- | \ F^+) \vee G^- = F_f$, $H^- = \sigma(F^+ | \ F^-) \vee G^- = \sigma(F^+ | \ F^-) = G^-$, and $(H^+, H^-) = (F_f, \ F^{+-})$.

7.3.8 Relations of Tuples of Minimal State Sigma-Algebras

How are minimal state σ-algebras related? The relation between minimal state σ-algebras is also a research issue of realization theory. The discussion motivates the following concept. The reader should keep in mind that, according to Subsection 7.3.5, the sets $FG^{+}_{min}(F^{+},F_f)$ and $FG^{-}_{min}(F^{-},F_f)$ are partially ordered and not totally ordered. The consequences of this conclusion show up in this subsection.

Definition 7.3.27. Consider the set of minimal triples of σ-algebras,

$$CI_3(F^{+},F^{-}) = \left\{ (F^{+},F^{-},G) \in \mathbf{F}\times\mathbf{F}\times\mathbf{F} |\ (F^{+},F^{-}|G) \in \mathrm{CI}_{\mathrm{min}},\ G \subseteq F_f \right\}.$$

Define the *σ-algebraic isomorphism relation* by the formula,

$$\mathrm{CI}_{\mathrm{isom}}(F^{+},F^{-}) = \left\{ \begin{array}{l} ((F^{+},F^{-},G_1),(F^{+},F^{-},G_2)) \in CI_3 \times CI_3 | \\ \sigma(G_1|G_2) = G_2 \text{ and } \sigma(G_2|G_1) = G_1 \end{array} \right\}.$$

Notation used also is $(G_1,G_2) \in \mathrm{CI}_{\mathrm{isom}}$ if the σ-algebras F^{+} and F^{-} are clear. One then says that the two triples of $\mathrm{CI}_{\mathrm{min}}$ are *isomorphic*; equivalently, the two σ-algebras (G_1,G_2), are *isomorphic*.

The concept of an isomorphy in algebra requires consistency of algebraic operations which in this case refers to the minimal conditional independence relation.

Example 7.3.28. *Non-isometry*. Consider Example 7.3.12 with a finite probability space. Recall from there that,

$$G_2,\ G_3 \subseteq F_f = F^{-+} \vee F^{+-} = F^{+} \vee F^{-},$$
$$(F^{+},F^{-}|G_2) \in \mathrm{CI}_{\mathrm{min}}, \text{ and } (F^{+},F^{-}|G_3) \in \mathrm{CI}_{\mathrm{min}}.$$

Yet $\sigma(G_2|G_3) \neq G_3$ and $\sigma(G_3|G_2) \neq G_2$ hence the two minimal state σ-algebras G_2 and G_3 are not isomorphic and $(G_2,\ G_3) \notin \mathrm{CI}_{\mathrm{isom}}(F^{+},F^{-})$.

Proposition 7.3.29. *The σ-algebraic isomorphism relation on* $\mathrm{CI}_{\mathrm{min}}$ *is a relation which is reflexive and symmetric. However, in general it is not a transitive relation.*

Proof. (1) Reflexivity means that $(F^{+},F^{-}|G),(F^{+},F^{-}|G)) \in \mathrm{CI}_{\mathrm{isom}}$ which follows because $\sigma(G|G) = G$.
(2) The symmetry follows directly from the definition which is symmetric in terms of G_1 and G_2.
(3) Example 7.3.28 shows that transitivity does not hold in general. Note that from $(F^{-+},F^{+-}|\ G_2) \in \mathrm{CI}_{\mathrm{min}}$, $(F^{-+},F^{+-}|\ G_3) \in \mathrm{CI}_{\mathrm{min}}$, and Theorem 7.3.30.(b) follows that $\sigma(G_2|\ F^{-+}) = F^{-+}$ and $\sigma(F^{-+}|\ G_3) = G_3$. Transitivity then implies that $\sigma(G_2|\ G_3)$. Yet, Example 7.3.28 claims that $\sigma(G_2|G_3) \neq G_3$. Hence transitivity does not hold. □

Theorem 7.3.30. Sufficient conditions for minimal state σ-algebras to be isomorphic. *Consider the σ-algebras $F^{+},\ F^{-} \subseteq F$ and recall that*
$F^{-+} = \sigma(F^{-}|F^{+})$ and $F^{+-} = \sigma(F^{+}|F^{-})$.

(a)$(F^{-+}, F^{+-}) \in \mathrm{CI}_{\mathrm{isom}}$ *are isomorphic* σ*-algebras.*

(b)If $(F^{-+}, F^{+-}|G) \in \mathrm{CI}_{\mathrm{min}}$ *and* $G \subseteq F_f = F^{-+} \vee F^{+-}$ *then* $(G, F^{-+}) \in \mathrm{CI}_{\mathrm{isom}}$ *and* $(G, F^{+-}) \in \mathrm{CI}_{\mathrm{isom}}$ *are isomorphic tuples.*

(c)If (1) $(F^{-+}, F^{+-}|G_1) \in \mathrm{CI}_{\mathrm{min}}$ *and* $(F^{-+}, F^{+-}|G_2) \in \mathrm{CI}_{\mathrm{min}}$*; (2)* $G_1, G_2 \subseteq F_f$*; (3) either* $G_2 \subseteq F^{-+} \vee G_1$ *or* $G_2 \subseteq F^{+-} \vee G_1$*; and (4) either* $G_1 \subseteq F^{-+} \vee G_2$ *or* $G_1 \subseteq F^{+-} \vee G_2$*; then* $(G_1, G_2) \in \mathrm{CI}_{\mathrm{isom}}$ *are isomorphic tuples. The formulation of the conditions (3) and (4) are related to the conclusion that the sets* FG^+_{min} *and* FG^-_{min} *are partially ordered sets which are in general not totally ordered. The conditions (3) and (4) refer to the case where the* σ*-algebras* $F^{-+} \vee G_1$ *and* $F^{-+} \vee G_2$ *belong to the same branch of totally ordered elements.*

(d)If $G_1,\ G_2 \in CI_3$ *are such that* $(G_1, G_2) \in \mathrm{CI}_{\mathrm{isom}}$ *then the four conditions of (c) hold.*

Proof. (a) It follows from Proposition 7.3.6.(a) that $(F^{-+}, F^{+-}|F^{-+}) \in \mathrm{CI}_{\mathrm{min}}$ and that $(F^{-+}, F^{+-}|F^{+-}) \in \mathrm{CI}_{\mathrm{min}}$. It follows from Proposition 19.8.6.(h) that $\sigma(F^{-+}|F^{+-}) = F^{+-}$ and that $\sigma(F^{-+}|F^{+-}) = F^{-+}$. Thus $(F^{-+}, F^{+-}) \in \mathrm{CI}_{\mathrm{isom}}$.
(b) Note that,

$$\begin{aligned}
& G \subseteq F_f = F^{-+} \vee F^{+-} \text{ and } (F^+, F^-|G) \in \mathrm{CI}, \\
\Rightarrow\ & \sigma(G|F^{+-}) = F^{+-} \text{ and } \sigma(G|F^{-+}) = F^{-+}, \text{ because of Proposition 7.3.8,} \\
& (F^{-+}, F^{+-}|G) \in \mathrm{CI}_{\mathrm{min}} \\
\Rightarrow\ & \sigma(F^{-+}|G) = G \text{ and } \sigma(F^{+-}|G) = G, \text{ by Proposition 7.3.15;} \\
& \sigma(G|F^{-+}) = F^{-+} \text{ and } \sigma(F^{-+}|G) = G \Rightarrow\ (G, F^{-+}) \in \mathrm{CI}_{\mathrm{isom}}; \\
& \sigma(G|F^{+-}) = F^{+-} \text{ and } \sigma(F^{+-}|G) = G \Rightarrow\ (G, F^{+-}) \in \mathrm{CI}_{\mathrm{isom}}.
\end{aligned}$$

(c) Note that,

$$\begin{aligned}
& (F^{-+}, F^{+-}|G_1) \in \mathrm{CI}_{\mathrm{min}} \text{ and } G_2 \subseteq F^{-+} \vee G_1 \\
\Rightarrow\ & (F^{-+}, F^{+-}|\sigma(G_2|G_1)) \in \mathrm{CI}, \text{ by Proposition 19.8.4,} \\
\Rightarrow\ & \sigma(G_2|G_1) = G_1, \text{ by } \sigma(G_2|G_1) \subseteq G_1 \text{ and by the minimality of } G_1.
\end{aligned}$$

If $G_2 \subseteq F^{+-} \vee G_1$ then a similar argument, with F^{-+} and F^{+-} interchanged, can be applied yielding that $\sigma(G_2|G_1) = G_1$. Condition (4) produces in a corresponding way in either case that $\sigma(G_2|\ G_1) = G_1$.
(d) Consider $G_1,\ G_2 \in CI_3$ such that $(G_1, G_2) \in \mathrm{CI}_{\mathrm{isom}}$. From Theorem 7.3.24.(e) follows that there exist $(H_1^+, H_1^-),\ (H_2^+, H_2^-) \in \mathbf{H}_{min}$ such that $G_1 = f_{sr,min}(H_1^+, H_1^-)$ and $G_2 = f_{sr,min}(H_2^+, H_2^-)$. Note that $H_1^+ = F^{-+} \vee G_1$ and $H_1^- = F^{+-} \vee G_1$ and corresponding relations for (H_2^+, H_2^-).

If $(H_1^+, H_1^-) \le (H_2^+, H_2^-)$ then if follows from Theorem 7.3.24.(d) that $G_1 \subsetneq G_2$ which contradicts the minimality of G_1. Consequently, if $H_1^+ \subseteq H_2^+$ then $H_1^- \subseteq H_2^-$ cannot hold hence $H_2^- \le H_2^+$. Note that $H_1^+ = F^{-+} \vee G_1 \subseteq (F^{-+} \vee G_2) = H_2^+$ if and only if $G_1 \subseteq (F^{-+} \vee G_2)$; and $H_1^- = F^{+-} \vee G_1 \subseteq (F^{+-} \vee G_2) = H_2^-$ if and only if $G_1 \subseteq (F^{+-} \vee G_2) = H_2^-$. Thus the conditions (3) and (4) hold. A similar conclusion holds for the converse inclusion relation. □

Example 7.3.31. Reconsider Example 7.3.12 and Example 7.3.28. Recall the σ-algebras F^+, F^-, G_2, and G_3 from there. A calculation shows that the following non-inclusions all hold,

$$G_3 \not\subseteq F^+ \vee G_2,\ G_3 \not\subseteq F^- \vee G_2,\ G_2 \not\subseteq F^+ \vee G_3,\ G_2 \not\subseteq F^- \vee G_3.$$

The σ-algebra $F^+ \vee G_2$ is calculated by taking as its atoms all intersections of the atoms of F^+ and of G_2. The inclusion $G_3 \subseteq F^+ \vee G_2$ holds if every atom of G_3 either is an atom of $F^+ \vee G_2$ or is a union of atoms of $F^+ \vee G_2$. The atom (5) is an atom of G_3 but neither an atom of $F^+ \vee G_2$ nor a union of atoms of $F^+ \vee G_2$; etc.

This example shows that neither condition Theorem 7.3.30.(c.3) holds nor condition Theorem 7.3.30.(c.4) holds for the σ-algebras G_2 and G_3.

Proposition 7.3.32. *The following statements are equivalent:*

(a)$(G_1, G_2) \in \mathrm{CI}_3$*;*
(b)(H_1^+, H_1^-), $(H_2^+, H_2^-) \in \mathbf{H}_{min}(F^{-+}, F^{+-})$ *are incomparable with respect to the order relation and such that either (1)* $H_1^+ \subseteq H_2^+$ *and* $H_2^- \subseteq H_1^-$ *or (2)* $H_2^+ \subseteq H_1^+$ *and* $H_1^- \subseteq H_2^-$ *hold.*

The relations between the above considered objects are by Theorem 7.3.24, $G_1 = H_1^+ \cap H_1^-$ *and* $G_2 = H_2^+ \cap H_2^-$.

Proof. This follows directly from Theorem 7.3.30.(c) and (d). □

Within the set of minimal state σ-algebras there exits thus tuples of minimal state σ-algebras which are not isomorphic. It is of interest to further investigate this relation.

7.4 Stochastic Realization of a Sigma-Algebra Family

In this section the theory of the previous section is extended to a family of σ-algebras.

The formulation of the theory has been inspired by the corresponding framework for Hilbert spaces formulated by G. Picci and A. Lindquist, and G. Ruckebusch and M. Metivier. These researchers were inspired by the Lax-Phillips scattering theory.

7.4.1 Problem

Problem 7.4.1. *Strong stochastic realization of a σ-algebra family.* Consider a σ-algebra family $\{F_t,\ \forall\, t \in T\}$. Define the subproblems:

(a) Does there exist a σ-algebra family $\{G_t \subseteq F,\ \forall\, t \in T\}$ such that $\{F_t, G_t,\ \forall\, t \in T\}$ is a σ-algebraic system and hence a stochastic realization of the family $\{F_t,\ t \in T\}$?

(b)Characterize those stochastic realizations of $\{F_t,\ G_t,\ t \in T\}$ which are minimal stochastic realizations.

(c)Classify or describe all σ-algebra families $\{G_t,\ \forall\ t \in T\}$ such that $\{F_t, G_t,\ \forall\ t \in T\}$ is a minimal stochastic realization. Relate tuples of minimal state σ-algebras.

It is possible to generalize the stochastic realization problem to the case in which the state σ-algebra is not restricted to belong to the frame σ-algebra but may depend on an additional σ-algebra.

The above problem will be solved using the concepts and results of Section 7.3.

7.4.2 Concepts

Recall the following concept from Def. 5.8.1 which is repeated her for the convenience of the readers.

Definition 7.4.2. A *σ-algebraic system.* The σ-algebra families $\{F_t,\ G_t \subseteq F,\ \forall\ t \in T = \mathbb{N}\}$ are said to form a *σ-algebraic system* if,

$$\forall\ t \in T,\ (F_t^+ \vee G_t^+,\ F_{t-1}^- \vee G_t^- \mid G_t) \in \text{CI};$$
$$\text{where } F_t^+ = \vee_{s \geq t}\ F_s,\ F_{t-1}^- = \vee_{s \leq t-1}\ F_s,\ G_t^+ = \vee_{s \geq t}\ G_s,\ G_t^- = \vee_{s \leq t}\ G_s.$$

The reader may think of the σ-algebra F_t as representing the output of a system at time $t \in T$ and of G_t as representing the state of a system at the same time. The definition imposes the requirement that, at any time $t \in T$, the current state σ-algebra G_t makes the combined future of the output process and of the state process conditionally independent from the combined past of the output process and of the state process. This definition is consistent with Def. 4.2.2.

In the definition below the adjective *strong* could have been added to the terms. This is not done not to overload the terminology.

Definition 7.4.3. *Stochastic realization of a σ-algebra family.* Consider a σ-algebra family $\{F_t \subseteq F,\ \forall\ t \in T\}$ with $T = \mathbb{N}$.

Call the joint σ-algebra family $\{F_t,\ G_t \in \mathbf{F},\ \forall\ t \in T\}$ a *stochastic realization* of the family $\{F_t,\ \forall\ t \in T\}$ if:
(1) $\{F_t,\ G_t,\ \forall\ t \in T\}$ satisfies the definition of a σ-algebraic system; equivalently, for all $t \in T$, $(F_t^+ \vee G_t^+, F_{t-1}^- \vee G_t^- \mid G_t) \in \text{CI}$; and
(2) for all $t \in T$, $G_t \subseteq F_t^+ \vee F_{t-1}^-$.

Call it a *minimal stochastic realization* if (1) it is a stochastic realization and (2) mimimality holds: if $\{F_t, \overline{G}_t,\ \forall\ t \in T\}$ is another stochastic realization of the σ-algebra family such that for all $t \in T$, $\overline{G}_t \subseteq G_t$, then $\overline{G}_t = G_t$.

Definition 7.4.4. Consider a joint σ-algebra family $\{F_t,\ G_t \in \mathbf{F},\ \forall\ t \in T\}$ Define the concepts,

(a) *stochastic controllability* by the condition that $\sigma(F_t^+ |\, F_{t-1}^-) = \sigma(G_t |\, F_{t-1}^-)$ for all $t \in T$;
(b) *stochastic co-controllability* by the condition $\sigma(F_{t-1}^- |\, F_t^+) = \sigma(G_t |\, F_t^+)$ for all $t \in T$;
(c) *stochastic observability* by the condition $\sigma(F_t^+ |\, G_t) = G_t$ for all $t \in T$; and
(d) *stochastic co-observability* by the condition $\sigma(F_{t-1}^- |\, G_t) = G_t$ for all $t \in T$.

The above defined concepts correspond to those of Def. 7.3.4 for a tuple of σ-algebras.

In case the output σ-algebra family $\{F_t,\ t \in T\}$ is generated by a stationary process, then the conditions of stochastic controllability etc need not be checked for all times $t \in T$ but for one time only.

Recall the notation for the past and the future of the output σ-algebra and the state σ-algebra, $F_t^+ = \vee_{s>t} F_s$, $F_{t-1}^- = \vee_{s \leq t} F_s$, $G_t^+ = \vee_{s>t} G_s$, and $G_{t-1}^- = \vee_{s \leq t} G_s$. Then $\{F_t,\ t \in T\}$ is not a filtration because in general $F_t \not\subseteq F_{t+1}$ but $\{F_t^-,\ t \in T\}$ is a filtration, $F_t^- \subseteq F_{t+1}^-$. Similary $\{F_t^+,\ t \in T\}$ is a backward filtration meaning that for all $t \in T$, $F_t^+ \supseteq F_{t+1}^+$. Therefore one refers to the notation $\{F_t,\ G_t,\ \forall\, t \in T\}$ as a *family* of a tuple of σ-algebras.

Definition 7.4.5. Consider a family of σ-algebras satisfying,

$$\{F_t,\ G_t,\ \forall\, t \in T\},\ \forall\, t \in T,\ (F_t^+,\ F_{t-1}^- |\, G_t) \in \text{CI},\ G_t \subseteq F_\infty^- = F_t^+ \vee F_{t-1}^-.$$

(a) Call this family a *transitive projection consistent family* of σ-algebras if

$$\sigma(F_{t-1}^- \vee G_t^- |\, F_t^+ \vee G_t^+) \subseteq F_t^+ \vee G_t,\ \ \forall\, t \in T,$$
$$\sigma(F_t^+ \vee G_t^+ |\, F_{t-1}^- \vee G_t^-) \subseteq F_{t-1}^- \vee G_t,\ \ \forall\, t \in T.$$

(b) Call this family a *transitive* family of σ-algebras if the following two conditions both hold: (1) $\{F_{t-1}^- \vee G_t,\ \forall\, t \in T\}$ is increasing in T; (2) $\{F_t^+ \vee G_t,\ \forall\, t \in T\}$ is decreasing in T; or, equivalently, if

$$(1)\ \ \forall\, s,\ t \in T,\ s < t \Rightarrow\ (F_{s-1}^- \vee G_s) \subseteq (F_{t-1}^- \vee G_t),\ \text{and}$$
$$(2)\ \ \forall\, s,\ t \in T,\ s < t \Rightarrow\ (F_s^+ \vee G_s) \supseteq (F_t^+ \vee G_t).$$

(c) Call this family a *current-state family* of σ-algebras if,

$$\forall\, t \in T,\ F_{t-1}^- \vee G_t = F_{t-1}^- \vee G_t^-,\ \ F_t^+ \vee G_t = F_t^+ \vee G_t^+.$$

Note that $\{F_{t-1}^- \vee G_t^-,\ t \in T\}$ is a filtration but $\{F_{t-1}^- \vee G_t,\ t \in T\}$ is not a filtration because in general $G_t \not\subseteq G_{t+1}$. However, if the condition of a current-state family holds then $\{F_{t-1}^- \vee G_t,\ t \in T\}$ is a filtration because of the equality of the condition.

(d) The difference between the three defined conditions is illustrated for the future σ-algebras by the formulas

$$(a)\ \ F_t^+ \vee G_t \supseteq \sigma(F_{t-1}^- \vee G_t^- |\, F_t^+ \vee G_t^+),\ \forall\, t \in T;$$
$$(b)\ \ F_t^+ \vee G_t \supseteq F_s^+ \vee G_s,\ \forall\, s,\ t \in T,\ t < s;$$
$$(c)\ \ F_t^+ \vee G_t = F_t^+ \vee G_t^+,\ \ \forall\, t \in T.$$

Lemma 7.4.6. *Consider the σ-algebra family $\{F_t,\ G_t,\ \forall\, t \in T\}$. Assume that for all $t \in T$, $(F_t^+,\ F_{t-1}^-|\ G_t) \in \mathrm{CI}$.*

(a)The family $\{F_t,\ G_t,\ \forall\ t \in T\}$ is transitive if and only if it is a current-state σ-algebra family.

(b)If the σ-algebra is transitive then it satisfies the condition of a transitive projection consistent σ-algebra family.

Proof. (a) ($\Rightarrow$) From the definition of transitivity follows that $\{F_{t-1}^- \vee G_t,\ \forall\, t \in T\}$ is increasing, hence, for $s,t \in T$ with $s < t$, $(F_{s-1}^- \vee G_s) \subseteq (F_{t-1}^- \vee G_t)$. Thus,

$$G_s \subseteq (F_{t-1}^- \vee G_t),\ \forall\ s \leq t,\ \Rightarrow\ G_t^- = \vee_{s \leq t}\, G_s \subseteq (F_{t-1}^- \vee G_t),$$
$$\Rightarrow (F_{t-1}^- \vee G_t^-) \subseteq (F_{t-1}^- \vee G_t).$$

From the definition of G_t^- follows that, for all $t \in T$, $(F_{t-1}^- \vee G_t) \subseteq (F_{t-1}^- \vee G_t^-)$, hence $(F_{t-1}^- \vee G_t) = (F_{t-1}^- \vee G_t^-)$. Similarly one proves that $(F_t^+ \vee G_t) = (F_t^+ \vee G_t^+)$.
($\Leftarrow$) Consider $s,t \in$ with $s < t$. Then,

$$F_{s-1}^- \vee G_s = F_{s-1}^- \vee G_s^-,\ \text{by assumption},$$
$$\subseteq F_{t-1}^- \vee G_t^-,\ \text{by definition of } F_{t-1}^- \text{ and } G_t^-,$$
$$= F_{t-1}^- \vee G_t,\ \text{by assumption;}$$
$$F_s^+ \vee G_s = F_s^+ \vee G_s^+ \supseteq F_t^+ \vee G_t^+ = F_t^+ \vee G_t,$$

hence $\{F_t,\ G_t,\ \forall\ t \in T\}$ is transitive.
(b) If the σ-algebra family is transitive then it follows from (a) that it is also a currrent state family. Then, for all $t \in T$,

$$\sigma(F_{t-1}^- \vee G_t^-\,|\ F_t^+ \vee G_t^+) = \sigma(F_{t-1}^- \vee G_t^-\,|\ F_t^+ \vee G_t) \subseteq F_t^+ \vee G_t;$$
$$\sigma(F_t^+ \vee G_t^+|\ F_{t-1}^- \vee G_t^-) = \sigma(F_t^+ \vee G_t^+|\ F_{t-1}^- \vee G_t) \subseteq F_{t-1}^- \vee G_t.$$

Hence it satisfies the condition of the transitive projection consistent family. □

7.4.3 Characterization of a Stochastic Realization

Consider a family of a tuple of σ-algebras $\{F_t,\ G_t,\ \forall\ t \in T\}$ such that, $\forall\ t \in T$, σ-algebra $G_t \subseteq F_\infty^-$ and $(F_t^+,\ F_{t-1}^-|\ G_t) \in \mathrm{CI}$. Such a σ-algebra G_t can be constructed for every $t \in T$ according to the theory of Section 7.3. Such a family
$\{F_t,\ G_t,\ \forall\ t \in T\}$ does not satisfy the condition for being a σ-algebraic system. A condition is needed.

Theorem 7.4.7. Equivalent condition of a σ-algebraic system.
Consider a σ-algebra family $\{F_t,\ G_t,\ \forall\ t \in T\}$ with for all $t \in T$, $G_t \subseteq F_t^+ \vee F_{t-1}^-$. The following statements are equivalent:

(a)The σ-algebra family is a σ-algebraic system;

(b)the σ-algebra family satisfies that (b.1) for all $t \in T$, $(F_t^+,\, F_{t-1}^- |\, G_t) \in \text{CI}_{\min}$; and (b.2) the condition holds of a transitive projection consistent σ-algebra family.

Proof. (b) ⇒ (a). Fix $t \in T$. Note that by condition (b.1) and Proposition 19.8.2.(f)

$$(F_t^+,\, F_{t-1}^- |\, G_t) \in \text{CI}, \text{ hence } (F_t^+ \vee G_t,\, F_{t-1}^- \vee G_t |\, G_t) \in \text{CI}. \tag{7.5}$$

From condition (b.2) and Def. 7.4.5(a) follows that

$$\begin{aligned} &\sigma(F_{t-1}^- \vee G_t^- \,|\, F_t^+ \vee G_t^+) \subseteq (F_t^+ \vee G_t),\ F_{t-1}^- \vee G_t \subseteq F_{t-1}^- \vee G_t^-, \\ \Rightarrow\ &\sigma(F_{t-1}^- \vee G_t |\, F_t^+ \vee G_t^+) \subseteq \sigma(F_{t-1}^- \vee G_t^- \,|\, F_t^+ \vee G_t^+) \subseteq F_t^+ \vee G_t. \end{aligned}$$

The latter inclusion, Eqn. (7.5), and Theorem 19.8.1 imply that

$$(F_t^+ \vee G_t^+,\, F_{t-1}^- \vee G_t |\, G_t) \in \text{CI}. \tag{7.6}$$

Correspondingly,

$$\begin{aligned} &\sigma(F_t^+ \vee G_t^+ |\, F_{t-1}^- \vee G_t^-) \subseteq\ F_{t-1}^- \vee G_t, \text{ by (b.2) and Def.7.4.5.(a)}, \\ &(F_t^+ \vee G_t^+,\, F_{t-1}^- \vee G_t |\, G_t) \in \text{CI}, \text{ by Eqn. (7.6)}, \\ &\text{and Theorem 19.8.1 imply that } (F_t^+ \vee G_t^+,\, F_{t-1}^- \vee G_t^- \,|\, G_t) \in \text{CI}. \end{aligned}$$

Because this holds for all $t \in T$, the family is a σ-algebraic system.
(a) ⇒ (b). Because $\{F_t,\, G_t,\, \forall\, t \in T\}$ is a σ-algebraic system, it follows from the definition of such a system that

$$\begin{aligned} &(F_t^+ \vee G_t^+,\, F_{t-1}^- \vee G_t^- \,|\, G_t) \in \text{CI},\ \forall\, t \in T; \\ \Rightarrow\ &(F_t^+ \vee G_t,\, F_{t-1}^- \vee G_t |\, G_t) \in \text{CI},\ \forall\, t \in T, \text{ by restriction hence (b.1) holds;} \\ &G_t \subseteq G_t^+ = \vee_{s \le t}\, G_s\ \Rightarrow\ (F_t^+ \vee G_t,\, F_{t-1}^- \vee G_t^- \,|\, G_t) \in \text{CI},\ \forall\, t \in T, \\ &\sigma(F_{t-1}^- \vee G_t^- \,|\, F_t^+ \vee G_t^+) \subseteq\ (F_t^+ \vee G_t), \text{ by Theorem 19.8.1.} \end{aligned}$$

Correspondingly,

$$\begin{aligned} &(F_t^+ \vee G_t^+,\, F_{t-1}^- \vee G_t^- \,|\, G_t) \in \text{CI},\ (F_t^+ \vee G_t,\, F_{t-1}^- \vee G_t |\, G_t) \in \text{CI}, \\ &(F_t^+ \vee G_t^+,\, F_{t-1}^- \vee G_t |\, G_t) \in \text{CI}, \text{ both by restriction} \\ \Rightarrow\ &\sigma(F_t^+ \vee G_t^+ |\, F_{t-1}^- \vee G_t^-) \subseteq\ F_{t-1}^- \vee G_t, \end{aligned}$$

by Theorem 19.8.1. Thus by Def. 7.4.5(a) the family satisfies the transitive projection-consistency condition and (b.2) holds. □

Needed is a construction of a transitive σ-algebraic family. Such a construction is defined next.

Definition 7.4.8. *Hamiltonian σ-algebra families.* Consider a σ-algebraic family $\{F_t,\ \forall\ t \in T\}$. Call the σ-algebra family $\{H_t^+,\ H_t^- \subseteq F,\ \forall\ t \in T\}$ a *Hamiltonian σ-algebra family* for the σ-algebra family $\{F_t,\ t \in T\}$ if the following conditions all hold:

1. $\{H_t^+ \subseteq F,\ \forall\, t \in T\}$ is a decreasing family and $\{H_t^- \subseteq F,\ \forall\, t \in T\}$ is an increasing family;

2. for all $t \in T$, $H_t^+ \vee H_t^- = F_t^+ \vee F_{t-1}^- = F_\infty^-$;
3. for all $t \in T$, $F_{t-1}^- \subseteq H_t^-$ and $F_t^+ \subseteq H_t^+$;
4. for all $t \in T$, $H_t^+ \cap H_t^- = \sigma(H_t^+|H_t^-) = \sigma(H_t^-|H_t^+)$.

Theorem 7.4.9. A sufficient condition for a σ-algebraic system.
Consider a σ-algebraic family $\{F_t \subseteq F, \ \forall\, t \in T\}$.

(a)Consider a stochastic realization in the form of a σ-algebraic system, $\{F_t, \ G_t, \ \forall\, t \in T\}$, such that for all $t \in T$, $G_t \subseteq F_\infty^-$. Define,

$$H_t^+ = F_t^+ \vee G_t, \ H_t^- = F_{t-1}^- \vee G_t, \ \{H_t^+, \ H_t^-, \ \forall\, t \in T\}.$$

Assume that the σ-algebra family $\{F_t, \ G_t, \ \forall\, t \in T\}$ is transitive.
Then the family $\{H_t^+, \ H_t^-, \ \forall\, t \in T\}$ satisfies the conditions of a Hamiltonian σ-algebra family.

(b)Consider a Hamiltonian σ-algebra family for $\{F_t, \ \forall\, t \in T\}$. Define,

$$G_t = H_t^+ \cap H_t^-, \ \forall\, t \in T.$$

Then $\{F_t, \ G_t, \ \forall\, t \in T\}$ is a σ-algebraic system.

Note that the necessary condition of (a) and the sufficiency condition of (b) are not equal. Condition (1) of a Hamiltonian σ-algebra family is too strong for this setting.

Proof. (a) From the definitions of H_t^+ and H_t^- and from the assumption of transitivity follows that $\{H_t^+, \ \forall\, t \in T\}$ is a decreasing family and that $\{H_t^-, \ \forall\, t \in T\}$ is an increasing family. From the definition of H_t^+ and of H_t^- follows that $F_{t-1}^- \subseteq H_t^-$ and $F_t^+ \subseteq H_t^+$. From the assumption of a σ-algebraic system follows that

$$\begin{aligned} & (F_t^+, F_{t-1}^- | G_t) \in \text{CI} \\ \Rightarrow \ & (F_t^+ \vee G_t, F_{t-1}^- \vee G_t | G_t) = (H_t^+, H_t^- | G_t) \in \text{CI}. \end{aligned}$$

From Proposition 7.3.21 follows that for all $t \in T$, $H_t^+ \cap H_t^- = \sigma(H_t^+|H_t^-) = \sigma(H_t^-|H_t^+)$. Hence $\{H_t^+, H_t^-, \ t \in T\}$ is a Hamiltonian family.

Note that the assumption of a σ-algebraic system implies by Theorem 7.4.7 that the condition of transitive projection-consistency holds. However, that condition is according to Lemma 7.4.6(b) weaker than transitivity which is needed for the Hamiltonian family.

(b) Fix a $t \in T$. From the definition of G_t and from Proposition 7.3.6 follows that,

$$\begin{aligned} G_t \ &= H_t^+ \cap H_t^- = \sigma(H_t^+|H_t^-), \\ & (H_t^+, H_t^- | \ G_t) = (H_t^+, H_t^- | \ \sigma(H_t^+|H_t^-)) \in \text{CI}, \\ & \Rightarrow (H_t^+ \vee G_t, H_t^- \vee G_t | \ G_t) \in \text{CI}. \end{aligned}$$

Because for all $t \in$T, $G_t = H_t^+ \cap H_t^-$, $H_t^+ \vee G_t = H_t^+$. Because by definition $\{H_t^+, \ H_t^-, \ \forall\, t \in T\}$ is a Hamiltonian family, $\{H_t^+ \vee G_t, \ \forall\, t \in T\}$ is a decreasing family. Correspondingly, $\{H_t^- \vee G_t, \ \forall\, t \in T\}$ is an increasing family. Thus transitivity holds. From Lemma 7.4.6 then follows that the condition of a current state σ-algebra family holds, hence $H_t^+ \vee G_t = H_t^+ \vee G_t^+$ and $H_t^- \vee G_t = H_t^- \vee G_t^-$. Then

$$(H_t^+ \vee G_t, H_t^- \vee G_t |\, G_t) \in \text{CI} \;\Rightarrow\; (H_t^+ \vee G_t^+, H_t^- \vee G_t^- \,|\, G_t) \in \text{CI}$$
$$\Rightarrow (F_t^+ \vee G_t^+, F_{t-1}^- \vee G_t^- \,|\, G_t) \in \text{CI, because } F_t^+ \subseteq H_t^+,\ F_{t-1}^- \subseteq H_t^-.$$

Hence $\{F_t, G_t,\ t \in T\}$ is a σ-algebraic system. □

7.4.4 Stochastic Realization as a Filter System

Within stochastic realization theory the forward and backward filter systems play an important role. These are described next.

Definition 7.4.10. *(forward) σ-algebraic filter system and the backward σ-algebraic filter system*. Consider a family of σ-algebras $\{F_t,\ \forall\, t \in T\}$.

Define the *σ-algebraic filter system* associated with the above family of σ-algebras by the condition that,

$$\{F_t,\ G_t,\ \forall\, t \in T\},\ \ G_t = \sigma(F_t^+ |\, F_{t-1}^-),\ \forall\, t \in T.$$

Define the *backward σ-algebraic filter system* associated with the above family of σ-algebras by the condition that,

$$\{F_t,\ G_t,\ \forall\, t \in T\},\ G_t = \sigma(F_t^- \,|\, F_{t+1}^+),\ \forall\, t \in T.$$

That the above filter systems are indeed σ-algebraic systems is proven next.

Theorem 7.4.11. Stochastic realization of the σ-algebraic filter system. *Consider a σ-algebra family $\{F_t,\ \forall\, t \in T\}$.*

(a)The σ-algebraic filter system defined in Def. 7.4.10 is a σ-algebraic system;

$$\Leftrightarrow (F_t^+ \vee G_t^+,\ F_{t-1}^- \vee G_t^- \,|\, G_t) \in \text{CI},\ \forall\, t \in T.$$

(b)The backward σ-algebraic filter system defined in Def. 7.4.10 is a σ-algebraic system.

Proof. (a) It is to be proven that $(F_t^+ \vee G_t^+,\ F_{t-1}^- \vee G_t^- \,|\, G_t) \in \text{CI}$. It follows from Proposition 19.8.2 that

$$(F_t^+, F_{t-1}^- |\, G_t) = (F_t^+, F_{t-1}^- | \sigma(F_t^+ | F_{t-1}^-)) \in \text{CI}.$$

It will be proven that transitivity of the family holds.

Consider $s,t \in T$ with $s < t$. Then $F_{s-1}^- \vee G_s = F_{s-1}^- \subseteq F_{t-1}^- = F_{t-1}^- \vee G_t$ hence $\{F_{t-1}^- \vee G_t,\ \forall\, t \in T\}$ is increasing. Next it is to be proven that for all $s,t \in T$ with $s < t$ that $F_s^+ \vee G_s \supseteq F_t^+ \vee G_t$. Note that, with $s \le t$,

$$
\begin{aligned}
& (F_s^+, F_{s-1}^- |\, G_s) \in \text{CI}, \text{ by } G_s = \sigma(F_s^+ | F_{s-1}^-) \text{ and Proposition 19.8.2}, \\
\Rightarrow\ & (F_t^+ \vee F_{\{s:t-1\}}, F_{s-1}^- |\, G_s) \in \text{CI} \\
\Rightarrow\ & (F_t^+ \vee F_{\{s:t-1\}}, F_{s-1}^- |\, F_{\{s:t-1\}} \vee G_s) \in \text{CI}, \text{ by Proposition 19.8.4.(e)}, \\
\Rightarrow\ & (F_t^+ \vee F_{\{s:t-1\}}, F_{s-1}^- \vee F_{\{s:t-1\}} |\, F_{\{s:t-1\}} \vee G_s) \in \text{CI}, \\
& \text{by Proposition 19.8.2.(a)}, \\
\Rightarrow\ & (F_t^+, F_{t-1}^- |\, F_{\{s:t-1\}} \vee G_s) \in \text{CI}, \text{ by restriction}, \\
\Rightarrow\ & G_t = \sigma(F_t^+ | F_{t-1}^-) = \sigma(F_t^+ | F_{t-1}^- \vee F_{\{s:t-1\}} \vee G_s) \\
\subseteq\ & F_{\{s:t-1\}} \vee G_s \subseteq F_s^+ \vee G_s, \\
G_t =\ & \sigma(F_t^+ | F_{t-1}^-) = \sigma(F_t^+ | F_{t-1}^- \vee F_{\{s:t-1\}} \vee G_s), \\
& \text{because } G_s = \sigma(F_s^+ | F_{s-1}^-) \subseteq F_{s-1}^- \subseteq F_{t-1}^-, \\
& \text{hence } F_{t-1}^- \vee F_{\{s:t-1\}} \vee G_s = F_{t-1}^-, \\
=\ & \sigma(F_t^+ | F_{\{s:t-1\}} \vee G_s) \subseteq F_{\{s:t-1\}} \vee G_s \subseteq F_s^+ \vee G_s.
\end{aligned}
$$

But $s < t$ implies that $F_t^+ \subseteq F_s^+ = \vee_{r \geq s} F_r$. Thus $F_t^+ \vee G_t \subseteq F_s^+ \vee G_s$ and $\{F_t^+ \vee G_t,\ t \in T\}$ is decreasing. Hence $\{F_t, G_t,\ \forall\, t \in T\}$ is transitive. From Lemma 7.4.6(b) follows that the σ-algebra family satisfies the transitive projection-consistency condition hence condition (b.2) holds. From Theorem 7.4.7 then follows that $\{F_t, G_t,\ \forall\, t \in T\}$ is a σ-algebraic system.
(b) The proof is analogous to that of (a). □

7.4.5 Minimality of a Stochastic Realization

Consider a σ-algebra family $\{F_t \subseteq F,\ \forall\, t \in T\}$ called the *output σ-algebra family*. The reader may recall the definition of a minimal stochastic realization of the above family in the form of a σ-algebraic system of Def. 7.4.3, defined by the conditions that,

(1) $(F_t^+ \vee G_t^+,\ F_{t-1}^- \vee G_t^- \,|\, G_t) \in \text{CI}_{\min}$,

(2) $\forall\, t \in T,\ G_t \subseteq F_f = F_t^+ \vee F_{t-1}^- = F_\infty$.

Needed is a characterization of when a σ-algebraic system is a minimal stochastic realization of its output σ-algebra family.

Proposition 7.4.12. Necessary and sufficient conditions for a minimal σ-algebraic stochastic realization. *Consider a σ-algebraic stochastic system specified by the filtrations $\{F_t,\ G_t \in \mathbb{F},\ \forall\, t \in T\}$.*

If this system is a minimal stochastic realization of its output σ-algebra family then the following conditions all hold: stochastic controllability, stochastic co-controllability, stochastic observability, stochastic co-observability, as stated in Def. 7.4.4.

A sufficient condition for minimality can be formulated analogously to those of Theorem 7.3.17.(b).

Proof. From Def. 7.4.3 and the assumption that the σ-algebraic system is a minimal realization follows that,

$$\begin{aligned}
&(F_t^+ \vee G_t^+,\, F_{t-1}^- \vee G_t^- \,|\, G_t) \in \mathrm{CI}_{\mathrm{min}},\; G_t \subseteq F_f = F_\infty = F_t^+ \vee F_{t-1}^-,\; \forall\, t \in T;\\
\Rightarrow\, &(F_t^+ \vee G_t,\, F_{t-1}^- \vee G_t |\, G_t) \in \mathrm{CI}_{\mathrm{min}},\; G_t \subseteq F_f = F_\infty,\; \forall\, t \in T;\\
\Rightarrow\, &(F_t^+,\, F_{t-1}^- |\, G_t) \in \mathrm{CI}_{\mathrm{min}},\; G_t \subseteq F_f = F_\infty,\; \forall\, t \in T.
\end{aligned}$$

Then the four claims follow from the above last condition of $\mathrm{CI}_{\mathrm{min}}$, from Theorem 7.3.8 and from Theorem 7.3.15.

Consider a stochastic realization in the form of a σ-algebraic system. That then the above four conditions and condition Theorem 7.3.17.(a.3) imply that the σ-algebraic system is a minimal stochastic realization follows from Theorem 7.3.17. □

7.5 Stochastic Realization of Output-Finite Stochastic Systems

The reader finds in this section a discussion of the stochastic realization problem for output-finite stochastic processes. The problem is not satisfactorily solved despite the fact that it was formulated around 1957.

Stochastic systems with a finite output set are frequently used in the research areas of signal processing and of control and system theory. The models are simple to formulate while a filter system has been derived and is simple to implement. The associated stochastic control system is frequently applied. The associated stochastic control systems are also used for approximating arbitrary stochastic systems. The motivation is therefore clear for the stochastic realization problem of an output-finite stochastic system.

The reader finds in Section 5.7 the definitions of an output-finite-state-finite stochastic system and of a output-finite-state-polytopic stochastic system.

Problem 7.5.1. *Weak stochastic realization of an output-finite-state-polytopic stochastic system.* Consider a family of finite-dimensional probability distributions of a finite-valued observed stochastic process.

The main research issues of stochastic realization of an output-finite-state-polytopic stochastic system are:

1. *Existence of a stochastic realization.* Does the family of finite-dimensional distributions of a finite-valued stochastic process admit a representation as an output-finite-state-polytopic stochastic system such that the output of this system has the same family of finite-dimensional distributions as those of the considered process? Call any such system a *weak stochastic realization* of the considered process. This existence subproblem has been satisfactorily solved.
2. *Minimality of a stochastic realization.* What is the concept of a minimal weak stochastic realization? How to characterize a minimal weak stochastic realization? This subproblem has not been satisfactorily solved.

3. *Classification and parametrization of minimal weak stochastic realizations.* How to parametrize the set of all minimal weak stochastic realizations? In general, the set of all minimal weak stochastic realizations contains two or more but in general many members. How to construct a canonical form for this subset of minimal weak stochastic realizations? The solution to this problem part is needed for the system identification of the considered set of stochastic systems. This subproblem can only be solved satisfactorily after Subproblem 2 is fully solved.

The set of output-finite-state-polytopic stochastic systems requires an explanation. Initially the above stochastic realization problem was formulated for output-finite-state-finite stochastic systems. But the filter system of an output-finite-state-finite stochastic system is an output-finite-state-polytopic stochastic system. Because the filter system is also a stochastic realization, the set of stochastic systems has to be enlarged to the set of output-finite-state-polytopic stochastic systems. This more general set contains of course the output-finite-state-finite stochastic systems as a subset. This issue is not well known, the literature is not clear on this issue.

The framework for the problem proposed by G. Picci is briefly mentioned. The existence is proven to be equivalent to the existence of a polytope of conditional probability measures which: (1) contains all conditional probability measures of finite output sequences and (2) which is invariant with respect to the system dynamics. This formulation has been used in Def. 5.7.1 to formulate the concept of an output-finite-state-polytopic stochastic system. The condition that the subset of conditional probability measures on the state set is actually a polytope which is invariant with respect to the state dynamics has been used in Theorem 5.7.11

The discussions in the literature on the minimality of a stochastic realization of an output-finite-state-finite stochastic system are not clear to the author. Needed for the state set of such a system is the set of tuples of the positive real numbers, the concept of a polyhedral cone, of an extremal polyhedral cone, and the concept of positive rank.

Both the number of states necessary to describe the state set and the number of vertices of the polytope, play a role in the concept of minimality. This research issue is partly clarified in Chapter 18 on positive matrices, in particular on the sections on factorizations of positive matrices and on the concept of an extremal cone in the positive orthant. A satisfactory solution to this issue is not clear yet to the author.

A problem issue with the minimality of a stochastic realization is illustrated by the following matrix,

$$H = \begin{pmatrix} 1\,0\,0\,1 \\ 1\,1\,0\,0 \\ 0\,1\,1\,0 \\ 0\,0\,1\,1 \end{pmatrix}, \ \mathrm{rank}(H) = 3 < 4 = \mathrm{pos} - \mathrm{rank}(H).$$

The reader finds the concept of positive rank of a matrix in Chapter 18 The example shows that the theory of realization of a finite-dimensional linear system is not applicable to a stochastic realization of an output-finite-state-polytopic stochastic system.

The strong stochastic realization problem for an output-finite-state-finite stochastic systems has not been investigated deeply yet. Example 7.3.7 of earlier in this chapter shows that there is also for this stochastic realization problem a research issue as to the concept of minimality of a stochastic realization.

7.6 Further Reading

History of stochastic realization. In the proceedings paper [38] where a review of the Kalman filter is discussed, R.E. Kalman is preparing research of stochastic realization theory, if one uses hindsight knowledge. The initial research on the weak Gaussian stochastic realization problem was carried out by P. Faurre working with R.E. Kalman as his research advisor at Stanford University. Reference [22] is a report in the French language with the results of that investigation. The complete story of Faurre's investigation is written in the French language book [24]. A brief summary of his theory in English is provided as a book chapter, [23].

The contribution of H. Akaike was to enrich stochastic realization theory with the canonical variable decomposition of a tuple of Gaussian random variables and to show how that decomposition could be used for system identification of Gaussian systems. Akaike's research was motivated by a discussion of H. Akaike with R.E. Kalman during a visit of the latter to Japan. See Akaike's publications in [1, 2, 3, 4, 5, 6, 7]. Akaike's research is the basis for stong Gaussian stochastic realization theory.

The focus on the strong Gaussian stochastic realization theory started with the paper of G. Picci, [48]. Conceptually the theory was much advanced by this paper with its focus on conditional independence and the relation of stochastic realization with realization theory of deterministic systems. The Hilbert space framework was investigated by G. Ruckebusch, [50, 51, 52]. A. Lindquist and G. Picci developed the Hilbert space framework which is fully described in their book [41]. The journal papers of those authors are not mentioned due to the availability of their book. The joint paper of the three researcher on conditional independence in the Hilbert space framework is of particular interest, [42].

Stochastic realization of a tuple of Gaussian random variables. These results were published in [57] as a special case of the strong stochastic realization problem of Gaussian processes. The staments and the proofs of this book are extensions of those of the quoted reference.

Stochastic realization of a tuple of σ-algebras. The author has formulated the problems of Section 7.3 and of Section 7.4 in the period 1975-1979. He was inspired by the paper of G. Picci, [48]. Later he learned of the approach of G. Ruckebush with his advisor M. Metivier, [50, 51, 52] and of A. Lindquist and G. Picci, for the approach to conditional independence of Hilbert spaces. A model for a stationary Gaussian stochastic process is a triple of Hilbert spaces for which the conditional-independence relation for Hilbert spaces can be defined. The reader is referred to the book of the latter authors, [41] and to the references quoted in that book.

The conditional-independence relation for sets was detailed by J.C. Willems for behaviors from the 1970's onward, [61, 62, 63].

The first publications of the author with a co-author on this subject date from 1979 and 1982, [56, 59].

At about that time there appeared reports and papers of others on the conditional independence relation and on the σ-algebraic realization problem, [18, 19, 44, 45, 46, 26]. The book [26, Section 5,4] published in 1990 treats a related problem as that of this section but with a different concept. The reader may also want to read Section 19.8 and the references quoted there.

Subsequently other papers of a co-author and the author were published, [57, 58, 60]. The conference paper [60] provides an overview of the results but without proofs. The proofs of that reference are now provided in the Section 7.3 and Section 7.4. The author regrets not having published earlier the results of those sections.

Proposition 7.3.6.(b) is due to H.P. McKean, [43].

Several of the results of Section 7.3 were first published in papers and a report of M. Mouchart and J.-M. Rolin, [44, 45, 46, 26]. However, there are also differences in notation, in concepts, and in results between those reference and this chapter. See also detailed information below.

There are concepts and results for the conditional independence relation which are related to the stochastic realization problem. These results are included the book [26] and in papers and reports, [46].

In particular the following results of Section 7.3 were published in the following references: Proposition 7.3.6.(c) is related to [46, 4.3]. Theorem 7.3.17 is related to the reference [46, Thm. 2.4]. Proposition 7.3.22.(b) and .(c) are related to [46, Prop. 3.1, Th. 3.2, Cor. 3.3].

Example 7.3.7 is a generalization of [59, Ex. 4.5]. Proposition 7.3.8 is adjusted from [59, Prop. 4.8]. Example 7.3.9 is adjusted from [59, Ex. 4.7]. Proposition 7.3.10 is adjusted from [59, Prop. 4.9]. Theorem 7.3.15 is adjusted from [59, Prop. 4.3]. Example 7.3.16 is adjusted from [59, Ex. 4.4]. The concept of a pair of Hamiltonian σ-algebras, Def. 7.3.23, is adjusted from the author's publication [60, Def. 3.4]. Theorem 7.3.24 is stated in [59, Thm. 4.11] without proof, it is related to [60, Prop. 3.5, Th. 3.6], and it is related to [46, Prop. 3.1, Thm. 3.2]. Theorem 7.4.11 is adjusted from [28, Lemma 4 and Theorem 1] and from [55].

Stochastic realization of a family of σ-algebras. Most results are generalizations of the paper [60] which paper does not include any proofs. The proofs are stated in this book.

For a stationary Gaussian process, the concept of an invariant measure and of an invariant measurable map are relevant, [41] and [26, Section 8.2].

There is a related paper by J.B. Gill, [28], but that papers differs from the approach of this chapter because (1) it is limited to the stochastic realization of a filtering realization; (2) an independence assumption is used; and (3) the existence of an independent complement is assumed.

Causality of two Gaussian processes has been investigated by statisticians and researchers of system theory. Well known is Granger causality, [30, 31, 32, 33]. See the publications of P.E. Caines, [12, 13, 14] Causality has been investigated also from the view of Bayesian statistics, see [25]. See also the publications of M. Józsa for the Hilbert space framework, [36]. Causality in terms of a system in a Hilbert space was formulated in [29].

The stochastic realization problem for a finite stochastic system was indirectly formulated in a paper published in 1957 of D. Blackwell and L. Koopmans, [11]. Koopmans was an economist or econometrician who pursued the use of mathematical models for economic phenomena. The problem statement is: Is it possible to uniquely determine the parameters of a state-finite Markov process or of an output-finite-state-finite stochastic system? In control and system theory this problem would now be called the structural identifiability problem of an output-finite-state-finite stochastic system. A second paper by E.J. Gilbert, advised by D. Blackwell, is [27].

There followed in the period 1957-1970 papers in which the existence problem was solved. The existence of a stochastic realization was solved by A. Heller, [34, 35] in an abstract setting. Other references include [9, 16, 17, 21]. A generalization of the existence result was formulated by M. Arbib, [8]. The theory as of 1971 is described in the book of A. Paz, [47].

A paper of G. Picci published in 1978 describes the approach to the existence problem analogously to that of realization of a finite-dimensional linear system, [49]. The author strongly recommends this paper. The set of states is isomorphic to a polytope within the probability simplex rather than an arbitrary set of probability measures. The condition of a polytope, hence it is a polyhedral set with a finite number of vertices, corresponds to the condition of a finite-rank Hankel matrix in realization theory of a linear system, see Section 21.8.

Minimality of a realization of a finite-valued output process. For the output-finite-state-polytopic stochastic realization problem the characterization of the minimal state set is an open problem though there is progress on the stochastic observability and stochastic controllability. Early references on the minimality of such a stochastic realization include [9, 16].

Recent books on output-finite-state-finite stochastic systems include [15, 39].

Abstract realization theory. Realization theory for systems in sets, [20], [53, Section III.7]. For algebraic systems see the book [54].

References

1. H. Akaike. On a decision procedure for system identification. In *Proceedings IFAC Symposium System Engineering Approach to Computer Control*, page Paper 30.1. IFAC, Japan, 1970. 275
2. H. Akaike. *Use of an information theoretic quantity for statistical model identification*, pages 249–250. Western Periodicals Co., 1972. 275
3. H. Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csaki, editors, *Proceedings 2nd International Symposium Information Theory*, pages 267–281. Akademia Kiado, Budapest, 1973. 275
4. H. Akaike. Stochastic theory of minimal realization. *IEEE Trans. Automatic Control*, 19:667–674, 1974. 217, 275

5. H. Akaike. Markovian representation of stochastic processes by canonical variables. *SIAM J. Control*, 13:162–173, 1975. 217, 275
6. H. Akaike. Canonical correlation analysis of time series and the use of an information criterion. In R.K. Mehra and D.G. Lainiotis, editors, *System identification - Advances and case studies*, pages 27–96. Academic Press, New York, 1976. 275
7. H. Akaike and T. Nakagawa. *Statistical analysis and control of dynamic systems*. Kluwer Academic Publishers, Dordrecht, 1988. 275
8. M. Arbib. Realization of stochastic systems. *Ann. Math. Statist.*, 38:927–933, 1967. 277
9. Glenn C. Bacon. Minimal-state stochastic finite state systems. *IEEE Trans. CT*, 11:307–308, 1964. 277
10. R.R. Bahadur and E.L. Lehmann. Two comments on 'Sufficiency and statistical decision functions'. *Ann. Math. Statist.*, 26:139–142, 1955. 31
11. D. Blackwell and L. Koopmans. On the identifiability problem for functions of finite Markov chains. *Ann. Math. Statist.*, 28:1011–1015, 1957. 175, 277
12. P.E. Caines. *Linear stochastic systems*. John Wiley & Sons, New York, 1988. 120, 276, 302, 310, 575
13. P.E. Caines and C.W. Chan. Feedback between stationary stochastic processes. *IEEE Trans. Automatic Control*, 20:498–508, 1975. 276
14. P.E. Caines and S.P. Chan. Recursiveness, causality and feedback. *IEEE Trans. Automatic Control*, 24:113–115, 1979. 276
15. O. Cappé, E. Moulines, and T. Rydén. *Inference in hidden Markov models*. Springer, Berlin, 2005. 169, 277, 353
16. J.W. Carlyle. Reduced forms for stochastic sequential machines. *J. Math. Anal. Appl.*, 7:167–175, 1963. 277
17. J.W. Carlyle. State-calculable stochastic sequential machines, equivalences, and events. In *Proc. IEEE Sixth Annual Symposium on Switcing Circuit Theory and Logical Design*, New York, 1965. IEEE Press. 277
18. A.P. Dawid. Conditional independence for statistical operations. *Ann. Math. Statist.*, 8:598–617, 1980. 49, 276, 742
19. R. Döhler. On the conditional independence of random events. *Theory Probab. Appl.*, 25:628–634, 1980. 276, 742
20. S. Eilenberg. *Automata, languages, and machines (Volumes A and B)*. Academic Press, New York, 1974, 1976. 277, 808
21. Shimon Even. Commment on the minimization of stochastic machines. *IEEE Trans. EC*, 14:634–637, 1965. 277
22. P. Faurre. Réalization markoviennes de processus stationnaires. Rapport de recherche 13, IRIA, Rocquencourt, 1973. 175, 180, 217, 275
23. P. Faurre. Stochastic realization algorithms. In R.K. Mehra and D.G. Lainiotis, editors, *System Identification - Advances and Case Studies*, pages 1–25. Academic Press, New York, 1976. 180, 217, 275
24. P. Faurre, M. Clerget, and F. Germain. *Opérateurs rationnels positifs*. Dunod, Paris, 1979. 175, 180, 217, 275, 292, 310, 850, 865, 867, 877, 885
25. J.P. Florens and M. Mouchart. A note on noncausality. *Econometrica*, 50:583–591, 1982. 277
26. J.P. Florens, M. Mouchart, and J.M. Rolin. *Elements of Bayesian statistics*. Routeldge, Baton Rouge, FL, 1990. 253, 276, 723, 741, 742
27. E.J. Gilbert. On the identifiability problem for functions of finite Markov chains. *Ann. Math. Statist.*, 30:688–697, 1959. 277
28. J.B. Gill. Markovian extensions and reductions of a family of σ-algebras. *Int. J. Math.*, 7:523–528, 1984. 276
29. J.B. Gill and L. Petrovic. Causality and stochastic dynamic systems. *SIAM J. Appl. Math.*, 47:1361–1366, 1987. 277
30. C.W.J. Granger. Economic processes involving feedback. *Info. Control*, 6:28–48, 1963. 276
31. C.W.J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424–438, 1969. 276

32. C.W.J. Granger. Testing for causality - a personal viewpoint. *J. Econ. Dyn. Control*, 2:329–352, 1980. 276
33. C.W.J. Granger. Some recent development in a concept of causality. *J. Econometrics*, 39:199–211, 1988. 276
34. A. Heller. On stochastic processes derived from Markov chains. *Ann.Math. Statist.*, 36:1286–1291, 1965. 277
35. A. Heller. Probabilistic automata and stochastic transformations. *Math. Syst. Theory*, 1:197–208, 1967. 277
36. Mónika Józsa. *Relationship between Granger-causality and network graphs of state-space represenations*. PhD thesis, University of Groningen, Groningen, 25 February 2019. 277
37. R.E. Kalman. Mathematical description of linear dynamical systems. *SIAM J. Control*, 1:152–192, 1963. 174, 761, 807, 808, 809
38. R.E. Kalman. New methods in Wiener filtering theory. In J.L. Bogdanoff and F. Kozin, editors, *Proceedings 1st Symposium Engineering Applications of Random Function Theory and Probability*, pages 270–388, New York, 1963. Wiley. 175, 275, 310
39. V. Krishnamurthy. *Partially observed Markov decision processes*. Cambridge University Press, Cambridge, 2016. 169, 277, 353, 575
40. A. Lindquist and G. Picci. On the stochastic realization problem. *SIAM J. Control & Opt.*, 17:365–389, 1979. 217, 253
41. A. Lindquist and G. Picci. *Stochastic realization of Gaussian processes – A geometric approach to modeling, estimation and identification*. Springer, Heidelberg, 2015. 120, 121, 175, 176, 180, 217, 246, 253, 254, 275, 276
42. A. Lindquist, G. Picci, and G. Ruckebusch. On minimal splitting subspaces and Markovian representations. *Math. Systems Th.*, 12:271–279, 1979. 120, 175, 217, 275
43. H.P. McKean Jr. Brownian motion with a several dimensional time. *Theory Probab. Appl.*, 8:335–354, 1963. 276, 741
44. M. Mouchart and J.-M. Rolin. A note on conditional independence (with statistical applications). Report 129, Institut de Mathématique Pure et Appliquée, Université Catholique de Louvain, Louvain-la-Neuve, 1979. 49, 276, 723, 742
45. M. Mouchart and J.-M. Rolin. A note on conditional independence with statistical applications. *Statistica*, 44:557–584, 1984. 49, 276, 723, 742
46. M. Mouchart and J.-M. Rolin. On the σ-algebraic realization problem. Report 8604, Center for Operations Research and Econometrics, Université de Louvain, Louvain-la-Neuve, 1986. 253, 276, 723, 742
47. A. Paz. *Introduction to probabilistic automata*. Academic Press, New York, 1971. 169, 277
48. G. Picci. Stochastic realization of gaussian processes. *Proc. IEEE*, 64:112–122, 1976. 120, 175, 217, 275
49. G. Picci. On the internal structure of finite-state stochastic processes. In *Proc. of a U.S.-Italy Seminar*, volume 162 of *Lecture Notes in Economics and Mathematical Systems*, pages 288–304. Springer-Verlag, Berlin, 1978. 120, 150, 277
50. G. Ruckebusch. Représentations markoviennes de processus gaussiens stationnaires. *C. R. Acad. Sc. Paris, Série A*, 282:649–651, 1976. 120, 175, 217, 248, 253, 275
51. G. Ruckebusch. Représentations markoviennes de processus gaussiens stationnaires et applications statistiques. Rapport interne 18, Ecole Polytechnique, Centre de Mathématiques Appliquées, 1977. 120, 175, 248, 253, 275, 291, 310
52. G. Ruckebusch. Théorie géométrique de la représentation markovienne. *Ann. Inst. Henri Poincaré*, 16:225–297, 1980. 120, 175, 217, 248, 253, 275
53. E.D. Sontag. Realization theory of discrete-time nonlinear systems: I. The bounded case. *IEEE Trans. Circuits & Systems*, 26:342–356, 1979. 277
54. E.D. Sontag. *Mathematical control theory: Deterministic finite dimensional systems (2nd. Ed.)*. Number 6 in Graduate Text in Applied Mathematics. Springer, New York, 1998. 217, 277, 808
55. W.J. Stronegger. When is a sequence of sufficient σ-algebras a filter system. *Systems & Control Lett.*, 16:473–477, 1991. 276

56. C. van Putten and J.H. van Schuppen. On stochastic dynamical systems. In *Proceedings Fourth International Symposium on the Mathematical Theory of Networks and Systems (MTNS79)*, pages Volume 3, 350–355, North Hollywood, CA, 1979. Western Periodical. 120, 276
57. C. van Putten and J.H. van Schuppen. The weak and strong Gaussian probabilistic realization problem. *J. Multivariate Anal.*, 13:118–137, 1983. 49, 237, 275, 276
58. C. van Putten and J.H. van Schuppen. Invariance properties of the conditional independence relation. *Ann. Probab.*, 13:934–945, 1985. 49, 276, 723, 741, 742
59. J.H. van Schuppen. The strong finite stochastic realization problem - Preliminary results. In A. Bensoussan and J.L. Lions, editors, *Analysis and optimization of systems*, volume 44 of *Lecture Notes in Control and Information Sciences*, pages 179–190, Berlin, 1982. Springer-Verlag. 120, 276, 742
60. J.H. van Schuppen. Stochastic realization of σ-algebras. In *Proc. 14th Int. Symposium MTNS (published on CD-ROM only)*, Perpignan, 2000. Université de Perpignan. 276
61. J.C. Willems. From time series to linear systems - Part I. Finite dimensional linear time invariant systems. *Automatica J. IFAC*, 22:561–580, 1986. 174, 276
62. J.C. Willems. From time series to linear systems - Part II. Exact modelling. *Automatica J. IFAC*, 22:675–694, 1986. 174, 276
63. J.C. Willems. From time series to linear systems - Part III. Approximate modelling. *Automatica J. IFAC*, 23:87–115, 1987. 174, 276

Chapter 8
Filtering of Gaussian Systems

Abstract The filter problem is to derive an expression for the conditional distribution of the state of a stochastic system conditioned on the past outputs of the considered system and a recursion of the parameters of that conditional distribution. In this chapter the filter problem for a Gaussian stochastic system is solved. The Kalman filter is derived, including the time-varying, the time-invariant Kalman filter, and the conditional Kalman filter. The relation of the Kalman filter with stochastic realization is discussed. The Kalman filter is used in a very large number of research areas including signal processing, information theory, communication theory, weather prediction, hydrology, etc.

Key words: Kalman filter. Filter algebraic Riccati equation.

The filter problem is to derive an expression for the conditional distribution of the state of a stochastic system conditioned on the past outputs of the considered system and a recursion of the parameters of that conditional distribution. In this chapter the filter problem for a Gaussian stochastic system is solved. The Kalman filter is derived, including the time-varying, the time-invariant Kalman filter, and the conditional Kalman filter.

It will be argued that the Kalman filter, rewritten as a filter system, is a stochastic realization of the output process of which the state is measurable with respect to the past outputs and which stochastic realization has a finite-dimensional state set.

8.1 Problems of Filtering, Prediction, Smoothing, and Interpolation

In engineering and in econometrics there are problems of prediction or forecasting. The prediction problem is to predict the value of a stochastic process based on the observations of a related observed process. The expression *prediction* is primarily used in engineering while the expression *forecasting* is used in econometrics.

In this book the term *filtering* is used for the research area which encompasses state estimation, prediction, tracking, smoothing, interpolation, and the related signal processing issues. The *filter problem* is in general the problem to estimate the state of a stochastic system from past outputs, though it is generalized to encompass the Wiener filter problem.

Examples of problems of the research area of filtering are:

- Radar tracking, see [35].
- Prediction of electric power demand, see [5, 12, 41].
- Channel equalization, see [9, 22].
- Prediction of water levels near coastal areas, see [11].
- Air pollution prediction, see [2, 39].
- Adaptive prediction of traffic flow at the boundary of a road network, [43].

8.2 Problem of Filtering

The first theoretical approaches to the filter problem were formulated by N. Wiener in the U.S.A. and, apparently independently, by A. Kolmogorov in Russia, [23]. The report of Wiener on the subject was only published in 1949 due to the report being held confidential during the second world war, 1939–1945.

Problem 8.2.1. The *Wiener filter problem*. Consider two stochastic processes on a probability space (Ω, F, P),

$$y : \Omega \times T \to Y = \mathbb{R}^{m_y}, \ z : \Omega \times T \to Z = \mathbb{R}^{m_z},$$

where y represents the *observed process* and z the *to-be-estimated process*. Determine for all $t \in T$, the *linear least-squares estimate* of $z(t)$ given the past observations of y by,

$$\begin{aligned} LLS[z(t)|F^y_{t-1}] &= \sum_{s=0}^{t-1} K(t,s)y(s), \ K : T \times T \to \mathbb{R}^{m_z \times m_y}, \\ \hat{z}(t) &= LLS[z(t)|F^y_{t-1}], \ \text{where, } F^y_t = \sigma(\{y(s), \forall s \leq t\}), \\ LLS[z(t)|F^y_{t-1}] &= \text{argmin}_{\bar{z}(t) \in L_2(F^y_{t-1})} \, E\|z(t) - \bar{z}(t)\|^2. \end{aligned}$$

Thus the linear least-squares estimate is a linear function of the past observations which minimizes the estimation error over all such linear functions.

The Wiener filter problem covers as special cases the *prediction problem*, where $z(t) = y(t+s)$ for $s \in \mathbb{Z}_+$, and the case where the process z is different from but dependent on the observed process y.

The contribution of R.E. Kalman to filter theory is to restrict the problem formulation to the determination of the state of a stochastic system conditioned on the past of the observed process and this for Gaussian processes. This is described next.

Problem 8.2.2. *The Kalman filter problem for a Gaussian system.* Consider a time-varying Gaussian system with representation,

$$\begin{aligned} x(t+1) &= A(t)x(t)+M(t)v(t), \;\; x(0)=x_0 \in G(m_0,Q_0), \\ y(t) &= C(t)x(t)+N(t)v(t), \;\; v(t) \in G(0,I). \end{aligned}$$

Determine the following conditional distribution or the conditional characteristic function for all times,

$$\begin{aligned} &\text{cpdf}(.;x(t+1)|\, F_t^y) \; \forall\, t \in T; \text{ or, equivalently,} \\ &E[\exp(iw_x^T x(t+1))|F_t^y], \; \forall\, w_x \in \mathbb{R}^{n_x}, \; \forall\, t \in T. \end{aligned}$$

Determine how to calculate recursively in time the parameters of the requested conditional distribution.

The reader is alerted of the fact that the above formulated filter problem differs from the problem in which one determines,

$$\begin{aligned} &\text{cpdf}(.;x(t)|\, F_t^y) \; \forall\, t \in T; \text{ or, equivalently,} \\ &E[\exp(iw_x^T x(t))|F_t^y], \; \forall\, w_x \in \mathbb{R}^{n_x}, \; \forall\, t \in T. \end{aligned}$$

In the literature this second case is occasionally discussed, the resulting formulas are different though closely related to those of the case formulated.

Kalman [17] has formulated the problem in terms of least-squares estimation which is generalized above to the determination of the conditional distribution.

In the Kalman filter problem attention is first restricted to a Gaussian system in state-space form and secondly to determining the conditional distribution of the state conditioned on the past outputs. This problem is solvable and the solution admits recursions for the parameters of the conditional distribution as is shown in the next section. This formulation is a major advance with respect to the Wiener filter problem.

In addition, the Kalman filter problem also provides the solution to a special case of the Wiener filter problem as will be discussed in Section 8.4. The Kalman filter problem covers time-varying Gaussian systems while Wiener filter problem covers only stationary square-integrable processes, not necessarily Gaussian. However, the intersection of the Wiener filter problem and of the Kalman filter problem is neither equal to the Wiener filter problem nor equal to the Kalman filter problem.

8.3 Time-Varying Kalman Filter

The following result is an extension of the result of R.E.Kalman, [17], for the Kalman filter problem from linear least-squares estimation to determination of the conditional distribution of a time-varying Gaussian stochastic system.

Assumption 8.3.1 *Consider a time-varying Gaussian system of Problem 8.2.2. Define the assumptions:*

1. $n_y \leq n_v$ *and, for all* $t \in T$, $\text{rank}(N(t)) = n_y$. *Consequently,* $N(t)N(t)^T \in \mathbb{R}_{spds}^{n_y \times n_y}$ *hence* $N(t)N(t)^T \succ 0$.
2. *the system is supportable in regard to the relation of the noise process v to the state process x;*
3. *the system is stochastically observable in regard to the relation from the state to the observed output y; and*
4. *the system is stochastically co-observable.*

Theorem 8.3.2. *Consider the filter problem for a time-varying Gaussian system, Problem 8.2.2. Let the conditions of Assumption 8.3.1 all hold.*

(a)The conditional distribution of $x(t)$ conditioned on F_{t-1}^y, for any $t \in T$, is Gaussian and specified by the characteristic function,

$$E[\exp(iw_x^T x(t))|F_{t-1}^y] = \exp(iw_x^T \hat{x}(t) - \frac{1}{2} w_x^T Q_f(t) w_x), \ \forall w_x \in \mathbb{R}^{n_x}. \tag{8.1}$$

Call then $\hat{x} : \Omega \times T \to \mathbb{R}^n$ *the* conditional mean process *and* $Q_f : T \to \mathbb{R}^{n \times n}$ *the* error-variance process.

(b)The parameters of the characteristic function (8.1) can be recursively calculated according to the recursions,

$$\hat{x}(t+1) = A(t)\hat{x}(t) + K(t)[y(t) - C(t)\hat{x}(t)], \ \hat{x}(0) = E[x_0] = m_0, \tag{8.2}$$

$$\begin{aligned} Q_f(t+1) &= A(t)Q_f(t)A(t)^T + M(t)M(t)^T + \\ &\quad -[A(t)Q_f(t)C(t)^T + M(t)N(t)^T] \times \\ &\quad \times [C(t)Q_f(t)C(t)^T + N(t)N(t)^T]^{-1} \times \\ &\quad \times [A(t)Q_f(t)C(t)^T + M(t)N(t)^T]^T, \qquad (8.3) \\ &\quad Q_f(0) = E[(x_0 - E[x_0])(x_0 - E[x_0])^T] = Q_{x_0}, \ Q_f(t) \in \mathbb{R}_{pds}^{n_x \times n_x}, \end{aligned}$$

$$\begin{aligned} K(t, Q_f(t)) &= [A(t)Q_f(t)C(t)^T + M(t)N(t)^T] \times \\ &\quad \times [C(t)Q_f(t)C(t)^T + N(t)N(t)^T]^{-1}. \end{aligned} \tag{8.4}$$

The recursion of equation (8.2) is called the Kalman filter, *that of equation (8.3) is called the* Filter Riccati Recursion, *and the matrix of equation (8.4) is called the* Kalman gain matrix.

(c)From (a) and (b) follows that for all $t \in T$ a.s.,

$$\begin{aligned} \hat{x}(t) &= E[x(t)|F_{t-1}^y], \\ Q_f(t) &= E[(x(t) - \hat{x}(t))(x(t) - \hat{x}(t))^T|F_{t-1}^y] = E[(x(t) - \hat{x}(t))(x(t) - \hat{x}(t))^T]. \end{aligned}$$

Note that the expression of $Q_f(t)$ does not depend explicitly on the values of the output process. This is a particular property of the conditional expectation of Gaussian random variables.

(d)Define the innovation process *as the stochastic process,*

$$\bar{v} : \Omega \times T \to \mathbb{R}^{n_y}, \ \bar{v}(t) = y(t) - C(t)\hat{x}(t).$$

Then $\bar{v}$ is a Gaussian white noise process, hence a squence of independent random variables, such that for all $t \in T$,

$$\bar{v}(t) \in G(0, Q_{\bar{v}}(t)), \tag{8.5}$$

$$Q_{\bar{v}}(t) = C(t)Q_f(t)C(t)^T + N(t)N(t)^T \in \mathbb{R}^{n_y \times n_y}_{pds}. \tag{8.6}$$

(e)For all $t \in T$, $F^y_t = F^{\bar{v}}_t$.
(f) Define the filter-error process *or just the* error process *as,*

$$\begin{aligned} e(t) &= x(t) - \hat{x}(t),\ e : \Omega \times T \to \mathbb{R}^{n_x};\ \textit{then}, \\ e(t+1) &= (A(t) - K(t)C(t))e(t) + (M(t) - K(t)N(t))v(t),\ e(0) = x_0 - m_0, \\ &E[\exp(iw^T e(t))|F^y_{t-1}] \\ &= \exp\left(-\frac{1}{2}w^T Q_f(t))w\right),\ \forall\, w \in \mathbb{R}^{n_x},\ e(t) \in G(0, Q_f(t)). \end{aligned}$$

(g)The variances of the system, the Kalman filter, and the solution of the Riccati recursion satisfy,

$$\begin{aligned} Q_x(t) &= Q_{\hat{x}}(t) + Q_f(t),\ \ \forall\, t \in T; \\ Q_{\hat{x}}(t) &\leq Q_x(t),\ \forall\, t \in T; \\ Q_{\hat{x}}(t+1) &= AQ_{\hat{x}}(t)A^T + K(t)Q_{\bar{v}}(t)K(t)^T,\ \forall\, t \in T. \end{aligned}$$

Proof. Below use is made of the definition and theorems for conditional Gaussian random variables, see Section 19.7.
(1) Define the σ-algebra families,

$$\begin{aligned} &\{F_t, t \in \{-1\} \cup T\},\ F_{-1} = F^{x_0},\ F_t = F^{x_0} \vee F^v_t, \forall\, t \in T, \\ &\{F^y_t, t \in \{-1\} \cup T\},\ F^y_{-1} = \{\emptyset, \Omega\},\ F^y_t = \sigma(\{y(s), \forall s \leq t\}),\ \forall\, t \in T. \end{aligned}$$

From the equations,

$$\begin{aligned} x(t+1) &= A(t)x(t) + M(t)v(t), x(0) = x_0, \\ y(t) &= C(t)x(t) + N(t)v(t), \end{aligned}$$

it follows by induction that for all $t \in T$, $x(t+1)$ and $y(t)$ are F_t measurable. For example, x_0 is F_{-1} measurable by definition of that σ-algebra. Suppose by induction that $x(t)$ is F_{t-1} measurable. Then it follows from the recursion for the state process displayed above and from the definition of the filtration $\{F_t, t \in T\}$, that $x(t+1)$ and $y(t)$ are F_t measurable.

(2) For any $t \in T$,

$$\begin{aligned} &E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix}\right)|F_{t-1}\right] \\ &= \exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} A(t)x(t) \\ C(t)x(t) \end{pmatrix}\right) E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} M(t) \\ N(t) \end{pmatrix} v(t)\right)|F_{t-1}\right] \\ &= \exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{bmatrix} A(t)x(t) \\ C(t)x(t) \end{bmatrix} - \frac{1}{2}\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{bmatrix} M(t)M(t)^T & M(t)N(t)^T \\ N(t)M(t)^T & N(t)N(t)^T \end{bmatrix} \begin{pmatrix} w_x \\ w_y \end{pmatrix}\right), \end{aligned}$$

where the first equality follows because $x(t)$ is measureable with respect to the indicated σ-algebras and the second equality follows because v is a Gaussian white noise process; hence $(x(t+1), y(t))$ are conditionally Gaussian given F_{t-1}.
(3) By induction in $t \in T$ the two equations (8.7) and (8.8) will be proven to hold,

$$
\begin{aligned}
&(\forall\, s = 0,1,\ldots,t_1-1) \\
&E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x(s+1) \\ y(s) \end{pmatrix}\right)|F^y_{s-1}\right] \qquad (8.7)\\
&= \exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} A(s)\hat{x}(s) \\ C(s)\hat{x}(s) \end{pmatrix} - \frac{1}{2}\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T H(s) \begin{pmatrix} w_x \\ w_y \end{pmatrix}\right), \\
&E[\exp(iw^T x(s+1))|F^y_s] \\
&= \exp(iw^T\hat{x}(s+1) - \frac{1}{2}w^T Q_f(s+1)w), \qquad (8.8)
\end{aligned}
$$

$$
\begin{aligned}
H(t) &= \begin{pmatrix} H_{11}(t) & H_{12}(t) \\ H_{12}(t)^T & H_{22}(t) \end{pmatrix} \in \mathbb{R}^{(n_x+n_y)\times(n_x+n_y)}_{pds}, \qquad (8.9)\\
H_{11}(t) &= A(t)Q_f(t)A(t)^T + M(t)M(t)^T, \\
H_{12}(t) &= A(t)Q_f(t)C(t)^T + M(t)N(t)^T, \\
H_{22}(t) &= C(t)Q_f(t)C(t)^T + N(t)N(t)^T.
\end{aligned}
$$

For $t = 0$,

$$
\begin{aligned}
x(1) &= A(0)x_0 + M(0)v(0), \\
y(0) &= C(0)x_0 + N(0)v(0).
\end{aligned}
$$

By the assumptions on a Gaussian system, $(x_0, v(0))$ are independent random variables with $x_0 \in G(m_0, Q_0)$ and $v(0) \in G(0, I)$. Thus $(x_0, v(0))$ are jointly Gaussian random variables. Then,

$$
\begin{aligned}
&E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x(1) \\ y(0) \end{pmatrix}\right)|F^y_{-1}\right] \\
&= E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x(1) \\ y(0) \end{pmatrix}\right)\right], \text{ because } F^y_{-1} = \{\emptyset, \Omega\}, \\
&= E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} A(0) & M(0) \\ C(0) & N(0) \end{pmatrix}\begin{pmatrix} x(0) \\ v(0) \end{pmatrix}\right)\right], \\
&= \exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} A(0)\hat{x}_0 \\ C(0)\hat{x}_0 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T H(0)\begin{pmatrix} w_x \\ w_y \end{pmatrix}\right) \\
&\quad \text{because of Theorem 2.8.3 and} \\
&\quad x_0 \in G(m_{x_0}, Q_{x_0}),\ v(0) \in G(0, I),\ \hat{x}_0 = m_{x_0},\ Q_f(0) = Q_{x_0}.
\end{aligned}
$$

From Theorem 2.8.3 and equation (8.7) follows equation (8.8) for $t = 0$.

Suppose that the equations (8.7,8.8) hold for $s = 0,1,\ldots,t \in T$. They will then be proven to hold for $s = t+1$. Note that for all $(w_x, w_y) \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_y}$,

$$
\begin{aligned}
& E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix}\right)|F^y_{t-1}\right] \\
& = E\left[E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix}\right)|F_{t-1}\right]|F^y_{t-1}\right], \text{ because } F^y_{t-1} \subseteq F_{t-1}, \\
& = E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} A(t) \\ C(t) \end{pmatrix} x(t)\right)|F^y_{t-1}\right] \\
& \quad \times \exp\left(-\frac{1}{2}\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} M(t)M(t)^T & M(t)N(t)^T \\ N(t)M(t)^T & N(t)N(t)^T \end{pmatrix}\begin{pmatrix} w_x \\ w_y \end{pmatrix}\right), \\
& \quad \text{because of Step (2) of the proof,} \\
& = \exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} A(t) \\ C(t) \end{pmatrix}\hat{x}(t) - \frac{1}{2}\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T H(t)\begin{pmatrix} w_x \\ w_y \end{pmatrix}\right), \\
& \quad \text{because of (8.8) of the induction step for } s = t, \text{ and,}
\end{aligned}
$$

$$
H(t) = \begin{pmatrix} A(t) \\ C(t) \end{pmatrix} Q_f(t) \begin{pmatrix} A(t) \\ C(t) \end{pmatrix}^T + \begin{pmatrix} M(t) \\ N(t) \end{pmatrix}\begin{pmatrix} M(t) \\ N(t) \end{pmatrix}^T.
$$

Use is made of Proposition 19.7.4.

$$
\begin{aligned}
& E[\exp(iw_x^T x(t+1))|F_t^y] = E[\exp(iw_x^T x(t+1))|F^{y(t)} \vee F^y_{t-1}] \\
& = \exp(iw_x^T[A\hat{x}(t) + H_{12}(t)H_{22}^{-1}(y(t) - C\hat{x}(t))]) \times \\
& \quad \times \exp(-\frac{1}{2}w_x^T[H_{11}(t) - H_{12}(t)H_{22}^{-1}(t)H_{12}^T(t)]w_x), \\
& \quad \text{by the induction step (8.8) for } s = t \in T, \text{ and by Proposition 19.7.4,} \\
& = \exp\left(iw_x^T\hat{x}(t+1) - \frac{1}{2}w_x^T Q_f(t+1)w_x\right), \\
& \quad \text{with equation (8.8) and the equations (8.2,8.3).}
\end{aligned}
$$

By the principle of induction the equations (8.7,8.8) then hold for all $t \in T$. Then (a) and (b) are proven.

(c) From (a) follows that the conditional distribution is Gaussian and from this and Proposition 19.7.4, follows (c).

(d) From Step (3) and from equation (8.7) with $w_x = 0$ follows that for all $w_y \in \mathbb{R}^{n_y}$,

$$
\begin{aligned}
E\left[\exp(iw_y^T y(t)|F^y_{t-1}\right] &= \exp(iw_y^T C(t)\hat{x}(t) - \frac{1}{2}w_y^T Q_{\bar{v}}(t)w_y), \\
E\left[\exp(iw_y^T \bar{v}^T(t))|F^y_{t-1}\right] &= \exp(-\frac{1}{2}w_y^T Q_{\bar{v}}(t)w_y),\ Q_{\bar{v}}(t) = H_{22}(t).
\end{aligned}
$$

Because $Q_{\bar{v}}(t)$ is a deterministic function it follows from Theorem 2.8.2.(f) that $F^{\bar{v}(t)}$ is independent of F^y_{t-1}. By definition of $\bar{v}(t)$, $F^{\bar{v}}_{t-1} \subseteq F^y_{t-1}$. Hence $\bar{v}$ is an independent sequence.

(e) By induction it will be proven that $F_t^{\bar{v}} = F_t^y$ for all $t \in T$. Because $\bar{v}(0) = y(0) -$

$C(0)\hat{x}_0$ and $\hat{x}_0 = m_0 \in \mathbb{R}^{n_x}$, $F_0^{\bar{v}} = F_0^y$. Suppose that for all $s,t \in T$, $s \leq t$, $F_s^{\bar{v}} = F_s^y$. Then $\bar{v}(t) = y(t) - C(t)\hat{x}(t)$ is F_t^y measurable, hence $F_t^{\bar{v}} \subseteq F_t^y$. Because $y(t) = \bar{v}(t) + C(t)\hat{x}(t)$, $\hat{x}(t)$ is F_{t-1}^y measurable, and, by the induction step, $F_{t-1}^y = F_{t-1}^{\bar{v}}$, $y(t)$ is $F_{t-1}^{\bar{v}}$ measurable, hence $F_t^y \subseteq F_t^{\bar{v}}$.
(f) From the problem formulation and the statement of (b) follows that,

$$\begin{aligned}
e(t+1) &= x(t+1) - \hat{x}(t+1) \\
&= A(t)x(t) + M(t)v(t) - A(t)\hat{x}(t) - K(t)[y(t) - C(t)\hat{x}(t)] \\
&= Ae(t) + M(t)v(t) - K(t)[C(t)e(t) + N(t)v(t)] \\
&= (A(t) - K(t)C(t))e(t) + (M(t) - K(t)N(t))v(t).
\end{aligned}$$

From equation (8.2) follows by multiplication of the term with $\hat{x}(t)$ that,

$$E[\exp(iw_x^T e(t))|F_{t-1}^y] = \exp(-\frac{1}{2}w_x^T Q_f(t) w_x), \; \forall \; w_x \in \mathbb{R}^{n_x}.$$

(g) Note that for all $t \in T$,

$$\begin{aligned}
E[x(t) - \hat{x}(t)] &= E[E[x(t)|F_{t-1}^y] - \hat{x}(t)] = 0, \\
Q_x(t) &= E[(x(t) - E[x(t)])(x(t) - E[x(t)])^T] \\
&= E[x(t)x(t)^T] - E[x(t)]E[x(t)]^T, \\
Q_{\hat{x}}(t) &= E[(\hat{x}(t) - E[\hat{x}(t)])(\hat{x}(t) - E[\hat{x}(t)])^T] \\
&= E[\hat{x}(t)\hat{x}(t)^T] - E[\hat{x}(t)]E[\hat{x}(t)]^T = E[\hat{x}(t)\hat{x}(t)^T] - E[x(t)]E[x(t)]^T, \\
Q_f(t) &= E[(x(t) - \hat{x}(t))(x(t) - \hat{x}(t))^T] \\
&= E[x(t)x(t)^T] - 2E[x(t)\hat{x}(t)^T] + E[\hat{x}(t)\hat{x}(t)^T] \\
&= E[x(t)x(t)^T] - 2E[E[x(t)|F_{t-1}^y]\hat{x}(t)^T] + E[\hat{x}(t)\hat{x}(t)^T] \\
&= E[x(t)x(t)^T] - E[\hat{x}(t)\hat{x}(t)^T] = Q_x(t) - Q_{\hat{x}}(t), \;\; Q_f(t) \geq 0, \\
Q_{\hat{x}}(t) &= Q_x(t) - Q_f(t) \leq Q_x(t), \\
\hat{x}(t+1) &= A(t)\hat{x}(t) + K(t, Q_f(t))\bar{v}(t), \\
Q_{\hat{x}}(t+1) &= A(t)Q_{\hat{x}}(t)A(t)^T + K(t, Q_f(t))Q_{\bar{v}}(t)K(t, Q_f(t))^T, \\
Q_{\bar{v}}(t) &= C(t)Q_f(t)C(t)^T + N(t)N)t)^T.
\end{aligned}$$

□

Computational Aspects

A direct implementation of the time-varying Kalman filter may run into numerical problems. Consider the Riccati recursion,

$$Q_f(t+1) = f_{FARE}(Q_f(t)), \; Q_f(0) = Q_{x_0}, \; Q_f : T \to \mathbb{R}_{pds}^{n_x \times n_x}.$$

In particular cases, depending on the initial condition, on the system matrices, and on the numerical implementation, there may exist a time $t \in T$ such that $0 \not\leq Q_f(t)$.

The approach to this problem is to use a square-root filter. In such a filter, the matrix function of the filter Riccati recursion is defined such that,

$$\begin{aligned}
&Q_r(t+1) = f_r(Q_r(t)),\ Q_r(0) = Q_{r,0} \in \mathbb{R}^{n_x \times n_x},\ Q_r : T \to \mathbb{R}^{n_x \times n_x},\ \text{such that,}\\
&\quad Q_f(t) = Q_r(t)Q_t(t)^T,\ \text{satisfies,}\\
&Q_f(t+1) = f_{FARE}(Q_f(t)),\ Q_f(0) = Q_{x_0}.
\end{aligned}$$

The approach was developed by J. Bierman, see his book [4]. The approach is quite effective and has been used for several Gaussian systems.

8.4 Time-Varying Kalman Filter and Stochastic Realization

The understanding of the filter problem may increase by relating it to stochastic realization of the output process.

The Kalman Filter and the Wiener Filter

Problem 8.4.1. Consider the *Wiener filter problem*. Assume that (1) the processes (y,z) are jointly stationary and Gaussian and (2) there exists a Gaussian stochastic realization of the combined process (y,z) in the form of a time-invariant Gaussian system,

$$\begin{aligned}
x(t+1) &= Ax(t) + Mv(t),\ x(0) = x_0 \in G(m_0, Q_0),\\
y(t) &= Cx(t) + Nv(t),\ \ v(t) \in G(0, Q_v),\\
z(t) &= Hx(t) + Jv(t),\ \ H \in \mathbb{R}^{n_z \times n_x},\ J \in \mathbb{R}^{n_z \times n_v}.
\end{aligned}$$

Determine for all $t \in T$ the conditional distribution of $z(t)$ conditioned on F_t^y.

Theorem 8.4.2. Relation of Wiener filter problem to Kalman filter problem. *Consider the Wiener filter problem for the specific case of Problem 8.4.1. Then the conditional distribution of $z(t)$ conditioned on F_{t-1}^y is Gaussian with,*

$$\begin{aligned}
E[\exp(iw^T z(t))|F_{t-1}^y] &= \exp(iw^T \hat{z}(t) - \frac{1}{2} w^T Q_z(t) w),\ \ \forall\, w \in \mathbb{R}^{n_z},\\
\hat{x}(t+1) &= A\hat{x}(t) + K(t)[y(t) - C\hat{x}(t)],\ \ \hat{x}(0) = m_0,\\
\hat{z}(t) &= E[z(t)|F_{t-1}^y] = H\hat{x}(t),\\
Q_z(t) &= HQ_f(t)H^T + JQ_vJ^T,
\end{aligned}$$

where the other matrices are as in the Kalman filter, see Theorem 8.3.2.

Proof. Because by Theorem 8.3.2 the conditional distribution of $x(t)$ conditioned on the σ-algebra F_{t-1}^y is Gaussian, and so is that of $z(t) = Hx(t) + Jv(t)$ when conditioned on the same σ-algebra. Then,

$$\begin{aligned}
\hat{z}(t) &= E[z(t)|F^y_{t-1}] = HE[x(t)|F^y_{t-1}] + JE[v(t)|F^y_{t-1}] = H\hat{x}(t),\\
z(t) - \hat{z}(t) &= H(x(t) - \hat{x}(t)) + Jv(t),\\
Q_z(t) &= E[(z(t) - \hat{z}(t))(z(t) - \hat{z}(t))^T] = HQ_f(t)H^T + JQ_vJ^T.
\end{aligned}$$

The result then follows from Theorem 8.3.2. □

The Kalman Filter and Stochastic Realization

The Kalman filter admits an interpretation as a stochastic realization. This interpretation is not surprising considering the approach of Wiener filtering.

Proposition 8.4.3. *Consider a time-varying Gaussian system and its associated time-varying Kalman filter,*

$$x(t+1) = A(t)x(t) + M(t)v(t), \tag{8.10}$$

$$y(t) = C(t)x(t) + N(t)v(t), \tag{8.11}$$

$$\hat{x}(t+1) = A(t)\hat{x}(t) + K(t)[y(t) - C(t)\hat{x}],\ \hat{x}(0) = m_0, \tag{8.12}$$

$$\overline{v}(t) = y(t) - C(t)\hat{x}(t). \tag{8.13}$$

The time-varying Kalman filter can be rewritten as a Gaussian system of the form, using the innovation process,

$$\hat{x}(t+1) = A(t)\hat{x}(t) + K(t)\overline{v}(t),\ \ \hat{x}(0) = m_0, \tag{8.14}$$

$$y(t) = C(t)\hat{x}(t) + \overline{v}(t). \tag{8.15}$$

(a)Then the Gaussian system (8.14,8.15) is analogous to that of (8.10,8.11), the innovation process is a Gaussian white noise process, and hence the Gaussian system (8.14,8.15) is also a weak Gaussian stochastic realization of the output process y.
(b)The variance of the Gaussian system (8.14,8.15) is less than or equal to the variance of the Gaussian system representation (8.10, 8.11), $Q_{\hat{x}}(t) \leq Q_x(t)$ for all times $t \in T$.
(c)The Gaussian system (8.14,8.15) corresponds to the Kalman realization of weak Gaussian stochastic realization theory.

Proof. (a) This follows from the above Gaussian system representation with the property that the innovation process is Gaussian white noise by Theorem 8.3.2.(d).
(b) This follows from Theorem 8.3.2.(g).
(c) See Def. 6.4.6. □

Derivation of the Kalman Filter via Stochastic Realization

The interpretation of the Kalman filter as a particular weak Gaussian stochastic realization of the output process then leads to the following problem.

Problem 8.4.4. *Kalman filter via stochastic realization*. Consider the Wiener filter problem for jointly stationary Gaussian stochastic processes (y,z) and assume that these processes satisfy the conditions for the existence of a weak Gaussian stochastic realization. Construct then the weak Gaussian stochastic realization with the condition that the state process at any time is measurable with respect to the σ-algebra of the output process up to the preceding time,

$$\begin{aligned}\hat{x}(t+1) &= A(t)\hat{x}(t)+K(t)\overline{v}(t),\ \ \hat{x}(0)=\hat{x}_0 \in G(\overline{m}_0,\overline{Q}_0),\\ y(t) &= C(t)\hat{x}(t)+\overline{v}(t),\ \overline{v}(t)\in G(0,\overline{Q_v}),\\ &\quad \hat{x}(t) \text{ is } F^y_{t-1} \text{ measurable. Hence,}\\ \hat{x}(t+1) &= A(t)\hat{x}(t)+K(t)[y(t)-C(t)\hat{x}(t)],\ \forall\, t\in T.\end{aligned}$$

The approach to the filter problem via realization theory is universal, it holds not only for Gaussian systems but also for other stochastic systems and for deterministic systems. Special cases are [42, 8] for finite state systems and [33] for rational systems.

Levinson Filter

N. Levinson was a colleague of N. Wiener at MIT during the period of about 1935 to 1950. They cooperated. N. Levinson published a derivation of the Wiener filter in [25, 24]. The approach is actually a form of stochastic realization though Levinson did not use that term. There follows a brief summary of the approach.

Levinson considered a stationary second-order process, the reader can also consider a stationary Gaussian process, and note the backward recursive estimates,

$$\begin{aligned}& y:\Omega\times T\to\mathbb{R},\ t\in T,\ E[y(t)]=0,\ Q_{y(t)}\succ 0,\\ & E[y(t)|F(y(t-1))]=Q_{y(t),y(t-1)}Q^{-1}_{y(t)}y(t-1),\\ & E[y(t)|F(y(t-1),\ y(t-2))]\\ =\ & \begin{pmatrix} Q_{y(t),y(t-1)} & Q_{y(t),y(t-2)}\end{pmatrix} Q^{-1}_{(y(t-1),y(t-2)}\begin{pmatrix} y(t-1)\\ y(t-2)\end{pmatrix},\\ & E[y(t)|F(y(t-1),\ldots,y(t-k))]=L(t,\ldots,t_k)\begin{pmatrix} y(t-1)\\ \vdots\\ y(t-k)\end{pmatrix},\ \forall\, k\in\mathbb{Z}_+.\end{aligned}$$

It is assumed that the variance matrices of the vectors of the output are nonsingular. The formulas for the conditional expectations follow from Theorem 2.8.3. In case the observation process y is the output of a Gaussian system with state-space dimension $n_x\in\mathbb{Z}_+$ then a vector of outputs of n_x elements is sufficient for a recursion as in the Kalman filter.

The above formulated approach produces what is called the *Levinson filter*. The approach has been generalized to the multivariable case by G. Ruckebusch, [38], see also [7, Annexe 8.A], and then the Levinson filter becomes a procedure to construct the stochastic realization corresponding to the Kalman filter.

The approach can also be formulated to construct the backward Kalman filter using the approach sketched by,

$$E[y(t-1)|F(y(t),y(t+1),\ldots,y(t+k))] = L \begin{pmatrix} y(t) \\ y(t+1) \\ \vdots \\ y(t+k) \end{pmatrix}, \ \forall \, k \in \mathbb{Z}_+.$$

8.5 Time-Invariant Kalman Filter

In control engineering, time-invariant Kalman filters are mostly used. How are these formulated?

Problem 8.5.1. *The filter problem of a time-invariant Gaussian system.*
Consider the time-invariant Gaussian system with representation,

$$\begin{aligned} x(t+1) &= Ax(t) + Mv(t), \ x(0) = x_0, \\ y(t) &= Cx(t) + Nv(t), \ v(t) \in G(0,I), \\ &N \in \mathbb{R}^{n_y \times n_v}, \ n_y \leq n_v, \ \text{rank}(N) = n_y \ \Rightarrow \ NN^T \succ 0. \end{aligned}$$

Determine the conditional distribution of the state based on partial observations, or an approximation of this conditional distribution.

Theorem 8.3.2 applied to the time-invariant Gaussian system of Problem 8.5.1 leads to a time-varying Kalman filter of the form,

$$\begin{aligned} \hat{x}(t+1) &= A\hat{x}(t) + K(Q_f(t))[y(t) - C\hat{x}(t)], \ \hat{x}(0) = m_{x_0}, \\ Q_f(t+1) &= f_{FARE}(Q_f(t)), \ Q_f(0) = Q_{x_0}. \end{aligned}$$

Note that the functions $K(.)$ and f_{FARE} do not depend explicitly on the time variable $t \in T$ due to the Gaussian system being time-invariant.

Engineers noted that there are examples for which the time-varying Kalman filter after a while becomes a time-invariant system. Hence the following limits existed for examples,

$$\lim_{t \to \infty} Q_f(t) = Q_f(\infty), \ \ \lim_{t \to \infty} K(Q_f(t)) = K(Q_f(\infty)).$$

Therefore engineers almost always implement the time-invariant Kalman filter defined below and based on the limit values. The little extra cost for use of a time-invariant filter was much less than the computational cost of implementing a filter with a time-varying gain matrix.

The definition of the time-invariant Kalman filter requires the definition of assumptions and of a result for the filter algebraic Riccati equation.

Definition 8.5.2. Consider Problem 8.5.1. The system matrices of the Gaussian system are denoted by $(A,\ C,\ M,\ N)$.

Define the matrices,

$$A_f \in \mathbb{R}^{n_x \times n_x},\ M_f \in \mathbb{R}^{n_x \times n_x},\ \text{such that,}\ \begin{pmatrix} M \\ N \end{pmatrix} \begin{pmatrix} M \\ N \end{pmatrix}^T \succeq 0 \Rightarrow \tag{8.16}$$

$$\begin{aligned} A_f &= A - MN^T(NN^T)^{-1}C, \\ M_f M_f^T &= MM^T - MN^T(NN^T)^{-1}NM^T \succeq 0. \end{aligned} \tag{8.17}$$

It follows from the result on the Schur complement, Proposition 17.4.33, for the matrix of equation (8.16) that $M_f M_f^T \succeq 0$.

The *controllability and the observability assumptions* (a) hold if: (1) (A,C) is an observable pair which is equivalent with the Gaussian system being stochastically observable; (2) (A_b,C_b) is an observable pair which is equivalent with the Gaussian system being stochastically co-observable; (3) (A_f,M_f) is a supportable pair. If $MN^T = 0$ then $A_f = A$ and M_f can be chosen as $M_f = M$ hence $(A_f,M_f) = (A,M)$ and the condition (2) is then equivalent to (A,M) being a supportable pair.

The *stabilizabilty and the detectability assumptions* (b) hold if: (1) (A,C) is a detectable pair; (1) (A_b,C_b) is a detectable pair; (3) (A_f,M_f) is a supportable-stable pair. Again, if $MN^T = 0$ then a corresponding implication as in (a) holds.

Assumptions (b) are strictly weaker than Assumptions (a).

The conditions above for the system matrices (A_b,C_b) of the backward system representation are not needed for the solution of the filter algebraic Riccati equation but if they are not satisfied then the Gaussian system is a nonminimal realization and then the state-space dimension is unnecessarily high.

Below part of Theorem 22.2.2 is stated for the information of the reader.

Theorem 8.5.3. *Consider the time-invariant Gaussian system with representation*

$$\begin{aligned} x(t+1) &= Ax(t) + Mv(t), x(0) = x_0, \\ y(t) &= Cx(t) + Nv(t). \end{aligned}$$

Consider Problem 8.5.1. Assume that Assumptions Def. 8.5.2.(b) hold. It follows from Theorem 22.2.2.(a) and (b) that there exists a unique solution Q_f of the filter algebraic Riccati equation*, (8.19), with the side conditions (8.18,8.20),*

$$Q_f \in \mathbb{R}_{pds}^{n_x \times n_x}, \tag{8.18}$$

$$\begin{aligned} Q_f &= f_{FARE}(Q_f) = AQ_fA^T + MM^T \\ &\quad -[AQ_fC^T + MN^T][CQ_fC^T + NN^T]^{-1}[AQ_fC^T + MN^T]^T. \end{aligned} \tag{8.19}$$

$$\text{spec}(A(Q_f)) \subset \text{D}_o;\ \textit{where,} \tag{8.20}$$

$$K(Q_f) = [AQ_fC^T + MN^T][CQ_fC^T + NN^T]^{-1} \in \mathbb{R}^{n_x \times n_y}, \tag{8.21}$$

$$A(Q_f) = A - K(Q_f)C \in \mathbb{R}^{n_x \times n_x}. \tag{8.22}$$

The conditions of the equations (8.18,8.20), are related, see Theorem 22.2.2. Nevertheless, the above formulation is preferred.

Definition 8.5.4. Consider the time-invariant Gaussian system with representation

$$\begin{aligned} x(t+1) &= Ax(t)+Mv(t), x(0)=x_0, \\ y(t) &= Cx(t)+Nv(t). \end{aligned}$$

Consider Problem 8.5.1. Assume that Assumptions Def. 8.5.2.(b) hold. It follows from Theorem 8.5.3 that there exists a matrix $Q_f \in \mathbb{R}_{pds}^{n_x \times n_x}$ which is a solution of the filter algebraic Riccati equation with side conditions.

Define respectively the *time-invariant Kalman gain matrix* and the *time-invariant Kalman filter* by the formulas,

$$K = K(Q_f) = [AQ_fC^T + MN^T][CQ_fC^T + NN^T]^{-1} \in \mathbb{R}^{n_x \times n_y}, \tag{8.23}$$

$$\bar{x}(t+1) = A\bar{x}(t) + K[y(t) - C\bar{x}(t)],\ \bar{x}(0) = 0,\ \bar{x} : \Omega \times T \to \mathbb{R}^{n_x}, \tag{8.24}$$

$$Q_{y,kf} = CQ_fC^T + NN^T. \tag{8.25}$$

Note that the state of the time-invariant Kalman filter is denoted by $\bar{x}(t)$ and not by $\hat{x}(t)$ because the hat notation implies that $\hat{x}(t) = E[x(t)|F_{t-1}^y]$ which does not hold in general for $\bar{x}(t)$. The choice $\bar{x}_0 = 0$ is for convenience. If the reader prefers another real vector then that is possible but it influences the initial behavior of the process.

A comparison follows between the time-varying Kalman filter and the time-invariant Kalman filter.

Proposition 8.5.5. *Consider Problem 8.5.1. Consider for a time-invariant Gaussian system the time-varying Kalman filter and the time-invariant Kalman filter, denoted respectively by,*

$$\begin{aligned} \hat{x}(t+1) &= A\hat{x}(t) + K(Q_{f,1}(t))[y(t) - C\hat{x}(t)],\ \hat{x}(0) = E[x_0], \\ Q_{f,1}(t+1) &= f_{FARE}(Q_{f,1}(t)),\ Q_{f,1}(0) = Q_{x_0}; \\ Q_f^* &= \mathrm{f}_{\mathrm{FARE}}(Q_f^*),\ Q_f^* \in \mathbb{R}_{pds}^{n_x \times n_x},\ \mathrm{spec}(A - K(Q_f^*)C) \subset \mathrm{D}_o; \\ \bar{x}(t+1) &= A\bar{x}(t) + K(Q_f^*)[y(t) - C\bar{x}(t)],\ \bar{x}(0) = 0. \end{aligned}$$

(a)The error system and the combined original system and error system, are represented by,

$$\begin{aligned} e(t) &= \hat{x}(t) - \bar{x}(t),\ \ e : \Omega \times T \to \mathbb{R}^{n_x}, \\ e(t+1) &= (A - K(Q_f^*)C)e(t) + [K(Q_f(t)) - K(Q_f^*)]\bar{v}(t), \\ & e(0) = \hat{x}(0) - \bar{x}(0) = m_{x_0} - \bar{x}(0), \\ Q_f(t+1) &= F_{FARE}(Q_f(t)),\ Q_f(0) = Q_{x_0}. \end{aligned}$$

(b)If $Q_{x_0} = Q_f^*$, where Q_f^* is the solution of the filter algebraic Riccati equation (FARE), then for all $t \in T$, $Q_f(t) = Q_f^*$.

(c)If $\bar{x}_0 = m_{x_0}$ and if $Q_{x_0} = Q_f^*$ then $e(0) = \hat{x}(0) - \bar{x}(0) = m_{x_0} - m_{x_0} = 0$ and, by (a) and (b), for all $t \in T$, $0 = e(t) = \hat{x}(t) - \bar{x}(t)$, hence $\bar{x}(t) = \hat{x}(t)$. Thus the time-invariant Kalman filter produces the same state estimates as the time-varying Kalman filter!

(d)Assume that Assumption Def. 8.5.2.(b) holds. If either $e(0) = \bar{x}_0 - m_{x_0} \neq 0$ *or if* $Q_{x_0} \neq Q_f$ *or if both these conditions hold then,*

$$\begin{aligned}
&\lim_{t\to\infty} Q_{f,1}(t) = Q_{f,1}(\infty) = Q_f^*, \textit{ where}\\
&Q_f^* = f_{FARE}(Q_f^*) \textit{ with the side conditions of Def. 8.5.4,}\\
&\lim_{t\to\infty} K(Q_{f,1}(t)) = K(Q_f^*) = K,\\
0 = L_2 - \lim(e(t)) = &\lim_{t\to\infty} E[(\hat{x}(t) - \bar{x}(t))(\hat{x}(t) - \bar{x}(t))^T].
\end{aligned}$$

The asymptotic convergence rate is that of the spectrum of $\mathrm{spec}(A - K(Q_f^*)C)$.

Proof. (a & b & c)

$$\begin{aligned}
e(t) &= \hat{x}(t) - \bar{x}(t), \;\; e : \Omega \times T \to \mathbb{R}^{n_x},\\
e(t+1) &= A\hat{x}(t) + K(Q_{f,1}(t))[y(t) - C\hat{x}(t)]\\
&\quad -A\bar{x}(t) - K(Q_f^*)[y(t) - C\hat{x}(t) + C\hat{x}(t) - C\bar{x}(t)]\\
&= (A - K(Q_f^*)C)e(t) + [K(Q_{f,1}(t)) - K(Q_f^*)]\bar{v}(t),\\
&\quad e(0) = E[x_0] - \bar{x}_0,\\
Q_{f,1}(t+1) &= f_{FARE}(Q_{f,1}(t)),\; Q_{f,1}(0) = Q_{x_0};\\
&\quad Q_{f,1}(0) = Q_{x_0} = Q_f^* \;\Rightarrow\; Q_{f,1}(t) = Q_f^*,\; \forall\, t \in T,\\
K(Q_{f,1}(t)) &= K(Q_f^*),\; K(Q_{f,1}(t)) - K(Q_f^*) = K(Q_f^*) - K(Q_f^*) = 0,\\
&\quad e(0) = \hat{x}_0 - \bar{x}_0 = 0 \;\Rightarrow\; e(t) = 0,\; \forall\, t \in T.
\end{aligned}$$

(d) Note that the assumption holds. It follows from Theorem 22.2.2.(c) that $\lim_{t\to\infty} Q_{f,1}(t) = Q_\infty$ and that $Q_\infty = f(Q_\infty) = Q_f^*$ is a solution of the filter algebraic Riccati equation. It follows from Theorem 22.2.2.(d) that $\mathrm{spec}(A - K(Q_\infty)C) \subset \mathrm{D}_o$, and from Theorem 22.2.2.(e) that the algebraic Riccati equation $Q_\infty = f(Q_\infty)$ has a unique solution. According to Definition 8.5.4, there exists a unique solution $Q_f^* = f_{FARE}(Q_f^*)$. From the uniqueness then follows that $Q_\infty = Q_f^*$. Thus,

$$\begin{aligned}
&\lim_{t\to\infty} Q_{f,1}(t) = Q_\infty = Q_f^*, \;\; \lim_{t\to\infty} K(Q_{f,1}(t)) = K(Q_\infty) = K(Q_f^*);\\
K(Q) &= [AQC^T + MN^T][CQC^T + NN^T]^{-1},\; K : \mathbb{R}^{n_x \times n_x} \to \mathbb{R}^{n_x \times n_y}.
\end{aligned}$$

where use is made of the fact that the function $Q \mapsto K(Q)$ is a well defined continuous function of Q.

Then, using the convergence result of Proposition 4.8.1,

$$\begin{aligned}
e(t+1) &= (A - K(Q_f^*)C)e(t) + [K(Q_{f,1}(t)) - K(Q_f^*)]\bar{v}(t),\\
&\quad e(0) = E[x_0] - \bar{x}_0,\;\; \mathrm{spec}(A - K(Q_f^*)C) \subset \mathrm{D}_o,\\
0 &= \lim_{t\to\infty}(K(Q_{f,1}(t)) - K(Q_f^*)),\\
\lim_{t\to\infty} Q_{\bar{v}}(Q_{f,1}(t)) &= \lim_{t\to\infty} [CQ_{f,1}(t)C^T + NN^T] = Q_{\bar{v}}(Q_f^*) \in \mathbb{R}^{n_y \times n_y},\\
&\Rightarrow \lim_{t\to\infty} E[(\hat{x}(t) - \bar{x}(t))(\hat{x}(t) - \bar{x}(t))^T] = 0.
\end{aligned}$$

□

An example is to apply the time-invariant Kalman filter to the combined system of a control system and a noise-shaping filter as described in Example 4.7.1. Then the Kalman filter estimates jointly the state of the control system and that of the noise-shaping filter.

8.6 Approximations of a Time-Invariant Kalman Filter

Engineers are faced with difficulties when constructing Kalman filters. For example, if the state-space dimension is very large to extremely large. Based on the structure of the Kalman filter, other linear filters are proposed. Such filters are simple to formulate but their performance determined by the variance of the estimation error process is in general less than that of the Kalman filter. There is then an engineering trade-off between simplicity of filter design and filter performance. In each case an engineer has to make a choice. The results of this section provide guidance for this engineering trade-off.

The time-invariant Kalman filter has several optimality properties. It has been proven in Section 8.3 that, for any $t \in T$, the conditional expection $E[x(t)|F_{t-1}^y]$ is the estimate of $x(t)$ conditioned in the σ-algebra F_{t-1}^y. In addition, the time-invariant Kalman filter has the lowest variance among all linear filters with a particular structure of the filter, as shown below.

Proposition 8.6.1. Optimality of the time-invariant Kalman filter over the set of time-invariant linear filters.
Consider Problem 8.5.1, the time-invariant Kalman filter,

$$\bar{x}(t+1) = A\bar{x}(t) + K(Q_f^*)[y(t) - C\bar{x}(t)],\ \bar{x}(0) = 0,$$

and an arbitrary time-invariant linear filter of the form,

$$\begin{aligned} x_a(t+1) &= Ax_a(t) + K_a[y(t) - Cx_a(t)],\ x_a(0) = 0, \\ &\quad \textit{with } K_a \in \mathbb{R}^{n_x \times n_y} \textit{ such that } \mathrm{spec}(A - K_aC) \subset \mathrm{D}_o. \end{aligned}$$

Then,

$$\begin{aligned} Q_f^* &= \lim_{t\to\infty} Q_f(t) = \lim_{t\to\infty} E[(x(t) - \bar{x}(t))(x(t) - \bar{x}(t))^T], \\ Q_{f,e_a} &= \lim_{t\to\infty} Q_{f,e_a}(t) = \lim_{t\to\infty} E[(x(t) - x_a(t))(x(t) - x_a(t))^T], \\ Q_{f,e_a} &= (A - K_aC)Q_{f,e_a}(A - K_aC)^T + (M - K_aN)(M - K_aN)^T, \\ Q_f^* &\leq Q_{f,x_a}, \end{aligned}$$

hence the asymptotic variance of the time-invariant linear filter is always larger than or equal to that of the time-invariant Kalman filter. Q_{f,e_a} is the unique solution of the Lyapunov equation listed above.

Proof. (a) Define the processes $e,\ e_a : \Omega \times T \to \mathbb{R}^{n_x}$,

$$\begin{aligned}
e(t) &= x(t) - \overline{x}(t), \; e_a(t) = x(t) - x_a(t). \\
e(t+1) &= (A - KC)e(t) + (M - KN)v(t), \; e(0) = x_0, \\
e_a(t+1) &= (A - K_aC)e_a(t) + (M - K_aN)v(t), \; e_a(0) = x_0.
\end{aligned}$$

It follows from Theorem 22.2.2 that,

$$\begin{aligned}
&\lim_{t\to\infty} E\left[(x(t) - \overline{x}(t) - E[x(t) - \overline{x}(t))(x(t) - \overline{x}(t) - E[x(t) - \overline{x}(t))^T\right] \\
&= \lim_{t\to\infty} E[(e(t) - E[e(t)])(\ldots)^T] = \lim Q_f(t) = Q_f^* \succeq 0,
\end{aligned}$$

and Q_f^* is the solution of the FARE $Q_f^* = f(Q_f^*)$ with the side conditions. Define as in Section 22.2,

$$\begin{aligned}
L_o &: \mathbb{R}^{n_x \times n_x} \times \mathbb{R}^{n_x \times n_y} \to \mathbb{R}^{n_x \times n_x}, \\
L_o(Q,K) &= [A - KC]Q[A - KC]^T + [M - KN][M - KN]^T.
\end{aligned}$$

From Proposition 22.2.7.(d) follows that,

$$\begin{aligned}
Q_f^* &= f(Q_f^*) = L_o(Q_f^*, K(Q_f^*)); \text{ note that,} \\
e_a(t+1) &= x(t+1) - x_a(t+1) = Ax(t) + Mv(t) - Ax_a(t) - K_a(y(t) - Cx_a(t)) \\
&= (A - K_aC)e_a(t) + (M - K_aN)v(t), \; e_a(0) = x_0, \\
Q_{e_a}(t) &= E\left[(e_a(t) - E[e_a(t)])^T(e_a(t) - E[e_a(t)])^T\right], \\
Q_{x_a}(t+1) &= (A - K_aC)Q_{x_a}(t)(A - K_aC)^T + (M - K_aN)(M - K_aN)^T, \\
&\quad Q_{e_a}(0) = Q_{x_0}, \text{ by Theorem 4.3.5,} \\
\lim_{t\to\infty} Q_{e_a}(t) &= Q_{f,e_a}, \text{ because of } \mathrm{spec}(A - K_aC) \subset \mathbb{D} \text{ and of Theorem 4.4.5,} \\
Q_{f,e_a} &= (A - K_aC)Q_{f,x_a}(A - K_aC)^T + (M - K_aN)(M - K_aN)^T \\
&= L_o(Q_{f,e_a}, K_a),
\end{aligned}$$

and Q_{f,e_a} is the unique solution of this equation. It follows from Proposition 22.2.7.(d) that

$$\begin{aligned}
Q_{f,e_a} - Q_f^* &= L_o(Q_{f,e_a}, K_a) - L_o(Q_f^*, K(Q_f^*)) \\
&= (A - K_aC)(Q_{f,e_a} - Q_f^*)(A - K_aC)^T + \\
&\quad + (K_a - K(Q_f^*))(CQ_f^*C^T + NN^T)(K_a - K(Q_f^*))^T.
\end{aligned}$$

Note that $CQ_f^*C^T + NN^T \succeq NN^T \succ 0$. From Theorem 22.1.2.(b) follows that $Q_{f,e_a} - Q_f^* \succeq 0$ or $Q_f^* \preceq Q_{f,e_a}$. □

The trade-off in the filter design is described next. One of the control objectives of filter design is to achieve a low variance of the estimation error Q_e and a fast convergence of Q_e to the limit value. If the Gaussian system is such that (A,C) is an observable pair, then the control objective of fast convergence can be achieved by selection of a Kalman gain matrix K_s such that the eigenvalues of the system matrix of the error system, $\mathrm{spec}(A - K_sC) \subset \mathrm{D}_o$, are placed arbitrarily close to zero.

Another control objective is keep the magnitude of the correction term in the Kalman filter, $K_s[y(t) - C\hat{x}(t)]$, small. Note that the above variable describes the term which has to be supplied to the model of the system in the Kalman filter to achieve a state estimate. This can be achieved by keeping the magnitude of the Kalman gain matrix K_s small.

In signal processing one searches for a trade-off between the two control objectives described above: on one hand a small error variance and fast convergence, on the other hand a small Kalman gain matrix. The Kalman filter achieves such a trade-off. In the Kalman filter the matrices (MM^T, NN^T, MN^T) describing the variance of the combined noise components, determine the optimal Kalman gain matrix $K(Q_f^*)$ which achieves the minimal error variance. In practice, the values of the matrices (M, N) are only known approximately hence one has to experiment with the Kalman filter in the design process.

Data assimilation is a subarea of filtering in which a specific form of filters are used that are inspired by the time-invariant Kalman filter [1]. The motivation comes from applying the time-invariant filter to stochastic systems with a very high state-space dimension say in the order of 10^2, 10^3, 10^4, or higher. The system matrices A and C in such a case are sparse matrices, meaning that in every row and every column only a few % of the entries are nonzero. If that happens to be the case then it does not make sense to carry out the computation of $A\hat{x}(t)$ as a matrix vector multiplication. In stead the software specified to only compute expressions like the following formula,

$$(A\hat{x}(t))_i = A_{i,j_1}\hat{x}(t)_{j_1} + A_{i,j_2}\hat{x}(t)_{j_2} + A_{i,j_3}\hat{x}(t)_{j_3} + \text{etc.},$$

for particular indices $j_1, j_2, j_3 \in \mathbb{Z}_n$.

A further saving on the computational complexity can be made by not computing the solution of the algebraic Riccati equation. Thus one selects a time-invariant gain matrix $K \in \mathbb{R}^{n_x \times n_y}$ such that the stability condition holds, $\text{spec}(A - KC) \subset \text{D}_o$. For subclass of stochastic systems, design techniques for the choice of the gain matrix have been developed. Such techniques are best learned from application papers.

Sensor Allocation

The sensor allocation problem arises at various domains of engineering. Where along a river should observation stations be situated so as to best estimate the pollution concentrations in the water? The aim of that investigation is to measure water quality. Where in a large lake, are the observation stations best situated? Which sensors have to be placed in a car to quickly estimate the state of the car for use in control of the vehicle? Where should the sensors be based on a rotating satellite to observe the orientation which measurements are then used for estimating the state of the body?

The decision of how many sensors to use and of where to place them is also an economic issue. At the end, the owner of the problem has to take a decision. A control engineer can advise the owner of the problem.

Example 8.6.2. *A sensor allocation problem.* Consider a time-invariant Gaussian system with representation,

$$\begin{aligned} x(t+1) &= Ax(t)+Mv(t),\ x(0)=x_0, \\ y(t) &= C_1x(t)+Nv(t),\ \ T=\mathbb{N},\ n_x,\ n_y=1\in\mathbb{Z}_+, \\ C_1 &= \begin{pmatrix} 1\ 0\ 0\ 0 \end{pmatrix}. \end{aligned}$$

Suppose that the performance of the time-invariant Kalman filter is below ones expectations. Therefore the selection of another sensor is considered. A sensor will be associated with another row of the C matrix, for example,

$$C_2 = \begin{pmatrix} 0\ 0.3\ 0.3\ 0.4 \end{pmatrix},\ C = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}.$$

A comparison of the performances is then to be made of the time-invariant Kalman filter in terms of the error variances and in terms of the singular values of the observability matrices for the cases (1) C_1 only and (2) $(C_1,\ C_2)$.

$$Q_f^*(C_1,C_2) \preceq Q_f^*(C_1)?;\ \ \text{SVD}(\text{obsmat}(A,(C_1\ C_2))),\ \ \text{SVD}(\text{obsmat}(A,C_1)).$$

An option is always to choose $C=I_{n_x}$ though this is often physically not possible. One can formulate a simple optimization problem to determine a solution.

Criteria for the selection of the sensors often include the following.

1. There should not be redundant sensors in general. Mathematically formulated, one requires that $\text{rank}(C)=\min\{n_y,\ n_x\}$. If $n_y \leq n_x$ and if $\text{rank}(C) < n_y$ then the row vectors of the C matrix of two or more sensors are dependent and the number of sensors can then be reduced.
2. The condition of stochastic observability condition and of stochastic co-observability should hold, either for the (A,C) pair or for the (A_f,M_f) pair, see Def. 8.5.2. Mathematically this can be formulated, for example, as the rank condition on the observability matrix, $\text{rank}(\text{obsmat}(A,C))=n_x$. An advice is to also investigate the singular values of the observability matrix, possibly also the matrix of the singular vectors.
3. The asymptotic Kalman filter should converge relatively fast when started at an arbitrary condition or when moved away from the state estimate by a noise perturbation. In practice this amounts to a trade-off between on one hand the magnitude of the Kalman gain matrix and on the other hand the spectrum of the system matrix of the error system, $\text{spec}(A-KC)$, and the variance of the error, Q_e. See below for additional information.

In practice a combination of these criteria is used.

The sensor allocation problem gets more complicated if one starts to add rows to the observation matrix C or modify the rows of this matrix. It is possible to define an optimization problem, compute solutions, and evaluate those. Such a control design procedure can be carried out by the reader.

8.7 Prediction

Prediction problems arise in several areas of engineering. Examples are the prediction of power demand by households and by companies for tomorrow. Similarly, prediction of gas demand, demand for drinking water, etc. In road traffic management, prediction of tomorrows traffic intensity is used for control purposes. Finally weather predictions are based on Kalman predictors of pressures over a large geographical area. Examples of prediction problems include prediction of weather variables, demand for water supply in a city, water levels at the North Sea near the harbor of Rotterdam, power demand, etc.

Prediction problems for Gaussian systems are based on the Kalman filter though with a new element. The Kalman predictor is derived below.

Problem 8.7.1. *Prediction problem of a Gaussian system*. Consider a time-varying Gaussian system with forward representation

$$\begin{aligned} x(t+1) &= A(t)x(t)+M(t)v(t),\ x(0)=x_0, \\ y(t) &= C(t)x(t)+N(t)v(t). \end{aligned}$$

The *prediction problem* is to determine the conditional probability distribution of $x(t)$ conditioned on the σ-algebra F^y_{s-1} for $s,t \in T$, $s<t$, with:

(a)*fixed data: s* is fixed and t is increasing;
(b)*fixed increment:* $t-s$ is fixed and s is increasing;
(c)*fixed prediction time: t* is fixed and s is increasing.

In particular, determine a system for the parameters of the conditional probability distribution.

Theorem 8.7.2. *Consider the prediction Problem 8.7.1.*

*(a)*The case of fixed data *with a prediction horizon of $t_p \in \mathbb{Z}_+$. The conditional probability distribution of $x(t)$ conditioned on F^y_{s-1} for all $s,t \in T$ with $s<t$, is Gaussian with conditional mean and conditional variance specified by,*

$$\begin{aligned} & E[\exp(iw^T x(t))|F^y_{s-1}] \\ & = \exp(iw^T \hat{x}(t|s-1) - \frac{1}{2} w^T Q_p(t|s-1)w),\ \forall\, w \in \mathbb{R}^n, \\ & \forall\, t = s,\ s+1,\ldots,t_1, \\ \hat{x}(t+1|s-1) &= E[x(t+1)|F^y_{s-1}], \\ Q_p(t+1|s-1) &= E[(x(t+1)-\hat{x}(t+1|s-1))(x(t+1)-\hat{x}(t+1|s-1))^T]. \end{aligned}$$

(b)The conditional mean and the conditional variance are determined by the recursions, where $\hat{x}(s)$, $Q_f(s)$ are generated by the time-varying Kalman filter,

$$\forall\, t = s,\ s+1,\ldots,t_1-1,$$

$$\hat{x}(t+1|s-1) = A(t)\hat{x}(t|s-1),\ \hat{x}(s|s-1) = \hat{x}(s), \tag{8.26}$$

$$\begin{aligned} Q_p(t+1|s-1) &= A(t)Q_p(t|s-1)A(t)^T + M(t)M(t)^T, \\ & Q_p(s|s-1) = Q_f(s). \end{aligned} \tag{8.27}$$

*(c)*The case of a fixed increment *with as increment* $t_{incr} \in \mathbb{Z}_+$. *Then,*

$$\begin{aligned}\hat{x}(t+t_{incr}) &= \phi(t+t_{incr},t)\hat{x}(t),\ \forall\, t \in T,\\ &\phi(t+t_{incr},t) = A(t+t_{incr}-1)\ldots A(t).\end{aligned}$$

*(d)*The case of a fixed-terminal-time of prediction *with as time* $t \in T$ *and as starting time* $s \in T$. *Then,*

$$\begin{aligned}\hat{x}(t) &= \phi(t,r)\hat{x}(r),\ \forall\, r = s,\ s+1,\ s+2,\ldots,t-1,\\ &\phi(t,r) = A(t-1)\ldots A(r).\end{aligned}$$

Proof. (a) It follows similarly as in the proof of Theorem 8.3.2 that the conditional probability distribution of $x(t)$ conditioned on F^y_{s-1} is Gaussian for $s,\ t \in T$, $s < t$. Denote of this conditional distribution the conditional mean by $\hat{x}(t|s-1)$ and the conditional variance by $Q_f(t|s-1)$. It follows from Theorem 2.8.3 that,

$$\begin{aligned}\hat{x}(t|s-1) &= E[x(t)|F^y_{s-1}],\\ Q_p(t|s-1) &= E[(x(t)-\hat{x}(t|s-1))^T(x(t)-\hat{x}(t|s-1))].\end{aligned}$$

(b) Next the recursions are derived.

$$\begin{aligned}\hat{x}(t+1|s-1) &= E[x(t+1)|F^y_{s-1}] = E[A(t)x(t)+M(t)v(t)|F^y_{s-1}]\\ &= A(t)\hat{x}(t|s-1);\ \text{ because by reconditioning,}\\ E[v(t)|F^y_{s-1}] &= E[E[v(t)|F_{t-1}]|F^y_{s-1}] = E[E[v(t)]|F^y_{s-1}] = 0,\end{aligned}$$

because $v(t)$ is independent of F_{t-1}. In addition, $\hat{x}(s|s-1) = E[x(s)|F^y_{s-1}] = \hat{x}(s)$ where $\hat{x}(s)$ is produced by the Kalman filter. For the conditional variance one calculates,

$$\begin{aligned}& x(t+1)-\hat{x}(t+1|s-1)\\ &= A(t)x(t)+M(t)v(t)-A(t)\hat{x}(t|s-1) = A(t)[x(t)-\hat{x}(|s-1)]+M(t)v(t),\\ & E[(x(t)-\hat{x}(t|s-1))v(t)^T] = E[E[(x(t)-\hat{x}(t|s-1))v(t)^T|F_{t-1}]]\\ &= E[(x(t)-\hat{x}(t|s-1))\ E[v(t)^T|F_{t-1}]] = 0;\\ & Q_p(t+1|s-1) = E[(x(t+1)-\hat{x}(t+1|s-1))\ (\ldots)^T]\\ &= A(t)Q_p(t|s-1)A(t)^T+M(t)M(t)^T,\\ & Q_p(s|s-1) = Q_f(s),\ \text{where } Q_f(s) \text{ is produced by the Kalman filter.}\end{aligned}$$

(c) and (d). The proofs are similar to those of (a) and (b). □

The performance of the Kalman predictor for a time-invariant Gaussian system is discussed. According to Theorem 8.7.2 and at any time $tinT$, the conditional mean of the prediction starts from the state estimate of the Kalman filter $\hat{x}(s+1|s) = \hat{x}(s+1)$ and along the prediction horizon proceeds by the recursion $\hat{x}(t+1+s|s) = A\hat{x}(t+s|s)$, for $t = 1,\ 2,\ \ldots,\ t_1$, hence converges to zero if the system matrix A is exponentially stable. The speed with which the predictions go to zero determines the quality of the predictions: if the convergence is slow, say with eigenvalues 0.9 or 0.85, then the predictions are valuable for about 5 time steps. if the convergence is fast, say with an eigenvalue of 0.4 then the prediction of only one time step seems valuable. Note that the eigenvalues of the system matrix A are relevant for this performance, not those of the error system $A - KC$.

The variance of a prediction starts at the estimate produced by the Riccati recursion $Q_f(t)$ and progresses along the prediction horizon by the recursion,

$$Q_p(t+1+s|s) = AQ_p(t+s|s)A^T + MM^T.$$

From the formula of the filter Riccati recursion it is clear that the magnitude of the variance matrix $Q_p(t+s|s)$ is larger than that of $Q_f(s)$ hence the quality of the error variance decreases as the prediction horizon advances.

In practice, the predictions are useful for a horizon which is about as large as the dimension of the state vector of the system while being more accurate for low prediction horizons and less accurate for longer and longer prediction horizons.

Smoothing problems can be solved analogously to prediction problems. Smoothing is like prediction but then in the backward direction using at any time output values from future times. The reader may find results on smoothing problems in [6, Sec. 3.6].

8.8 Interpolation

There is a signal processing problem for which concepts and procedures of the Kalman filter have been used. The problem is called interpolation.

Consider a CD player for music CDs. CD players were developed during the 1980’s based on laser technology. The music of a CD player provided a higher degree of signal reproduction than the LP records of an earlier generation. At the time this book is published, the CD technology is slowly phased out though many older persons still have a CD player and CDs. The current access to music is via a provider which delivers a stream of signals to one’s house or to one’s phone.

With a CD there can be a problem. In case of a skratch on the CD part of the music signal is missing. This is supposed not to happen but humas drop CDs or accidentally skratch them. Therefore a CD player has to have a recovery procedure to restore missing music if there is a skratch.

Researchers of the company Philips Research solve the interpolation problem for the missing music signal. In addition, they online estimated the parameters of the stochastic system which described the local part of the music, where, due to the changing music, the parameters change in time. The solution is encoded in a standard that was then implemented in the CD players of the Philips company and the standard was used and possibly is still used by other companies. The implementation is such that at any time, there is about a minute or so of output signal from the CD in the CD player to which an electronic circuit applies the procedure for the problem. If a skratch is present and detected then the estimated signal replaces the missing music which is then sent to the speakers for the listener.

The research was carried out by A.J.E.M. Janssen, R.N.J. Veldhuis, and L.B. Vries, [14, 13].

The term *interpolation* is used for the problem of reconstructing the missing music signal. The word *interpolation* is used also in other parts of mathematics and of engineering. The term is not standardized.

The engineering model of the interpolation problem follows. The signal is assumed to be modelled as the output of a time-varying Gaussian system. There is a finite horizon. The output signal is not available for a middle part of the horizon. The interpolatioin problem is to determine the conditional probability distribution of the state and of the output of the signal for the times in the interval of the missing outputs.

Problem 8.8.1. *Interpolation problem for a time-varying Gaussian system.* Consider a time-varying Gaussian system with the forward represenation, Def. 4.3.1,

$$\begin{aligned} x(t+1) &= A(t)x(t) + M(t)v(t),\ x(0) = x_0,\\ y(t) &= C(t)x(t) + N(t)v(t),\\ &\quad x_0 \in G(0, Q_{x_0}),\ \forall\, t \in T,\ v(t) \in G(0, I_{n_v}),\\ T &= \{0, 1, \ldots, t_1\} = T(0 : t_1) \subset \mathbb{N}. \end{aligned}$$

Suppose that the observations are not available on the interval,

$$T_m = \{t_2+1, t_2+2, \ldots, t_3-1\} \subset T,\ t_2,\ t_3 \in T,$$

The problem is to determine the conditional probability distribution of the states and of the outputs for times in the interval of missing output values,

$$\begin{aligned} &E[\exp(iw_x^T x(s)) |\, F_{t_2}^{y-} \vee F_{t_3}^{y+}],\ \forall\, w_x \in \mathbb{R}^{n_x},\\ &E[\exp(iw_y^T y(s)) |\, F_{t_2}^{y-} \vee F_{t_3}^{y+}],\ \forall\, w_y \in \mathbb{R}^{n_y},\ \forall\, s \in T_m. \end{aligned}$$

This problem was solved by the researchers quoted. Their solution is in terms of an autoregressive system representation which is much used in the research area of signal processing. Another publication by M. Pavon is formulated in terms of Hilbert spaces, [34]

The author prefers a different solution formulation which is summarized below. The proof is still not complete at this time of writing. The conjectured solution employs the forward Kalman filter and the backward Kalman filter. In addition, it uses the forward predictor based on the past output and the backward smoother based on the future output.

Proposition 8.8.2. *Consider the interpolation Problem 8.8.1. The solution of the interpolution problem is,*

$$\begin{aligned} \hat{x}_-(t+1) &= A(t)\hat{x}_-(t) + K_-(t, Q_{f-}(t))[y(t) - C(t)\hat{x}_-(t)],\ \hat{x}_-(0) = 0,\\ \hat{x}_+(t-1) &= A(t)\hat{x}_+(t) + K_+(t, Q_{f+}(t))[y(t) - C(t)\hat{x}_+(t)],\ \hat{x}_-(t_1) = 0,\\ &E[x(t_2+1) |\, F_{t_2}^{y-}] = \hat{x}_-(t_2+1|t_2) = \hat{x}_-(t_2+1),\\ &E[x(t_3-1) |\, F_{t_3}^{y+}] = \hat{x}_+(t_3-1|t_3) = \hat{x}_-(t_3-1), \end{aligned}$$

$$\begin{aligned}
&\forall\ s \in T_m = \{t_2+1,\ldots,t_3-1\},\\
&E[\exp(iw_x^T x(s))|\ F_{t_2}^{y-} \vee F_{t_3}^{y+}] = E[\exp(iw_x^T x(s))|\ F^{\hat{x}_-(t_2)} \vee F^{\hat{x}_+(t_3)}],\\
&E[x(s)|\ F_{t_2}^{y-} \vee F_{t_3}^{y+}] = L_1(s,t_2)\hat{x}_-(t_2) + L_+(s,t_3)\hat{x}_+(t_3),\\
&E[\exp(iw_y^T y(s))|\ F_{t_2}^{y-} \vee F_{t_3}^{y+}].
\end{aligned}$$

Proof. (1) First it is proven using conditional independence that,

$$(F^{x(s)}, F_{t_2}^{y-}|\ F^{\hat{x}_-(s|t_2)}) \in \text{CI},\ \ (F^{x(s)}, F_{t_3}^{y+}|\ F^{\hat{x}_+(s|t_3)}) \in \text{CI}.$$

(2) The projection is calculated of,

$$E[\exp(iw_x^T x(s))|\ F^{\hat{x}_-(t_2|s)} \vee F^{\hat{x}_+(s|t_3)}.$$

with the formulas for the representation. □

8.9 Conditional Kalman Filter

The reader finds in this section the problem formulation, the theorem, and the proof of the conditional Kalman filter. The result is used in stochastic control with partial observations, see Chapter 14.

The reader is assumed to be familiar with the concept of a conditionally Gaussian process and a conditionally Gaussian system. See for these concepts Section 19.7.

Problem 8.9.1. Consider a conditional Gaussian system of the form,

$$x(t+1) = A(t)x(t) + b(t) + M(t)v(t),\ \ x(0) = x_0 \in G(0,Q_0), \tag{8.28}$$

$$y(t) = C(t)x(t) + c(t) + N(t)v(t), \tag{8.29}$$

$$\begin{aligned}
&v: \Omega \times T \to \mathbb{R}^{m_v},\ v(t) \in G(0,I),\ \text{Gaussian white noise},\\
&F^{x_0},\ F_\infty^v,\ \text{are independent},\\
&A: \Omega \times T \to \mathbb{R}^{n_x \times n_x},\ (A(t), F_{t-1}^y, t \in T),\ \text{adapted},\\
&b: \Omega \times T \to \mathbb{R}^{n_x},\ (b(t), F_{t-1}^y, t \in T),\ \text{adapted},\\
&C: \Omega \times T \to \mathbb{R}^{n_y \times n_x},\ (C(t), F_{t-1}^y, t \in T),\ \text{adapted},\\
&c: \Omega \times T \to \mathbb{R}^{n_y},\ (c(t), F_{t-1}^y, t \in T),\ \text{adapted},\\
&M: T \to \mathbb{R}^{n_x \times n_v},\ N: T \to \mathbb{R}^{n_y \times n_v}.
\end{aligned}$$

Determine the conditional characteristic function of the current state based on past observations,

$$\begin{aligned}
&E[\exp(iw_x(t)^T x(t))|F_{t-1}^y],\ \forall\ t \in T,\\
&\forall\ w_x: \Omega \times T \to \in \mathbb{R}^{n_x},\ w_x(t)\ \text{is}\ F_{t-1}^y\ \text{measurable}.
\end{aligned}$$

Note that the state and output process are not necessarily Gaussian in general. Hence the theorem of the unconditional Kalman filter does not apply.

Theorem 8.9.2. *Consider the conditional Kalman filter Problem 8.9.1. Assume that, $\forall\, t \in T,\ \ N(t)N(t)^T \succ 0$.*

(a)The conditional distribution of $x(t)$ conditioned on F^y_{t-1}, for any $t \in T$, is conditionally Gaussian and specified by the characteristic function,

$$E[\exp(iw_x^T x(t))|F^y_{t-1}] = \exp(iw_x^T \hat{x}(t) - \frac{1}{2} w_x^T Q_f(t) w_x),\ \forall\, t \in T, \tag{8.30}$$
$$\forall w_x : \Omega \times T \to \mathbb{R}^{n_x},\ w_x(t),\ F^y_{t-1} \text{ measureable.}$$

Call then $\hat{x} : \Omega \times T \to \mathbb{R}^{n_x}$ the conditional mean process *and $Q_f : T \to \mathbb{R}^{n_x \times n_x}$ the* conditional error-variance process. *Note that the conditional error-variance process is stochastic and will in general depend on the output process. This property makes the conditional Kalman filter different from the time-varying Kalman filter.*

(b)The parameters of the characteristic function (8.30) can be recursively calculated according to the recursions,

$$\begin{aligned}
\hat{x}(t+1) &= A(t)\hat{x}(t) + b(t) + K(t)[y(t) - c(t) - C(t)\hat{x}(t)], \\
\hat{x}(0) &= E[x_0] = m_0, && (8.31) \\
H_{11}(t) &= A(t)Q_f(t)A(t)^T + M(t)M(t)^T, \\
H_{12}(t) &= A(t)Q_f(t)C(t)^T + M(t)N(t)^T, \\
H_{22}(t) &= C(t)Q_f(t)C(t)^T + N(t)N(t)^T, \\
K(t) &= H_{12}(t)H_{22}^{-1}(t) \in \mathbb{R}^{n_x \times n_y}, && (8.32) \\
Q_f(t+1) &= H_{11}(t) - H_{12}(t)H_{22}^{-1}H_{12}^T(t) \in \mathbb{R}^{n_x \times n_x}_{pds}, && (8.33) \\
Q_f(0) &= E[(x_0 - E[x_0])(x_0 - E[x_0])^T] = Q_{x_0},
\end{aligned}$$

The recursion of equation (8.31) is called the conditional Kalman filter *and that of equation (8.33) is called the* conditional Filter Riccati recursion.

(c)From (a) and (b) follows that for all $t \in T$ a.s.,

$$\begin{aligned}
\hat{x}(t) &= E[x(t)|F^y_{t-1}], && (8.34) \\
Q_f(t) &= E[(x(t) - \hat{x}(t))(x(t) - \hat{x}(t))^T|F^y_{t-1}]. && (8.35)
\end{aligned}$$

(d)Define the innovation process *as the stochastic process,*

$$\overline{v} : \Omega \times T \to \mathbb{R}^{n_y},\ \overline{v}(t) = y(t) - C(t)\hat{x}(t).$$

Then $\overline{v}$ is a conditional Gaussian process such that for all $t \in T$,

$$\begin{aligned}
&E[\exp(iw^T \overline{v}(t))|F^y_{t-1}] = \exp(-w^T Q_{\overline{v}}(t) w),\ \forall\, w \in \mathbb{R}^{n_y}, \\
&\overline{v}(t) \in G(0, Q_{\overline{v}}(t)),\ Q_{\overline{v}} : \Omega \times T \to \mathbb{R}^{n_y \times n_y}_{pds}, \\
Q_{\overline{v}}(t) &= C(t)Q_f(t)C(t)^T + N(t)N(t)^T.
\end{aligned}$$

(f) Define the estimation error process,

$$\begin{aligned}
e(t) &= x(t) - \hat{x}(t),\ e : \Omega \times T \to \mathbb{R}^{n_x};\ \text{then,} \\
e(t+1) &= (A(t) - K(t)C(t))e(t) + (M(t) - K(t)N(t))v(t),\ e(0) = x_0 - m_0.
\end{aligned}$$

Proof. Below use is made of the definition and theorems for conditional Gaussian random variables, see Section 19.7.
(1) Define the σ-algebra families,

$$\begin{aligned}
&\{F_t, t \in \{-1\} \cup T\},\ F_{-1} = F^{x_0},\ F_t = F^{x_0} \vee F_t^v, \forall t \in T, \\
&\{F_t^y, t \in \{-1\} \cup T\},\ F_{-1}^y = \{\emptyset, \Omega\},\ F_t^y = \sigma(\{y(s), \forall s \leq t\}),\ \forall t \in T.
\end{aligned}$$

From the equations,

$$\begin{aligned}
x(t+1) &= A(t)x(t) + M(t)v(t), x(0) = x_0, \\
y(t) &= C(t)x(t) + N(t)v(t),
\end{aligned}$$

it follows by induction that for all $t \in T$, $x(t+1)$ and $y(t)$ are F_t measurable. For example, x_0 is F_{-1} measurable by definition of that σ-algebra. Suppose by induction that $x(t)$ is F_{t-1} measurable. Then it follows from the recursion for the state process displayed above and from the definition of the filtration $\{F_t, t \in T\}$ that $x(t+1)$ and $y(t)$ are F_t measurable.

(2) For any $t \in T$,

$$\begin{aligned}
&E[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix}\right)|F_{t-1}] \\
&= \exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} A(t)x(t) \\ C(t)x(t) \end{pmatrix}\right) E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} M(t) \\ N(t) \end{pmatrix} v(t)\right)|F_{t-1}\right] \\
&= \exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} A(t)x(t) \\ C(t)x(t) \end{pmatrix} - \frac{1}{2}\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} M(t) \\ N(t) \end{pmatrix} \begin{pmatrix} M(t) \\ N(t) \end{pmatrix}^T \begin{pmatrix} w_x \\ w_y \end{pmatrix}\right),
\end{aligned}$$

hence $(x(t+1), y(t))$ are conditionally Gaussian given F_{t-1}.
(3) By induction in $t \in T$ the following two equations will be proven to hold,

$$\begin{aligned}
&E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix}\right)|F_{t-1}^y\right] \\
&= \exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} A(t)\hat{x}(t) \\ C(t)\hat{x}(t) \end{pmatrix} - \frac{1}{2}\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T H(t) \begin{pmatrix} w_x \\ w_y \end{pmatrix}\right),\ \forall t \in T,
\end{aligned} \tag{8.36}$$

$$\begin{aligned}
H(t) &= \begin{pmatrix} H_{11}(t) & H_{12}(t) \\ H_{12}(t)^T & H_{22}(t) \end{pmatrix}, \\
H_{11}(t) &= A(t)Q_f(t)A(t)^T + M(t)M(t)^T, \\
H_{12}(t) &= A(t)Q_f(t)C(t)^T + M(t)N(t)^T, \\
H_{22}(t) &= C(t)Q_f(t)C(t)^T + N(t)N(t)^T,
\end{aligned}$$

$$\begin{aligned}
&E[\exp(iw_x^T x(t+1))|F_t^y] \\
&= \exp(iw_x^T \hat{x}(t+1) - \frac{1}{2}w_x^T Q_f(t+1)w_x),\ \forall t \in T;,
\end{aligned} \tag{8.37}$$

$$x(1) = A(0)x_0 + M(0)v(0),\ \ y(0) = C(0)x_0 + N(0)v(0).$$

By the assumptions on a Gaussian system, $(x_0, v(0))$ are independent random variables with $x_0 \in G(m_0, Q_0)$ and $v(0) \in G(0, I)$. Thus $(x_0, v(0))$ are jointly Gaussian random variables. Then,

$$\begin{aligned}
& E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x(1) \\ y(0) \end{pmatrix}\right)|F^y_{-1}\right] \\
&= E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x(1) \\ y(0) \end{pmatrix}\right)\right], \text{ because } F^y_{-1} = \{\emptyset, \Omega\}, \\
&= \exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} A(0)\hat{x}_0 \\ C(0)\hat{x}_0 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T H(0) \begin{pmatrix} w_x \\ w_y \end{pmatrix}\right), \\
& \quad \text{because of Theorem 2.8.3 and } x_0 \in G(m_0, Q_0),\ v(0) \in G(0, I).
\end{aligned}$$

From Theorem 2.8.3 and equation (8.36) follows equation (8.37) for $t = 0$. Suppose that the equations (8.36,8.37) hold for $t \in T$. They will be proven to hold for $t + 1$. Note that,

$$\begin{aligned}
& E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix}\right)|F^y_{t-1}\right] \\
&= E\left[E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix}\right)|F_{t-1}\right]|F^y_{t-1}\right], \text{ because } F^y_{t-1} \subseteq F_{t-1}, \\
&= E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} A(t) \\ C(t) \end{pmatrix} x(t)\right)|F^y_{t-1}\right] \\
& \quad \times \exp\left(-\frac{1}{2}\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} M(t) \\ N(t) \end{pmatrix}\begin{pmatrix} M(t) \\ N(t) \end{pmatrix}^T \begin{pmatrix} w_x \\ w_y \end{pmatrix}\right), \\
& \quad \text{because of (2) of the proof,} \\
&= \exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} A(t) \\ C(t) \end{pmatrix}\hat{x}(t) - \frac{1}{2}\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T H(t) \begin{pmatrix} w_x \\ w_y \end{pmatrix}\right),
\end{aligned}$$

because of (8.36) of the induction step and of Proposition 19.7.4.

$$\begin{aligned}
& E[\exp(iw_x^T x(t+1))|F^y_{t-1}] = E[\exp(iw_x^T x(t+1))|F^{y(t)} \vee F^y_{t-1}] \\
&= \exp(iw_x^T[A\hat{x}(t) + H_{12}(t)H_{22}^{-1}(y(t) - C\hat{x}(t))) \times \\
& \quad \times \exp(-\frac{1}{2}w_x^T[H_{11}(t) - H_{12}(t)H_{22}^{-1}(t)H_{12}^T(t)]w_x), \\
&= \exp\left(iw_x^T\hat{x}(t+1) - \frac{1}{2}w_x^T Q_f(t+1)w_x\right),
\end{aligned}$$

where the successive steps are explained by the arguments: the induction step for $t \in T$, a result for conditional Gaussian random variables, see Section 19.7; with equation (8.37) and the equations (8.31,8.33).

By the principle of induction the equations (8.36,8.37) then hold for all $t \in T$. Then (a) and (b) are proven.
(c) From (a) follows that the conditional distribution is Gaussian and from this and Proposition 19.7.4, follows (c).
(d) From Step (3) and equation (8.36) follows that,

$$\begin{aligned} E\left[\exp(iw_y^T y(t))|F_{t-1}^y\right] &= \exp(iw_y^T C(t)\hat{x}(t) - \frac{1}{2}w_y^T Q_{\bar{v}}(t)w_y), \\ E\left[\exp(iw_y^T \bar{v}^T(t))|F_{t-1}^y\right] &= \exp(-\frac{1}{2}w_y^T Q_{\bar{v}}(t)w_y). \end{aligned}$$

Because $Q_{\bar{v}}$ is a stochastic process, the process $\bar{v}$ is in general not a white noise process.
(e) By induction it will be proven that $F_t^{\bar{v}} = F_t^y$ for all $t \in T$. Because $\bar{v}(0) = y(0) - C(0)\hat{x}_0$ and $\hat{x}_0 = m_0 \in \mathbb{R}^n$, $F_0^{\bar{v}} = F_0^y$. Suppose that for all $s,t \in T$, $s \leq t$, $F_s^{\bar{v}} = F_s^y$. Then $\bar{v}(t) = y(t) - C(t)\hat{x}(t)$ is F_t^y measurable, hence $F_t^{\bar{v}} \subseteq F_t^y$. Because $y(t) = \bar{v}(t) + C(t)\hat{x}(t)$, $\hat{x}(t)$ is F_{t-1}^y measurable, and, by the induction step, $F_{t-1}^y = F_{t-1}^{\bar{v}}$, $y(t)$ is $F_{t-1}^{\bar{v}}$ measurable, hence $F_t^y \subseteq F_t^{\bar{v}}$. □

8.10 Exercises

Problem 8.10.1. Observability versus detectability.
Consider a time-invariant Gaussian system and its associated time-invariant Kalman filter.

(a) Explain the difference between the following two statements:

(a.1) The system is observable, say (A,C) is an observable pair.
(a.2) The system is detectable, say (A,C) is a detectable pair.

(b) Write down the recursion for the error system of the filter in such a way that the unobservable part of the system is explicitly displayed.
(c) Suppose that you are asked to advise an engineer on the choice between the following two statements:

(c.1) A Gaussian system with (A,C) a detectable pair but not an observable pair; and
(c.2) Acquiring at a cost another sensor after which the new set of system matrices (A_{new}, C_{new}) is an observable pair.

Which arguments would you use for your advice?

Problem 8.10.2. Design of a filter. Consider a scalar time-invariant Gaussian system with $n_x = 1$, $n_y = 1$, and $n_v = 1$, denoted by

$$\begin{aligned} x(t+1) &= ax(t) + mv(t),\ x(0) = x_0, \\ y(t) &= cx(t) + nv(t),\ v(t) \in G(0,q_v),\ |a| < 1. \end{aligned}$$

Consider the simple filter,

$$\bar{x}(t+1) = a\bar{x}(t) + k[y(t) - c\bar{x}(t)],\ \bar{x}(0) = \bar{x}_0 = E[x_0] = m_0,\ k \in \mathbb{R}.$$

The problem is: How to choose k?

(a) Write down a recursion for the variance of the filter error

$$q_f(t) = E[(x(t) - \bar{x}(t))^2].$$

(b) Discuss the advantages and the disadvantages of the following choices for k:

(b.1) k very small;
(b.2) $a - kc$ almost zero;
(b.3) the choice of the time-invariant Kalman filter.

Problem 8.10.3. Robustness of the Kalman filter. Consider a time-invariant Gaussian system to be called the *real system,*

$$\begin{aligned} x_r(t+1) &= A_1 x_r(t) + M_1 v_r(t),\ x_r(0) = x_{r,0}, \\ y_r(t) &= C_1 x_r(t) + N_1 v_r(t),\ v_r(t) \in G(0, Q_{v_r}). \end{aligned}$$

Consider a model for the above system in the form of a time-invariant Gaussian system called the *imaginary system*,

$$\begin{aligned} x_i(t+1) &= A_2 x_i(t) + M_2 v_i(t),\ x_i(0) = x_{i,0}, \\ y_i(t) &= C_2 x_i(t) + N_2 v_i(t),\ v_i(t) \in G(0, Q_{v_i}). \end{aligned}$$

Consider the time-invariant Kalman filter for the imaginary system
which filter is assumed to be well defined,

$$\hat{x}_i(t+1) = A_2 \hat{x}_i(t) + K_2[y_i(t) - C_2 \hat{x}_i(t)],\ \hat{x}_i(0) = m_{i,0}.$$

Assume that the filter for the imaginary system is now provided the output of the real system, hence the real filter is represented by the recursion,

$$\bar{x}_r(t+1) = A_2 \bar{x}_r(t) + K_2[y_r(t) - C_2 \bar{x}_r(t)],\ \bar{x}_r(0) = m_{i,0}.$$

Define the *error* between the state of the real system and that of the time-invariant Kalman filter as $e(t) = x_r(t) - \bar{x}_r(t)$.

(a) Derive the recursion for the error system of the error process defined above.
(b) Which conditions do you recommend to keep the error small? Focus attention on the stability properties of the error system derived in (a).

8.11 Further Reading

History. The filter problem was apparently first formulated, solved, and published by A. Kolmogorov in 1941, [23], and, supposedly independently, by Norbert Wiener in a book published in 1949, [44]. The results of Wiener were derived during the second world war and not published during that period. See also the papers [45, 46]. The Levinson filter algorithm and its relation with the Wiener filter problem are described in [24, 25]. For text books with the topic of Wiener filtering see [16, 47, 48].

The Kalman filter was published for discrete-time Gaussian systems by R.E. Kalman, [17], in 1960 for a least-squares setting. The filter problem of a continuous-time Gaussian system was published by R.E. Kalman and R.S Bucy in 1961, [21]. Articles with a review of Kalman filtering by R.E. Kalman are [18, 19, 20].

Books on the Kalman filter. Books on the Kalman filter there are many, for example [30, 15]. The viewpoint of time series models, [10]. In the French language, [7, Ch. 8],

The square-root implementation of a time-invariant Kalman filter is described in the book by G.J. Bierman, [4].

For the history of the development and the application of Kalman filters in aerospace, [28, 32, 40]. For a text book on applications of Kalman filters and its extensions to problems of tracking and navigation, see [3].

The March 1983 issue of the journal *IEEE Transactions on Automatic Control* presents papers on applications of Kalman filters.

Kalman filters and stochastic realization. [18, 25, 24, 37, 38].

Prediction. [12, 36]. *Smoothing*. [6, 29, 31]. *Interpolation*. [14, 13].

The analogue of the Kalman filter can also be formulated for a linear system defined in a finite field as used in information theory and in communication theory, [42, 8]

For deterministic linear systems the concept of an observer has been formulated by D.G. Luenberger based on the formulas of the Kalman filter, [26, 27] and [49].

References

1. Mark Asch, Marc Bocquet, and Maëlle Nodet. *Data assimilation: Methods, algorithms, and applications*. SIAM, Philadelphia, PN, USA, 2017. 298
2. S.G. Bankoff and E.L. Hanzevack. The adaptive-filtering transport model for prediction and control of pollutant concentration in an urban airshed. *Atmospheric Environment*, 9:793–808, 1975. 282
3. Y. Bar-Shalom, X.R. Li, and T. Kirubarajan. *Estimation with applications to tracking and navigation*. John Wiley & Sons Inc., New York, 2001. 310, 353
4. G.J. Bierman. *Factorization methods for discrete sequential estimation*. Academic Press, New York, 1977. 289, 310, 833, 849, 850
5. T. Bohlin. Four cases of identification of changing systems. In R.K. Mehra and D.G. Lainiotis, editors, *System identification - Advances and case studies*, pages 441–518. Academic Press, New York, 1976. 78, 282
6. P.E. Caines. *Linear stochastic systems*. John Wiley & Sons, New York, 1988. 120, 276, 302, 310, 575
7. P. Faurre, M. Clerget, and F. Germain. *Opérateurs rationnels positifs*. Dunod, Paris, 1979. 175, 180, 217, 275, 292, 310, 850, 865, 867, 877, 885
8. G. David Forney Jr. The Viterbi algorithm. *Proc. IEEE*, 61:268–278, 1973. 291, 310

9. D. Godard. Channel equalization using a Kalman filter for fast data transmission. *Communication*, pages 267–273, 1974. 282
10. A.C. Harvey. *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press, Cambridge, UK, 1989. 310
11. A.W. Heemink. *Storm surge prediction using Kalman filtering*. Thesis, Twente University, Enschede, 1986. 78, 282
12. J. Holst. Adaptive prediction and recursive estimation. Ph.d. thesis, Lund Institute of Technology, Lund, 1977. 282, 310
13. A.J.E.M. Janssen, R.N.J. Veldhuis, and L.B. Vries. Adaptive interpolation of discrete-time signals that can be modelled as autoregressive processes. *IEEE Trans. Acoustics, Speech & Signal Processing*, 34:317–330, 1986. 78, 303, 310
14. A.J.E.M. Janssen and L.B. Vries. *Interpolation of band-limited discrete-time signals by minimizing out-of-band energy*, page Paper 12B.2. IEEE, 1984. 303, 310
15. A.H. Jazwinski. *Stochastic processes and filtering theory*. Academic Press, New York, 1970. 310, 353
16. W.B. Davenport Jr. and W.L. Root. *An introduction to the theory of random signals and noise*. McGraw-Hill Book Co., New York, 1958. 9, 72, 310
17. R.E. Kalman. A new approach to linear filtering and prediction problems. *J. Basic Eng.*, 82:35–45, 1960. 283, 310, 808
18. R.E. Kalman. New methods in Wiener filtering theory. In J.L. Bogdanoff and F. Kozin, editors, *Proceedings 1st Symposium Engineering Applications of Random Function Theory and Probability*, pages 270–388, New York, 1963. Wiley. 175, 275, 310
19. R.E. Kalman. Linear stochastic filtering - Reappraisal and outlook. In J. Fox, editor, *Proceedings Symposium on System Theory*, pages 197–205, New York, 1965. Polytechnic Press. 217, 310
20. R.E. Kalman. A retrospective after twenty years: from the pure to the applied. In Chao-Lin Chiu, editor, *Applications of Kalman filter to hydryology, hydraulics, and water resources*, pages 31–54. University of Pittsburgh, Pittsburgh, 1978. 310
21. R.E. Kalman and R.S. Bucy. New results in linear filtering and prediction theory. *J. Basic Eng.*, 83:95–107, 1961. 310
22. S.B. Kleibanov, V.B. Prival'skii, and I.V. Tinre. Kalman filtering for equalizing of digital communications. *Automation and Remote Control*, 35:1097–1102, 1974. 282
23. A.N. Kolmogorov. Interpolation and extrapolation of stationary random sequences (in russian). *Bull. Moscow University, USSR, Ser. Math.*, 5:3–14, 1941. 282, 309
24. N. Levinson. A heuristic exposition of Wiener's mathematical theory of prediction and filtering. *J. Mathematics and Physics*, 26:110–119; see also Appendix C. in the book titled, 1947. 291, 310
25. N. Levinson. The Wiener rms (root mean square) error criterion in filter design and prediction. *J. Mathematics and Physics*, 25:261–278; see also Appendix B. in the book titled, 1947. 291, 310
26. D.G. Luenberger. Observing the state of a linear system. *IEEE Trans. Military Electronics*, 23:119–125, 1964. 310
27. D.G. Luenberger. Observers for multivariable systems. *IEEE Trans. Automatic Control*, 11:190–197, 1966. 310
28. Leonard A. McGee and Stanley F. Schmidt. Discovery of the Kalman filter as a practical tool for aerospace and industry. Report 86847, NASA, Moffett Field, CA, 1985. 310
29. J.S. Meditch. On optimal linear smoothing theory. *J. Info. Control*, 10:598–615, 1967. 310
30. J.S. Meditch. *Stochastic optimal estimation and control*. McGrawHill, New York, 1969. 310, 467
31. J.S. Meditch. A survey of data smoothing for linear and nonlinear dynamic systems. *Automatica*, 9:151–162, 1973. 310
32. NASA. Discovery of the kalman filter as a practical tool for aerospace and industry. Report, NASA, Langley, 1985. 310

33. Jana Němcová, Mihály Petreczky, and Jan H. van Schuppen. Rational observers of rational systems. In *Proc. 55th IEEE Conference on Decision and Control (CDC.2016)*, pages 6252–6257, New York, 2016. IEEE, IEEE Press. 291

34. Michele Pavon. Optimal interpolation for linear stochastic systems. *SIAM J. Control & Opt.*, 22:618–629, 1984. 303

35. J.B. Pearson. Kalman filter applications in airborne rader tracking. *IEEE Trans. Aerospace and Electronic Systems*, 10:319–329, 1974. 282

36. J. Rissanen. A fast algorithm for optimal predictors. *IEEE Trans. Automatic Control*, 18:555, 1973. 310

37. G. Ruckebusch. *Répresentations markoviennes de processus gaussiens stationnaires*. PhD thesis, Thèse de 3e cycle, Paris VI, Paris, 1975. 310

38. G. Ruckebusch. Représentations markoviennes de processus gaussiens stationnaires et applications statistiques. Rapport interne 18, Ecole Polytechnique, Centre de Mathématiques Appliquées, 1977. 120, 175, 248, 253, 275, 291, 310

39. Y. Sawaragi, T. Soeda, H. Tamura, T. Yoshimura, S. Ohe, Y. Chujo, and H. Ishihara. Statistical prediction of air pollution levels using non-physical models. *Automatica*, 15:441–451, 1979. 282

40. S.F. Schmidt. The Kalman filter: Its recognition and development for aerospace applications. *AIAA J. Guidance Control*, 4:4, 1981. 310

41. J.D. van der Bij and J.H. van Schuppen. Prediction of railway power demand. *Automatica J. IFAC*, 19:487–494, 1983. 282

42. A.J. Viterbi. Error bounds for convolutional codes and an asymptotic optimum decoding algorithm. *IEEE Trans. Information Theory*, 13:260–269, 1967. 291, 310

43. Yubin Wang, Jan H. van Schuppen, and Jos Vrancken. Prediction of traffic flow at the boundary of a motorway network. *IEEE Trans. Intelligent Transportation Systems*, 15:214–227, 2014. 282

44. N. Wiener. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. Technology Press of the MIT, Cambridge, MA, U.S.A., 1949. MIT Press, Cambridge, MA, U.S.A., 1966. 72, 310

45. N. Wiener and P. Masani. The prediction theory of multivariate stochastic processes – part i. *Acta Math.*, 98:111–150, 1957. 310

46. N. Wiener and P. Masani. The prediction theory of multivariate stochastic processes – part ii. *Acta Math.*, 99:93–137, 1958. 72, 310

47. E. Wong. *Stochastic processes in information and dynamical systems*. McGraw-Hill Book Co., New York, 1971. 72, 73, 310, 352

48. E. Wong and B. Hajek. *Stochastic processes in engineering systems*. Springer-Verlag, Berlin, 1985. 73, 310, 352

49. W.M. Wonham. *Linear multivariable control: A geometric approach*. Springer-Verlag, Berlin, 1979. 310, 779

Chapter 9
Filtering of Stochastic Systems

Abstract Filter problems are formulated for stochastic systems which are not Gaussian systems. Both the estimation problem, the sequential estimation problem, and the filter problem are treated. A sufficient condition for the existence of a finite-dimensional filter system is formulated. The concept of a family of invariant conditional distributions is defined. It is described how to solve the filter problem by a measure transformation method. Cases treated include the Poisson-Gamma filter and the filter of an output-finite-state-finite stochastic system.

Key words: Filter problem. Poisson-Gamma filter.

The approach to the filter problem is the determination of a stochastic realization of the output process of a stochastic system, in which the stochastic realization has a state process which is measurable with respect to the past of the output process and a finite-dimensional or finite state set.

9.1 Problems of Estimation, Sequential Estimation, and of Filtering

The *filter problem* is to determine for a stochastic system with partial observations, the conditional distribution of the state conditioned on the past observations. A formal definition follows below in this section.

Filter problems arise for example in engineering, in economics, and in the life sciences. The filter problem is often motivated by a related control problem. But it can also be based on a prediction, smoothing, or interpolation problem.

In the literature the filter problem is often described as a nonlinear filter problem. More precisely, a filter problem for a stochastic system of which the equations are not necessarily linear in terms of the state. But it also covers the case where the stochastic systems does not have Gaussian probability distributions even if the system is linear in the state. So as to avoid confusion between the many cases, the term *nonlinear filter problem* will not be used further in this book.

Problem 9.1.1. *Filter problem for a stochastic system.* Consider a stochastic system with partial observations (without inputs) described by the mathematical objects,

$$\begin{aligned}&((\Omega,F,P),\ T,(X,B(X)),(Y,B(Y)),\ x:\Omega\times T\to X,\ y:\Omega\times T\to Y),\\&\text{cpdf}((x(t+1),y(t))|F_t^x\vee F_{t-1}^y)=\text{cpdf}((x(t+1),y(t))|F^{x(t)}),\ \forall\ t\in T.\end{aligned}$$

Note that the latter condition of the above displayed formulas is the condition which defines a stochastic system, see Def. 4.2.2. Call x the *state process* and y the *output process* of the stochastic system.

The *filter problem* is to determine, for every time $t\in T$, the conditional distribution of the state $x(t+1)$, conditioned on the past of the observations, F_t^y,

$$\text{cpdf}(x(t+1)|F_t^y),\ \ \forall\ t\in T. \tag{9.1}$$

Below the term *conditional distribution* refers to that particular conditional distribution specified above. In particular cases, one determines the conditional characteristic function in stead of the conditional probability distribution because the two are bijectively related.

Equivalently, determine a strong stochastic realization of the output process of which the state is a function of the past outputs; thus $F^{\hat{x}(t)}\subset F_{t-1}^y$ for all $t\in T$. See Subsection 8.4 for a discussion of this equivalence.

Recall from Theorem 8.3.2 that for a Gaussian system with partial observations, the above defined conditional distribution is Gaussian with a conditional mean and a conditional variance denoted by $(\hat{x}(t+1),Q_f(t+1))$. Equivalently, it is described by the conditional characteristic function,

$$E[\exp(iw_x^Tx(t+1))|F_t^y]=\exp(iw_x^T\hat{x}(t+1)-\frac{1}{2}w_x^TQ_f(t+1)w_x),\ \forall\ w_x\in\mathbb{R}^{n_x}.$$

For a general stochastic system as defined above, it is not clear to which subset of probability distributions, the conditional probability distribution will belong.

The reader is alerted of the fact that in the literature outside control theory, for example in the area of operations research and of signal processing, the following problem has been investigated where one wants to determine,

$$E[\exp(iw_x^Tx(t))|F_t^y],\ \forall\ w_x\in\mathbb{R}^{n_x}.$$

The formulas of this case are different than those of Problem 9.1.1. The control theoretic literature has standardized on the representation of a stochastic system and on the filter problem as stated in Problem 9.1.1.

The reader is informed that in the literature there exist side by side two approaches to estimation and filtering problems: (1) the *conditional distribution approach*: the approach to determine the conditional probability distribution of the state based on past observations as used in Problem 9.1.1. (2) the *optimization approch*: in which one determines a function in terms of the observations, in a prespecified set of functions for example linear functions, which best approximates the state variable, or a function of the state, or the conditional distribution, according to a loss function. The optimization approach was advanced by the book of A. Wald, [42], which is a major source of statistical decision theory. However, in this book and in this chapter, the conditional distribution approach to estimation problems is used.

Not much progress can be made on this problem without further specification of the stochastic system which generates the state and the output process.

The filter problem is best seen as an extension of the estimation problem and the sequential estimation problem. The latter problems have been investigated in the research areas of statistics, probability, and statistical decision theory for a long while. The specifications of the estimation model, the sequential estimation model, and the filter problem follow. Subsequently the corrsponding problems are formulated. In the subsequent sections of this chapter the reader finds successively: estimation – theory and examples, sequential estimation – theory and examples, and filter theory and examples.

Definition 9.1.2. Define the following models.

(a) Define the *estimation model* by the mathematical objects and relations,

$$x : \Omega \to X,\ y : \Omega \to Y,\ (\text{cpdf}(y|F^x),\ \text{pdf}(x)).$$

One calls the random variable x the *state* and the random variable y the *output* or the *observation* of the estimation model.

(b) Define the *sequential estimation system* by the mathematical objects and relations,

$$x : \Omega \to X,\ y : \Omega \times T \to Y,\ \ \left(\text{cpdf}(y(t)|F^x \vee F^y_{t-1}), \text{pdf}(x)\right),\ \forall\, t \in T.$$

One calls the random variable x the *state* and the stochastic process y the *output process* or the *observation process* of the sequential estimation system. In the above, the understanding is that for time $t = 0$ the conditional distribution of the problem statement is $\text{cpdf}(y(0)|\ F^x)$.

(c) Denote a *stochastic system* as in Problem 9.1.1 and denote the mathematical objects and relations by,

$$x : \Omega \times T \to X,\ y : \Omega \times T \to Y,$$
$$\text{cpdf}(x(t+1),\ y(t))|F^x_t \vee F^y_{t-1}) = \text{cpdf}(x(t+1),\ y(t))|F^{x(t)}),\ \forall\, t \in T.$$

Problem 9.1.3. Consider the estimation model, the sequential estimation model, and the stochastic system of Def. 9.1.2.

(a) The *estimation problem* for an estimation model is to determine the conditional probability distribution of the state conditioned on the observation,

$$\text{cpdf}(x|F^y). \tag{9.2}$$

(b) The *sequential estimation problem* for a sequential estimation system is to determine the conditional probability distribution of the state conditioned on the past observations,

$$\text{cpdf}(x|F^y_{t-1}),\ \ \forall\, t \in T\backslash\{0\}. \tag{9.3}$$

(c) The *filter problem* is to determine the conditional probability distribution of the current state conditioned on the past observations,

$$\text{cpdf}(x(t)|F^y_{t-1}), \ \ \forall\, t \in T\backslash\{0\}. \tag{9.4}$$

9.2 Finite-Dimensional Filter Systems

The reader who has learned about the Kalman filter will not have failed to notice that the usefulness of the Kalman filter is based on the property that the conditional distribution of the filter problem is for every time a Gaussian conditional probability distribution, that this distribution has only two parameters, and that the parameters of the conditional distribution, $(\hat{x}(t), Q_f(t))$, can be recursively calculated. Therefore this concept should be used to develop a theory for the filter problem. Basically this concept has to be combined with stochastic realization theory.

Definition 9.2.1. Consider the filter problem 9.1.1. The filter problem is said to admit a *finite-dimensional filter system* if there exists a finite-dimensional system, to be called the *filter system*, which characterizes the conditional distribution according to,

$$n_f \in \mathbb{Z}_+,\ X_f \subset \mathbb{R}^{n_f},\ x_f : \Omega \times T \to X_f,$$

$$P(\{x(t) \in A\}|F^y_{t-1}) = p_f(A, x_f(t), y(t-1)),\ A \in B(X_f), \tag{9.5}$$

$$x_f(t+1) = f_{fs}(x_f(t), y(t)),\ x_f(0) = x_{f,0} \in X_f, \tag{9.6}$$

Call x_f the *state of the filter system*. A direct consequence of equation (9.5) is that for all $t \in T$, $x_f(t)$ is F^y_{t-1} measurable.

Comments on the above definition follow. A simple attempt to define a filter system is to take as the state of a filter system the vector of past outputs,

$$x_f(t) = \begin{pmatrix} y(t-1) \\ y(t-2) \\ \vdots \\ y(0) \end{pmatrix}.$$

But the dimension of that vector $x_f(t)$ grows to infinity when time progresses to infinity, $t \to \infty$. Thus that approach will not produce a finite-dimensional filter system. Any computer which implements a filter system has finite memory though it can be large, and therefore the restriction to a finite-dimensional filter system is motivated.

A second point to notice is that the function p_f describing the conditional distribution may not depend on time. Therefore, the conditional distribution is of the same analytic form for all times. The researcher J. Bather, in the paper [7], has termed this property that of an *invariant conditional distribution* though, more accurately, that of an *invariant set of conditional probability distributions*. This concept is useful for filter theory. This restriction is imposed to limit the complexity of the filter system.

If the conditional distribution belongs at each time to a different subset of probability distributions, often totally disjoint from that of other time moments, then the complexity of the filter system becomes large and that is not practical for control engineering.

The operation of the filter system can be seen as described next. At time $t = 0$, the current state of the filter system $x_f(0)$ and the current output $y(0)$ determine by the function p_f the conditional distribution of $x_f(1)$ conditioned on F_0^y according to equation (9.5). The next state $x_f(1)$ of the filter system is then calculated by equation (9.6). Once $y(1)$ is observed, the one calculates the conditional distribution $\text{cpdf}(x(2)|\,F_1^y)$ by the function p_f and the next state $x_f(2)$ of the filter system by f_{fs}. Similarly for any time $t \in T$, the conditional distribution is determined by equation (9.5) and the next state $x_f(t+1)$ of the filter system by equation (9.6).

The reader may think of the Kalman filter as an illustration of a finite-dimensional filter system. In the Kalman filter, the conditional distribution is determined by the conditional mean and the conditional variance, denoted by $(\hat{x}(t), Q_f(t))$, which combined is the state of the filter system.

Problem 9.2.2. Consider the Filter Problem 9.1.1. Which stochastic systems admit a finite-dimensional filter system? Derive sufficient and necessary conditions on the considered stochastic system for the existence of a finite-dimensional filter system.

The concept has been motivated above.

Definition 9.2.3. Consider the filter problem of Problem 9.1.3. Call a set $F_{cinvpdf}$ the *set of invariant conditional distribution functions* of the filter problem if the conditional probability distributions,

$$\text{cpdf}(x(t+1)|\,F_{t-1}^y),\ \forall\, t \in T\backslash\{0\},$$

belong for all times to the considered set $F_{cinvpdf}$.

There exist algebraic conditions of a stochastic system which imply the existence of an invariant conditional distribution. These conditions are explored in this chapter.

The existence of a finite-dimensional filter system of Def. 9.2.1 is conjectured to depend on the following two sufficient conditions:

1. that there exists a set of invariant conditional probability distributions for the problem;
2. that the set of invariant conditional probability distribution functions is characterized by a finite number of parameters.

Then the vector of parameters of the obtained set of the conditional probability distribution functions is a candidate for the state of the filter system.

It will be described later in this chapter for several stochastic systems that the following two conditions are sufficient conditions for the existence of a finite-dimensional filter system:

1. that the conditional distribution of the current observation conditioned on the current state, and the unconditional distribution of the state, form *a pair of conjugate*

probability distributions (see Section 9.2); the specification includes which parameter of the conditional distribution of the output and the state, is regarded as the state variable; and

2. that the conditional probability distribution of the next state conditioned on past states and past outputs has a representation in terms of the current state which is of the same analytic form as that of the variable of the conditional probability distribition function as mentioned in the previous condition.

So far there are few stochastic systems for which a finite-dimensional filter system have been proven to exist. The primary case for which a finite-dimensional filter exist, is that of the Gaussian system with partial observations which admits the Kalman filter. Additional stochastic systems which admit a finite-dimensional filter system are: a Poisson-Gamma system and a output-finite-state-finite stochastic system.

The family of exponential probability measures is often mentioned in regard to the existence of finite-dimensional filter systems. But that family satisfies only the second condition listed above, it is stable with respect to addition of independent random variables of which the random variables have a probability distribution in the set of exponential probability distribution functions. In general it does not satisfy the first condition.

Filter Theory with Estimation-Conjugate Probability Distributions

In the theory of estimation and filtering, the concept of estimation-conjugate probability function plays a major role. The concept is mentioned in the book of H. Raiffa and R. Schlaifer, [32]. Below the term *estimation-conjugate* probability distribution functions and *filter-conjugate* probability distribution functions are used because there also exists a term *control-conjugate functions*.

In this section the concept of estimation-conjugate probability distributions is defined. In the Sections 9.3, 9.4, and 9.6, the reader finds particular examples.

Examples of conjugate distributions are provided in Table 9.1. That these are indeed such pairs is proven in Section 9.3 for the indicated special cases.

Y	$f(.;y\mid x)$	State	X	$f(.;x)$	$f(.;x\mid y)$
	$E[\exp(iw_y^T y)\mid F^x]$	x		$E[\exp(iw_x^T x)]$	$E[\exp(iw_x^T x)\mid F^y]$
$\mathbb{N}_m$	Bernouli (q)	q	$(0,1)$	Beta	Beta
$\mathbb{N}$	Poisson (λ)	λ	$\mathbb{R}_+$	Gamma	Gamma
$\mathbb{R}_+$	Gamma (γ_1,γ_2)	γ_1	$\mathbb{R}_+$	Gamma	Gamma
$\mathbb{R}$	Gaussian (m,q)	m	$\mathbb{R}$	Gaussian	Gaussian
$\mathbb{R}$	Gaussian (m,q)	q	$\mathbb{R}_+$	Gamma	Gamma

Table 9.1 Table of pairs of conjugate probability distributions and their associated conditional distribution.

Definition 9.2.4. Consider a tuple consisting of: (1) a conditional probability distribution representing the conditional distribution of the output conditioned on the state variable, and select a particular parameter of that distribution function; and (2) a probability distribution representing the unconditional probability distribution of the state variable which appears as the selected parameter of the probability distribution function in (1).

The tuple is called a *pair or a tuple of estimation-conjugate (probability) distributions* if the unconditional a priori probability distribution of the state belongs to the same family of probability distribution as the a-posterori conditional probability distribution of the state conditioned on the past observations. The parameters of the apriori probability distribution and those of the a posteriori probability distribution will in general be different.

In terms of mathematical notation, consider a probability space (Ω, F, P) and random variables $x : \Omega \rightarrow X \subseteq \mathbb{R}^{n_x}$ and $y : \Omega \rightarrow Y \subseteq \mathbb{R}^{n_y}$ for $n_x,\ n_y \in \mathbb{Z}_+$. The ordered pair of a conditional characteristic function and of an unconditional characteristic function,

$$(E[\exp(iw_y^T y)|F^x] = f(x, p_2, \ldots, p_k),\ E[\exp(iw_x^T x)]),$$

is called a *pair of estimation-conjugate characteristic functions* if the characteristic functions,

$$E[\exp(iw_x^T x)|F^y],\ E[\exp(iw_x^T x)],$$

belong to the same subset of characteristic functions. In terms of probability density functions, the tuple is a pair of conjugate probability density functions if $(p(.;y|x)$ and $p(.;x))$ belong to the same subset of probability density functions.

The concept of conjugate probability distributions was introduced in the book of H. Raiffa and R. Schlaifer, for the context of sequential estimation problems, see [32]. See also its recent edition, [30]. It is not clear to the author whether for any conditional distribution $f(.;y|x)$ and any parameter of that distribution which equals the random variable x, there exists a probability distribution which forms a pair of conjugate distributions.

9.3 Estimation Theory

The estimation problem and its solution method are briefly described in words before the mathematical details. The results of this section are mostly known and published though the formulation of this section is more formal than the literature.

Consider two random variables, the *observed random variable* or *output y* and the unobserved random variable x, often referred to as the *state* of the estimation model. The estimation model is completely specified by two probability distributions: (1) the conditional distribution of the output conditioned on the state and (2) the unconditional distribution of the state variable.

The estimation problem, Problem 9.1.3, is to determine the conditional distribution of the state variable conditioned on the output variable.

The solution procedure of the estimation problem presented in this chapter requires as the first step the construction of a new probability measure, absolutely continuous with respect to the probability measure of the problem formulation, such that with respect to the new probability measure the output and the state are independent. The conditional distribution of the state conditioned on the output can then be calculated as a conditional expectation of the product of the state and of the associated Radon-Nikodym derivative conditioned on the output.

Readers not familiar with the transformation of probability measures are advised to read Section 19.9 before proceeding. A short summary of the theory of measure transformations is provided below.

Consider a measurable space (Ω, F) and two probability measures P_0, P_1 defined on it. The probability measure P_1 is said to be *absolutely continuous* with respect to P_0 if for all $A \in F$, $P_0(A) = 0$ implies that $P_1(A) = 0$. This is denoted by $P_1 \ll P_0$. The two probability measures are said to be *equivalent* if they are mutually absolutely equivalent; equivalently, if $P_1 \ll P_0$ and $P_0 \ll P_1$. This is denoted by $P_1 \sim P_0$.

The relation between two absolutely continuous measures is fully characterized by their Radon-Nikodym derivative as described next, see Theorem 19.9.4. Let (Ω, F, P_0) be a probability space. Then P_1 is a probability measure on (Ω, F) and $P_1 \ll P_0$ on F if and only if,

$$\begin{aligned} &\exists\, r_{1|0} : \Omega \to \mathbb{R}_+, \text{ such that } E_0[r_{1|0}] = 1, \\ P_1(A) = E_0[r_{1|0}\, I_A] = &\int_\Omega I_A(\omega)\, r_{1|0}(\omega) dP_0(\omega), \;\; \forall\, A \in F; \\ &\text{which definition is denoted by } dP_1/dP_0 = r_{1|0}. \end{aligned}$$

The real-valued random variable $r_{1|0}$ is called the *Radon-Nikodym derivative* of P_1 with respect to P_0. When P_1 is specified then the random variable $r_{1|0}$ is unique upto an almost sure modification. If in addition $P_1 \sim P_0$ then $r_{1|0} > 0$ almost surely P_0 and P_1 and then also $dP_0/dP_1 = r_{0|1} = r_{1|0}^{-1}$. Correspondingly, expectation is denoted by E_1 with respect to P_1 and by E_0 with respect to P_0.

The usefulness of the above result is that, instead of working with a probability measure P_1, one can work with the associated Radon-Nikodym derivative $r_{1|0}$ which is a real-valued random variable and therefore easier to analyse.

See Section 9.3, for the way to construct the Radon-Nikodym derivative in particular cases.

In the remainder of this chapter the probability measure P_1 denotes the one used for the probability space of the model while the probability measure P_0 denotes the new measure to be constructed in each case. Below the estimation model and the estimation problem are repeated for ease of reference.

Definition 9.3.1. The *estimation model*. Consider a probability space (Ω, F, P_1) and two measurable spaces $(X, B(X))$ and $(Y, B(Y))$ where $X \subseteq \mathbb{R}^{n_x}$ and $Y \subseteq \mathbb{R}^{n_y}$ for n_x, $n_y \in \mathbb{Z}_+$, and the associated σ-algebras are the Borels sets of the indicated sets. Define two random variables, $x : \Omega \to X$ and $y : \Omega \to Y$. The *estimation model* is completely specified by the two probability distributions, or the characteristic functions,

$$f(.;y|x), \text{ or, equivalently, } E_1[\exp(iw_y^T y)|F^x], \ \forall \, w_y \in \mathbb{R}^{n_y};$$
$$f(.;x), \text{ or, equivalently, } E_1[\exp(iw_x^T x)], \ \forall \, w_x \in \mathbb{R}^{n_x}.$$

Problem 9.3.2. The *state estimation problem* of an estimation model. Consider the estimation model of Def. 9.3.1. Determine the conditional distribution $p(.;x|y)$ or, equivalently, the conditional characteristic function,

$$E_1[\exp(iw_x^T x)|F^y] = f_e(y, w_x), \ \ \forall \, w_x \in \mathbb{R}^{n_x}. \tag{9.7}$$

Define the *estimator* as the function from the output y to the parameters of the conditional distribution or to the parameters of the conditional characteristic function.

Below a new measure P_0 is constructed such that the following two conditions both hold: (1) F^x and F^y are independent σ-algebras with respect to P_0; and (2) $P_1(A) = P_0(A)$ for all $A \in F^x$. In each application of the next theorem, the measure P_0 has to be constructed of which the particular definition depends on the estimation model.

Theorem 9.3.3. *Consider Def. 9.3.1 and Problem 9.3.2. Assume that there exists a random variable,*

$$\begin{aligned} r_{0|1} > 0, \ \ &\textit{a.s. } P_1, \ \ r_{0|1} : \Omega \to \mathbb{R}_+, \textit{ such that,} \\ E_1[r_{0|1}|F^x] &= 1, && (9.8) \\ E_1[\exp(iw_y^T y) r_{0|1}|F^x] &= E_1[\exp(iw_y^T y) r_{0|1}], \ \forall \, w_y \in \mathbb{R}^{n_y}. && (9.9) \end{aligned}$$

(a) 1. The formula $dP_0/dP_1 = r_{0|1}$ defines a probability measure P_0 on $(\Omega, F^x \vee F^y)$;

2. P_0 and P_1 are equivalent probability measures with $dP_1/dP_0 = r_{1|0} = r_{0|1}^{-1}$;

3. the σ-algebras F^x and F^y are independent with respect to the probability measure P_0;

4. the restrictions of P_1 and P_0 to the σ-algebra F^x are identical; equivalently, $\forall \, A \in F^x$, $P_0(A) = P_1(A)$;

5. $E_0[\exp(iw_y^T y)] = E_1[\exp(iw_y^T y) r_{0|1}]$ for all $w_y \in \mathbb{R}^{n_y}$.

(b)The procedure to solve the state estimation problem is provided by the formula,

$$E_1[\exp(iw_x^T x)|F^y] = \frac{E_0[\exp(iw_x^T x) r_{1|0}|F^y]}{E_0[r_{1|0}|F^y]}, \ \textit{a.s. } P_1. \tag{9.10}$$

The numerator of equation (9.10) can be obtained by integration over the probability distribution of x while treating the random variable y as a known variable, due to the property that, with respect to P_0, the σ-algebras F^x and F^y are independent by (a.3), and by Theorem 2.8.6. The denominator can be obtained from the numerator by setting $w_x = 0$.

Proof. Proof of Theorem 9.3.3,
(a.1) From condition (9.8) it follows by reconditioning that,

$$E_1[r_{0|1}] = E_1[E_1[r_{0|1}|F^x]] = 1.$$

Then the formula $dP_0/dP_1 = r_{0|1}$ defines a probability measure P_0 on (Ω, F) by Theorem 19.9.4.
(a.2) The assumption that $r_{0|1} > 0$ implies that the measures P_0 and P_1 are equivalent. It then follows from Theorem 19.9.4 that $r_{1|0} = r_{0|1}^{-1}$.
(a.3) If $A \in F^x$ then,

$$P_0(A) = E_0[I_A] = E_1[I_A r_{0|1}] = E_1[I_A E_1[r_{0|1}|F^x]] = E_1[I_A] = P_1(A),$$

because of condition (9.8).
(a.4) Then, for all $w_y \in \mathbb{R}^{n_y}$,

$$\begin{aligned} E_0[\exp(iw_y^T y)|F^x] &= \frac{E_1[\exp(iw_y^T y)r_{0|1}|F^x]}{E_1[r_{0|1}|F^x]}, \text{ by Theorem 19.9.10,} \\ &= E_1[\exp(iw_y^T y)r_{0|1}|F^x], \text{ by equation (9.8) and,} \\ &= E_1[\exp(iw_y^T y)r_{0|1}], \text{ by equation (9.9),} \\ &= E_0[\exp(iw_y^T y)], \end{aligned}$$

and with Theorem 2.8.2.(f) follows that, with respect to the probability measure P_0, the σ-algebras F^x and F^y are independent.
(a.5) This follows from the definition of $r_{0|1}$ and the definition of the probability measure P_0.
(b) This follows from the formula for conditional expectation under a measure transformation, see Theorem 19.9.10. In addition, the evaluation of the conditional expectation in the numerator follows from Theorem 2.8.6. □

Estimator Binomial-Beta

The model of this section has as output a random variable taking values either in the set $\mathbb{N}_1 = \{0,1\}$ or in the finite set $\mathbb{N}_{n_y} = \{0,1,\ldots,n_y\}$ for an integer $n_y \in \mathbb{Z}_+$. The state variable is the probability of obtaining the value 1 for the output variable. That state variable will have a Beta conditional probability distribution function when conditioned on the state.

For the benefit of the readers the two probability distributions are recalled. The *Binomial* probability frequency function with parameters $(n_y, p_1) \in \mathbb{Z}_+ \times (0,1)$ is defined by the equations,

$$\begin{aligned}
& n_y \in \mathbb{Z}_+, \; p_1 \in (0,1), \; \binom{n_y}{k} = \frac{n_y!}{k! \; (n_y - k)!}, \\
& y : \Omega \to \mathbb{N}_{n_y}, \text{ a random variable}, \\
p_y(k) &= \binom{n_y}{k} p_1^k \; (1-p_1)^{n_y - k}, \; p_y : \mathbb{N}_{n_y} \to \mathbb{R}_+; \\
& P(\{y=1\}) = p_y(1) = n_y p_1 (1-p_1)^{n_y - 1}, \\
m_y &= E[y] = n_y \; p_1, \\
q_y &= E[(y-m_y)^2] = n_y \; p_1(1-p_1), \\
c_y(w_y) &= E[\exp(iw_y y)] = (1 - p_1 + p_1 \exp(iw_y))^{n_y}.
\end{aligned}$$

The Beta probability density distribution with parameters $(\beta_1, \beta_2) \in (0,\infty)^2$ of a random variable x is defined by the formulas and relations,

$$\begin{aligned}
& x : \Omega \to (0,1), \\
p_x(v) &= v^{\beta_1 - 1} \; (1-v)^{\beta_2 - 1} / B(\beta_1, \beta_2), \; p_x : (0,1) \to \mathbb{R}_+, \\
B(\beta_1, \beta_2) &= \int_0^1 v^{\beta_1 - 1} \; (1-v)^{\beta_2 - 1} dv = \frac{\Gamma(\beta_1)\Gamma(\beta_2)}{\Gamma(\beta_1 + \beta_2 - 1)}, \\
m_x &= \frac{\beta_1}{\beta_1 + \beta_2}, \; q_x = \frac{\beta_1 \beta_2}{(\beta_1+\beta_2)^2 (\beta_1 + \beta_2 + 1)}, \\
c_x(w_x) &= \frac{\Gamma(\beta_1 + \beta_2)}{\Gamma(\beta_1)} \sum_{k=0}^{\infty} \frac{\Gamma(\beta_1 + k)}{\Gamma(\beta_1 + \beta_2 + k)} \frac{(iw_x)^k}{k!}, \\
\Gamma(\beta) &= \int_0^{\infty} v^{\beta - 1} \exp(-v) dv, \; \Gamma : (0,\infty) \to \mathbb{R}_+.
\end{aligned}$$

Definition 9.3.4. Define the *Binomial-Beta model* with representation,

$$\begin{aligned}
& (\Omega, F, P_1), \; y : \Omega \to \mathbb{N}_{n_y} = \{0, 1, \ldots, n_y\}, \; n_y \in \mathbb{Z}_+, \; x : \Omega \to (0,1), \\
& E_1[I_{\{y(\omega)=k\}} | F^x] = \binom{n_y}{k} x^k (1-x)^{n_y - k}, \; \forall \, k \in \mathbb{N}_{n_y}, \\
& x \text{ having a Beta-pdf with parameters, } (\beta_1, \beta_2) \in (0,\infty)^2.
\end{aligned}$$

The model has the following representation,

$$y = n_y \, x + v, \; E_1[v|F^x] = 0, \; v : \Omega \to \mathbb{R}.$$

The special case of $n_y = 1$ is called the *Bernoulli-Beta model* and it has the representation,

$$\begin{aligned}
& y : \Omega \to \mathbb{N}_1 = \{0,1\}, \\
E_1[I_{\{y(\omega)=k\}} | F^x] &= x^k (1-x)^{1-k} = \begin{cases} x, & \text{if } \; k = 1, \\ 1 - x, & \text{if } \; k = 0, \end{cases} \\
& y = x + v, \; E_1[v|F^x] = 0.
\end{aligned}$$

Note that the tuple (Bernoulli, Beta) is a tuple of estimation-conjugate probability distributions, see Def. 9.2.4

The Radon-Nikodym derivative of the above model is calculated. With respect to the probability measure P_1, the output has a Binomial distribution with parameters (n_y, x) while with respect to the probability measure P_0 it should have the parameters $(n_y,\ p_0)$ with the real number $p_0 \in (0,1)$. The Radon-Nikodym derivative can then be calculated as the quotient of the frequency functions for the respective measures,

$$\begin{aligned}
& p_0 \in (0,1), \\
r_{1|0} &= \left(\binom{n_y}{y} x^y (1-x)^{(n_y-y)} \right) / \left(\binom{n_y}{y} (p_0)^y (1-p_0)^{(n_y-y)} \right) \\
&= \left(\frac{x}{p_0} \right)^y \left(\frac{1-x}{1-p_0} \right)^{(n_y-y)},\ r_{0|1} = r_{1|0}^{-1}.
\end{aligned}$$

Proposition 9.3.5. *Consider the Binomial-Beta model of Def. 9.3.4. Define the random variable,*

$$r_{0|1} = \left(\frac{x}{p_0} \right)^{-y} \left(\frac{1-x}{1-p_0} \right)^{y-n_y}.$$

(a)Then,

$$r_{0|1} > 0,\ a.s.(P_1),\ E_1[\exp(iwy) r_{0|1} | F^x] = E_1[\exp(iwy) r_{0|1}],\ E_1[r_{0|1}|F^x] = 1.$$

The conditions of Theorem 9.3.3 are then satisfied. From that theorem then follows that the next formula defines a probability measure P_0 on $(\Omega, F^x \vee F^y)$ such that P_0 and P_1 are equivalent probability measures,

$$\begin{aligned}
& \frac{dP_0}{dP_1} = r_{0|1}, r_{1|0} = r_{0|1}^{-1}, \\
& E_0[\exp(iw_y y)] = E_1[\exp(iw_y y) r_{1|0}],\ \forall\ w_y \in \mathbb{R}, \\
& E_0[\exp(iw_x x)] = E_1[\exp(iw_x x)],\ \forall\ w_x \in \mathbb{R}, \\
& F^x,\ F^y \text{ are independent with respect to } P_0.
\end{aligned}$$

(b)The conditional distribution of the state x conditioned on the observation y equals,

$$\begin{aligned}
& E_1[\exp(iw_x x)|F^y] \\
&= \frac{\Gamma(\beta_1+\beta_2+n_y)}{\Gamma(\beta_1+y)} \sum_{k=0}^{\infty} \frac{\Gamma(\beta_1+n_y+k)}{\Gamma(\beta_1+\beta_2+n_y+k)} \frac{(iw_x)^k}{k!},
\end{aligned} \tag{9.11}$$

which is a Beta characteristic function with parameters $(\hat{\beta}_1, \hat{\beta}_2)$;

$$\hat{\beta}_1 = \beta_1 + y,\ \hat{\beta}_1 : \Omega \to \mathbb{R}_{s+}, \tag{9.12}$$

$$\hat{\beta}_2 = \beta_2 + n_y - y,\ \hat{\beta}_2 : \Omega \to \mathbb{R}_{s+}, \tag{9.13}$$

$$\begin{aligned}
\hat{x} &= m_{x|y} = E_1[x|F^y] = m_{x|y} = \frac{\hat{\beta}_1}{\hat{\beta}_1+\hat{\beta}_2} = \frac{\beta_1+y}{\beta_1+\beta_2+n_y} \\
&= E_1[x] + \frac{1}{\beta_1+\beta_2+n_y}(y - n_y E_1[x]),\ E_1[x] = \frac{\beta_1}{\beta_1+\beta_2},
\end{aligned} \tag{9.14}$$

$$q_{x|y} = E_1[(x - E_1[x|F^y])^2|F^y] = \frac{(\beta_1 + y)(\beta_2 + n_y - y)}{(\beta_1 + \beta_2 + n_y)(\beta_1 + \beta_2 + n_y + 1)}. \tag{9.15}$$

The Bernoulli-Beta case is specified by setting in the above formula $n_y = 1$.

$$\hat{x} = E_1[x] + \frac{1}{\beta_1 + \beta_2 + 1}(y - E_1[x]), \;\; E_1[x] = \frac{\beta_1}{\beta_1 + \beta_2}.$$

Both $\hat{\beta}_1$ and $\hat{\beta}_2$ depend on the output y. Also, both $\hat{x}$ and $q_{x|y}$ depend on the output.

Note that the conditional probability distribution $\mathrm{cpdf}(x|y)$ *is again a Beta probability distribution function because of the tuple of estimation-conjugate distributions (Binomial, Beta) chosen for the model.*

Proof. (a) It is proven that the random variable $r_{0|1}$ has the required properties.

$$x : \Omega \to (0,1), \;\; y : \Omega \to \mathbb{N}_{n_y} = \{0,1,\ldots,n_y\},$$

$$r_{0|1} = \left(\frac{x}{p_0}\right)^{-y} \left(\frac{1-x}{1-p_0}\right)^{y-n_y} > 0 \; a.s. \; P_1; \text{ because, } p_0 \in (0,1),$$

$$\begin{aligned}
&E_1[\exp(iw_y y) r_{0|1}(x,y)|F^x] \\
&= E_1[\exp(iw_y y) \left(\frac{x}{p_0}\right)^{-y} \left(\frac{1-x}{1-p_0}\right)^{y-n_y} |F^x] \\
&= E_1[\sum_{k=0}^{n_y} I_{\{y=k\}} \; \exp(iw_y y) \left(\frac{x}{p_0}\right)^{-k} \left(\frac{1-x}{1-p_0}\right)^{k-n_y} |F^x] \\
&= \sum_{k=0}^{n_y} \exp(iw_y k) \left(\frac{x}{p_0}\right)^{-k} \left(\frac{1-x}{1-p_0}\right)^{k-n_y} \binom{n_y}{k} x^k (1-x)^{n_y-k} \\
&= \sum_{k=0}^{n_y} \binom{n_y}{k} (p_0 \exp(iw_y))^k (1-p_0)^{n_y-k} \\
&= (1 - p_0 + p_0 \exp(iw_y))^{n_y} = E_1[\exp(iw_y y) r_{0|1}(x,y)],
\end{aligned}$$

because the previous expression is deterministic due to $p_0 \in (0,1)$ and $w_y \in \mathbb{R}$. This implies, by setting $w_y = 0$ in the above formula, that,

$$E_1[r_{0|1}(x,y)|F^x] = E_1[\exp(iw_y y) r_{0|1}(x,y)|F^x]|_{w_y=0} = (1 - p_0 + p_0) = 1.$$

Hence the conditions of Theorem 9.3.3 are satisfied. It follows from that theorem that the random variables y and x are independent with respect to the measure P_0 where $dP_0/dP_1 = r_{0|1}$. One can thus calculate with respect to P_0 the conditional characteristic function. This is done only for the case $n_y = 1$.

$$\begin{aligned}
&E_0[\exp(iw_x x) r_{1|0}|F^y] \\
&= E_0[\exp(iw_x x) \left(\frac{x}{p_0}\right)^{y} \left(\frac{1-x}{1-p_0}\right)^{1-y} |F^y] \\
&= \int_0^1 \exp(iw \, v) \; (\frac{v}{p_0})^y \; (\frac{1-v}{1-p_0})^{1-y} \; v^{\beta_1 - 1} \; (1-v)^{\beta_2 - 1} \; dv/B(\beta_1, \beta_2)
\end{aligned}$$

because wrt P_0, which equals wrt P_1, x has a Beta(β_1, β_2) pdf,

$$= \int_0^1 \exp(iw\ v)\ v^{\beta_1+y-1}\ (1-v)^{(\beta_2+1-y)-1}\ dv\ (p_0)^{-y}\ (1-p_0)^{y-1} / \ B(\beta_1,\beta_2)$$

which is a characteristic function of a Beta pdf,

except for terms not depending on the variable w_x,

$$E_1[\exp(iw_x x)|F^y] = \frac{E_0[\exp(iw_x x) r_{1|0}(x,y)|F^y]}{E_0[r_{1|0}(x,y)|F^y]}$$

$$= \int_0^1 \exp(iw\ v)\ v^{\beta_1+y-1}\ (1-v)^{(\beta_2+1-y)-1}\ dv / B(\beta_1+y, \beta_2+1-y)$$

$$= \frac{\Gamma(\beta_1+\beta_2+1)}{\Gamma(\beta_1+y)} \sum_{s=0}^{\infty} \frac{(iw_x)^k}{k!} \frac{\Gamma(\beta_1+y+k)}{\Gamma(\beta_1+\beta_2+1+k)},$$

which is a Beta-pdf $(\beta_1+y,\ \beta_2+1-y)$. The formulas for the conditional mean and the conditional variance then follow directly from the formulas above Def. 9.3.4.

□

Estimator Poisson-Gamma

Definition 9.3.6. Define the *Poisson-Gamma estimation model* by the following mathematical structure.

$$(\Omega, F, P_1),\ y: \Omega \to \mathbb{N},\ x: \Omega \to \mathbb{R}_+,$$

$$E_1[I_{\{y(\omega)=k\}}|F^x] = \frac{x^k}{k!}\exp(-x),\ \forall\ k \in \mathbb{N},$$

x has a Gamma-pdf with parameters $(\gamma_1, \gamma_2) \in (0,\infty)^2$;

$$y = x + w,\ \ E_1[w|F^x] = 0.$$

The conditional probability distribution cpdf$(.;y|F^x)$ belongs to the Poisson set of probability frequency functions. Note that the tuple (Poisson, Gamma) is a tuple of estimation-conjugate probability distributions.

That the random variable w has the stated property follows from,

$$E_1[y|F^x] = \sum_{k=0}^{\infty} k E_1[I_{\{y=k\}}|F^x] = [\sum_{k=1}^{\infty} \frac{x^k}{(k-1)!}\exp(-x)]$$

$$= x\exp(-1) \sum_{m=0}^{\infty} x^m/m! = x,\ \ w = y - x.$$

Below the Radon-Nikokym derivative is derived. The measure P_1 is defined above and with respect to that measure, the measure of the output condition on the state is a Poisson measure with parameter x. The measure P_0 has to be such that with respect to that measure the measure of the output is a Poisson measure with parameter $1 \in \mathbb{R}$. The quotient of the two measure is thus,

$$\sum_{k=0}^{\infty} I_{\{y=k\}} \frac{x^k \exp(-x)/k!}{1^k \exp(-1)/k!} = \sum I_{\{y=k\}} x^k \exp(-x+1) = x^y \exp(-x+1).$$

Thus the Radon-Nikodym derive will be taken to be the above formula. For this random variable the result will be derived.

Proposition 9.3.7. *Consider the Poisson-Gamma estimation model of Def. 9.3.6. Define the random variable,*

$$r_{0|1} = x^{-y} \exp(x-1),\ r_{0|1} : \Omega \to \mathbb{R}_+.$$

(a)Then

$$r_{0|1} > 0,\ a.s.,\ 1 = E_1[r_{0|1}|F^x],\ E_1[\exp(iwy) r_{0|1}|F^x] = E_1[\exp(iwy) r_{0|1}].$$

Consequently the conditions of Theorem 9.3.3 hold. Define the probability measure,

$$P_0,\ dP_0/dP_1 = r_{0|1} = x^{-y}\exp(x-1);\ \textit{then,}$$
$$E_0[\exp(iw_y y)] = \exp(\exp(iw_y) - 1),\ E_0[\exp(iw_x x)] = E_1[\exp(iw_x x)].$$

(b)The conditional characteristic function of the state variable conditioned on the output is of Gamma type with,

$$E_1[\exp(iw_x x)|F^y] = (1 - iw_x \hat{\gamma}_2)^{-\hat{\gamma}_1};\ \text{cpdf}(x|F^y) \in \Gamma(\hat{\gamma}_1, \hat{\gamma}_2);$$
$$\hat{\gamma}_1 = \gamma_1 + y,\ \hat{\gamma}_2 = \frac{\gamma_2}{1+\gamma_2} = \frac{1}{1+1/\gamma_2};$$
$$E_1[x|F^y] = \hat{\gamma}_1 \hat{\gamma}_2 = E_1[x] + \hat{\gamma}_2 (y - E_1[x]),\ E_1[x] = \gamma_1 \gamma_2.$$

Note that the conditional probability distribution derived above is again of Gamma type according to the tuple of estimation-conjugate probability distributions chosen.

Proof. (a) The result of Theorem 9.4.1 will be used. Because x has a Gamma probability distribution, $x > 0\ a.s.(P_1)$. Then,

$$r_{0|1} = x^{-y} \exp(-x+1) = \sum_{k=0}^{\infty} x^{-k} \exp(-x+1) I_{\{y=k\}} > 0,\ a.s.(P_1)$$

From the formula for the conditional distribution of the output conditioned on the state follows that,

$$\begin{aligned}
& E_1[\exp(iw_y y) r_{0|1}|F^x] = E_1[\exp(iwy)\ x^{-y}\exp(x-1)|F^x] \\
& = E_1[\sum_{k=0}^{\infty} I_{\{y=k\}}\ \exp(iwk) x^{-k} \exp(x-1)|F^x] \\
& = \exp(x-1) \sum \exp(iwk) x^{-k} E_1[I_{\{y=k\}}|F^x] \\
& = \exp(x-1) \sum \exp(iwk) x^{-k} x^k \exp(-x)/k! \\
& = \exp(-1) \sum (\exp(iw))^k\ k! = \exp(\exp(iw) - 1) \\
& = E_1[\exp(iwy) r_{0|1}];\ \ E_1[r_{0|1}|F^x] = 1;
\end{aligned}$$

where the next to last expression follows because the obtained formula is deterministic and by taking expectation; and the last formula follows from the first by setting $w = 0$.

Hence the conditions of Theorem 9.3.3 are satisfied. One concludes from that theorem that, with respect to P_0, the σ-algebras F^x and F^y are independent and that

$$E_1[\exp(iwx)] = E_0[\exp(iwx)].$$

(b) Thus one can calculate,

$$\begin{aligned}
&E_0[\exp(iw_x x) r_{1|0} | F^y] = E_0[\exp(iw_x x) x^y \exp(-x+1) | F^y] \\
&= e \int_0^\infty v^y \exp(-v(1 - iw_x)) v^{\gamma_1 - 1} \exp(-v/\gamma_2) \gamma_2^{-\gamma_1} dv / \Gamma(\gamma_1) \\
&= e^1 \int v^{(\gamma_1 + y) - 1} \exp(-v/((1/\gamma_2) + 1 - iw_x)) \gamma_2^{-\gamma_1} dv / \Gamma(\gamma_1) \\
&= e^1 ((1/\gamma_2) + 1 - iw_x)^{-(\gamma_1 + y)} \gamma_2^{-\gamma_1} \frac{\Gamma(\gamma_1 + y)}{\Gamma(\gamma_1)} \\
&E_1[\exp(iw_x x) | F^y] = \frac{E_0[\exp(iw_x x) r_{1|0} | F^y]}{E_0[r_{1|0} | F^y]} = \frac{(1 + 1/\gamma_2 - iw_x \gamma_2)^{-(\gamma_1 + y)}}{(1 + 1/\gamma_2)^{-(\gamma_1 + 1)}} \\
&= \left(1 - iw_x \frac{1}{1 + 1/\gamma_2}\right)^{-(\gamma_1 + y)} = (1 - iw_x \hat{\gamma}_2)^{-\hat{\gamma}_1}.
\end{aligned}$$

□

Estimator Gamma-Gamma

Definition 9.3.8. Define the *Gamma-Gamma estimation model* by the objects and relations,

$$\begin{aligned}
&(\Omega, F, P_1),\ y : \Omega \to \mathbb{R}_+,\ x : \Omega \to \mathbb{R}_+, \\
E_1[I_A(y) | F^x] = &\int_A v^{q-1} \exp(-v/x^{-1}) (1/x)^{-q} dv / \Gamma(q), \\
&\text{Gamma pdf } (q,\ 1/x),\ x : \Omega \to \mathbb{R}_{s+},\ q \in (0, \infty), \\
&x : \Omega \to \mathbb{R}_+,\ \text{Gamma pdf } (\gamma_1, \gamma_2);
\end{aligned}$$

thus with cpdf$(y|x)$ a Gamma conditional pdf with parameters $(q,\ 1/x)$ and with the random variable x having a Gamma-pdf with parameters $(\gamma_1, \gamma_2) \in \mathbb{R}_{s+}^2$. Note that the tuple (Gamma, Gamma) is a tuple of estimation-conjugate probability distribution functions though with the choice of the parametrization $(q, 1/x)$ for the parameters of the conditional distribution of the output conditioned on the state. The model has the variable representation,

$$y = \frac{q}{x} + w,\ \text{ with } E_1[w|F^x] = 0,\ \text{ which follows from } E_1[y|F^x] = q/x.$$

The Radon-Nikodym derivative is derived. The measure P_0 is chosen such that with respect to that measure the random variable y has a Gamma pdf with parameters $(q,1)$. The Radon-Nikodym derivative with respect to Lebesgue density is then the quotient of the two densities,

$$\begin{aligned} r_{1|0} &= \frac{y^{q-1}\ \exp(-y/x^{-1})(1/x)^{-q}/\Gamma(q)}{y^{q-1}\exp(-y)1^{-q}/\Gamma(q)} = x^q\exp(-y[x-1]), \\ r_{0|1} &= r_{1|0}^{-1} = x^{-q}\exp(y(x-1)). \end{aligned}$$

The random variable $r_{0|1}$ is the candidate Radon-Nikodym derivative.

Proposition 9.3.9. *Consider the Gamma-Gamma observation model of Def. 9.3.8. Define the random variable, $r_{0|1} = x^{-q}\exp(-y(x-1))$.*

(a)Then

$$\begin{aligned} &r_{1|0} > 0\ a.s.(P_1), \\ &E_1[\exp(iwy)r_{0|1}|F^x] = (1-iw)^{-q} = E_1[\exp(iwy)r_{0|1}],\ E_1[r_{0|1}|F^x] = 1. \end{aligned}$$

It then follows from Theorem 9.3.3 that the following formula defines a probability measure P_0 on $(\Omega, F^x \vee F_y)$ such that,

$$\begin{aligned} &\frac{dP_0}{dP_1} = r_{0|1},\ E_0[\exp(iw_xy)] = (1-iw_x)^{-q},\ E_0[\exp(iwx)] = E_1[\exp(iwx)], \\ &F^x,\ F^y \text{ are independent with respect to } P_0;\ P_0 \sim P_1. \end{aligned}$$

(b)The conditional characteristic function of the state variable x conditioned on the observation y equals,

$$\begin{aligned} E_1[\exp(iw_xx)|F^y] &= (1-iw_x\hat{\gamma}_2)^{-\hat{\gamma}_1}\ \forall\ w_x \in \mathbb{R}; \\ \hat{\gamma}_1 &= \gamma_1 + q,\ \hat{\gamma}_2 = \frac{\gamma_2}{1+y\gamma_2} = \frac{1}{y+1/\gamma_2}, \\ \hat{x} &= E_1[x|F^y] = \hat{\gamma}_1\hat{\gamma}_2 = \frac{\gamma_1+q}{y+1/\gamma_2}, \\ q_{x|y} &= E_1[(x-E_1[x|F^y])^2|F^y] = \frac{(\gamma_1+q)}{(y+1/\gamma_2)^2}. \end{aligned}$$

The variable $\hat{\gamma}_2$ depends on the output y but $\hat{\gamma}_1$ does not depend on the output. Both $\hat{x}$ and $q_{x|y}$ depend on the output.
The obtained conditional probability distribution is again of Gamma type because of the tuple of conjugate probability distrubitions (Gamma, Gamma) which was chosen.

Proof. (a) The conditions of Theorem 9.3.3 are used. Note that,

$$x > 0\ a.s.\ \Rightarrow\ r_{0|1} = x^{-q}\exp(y(x-1)) > 0.$$

Then,

$$
\begin{aligned}
& E_1[\exp(iwy)r_{0|1}|F^x] = E_1[\exp(iwy)x^{-q}\exp(y(x-1))|F^x] \\
& = \int_{\mathbb{R}_+} \exp(iwv)x^{-q}\exp(v(x-1))\; v^{q-1}\exp(-v/x^{-1})(1/x)^{-q}dv/\Gamma(q) \\
& = \int v^{q-1}x^{-q}\exp(-v[-iw-(x-1)+x])\;(1/x)^{-q}\;dv/\Gamma(q) \\
& = \int v^{q-1}\;\exp(-v(1-iw))dv/\Gamma(q) = (1-iw)^{-q} \\
& \quad w=0 \;\Rightarrow\; E_1[r_{0|1}|F^x] = 1.
\end{aligned}
$$

Thus the conditions of Theorem 9.3.3 are satisfied. It follows that P_0 is a probability measure and that $P_1 \sim P_0$ with $dP_1/dP_0 = r_{0|1}^{-1} = r_{1|0}$. Then,

$$
\begin{aligned}
& E_0[\exp(iwx)\; r_{1|0}|F^y] \\
& = \int \exp(iwv)\; v^q\exp(-y(v-1))\; v^{\gamma_1-1}\exp(-v/\gamma_2)\gamma_2^{-\gamma_1}dv/\Gamma(\gamma_1) \\
& = \exp(y)\int v^{q+\gamma_1-1}\exp(-v[y+1/\gamma_2-iw])\gamma_2^{-\gamma_1}dv/\Gamma(q) \\
& = \exp(y)\int v^{q+\gamma_1-1}\exp(-v/(y+1/\gamma_2-iw))^{-1})\gamma_2^{-\gamma_1}dv/\Gamma(q) \\
& = [y+1/\gamma_2-iw]^{-(\gamma_1+q)}\exp(y)\gamma_2^{-\gamma_1}\Gamma(\gamma_1+q)/\Gamma(\gamma_1), \\
& E_1[\exp(iwx)|F^y] \\
& = \frac{E_0[\exp(iwx)r_{1|0}|F^x]}{E_0[r_{1|0}|F^x]} = \frac{[y+(1/\gamma_2)-iw]^{-(\gamma_1+q)}}{[y+(1/\gamma_2)]^{-\gamma_1+q}} \\
& = [1-iw\frac{1}{y+1/\gamma_2}]^{-(\gamma_1+q)} = (1-iw\hat{\gamma}_2)^{-\hat{\gamma}_1}.
\end{aligned}
$$

□

Estimator Gauss-Gauss

A Gaussian system representation has been defined in Def. 4.3.1. One can define the estimation model and the sequential estimation model accordingly. The filter problem for a Gaussian system has been solved in Chapter 8.

One can also derive the Kalman filter of a Gaussian system by the method used in this chapter. This will not be written in this book. Only the Gauss-Gauss estimation model will be stated with the solution to the estimation problem and its proof. The outcome of the proposition is identical to that provided by Theorem 2.8.3. The reader may appreciate the simple proof of Theorem 2.8.3 compared to the proof of Proposition 9.3.11 provided below.

Definition 9.3.10. Define the *Gauss-Gauss estimation model* by the objects and relations,

$$\begin{aligned}
&(\Omega, F, P_1),\ y : \Omega \to \mathbb{R},\ x : \Omega \to \mathbb{R},\ q_v \in (0, \infty),\\
E_1[\exp(iw_y y)|F^x] &= \exp(iw_y x - w_y^2 q_v/2),\\
&x : \Omega \to \mathbb{R},\ \ x \in G(m_x, q_x),\ (m_x, q_x) \in \mathbb{R} \times \mathbb{R}_{s+}.
\end{aligned}$$

There exists a random variable $v = y - x$ such that,

$$\begin{aligned}
&v : \Omega \to \mathbb{R},\ v \in G(0, q_v),\ F^x,\ F^v \text{ independent such that,}\\
&y = x + v.
\end{aligned}$$

Note that the tuple (Gauss, Gauss) is a tuple of estimation-conjugate probability distribution functions with the representation in terms of the specified parameters chosen.

Proposition 9.3.11. *Consider the Gauss-Gauss estimation model of Def. 9.3.10.*

(a)The formula $r_{0|1}$ displayed below defines a probability measure P_0 on the measurable space (Ω, F) such that P_0 and P_1 are equivalent probability measures with the properties that,

$$\begin{aligned}
r_{0|1} &= \exp(-xy/q_v + x^2/(2q_v)),\\
dP_0/dP_1 &= r_{0|1},\\
E_1[r_{0|1} \exp(iw_y y)|F^x] &= E_1[r_{0|1} \exp(iw_y y)],\\
E_0[\exp(iw_y y)] &= E_1[r_{0|1} \exp(iw_y y)] = \exp(-w_y^2 q_v/2).
\end{aligned}$$

(b)The conditional distribution of the state variable x conditioned on the observation y equals,

$$\begin{aligned}
E_1[\exp(iw_x x)|F^y] &= \exp\left(iw_x \hat{x} - \frac{1}{2} w_x^2 q_{x|y}\right),\ \forall\ w_x \in \mathbb{R};\\
E_1[x|F^y] &= \hat{x} = (1/q_x + 1/q_v)^{-1}[y/q_v + m_x/q_x]\\
&= m_x + q_x q_y^{-1}(y - m_y),\\
q_{x|y} &= q_x - q_x^2 q_y^{-1},\ q_y = q_x + q_v > 0.
\end{aligned}$$

That the obtained conditional probability distribution is again of Gaussian type is a direct consequence of the chosen tuple of conjugate probability distributions (Gauss, Gauss).

Proof. Theorem 9.3.3 will be used to establish the result.
(a) Consider the random variable $r_{0|1}$ as defined (a). One calculates,

$$\begin{aligned}
&E_1[r_{0|1} \exp(iw_y y)|F^x]\\
&= \int \exp\left(-2xv_y/(2q_v) + x^2/(2q_v) + iw_y v_y - (v_y - x)^2/(2q_v)\right)(2\pi q_v)^{-1/2} dv_y\\
&= \int \exp(iw_y v_y)\ \exp(-v_y^2/(2q_v)) dv_y/(2\pi q_v)^{1/2}\\
&= \exp(-w_y^2 q_v/2) = E_1[r_{0|1} \exp(iw_y y)];\\
&\quad E_1[r_{0|1}|F^x] = 1,\ \text{ obtained by setting } w_y = 0 \text{ in the above formula.}
\end{aligned}$$

Thus $dP_0/dP_1 = r_{0|1}$ defines a probability measure P_0 on (Ω, F). Because $r_{0|1} > 0$ *a.s.*, the probability measures P_0 and P_1 are equivalent. Then

$$dP_1/dP_0 = r_{1|0} = r_{0|1}^{-1} = \exp(xy/q_v - x^2/(2q_v)).$$

Moreover, by Theorem 9.3.3,

$$E_0[\exp(iw_x x)] = E_1[\exp(iw_x x)] = \exp(iw_x m_x - w_x^2 q_x/2).$$

Define the random variable v as,

$$\begin{aligned}
v &= y - x,\\
&\quad E_1[\exp(iw_y y)|F^x] = \exp(iw_y x - w_y^2/(2q_v)),\\
&\Rightarrow E_1[\exp(iw_y v)|F^x] = \exp(-w_y^2/(2q_v)) = E_1[\exp(iw_y y)],\\
&\Rightarrow v \in G(0, q_v), \ \text{ and } F^v,\ F^x \ \text{ are independent.}
\end{aligned}$$

(b) One calculates,

$$\begin{aligned}
&E_0[r_{1|0}\exp(iw_x x)|F^y] = E_0[\exp(\frac{1}{2q_v}[2xy - x^2] + iw_x x)|F^y]\\
&= \int \exp(-\frac{1}{2}[-2v_x y/q_v + v_x^2/q_v + (v_x - m_x)^2/q_x - iw_x v_x]) \times\\
&\quad \times\ (2q_x)^{-1/2} dv_x,\\
[\ldots] &= -2v_x y/q_v + v_x^2/q_v + v_x^2/q_x - 2v_x m_x/q_x + m_x^2/q_x\\
&= (1/q_x + 1/q_v)v_x^2 - 2v_x[y/q_v + m_x/q_x] + m_x^2/q_x\\
&= (1/q_x + 1/q_v)(v_x - [1/q_x + 1/q_v]^{-1}[y/q_v + m_x/q_x])^2 +\\
&\quad + m_x^2 q_x - (1/q_x + 1/q_v)^{-1}[y/q_v + m_x/q_x]^2,\\
&E_0[r_{1|0}\exp(iw_x x)|F^y] = \int \exp(iw_x v_x) p_{x|y}(dv_x)\\
&= \exp(iw_x E[x|F^y] - w_x^2 q_{x|y}/2) \times f(y, m_x, q_x, q_v),
\end{aligned}$$

$$\begin{aligned}
&E_1[\exp(iw_x x)|F^y]\\
&= \frac{E_0[r_{1|0}\exp(iw_x x)|F^y]}{E_0[r_{1|0}|F^y]} = \exp(iw_x E[x|F^y] - w_x^2 q_{x|y}/2),\\
&(1/q_x + 1/q_v)^{-1} = \frac{q_x q_v}{q_x + q_v} = \frac{q_x q_v}{q_y},\\
&E_1[x|F^y] = (1/q_x + 1/q_v)^{-1}[y/q_v + m_x/q_x] = m_x + q_x q_y^{-1}[y - m_y],\\
q_{x|y} &= \frac{q_x q_v}{q_y} = \frac{q_x q_y}{q_y} - \frac{q_x q_y - q_x q_v}{q_y} = q_x - q_x^2 q_y^{-1}.
\end{aligned}$$

Note that the function f used above does not depend on the variable w_x hence it cancels in the quotient of numerator and denominator. □

Estimation of a Finite-Valued Random Variable

The formula for the conditional expectation of a tuple of finite-valued random variables is stated in Proposition 2.8.4. Below another formulation and a deriviation by the measure transformation method is described. The estimation problem is an introduction to the derivation of the filter system of a finite stochastic system.

Problem 9.3.12. *Conditional expectation of finite-valued random variables.* Consider the two finite-valued random variables both in their indicator representation Def. 2.5.9,

$$x : \Omega \to \mathbb{R}^{n_x},\; y : \Omega \to \mathbb{R}^{n_y},$$
$$Q_{x,y} = E_1[xy^T] \in \mathbb{R}_+^{n_x \times n_y},\; Q_{x,y} \neq 0,\; 1^T Q_{x,y} = (p_y)^T;$$
$$\text{assume } p_y \in \mathbb{R}_{s+}^{n_y}\; (\Rightarrow\; \forall\, i \in \mathbb{Z}_{n_y},\; p_{y,i} > 0).$$

Determine an expression for the conditional expectation $E_1[x|\; F^y]$.

The measure transformation will be used to solve the problem. With respect to the probability measure of the problem, P_1, the probability of the components of y are,

$$I_{\{y=e_i\}} \sum_{j=1}^{n_x} C_{i,j}\, x_j,\;\; \forall\, i \in \mathbb{Z}_{n_y}.$$

A new measure P_0 is to be constructed such that with respect to P_0, the probability of the components of y are,

$$I_{\{y=e_i\}}\; 1/n_y,\; \forall\, i \in \mathbb{Z}_{n_y}.$$

The Radon-Nikodym derivative is then,

$$r_{1|0} I_{\{y=e_i\}} = \frac{\sum_{j=1}^{n_x} C_{i,j}\, x_j}{1/n_y} = \frac{(Cx)_i}{1/n_y}\, I_{\{y=e_i\}},\;\; r_{1|0} = n_y (Cx)^T y = n_y x^T C^T y.$$

Proposition 9.3.13. Conditional expectation of finite-valued random variables by the measure transformation method. *Consider Problem 9.3.12. Define the random variable* $r_{1|0} = n_y (Cx)^T y = n_y x^T C^T y = n_y y^T Cx.$

(a)Then,

$$y^T Cx > 0,\; r_{0|1} = \frac{1}{n_y y^T Cx} > 0,\; a.s.(P_1),\; \textit{define}\; r_{1|0} = r_{0|1}^{-1},$$
$$E_1[\exp(iwy) r_{0|1} | F^x] = E_1[\exp(iwy) r_{0|1}],\; E_1[r_{0|1}|F^x] = 1.$$

The conditions of Theorem 9.3.3 are then satisfied. From that theorem then follows that the next formula defines a probability measure P_0 *on* $(\Omega, F^x \vee F^y)$ *such that* P_0 *and* P_1 *are equivalent probability measures,*

$$\frac{dP_1}{dP_0} = r_{1|0},$$
$$E_0[b(x)] = E_1[b(x)],$$
$$F^x,\; F^y \textit{ are independent with respect to } P_0.$$

(b)The conditional expectation requested in the problem equals,

$$
\begin{aligned}
x_{un} &= \mathrm{Diag}(C^T y) p_x = \mathrm{Diag}(p_x) C^T y,\ x_{un} : \Omega \to \mathbb{R}_+^{n_x},\\
\hat{x} &= E[x|\ F^y] = x_{un}/(1_{n_x}^T x_{un}) = \mathrm{Diag}(C^T y) p_x/[p_y^T y] = \mathrm{Diag}(p_x) C^T y/[p_y^T y].
\end{aligned}
$$

One calls x_{un} the unnormalized conditional expectation *of x conditioned on or given F^y.*

Proof. (a) Because $Q_{x,y} \neq 0$, $r_{1|0} = n_y y^T C x > 0$. Define then, $r_{0|1} = r_{1|0}^{-1}$. One calculates,

$$
\begin{aligned}
& E_1[I_{\{y=e_i\}}\ r_{0|1}|\ F^x] = E_1[I_{\{y=e_i\}}\ n_y^{-1} 1/[x^T C^T y|\ F^x]\\
&= E_1[I_{\{y=e_i\}}\ n_y^{-1} 1/[x^T C^T e_i|\ F^x] = n_y^{-1}\ \frac{1}{x^T C^T e_i}\ E_1[I_{\{y=e_i\}}|\ F^x]\\
&= n_y^{-1}\ \frac{1}{(Cx)_i}\ (Cx)_i = n_y^{-1}\\
&= E_1[I_{\{y=e_i\}}\ r_{0|1}], \text{ because the obtained expression is deterministic,}\\
& E_1[r_{0|1}] = \sum_{i=1}^{n_y} E_1[I_{\{y=e_i\}} r_{0|1}] = \sum_{i=1}^{n_y} 1/n_y = 1.
\end{aligned}
$$

The conclusions then follow from Theorem 9.3.3. Note that then,

$$
\begin{aligned}
& r_{1|0} = r_{0|1}^{-1} = n_y\ x^T C^T y,\\
& \qquad E_0[x] = E_1[x],\ F^x,\ F^y \text{ are independent with respect to } P_0.
\end{aligned}
$$

(b) Note the calculations,

$$
\begin{aligned}
& E_0[I_{\{x=e_k\}} r_{1|0}|\ F^y]\\
&= E_0[I_{\{x=e_k\}} n_y x^T C^T y|\ F^y] = E_0[I_{\{x=e_k\}}|\ F^y]\ n_y e_k^T C^T y\\
&= E_0[I_{\{x=e_k\}}]\ n_y e_k^T C^T y, \text{ because wrt } P_0\ F^x \text{ and } F^y \text{ are independent,}\\
&= n_y p_x(k) e_k^T C^T y = n_y (C^T y)_k p_x(k) = n_y [\mathrm{Diag}(C^T y) p_x]_k,\\
E_0[x r_{1|0}|F^y] &= n_y \mathrm{Diag}(C^T y) p_x,\ x_r = \mathrm{Diag}(C^T y) p_x,\\
E_1[x|\ F^y] &= \frac{E_0[x r_{1|0}|\ F^y]}{E_0[r_{1|0}|\ F^y]} = \frac{n_y \mathrm{Diag}(C^T y) p_x}{n_y 1_{n_x}^T \mathrm{Diag}(C^T y) p_x} = \frac{x_{un}}{1_{n_x}^T x_{un}}, \text{ where,}\\
& 1_{n_x}^T x_{un} = (C^T y)^T p_x = y^T C p_x = y^T p_y = p_y^T y.
\end{aligned}
$$

□

9.4 Sequential Estimation

Below it is shown how the sequential estimation problem can be solved analogously to the estimation problem. This section makes use of elementary properties of a martingale process. See Section 20.2 for the definition of a martingale. In this section, a martingale depends on the particular probability measure used. Thus an integrable martingale with respect to the probability measure P_0 is denoted by $M \in M_1(P_0)$ while one with respect to the probability measure P_1 is denoted by $M \in M_1(P_1)$.

The problem formulation is with respect to a probability measure P_1 on the measurable space (Ω, F). A probability measure P_0 is constructed on this probability space such that with respect to the measure P_0:

1. the random variable x to be estimated is independent of the output process $\{y(s) \in \mathbb{R}_{n_y}, \forall\, s \in T\}$; and
2. the observed stochastic process is a sequence of independent random variables; and
3. the probability measures P_1 and P_0 restricted to the σ-algebra F^x are identical.

Below these properties will be proven by first defining the Radon Nikodym derivative of dP_0/dP_1 and secondly proving that P_0 is a probability measure with the required properties.

Theorem 9.4.1. *Consider the sequential estimation model of Def. 9.1.2 and the associated sequential estimation problem of Problem 9.1.3. Assume that there exists a martingale,*

$$\{r_{0|1}(t), F^x \vee F_t^y, t \in T\} \in M_1(P_1),\ r_{0|1}(0) = 1,\ r_{0|1}(t) > 0,\ \forall\, t \in T.$$

Assume in addition that,

$$\begin{aligned} & E_1[\exp(iw_y y(t+1)) \frac{r_{0|1}(t+1)}{r_{0|1}(t)} | F^x \vee F_t^y] \\ & = E_1[\exp(iw_y y(t+1)) \frac{r_{0|1}(t+1)}{r_{0|1}(t)}],\ \forall\, t \in T. \end{aligned} \tag{9.16}$$

(a.a) Then the formula,

$$\frac{dP_0}{dP_1} = r_{0|1}(t_1), \tag{9.17}$$

defines a probability measure P_0 on the probability space $(\Omega, F^x \vee F_{t_1}^y)$ such that P_0 is equivalent to P_1;

(a.b) Define the process $r_{1|0}(t) = r_{0|1}(t)^{-1}$, which will then be a Radon-Nikodym martingale,

$$\{r_{1|0}(t), F^x \vee F_t^y, t \in T\} \in M_1(P_0); \tag{9.18}$$

(a.c) F^x and $F_{t_1}^y$ are independent with respect to the probability measure P_0;

(a.d) with respect to the probability measure P_0 the stochastic process y is a sequence of independent random variables; and

(a.e) the restrictions of P_1 and P_0 to F^x are identical.

(b)The conditional characteristic function of the state x conditioned on the past of the outputs equals the formula,

$$E_1[\exp(iw_x^T x) | F_t^y] = \frac{E_0[\exp(iw_x^T x) r_{1|0}(t) | F_t^y]}{E_0[r_{1|0}(t) | F_t^y]},\ \forall\, t \in T,\ \forall\, w_x \in \mathbb{R}^{n_x}. \tag{9.19}$$

The numerator of this function can be calculated according to Theorem 2.8.6 by integrating over the probability distribution function of x while treating the stochastic process y as a known variable. The denominator can be obtained from the numerator by setting in the numerator $w_x = 0$.

Proof. (a.a) By the assumption that $r_{0|1}$ is a martingale with respect to P_1, it follows that,

$$E_1[r_{0|1}(t_1)|F^x] = E_1[E_1[r_{0|1}(t_1)|F^x \vee F_0^y]|F^x] = E_1[r_{0|1}(0)|F^x] = 1,$$
$$E_1[r_{0|1}(t_1)] = E_1[E_1[r_{0|1}(t_1)|F^x]] = E_1[r_{0|1}(0)] = r_{0|1}(0) = 1.$$

The formula $dP_0/dP_1 = r_{0|1}(t_1)$ then defines a probability measure P_0 on (Ω, F). The assumption that $r_{0|1}(t_1) > 0$ implies that the probability measures P_1 and P_0 are equivalent. The assumption that $r_{0|1}(t_1) > 0$ almost surely with respect to P_1 implies by [28, VI T 15] that for all $t \in T$, $r_{0|1}(t) > 0$ a.s. P_1, hence a.s. P_0. Then,

$$\frac{dP_1}{dP_0} = r_{0|1}(t_1)^{-1} = r_{1|0}(t_1).$$

(a.b) Then, $\forall\, A \in F^x$,

$$P_0(A) = E_0[I_A] = E_1[I_A r_{0|1}(t_1)] = E_1[I_A E_1[r_{0|1}(t_1)|F^x \vee F_0^y]]$$
$$= E_1[I_A r_{0|1}(0)] = E_1[I_A] = P_1(A),$$

from which follows that the restrictions of the measures P_1 and P_0 to F^x are identical.
(a.c) and (a.d) Let $t \in T$.

$$\begin{aligned}
&E_0[\exp(iw_y^T y(t+1))|F^x \vee F_t^y]\\
&= \frac{E_1[\exp(iw_y^T y(t+1)) r_{0|1}(t_1)|F^x \vee F_t^y]}{E_1[r_{0|1}(t_1)|F^x \vee F_t^y]}, \text{ by Theorem 19.9.10,}\\
&= \frac{E_1[\exp(iw_y^T y(t+1)) E_1[r_{0|1}(t_1)|F^x \vee F_{t+1}^y]|F^x \vee F_t^y]}{r_{0|1}(t)}\\
&= E_1[\exp(iw_y^T y(t+1)) \frac{r_{0|1}(t+1)}{r_{0|1}(t)}|F^x \vee F_t^y]\\
&= E_1[\exp(iw_y^T y(t+1)) \frac{r_{0|1}(t+1)}{r_{0|1}(t)}], \text{ by equation (9.16), hence}\\
&E_0[\exp(iw_y^T y(t+1))|F^x \vee F_t^y] = E_0[\exp(iw_y^T y(t+1))].
\end{aligned}$$

Hence the σ-algebra $F^{y(t+1)}$ is independent of $F^x \vee F_t^y$ for all $t \in T$. Thus, with respect to P_0, the sequence of random variables $\{y(t),\ t \in T\}$ is an independent sequence and the σ-algebras F^x and F_t^y are independent for all $t \in T$.
(a.e) Note that,

$$\begin{aligned}
&E_0[r_{1|0}(t+1)|F^x \vee F_t^y] = \frac{E_1[r_{1|0}(t+1) r_{0|1}(t_1)|F^x \vee F_t^y]}{E_1[r_{0|1}(t_1)|F^x \vee F_t^y]}\\
&= E_1[r_{1|0}(t+1) E_1[r_{0|1}(t_1)|F^x \vee F_{t+1}^y]|F^x \vee F_t^y]/r_{0|1}(t),\\
&= E_1[r_{1|0}(t+1) r_{0|1}(t+1)|F^x \vee F_t^y] r_{0|1}(t)^{-1} = r_{1|0}(t), \text{ hence,}\\
&\{r_{1|0}(t), F^x \vee F_t^y,\ t \in T\} \in M_1(P_0).
\end{aligned}$$

(b) From Theorem 19.9.10 follows directly that for all $t \in T$,

$$E_1[\exp(iw_x^T x)|F_t^y] = \frac{E_0[\exp(iw_x^T x) r_{1|0}(t_1)|F_t^y]}{E_0[r_{1|0}(t_1)|F_t^y]}$$
$$= \frac{E_0[\exp(iw_x^T x) E_0[r_{1|0}(t_1)|F^x \vee F_t^y]|F_t^y]}{E_0[E_0[r_{1|0}(t_1)|F^x \vee F_t^y]|F_t^y]} = \frac{E_0[\exp(iw_x^T x) r_{1|0}(t)|F_t^y]}{E_0[r_{1|0}(t)|F_t^y]}.$$

□

Sequential Estimator Binomial-Beta

In this subsection the solution is provided to the sequential estimation problem for an output process which takes values in a finite set $\mathbb{N}_n = \{0,1,2,\ldots,n\}$ or in the binary set $\mathbb{N}_1 = \{0,1\}$. The conditional distribution of the output conditioned on the state is either binomial or the special case of a Bernoulli distribution while the state has a beta distribution. Hence the name of Binomial-Beta estimation model and of Binomial-Beta sequential estimation system.

Definition 9.4.2. Define the *Binomial-Beta sequential estimation system* by the objects and relations,

$$y : \Omega \times T \to \mathbb{N}_{n_y},\ x : \Omega \to (0,1),$$
$$E_1[I_{\{y(t)=k\}}|F^x \vee F_{t-1}^y] = \binom{n_y}{k} x^k (1-x)^{n_y - k},\ \ \forall\, k \in \mathbb{N}_{n_y},$$
$$x \text{ has a Beta-pdf } (\beta_1, \beta_2) \in \mathbb{R}_{s+}^2 = (0,\infty)^2.$$

The *Bernoulli-Beta sequential estimation system* is the special case with $n_y = 1$. Note that the tuples (Binomial, Beta) and (Bernoulli, Beta) are pairs of conjugate probability distributions. The model has the representation,

$$y(t) = n_y\, x + \Delta\ w(t),$$
$$w : \Omega \times T \to \mathbb{R},\ \Delta w : \Omega \times T \to \mathbb{R},\ w(0) = 1,\ w(t) = 1 + \sum_{s=1}^{t} \Delta w(s),$$
$$(w(t), F^x \vee F_t^y,\ t \in T\} \in M_1(P_1).$$

Proposition 9.4.3. *Consider the Binomial–Beta sequential estimation system of Def. 9.4.2. The conditional distribution of the state conditioned on the past of the observations at any time $t \in T$, is of Beta type with the representation,*

$$E_1[\exp(iw_x x)|F_t^y] = \frac{\Gamma(\beta_1 + \beta_2 + tn_y)}{\Gamma(.)\Gamma(.)} \times$$
$$\times \sum_{k=0}^{\infty} \frac{\Gamma(\beta_1 + k + \sum_{s=1}^{t} y(s))}{\Gamma(\beta_1 + \beta_2 + tn_y + k)} \frac{(iw_x)^k}{k!} I_{\{y(t)=k\}},\ \forall\, t \in T, \quad (9.20)$$

which is a conditional characteristic function of Beta type with parameters, due to the conjugate probability distributions chosen,

$$
\begin{aligned}
\text{cpdf}(x|F^y_{t-1}) &= \beta\ (\hat{\beta}_1(t), \hat{\beta}_2(t)),\\
\hat{\beta}_1(t) &= \beta_1 + \sum_{s=0}^{t} y(s),\ \hat{\beta}_1 : \Omega \times T \to \mathbb{R}_+,\\
\hat{\beta}_2(t) &= \beta_1 + \beta_2 + t \times n_y,\ \hat{\beta}_2 : \Omega \times T \to \mathbb{R}_+,\\
\hat{\beta}_1(t+1) &= \hat{\beta}_1(t) + y(t),\ \hat{\beta}_1(0) = \beta_1,\\
\hat{\beta}_2(t+1) &= \hat{\beta}_2(t) + n_y,\ \hat{\beta}_2(0) = \beta_1 + \beta_2;\\
\hat{x}(t) &= E_1[x|F^y_t] = \frac{\hat{\beta}_1(t)}{\hat{\beta}_2(t)},\\
\hat{x}(t+1) &= \hat{x}(t) + \frac{1}{(\beta_1 + \beta_2 + tn_y + 1)}(y(t+1) - n_y\hat{x}(t)),\ \hat{x}(0) = \frac{\beta_1}{\beta_1+\beta_2}.
\end{aligned}
$$

Proof. The result of Theorem 9.4.1 will be used. The conditions of that proposition are verified. It follows from the definition of the random variable x and from $q_y \in (0,1)$ that the candidate Radon-Nikodym process equals,

$$
\begin{aligned}
r_{0|1}(t) &= \binom{n_y}{y(t)} x^{y(t)}(1-x)^{n_y - y(t)} / \binom{n_y}{y(t)} q_y^{y(t)}(1-q_y)^{n_y-y(t)}\\
&= \left(\frac{x}{q_y}\right)^{y(t)} \left(\frac{1-x}{1-q_y}\right)^{n_y - y(t)},\\
r_{0|1}(t) &= r^{-1}_{1|0}(t) = \left(\frac{x}{q_y}\right)^{-y(t)} \left(\frac{1-x}{1-q_y}\right)^{y(t)-n_y} > 0,\ a.s.\ P_1.
\end{aligned}
$$

As in the proof of Proposition 9.3.5 it follows that,

$$
\begin{aligned}
&E_1\left[\left(\frac{x}{q_y}\right)^{-y(t)} \left(\frac{1-x}{1-q_y}\right)^{y(t)-n_y} |F^x\right] = 1,\\
&E_1[\exp(iw_y y(t)) \left(\frac{x}{q_y}\right)^{-y(t)} \left(\frac{1-x}{1-q_y}\right)^{y(t)-n_y} |F^x]\\
&= (1 - q_y + q_y \exp(iw_y))^{n_y}.
\end{aligned}
$$

In addition, the following formula holds,

$$
E_1[\exp(iw_y y(t))|F^x] = (1 - x + x\exp(iw_y))^{n_y}.
$$

The latter characteristic function is stable with respect to finite additions.

The result then follows from Theorem 9.4.1 and a calculation. □

Sequential Estimator Gamma-Gamma

Definition 9.4.4. Define the *Gamma-Gamma recursive estimation model* by the objects and relations,

$T = T(0:t_1) = \{0,1,\ldots,t_1\},\ (\Omega,F,P_1),\ y:\Omega\times T\to\mathbb{R}_+,\ x:\Omega\to\mathbb{R}_+,$
$y(0)=0,$
$E_1[I_A(y(t+1))|F^x\vee F_t^y] = \int_A x^q v^{q-1}\exp(-v/x^{-1})(1/x)^{-q}dv/\Gamma(q),$
$\forall\, t\in T(0:t_1-1),$
which is a Gamma pdf with parameters, $(q,1/x),\ q\in(0,\infty),$
$x:\Omega\to\mathbb{R}_+,$ having a Gamma-pdf $(\gamma_1,\gamma_2)\in\mathbb{R}_{s+}.$

The model has the variable representation,

$$y(t) = \frac{q}{x} + w(t),\ \ (w(t),F^x\vee F_t^y, t\in T)\in M_1.$$

Proposition 9.4.5. *Consider the Gamma-Gamma recursive estimation model of Def. 9.3.8.*

(a)The Radon-Nikodym derivative is specified by,

$$r_{1|0} = x^q\ \exp(-y(t)(x-1)),$$
$$E_0[\exp(iw_xy)] = (1-iw_x)^{-q}.$$

(b)The conditional distribution of the state variable x conditioned on the past observation equals,

$$E_1[\exp(iw_xx)|F_t^y] = (1-iw_x\hat{\gamma}_2(t))^{-\hat{\gamma}_1(t)},\ \forall\, w_x\in\mathbb{R};$$
$$\hat{\gamma}_1(t) = \gamma_1 + qt,$$
$$\frac{1}{\hat{\gamma}_2(t)} = \frac{1}{\gamma_2(t)} + \sum_{s=1}^{t} y(s),$$
$$\hat{x}(t) = E_1[x|F_t^y] = \hat{\gamma}_1(t)\hat{\gamma}_2(t).$$

The obtained conditional probability distribution is again of Gamma type because of the tuple of conjugate probability distrubitions (Gamma, Gamma) which was chosen.

Proof. By now the reader can construct the details of the proof using the previous example.

9.5 Filtering Theory

In this section is described how to solve a filter problem along the line of the solution methods for the estimation problem and for the sequential estimation problem.

For a filter problem a measure transformation will be used. The problem is defined with respect to the probability measure P_1 on the measurable space (Ω,F). Then a new probability measure P_0 is constructed on the same probability space such that with respect to P_0 the following conditions all hold:

1. the observation process $\{y(s),\ \forall\ s \in T\}$ is an independent sequence;
2. the state and the output process satisfy the following independence condition: $\forall\ t \in T$, $F^{y(t)}$ is independent of $F_t^x \vee F_{t-1}^y$.

Below a sufficient condition for the existence of the probability measure P_0 is provided. For each special case of a stochastic system, it then has to be proven that the sufficient condition is satisfied.

Definition 9.5.1. Consider a stochastic system (without input) with the mathematical structure,

$$
\begin{aligned}
& \{(\Omega, F, P_1), T, (Y, B(Y)), (X, B(X)), y, x\}, \\
& (\{\text{cpdf}((x(t+1), y(t)) | F_t^x \vee F_{t-1}^y),\ \forall\ t \in T\},\ p_{x_0}), \\
= & (\{\text{cpdf}((x(t+1), y(t)) | F^{x(t)}),\ \forall\ t \in T\},\ p_{x_0}).
\end{aligned}
$$

Problem 9.5.2. Consider the stochastic system of Def. 9.5.1. Determine for each time $t \in T$, the conditional probability distribution of the state conditioned on the past outputs, described by the conditional probability distribution function, or, equivalently, by the conditional characteristic function,

$$
\begin{aligned}
& \text{cpdf}(x(t+1) |\ F_t^y), \text{ or} \\
& E_1[\exp(iw_x^T x(t+1)) | F_t^y],\ \forall\ w_x \in X \subseteq \mathbb{R}^{n_x},\ \forall\ t \in T.
\end{aligned}
$$

Call the recursions of all parameters of a conditional distribution, the *filter system* of the considered stochastic system.

Below use is made of a bounded and measurable function of a random variable taking values in a measurable space (X, G), $X \subseteq \mathbb{R}^n$, defined for the above formulated problem as,

$$b : X \to \mathbb{R},\ \exists\ c \in \mathbb{R}_{s+},\ \forall\ x_b \in X,\ |b(x_b)| \leq c.$$

The examples mainly used as a bounded and measurable function are: (1) $b(x(t)) = \exp(iw^T x(t))$ for $w \in \mathbb{R}^{n_x}$ in which case $|b(x(t))| = 1$; and (2) $x(t)$ is in the form of an indicator representation because then, for every component $i \in \mathbb{Z}_{n_x}$, there holds $|x_i(t)| \leq 1$. The reason to impose the boundedness assumption on the function b is to guarantee the existence of the conditional expectation for the random variable $b(x(t+1))$ conditioned on F_t^y.

Theorem 9.5.3. *Consider the stochastic system of Def. 9.5.1. Consider any bounded and measurable function $b : X \to \mathbb{R}_+$.*

Assume that there exists a martingale with respect to the probability measure P_1 such that,

$$
\begin{aligned}
& (r_{0|1}(t), F_t^x \vee F_t^y, t \in T) \in M_1(P_1),\ r_{0|1}(0) = 1,\ r_{0|1}(t_1) > 0,\ a.s.\ P_1; \\
& E_1[b(y(t+1)) \frac{r_{0|1}(t+1)}{r_{0|1}(t)} | F_t^x \vee F_t^y] = E_1[b(y(t+1)) \frac{r_{0|1}(t+1)}{r_{0|1}(t)}]. \qquad (9.21)
\end{aligned}
$$

*(**a**)a) The formula,*

$$\frac{dP_0}{dP_1} = r_{0|1}(t_1), \tag{9.22}$$

defines a probability measure P_0 on the probability space $(\Omega, F^x_{t_1} \vee F^y_{t_1-1})$ such that P_0 and P_1 are equivalent probability measures; define then,

$$r_{1|0}(t) = r_{0|1}(t)^{-1}, \ \forall\, t \in T.$$

(a.b) $\{r_{1|0}(t), F^x_t \vee F^y_t, t \in T\} \in M_1(P_0)$.

(a.c) The following relation holds,

$$E_1[b(x(t+1))|F^x_t \vee F^y_t] = \frac{E_0[b(x(t+1))r_{1|0}(t+1)|F^x_t \vee F^y_t]}{r_{1|0}(t)}, \ \forall\, t \in T.$$

(a.d) with respect to P_0, $\forall\, t \in T\backslash\{t_1\}$, $F^{y(t)}$ is independent of $F^x_t \vee F^y_{t-1}$;

(a.e) with respect to P_0, the process $\{y(t),\ t \in T\}$ is a sequence of independent random variables.

(b)The conditional expectation of the state conditioned on the past outputs equals,

$$E_1[b(x(t+1))|F^y_t] = \frac{E_0[E_1[b(x(t+1))|\ F^x_t \vee F^y_t]\ r_{1|0}(t)|F^y_t]}{E_0[r_{1|0}(t)|F^y_t]}, \ \forall\, t \in T. \tag{9.23}$$

The denominator can be obtained from the numerator by selection of the function $b(x_b) = 1$. Without further assumptions on the Radon-Nikodym process $r_{1|0}$, it is not possible to derive a more detailed expression.

Proof. (a.a) By assumption on the process $r_{0|1}$,

$$E_1[r_{0|1}(t_1)] = E_1[E_1[r_{0|1}(t_1)|F^x_0 \vee F^y_0]] = E_1[r_{0|1}(0)] = 1.$$

Because by assumption $(r_{0|1}(t), F^x_t \vee F^y_t,\ t \in T) \in M_1(P_1)$ and $r_{0|1}(t_1) > 0$ *a.s.*, it follows from [29, VI.T.15] that $r_{0|1}(t) > 0$, *a.s.* for all $t \in T$.

It follows from Theorem 19.9.4 that the formula (9.22) defines a probability measure P_0 on $(\Omega, F^x_{t_1} \vee F^y_{t_1})$. Moreover, P_1 and P_0 are equivalent on $F^x_t \vee F^y_t$ for all $t \in T$. Define for all $t \in T$, $r_{1|0}(t) = r_{0|1}(t)^{-1}$.

(a.b) Because P_0 and P_1 are equivalent, $dP_1/dP_0 = r_{1|0}(t_1)$. Then,

$$E_1[r_{0|1}(t_1)|F^x_t \vee F^y_t] = r_{0|1}(t), \text{ because } \{r_{0|1}(t), F^x_t \vee F^y_t,\ t \in T\} \in M_1(P_1),$$

$$E_0[r_{1|0}(t+1)|F^x_t \vee F^y_t] = \frac{E_1[r_{1|0}(t+1)r_{0|1}(t+1)|F^x_t \vee F^y_t]}{E_1[r_{0|1}(t+1)|F^x_t \vee F^y_t]},$$

by Theorem 19.9.10,

$$= r_{0|1}(t)^{-1} = r_{1|0}(t), \text{ because } r_{1|0}(t+1) = r_{0|1}(t+1)^{-1},$$

hence, $(r_{1|0}(t), F^x_t \vee F^y_t, t \in T) \in M_1(P_0)$.

(a.c) Note that then,

$$\begin{aligned}
& E_1[b(x(t+1))|F_t^x \vee F_t^y] \\
& = E_0[b(x(t+1))r_{1|0}(t_1)|F_t^x \vee F_t^y]/\ E_0[r_{1|0}(t_1)|F_t^x \vee F_t^y] \\
& = E_0[b(x(t+1))E_0[r_{1|0}(t_1)|F_{t+1}^x \vee F_{t+1}^y]|\ F_t^x \vee F_t^y]/\ r_{1|0}(t) \\
& = E_0[b(x(t+1))r_{1|0}(t+1)|\ F_t^x \vee F_t^y]/\ r_{1|0}(t).
\end{aligned}$$

(a.d) and (a.e) For $t \in T$, using the assumptions,

$$\begin{aligned}
& E_0[b(y(t+1))|F_t^x \vee F_t^y] \\
& = \frac{E_1[b(y(t+1))\ r_{0|1}(t_1)|F_t^x \vee F_t^y]}{E_1[r_{0|1}(t_1)|F_t^x \vee F_t^y]}, \text{ by Theorem 19.9.10}, \\
& = \frac{E_1[b(y(t+1))E_1[r_{0|1}(t_1)|F_{t+1}^x \vee F_{t+1}^y]|F_t^x \vee F_t^y]}{r_{0|1}(t)}, \text{ because } r_{1|0} \in M_1(P_1), \\
& = E_1[b(y(t+1))\frac{r_{0|1}(t+1)}{r_{0|1}(t)}|F_t^x \vee F_t^y] \\
& = E_1[b(y(t+1))\frac{r_{0|1}(t+1)}{r_{0|1}(t)}], \text{ by the assumption of (9.21)}, \\
& = E_0[b(y(t+1))],
\end{aligned}$$

because the next to last formula is a deterministic function after which one takes expectations of the obtained expressions and equates these expressions. The above result and Theorem 2.8.2.(f) prove that, with respect to the probability measure P_0, $F^{y(t+1)}$ is independent of $F_t^x \vee F_t^y$ for all $t \in T\backslash\{t_1\}$, hence that $\{y(t),\ t \in T\}$ is an independent sequence.

(b) Note that then,

$$\begin{aligned}
& E_1[b(x(t+1))|F_t^y] = \frac{E_0[b(x(t+1))r_{1|0}(t_1)|F_t^y]}{E_0[r_{1|0}(t_1)|F_t^y]}, \text{ by Theorem 19.9.10}, \\
& = \frac{E_0[b(x(t+1))E_0[r_{1|0}(t_1)|F_{t+1}^x \vee F_{t+1}^y]|F_t^y]}{E_0[E_0[r_{1|0}(t_1)|F_t^x \vee F_t^y]|F_t^y]} = \frac{E_0[b(x(t+1))r_{1|0}(t+1)|F_t^y]}{E_0[r_{1|0}(t)|F_t^y]} \\
& = \frac{E_0[E_0[b(x(t+1))r_{1|0}(t+1)|\ F_t^x \vee F_t^y]|\ F_t^y]}{E_0[r_{1|0}(t)|F_t^y]} \\
& = \frac{E_0[E_1[b(x(t+1))|\ F_t^x \vee F_t^y]\ r_{1|0}(t)|\ F_t^y]}{E_0[r_{1|0}(t)|F_t^y]} \text{ by (a.c).}
\end{aligned}$$

□

The derivation of a filter for a particular stochastic system requires a discussion. The reader may find examples of several filters in the subsequent sections. In particular, it has to be explained which stochastic systems can admit a finite-dimensional filter system.

The calculation of a filter system differs significantly from the calculation of an estimator and of a sequential estimator as discusses in the previous sections.

To derive the conditional distribution of the next state conditioned on past observations, Theorem 9.5.3 prescribes to calculate the following expression,

$$\begin{aligned} & E_0[\, E_1[b(x(t+1))|\, F_t^x \vee F_t^y]\; r_{1|0}(t)|\, F_t^y] \\ & = E_0[\, E_1[b(x(t+1))|\, F^{x(t)}]\; r_t(x(t),y(t)) r_{1|0}(t-1)|\, F_t^y]; \qquad (9.24) \\ & r_{1|0}(t) = r_t(x(t),y(t))\; r_{1|0}(t-1). \end{aligned}$$

The above equality holds because the stochastic system with the state and output process $(x,\ y)$ is a stochastic system, and because of the definition of the Radon-Nikodym martingale $r_{1|0}$.

The calculation amounts to the steps,

$$b_2(x(t),y(t)) = E_1[\exp(iw\, x(t+1))|\, F^{x(t)}]\; r_t(x(t),y(t)); \qquad (9.25)$$

$$E_0[b_2(x(t),y(t))\; r_{1|0}(t-1)|\, F_t^y], \qquad (9.26)$$

where the expression of equation (9.25) is a function of the random variables $(x(t),\ y(t))$ only. That expression is to be integrated over the conditional probability distribution of $x(t)|\, F_t^y$ with respect to the probability measure P_0 which is known from the previous step. It can be proven that it is identical to the conditional probability distribution of $x(t)|\, F_{t-1}^y$.

To be able to calculate the expression (9.24) one has (1) to first calculate the product of the term of equation (9.25) and (2) integrate the result over the conditional distribution of the previous step according to equation 9.26. In the second calculation the random variable $y(t)$ is considered known due to the conditioning on F_t^y while $x(t)$ and $y(t)$ are independent with respect to P_0. If one wants invariance of the conditional distribution then algebraic properties have to be satisfied.

In sequential estimation, according to Theorem 9.3.3.(b), one has to calculate the expression,

$$E_0[E_1[\exp(iw\, x)]\; r_{1|0}(x,y)|\, F^y]. \qquad (9.27)$$

In this case, the expression $E_1[\exp(iw\, x)]$ is a particular characteristic function depending only on deterministic parameters whose values are assumed known. The integration of equation (9.27) is therefore simpler to calculate than equation (9.26).

9.6 Filter of a Poisson-Gamma System

In this section the filter problem is solved for an output process of which at each time the output value belongs to the natural numbers, $\mathbb{N} = \{0,1,\ldots\}$. The conditional distribution of the output conditioned on the current state is Poisson while the conditional distribution of the next state conditioned on past states is a Gamma pdf. Hence the associated system is called a *Poisson-Gamma system*, see Def. 5.3.3.

For the Poisson-Gamma models, one can solve the estimation problem, the sequential estimation problem, and a filter problem.

Definition 9.6.1. Define the *Poisson-Gamma stochastic system* as a stochastic system with the following objects and relations.

$$\begin{aligned}
& (\Omega, F, P_1),\ T(0:t_1) = \{0,1,\ldots,t_1\},\ t_1 \in \mathbb{Z}_+, \\
& x : \Omega \times T \to \mathbb{R}_+,\ y : \Omega \times T \to \mathbb{N}, \\
x(t+1) &= a\,x(t) + v(t),\ x(0) = x_0, \\
& E_1[I_{\{y(t)=k\}} | F^x_t \vee F^y_{t-1}] = \frac{x(t)^k}{k!} \exp(-x(t)) \\
& \{v(t),\ t \in T\},\ v : \Omega \times T \to \mathbb{R}_+,\ \text{a sequence of independent rvs}, \\
& \text{pdf}(v(t)) \in \Gamma(\gamma_{v,1}(t), \gamma_{v,2}(t)),\ \forall\, t \in T, \\
\gamma_{v,2}(t+1) &= \frac{a\gamma_{v,2}(t)}{(1+\gamma_{v,2}(t))},\ \gamma_{v,2}(0) = 1;\ a \in (0,\infty), \\
& \text{pdf}(x_0) \in \Gamma(\gamma_{x_0,1}, \gamma_{x_0,2}),\ F^{x_0},\ F^v_{t_1}\ \text{are independent } \sigma\text{-algebras}, \\
y(t) &= x(t) + w(t), \\
& w : \Omega \times T \to \mathbb{R},\ (\sum_{s=0}^{t} w(s), F^x_t \vee F^y_{t-1},\ t \in T\}) \in M_1(P_1).
\end{aligned}$$

The assumed recursion of the function $\gamma_{v,2} : T \to \mathbb{R}_+$ is rather special but without it the filter system does not have the nice formulation as stated below.

The relevant Radon-Nikodym derivative is derived. If with respect to P_1 the output $y(t)$ has a conditional Poisson distribution with parameter $x(t)$ and with respect to P_0 the output $y(t)$ has a conditional Poisson distribution with parameter 1, then,

$$\begin{aligned}
r_{1|0}(t) &= r_{1|0}(t-1) \sum_{k=0}^{\infty} I_{\{y(t)=k\}} \frac{x(t)^k \exp(-x(t))/k!}{1^k \exp(-1)/k!} \\
&= r_{1|0}(t-1)\, x(t)^{y(t)}\, \exp(-x(t)+1),\ r_{1|0}(0) = 1, \\
r_{0|1}(t) &= r_{1|0}(t)^{-1} = r_{0|1}(t-1)\, x(t)^{-y(t)}\, \exp(x(t)-1),\ \forall\, t \in T,\ r_{0|1}(0) = 1.
\end{aligned}$$

Theorem 9.6.2. *Consider the Poisson-Gamma system of Def. 9.6.1. Define the stochastic process,*

$$\begin{aligned}
& \{r_{0|1}(t), F^x_t \vee F^y_t, t \in T\},\ r_{0|1}(0) = 1, \\
r_{0|1}(t) &= r_{0|1}(t-1)\, x(t)^{-y(t)} \exp(x(t)-1).
\end{aligned}$$

The conditional characteristic function of the state based on past observations is of Gamma type with representation,

$$
\begin{aligned}
& E_1[\exp(iw_x\, x(t+1))|F_t^y] \\
&= (1 - iw_x\, \gamma_2(t+1))^{-\hat{\gamma}_1(t+1)},\ \forall\, t \in T, && (9.28)\\
& \textit{which is a Gamma cpdf with parameters } (\hat{\gamma}_1(t+1), \gamma_2(t+1)),\\
\hat{\gamma}_1(t+1) &= \hat{\gamma}_1(t) + \gamma_{v,1}(t) + y(t),\ \gamma_1(0) = \gamma_{x_0,1}, && (9.29)\\
\gamma_2(t+1) &= \frac{a\gamma_2(t)}{1+\gamma_2(t)},\ \gamma_2(0) = \gamma_{x_0,2}, && (9.30)\\
\hat{x}(t) &= \hat{\gamma}_1(t)\gamma_2(t);\\
\hat{x}(t+1) &= \frac{a\gamma_2(t)}{1+\gamma_2(t)}\hat{x}(t) + \frac{a\gamma_2(t)}{1+\gamma_2(t)}\ [y(t) - \hat{x}(t)] + [\hat{\gamma}_1(t) + \gamma_{v,1}(t)]\frac{a\gamma_2(t)}{1+\gamma_2(t)}\\
& \hat{x}(0) = \gamma_1(0)\gamma_2(0). && (9.31)
\end{aligned}
$$

Proof. From the system definition it follows that for all $t \in T$, $x(t) > 0$. Then one proves by induction that $r_{0|1}(t) > 0$ for all $t \in T$. By definition of the Radon-Nikodym expression $r_{0|1}$ it follows that $r_{0|1}(t)$ is $F_t^x \vee F_t^y$ measurable for all $t \in T$.

It is proven that $r_{0|1}$ is a martingale with respect to P_1. By definition, $E_1[r_{0|1}(0)] = 1$. Suppose that $E_1[r_{0|1}(s) = 1$ for $s = 0, 1, \ldots, t$. Then it will be proven for $t+1$. Consider any bounded measurable function $b : \mathbb{N} \to \mathbb{R}_+$.

$$
\begin{aligned}
& E_1[b(y(t+1))\ r_{0|1}(t+1)|\ F_t^x \vee F_t^y]\\
&= E_1[b(y(t+1))\ x(t+1)^{-y(t+1)} \exp(x(t+1)-1)|\ F_t^x \vee F_t^y] r_{0|1}(t)\\
&= E_1[\sum_{k=0}^{\infty} E_1[I_{\{y(t+1)=k\}}|\ F_{t+1}^x \vee F_t^y]\ b(k)\ x(t+1)^{-k}\ \times\\
&\quad \times\ \exp(x(t+1)-1)|\ F_t^x \vee F_t^y]\ r_{0|1}(t)\\
&= \sum_{k=0}^{\infty} x(t+1)^k \exp(-x(t+1))\ b(k)\ x(t+1)^{-k} \exp(x(t+1)-1)/k!\ r_{0|1}(t)\\
&= \exp(-1) \sum_{k=0}^{\infty} b(k)/k!\ r_{0|1}(t);\\
& E_1[b(y(t+1))\ r_{0|1}(t+1)]\\
&= \exp(-1) \sum_{k=0}^{\infty} b(k)/k!\ E_1[r_{0|1}(t)]\\
&= \exp(-1) \sum_{k=0}^{\infty} b(k)/k! = E_1[b(y(t+1))\ r_{0|1}(t+1)|\ F_t^x \vee F_t^y].
\end{aligned}
$$

The last equality follows because the expression of the preceding line is deterministic. If one sets in the above equality $b(.) = 1$ then it follows that $E_1[r_{0|1}(t+1)|F_t^x \vee F_t^y] = r_{0|1}(t)$ hence $\{r_{0|1}(t),\ F_t^x \vee F_t^y,\ t \in T\}$.

Then the conditions of Theorem 9.5.3 are satisfied. Then it follows from that theorem that the formula $dP_0/dP_1 = r_{0|1}(t_1)$ defines a probability measure P_0 on $(\Omega, F_{t_1}^x \vee F_{t_1}^y)$, that P_0 and P_1 are equivalent probability measures, that with respect to P_0 the output process is a sequence of independent random variables, and that,

$$\begin{aligned}
&\{r_{1|0}(t),\ F^x_{t_1} \vee F^y_t,\ t \in T\} \in M_1(P_0),\\
&r_{1|0}(t) = r_{0|1}(t)^{-1} = r_{1|0}(t-1) r_t(x(t), y(t)),\\
&r_t(x(t), y(t)) = x(t)^{y(t)} \exp(-x(t)+1),\\
&E_0[\exp(iw\, y(t))] = \exp(\exp(iw) - 1),\ \forall\, t \in T(0 : t_1 - 1),\\
&E_0[\exp(iw\, x(t))] = E_1[\exp(iw\, x(t))].
\end{aligned}$$

From the specification of the Poisson-Gamma system follows that, $F^{v(t)}$ is independent of $F^{x_0} \vee F^v_{t-1}$, that $F^{x(t)} \subseteq F^{x_0} \vee F^v_{t-1}$ for all $t \in T$, hence $F^{v(t)}$ is independent of F^x_t. Similarly, $F^{y(t)} \subseteq F^{x_0} \vee F^v_{t-1} \vee F^w_t$. By induction one then proves that for all $t \in T$, $F^{v(t)}$ independent of $F^x_t \vee F^y_t$. Note further that,

$$\begin{aligned}
&E_1[\exp(iw\, x(t+1))|\, F^x_t \vee F^y_t] = E_1[\exp(iw\, v(t))|\, F^x_t \vee F^y_t]\ \exp(iwa\, x(t))\\
&\text{because of the recursion of the state process,}\\
&= E_1[\exp(iw\, v(t))]\ \exp(iwa\, x(t)) \text{ because } F^{v(t)},\ F^x_t \vee F^y_t \text{ are independent,}\\
&= \exp(iwa\, x(t))\ (1 - iw\, \gamma_{v,2}(t))^{-\gamma_{v,1}(t)},\ \forall\, t \in T. \qquad (9.32)
\end{aligned}$$

The proof proceeds by induction. The initial step is calculated by,

$$\begin{aligned}
&E_1[\exp(iw\, x(0))] = (1 - iw\gamma_2(0))^{-\hat{\gamma}_1(0)},\\
&E_1[\exp(iw\, x(1))|\, F^y_0] = \frac{E_0[\exp(iw\, x(1)) r_{1|0}(0)|\, F^y_0]}{E_0[r_{1|0}(0)|\, F^y_0]} = E_0[\exp(iwx(1))],\\
&\text{because } r_{1|0}(0) = 1 \text{ and } F^{x_1} \text{ and } F^y_0 \text{ are independent wrt to } P_0,\\
&= E_0[\exp(iw\, ax(0))]\ E_0[\exp(iw\, v(0))], \text{ by independence of } x_0 \text{ and } v(0),\\
&= (1 - iw\, a\gamma_2(0))^{-\gamma_1(0)} (1 - iw\gamma_{v,2}(0))^{-\gamma_{v,1}(0)}\\
&= (1 - iw\, \gamma_2(1))^{-\gamma_1(1)}.
\end{aligned}$$

because $x(0)$ and $v(1)$ are independent wrt P_1 hence wrt to P_0, $\gamma_{2,v}(1) = a\gamma_2(1)$ and $\gamma_1(1) = \gamma_1(0) + \gamma_{v,1}(1)$.

Suppose that for $s = 1, 2, \ldots, t$

$$\begin{aligned}
&E_1[\exp(iw\, x(s))|F^y_{s-1}] = (1 - iw\, \gamma_2(s))^{-\hat{\gamma}_1(s)},\\
&E_0[\exp(iw\, x(s)) r_{1|0}(s)|F^y_{s-1}]\\
&= E_1[\exp(iw\, x(s))|F^y_{s-1}] E_0[r_{1|0}(s)|\, F^y_{s-1}] \text{ hence by Theorem 9.5.3.(b),}\\
&= (1 - iw\, \gamma_2(s))^{-\hat{\gamma}_1(s)}\ E_0[r_{1|0}(s)|\, F^y_{s-1}],\\
&= E_0[r_{1|0}(s)|\, F^y_{s-1}]\ \times\\
&\times \int_0^\infty \exp(iw\, v)\ v^{\hat{\gamma}_1(s)-1}\ \exp(-v/\gamma_2(s)) dv \gamma_2(s)^{\hat{-}\gamma_1(s)} / \Gamma(\hat{\gamma}_1(s)). \qquad (9.33)
\end{aligned}$$

Next one proves that the latter formulas also holds for $s = t+1$. One calculates using Theorem 9.5.3.(b),

$$
\begin{aligned}
& E_0[E_1[\exp(iw\, x(t+1))|\ F_t^x \vee F_t^y]\ r_{1|0}(t)|\ F_t^y] \\
&= E_0[\exp(iwa\, x(t))\ x(t)^{y(t)}\ \exp(-x(t)+1) r_{1|0}(t-1)|\ F_t^y] \times \\
&\quad \times\ (1-iw\ \gamma_{v,2}(t))^{-\gamma_{v,1}1(t)} \text{ by equation (9.32),} \\
&= \exp(1)\ E_0[x(t)^{y(t)} \exp(-x(t)\ [1-iwa])\ r_{1|0}(t-1)|\ F_t^y] \times \\
&\quad \times\ (1-iw\ \gamma_{v,2}(t))^{-\gamma_{v,1}(t)} \\
&= \exp(1) \int_0^\infty\ v^{y(t)}\ \exp(-v[1-iwa])\ v^{\hat{\gamma}_1(t)-1}\ \exp(-v/\gamma_2(t))\ dv \\
&\quad (1-iw\ \gamma_{v,2}(t))^{-\gamma_{v,1}(t)} \gamma_2(t)^{-\hat{\gamma}_1(t)} E_0[r_{1|0}(t)|F_{t-1}^y])/\Gamma(\hat{\gamma}_1(t)) \\
&\quad \text{by the induction step and (9.33),} \\
&= \exp(1) \int_0^\infty\ v^{\hat{\gamma}_1(t)+y(t)-1}\ \exp(-v[1-iwa+1/\gamma_2(t)]) dv \\
&\quad (1-iw\ \gamma_{v,2}(t))^{-\gamma_{v,1}(t)} \gamma_2(t)^{-\hat{\gamma}_1(t)} E_0[r_{1|0}(t)|F_{t-1}^y])/\Gamma(\hat{\gamma}_1(t)) \\
&= \exp(1)\ [1+\frac{1}{\gamma_2(t)}-iwa]^{-[\hat{\gamma}_1(t)+y(t)]} (1-iw\gamma_{v,2}(t))^{-\gamma_{v,1}(t)}\ \times \\
&\quad \times\ (\gamma_2(t))^{-\hat{\gamma}_1(t)} E_0[r_{1|0}(t)|F_{t-1}^y]) \Gamma(\hat{\gamma}_1(t)+y(t))/\Gamma(\hat{\gamma}_1(t)) \\
&\quad \text{using the integration formula (2.5);} \\
&= \exp(1)\ \left[1-iw\frac{a\gamma_2(t)}{1+\gamma_2(t)}\right]^{-[\hat{\gamma}_1(t)+y(t)]} (1-iw\gamma_{v,2}(t))^{-\gamma_{v,1}(t)}\ \times \\
&\quad \times\ \left[\frac{1+\gamma_2(t)}{\gamma_2(t)}\right]^{-[\hat{\gamma}_1(t)+y(t)]} (\gamma_2(t))^{-\hat{\gamma}_1(t)} \Gamma(\hat{\gamma}_1(t)+y(t)) E_0[r_{1|0}(t)|F_{t-1}^y]) \times \\
&\quad /\Gamma(\hat{\gamma}_1(t)) \\
&= \exp(1)\ [1-iw\gamma_2(t+1)]^{-[\hat{\gamma}_1(t)+\gamma_{v,1}(t)+y(t)]} \\
&\quad \times\ \left[\frac{1+\gamma_2(t)}{\gamma_2(t)}\right]^{-[\hat{\gamma}_1(t)+y(t)]} (\gamma_2(t))^{-\hat{\gamma}_1(t)}\ \Gamma(\hat{\gamma}_1(t)+y(t))\ E_0[r_{1|0}(t)|F_{t-1}^y]) \times \\
&\quad /\Gamma(\hat{\gamma}_1(t)),\ \text{if } \gamma_2(t+1) = \frac{a\gamma_2(t)}{1+\gamma_2(t)} = \gamma_{v,2}(t), \\
&\quad E_1[\exp(iw\, x(t+1))|\ F_t^y] = \frac{E_0[E_1[\exp(iw\, x(t+1))|\ F_t^x \vee F_t^y] r_{1|0}(t)|\ F_t^y]}{E_0[r_{1|0}(t)|\ F_t^y]} \\
&= [1-iw\ \gamma_2(t+1)]^{-\hat{\gamma}_1(t+1)}, \\
&\quad \hat{\gamma}_1(t+1) = \hat{\gamma}_1(t) + \gamma_{v,1}(t) + y(t),\ \gamma_2(t+1) = \frac{a\gamma_2(t)}{1+\gamma_2(t)} = \gamma_{v,1}(t).
\end{aligned}
$$

□

9.7 Filter of an Output-Finite-State-Finite Stochastic System

The reader finds in this section the solution to the filter problem of a finite stochastic system by the measure transformation method. The result is known for several decades and used in signal processing. The formulation and the proof below are different from the literature.

Problem 9.7.1. Consider a time-invariant output-finite-state-finite stochastic system. Assume that state-output conditional independence holds when conditioned on past states and outputs as defined in Def. 5.7.4. The indicator representation of the system is used. Assume that the system transition matrix A is irreducible and nonperiodic.

The system representation is then,

$$E_1\left[\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix} \mid F_t^x \vee F_{t-1}^y\right] = \begin{pmatrix} A \\ C \end{pmatrix} x(t),\ x(0) = x_0,$$
$$T = T(0:t_1) = \{0,1,\ldots,t_1\},$$
$$X_e = \{e_1, e_2, \ldots, e_{n_x} \in \mathbb{R}_{st}^{n_x}\},\ Y_e = \{e_1, e_2, \ldots, e_{n_y} \in \mathbb{R}_{st}^{n_y}\},$$
$$A \in \mathbb{R}_{st}^{n_x \times n_x},\ C \in \mathbb{R}_{st}^{n_y \times n_x},\ \exists\ p_s \in \mathbb{R}_{s+}^{n_x} \text{ such that } Ap_s = p_s;$$
$$p_x(0) = p_s\ \Rightarrow\ \forall\ t \in T,\ p_x(t) = p_s.$$

Because the state transition matrix A is irreducible and nonperiodic, there exists by Theorem 18.8.2.(b) a steady-state stochastic vector $p_s \in \mathbb{R}_{s+}^{n_x}$ such that $Ap_s = p_s$, hence, for all $i \in \mathbb{Z}_{n_x}$, $p_{s,i} > 0$.

The problem is to determine the conditional probability distribution,

$$\hat{x}(t+1) = E_1[x(t+1)|\ F_t^y],\ \forall\ t \in T.$$

The vector $\hat{x}(t+1) = E_1[x(t+1)|F_t^y]$ is the conditional distribution of $x(t+1)$ conditioned on F_t^y because the state set is finite and because the vector $x(t)$ is in the indicator representation.

The reader is again alerted of the fact that in the literature on the Hidden Markov Model (HMM) the following filter problem is mainly treated,

$$\hat{x}(t|t) = E_1[x(t)|\ F_t^y],\ \forall\ t \in T.$$

The formulas for the latter expression differs from that of Problem 9.7.1 though there is of course a relation.

Theorem 9.7.2. *Consider Problem 9.7.1 with the assumptions stated there.*

(a)Define the stochastic process,

$$r_{0|1}(t) = \prod_{s=1}^{t} \frac{1/n_y}{x(s)^T C^T y(s)},\ \forall\ t \in T\backslash\{0\},\ r_{0|1}(0) = 1,\ r_{0|1} : \Omega \times T \to \mathbb{R}_+.$$

Then the assumptions of Theorem 9.5.3 and hence the conclusions of that theorem all hold. Note that then,

$$r_{1|0}(t+1) = r_{0|1}^{-1}(t+1) = x(t+1)^T C^T y(t+1) n_y\ r_{1|0}(t),\ r_{1|0}(0) = 1,$$
$$r_{1|0} : \Omega \times T \to \mathbb{R}_+.$$

(b)The conditional distribution can be recursively calculated according to,

$$\begin{aligned}
\hat{x}_{un}(t+1) &= A\ \mathrm{Diag}(C^T y(t))\ \hat{x}_{un}(t),\ \hat{x}_{un}(0) = A p_x(0),\\
\hat{x}(t+1) &= E_1[x(t+1)|\ F_t^y] = \hat{x}_{un}(t+1)/[1_{n_x}^T \hat{x}_{un}(t+1)],\ \forall\, t \in T,\\
&\hat{x}_{un} : \Omega \times T \to \mathbb{R}_+^{n_x},\ \hat{x} : \Omega \times T \to \mathbb{R}_{st}^{n_x}.
\end{aligned}$$

(c)The filter error satisfies the equations,

$$\begin{aligned}
e(t) &= x(t) - \hat{x}(t),\ \ E_1[e(t+1)|\ F_t^y] = 0,\\
&E_1[e(t+1)e(t+1)^T|\ F_t^y] = \mathrm{Diag}(\hat{x}(t+1)) - \hat{x}(t+1)\hat{x}(t+1)^T\\
&= \begin{cases} \hat{x}_i(t+1)[1-\hat{x}_i(t+1)],\ i = j, \\ -\hat{x}_i(t+1)\hat{x}_j(t+1), \quad\ \ i \neq j, \end{cases} \ \forall\, t \in T,\\
&Q_{\hat{x}}(t) = E[\hat{x}(t)\hat{x}(t)^T] = Q_x(t) - Q_e(t) \preceq Q_x(t) = E_1[x(t)x(t)^T].
\end{aligned}$$

Proof. (a) Because the system is started in the invariant probability measure, it follows from Corollary 5.7.10 that,

$$\begin{aligned}
&\forall\, t \in T,\ p_x(t) = E_1[x(t)] = A^t E_1[x(0)] = A^t p_x(0) = A^t p_s = p_s \in \mathbb{R}_{sst}^{n_x};\\
&\Rightarrow \forall\, i \in \mathbb{Z}_{n_x},\ p_{s,i}(t) > 0.
\end{aligned}$$

Because, for any time $t \in T$, $y(t)$ is in the indicator representation, there exists an integer $k \in \mathbb{Z}_{n_y}$ such that $y_k(t) = 1$ while for all $j \in \mathbb{Z}_{n_y}\backslash\{k\}$, $y_j(t) = 0$. Because C is a stochastic matrix and because $x(t)$ is also in the indicator representation, $Cx(t) \in \mathbb{R}_{st}^{n_x}$ is a probability measure for $y(t)$ hence $1_{n_y}^T Cx(t) = 1$. It will be argued that then $x(t)^T C^T y(t) > 0\ a.s.$ If $0 = x(t)^T C^T y(t) = (Cx(t))^T y(t) = (Cx(t))_k$, by the above assumption of the existence of k, then the probability that $y(t) = e_k$ is zero by the definition of a finite stochastic system. But this contradicts that $y(t) = e_k\ a.s.$ Thus $0 < x(t)^T C^T y(t)$ for all $t \in T$. Hence,

$$\forall\, t \in T,\ \ r_{0|1}(t) = \prod_{s=0}^{t} \frac{1}{n_y x(s)^T C^T y(s)} > 0\ a.s.$$

Note that,

$$\begin{aligned}
&E_1[r_{0|1}(t+1)|\ F_t^x \vee F_t^y]\\
&= E_1[E_1[r_{0|1}(t+1)|\ F_{t+1}^x \vee F_t^y]|\ F_t^x \vee F_t^y]\\
&= E_1[E_1[\frac{1}{n_y x(t+1)^T C^T y(t+1)}|F_{t+1}^x \vee F_t^y] r_{0|1}(t)|\ F_t^x \vee F_t^y]\\
&= E_1[E_1[\sum_{k=1}^{n_y} I_{\{y(t+1)=e_k\}}\ \frac{1}{n_y(Cx(t+1))_k}|F_{t+1}^x \vee F_t^y] r_{0|1}(t)|\ F_t^x \vee F_t^y]\\
&= \sum_{k=1}^{n_y} E_1[E_1[I_{\{y(t+1)=e_k\}}\ |F_{t+1}^x \vee F_t^y] \frac{1}{n_y(Cx(t+1))_k} r_{0|1}(t)|\ F_t^x \vee F_t^y]\\
&= \sum_{k=1}^{n_y}\ (1/n_y)\ r_{0|1}(t) = r_{0|1}(t),\ \text{because } E_1[y(t+1)|F_{t+1}^x \vee F_t^y] = Cx(t+1),\\
&\Rightarrow (r_{0|1}(t),\ F_t^x \vee F_t^y,\ t \in T) \in M_1(P_1).
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&E_1[y(t+1)r_{0|1}(t+1)/r_{0|1}(t)|F_t^x \vee F_t^y] \\
&E_1[E_1[y(t+1)r_{0|1}(t+1)/r_{0|1}(t)|F_{t+1}^x \vee F_t^y]|\ F_t^x \vee F_t^y] \\
= {} & E_1[E_1[y(t+1)\frac{1}{n_y x(t+1)^T C^T y(t+1)}|F_{t+1}^x \vee F_t^y]|\ F_t^x \vee F_t^y] \\
= {} & E_1[E_1[\sum_{k=1}^{n_y} I_{\{y(t+1)=e_k\}}\ e_k \frac{1}{n_y(Cx(t+1))_k}|\ F_{t+1}^x \vee F_t^y]|\ F_t^x \vee F_t^y] \\
= {} & E_1[E_1[\sum_{k=1}^{n_y} I_{\{y(t+1)=e_k\}}|\ F_{t+1}^x \vee F_t^y] e_k \frac{1}{n_y(Cx(t+1))_k}|\ F_t^x \vee F_t^y] \\
= {} & \sum_{k=1}^{n_y} e_k/n_y = 1_{n_y}\ /n_y = E_1[y(t+1)r_{0|1}(t+1)/r_{0|1}(t)],
\end{aligned}$$

because the next to last expression is deterministic, the conditional expectation equals its expectation. The conditions of Theorem 9.5.3 are then satisfied. It follows from that theorem that there exists a probability measure P_0 on $(\Omega, F_{t_1}^x \vee F_{t_1}^y)$ such that $P_0 \sim P_1$ with

$$r_{1|0}(t) = r_{0|1}(t)^{-1} = \prod_{s=1}^{t_1} x(s)^T C^T y(s) n_y,\ (r_{1|0}(t),\ F_t^x \vee F_t^y,\ t \in T) \in M_1(P_0),$$

that for all $t \in T \backslash \{t_1\}$, $F^{y(t+1)}$ and $F_{t+1}^x \vee F_t^y$ are independent with respect to P_0, and that, with respect to P_0, the $\{y(t),\ t \in T\}$ is a sequence of independent random variables.

(b) It follows from Theorem 9.5.3.(b) that the following conditional expectation satisfies the specified relations,

$$E_1[x(t+1)|F_t^Y] = \frac{E_0[E_1[x(t+1)|\ F_t^x \vee F_t^y]\ r_{1|0}(t)|F_t^y]}{E_0[r_{1|0}(t)|F_t^y]}.$$

It will be proven by induction that, if,

$$\begin{aligned}
\hat{x}_{un}(t) &= E_0[x(t)r_{1|0}(t)|F_{t-1}^y],\ \ \hat{x}_{un} : \Omega \times T \rightarrow \mathbb{R}^{n_x},\ \text{then,} \\
\hat{x}_{un}(t+1) &= A\ \text{Diag}(\hat{x}_{un}(t))\ C^T y(t) = A\text{Diag}(C^T y(t))\hat{x}_{un}(t), \\
\hat{x}(t+1) &= E_1[x(t+1)|F_t^y] = \frac{\hat{x}_{un}(t+1)}{1_{n_x}^T \hat{x}_{un}(t+1)},\ \forall\, t \in T.
\end{aligned}$$

The initial step proceeds by,

$$\begin{aligned}
\hat{x}_{un}(1) &= E_0[x(1)r_{1|0}(1)|\ F_0^y] = E_0[E_0[x(1)r_{1|0}(1)|\ F_0^x \vee F_0^y]|\ F_0^y] \\
&= E_0[E_1[x(1)|\ F_0^x \vee F_0^y]\ r_{1|0}(0)|\ F_0^y],\ \text{by Theorem 9.5.3.(a.c),} \\
&= E_0[Ax(0)|\ F_0^y],\ \text{by the system representation and by } r_{1|0}(0) = 1, \\
&= AE_0[x(0)|F_0^y] = AE_0[x(0)], \\
&\quad \text{because by Theorem 9.5.3.(a.c), wrt } P_0,\ F_0^x \text{ and } F_0^y \text{ are independent,} \\
&= AE_1[x(0)r_{0|1}(0)] = Ap_x(0),\ \text{by } r_{0|1}(0) = 1, \\
\hat{x}_{un}(1) &= Ap_x(0).
\end{aligned}$$

For the induction step one calculates,

$$
\begin{aligned}
&E_0[x(t+1)r_{1|0}(t+1)|F_t^y] \\
&= E_0[E_0[x(t+1)r_{1|0}(t+1)|F_t^x \vee F_t^y]|\; F_t^y] \\
&= E_0[E_1[x(t+1)|F_t^x \vee F_t^y]\; r_{1|0}(t)|\; F_t^y],\text{ by Theorem 9.5.3.(b),} \\
&= E_0[Ax(t)\; r_{0|1}(t)|F_t^y],\text{ by the system representation,} \\
&= E_0[Ax(t)\; x(t)^T C^T y(t)\; n_y r_{1|0}(t-1)|\; F_t^y],\text{ by } r_{1|0}(t), \\
&= AE_0[\text{Diag}(x(t))r_{1|0}(t-1)|\; F^{y(t)} \vee F_{t-1}^y]C^T y(t) n_y, \\
&\quad \text{by the indicator representation } x(t)x(t)^T = \text{Diag}(x(t)), \\
&= AE_0[\text{Diag}(x(t))r_{1|0}(t-1)|\; F_{t-1}^y]C^T y(t) n_y, \\
&\quad \text{because } (F^{x(t)}, F^{y(t)}|\; F_{t-1}^y) \in \text{CI}(P_0),\text{ by Theorem 9.5.3.(c),} \\
&\quad \text{and because } r_{1|0}(t-1) \text{ is } F_{t-1}^x \vee F_{t-1}^y \text{ measurable,} \\
&= A\;\text{Diag}(\hat{x}_{un}(t))\; C^T y(t)\; n_y,\text{ by the induction step,} \\
&= A\;\text{Diag}(C^T y(t))\; \hat{x}_{un}(t)\; n_y, \\
&E_1[x(t+1)|F_t^y] = \frac{A\text{Diag}(C^T y(t))\; \hat{x}_{un}(t)}{1_{n_x}^T \hat{x}_{un}(t+1)} = \frac{\hat{x}_{un}(t+1)}{1_{n_x}^T \hat{x}_{un}(t+1)}, \\
\hat{x}_{un}(t+1) &= A\;\text{Diag}(C^T y(t))\; \hat{x}_{un}(t).
\end{aligned}
$$

(c) The formulas for the filter error system follow from,

$$
\begin{aligned}
&E[e(t+1)|\; F_t^y] = E[x(t+1) - \hat{x}(t+1)|\; F_t^y] \\
&= E[x(t+1)|\; F_t^y] - \hat{x}(t+1) = 0, \\
&E[e(t+1)e(t+1)^T|\; F_t^y] = E[x(t+1)x(t+1)^T|\; F_t^y] - \hat{x}(t+1)\hat{x}(t+1)^T \\
&= \text{Diag}(\hat{x}(t+1)) - \hat{x}(t+1)\hat{x}(t+1)^T; \\
Q_x(t) &= E[x(t)x(t)^T] = E[(x(t) - \hat{x}(t) + \hat{x}(t))(\ldots)^T] \\
&= Q_e(t) + Q_{\hat{x}}(t) \;\rightarrow\; Q_{\hat{x}}(t) = Q_x(t) - Q_e(t) \prec Q_x(t),\; \forall\, t \in T.
\end{aligned}
$$

□

Theorem 9.7.3. *Consider the filter problem for the finite stochastic system of Problem 9.7.1 The* filter realization *of this finite stochastic system is specified by the representation,*

$$
\begin{aligned}
&E_1\left[\begin{pmatrix} \hat{x}_{un}(t+1) \\ y(t) \end{pmatrix} |\; F_{t-1}^y\right] \\
&= \begin{pmatrix} A\;\text{Diag}\left(C^T C\hat{x}_{un}(t)[1^T \hat{x}_{un}(t)]^{-1}\right)\; \hat{x}_{un}(t), \\ C\hat{x}_{un}(t)[1^T \hat{x}_{un}(t)]^{-1} \end{pmatrix},\; \hat{x}_{un}(0) = Ap_s = p_s, \\
\hat{x}(t+1) &= E_1[x(t+1)|\; F_t^y] = \hat{x}_{un}(t+1)/[1_{n_x}^T \hat{x}_{un}(t+1)], \\
\hat{X}_{un}(t+1) &= \{A\text{Diag}(C^T e_i)\hat{x}_{un}(t) \in \mathbb{R}_+^{n_x},\; \forall\, i \in \mathbb{Z}_{n_y}\},\;\; E_1[x(t+1)|\; F_t^y] \in Y_e.
\end{aligned}
$$

The understanding of the above defined stochastic system is that the generation of the random variable $y(t)$ condition on the observation filtration $\{F_{t-1}^y,\; t \in T\}$ by a *randomization mechanism selects a value in the set Y_e of unit vectors of $\mathbb{R}_+^{n_y}$ based on the conditional distribution while the generation of $\hat{x}_{un}(t+1)$ selects a value in the subset $\hat{X}_{un}(t+1) \subset \mathbb{R}_+^{n_x}$ based on the conditional distribution.*

Proof. This follows from Theorem 9.7.2 and from,

$$\begin{aligned} E[y(t)|\, F_{t-1}^y] &= E[E[y(t)|F_t^x \vee F_{t-1}^y]|\, F_{t-1}^y] = E[Cx(t)|\, F_{t-1}^y] \\ &= C\hat{x}(t) = C\hat{x}_{un}(t)/[1^T \hat{x}_{un}(t)]. \end{aligned}$$

□

9.8 Further Reading

History. Eugene Wong of the University of California at Berkeley, has introduced the author to the measure transformation method for filter problems. The method for continuous-time systems is described in the book [46, Section 6.5] and its second edition [47]. A source of inspiration for the author was a conference paper by P. Frost, [16]

The method of this chapter to solve filter problems makes use of the measure transformation method. That method originates in statistics. It was generalized to continuous-time processes by I.V. Girsanov [17], and E. Wong etal, see [41]. The reader may want to read Section 19.9 for an exposition of the measure transformation method.

The main research issue of the filter problem is to formulated algebraic conditions which result in a finite-dimensional filter system. Only finite-dimensional filter systems or filter systems with a finite state set can be implemented in practice. To solve filter problems for a wider class than Gaussian systems and finite stochastic systems, one has to investigate algebraic conditions. The concept of an *invariant conditional distribution* was formulated by J. Bather, [7].

It was later discovered that the concept of conjugate probability distributions is a sufficient condition for the existence of conditional invariant distributions. The concept of a tuple of conjugate probability distributions was formulated in the book, [32]. This concept is also treated in [15, Ch. 9]. Conjugate priors for probability distributions which are members of the exponential family are described in [2].

The examples of filter problems presented in this chapter were then formulated by the author. The formulation of the problem and most of the results of this chapter are from the publications, [38, 39, 40]. Studies built on the papers quoted above include [35, 34]. An algebraic approach to the existence of finite-dimensional filter system was explored by S.I. Marcus and A. Willsky, [27].

References with a framework of finite-dimensional filters include [25].

Books and papers on estimation and filtering. Estimation problems are usually described under Bayesian estimation, see the books [9, 10, 21, 23, 22, 24, 36, 43, 48]. A survey of estimation techniques is the report, [37].

The approach to estimation in which one infimizes a loss function, is due to A. Wald, [42] and is also described in [11]. Bayesian inference in econometrics is presented in [48].

A signal processing approach to estimation is provided by S.M. Kay in [19]. A reference on the relation between digital signal processing and estimation is by A.S. Willsky, [44].

Statistical decision theory was developed since the 1940's and is well described in the books, [30, 32].

Exponential probability densities were studied in depth by
Barndorff-Nielsen, see his books [6, 5].

Filter systems for specific stochastic systems. Filter of a Bernoulli system. A tutorial on filtering and prediction of stochastic systems with a Bernoulli conditional output process is [33].

Filter of a Poisson-Gamma system. The estimation problem and the sequential estimation problem were treated in [16]. The extension to filtering of a Poisson-Gamma system is due to J.A. Bather, [7]. It was rediscovered without this knowledge in [38, 39]. The formulations used in these references differ slightly.

Filter of a stochastic system with an output in a bounded interval of the real numbers. A filter problem for a stochastic system with a beta-distributed output was formulated in [14]; for an estimation problem see also the book
[8, Subsubsection 5.5.5].

Filter of an output-finite-state-finite stochastic system. See the tutorial paper of L. Rabiner, [31]. The books [20, Sec. 3.5] and [13, Ch. 3] treat the filter problem for this stochastic system.

An approximation of a probability distribution by a finite sum of Gaussian probability distribution is treated in [1].

Books on filter problems for various domains. Navigation and tracking, [3, 4, 18]. Multi-target tracking using approximate filters, [26].

Filter problems on geometric spaces require a background in algebraic geometry, differential geometry, and in algebra. Little of this approach has been developed though the theory seems doable. Algebraic conditions will have to be formulated and solved for the existence of finite-dimensional filters. These problems are best treated for continuous-time stochastic systems. References on the filter problem for a system on a circle or on a sphere are [12, 45].

References

1. D.L. Alspach and H.W. Sorenson. Nonlinear bayesian estimation using gaussian sum approximations. *IEEE Trans. Automatic Control*, 17:439–448, 1972. 353
2. P. Diaconis anad D. Ylvisaker. Conjugate priors for exponential families. *Ann. Statist.*, 7:269–281, 1979. 352, 742
3. Y. Bar-Shalom and X.R. Li. *Multitarget–multisensor tracking: Principles and techniques*. YBS, Urbana, IL, USA, 1995. 353
4. Y. Bar-Shalom, X.R. Li, and T. Kirubarajan. *Estimation with applications to tracking and navigation*. John Wiley & Sons Inc., New York, 2001. 310, 353

5. O.E. Barndorf-Nielsen. *Parametric statistical methods and likelihood*. Springer, New York, 1988. 353
6. O.E. Barndorff-Nielsen. *Information and exponential families in statistical theory*. Wiley, London, 1978. 353, 742
7. J.A. Bather. Invariant conditional distributions. *Ann. Math. Statist.*, 36:829–846, 1965. 316, 352, 353
8. A. Bensoussan. *Estimation and control of dynamical systems*. Springer, Berlin, 2018. 353
9. J.O. Berger. *Statistical decision theory and Bayesian analysis*. Springer-Verlag, Berlin, 1985. 352
10. J.M. Bernardo and A.F.M. Smith. *Bayesian theory*. John Wiley, Chichester, 1994. 352
11. D. Blackwell and M.A. Girshick. *Theory of games and statistical decisions*. Wiley, New York, 1954. 353, 410, 467
12. R.W. Brockett and A.S. Willsky. Finite group homomorphic sequential systems. *IEEE Trans. Automatic Control*, 17:483–490, 1972. 353, 786
13. O. Cappé, E. Moulines, and T. Rydén. *Inference in hidden Markov models*. Springer, Berlin, 2005. 169, 277, 353
14. C.Q. da Silva, H.S. Mignon, and L.T. Correia. Dynamic Bayesian beta models. *Computational Statistics and Data Analysis*, 55:0–0, 2011. 353
15. M.H. DeGroot. *Optimal statistical decisions*. McGrawHill, New York, 1970. 352
16. P.A. Frost. Examples of linear solutions to nonlinear estimation problems. In *Proceedings 5th Annual Princeton Conference on Information Sciences*, pages 20–24, 1971. 352, 353
17. I.V. Girsanov. On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. *Th. Probab. Appl.*, 5:285–301, 1960. 352, 604, 742
18. A.H. Jazwinski. *Stochastic processes and filtering theory*. Academic Press, New York, 1970. 310, 353
19. S.M. Kay. *Fundamentals of statistical signal processing - Volume 1, Estimation theory*. Prentice Hall, 1993. 353
20. V. Krishnamurthy. *Partially observed Markov decision processes*. Cambridge University Press, Cambridge, 2016. 169, 277, 353, 575
21. R. Kulhavy. Recursive nonlinear estimation: A geometric approach. *Automatica*, 26:545–555, 1990. 352
22. R. Kulhavy. On design of approximate finite-dimensional estimators: The bayesian view. In X, editor, *IFAC Workshop on Mutual Impact of Computing Power and Control Theory (September 1-2, 1993 Prague, Czechoslovakia)*, pages x–y, X, 1992. X. 352
23. R. Kulhavy. Recursive nonlinear estimation: Geometry of a space of posterior densities. *Automatica*, 28:313–323, 1992. 352
24. R. Kulhavy. Can approximate Bayesian estimation be consistent with the ideal solution. In X, editor, *Proceedings 12th IFAC World Congress*, pages x–y, London, 1993. Pergamon Press. 352
25. J. Levine and G. Pignie. Exact finite dimensional filters for a class of nonlinear discrete time systems. *Stochastics*, 18:97–132, 1986. 352
26. R. Mahler. *Statistical multisource multitarget information fusion*. Artech House, Norwood, MA, USA, 2007. 353
27. S.I. Marcus and A.S. Willsky. Algebraic structure and finite dimensional nonlinear estimation. *SIAM J. Math. Anal.*, 9:312–327, 1978. 352
28. P.A. Meyer. *Probability and Potentials*. Blaisdell Publishing Company, Waltham, MA, 1966. 49, 336, 741, 758
29. P.A. Meyer. *Processus de Markov*, volume 26 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1967. 49, 73, 341
30. J.W. Pratt, H. Raiffa, and R. Schlaifer. *Introduction to statistical decision theory*. MIT Press, Cambridge, MA, 2008. 319, 353, 410
31. L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77:257–286, 1989. 169, 353
32. H. Raiffa and R. Schlaifer. *Applied statistical decision theory*. Harvard Business School, Boston, MA, USA, 1961. 318, 319, 352, 353

33. Branko Ristic, Ba-Tuong Vo, Ba-Ngu Vo, and A. Farina. A tutorial on Bernoulli filters: Theory, implementation, and applications. *IEEE Trans. Sign. Proc.*, 61:3406–3410, 2013. 353
34. W.J. Runggaldier and F. Spizzichino. Finite dimensionality in discrete time nonlinear filtering from a bayesian statistics viewpoint. In A. Germani, editor, *Stochastic modeling and filtering*, volume 91 of *Lecture Notes in Control and Information Sciences*, pages 161–184. Springer-Verlag, Berlin, 1987. 352
35. G. Sawitzki. Exact filtering in exponential families: Discrete time. Report X, Institut für Mathematik, Ruhr-Universität, Bochum, 1979. 352
36. J.C. Spall, editor. *Bayesian analysis of time series and dynamic models*. Marcel Dekker, Inc., New York, 1988. 352
37. Pawel Stano, Zsófia Lendek, Jelmer Braaksma, Robert Bobuška, Cees de Keizer, and Arnold J. den Dekker. Parametric Bayesian filters for nonlinear stochastic dynamical systems: A survey. Report, Delft Center for Systems and Control, Delft University of Technology, Delft, 2012. 352
38. J.H. van Schuppen. A study of estimation and filtering by the bayesian method. SSM-Report 7607, Systems Science and Mathematics Department, Washington University, St. Louis, 1976. 352, 353
39. J.H. van Schuppen. Representations and filtering problems for discrete time stochastic processes. In *Proc. 1977 Joint Automatic Control Conference (ACC.1977)*, pages 1044–1048, 1977. 352, 353
40. J.H. van Schuppen. Stochastic filtering theory: A discussion of concepts, methods and results. In W. Vogel M. Kohlmann, editor, *Stochastic control theory and stochastic differential systems*, volume 16 of *Lecture Notes in Control and Information Sciences*, pages 209–226, Berlin, 1979. Springer-Verlag. 120, 352
41. J.H. van Schuppen and E. Wong. Transformations of local martingales under a change of law. *Ann. Probab.*, 2:879–888, 1974. 352, 742
42. A. Wald. *Statistical decision functions*. John Wiley, New York, 1950. 315, 353, 410
43. M. West. A Bayesian approach to non-linear filtering. Report, U. of Nothingham, Nothingham, 1980? 352
44. Alan S. Willsky. Relationship between digital signal processing and control and estimation theory. *Proceedings of the IEEE*, 66:996–1017, 1978. 353
45. Alan S. Willsky and James Ting-Ho Lo. Estimation for rotational processes with one degree of freedom - Part II: Discrete-time processes. *IEEE Trans. Automatic Control*, 20:22–30, 1975. 353
46. E. Wong. *Stochastic processes in information and dynamical systems*. McGraw-Hill Book Co., New York, 1971. 72, 73, 310, 352
47. E. Wong and B. Hajek. *Stochastic processes in engineering systems*. Springer-Verlag, Berlin, 1985. 73, 310, 352
48. A. Zellner. *An introduction to Bayesian inference in econometrics*. Wiley, New York, 1971. 352, 353

Chapter 10
Stochastic Control Systems

Abstract The concept of a stochastic control system is defined as a map from a tuple of the current state and the current input to the conditional probability distribution of the tuple of the next state and the current output. A Gaussian stochastic control system representation is defined which represents such a stochastic system. Stochastic controllability is defined as the property that there exists an input trajectory such that the probability distribution of a future state equals a prespecified particular probability distribution. A characterization of stochastic controllability of a Gaussian stochastic control system is provided.

Key words: Stochastic control system. Stochastic controllability. A stochastic control system was used in Example 4.1.1 for control of a paper machine and in Example 12.1.1 for control of steering of a ship.

10.1 Stochastic Control System

There follows a gentle introduction to the definition of a stochastic control system which is followed by a formal definition.

A *stochastic control system*, in a preliminary formulation, with a state process x, an input process u, and an output process y, is herewith defined by (1) the probability measure of the initial state and by (2) the probabilistic transition function of the system,

$$(x(t), u(t)) \mapsto \text{cpdf}\left(\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix} | F_t^x \vee F_{t-1}^y \vee F_t^u \right), \ \forall \ t \in T,$$

where cpdf denotes the conditional probability distribution of the indicated random variables conditioned on the specified σ-algebras. The probabilistic transition function at any time $t \in T$, maps the current state and the current input values $(x(t), u(t))$ to the conditional probability distribution function of the next state and the current output $(x(t+1), y(t))$ conditioned on the past states, outputs, and inputs. This definition applies to any stochastic control system regardless of its probability distributions.

It will assumed in this section that the input process is a sequence of independent random variables. Moreover, the input process will be assumed independent of the initial state and of the noise process. This assumption is particular to this chapter, it does not hold in the chapters on control synthesis which follow this chapter. Needed in this chapter is the open loop formulation of a stochastic control system in which the input at any time does not depend on the past of the state and of the output. The assumption is needed to relate various representations of stochastic control systems as will be pointed out at various places.

The above definition is equivalent with the statement that,

$$(F^{x(t+1)} \vee F^{y(t)}, F_t^x \vee F_{t-1}^y \vee F_t^u | F^{x(t),u(t)}) \in \text{CI}, \; \forall \, t \in T.$$

The definition of a stochastic control system is informally illustrated for a time-invariant Gaussian stochastic control system defined by the representation,

$$x(t+1) = Ax(t) + Bu(t) + Mv(t), \; x(0) = x_0, \tag{10.1}$$

$$y(t) = Cx(t) + Du(t) + Nv(t), \tag{10.2}$$

where $n_x, \; n_y, \; n_u, \; n_v \in \mathbb{Z}_+$, $x_0 : \Omega \to \mathbb{R}^{n_x}$, $x_0 \in G(m_{x_0}, Q_{x_0})$, $v : \Omega \times T \to \mathbb{R}^{n_v}$ is a standard Gaussian white noise process, such that for all $t \in T$, $F^{v(t)}$ is independent of $F^{x_0} \vee F_{t-1}^u$; $u : \Omega \times T \to \mathbb{R}^{n_u}$ is a sequence of independent random variables; $A \in \mathbb{R}^{n_x \times n_x}$, $B \in \mathbb{R}^{n_x \times n_u}$, $M \in \mathbb{R}^{n_x \times n_v}$, $C \in \mathbb{R}^{n_y \times n_x}$, $D \in \mathbb{R}^{n_y \times n_u}$, $N \in \mathbb{R}^{n_y \times n_v}$, and $x : \Omega \times T \to \mathbb{R}^{n_x}$ and $y : \Omega \times T \to \mathbb{R}^{n_y}$ are stochastic processes defined by the above equations. It may be shown that the stochastic control system described above is equivalent with the object specified by the conditions: (1) the initial state has a Gaussian probability distribution, and (2) the conditional characteristic function,

$$\begin{aligned} & x_0 \in G(m_0, Q_0), \; u_0 \in \mathbb{R}^{n_{u_0}}; \\ & E\left[\exp\left(i \begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix}\right) | F_t^{x-} \vee F_{t-1}^{y-} \vee F_t^{u-}\right] \\ & = \exp\left(i \begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} Ax(t) + Bu(t) \\ Cx(t) + Du(t) \end{pmatrix} - \frac{1}{2} \begin{pmatrix} w_x \\ w_y \end{pmatrix}^T Q_r \begin{pmatrix} w_x \\ w_y \end{pmatrix}\right), \end{aligned}$$

for all $t \in T$, $(w_x, w_y) \in (\mathbb{R}^{n_x} \times \mathbb{R}^{n_y})$, and a matrix $Q_r \in \mathbb{R}_{pds}^{(n_x+n_y)\times(n_x+n_y)}$.

Note that the probabilistic transition function of the above defined time-invariant Gaussian stochastic control system is a conditional Gaussian measure of the form,

$$G\left(\begin{pmatrix} Ax(t) + Bu(t) \\ Cx(t) + Du(t) \end{pmatrix}, Q_r | \; F_t^{x-} \vee F_{t-1}^{y-} \vee F_t^{u-}\right). \tag{10.3}$$

The input therefore affects of the conditional probability transition function only the conditional mean and does not affect the conditional variance. This fact restricts the usefulness of the system. There are other stochastic systems for which the variance depends on the current state of the system. A Gaussian stochastic control system is generally accepted in the research community of control engineering as a realistic model of a engineering phenomenon with a control input.

The following definition is therefore motivated.

Definition 10.1.1. A discrete-time *stochastic control system* is a collection of spaces and processes defined by the conditions that,

$$\sigma = \{\Omega, F, P, T, Y, B_Y, X, B_X, U, B_U, y, x, u\} \in \text{StocConS},$$

where (Ω, F, P) is a complete probability space; $T = \mathbb{N} = \{0, 1, \ldots\}$, is called the *time index set;* (Y, B_Y) is a measurable space is called the *output space;* (X, B_X) is a measurable space is called the *state space;* (U, B_U) is a measurable space is called the *input space;* $y : \Omega \times T \to Y$ is a stochastic process, is called the *output process;* $x : \Omega \times T \to X$ is a stochastic process, is called the *state process*; and $u : \Omega \times T \to U$ is a stochastic process, is called the *input process;* such that, the probability distribution of the initial state x_0 is specified, the input u is a sequence of independent random variables which are also independent of x_0, the noise process v is a sequence of independent random variables such that for any time $t \in T$, the σ-algebra $F^{v(t)}$ is independent of the σ-algebra $F^{x_0} \vee F^{v}_{t-1} \vee F^{u}_{t}$; and the following conditional independence property holds in terms of the conditional probability distribution,

$$\begin{aligned}
&\text{cpdf}((x(t+1), y(t))|F^{x-}_{t} \vee F^{y-}_{t-1} \vee F^{u}_{t}) \\
&= \text{cpdf}((x(t+1), y(t))|F^{x(t)} \vee F^{u(t)}); && (10.4) \\
&\Leftrightarrow (F^{x(t+1)} \vee F^{y(t)}, F^{x-}_{t} \vee F^{y-}_{t-1} \vee F^{u}_{t}|F^{x(t),u(t)}) \in \text{CI}, \ \ \forall\, t \in T. && (10.5)
\end{aligned}$$

Note that no restriction is imposed on the measurability of the state process with respect to a filtration. There exists a set of state processes for a fixed tuple of input-output processes.

The class of stochastic control systems is denoted by StocConS. A stochastic control system is called *time invariant* if the map (10.4) does not depend on time explicitly.

It is called a *Gaussian stochastic control system* if $Y = \mathbb{R}^{n_y}$, $X = \mathbb{R}^{n_x}$, $U = \mathbb{R}^{n_u}$, if the probability distribution of the initial state is Gaussian, and if the conditional distribution (10.4) is conditionally Gaussian.

It is called a *finite stochastic control system* if X, Y are finite sets. The input set U need not be finite though in many cases it will be so.

In case there is no output process the following representation is often used.

Definition 10.1.2. A *recursive (state-observed) stochastic control system representation* is a collection in which the state and the input processes are related by the equation,

$$\begin{aligned}
&\{\Omega, F, P, T, X, B_X, U, B_U, x, u\}, \\
x(t+1) &= f(t, x(t), u(t), v(t)), x(0) = x_0,
\end{aligned}$$

where $x_0 : \Omega \to X$ is a random variable; $v : \Omega \times T \to V$ is a stochastic process consisting of an independent sequence; $u : \Omega \times T \to U$ is a stochastic process, called the *input process* which is an independent sequence of random variables; x_0 and v are independent objects, and for all $t \in T$, $F^{v(t)}$ is independent of $F^{x_0} \vee F^{v}_{t-1} \vee F^{u}_{t}$;

$f : T \times X \times U \to X$ is a measurable map; and $x : \Omega \times T \to X$ is defined by the above recursion.

Note that for a recursive state-observed stochastic control system, at any $t \in T$, $F_t^{v+} = \sigma(\{v(s), s \geq t\})$ is independent of $F^{x_0} \vee F_{t-1}^{v-}$. Also
$F^{x(t)} \subset (F^{x_0} \vee F_{t-1}^{v-} \vee F_{t-1}^{u})$. Hence there exists a probabilistic transition function,

$$\begin{aligned} &(x(t), u(t)) \mapsto \text{conditional distribution}(x(t+1)|F_t^{x-} \vee F_t^{u}) \\ &= \text{cpdf}(x(t+1))|F^{x(t)} \vee F^{u(t)}) = \text{cpdf}(f(t, x(t), u(t), v(t))|F^{x(t)} \vee F^{u(t)}). \end{aligned}$$

A recursive stochastic control system of Definition 10.1.2 is therefore a stochastic control system as defined in Definition 10.1.1.

The problem of optimal control of a stochastic system treated in the Chapters 12 to 15 is much clarified by the concept of a controlled output. In an optimal stochastic control problem, the cost rate will be a function of the controlled output.

Definition 10.1.3. Consider a recursive stochastic control system representation,

$$x(t+1) = f(t, x(t), u(t), v(t)), x(0) = x_0.$$

Define the *controlled output* of this system by the representation,

$$\begin{aligned} &z : \Omega \times T \to \mathbb{R}^{n_z},\ n_z \in \mathbb{Z}_+,\ T = T(0:t_1) = \{0, 1, 2, \ldots, t_1\}, \\ z(t) &= h(t, x(t), u(t)),\ \forall\, t \in T(0:t_1 - 1), \\ z(t_1) &= h_1(x(t_1)). \end{aligned}$$

The definition of $z(t_1)$ is motivated by the fact that there does not exist a variable $u(t_1)$ on the finite horizon $T(0:t_1)$ while there is a need to refer to $x(t_1)$.

10.2 Gaussian Stochastic Control Systems

As for stochastic systems without input, one defines a Gaussian stochastic control system representation. In control theory it is a frequently used model.

Definition 10.2.1. A *Gaussian stochastic control system representation* is a collection

$$\{\Omega, F, P,\ T,\ \mathbb{R}^{n_y}, B(\mathbb{R}^{n_y}),\ \mathbb{R}^{n_x}, B(\mathbb{R}^{n_x}),\ \mathbb{R}^{n_u}, B(\mathbb{R}^{n_u}),\ y,\ x,\ u\},$$

$$x(t+1) = A(t)x(t) + B(t)u(t) + M(t)v(t),\ x(t_0) = x_0, \tag{10.6}$$

$$y(t) = C(t)x(t) + D(t)u(t) + N(t)v(t); \tag{10.7}$$

where (Ω, F, P) is a complete probability space, $T = \mathbb{N} = \{0,\ 1, \ldots\}$ is a time index set, $X = \mathbb{R}^{n_x}$, $B_X = B(\mathbb{R}^{n_x})$ denotes the Borel σ-algebra of X, $Y = \mathbb{R}^{n_y}$, $B_Y = B(\mathbb{R}^{n_y})$, $U = \mathbb{R}^{n_u}$, $B_U = B(\mathbb{R}^{n_u})$, $x_0 : \Omega \to X$, $x_0 \in G(m_0, Q_{x_0})$, $v : \Omega \times T \to \mathbb{R}^{n_v}$ is a Gaussian white noise process with $v(t) \in G(0, I_{n_v})$, $u : \Omega \times T \to U$ is a stochastic process which is a sequence of independent random variables, with F^{x_0}, F_∞^v, independent; and, for all $t \in T$, $F^{v(t)}$ is independent of $F^{x_0} \vee F_{t-1}^{v} \vee F_t^{u}$; $A : T \to \mathbb{R}^{n_x \times n_x}$, $B :$

$T \to \mathbb{R}^{n_x \times n_u}$, $M : T \to \mathbb{R}^{n_x \times n_v}$, $C : T \to \mathbb{R}^{n_y \times n_x}$, $D : T \to \mathbb{R}^{n_y \times n_u}$, $N : T \to \mathbb{R}^{n_y \times n_v}$; and $x : \Omega \times T \to X$, $y : \Omega \times T \to Y$ are defined by the equations (10.6 & 10.7). A Gaussian stochastic control system representation is a Gaussian stochastic control system as defined in Def. 10.1.1.

A Gaussian stochastic control system representation is called a *time-invariant Gaussian stochastic control system representation* if the functions A, B, C, D, M, N do not depend on the time variable, hence may be represented by matrices denoted by the same symbols. The set of parameters of a time-invariant Gaussian stochastic control system representation is denoted by

$$\{n_y, n_x, n_u, n_v, A, B, M, C, D, N\} \in \text{GStocCSP}.$$

Define the following conditions for a time-invariant Gaussian stochastic control system,

$$n_y \leq n_v, \ \text{rank}(N) = n_y, \ \text{rank} \begin{pmatrix} M \\ N \end{pmatrix} = n_v;$$

$(A,\ M)$ is a supportable pair or a supportable-stable pair,

$(A,\ B)$ is a controllable pair or a stabilizable pair.

These conditions will occasionally be imposed and will then be explicitly stated. The interpretations of these conditions will be stated later.

Proposition 10.2.2. *Consider a Gaussian stochastic control system representation as defined in Def. 10.2.1. Then this system is a Gaussian stochastic control system as defined in Def. 10.1.1*

Proof. Note that by assumption, F^{v_0} is independent of $F^{x_0} \vee F_0^u$. Then $x(1) = A(0)x_0 + B(0)u_0 + M(0)v_0$ is a linear function of the three independent random variables x_0, u_0, v_0. By induction it follows that, for any time $t \in T$, with the expression $x(t+1) = A(t)x(t) + B(t)u(t) + M(t)v(t)$, $F^{v(t)}$ is independent of $F_t^x \vee F_t^u \vee F_{t-1}^v$.

Then a calculation shows that, for a Gaussian stochastic control system,

$$\begin{aligned} &E\left[\exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix}\right) | F_t^{x-} \vee F_{t-1}^{y-} \vee F_t^u\right] \\ &= \exp\left(i\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T \begin{pmatrix} A(t)x(t) + B(t)u(t) \\ C(t)x(t) + D(t)u(t) \end{pmatrix} - \frac{1}{2}\begin{pmatrix} w_x \\ w_y \end{pmatrix}^T Q_r(t) \begin{pmatrix} w_x \\ w_y \end{pmatrix}\right), \end{aligned}$$

$$Q_r(t) = \begin{pmatrix} M(t) \\ N(t) \end{pmatrix} \begin{pmatrix} M(t) \\ N(t) \end{pmatrix}^T \in \mathbb{R}_{pds}^{(n_x+n_y)\times(n_x+n_y)}.$$

This expression determines the probabilistic system transition map. Thus the representation determines a Gaussian stochastic control system. □

Definition 10.2.3. *Controlled output of a Gaussian stochastic control system representation.* In case of a time-varying Gaussian stochastic control system the controlled output has a linear representation formulated as,

$$\begin{aligned}
&z : \Omega \times T \to \mathbb{R}^{n_z},\ n_z \in \mathbb{Z}_+,\ T = T(0:t_1),\\
&C_z : T(0:t_1) \to \mathbb{R}^{n_z \times n_x},\ D_z : T(0:t_1-1) \to \mathbb{R}^{n_z \times n_u};\\
z(t) &= C_z(t)x(t) + D_z(t)u(t),\ \forall\, t \in T(0:t_1-1),\\
z(t_1) &= C_z(t)x(t_1).
\end{aligned}$$

In case of a time-invariant Gaussian stochastic control system the representation is,

$$\begin{aligned}
&z : \Omega \times T \to \mathbb{R}^{n_z},\ n_z \in \mathbb{Z}_+,\ T = T(0:t_1),\ C_z \in \mathbb{R}^{n_z \times n_x},\ D_z \in \mathbb{R}^{n_z \times n_u};\\
z(t) &= C_z x(t) + D_z u(t),\ \forall\, t \in T(0:t_1-1),\\
z(t_1) &= C_z x(t_1).
\end{aligned}$$

In the time-invariant case, call (1) $n_z < n_u$ the *over-actuated case*, (2) $n_z = n_u$ the *exact-actuated case*, and (3) $n_z > n_u$ the *under-actuated case*. The terminology is motivated by the fact that in the exact-actuated case there are as many components of the input as there are of the controlled output. In the under-actuated case, there are strictly fewer components of the input than in the controlled output, etc. These terms are used in control engineering.

Define the following conditions of the system matrices for the under-actuated and the exact-actuated cases,

$$\begin{aligned}
&n_z \geq n_u,\ D_z \in \mathbb{R}^{n_z \times n_u},\ \text{rank}(D_z) = n_u \ \Rightarrow\ 0 \prec D_z^T D_z,\\
&\text{rank} \begin{pmatrix} B \\ D_z \end{pmatrix} = n_u,\\
&(A, C_z) \text{ is an observable pair or a detectable pair},\\
&(A, B) \text{ is a controllable pair or a stabilizable pair}.
\end{aligned}$$

In the overactuated case, $n_z < n_u$, $\text{rank}(D_z) \leq n_z < n_u$ hence $\text{rank}(D_z^T D_z) < n_u$. This is called the singular optimal control case in optimal control theory. That case has been analyzed but requires a treatment with the inverse system that is not further discussed in this book.

The underactuated case, $n_z > n_u$, is regarded as the standard case to be discussed.

Often control engineers take $n_z = n_x$ and then most often $n_z = n_x \geq n_u$ thus one obtains the under-actuated case if $n_z > n_u$ or the exact-actuated case if $n_z = n_u$.

The condition $\text{rank}(D_z^T D_z) = n_u$ is a regularity condition often used for optimal control theory.

The column rank condition on the column matrix of B and D_z is used to eliminate dependence among the representation of the components of the input vector. It is also related to the existence of an inverse system.

In optimal stochastic control theory there is defined a cost rate often in the form of a quadratic cost function of the state and of the input. Below the cost rate will be defined as a quadratic form of the controlled output, $z(t)^T z(t)$. This allows a direct interpretation of the cost function in terms of a stochastic control system from the input u to the controlled output z. Note that,

$$z(t) = C_z x(t) + D_z u(t) \;\Rightarrow\; z(t)^T z(t) = \begin{pmatrix} x(t) \\ u(t) \end{pmatrix}^T \begin{pmatrix} C_z^T C_z & C_z^T D_z \\ D_z^T C_z & D_z^T D_z \end{pmatrix} \begin{pmatrix} x(t) \\ u(t) \end{pmatrix}^T ;$$

$$n_z \geq n_u \text{ and } \mathrm{rank}(D_z) = n_u \;\Rightarrow\; \mathrm{rank}(D_z^T D_z) = n_u .$$

10.3 Stochastic Controllability and Stochastic Co-Controllability

Controllability is a major concept of control and system theory. Controllability is often a necessary and sufficient condition for the existence of a control law. In this section the concept of controllability is defined and characterized both for deterministic systems and for Gaussian stochastic control systems.

10.3.1 Controllability of a Deterministic System

The concept of controllability is first defined in terms of sets and maps.

Definition 10.3.1. Consider a collection of sets and maps of the form,

$$\{U,\, X,\, Y,\, g : U \to X,\, h : X \to Y\}.$$

Consider a *control-objective subset* $X_{co} \subseteq X$. Define the map $g : U \to X$ to be a *controllable map* with respect to the control-objective subset X_{co} if the function $g : U \to X$ is surjective with respect to the control-objective subset $X_{co} \subseteq X$. It is called a *completely controllable map* if it is a controllable map with respect to $X_{co} = X$.

Recall that the map $g : U \to X$ is surjective or onto with respect to $X_{co} \subseteq X$ if for all $x_c \in X_{co}$ there exists a $u \in U$ such that $x_c = g(u)$. See Definition 17.1.11 for a formal definition of a surjective map.

Proposition 10.3.2. *Consider the sets $U = \mathbb{R}^{n_u}$ and $X = \mathbb{R}^{n_x}$, and the linear map $g : U \to X$, $g(u) = Lu$ for a matrix $L \in \mathbb{R}^{n_x \times n_u}$.*

(a)Consider a control-objective subset $X_{co} \subseteq X$. The map g is controllable with respect to the subset X_{co} if and only if X_{co} is contained in the image of g,

$$X_{co} \subseteq \mathrm{Im}(g) = \{x \in X |\; \exists\, u \in U \text{ such that } x = g(u)\}.$$

(b)Consider next a control-objective linear subspace $X_{co} \subseteq X$ of dimension $n_{X_{co}}$. The map g is controllable with respect to X_{co} and $X_{co} = \mathrm{Im}(g)$ if and only if (1) $X_{co} \subseteq \mathrm{Im}(g)$ and (2) $n_{X_{co}} = \mathrm{rank}(L)$.

Proof. (a) This set formulation is a reformulation of the definition.
(b) ($\Leftarrow$) Both X_{co} and $\mathrm{Im}(g)$ are linear subspaces of the linear space X. Condition (1) implies that $n_{X_{co}} = \dim(X_{co}) \leq \dim(\mathrm{Im}(g)) = \mathrm{rank}(L)$. If by (2) $\mathrm{rank}(L) = n_{X_{co}}$ then $\dim(X_{co}) = n_{X_{co}} = \mathrm{rank}(L) = \dim(\mathrm{Im}(g))$ and with (1) follows that $X_{co} = \mathrm{Im}(g)$ and controllability with respect to X_{co} holds.

($\Rightarrow$) If (1) does not hold then there exists a member $x_{co} \in X_{co}$ which is not in the image of g, $x_{co} \not\in \text{Im}(g)$. This contradicts controllabilty with respect to X_{co}. From (1) follows that $n_{X_{co}} = \dim(X_{co}) \leq \dim(\text{Im}(g)) = \text{rank}(L)$. Suppose that $n_{X_{co}} < \text{rank}(L)$. Then $\dim(X_{co}) = n_{X_{co}} < \text{rank}(L) = \dim(\text{Im}(g))$. Hence there exists a member $x_c \in X_{co}$ which is not in the image of g, $x_c \not\in \text{Im}(g)$. This contradicts controllability with respect to X_{co}. Thus $n_{X_{co}} = \text{rank}(L)$. □

Example 10.3.3. Consider the linear space $X = \mathbb{R}^2$. Consider the subspaces

$$\begin{aligned} X_c &= \left\{ \begin{pmatrix} x_1 \\ 0 \end{pmatrix} \mid \forall\, x_1 \in \mathbb{R} \right\} = \text{Im}(g),\ g(u) = Lu,\ L = \begin{pmatrix} 1\ 0 \\ 0\ 0 \end{pmatrix} \in \mathbb{R}^{2\times 2}, \\ X_{co1} &= \left\{ \begin{pmatrix} 0 \\ x_2 \end{pmatrix} \mid \forall\, x_2 \in \mathbb{R} \right\}. \end{aligned}$$

Then (1) $X_{co1} \not\subseteq X_c$ while (2) $\dim(X_{co1}) = 1 = \dim(X_c) = \dim(\text{Im}(g)) = \text{rank}(L)$. This shows that the map $g(u) = Lu$ is not controllable with respect to X_{co1}, condition (1) does not hold, but condition (2) holds.

Consider the new subspace $X_c = \mathbb{R}^2$ and X_{co1}. Then $X_{co1} \subset X_c$ holds hence the map L is controllable with respect to X_{co1}, condition (1) holds, but condition (2) does not hold, because $\dim(X_{co1}) = 1 < 2 = \dim(X_c) = \dim(\text{Im}(g)) = \text{rank}(L)$.

Consider again the first subspace $X_c = \text{Im}(g)$ and the subspace

$$X_{co2} = \left\{ \begin{pmatrix} x_1 \\ 0 \end{pmatrix} \mid \forall\, x_1 \in \mathbb{R} \right\}.$$

Then (1) $X_{co2} = X_c$, (2) $1 = \dim(X_{co2}) = n_{X_{co2}} = \text{rank}(L)$, and the map $g(u) = Lu$ is controllable with respect to the subspace X_{co2}.

Definition 10.3.4. Consider a time-varying linear system representation without output

$$x(t+1) = A(t)x(t) + B(t)u(t),\ x(t_0) = x_0.$$

Consider a linear subspace $X_{co} \subseteq X = \mathbb{R}^{n_x}$ which represents the subspace of control objective states.

(a) Call this linear system *controllable with respect to the control objective subspace X_{co} on the interval T_c* if the *future-input-to-state map* is surjective with respect to X_{co},

$$\begin{aligned} T_c &= \{t_0,\ t_0+1,\ \ldots,\ t_0+t_1-1\} \subseteq T, \\ &\quad \text{I2Smap}(x(t_0),\ u(t_0 : t_0+t_1-1)), \\ &\quad \{x(t_0),\ u(t_0),\ u(t_0+1),\ \ldots,\ u(t_0+t_1-1)\} \mapsto x(t_0+t_1), \\ x(t_0+t_1) &= \Phi(t_0+t_1,t_0)x(t_0) + \sum_{s=t_0}^{t_0+t_1-1} \Phi(t_0+t_1-1,s)\ B(s)u(s); \\ &\quad \forall\, x_c \in X_{co},\ \exists\, u(t_0 : t_0+t_1-1) \text{ such that} \\ x_c &= x(t_0+t_1) = \text{I2Smap}(x(t_0),\ u(t_0 : t_0+t_1-1)). \end{aligned}$$

(b)Consider a time-invariant linear system representation,

$$x(t+1) = Ax(t) + Bu(t),\ x(t_0) = x_0,\ \ A \in \mathbb{R}^{n_x \times n_x},\ B \in \mathbb{R}^{n_x \times n_u}.$$

Call this system *controllable with respect to the subspace X_{co}* if there exists a time $t_0 \in T$ and an integer $t_1 \in \mathbb{Z}_+$ such that the interval T_c satisfies $T_c = \{t_0,\ t_0+1,\ \ldots,\ t_0+t_1-1\} \subseteq T$ and the system is controllable with respect to the control objective space X_{co} on the interval T_c as defined in (a). If this holds then it holds for all initial times t_0 such that $T_c \subseteq T$. Call this system *completely controllable* if it is controllable with respect to the state space $X_{co} = X$. The term *completely controllable* will occasionally be abbreviated to only *controllable* if the context is clear.

Definition 10.3.5. Consider a time-varying linear system. Define the *controllability matrix of an interval* by the formula,

$$\begin{aligned} T_c &= \{t_0,\ t_0+1,\ \ldots,\ t_0+t_1-1\} \subseteq T, \\ &\quad \text{conmat}(A,\ B,\ t_0,\ t_0+t_1-1) = \begin{pmatrix} H_1\ H_2\ \ldots\ H_{t_1-1}\ H_{t_1} \end{pmatrix} \in \mathbb{R}^{n_x \times (t_1 n_u)}, \\ H_1 &= \Phi(t_0+t_1-1, t_0)B(t_0),\ H_2 = \Phi(t_0+t_1-1, t_0+1)B(t_0+1), \\ H_k &= \Phi(t_0+t_1-1, t_0+k)B(t_0+k),\ \forall\, k = 0,\ 1,\ \ldots,\ t_1-1. \end{aligned}$$

Consider a time-invariant linear system. Define the *controllability matrix* of this system as

$$\text{conmat}(A,\ B) = \begin{pmatrix} B\ AB\ \ldots\ A^{n_x-2}B\ A^{n_x-1}B \end{pmatrix} \in \mathbb{R}^{n_x \times (n_x n_u)}.$$

Note the slight abuse of notation for the controllability matrices.

Proposition 10.3.6. *(a)Consider a time-varying linear system. The system is controllable with respect to the control objective space $X_{co} = X$ on the interval $T_c = \{t_0,\ t_0+1,\ \ldots,\ t_0+t_1\} \subseteq T$ if and only if* $\text{rank}(\text{conmat}(A,\ B,\ t_0, t_0+t_1-1)) = n_x$.

(b)Consider a time-invariant linear system. The system is controllable with respect to the control objective space $X_{co} = X$ if and only if $\text{rank}(\text{conmat}(A,\ B)) = n_x$.

Proof. (a) Note the linear map,

$$x_c - \Phi(t_0+t_1, t_0)x(t_0) = \text{conmat}(A,\ B,\ t_0, t_0+t_1-1)u(t_0 : t_0+t_1-1).$$

In the vector $u(t_0 : t_0+t_1-1)$ the input vectors $u(t_0),\ u(t_0+1),\ \ldots,\ u(t_0+t_1-1)$ are arranged from top to bottom in the indicated order. Note that $X_{co} = X$ and controllability imply that $X = X_{co} \subseteq \text{Im}(\text{conmat}(.))$. Next Proposition 10.3.2 is applied. $(\Leftarrow)$

$$\begin{aligned} \dim(X_{co}) &= \dim(X) = n_X = \text{rank}(\text{conmat}(.)) = \dim(\text{Im}(\text{conmat}(.)) \\ &\quad \Rightarrow X = X_{co} = \text{Im}(\text{conmat}(.)). \end{aligned}$$

$(\Rightarrow)$ If the system is controllable with respect to $X_{co} = X$ on T_c then by the proposition $X = X_{co} \subseteq \text{Im}(\text{conmat}(.))$. Then

$$
\begin{aligned}
n_x &= \dim(X) = \dim(X_{co}) \le \dim(\text{Im}(\text{conmat}(.))) = \text{rank}(\text{conmat}(.)) \le n_x,\\
&\Rightarrow n_x = \text{rank}(\text{conmat}(.)).
\end{aligned}
$$

(b) The proof is similar to that of part (a) if one uses the Cayley-Hamilton theorem which states that for A^{n_x} and for higher powers A^m with $m > n_x$,

$$A^{n_x} = \sum_{k=0}^{n_x-1} a_k\, A^k, \;\; A^m = \sum_{k=0}^{n_x-1} c_k\, A^k.$$

□

Example 10.3.7. Consider a time-invariant linear system without output. Assume that the system is not controllable. Then there exists by Proposition 21.2.9 a linear state-space transformation such that the new system representation has the form,

$$
\begin{aligned}
x(t+1) = & \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} + \begin{pmatrix} B_1 \\ 0 \end{pmatrix} u(t),\; x(0) = x_0,\\
& \forall\, i,\; j \in \{1,\, 2\},\; A_{i,j} \in \mathbb{R}^{n_i \times n_j},\; B_1 \in \mathbb{R}^{n_1 \times n_u},\\
& (A_{11},\, B_1) \text{ a controllable pair, } n_1,\; n_2 \in \mathbb{N},\; n_x = n_1 + n_2.
\end{aligned}
$$

Note that (1) the second state component x_2 is not affected by the input at all; and (2) the second state component x_2 is not influenced by the first state component x_1 at all. In fact,

$$x_2(t+1) = A_{22} x_2(t),\; x_2(0) = x_{2,0}.$$

If $\text{spec}(A) \subset \text{D}_o$ then $\lim_{t\to\infty}\, x_2(t) = \lim\, A_{22}^t\, x_{2,0} = 0$. Attention can therefore be restricted to the first state component, hence one obtains the system representation,

$$
\begin{aligned}
x_1(t+1) = & A_{11} x_1(t) + B_1 u(t),\; x_1(0) = x_{1,0},\\
& (A_{11},\, B_1) \text{ a controllable pair.}
\end{aligned}
$$

In conclusion, noncontrollability of a time-invariant linear system leads to the conclusion that the system can be reduced in state-space dimension to a new system representation which is controllable.

10.3.2 The Concept of Stochastic Controllability

Stochastic controllability of a stochastic control system holds if the system achieves a set of prespecified probability measures on the state set of the system by an appropriately chosen input signals. Controllability of a stochastic control system is significantly different from controllability of a deterministic control system. An informal introduction precedes the formal definition.

Consider a stochastic control system. The inputs discussed below can be either deterministic or stochastic. Most of the discussion is restricted to an input of finite length. If an infinite length input is needed this will be stated. The relevant time interval of stochastic controllability starts with the initial time t_0 and ends at the terminal time t_0+t_1 for an integer $t_1 \in \mathbb{Z}_+$.

A finite input supplied to the stochastic control system determines a probability measure on the set X at the terminal time t_0+t_1. That measure is regarded as the conditional probability measure of $x(t_0+t_1)$ conditioned on the initial state $x(t_0)$ and on the input signal used on the interval. Mathematical notation follows.

$$\begin{aligned}
& \exists\, t_0 \in T,\ \exists\, t_1 \in \mathbb{Z}_+, \text{ such that} \\
T_c = {} & \{t_0,\ t_0+1,\ \ldots, t_0+t_1-2,\ t_0+t_1-1\} \subseteq T; \\
& u(t_0:t_0+t_1-1) \in U^{t_1}, \\
& u(t_0:t_0+t_1-1) \mapsto \text{cpm}\left(x(t_1)|\ F^{x(t_0)} \vee F^{u(t_0:t_0+t_1-1)}\right), \qquad (10.8) \\
& P_c(t_0+t_1,\ X,\ B(X)) \\
& = \left\{ \begin{array}{l} \text{cpm}\left(x(t_0+t_1)|\ F^{x(t_0)} \vee F^{u(t_0:t_0+t_1-1)}\right) \in P(X,B(X))\ | \\ \forall\ u(t_0:t_0+t_1-1) \in U^{t_1} \end{array} \right\}. \qquad (10.9)
\end{aligned}$$

Here the map of Eq. (10.8) is called the *stochastic (future-) input-to-state-measure map of the interval*. The set P_c of Eqn. (10.9) is called the *controllable set of state probability measures at the terminal time* or, for short, the *controllable set*. The symbol cpm denotes a *conditional probability measure*. The understanding of the term *state-measure map* is that it refers to the probability measure of the future state $x(t_0+t_1)$. The reader should note that there is a difference between controllability of a deterministic system and that of a stochastic system in that for the first system one refers to a state while for the second system one refers to a probability measure of the state. The term *future-input map* is to distinguish it from the past-input map which is used to define stochastic co-controllability.

If the initial state $x(t_0) \in X$ is fixed and if the input $u(t_0:t_0+t_1-1)$ on the interval is fixed, then the stochastic input-to-state-measure map of the interval determines a probability measure on the set $(X,B(X))$ for the terminal state $x(t_0+t_1)$. Control of a stochastic system on an interval thus produces, apriori, a probability measure for the state at the terminal time of the interval.

Recall the concept of support of a probability measure, see Def. 2.5.7. Define the support of all probability measures of the set P_c, with the understanding that $x(t_0)$ is fixed, as the set,

$$X_{c,support} = \cup \text{ support}\left(\text{ cpm}\left(x(t_1)|\ F^{x(t_0)} \vee F^{u(t_0:t_0+t_1-1)}\right)\right),$$

where the union is over all conditional probability measures of the set P_c and hence over all possible finite input sequences of the interval. Then $X_{c,support} \subseteq X$. The minimal state set of the interval is actually the subset $X_{c,support}$. If X is strictly larger than $X_{c,support}$ then the state set is unnecessary large for this interval.

Note that the controllable set of state probability measures at the terminal time is in general strictly smaller than the set of all probability measures on the set X. The stochastic control system imposes restrictions. For example, for a Gaussian stochastic control system it will be argued below that the controllable set of state probability measures are Gaussian probability measures of which the mean value can be affected by the input but the variance cannot at all be affected by the input. For example, for a finite-valued stochastic control system it will be argued below in this chapter that the controllable set of state probability measures are not all measures on the set of probability vectors but mostly a strictly smaller polytope of probability measures.

Controllability of a deterministic system refers to a control objective subset of the state set that should be contained in the controllable set. For a stochastic system one has therefore to specify a subset of probability measures on the state set $(X,B(X))$ which an input should attain. Introduce therefore the concept of a *set of control objective probability measures* denoted by $P_{co}(X,B(X)) \subseteq P(X,B(X))$.

Stochastic controllability of a stochastic control system on an interval will be defined such that the following inclusion relations holds,

$$P_{co}(X,B(X)) \subseteq P_c(t_0+t_1,X,B(X)).$$

Thus, every probability measure on the terminal state of the interval $x(t_0+t_1)$ which belongs to the subset P_{co} should be attainable by a particular input on the interval.

A consequence of the above formulation is that the set of control objective probability measures has to respect the restrictions of the set P_c. In general it is not clear whether for a particular P_{co} one can determine a set of inputs which achieves equality in $P_{co} = P_c$.

Definition 10.3.8. Consider a stochastic control system,

$$\begin{aligned}&\{\Omega,F,P,T,\mathbb{R}^{n_y},B(\mathbb{R}^{n_y}),\mathbb{R}^{n_x},B(\mathbb{R}^{n_x}),\mathbb{R}^{n_u},B(\mathbb{R}^{n_u}),y,x,u\} \in \text{StocConS},\\ &(X,B(X)) = (\mathbb{R}^{n_x},B(\mathbb{R}^{n_x})),\ P(X,B(X)) = P(\mathbb{R}^{n_x},B(\mathbb{R}^{n_x}));\end{aligned}$$

where the latter set denotes the set of probability measures on the space X.

(a) The stochastic control system is called *stochastically controllable* on the interval T_c with respect to the set of control objective probability measures $P_{co}(X,B(X))$ if the following inclusion for this set of measures holds,

$$\begin{aligned}P_{co}(X,B(X)) &\subseteq P_c(t_0+t_1,\ X,\ B(X)), \text{ where}\\ T_c &= \{t_0,\ t_0+1,\ \ldots,\ t_0+t_1-1\} \subseteq T.\end{aligned}$$

In words, every probability measure of the control objective set belongs to the controllable set of probability measures and hence is attainable at time $t_0+t_1 \in T$ with an input process u on the considered time interval T_c.

(b) This stochastic system is called *stochastically co-controllable* on the interval $T_{cb} = \{t_0-1,\ t_0-2,\ldots,t_0-t_1\} \subset T$ with respect to the set of control objective probability measures $P_{co}(X,B(X))$ if the following inclusion for this set of measures holds,

$$\begin{aligned}P_{co}(X,B(X)) &\subseteq P_c(t_0-t_1,\ X,\ B(X)), \text{ where}\\ T_{cb} &= \{t_0,\ t_0-1,\ \ldots,\ t_0-t_1+1\} \subseteq T.\end{aligned}$$

(c) Assume that the stochastic control system is time-invariant. Then this stochastic system is called *stochastically controllable* with respect to the control objective set of probability measures $P_{co}(X, B(X))$ if there exists an initial time $t_0 \in T$ and an integer $t_1 \in \mathbb{Z}_+$ such that $T_c = \{t_0, t_0+1, \ldots, t_1-1\} \subseteq T$ and such that the stochastic control system is stochastically controllable on the interval T_c with respect to $P_{co}(X, B(X))$ as defined in (a). By stationarity this then holds for all $t_0 \in T$.

(d)Assume that the stochastic control system is time-invariant. Then this stochastic system is called *stochastically co-controllable* with respect to the control objective set of probability measures $P_{co}(X, B(X))$ if there exists an initial time $t_0 \in T$ and an integer $t_1 \in \mathbb{Z}_+$ such that $T_{cb} = \{t_0, t_0-1, \ldots, t_0-t_1+1\} \subseteq T$ and such that the stochastic control system is stochastically co-controllable on the interval T_{cb} with respect to $P_{co}(X, B(X))$ as defined in (b). By stationarity this then holds for all $t_0 \in T$.

The term *controllability of a stochastic control system* would have been more descriptive but the term of stochastic controllability is now commonly used in the literature.

10.3.3 Stochastic Controllability of a Gaussian Control System

The stochastic controllability of a time-invariant Gaussian stochastic control system is characterized below. Consider a time-invariant Gaussian stochastic control system representation,

$$\begin{aligned} x(t+1) &= Ax(t)+Bu(t)+Mv(t), x(t_0) = x_0, \\ y(t) &= Cx(t)+Du(t)+Nv(t). \end{aligned}$$

Then, for $t_0 \in T$ and $t_1 \in \mathbb{Z}_+$ such that $T_c = \{t_0,\ t_0+1,\ \ldots, t_0+t_1-1\} \subseteq T$ and due to the assumed independence of x_0, v, and u,

$$\begin{aligned} x(t_0+t_1) &= A^{t_1}x(t_0) + \sum_{s=t_0}^{t_0+t_1-1} A^{t_0+t_1-s-1}Bu(s) + \\ &\quad + \sum_{s=t_0}^{t_0+t_1-1} A^{t_0+t_1-s-1}Mv(s); \qquad (10.10) \end{aligned}$$

$$\begin{aligned} &E[\exp(iw^T x(t_0+t_1))|F^{x(t_0)} \vee F^u_{t_0+t_1-1}] \qquad (10.11) \\ &= \exp\begin{pmatrix} iw^T[A^{t_1}x(t_0) + \sum_{s=t_0}^{t_0+t_1-1} A^{t_0+t_1-1-s}Bu(s)] + \\ -\frac{1}{2}w^T Q_c(t_0+t_1)w \end{pmatrix}, \end{aligned}$$

$$Q_c(t_0+t_1) = \sum_{s=t_0}^{t_0+t_1-1} A^{t_0+t_1-1-s}MM^T(A^T)^{t_0+t_1-1-s}. \qquad (10.12)$$

Consider for fixed initial state $x(t_0) = x_0$, the stochastic input-to-state-measure map,

$$\begin{aligned}
&(u(t_0),\ldots,u(t_0+t_1-1)) \mapsto E[\exp(iw^T x(t_0+t_1))|F^{x(t_0)} \vee F^u_{t_0+t_1-1}] \\
&= \exp\left(iw^T [A^{t_1} x_0 + \sum_{s=t_0}^{t_0+t_1-1} A^{t_0+t_1-1-s} Bu(s)] - \frac{1}{2} w^T Q_c(t_0+t_1) w \right). \qquad (10.13)
\end{aligned}$$

The expression (10.13) shows that by choosing $(u(t_0),\ldots,u(t_0+t_1-1))$ one can influence the mean of the conditional distribution $\text{cpdf}(x(t_0+t_1)|F^{x(t_0)} \vee F^u_{t_0+t_1-1})$ but one cannot influence the variance of the conditional distribution by the input signal. This shows that in the definition of stochastic controllability one has to restrict the set of measures on the state set for the terminal state $x(t_0+t_1)$ as the controllable set of probability measures does. Define the set of controllable probability measures and a set of control objective measures of this Gaussian stochastic control system respectively by the formulas,

$$\begin{aligned}
T_c &= \{t_0, t_0+1, \ldots, t_0+t_1-1\} \subset T, \\
&P_c(t_0+t_1,\, X,\, B(X)) \qquad (10.14) \\
&= \left\{ \begin{array}{l} G(m_c(t_0+t_1),\, Q_c(t_0+t_1)) \in P(X, B(X))| \\ m_c(t_0+t_1) \in \mathbb{R}^{n_x},\, Q_c(t_0+t_1) \in \mathbb{R}^{n_x \times n_x}_{pds} \\ \text{as specified below} \end{array} \right\}, \\
m_c(t_0+t_1) &= A^{t_1} x(t_0) + \sum_{s=t_0}^{t_0+t_1-1} A^{t_0+t_1-1-s} Bu(s), \\
Q_c(t_0+t_1) &= \sum_{s=t_0}^{t_0+t_1-1} A^{t_0+t_1-1-s} MM^T (A^T)^{t_0+t_1-1-s}, \\
P_{co}(X, B(X)) &= \{G(m_{co}, Q_c(t_0+t_1)) \in P(X, B(X))|\ \forall\, m_{co} \in \mathbb{R}^{n_x}\}. \qquad (10.15)
\end{aligned}$$

Recall from Theorem 21.2.7 that the deterministic time-invariant linear system with representation,

$$x(t+1) = Ax(t) + Bu(t),\ x(0) = x_0,$$

is controllable if and only if (A, B) is a controllable pair which is defined by the following rank condition, $n_x = \text{rank}(\text{conmat}(A, B)) = \text{rank}\left(B\ AB\ A^2B\ \ldots\ A^{n_x-1}B \right)$.

Proposition 10.3.9. *Consider the time-invariant Gaussian stochastic control system representation,*

$$\begin{aligned}
&\{n_y, n_x, n_u, n_v, A, B, C, D, M, N\} \in \text{GStocCSP}, \\
x(t+1) &= Ax(t) + Bu(t) + Mv(t),\ x(t_0) = x_0, \\
y(t) &= Cx(t) + Du(t) + Nv(t),\ v(t) \in G(0, I).
\end{aligned}$$

Assume that $\text{spec}(A) \subset \text{D}_o$ *and that* (A, M) *is a supportable pair.*

This time-invariant Gaussian stochastic control system is stochastically controllable with respect to the initial state $x(t_0) \in \mathbb{R}^{n_x}$ *on the interval* T_c *with respect to the control objective measures* $P_{co}(X, B_X)$ *as specified in equation (10.15) if and only if* $(A,\ B)$ *is a controllable pair.*

By symmetry, the time-invariant backward Gaussian stochastic control system representation is stochastically co-controlable with respect to probability measures in the subset $P_{co}(X,B_X)$ if and only if $(A_b,\ B_b)$ is a controllable pair.

The corresponding characterization of stochastic controllability of a time-varying Gaussian stochastic control system, is analogous.

Proof. Let $t_0 \in T$ and $t_1 \in \mathbb{Z}_+$ such that $T_c = \{t_0,\ t_0+1,\ \ldots, t_0+t_1-1\} \subseteq T$. Let $x_0 \in X$ and let $P_c(t_0+t_1, X, B(X))$ be as defined in (10.14). Let $P_a \in P_{co}(X,B(X))$, say $P_a = G(m_{co,a}, Q_c(t_0+t_1))$ with $Q_c(t_0+t_1)$ given by (10.15). Consider for fixed $x_0 \in X$, the map,

$$\begin{aligned}
&\{u(0),\ldots,u(t_1-1)\} \mapsto E[\exp(iw^T x(t_0+t_1))|F^{x(t_0)} \vee F^u_{t_0+t_1-1}] \\
&= \exp\left(iw^T[A^{t_1}x(t_0) + \sum_{s=t_0}^{t_0+t_1-1} A^{t_0+t_1-1-s}Bu(s)] - \frac{1}{2}w^T Q_c(t_0+t_1)w \right), \\
&\quad \forall\ w \in \mathbb{R}^{n_x},
\end{aligned}$$

where the expression for the conditional characteristic function follows from (10.13). There exists an input trajectory,

$$\begin{aligned}
&u(t_0 : t_0+t_1-1) = \{u(t_0),\ldots,u(t_0+t_1-1)\} \text{ such that,} \\
&\text{cpm}(x(t_0+t_1)|F^{x(t_0)} \vee F^u_{t_0+t_1-1}) \\
= \ & G(m_c(t_0+t_1), Q_1(t_0+t_1)|\ F^{x(t_0)} \vee F^u_{t_0+t_1-1}) = G(m_{co,a}, Q_c(t_0+t_1)), \\
\Leftrightarrow \ & \text{there exists an input trajectory } u(t_0 : t_0+t_1-1) \text{ such that} \\
&m_{co,a} = A^{t_1}x(t_0) + \sum_{s=t_0}^{t_0+t_1-1} A^{t_0+t_1-1-s}Bu(s), \text{by the expression (10.13),} \\
\Leftrightarrow \ & m_{co,a} - A^{t_1}x_0 = \sum_{s=t_0}^{t_0+t_1-1} A^{t_0+t_1-1-s}Bu(s), \\
\Leftrightarrow \ & \text{rank}\left(B\ AB\ A^2B\ \ldots\ A^{n_x-1}B \right) = n_x,
\end{aligned}$$

if and only if (A,B) is a controllable pair by Theorem 21.2.7. Thus, if (A,B) is a controllable pair then the stochastic control system is stochastically controllable on the interval T_c with respect to the set P_{co} hence the time-invariant Gaussian stochastic control system is stochastically controllable. □

10.4 State-Finite Stochastic Control Systems

10.4.1 Definition

A finite stochastic systems was defined in Def. 10.1.1. The reader finds in this section the definition of a finite stochastic control system representation in which both the state and the input are finite valued.

Definition 10.4.1. Define an *output-finite-state-finite stochastic control system representation* by the sets and relations,

$$\begin{aligned}
&\{\Omega,F,P,T,Y,X,U,y,x,u\},\\
&P(\{x(t+1)=i_x,y(t)=j_y\}|\,F_t^x\vee F_{t-1}^y\vee F_t^u)=f(t,i_x,j_y,x(t),u(t)),\\
&X_e=\{e_1,\,e_2,\,\ldots,\,e_{n_x}\in\mathbb{R}_+^{n_x}\},\ Y_e=\{e_1,\,e_2,\,\ldots,\,e_{n_y}\in\mathbb{R}_+^{n_y}\},\ U\subseteq\mathbb{R}^{n_u},\\
&n_x,\ n_y,\ n_u\in\mathbb{Z}_+,\ T=\mathbb{N}\text{ or }T=T(0:t_1),\ t_1\in\mathbb{Z}_+,\\
&f:T\times X_e\times Y_e\times X_e\times U\to[0,1];\\
&E[x(t+1)y(t)^T\,|\,F_t^x\vee F_{t-1}^y\vee F_t^u]=E[x(t+1)y(t)^T\,|\,F^{x(t),u(t)}],\ \forall\,t\in T.
\end{aligned}$$

Here X, Y, U are called respectively the *state set*, the *output set*, and the *input set*, x and y are finite-valued processes assumed to be in the indicator representation, and $u:\Omega\times T\to U$ is a stochastic process, This collection is a stochastic control system as defined in Definition 10.1.1 by definition of the above map.

Define a *state-finite stochastic control system representation* as a special case of the above representation in which only the state and the input processes are present and the output process is not present.

The term *output-finite-state-finite stochastic control system* may be explained as follows. If $u(t)=u^*$ is kept fixed for all time then x is a Markov process hence taking values in a finite set and often called a state-finite Markov process. The expression *controlled Markov chain* is used in the literature. Because the state process x is not observed but only the output process is observed and because the output process is also finite valued, one speaks of a *partially-observed Markov process*. In the literature it is not always clear whether the output process takes values in a finite set or in a noncountable subset of the real numbers. There are examples where the input set is not countable.

The main restriction of the above defined stochastic control system is that the state and the output are finite-valued processes. The input set may be finite but, in a generalization, could be a countable set or a subset of the real numbers. In this book the expression of an output-finit-state-finite stochastic control system representation will be used to keep the analogy with a Gaussian stochastic control system representation and other stochastic control systems.

Definition 10.4.2. Define a *time-invariant output-finite-state-finite stochastic control system representation*. by the sets and maps, using the indicator representation of finite-valued processes, see Section 3.5, and assuming that state-output conditional independence holds,

$$\begin{aligned}
&E\left[\begin{pmatrix}x(t+1)\\ y(t)\end{pmatrix}|\,F_t^{x-}\vee F_t^{u-}\right]=\begin{pmatrix}A(u(t))\\ C(u(t))\end{pmatrix}x(t),\\
&\text{such that, }\forall\,t\in T,\ \ (x(t),u(t))\mapsto\text{cpdf}(x(t+1),y(t))|F_t^{x-}\vee F_t^u);\\
&X_e=\{e_1,\,e_2,\,\ldots,\,e_{n_x}\in\mathbb{R}_{st}^{n_x}\},\ Y_e=\{e_1,\,e_2,\,\ldots,\,e_{n_y}\in\mathbb{R}_{st}^{n_y}\},\ U\subseteq\mathbb{R}^{n_u},\\
&n_x,\ n_y,n_u\in\mathbb{Z}_+,\ T=\mathbb{N}\text{ or }T=T(0:t_1),\ t_1\in\mathbb{Z}_+,\\
&x:\Omega\times T\to X_e,\ y:\Omega\times T\to Y_e,\ u:\Omega\times T\to U,\\
&\{A(u)\in\mathbb{R}_{st}^{n_x\times n_x},\ \forall\,u\in U\},\ \{C(u)\in\mathbb{R}_{st}^{n_y\times n_x},\ \forall\,u\in U\};
\end{aligned}$$

where X, Y, U are called respectively the *state set*, the *output set*, and the *input set*, x and y are finite-valued processes assumed to be in the indicator representation, $u : \Omega \times T \to U$ is a stochastic process, the matrix $A(u)$ is called the *state transition matrix* for the input value $u \in U$, and the matrix $C(u)$ is called the *output matrix* for the input value $u \in U$.

The system representation defined above satisfies the state-output conditional independence condition of Def. 5.7.4 see also Proposition 5.7.6 for the associated system representation of the state process.

Define a *time-invariant state-finite stochastic control system representation* as a special case of the above representation in which only the state and the input processes are present and the output process is not present.

An input of an output-finite-state-finite stochastic system representation acts differently on the state dynamics than in a time-invariant Gaussian system; the probability distribution of the tuple of the next state and the current output can be entirely affected by the input, depending on the form of the system matrices A and C.

A time-invariant output-finite-state-finite stochastic control system representation is a special case of the system representation of Def. 10.4.1.

Definition 10.4.3. Call an output-finite-state-finite stochastic control system *irreducible and nonperiodic* if, for any input value $u \in U$, the state transition matrix $A(u) \in \mathbb{R}_{st}^{n_x \times n_x}$ is an irreducible and nonperiodic matrix, Def. 18.8.3.

A sufficient conditiion is needed which implies that,

$$\forall\, k \in \mathbb{Z}_+,\ \forall\, u_1,\ u_2,\ \ldots,\ u_k \in U,\ A(u_k)\, A(u_{k-1})\ \ldots\ A(u_2)\, A(u_1) \in \mathbb{R}_{st}^{n_x \times n_x},$$

is an irreducible and nonperiodic matrix.

10.4.2 Stochastic Controllability

Is a time-invariant output-finite-state-finite stochastic control system representation a controllable system as defined for Gaussian stochastic control systems? Controllability of this set of stochastic systems requires a rather detailed investigation. Recall from Subsection 5.7.4 that in a state-finite stochastic system one distinghuishes initial subsystems, transient subsystems, and terminal subsystems which is a partition of the system. Can a similar distinction be made for stochastic control systems? Distinguish the cases:

1. The input can entirely destroy the partition of the stochastic system into initial subsystems, transient subsystems, and terminal subsystems. It could even be that the input changes for each time the partition of the system into subsystems. A way to proceed is to investigate every partition possible by the use of input values and then select those which satisfy best the control objectives. This is a complicated procedure but it may work in specific cases. The exploration complexity is finite but in practice it can be large.

2. Assume that the input values do never affect the partition of the stochastic system into initial subsystems, transient subsystems, and terminal subsystems. It depends on the way the input affects the finite stochastic system considered whether this assumption holds.

Below attention is restricted to the second case.

Example 10.4.4. Consider a state-finite stochastic control system with representation,

$$n_x = 6,\ n_u = 1,\ U = \{0,\ 1\},$$
$$A(u) = \begin{pmatrix} 1/3 & 3/4 & 0 & 0 & u & 0 \\ 2/3 & 1/4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/6 & 6/7 & 1-u & 0 \\ 0 & 0 & 5/6 & 1/7 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}_{st}^{6\times 6},\ x_0 = e_6 \in \mathbb{R}_{st}^{n_x}.$$

Thus the stochastic system starts at state $x_0 = e_6$. At time $t = 1$ it is at state $x(1) = e_5$. If $u(1) = 0$ then state $x(2) = e_1$ while if $u(1) = 1$ then $x(2) = e_3$.

The above example can be modified. Assume that $U = [0.2,\ 0.8]$. Then for any u the probability $P(x(2) = e_i)$ of the state $x(2)$ is the stochastic vector $p_x(2) = \begin{pmatrix} u & 0 & 1-u & 0 & 0 & 0 \end{pmatrix}^T$. Thus part of the probability mass goes to the terminal subsystem with states $(e_1,\ e_2)$ and part of the probability goes to the terminal subsystem with the states $(e_3,\ e_4)$.

The conclusion is that there exists a state-finite stochastic control system with one initial subsystem and two terminal subsystems which one can control by the input to go to either of the two terminal subsystems. The second part of the example shows the existence of another stochastic system where the probability mass is partitioned over the two terminal subsystems.

Example 10.4.5. Consider another state-finite stochastic control system specified by the represenation,

$$n_x = 4,\ n_u = 4,\ U = [0.2,\ 0.8]^4,\ x_0 = e_1 \in \mathbb{R}_{st}^{n_x},$$
$$A(u) = \begin{pmatrix} 1-u_1 & 0 & 0 & u_4 \\ u_1 & 1-u_2 & 0 & 0 \\ 0 & u_2 & 1-u_3 & 0 \\ 0 & 0 & u_3 & 1-u_4 \end{pmatrix} \in \mathbb{R}_{st}^{4\times 4},\ u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} \in U.$$

Thus the stochastic system starts at state $x_0 = e_1$. The system matrix is irreducible and nonperiodic. By a choice of the input vector u one obtains a steady state with a higher or lower probability at the four states and hence a longer sejourn times at the various states.

The formal definition of stochastic controllability of a state-finite stochastic control system is proceeded by an informal discussion. Recall from the discussion of stochastic controllability of a Gaussian stochastic control system the formulation to determine stochastic controllability. One first formulates a set of probability measures of the stochastic system, called the *set of control-objective state probability vectors*. Denote this set by $P(X_e)_{conobj} \subseteq P(X_e)$ Secondly, one determines the set of *attainable probability vectors* meaning all probability vectors on the state set which can be attained by the control system for a sequence of inputs or an input trajectory. In terms of notation, $P(X_e)_{attset} \subseteq P(X_e)$. The stochastic control system is then called stochastically controllable if $P(X_e)_{conobj} \subseteq P(X_e)_{attset}$.

Recall from Section 5.7.4 the concepts of a terminal subsystem, an initial subsystem, and a transient subsystem.

Definition 10.4.6. *Stochastic controllability of a time-invariant state-finite stochastic control system.*

1. Call a terminal subsystem of the considered stochastic control system *stochastically controllable* if the set of attainable probability vectors includes the set of control-objective probability vectors. Call then that terminal subsystem a *stochastically-controllable terminal subsystem*. The set of attainable probability vectors is defined as the set of probability vectors on the state set which can be obtained from an initial probability vector on the state set of the terminal subsystem by application of a set of input vectors. See below how to calculate this set of attainable probability vectors.
2. Call any tuple of an initial subsystem and of a terminal subsystem of the stochastic control system *transient stochastically-controllable* if there exists an initial state of the initial subsystem, if there exists a time $t_1 \in \mathbb{Z}_+$, and if there exists an input trajectory over the interval $T(0 : t_1 - 1)$ such that at time t_1 the set of attainable probability vectors of all states of the terminal subsystem includes the set of control-objective probability vectors. In general, there may be other terminal subsystems which also will attain a strictly positive probability vector. In general also, there may be two or more tuples of initial subsystems and terminal subsystems which are transient stochastically-controllable.

Definition 10.4.7. *Attainable subset of a terminal subsystem.* Consider a time-invariant state-finite stochastic control system. Assume that the system is irreducible and nonperiodic. Hence it is a terminal subsystem.

Define the *attainable subset* of probability vectors of this system by the formula,

$$\begin{aligned} P(X_e)_{attset} &= \text{cone}(\text{conmat}(A(.), p_{x_0})) \cap \mathbb{R}^{n_x}_{st} \subseteq \mathbb{R}^{n_x}_{st}, \\ \text{conmat}(A(.), p_{x_0}) &= \left\{ \begin{array}{l} A(u_1)A(u_2)\ldots\ A(u_{n_s})p_{x_0} \in \mathbb{R}^{n_x}_{st} \mid \\ \forall\, n_s \in \mathbb{Z}_+,\ \forall\, u_1, u_2, \ldots, u_{n_s} \in U. \end{array} \right\}. \end{aligned}$$

Note that the cone is intersected with the probability simplex $\mathbb{R}^{n_x}_{st}$.

Consistent with stochastic realization theory, the condition is imposed that the attainable set is a polytope of the set of probability measures on the state set $\mathbb{R}^{n_x}_{st}$. Needed is a characterization of those state-finite stochastic systems for which the attainable set is a polytope. Due to the fact that state-transition matrix depends on the input variable while the input space could be a subset of the real numbers, it is not clear how to obtain such a characterization. This research issue requires more investigation.

In the literature a weaker concept than stochastic controllability is proposed according to which the probability of any discrete state has to be strictly positive.

Definition 10.4.8. Consider a time-invariant state-finite stochastic control system. Assume that the system is irreducible and nonperiodic. Hence it is a terminal subsystem. Consider a subset of the probability simplex called the *control-objective probability measures* denoted by $P(X_e)_{conobj} \subseteq \mathbb{R}_{st}^{n_x}$.

Call the system *stochastically controllable* to the subset of control-objective probability measures if the attainable subset of probability vectors contains the subset of control objective probability vectors; equivalently, if the following inclusion holds,

$$P(X_e)_{conobj} \subseteq P(X_e)_{attset}. \tag{10.16}$$

A characterization is needed of stochastic controllability of a time-invariant state-finite stochastic control system. More research is needed.

10.5 Further Reading

History. Stochastic control systems with only state processes were first used in the operations research literature, see the books [9, 17]. Rather early books are those of H.J. Kushner, [12, 13]. Another early book on stochastic control systems is that of K.J. Aström, [1]. The formulation of a concept of a stochastic control system as described in this chapter, is not much discussed in the literature. The approach of approximating a continuous-time and continuous-space stochastic control system by a discrete-time and a discrete-space stochastic control system is very useful for control of nonlinear stochastic control systems, see the book of H.J. Kushner and P.G. Dupuis, [15].

Books which cover discrete-time stochastic control systems include, [1, 7, 11, 16, 18].

Stochastic control systems. The concept of a stochastic control system as formulated in this book is not common in the literature.

Stochastic controllability. An early paper on stochastic controllability using a closed-loop formulation, is that of M.M. Connors, [6] who, at the end of the paper, refers for the suggestion of the problem to R.E. Kalman. Another early reference on stochastic controllability is that of D. Blackwell, [3], where the term of producibility of a measure is used. D.P. Bertsekas proposes in [**?**, p. 338] the condition of weak accessibility which is weaker than stochastic controllability.

References on stochastic controllability of Gaussian stochastic control systems are [8, 24]. References on stochastic controllability of nonlinear stochastic system include [4, 10, 21, 20, 23].

Stochastic realization of a stochastic control system is not developepd in this book. However, see Chapter 14, in particular Section 14.3 for an application. See an approach to stochastic realization of a Gaussian stochastic control system motivated by system identification, [22].

Control system construction by system approximation. The synthesis approach of numerical approximation of a continuous-time and continuous-space stochastic control system by a discrete-time and a discrete-space stochastic control system is treated in the book by H.J. Kushner [14] and its extension by H.J. Kushner and P.G. Dupuis, [15]; see also the applications [2, 5, 19].

References

1. K.J. Aström. *Introduction to stochastic control*. Academic Press, New York, 1970. 376, 410, 467, 522, 575, 596
2. S. Bellizzi, R. Bouc, F. Campillo, and E. Pardoux. Contrôle optimal semi-actif de suspension de véhicule. In A. Bensoussan and J.L. Lions, editors, *Analysis and optimization of systems*, volume 111 of *Lecture Notes in Control and Information Sciences*, pages 689–699. Springer-Verlag, Berlin, 1988. 9, 377, 468
3. D. Blackwell. The stochastic processes of Borel gambling and dynamic programming. *Ann. Statist.*, 4:370–374, 1976. 49, 376, 419
4. A. Boyarski. Finite-dimensional attainable sets for stochastic control systems. *J. Optim. Th. Appl.*, 22:429–445, 1977. 376
5. F. Campillo. *Optimal ergodic control for a class of nonlinear stochastic systems - Application to semi-active vehicle suspensions*, pages 1190–1195. IEEE Press, New York, 1989. 377
6. M.M. Connors. Controllability of discrete, linear, random dynamical systems. *SIAM J. Control*, 5:183–209, 1967. 175, 376
7. M.H.A. Davis and R.B. Vinter. *Stochastic modelling and control*. Chapman and Hall, London, 1985. 120, 376, 410, 468, 575, 595
8. M. Ehrhardt and W. Kliemann. Controllability of linear stochastic systems. *Systems & Control Letters*, 2:145–153, 1982. 376
9. R. Howard. *Dynamic programming and Markov processes*. M.I.T. Press, Cambridge, 1960. 376, 467, 525
10. J. Klamka and L. Socha. Some remarks about stochastic controllability. *IEEE Trans. Automatic Control*, 22:880–881, 1977. 376
11. P.R. Kumar and P. Varaiya. *Stochastic systems: Estimation, identification, and adaptive control*. Prentice Hall Inc., Englewood Cliffs, NJ, 1986. 376, 410, 468, 525, 575, 595, 596
12. H.J. Kushner. *Stochastic stability and control*. Academic Press, New York, 1967. 121, 376, 410, 467
13. H.J. Kushner. *Introduction to stochastic control*. Holt, Rinehart and Winston Inc., New York, 1971. 121, 376, 410, 467, 525
14. H.J. Kushner. *Probability methods for approximations in stochastic control and for elliptic equations*. Academic Press, New York, 1977. 377
15. H.J. Kushner and P.G. Dupuis. *Numerical methods for stochastic control problems in continuous time (2nd Ed.)*. Number 24 in Applications of Mathematics. Springer, New York, 2001. 9, 376, 377, 466
16. H. Kwakernaak and R. Sivan. *Linear optimal control systems*. Wiley-Interscience, New York, 1972. 120, 376, 410, 467, 489, 593, 822, 823
17. R.D. Luce and H. Raiffa. *Games and decisions*. John Wiley & Sons, New York, 1957. 376, 410, 467
18. P.S. Maybeck. *Stochastic models, estimation and control: Volume 1,2 and 3*. Academic Press, New York, 1979. 120, 376, 410

19. J.N.T. Schuit. Numerical solution of the Hamilton-Jacobi-Bellman equation for a freeway traffic control problem. BS-N8901, CWI, Amsterdam, 1989. 377
20. Y. Sunahara, S. Aihara, and K. Kishino. On the stochastic observability and controllability for non-linear systems. *Int. J. Control*, 22:65–82, 1975. 376
21. Y. Sunahara, T. Kabeuchi, Y. Asada, and S. Aihara. On stochastic controllability for nonlinear systems. *IEEE Trans. Automatic Control*, 19:49–54, 1974. 376
22. J.H. van Schuppen. Stochastic realization of a gaussian stochastic control system. *Acta Appl. Math.*, 35:193–212, 1994. 377
23. C. Varsan. Controllability of nonlinear stochastic control systems. *Systems & Control Letters*, 2:243–247, 1982. 376
24. J. Zabczyk. Controllability of stochastic linear systems. *Systems & Control Letters*, 1:25–31, 1981. 376

Chapter 11
Stochastic Control Problems

Abstract A stochastic control problem is to determine a control law within a rather general set of control laws such that the closed-loop system meets prespecified control objectives. A stochastic control problem is motivated by control problem of engineering, economics, or other areas of the sciences. The concepts of an information pattern, a control law, and of a closed-loop stochastic system are defined. The main control objectives are stability, optimization of performance, and robustness in case of changes in the control system. Control synthesis at a theoretic level is described. Statistical decision problems are treated as elementary stochastic control problems in which there is no dynamics.

Key words: Stochastic control problems. Control law.

A reader who first learns about stochastic control may focus attention on the Sections 11.2, 11.3, 11.4, and 11.5.

11.1 Control Problems of Stochastic Control

Examples of stochastic control problems follow.

Example 11.1.1. *Overload control for switches in communication systems*. The purpose of this example is to show the role of stochastic control in designing a control gate in order to prevent congestion in a computer controlled telephone switch or a telephone exchange. The purpose of the controller is to maximize the long term rate of successful toll-paying connections. The limiting factors are the wastage of capacity in overload situations due to impatient callers terminating their call requests prematurely, the lack of memory space, and the experience that unsatisfied customers retry their call requests. Overload control is an actual problem in communication engineering, see the papers [16, 45, 46, 47] and the tutorial paper [19].

A brief summary of the operations of a telephone switch follows. After a calling party has requested a connection by lifting the phone off-hook, the telephone switch to which it is connected has to perform a large number of tasks in order to establish the connection with the requested phone. These tasks are: give a dial tone, read in the digits of the destination, check the validity of the destination number, establish a route through the network and through intermediate switches, give a busy or ringing tone etc. In modern electronic switches all these tasks for different call requests running simultaneously are handled by one or more processors.

Clearly call requests are in contention for the real time of the same processor and for the same processor memory. The more call requests there are, the longer the delay in executing the tasks, and hence the longer the delay in establishing the connection. This increases the probability that impatient callers will hang up when only part of their tasks have been carried out or that an admitted call request will be shut out for lack of memory space. Another type of impatient behavior arises from the use of time-out mechanisms, for instance in the search for free allocatable hardware resources such as senders and receivers. When the time limit that was set for such an operation expires, a call request will be abandoned before the connection is established. In each case, processor time and memory space have been wasted on tasks that do not lead to a successful connection. It has been recognized a long time ago that this phenomenon can lead to an instability of the communication system in the sense that an increase in the rate of call requests may lead to a decrease of the rate of successful connections. When this happens one says that there is *congestion.* The *overload control problem* is then to regulate admission to the switch to prevent congestion. A further complication is the positive feedback due to retrials. Unsuccessful call requests lead to unsatisfied customers. There is then a high probability that a new call request appears at the switch soon afterwards and a second or higher-number of call requests increases the arrival rate even further.

A simple model for a controlled switch consists of an access gate and a queueing system. The queueing system has one server and a finite waiting room. Assumptions must be made on the interarrival times of tasks and the service times of the server. Customers are for example served according to the First-Come-First-Served service rule. A newly arriving customer is for example admitted if the waiting space of the queue is not fully occupied upon arrival otherwise the call request is rejected and then the customer, he or she, is lost. Once a customer is admitted, his sojourn time limit is generated according to some probability distribution. This limit represents the time a customer is willing to wait for the processing of his call request, and the time-out mechanisms. The customer is served either immediately or, if the server is currently busy, joins a queue and waits for service. Subsequently the customer stays in the queueing system until his service is completed. The service completion of a customer is defined to be *successful* if the actual sojourn time of the customer has not exceeded his sojourn time limit.

The objective is to maximize the *call completion rate* which is defined as the mean number of successful service completions in equilibrium per time unit. This objective may be formulated as a stochastic control problem. The class of admissable control laws could consist of randomized control laws or randomized policies. According to such a randomized control law, a customer is admitted if the toss of coin comes out on head.

The optimal control problem is to maximize the call completion rate of the communication system.

The optimal control law of the control problem will, with respect to conditions, belong to the subclass of control laws of *bang-bang* type. A control law is said to be of bang-bang type if an arriving customer is either accepted or rejected and that no randomization of the input takes place, say a customer is accepted with probability a third. Often the control law is such that a newly arriving customer is accepted if and only if the current number of customers, waiting or being served, does not exceed a particular level. Once it is proven that the optimal control is of bang-bang type then control engineers can develop the threshold of the control law by simulations for a range of parameter values.

Interesting problems concern the model with two or more arrival streams of call requests from customers and from switches elsewhere in the network.

Example 11.1.2. *Maintenance of a machine* Consider a machine at a manufacturing company. The reader may think of a machine producing cookies, or metal products, or of a copy machine. The machine may be in either of two states: it is operating or it is down. The latter state means it is not operating at all. If the machine is down it may be repaired. After a repair the machine may move to the operating state. Since the company aims at making a profit, the production process should be running almost continuously, and the machine must be in the operating state as often as possible. To prevent the machine from going to the down state, preventive maintenance may be performed. It is possible to perform full maintenance, light maintenance, or no maintenance at all.

The control problem is to advise the company manager about on a maintenance control law or policy. The control law should specify when to execute no, light, or full maintenance so as to achieve overall minimal costs. The uncertainty in this problem is in the time moment the machine goes from the operating state to the down state, and in the uncertainty in how long a repaired machine will remain in the operating state. The reader may think of the maintenance of her or his car or of the laundry machine at home.

A mathematical model for this problem may be formulated as follows. Let (Ω,F,P) be a probability space, $T=\mathbb{N}$, $X=\{0,1\}$, and let $x:\Omega\times T\to X$ be the state process. Here the following representation is used:

$$x(t)=\begin{cases}0, & \text{machine in down state,}\\ 1, & \text{machine in operating state.}\end{cases}$$

Let $U=\{\ 0{,}1{,}2\ \}$, $u:\Omega\times T\to U$ with representation,

$$u(t)=\begin{cases}0, \text{ no maintenance or repair,}\\ 1, \text{ light maintenance,}\\ 2, \text{ full maintenance.}\end{cases}$$

The probability measure is specified by,

$$P(\{x(t+1)=i\}|\{x(t)=j,u(t)=u\})=q(i,j,u), \tag{11.1}$$

with $q : X \times X \times U \to [0,1]$. To be definite, one may take q of the form

$$q(1,1,0) = 0.6,\ q(1,1,1) = 0.8,\ q(1,1,2) = 0.95,$$

$$q(1,0,0) = 0,\ q(1,0,1) = 0.5,\ q(1,0,2) = 0.9.$$

Consider a control law $g : X \to U$. Let $u^g(t) = g(x(t))$, hence the transition measure is $q(x(t+1), x(t), g(x(t)))$. If the control objective is to maximize profit, then one may consider the cost criterion,

$$J(g) = E[\sum_{t=0}^{t_1}(c_1 x(t) - c_2 u^g(t))]. \tag{11.2}$$

The expression reflects that the profit goes up if the machine is in the operating state while maintenance reduces the profit. The parameters of the cost function, $c_1,\ c_2$, model respectively the usefulness of the machine in the operating state and cost of repair of the machine. It is possible to include in the model the speed of the repair process, including having a supply of spare parts of the machine for a speedy repair.

A stochastic control problem is then to determine the control law $g^* : X \to U$ such that the criterion J is maximized. Note that the control law g specifies in both the down state and the operating state whether to use full maintenance, light maintenance, or no maintenance. The control law g^* that maximizes J will depend on the function q and on the cost function J.

Example 11.1.3. *Inventory control.* Consider a shop that sells a particular good, say radios, to customers. The shop manager can order radios from a company. The aim of the shop is to operate at a profit. In this example the product is described as a radio so as to have an example of electrical engineering known to most readers. The inventory control problem applies to most products.

The control problem is: When to order new radios given knowledge of the current stock and past demand? The main uncertainty in this problem is in the future demand of customers for radios. The profit is the difference between income and cost. Income is determined by the demand that can be met. The costs consist of: (1) the radios ordered from the supplier; (2) a holding cost for radios not sold in a time period; (3) a cost for demand that cannot be met because it exceeds the available stock.

The stochastic control problem is to determine a control law that will specify how many radios to order at every time moment given information on the current stock.

A mathematical model may be formulated as follows. Let (Ω, F, P) be a probability space, $T = \{0,1,\ldots,t_1\}$, $X = U = W = \mathbb{N} = \{0,1,\ldots\}$, $x : \Omega \times T \to X$, $u : \Omega \times T \to U$, $w : \Omega \times T \to W$ with the representation,

$x(t)$	Stock available at beginning of the t-th period
$u(t)$	Stock ordered and immediately delivered at the beginning of the t-th period
$w(t)$	Demand in t-th period

The state dynamics are then described by the equation,

$$x(t+1) = max\{0, x(t) + u(t) - w(t)\}. \tag{11.3}$$

Assume for simplicity that $\{w(t), t \in T\}$ is a collection of independent random variables. A more realistic model may have a weekly and a seaonal pattern of demand for radios, but such a model is not discussed in this example. A control law or policy is a function $g : X \to U$. With $u(t) = g(x(t))$ the model becomes fully specified. It is assumed that excess demand, occuring if $x(t) + u(t) - w(t) < 0$, is lost to the shop. Again, the model can be extended to allow a number of disappointed customers to return in the next period.

Consider next the control objective of making profit. The profit is specified by the expression

$$\begin{aligned} J(g) = E[\sum_{t=0}^{t_1}(p\ \min\{w(t), x(t) + u(t)\} - (cu(t) + a\ \max\{w(t) - x(t) - u(t), 0\}) \\ -h\ \max\{x(t) + u(t) - w(t), 0\})] \end{aligned} \tag{11.4}$$

where per unit stock,
p Sale price,
c Price at which radios can be ordered,
h Holding price per period,
a Cost in case of excess demand.

The stochastic control problem is then to determine a control law g^* that maximizes the criterion J. Such a control law specifies how many radios to order at a specified stock level.

Stochastic control problems may be classified in regard to the control objectives of which several examples follow:

1. minimize the variance of the controlled output over the considered horizon;
2. schedule maintainance of service teams such that the uptime of equipment is maximized while the repair costs are minimized;
3. control the access to a service facility so that the success rate of service is maximized over time while the finite capacity of the server is respected;
4. route messages or tasks to one of two or more servers or service lines each having a finite capacity, so as to minimize their time to a destination;
5. control the distribution of an investment portfolio so as to maximize the expected profit and simultaneously minimize a risk measure;
6. how to maximize the growth rate of a national economy while minimizing the environmental impact.

In engineering, the main control objectives are: (1) to minimize the variance of the controlled output, and (2) to maintain stability and performance in case of constraints on state and on input variables. An example of minimization of the variance is the course keeping of a ship. An example of maintaining stability is control of motorway traffic, overload control of communication systems, and control of power systems. In economics, the main control objective is optimization of performance. Examples are those described earlier in this section.

11.2 Control Laws

In the examples of Section 11.1 it has been illustrated that one may specify the dependence of the input process on the state process by a control law. Below this specification is formalized.

The discussion is started by the concept of information structure. This concept will be useful for stochastic control with partial observations and it is necessary for control of networked stochastic systems. The term *information structure* was used by H.S. Witsenhausen in [53] and in the book by J. Marschak and R. Radner, [31, pp. 47–49].

Definition 11.2.1. Consider a stochastic control system as defined in 10.1.1.

An *information structure* is a a σ-algebra family, $\{G_t, t \in T\}$, possibly a filtration, such that, at any $t \in T$, the σ-algebra G_t specifies all the information available to the controller for the construction of the input value of $u(t)$.

An information structure is a measure theoretic way to describe on what information the input process is allowed to depend. Let $\{G_t, t \in T\}$ be an information structure for a stochastic control system. For any $t \in T$, $u(t)$ must be measurable with respect to G_t. In case $X = \mathbb{R}^n$, $t \in T$, and $G_t = F_t^{x-} = \sigma(\{x(s), s \leq t\})$, $u(t)$ is measurable with respect to G_t is equivalent with the existence of a function $g_t : X^{t+1} \to U$ such that $u(t) = g_t(x(0), x(1), \ldots, x(t))$.

In control theory one distinguishes:

- *control with complete observations* in which the information structure contains the complete past of the state process; and
- *control with partial observations* in which the information structure provides only partial information of the past of the state process; for example, less than half of the state components are observed;
- *decentralized control* of a stochastic system which consists of two or more inputs while each input is based on partial observations and the partial observations of the various inputs are different;
- *distributed control* of a stochastic system in which the control system is a network with two or more subsystems and in which there is a control law for every subsystem while each control law is based on partial observations.

Special cases of information structures follow.

Definition 11.2.2. Consider a stochastic control system as defined in 10.1.1.

(a) The *past-state information structure* is given by the filtration or the σ-algebra family $\{F_t^{x-}, t \in T\}$ where $F_t^{x-} = \sigma(\{x(0), x(1), \ldots, x(t)\})$.
(b) The *Markov information structure* is given by $\{F^{x(t)}, t \in T\}$.
(c) The *past-output information structure* is given by $\{F_{t-1}^{y-}, t \in T\}$. Thus at time $t \in T$, the controller has available the information of,

$$F_{t-1}^{y-} = \sigma(\{y(0), \ldots, y(t-1)\}).$$

(d)The *current-output information structure* is given by $\{F^{y(t)}, t \in T\}$.

(e)An information structure $\{H_t, t \in T\}$ is called a *classical information structure* if (1) there is only a single controller and one information structure for this controller; and (2) the information structure has *perfect recall*, or, equivalently, if for all $t \in T$ such that $t+1 \in T$, $H_t \subseteq H_{t+1}$. The reader may think for H_t of the *history* at time $t \in T$. It is called a *non-classical information structure* if either there is no perfect recall or there are two or more controllers.

Information structure play an important role in decision and control problems with two or more controllers where the information structures of the available controllers are in general different.

Note that a Markov information structure does not have perfect recall because $F^{x(t)}$ is in general not contained in $F^{x(t+1)}$. Thus, the Markov information structure $x(t)$ of time t is no longer available to the controller at time $t+1$.

Note further that the past-state information structure has perfect recall because $F_t^x \subset F_{t+1}^x$, for all $t \in T$. Similarly, the past-output information structure has perfect recall.

Definition 11.2.3. Consider a stochastic control system. Define the following control laws. See also Def. 11.3.1 of a closed-loop system for the use of the control law.

(a)Consider the past-state information structure. A *past-state control law* is a collection of measurable maps $g = \{g_0, g_1, g_2, \ldots\}$ such that for all $t \in T$, $g_t : X^{t+1} \to U$.

(b)Consider the Markov information structure. A *Markov control law*, is a measurable map $g : T \times X \to U$. Denote by G_M the set of Markov control laws.

(c)Consider the Markov information structure. A *time-invariant Markov control law* is a measurable map $g : X \to U$.

(d)Consider the past-output information structure. A *past-output control law* is a collection of measurable maps,

$$g = \{g_t,\ t \in T | g_0 \in U,\ \forall\ t \in T,\ g_{t+1} : Y^t \to U\}.$$

(e)Consider the current-output information structure. A *current-output control law* is a measurable map $g : T \times Y \to U$.

Definition 11.2.4. *Random control laws*. A *random control law* with the past-state information structure is defined as a family of maps so that,

$$\{g_t,\ \forall\ t \in T\},\ g_0 : \Omega \to U,\ \forall\ t \in T(1:\infty),\ g_t : X^{t+1} \to P(B(U)).$$

where $P(B(U))$ denotes the set of probability measures on the input space $(U, B(U))$.

The operation of a random control law is described next. If the control system at time $t \in$ has the state $x(t)$ then the random control law describes the probability measure $P_{u(t)} \in PM(U)$ on the input space U. A second mechanism to be specified then makes a choice of the actual input value $u(t)$ based on the measure $P_{u(t)}$ on U. The reader may think of the case of $U = \{1, 2\}$ in which the control law requires a toss of a coin to determine whether the input should be either 1 or 2. Or the reader can think of the space $U = \mathbb{R}^{n_u}$ in which a random number generator determines an input value with the measure $P_{u(t)}$.

Random control laws were first used in game theory. In control theory they have been considered to explore the use of this rather general set of control laws. In information theory there are also used to allow a general framework.

In game theory, the set of random control laws plays an essential role. For example, for the existence of a Nash equilibrium. In control theory, its usefulness is limited. It appears that most results are of the type that if the set of random control laws is considered, then the result is often that the optimal control law belongs to the set of deterministic control laws which map, for each state, to a deterministic value of the input space. This applies also to most cases of information theoretic problems where the set of random control laws is applied. Such results are of theoretical interest only.

11.3 Closed-Loop Stochastic Control Systems

Once a stochastic control system and information structure are specified and a control law is selected, then one can close the loop, see Figure 11.1. In this chapter are described only closed-loop systems based on complete observations.

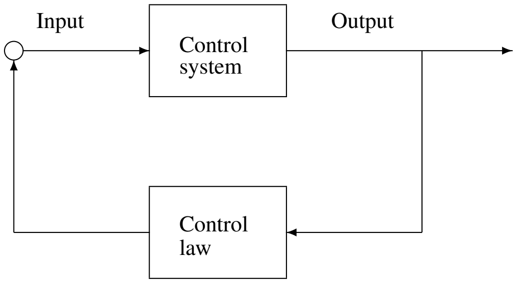


Fig. 11.1 A closed-loop stochastic control system.

Definition 11.3.1. A *closed-loop Gaussian stochastic control system representation*. Consider a time-varying Gaussian stochastic control system representation, as defined in Definition 10.2.1, with a controlled output,

$$x(t+1) = A(t)x(t) + B(t)u(t) + M(t)v(t), \quad (11.5)$$

$$z(t) = C_z(t)x(t) + D_z(t)u(t), \quad (11.6)$$

consider a past-state information structure $\{F_t^{x-},\ t \in T\}$, and consider a past-state control law $g \in G$, forall $t \in T$, $g_t : X^{t+1} \to U$.

Then define *recursively* according to the following equations the state process x^g and the input process u^g, of the closed-loop system by the equations,

$$\begin{aligned}
& x^g : \Omega \times T \rightarrow \mathbb{R}^{n_x},\ u^g : \Omega \times T \rightarrow \mathbb{R}^{n_u}, \\
x^g(t+1) &= A(t)x^g(t) + B(t)g_t(x^g(0:t)) + M(t)v(t),\ x^g(0) = x_0, && (11.7)\\
z^g(t) &= C_z(t)x^g(t) + D_z g_t(x^g(0:t)), && (11.8)\\
x^g(0:t) &= (x^g(0), x^g(1), x^g(2), \ldots, x^g(t)), && (11.9)\\
u^g(t) &= g_t(x^g(0:t)). && (11.10)
\end{aligned}$$

The stochastic system described by the equation (11.7) will be called the *closed-loop Gaussian stochastic control system* associated with the open-loop stochastic control system (11.5) and the control law g.

The order of the operations are specified for the understanding of the reader. The successive calculations are, for any time $t \in T$,

$$\begin{aligned}
& x^g(t),\ \text{assumed available}, \\
x^g(0:t) &= (x^g(0:t-1),\ x^g(t)), \\
u^g(t) &= g_t(x^g(0:t)), \\
x^g(t+1) &= A(t)x^g(t) + B(t)u^g(0:t) + M(t)v(t), \\
& x^g(t+1) \text{ is observed.}
\end{aligned}$$

Note that the state processes x^g is well defined by the recursion (11.7) because x_0, v, and the control law g are specified and the right-hand sides depend only on the variables already available or ‘known’ at time $t \in T$. Note also that in this particular case the closed-loop stochastic control system is such that the processes u and v are no longer independent because $u(t)$ depends on the past of the noise process v. However, the random variable $v(t)$ is independent of $u(t)$ and of the past noise variables $v(s)$ for $s = 0, 1, \ldots t-1$.

To emphasize the dependence of the state process on the control law $g \in G$ the notation x^g is used. The super index g will be omitted in later sections when the context is clear.

The reader should *distinguish* between (1) a control law g, for example $g : T \times X \rightarrow U$, and (2) the associated input process $u^g : \Omega \times T \rightarrow U$, that may for example be related by $u^g(t) = g(t, y(t-1))$. The control law is the most important object, the input at a particular time is less important than a control law. The literature is not always clear on this issue.

What properties does a closed-loop stochastic control system have? This question will be investigated in this chapter for the case in which there is no output process.

Proposition 11.3.2. *Consider a stochastic control system consisting of a state process only*

$$\begin{aligned}
& \{\Omega, F, P, T, X, B_X, U, B_U, x, u\}, \\
& (x(t), u(t)) \mapsto \textit{conditional distribution } (x(t+1)|F_t^{x-} \vee F^{u(t)}).
\end{aligned}$$

Let $g : T \times X \rightarrow U$ be a Markov control law. Then the closed-loop stochastic control system is such that $x^g : \Omega \times T \rightarrow X$ is a Markov process.

Proof. This follows directly from the transition map and the definition of a Markov process.

$$\begin{aligned} &(x^g(t), g(t, x^g(t)) \mapsto \text{conditional distribution } (x^g(t+1)|F_t^{x^g-}), \\ &\text{because, } u^g(t) = g(t,x(t)), \text{ implies that, } F^{u^g(t)} \subseteq F^{x(t)} \subset F_t^{x-}, \text{ hence} \\ &\ \text{conditional distribution } (x^g(t+1)|F_t^{x^g-} \vee F_t^{u^g}) \\ = &\ \text{conditional distribution } (x^g(t+1)|F^{x^g(t)}). \end{aligned}$$

Proposition 11.3.3. *Consider the recursive state-observed stochastic control system of Definition 10.1.2 with representation*

$$x(t+1) = f(t, x(t), u(t), v(t)), \ \ x(t_0) = x_0, \tag{11.11}$$

and a Markov control law $g : T \times X \to U$*. The closed-loop stochastic control system is then defined by*

$$x^g(t+1) = f(t, x^g(t), g(t, x^g(t)), v(t)), x(t_0) = x_0. \tag{11.12}$$

Then x^g *is a Markov process.*

Proof. This follows from Proposition 3.3.7 (a) ⇔ (d) and the assumption that v is a sequence of independent random variables.

Consider the stochastic control system of Proposition 11.3.3, and a past-state control law $g = \{g_0, g_1, \ldots\}$ where for any $t \in T$, $g_t : X^t \to U$. The closed-loop stochastic control system is then,

$$x^g(t+1) = f(t, x^g(t), g_t(x^g(0), \ldots, x^g(t)), v(t)), x(t_0) = x_0.$$

In general the process x^g will then not be a Markov process. Note that at time $t \in T$, the state of the system is the vector, $(x^g(0), x^g(1), \ldots, x^g(t))$ which grows in dimension with the time $t \in T$. Therefore x^g is not the state process of the closed-loop stochastic control system; the reader should not be misled by identifying a process denoted by x^g with a Markov process.

Distinction Between Open-Loop and Closed-Loop

Below it will be explained that in stochastic control, a closed-loop control may in general not be replaced by an open-loop control. However, in control of deterministic systems a closed-loop control is equivalent with an open-loop control.

Consider a stochastic control system of the form

$$x(t+1) = f(t, x(t), u(t), v(t)), \ x(t_0) = x_0, \tag{11.13}$$

where $v : \Omega \times T \to V$ is an independent sequence. Consider also a deterministic control system

$$z(t+1) = h(t, z(t), u(t)), \ z(t_0) = z_0. \tag{11.14}$$

An *open-loop control* will in this context be a function $u : T \to U$. A *closed-loop control* is for example a Markov information structure combined with a Markov control law, say $g : T \times X \to U$, such that the closed-loop control system is given by

$$x^g(t+1) = f(t, x^g(t), g(t, x^g(t)), v(t)), \ x^g(t_0) = x_0. \tag{11.15}$$

In general, a closed-loop control consists of the combination of an information structure and of a control law such that $u(t)$ effectively depends on the state $x^g(t)$ and, possibly, on past states $x^g(s)$ for $s < t$. The notions of open-loop and closed-loop control apply to both deterministic and stochastic systems.

Proposition 11.3.4. *For both a deterministic and a stochastic system, an open loop control is a special case of a closed-loop control.*

Proof. Let $u : T \to U$ be an open-loop control. Consider the Markov information structure and define $g : T \times X \to U$, $g(t,x) = u(t)$. Then g is a Markov control law defining a closed-loop control whose effect is the same as the open-loop control.

Proposition 11.3.5. *For a deterministic system a closed-loop control is a special case of an open-loop control. Hence for a deterministic system a closed-loop control and an open-loop control are special cases of each other.*

Proof. Consider the deterministic control system (11.14). Let $g : T \times X \to U$ be a control law. Define $u^g : T \to U$ by

$$z^g(t+1) = h(t, z^g(t), g(t, z^g(t))), z^g(t_0) = x_0, \tag{11.16}$$

$$u^g(t) = g(t, z^g(t)). \tag{11.17}$$

Then $u^g : T \to U$ is an open-loop control. Note that u^g has the same effect on the closed-loop system,

$$x^g(t+1) = h(t, x^g(t), u^g(t)), x^g(t_0) = x_0, \tag{11.18}$$

as the closed-loop control $g(t, x^g(t))$

$$x^g(t+1) = h(t, x^g(t), g(t, x^g(t))). \tag{11.19}$$

This result and Proposition 11.3.4 then imply that for a deterministic system an open-loop control and a closed-loop control are equivalent.

Proposition 11.3.6. *There exists a stochastic control system and a closed-loop control, for which there does not exist an open-loop control such that the state process of the closed-loop control system with a closed loop and the state process of the closed-loop control system with an open loop, have the same probability distributions.*

The proof of Proposition 11.3.6 is provided by the following example.

Example 11.3.7. *Closed-loop control* Consider the stochastic control system

$$x(t+1) = x(t) + u(t) + v(t), \;\; x(0) = x_0, \tag{11.20}$$

where $x_0 : \Omega \to \mathbb{R}$, $E[x_0] = 0$, $E[x_0^2] = q$, $v : \Omega \times T \to \mathbb{R}$ is an independent identically distributed sequence with for all $t \in T$, $E[v(t)] = 0$, $E[v(t)^2] = q$, $q > 0$. It is assumed that x_0 and v are independent. Consider the Markov information structure, and the Markov control law $g : \mathbb{R} \to \mathbb{R}$, $g(x) = -x$. The closed-loop stochastic control system is then given by

$$x^g(t+1) = x^g(t) + g(x^g(t)) + v(t) = v(t), \;\; x^g(0) = x_0. \tag{11.21}$$

Note that the state process $x^g : \Omega \times T \to \mathbb{R}$ is given by $x^g(t+1) = v(t)$. Also $u^g(t) = -x^g(t)$. Hence x^g and u^g each are a sequence of independent random variables with, for all $t \in T$, $E[x^g(t)] = 0$ and $E[(x^g(t))^2] = q$. Next consider any open-loop control $u : T \to R$. The closed-loop control system is then given by

$$x^u(t+1) = x^u(t) + u(t) + v(t) = x_0 + \sum_{s=0}^{t-1} u(s) + \sum_{s=0}^{t-1} v(s), x^u(0) = x_0. \tag{11.22}$$

Then $x^u : \Omega \times T \to \mathbb{R}$ is a stochastic process with

$$\begin{aligned} E[x^u(t)] &= \sum_{s=0}^{t-1} u(s), \\ E[(x^u(t) - E[x^u(t)])^2] &= E[x_0^2] + \sum_{s=0}^{t-1} E[v(s)^2] = (t+1)q. \end{aligned}$$

If one requires that for all $t \in T$, $E[x^g(t)] = E[x^u(t)]$ then $u(t) = 0$ for all $t \in T$. But then, for all $t \in T$, $t \geq 1$

$$E[x^g(t)^2] = q \neq (t+1)q = E[(x^u(t) - E[x^u(t)])^2]. \tag{11.23}$$

Hence for the closed-loop control g specified above, there does not exist an open-loop control achieving the same distribution of the state process of the closed-loop control system.

11.4 Stochastic Control Problems

Consideration of many engineering control problems leads for control theory to the following control theoretic problems.

Definition 11.4.1. Define the following subsets of engineering control problems.

1. *Set point control*. Keep the controlled output of the system as close as possible at a particular value called the *set point*.

2. *Minimal variance control*. Minimize the variance of the controlled output of the system over the considered horizon.
3. *Optimal control of a performance measure*. Minimize the performance measure of the system over the considered horizon.
4. *Tracking control*. Let a function of the state of the system or of the controlled output, follow a prespecified time function as closely as possible. The state function therefore *tracks* the prespecified time function.

Consider a stochastic control system, an information structure, and a class of control laws. The stochastic control problem is then to select a control law such that the closed-loop stochastic control system meets control objectives. But what are control objectives? And how to determine a control law?

Definition 11.4.2. A *control objective* is a property of the closed-loop control system that the control engineer strives to attain.

Particular control objectives of a control problem which have been used in control theory include:

1. *Stability*: of the state process, the controlled output process, or of the observation process;
2. *Optimal control*: Infimize a cost function over the set of control laws;
3. *Control in case of uncertain dynamics*: Satisfactory performance in case the dynamics of the system is different from what it was assumed to be; one refers to this case as one of uncertain dynamics;
4. *Robustness* of stability and of performance in case of uncertain dynamics and of disturbance signals: a satisfactory stability and a satisfactory performance if the model system, with which the control law is synthesized, is only an approximation of the actual engineering system or if it is affected by an additional disturbance signal not included in the model system;
5. *Control for online system identification*: use the observations produced by the control system to obtain an estimate of the parameter values of the system with the controlled output; a further going procedure is to probe the system with an extra excitation input and to deduce from the output of the system what the current values of the parameters of the system are; this way one obtains a currently realistic model of the control system;
6. *Adaptation*: Satisfactory performance if the actual engineering system changes slowly over time; the control objective may be achieved by successively: (1) carrying out control for online system identification; and (2) adjusting the stucture of the control law to the currently-estimated realistic model of the control engineering system.

It may not be possible to attain a control objective, although it is worth striving for a control objective. Often a stochastic control problem has several control objectives. Several or all of these objectives may be conflicting. One is then forced to strike a balance between these objectives. Optimal control theory is useful to find the optimal balance.

Definition 11.4.3. A *performance measure* of a closed-loop stochastic control system in combination with a control law, is a real-valued cost criterion which models the performance of the closed-loop system as well as possible. The definition has to be such that both (1) the control objectives and (2) the costs of the use of the inputs, are appropriately represented in the performance measure. The control problem is often to find the optimal balance between achieving the control objectives and the costs of inputs.

A performance measure will always be taken as a cost rate and a terminal cost, both of which are functions of the controlled output, see Def. 10.1.3. A performance measure is formulated in terms of the controlled output of a closed-loop stochastic control system,

$$z^g(t) = h(x^g(t), u^g(t)), \ \forall t \in T(0:t_1-1), \tag{11.24}$$

$$z^g(t_1) = h(x^g(t_1)). \tag{11.25}$$

As performance measure one then often chooses a norm of the controlled output, where the norm can be chosen based on the analytic structure of the system.

Problem 11.4.4. *The stochastic control problem.* Consider a stochastic control system, an information structure, a set of control laws, and control objectives. Determine a control law in the considered set of control laws such that the closed-loop stochastic control system satisfies the control objectives as well as possible.

There follows a discussion of the control objectives formulated above. These objectives may be used in connection with deterministic and with stochastic systems.

The control objective of minimization of a performance measure. An example of this control objective is that where the closed-loop control system exhibits a minimum variance of the controlled output.

Example 11.4.5. The control problem of minimizing a norm of the controlled output. Consider a time-invariant Gaussian stochastic control system representation

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) + Mv(t), \\ y(t) &= Cx(t) + Du(t) + Nv(t), \\ z(t) &= C_z x(t) + D_z u(t), \ \forall \, t \in T(0:t_1-1), \ z(t_1) = Cx(t_1). \end{aligned}$$

The process z denotes the controlled output. The controlled output represents a functional of the state and of the input process for which one may want to specify control objectives. For example, the control objective of minimizing the quadratic form,

$$J(g) = E\left[z(t_1)^T z(t_1) + \sum_{s=0}^{t_1} z(s)^T z(s) \right], \ J : G \to \mathbb{R}_+.$$

Example 11.4.6. *Minimum variance control* Consider a time-invariant Gaussian stochastic control system with representation

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) + Mv(t), \\ y(t) &= Cx(t) + Du(t) + Nv(t), \ v(t) \in G(0, I). \end{aligned}$$

Consider further the past-output information structure, and the set of past-output control laws G. Suppose that for any $g \in G$, the processes (x^g, y^g) are jointly stationary. Consider the variance of the output process

$$W_{y^g}(0) = E[(y^g(t) - E[y^g(t)])(y^g(t) - E[y^g(t)])^T].$$

The control objective of *minimum variance control* is to make the cost criterion of the trace of the output variance, $tr(W_{y^g}(0))$, as small as possible. This control objective has been used in connection with the control of a paper machine, see Example 4.1.1. It seems appropriate for a set of quality control problems of industrial firms. However, this control objective ignores the values of the input used to achieve the minimal variance.

In Example 11.1.2 on the maintenance problem for a machine, the objective is to maximize profit. This control objective may be modeled by the linear functional,

$$J = E[\sum_{t=0}^{t_1}(c_1x(t) - c_2u(t))].$$

In Example 11.1.3 on inventory control the objective is to maximize profit in the form of maximization of the criterion (11.4)

The control objective of control in case of uncertain dynamics. A stochastic control system is a mathematical model of a engineering control system exhiting uncertain dynamics. As such it is an approximation of reality of the phenomenon to be modeled. The model system is only an approximation of the actual engineering control system, say a car or an airplane.

Consider a deterministic control system. By analogy with Definition 10.1.1 one may define a deterministic control system for a state function $x : T \to X$ and an input function $u : T \to U$ by the transition map,

$$(x(t), u(t)) \mapsto x(t+1) = f(t, x(t), u(t)).$$

Suppose the dynamics is uncertain. To model this uncertainty the transition map should be redefined. One way to redefine it is to let $(x(t), u(t))$ map into a set, an uncertainty region, of which $x(t+1)$ will be a member. The latter approach is used in the research topic of differential inclusions. Another approach is to use a stochastic control system in which,

$$(x(t), u(t)) \mapsto \text{ conditional distribution } \{x(t+1)|F_t^{x-} \vee F_\infty^u\}$$

the transition map determines a measure on the state space. Therefore stochastic control systems account by definition already for the uncertainty in the dynamics. For deterministic control systems this control objective is more to the point.

The control objective of robust control: Satisfactory performance even when working with an approximate model or in case the control system is affected by additional disturbance signals.

There is a minor difference between this control objective and the one described directly above. Engineering practice shows that these control objectives are different and lead to different control laws.

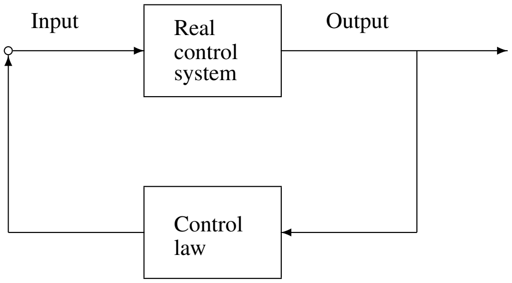


Fig. 11.2 The real control system and a control law.

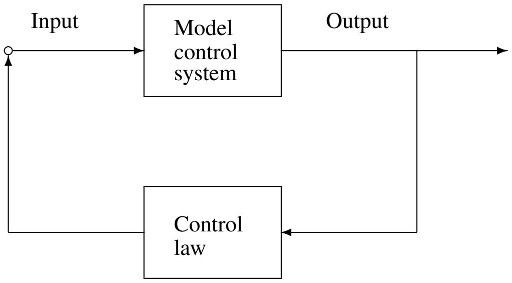


Fig. 11.3 The model control system and a control law.

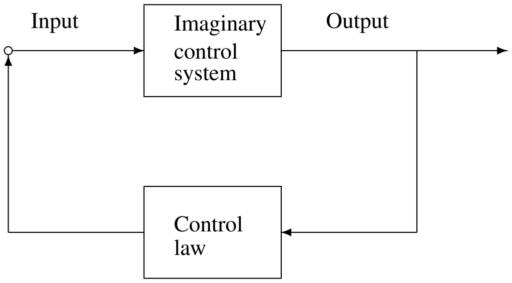


Fig. 11.4 The imaginary control system and a control law.

Definition 11.4.7. Consider next the closed-loop control system consisting of the control system and a control law. The closed-loop system is defined for the *real control system*, the *model control system*, and the *imaginary control system*, see the Figures 11.2, 11.3, and 11.4.

As mentioned before, a control system is as a mathematical model an approximation of a control process. For example, consider the problem of control of a paper machine, Example 4.1.1. A mathematical model for control of basis weight has been presented in the form of a Gaussian stochastic control system. The system identification procedure described in the paper yields values for the parameters of this system. However, the actual process of paper production can in general not be described by a Gaussian stochastic control system, and if it can then the values of the parameters corresponding to the process will differ from the estimated values. The conclusion has to be that one has to distinguish the *real control system* as it exists in the manufacturing company from the *model control system* in the form of a control system of an engineer.

However, in the case of robust control one also considers the *imaginary control system*. An imaginary control system is actually a set of control systems which include the model control system but also control systems with slightly different and possibly realistic other control systems. In several papers, the set of imaginary control systems is defined by a norm either on the input-output map or on the parameter values of the model system.

In Example 4.1.1 of the paper machine, the *real control system* is the actual chemical-mechanical control system at the paper company. The *model control system* is that formulated by the engineers who carried out the project. The *set of imaginary control systems* could be that of Gaussian control systems of the same state-space dimension as that of the model control system with a prespicified range of parameter values.

In the control problem with the control objective of robust control, one wants is that the corresponding closed-loop control system achieves the specified control objectives, not only for the real system, but also for the model system, and also for a set of imaginary systems.

Whether the control objectives of robust control are actually achieved by a determined control law has to be evaluated either by simulation, by analysis, by optimization, or by experiments on the actual real system.

There is a difference between the control objective of robust control and that of the control objective of control in case of uncertain dynamics. The latter control objective concerns satisfactory performance when faced with uncertainty in the dynamics that have effect on a relatively short time scale. The control objective of robust control concerns satisfactory performance due to changes that have effect on a relatively long time scale. The first objective may usually be satisfied by stochastic control techniques while the latter objective may require the use of robust control and of adaptive control. In engineering practice a control engineer will carry out extensive simulations for various situations to determine whether the control performance is robust against the expected changes in the dynamics and in the disturbance signals.

Example 11.4.8. *Example. Control paper machine* Consider again the problem of control of a paper machine. As stochastic control system an ARMAX representation is used, that is given by

$$y(t) = \sum_{i=1}^{n} a_i y(t-i) + \sum_{i=1}^{n} b_i u(t-i) + \sum_{i=1}^{n} c_i v(t-i).$$

This representation is equivalent to a Gaussian stochastic control system representation. As indicated in Example 4.1.1 the parameters of this representation $\{a_i, b_i, c_i, i \in Z_n\}$ may be estimated from data. It is realistic to say that this representation with parameter estimates is an approximation of the behavior of the paper machine. Moreover, the behavior of the paper machine may change in time, due to environmental conditions or aging of the machine parts. The control objective of robust control seems therefore quite appropriate for this problem. Thus the closed-loop control system should achieve say the control objective of minimum variance control for not just the model control system but also for a related set of imaginary control systems.

11.5 Control Synthesis and Control Design

In solution methods for stochastic control problems one distinguishes:

- *Control synthesis* which concerns the existence of a control law and the derivation of its structural properties;
- *Control design* which concerns the actual computation of the control law once its structure is known, including its numerical parameters and simulations to evaluate robustness.

An example of control synthesis is the optimal stochastic control problem for a Gaussian stochastic control system with past-state information structure and with a quadratic cost function. The solution to this problem is a linear control law $u(t) = Fx(t)$. An example of design is then the actual computation of the parameter values of this control law, here of the matrix $F \in \mathbb{R}^{n_u \times n_x}$. Control synthesis will provide an algorithm for this computation, but design involves more than this computation. Control design will have to take care of additional constraints or control objectives that were not specified in connection with control synthesis. Control design is nowadays often performed by programs for computer aided control system design in combination with simulation programs.

In these notes attention is focused on the control synthesis part of the stochastic control problem.

Definition 11.5.1. For a stochastic control problem the following *control synthesis approaches* for control of stochastic systems, are currently used:

- Control synthesis procedures of control of deterministic control systems;
- Optimal stochastic control;
- Establishing structural properties of control laws;
- First approximating the stochastic control system by a finite-state stochastic control system and, secondly, to apply control synthesis procedures for control of finite-state stochastic control systems.

To analyse the performance of a closed-loop stochastic control system the following concept has been defined. The *control tasks* of a controller in a stochastic control problem are:

- Regulation of the controlled variable so as to satisfy the performance measure;
- Gathering of information about the current value of the state process so as to reduce uncertainty about that value; and
- Use the input to excite the system state so as to allow a system identification procedure to produce realistic estimates of the parameters of the system with which one can compute a control law which achieves a lower cost.

These tasks are in general conflicting and in optimal stochastic control a choice is made between these conflicting tasks. For example, the input can be used to generate a large excursion of the state process so that the state can be estimated better after which the control brings the state to zero. But the generation of the state excursion may have a large effect on the actual cost. An optimal control law strikes a balance between the above tasks which overal achieves a minimal cost. In control synthesis there is no procedure to satisfactorily combine these tasks, though optimal control may be used to achieve it.

Control theory of stochastic systems is focused on the following research issues:

1. a formulation of the system as a stochastic control system with a state, an input, and a controlled output;
2. necessary and sufficient conditions for the existence of a control law such that the closed-loop system meets the control objectives; one expects stochastic controllability to be a necessary condition; in case of optimal control, one expects that a strict reduction of the cost function or finiteness of the cost function to require a condition of stochastic controllability;
3. decompositions of the stochastic control system will aid in the understanding of the necessary and sufficient conditions for the existence of the control law;
4. does the optimal control law possess particular structural properties?; such properties can often be proven even if the optimal control law cannot be determined analytically.

The subsequent chapters are structured by the above research issues for control of stochastic control systems.

Control synthesis approach of control theory of deterministic systems. In this approach one uses control techniques, usually inspired by control synthesis approaches for deterministic linear control systems, to arrive at a control law that satisfies the control objectives.

Example 11.5.2. *Example. Time-invariant control law* Consider a time-invariant Gaussian stochastic control system of the form

$$x(t+1) = Ax(t) + Bu(t) + Mv(t), x(t_0) = x_0, \tag{11.26}$$

the past-state information structure $\{F_t^{x-}, t \in T\}$, and the set of past-state control laws G. Suppose that the control objective is to determine a control law $g \in G$ such that the closed-loop control system has an invariant measure, or

$$\lim_{t\to\infty} \text{distribution } (x(t)) = \text{invariant measure such that,}$$
$$E[(x(t)-E[x(t)])^T(x(t)-E[x(t)])] < \infty.$$

If (A,B) is a controllable pair, then there exists a $F \in \mathbb{R}^{n_u \times n_x}$ such that $A+BF$ is an asymptotically stable matrix. Then the control law $g(x) = Fx$ results in the closed-loop stochastic control system

$$x(t+1) = (A+BF)x(t) + Mv(t), x(0) = x_0.$$

It then follows from Theorem 4.4.5 that asymptotically the state process x is stationary with $\mathrm{D}-\lim_{t\to\infty} x(t) = G(0,Q_x)$ and $\mathrm{tr}(Q_x) < \infty$. Thus any feedback matrix $F \in \mathbb{R}^{n_u \times n_x}$ such that $\mathrm{spec}(A+BF) \subset \mathrm{D}_o$ yields that $\mathrm{tr}(Q_x) < \infty$. But then the state variance is not necessarily minimized.

Control synthesis approach of optimal stochastic control. In this approach an optimal stochastic control problem is formulated. All of the most important control objectives are combined into an optimization criterion. This combination achieves a trade-off between the objectives and therefore the cost function has to be determined with due care. This approach is suitable if the control objective of transient response is easily satisfied. Then there is freedom left that may be used to satisfy other control objectives. This approach is also suitable to derive structural properties of control laws.

Consider a stochastic control system, with an information structure, a set of control laws, control objectives, and an optimization criterion. The *optimal stochastic control problem* is then to determine a control law in the given class such that the closed-loop stochastic control system optimizes the criterion.

For the optimal stochastic control problem the dynamic programming approach will be presented in Chapter 12. The Hamiltonian approach to optimal stochatic control problems is less used due to the required strong differentiability conditions.

Example 11.5.3. *Optimal stochastic control*. Consider again a time-invariant Gaussian stochastic control system including the controlled output process, say with the representation,

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) + Mv(t), \\ y(t) &= Cx(t) + Du(t) + Nv(t),\ v(t) \in G(0,Q_v), \\ z(t) &= C_z x(t) + D_z u(t). \end{aligned}$$

Consider further the past-output information structure $\{F^y_{t-1}, t \in T\}$ and a set of control laws G. Thus, for any $g \in G$ and $t \in T$, the closed-loop stochastic control system is given by,

$$\begin{aligned} x^g(t+1) &= Ax^g(t) + Bg_t(y^g(0),\ldots,y^g(t-1)) + Mv(t), && (11.27) \\ y^g(t) &= Cx^g(t) + Dg_t(y^g(0),\ldots,y^g(t-1)) + Nv(t). && (11.28) \\ z^g(t) &= C_z x^g(t) + D_z g_t(y^g(0),\ldots,y^g(t-1)), && (11.29) \end{aligned}$$

$$J_{ac}(g) = \limsup_{t_1\to\infty} \frac{1}{t_1} E\left[\sum_{s=0}^{t_1} (z^g(s)^T z^g(s))\right],\ J_{ac}: G \to \mathbb{R}_+.$$

In the displayed equations, $J_{ac}(g)$ denotes the average cost function of a control law $g \in G$.

An optimal stochastic control problem for this system is to determine a control law $g^* \in G$ such that,

$$J(g^*) \leq J(g), \ \forall \, g \in G.$$

This problem will be solved in Chapter 14 of these notes.

But will the control law $g^* \in G$ requested in the above problem and in Example 11.5.3 be satisfactorily in practice? Note that for any $t \in T$,

$$u^g(t) = g_t(y^g(0), \ldots, y^g(t-1)),$$

hence this control law may require that at $t \in T$ the set $\{y^g(0), \ldots, y^g(t-1)\}$ is available, say stored in the control computer's memory. If t gets large, a large memory is needed! Practical considerations seem to dictate that the control law has finite memory.

It will be useful if the control law consisted of say a nonlinear stochastic system of the form,

$$\bar{x}(t+1) = f(t, \bar{x}(t), y(t-1)), \ \bar{x}(0) = \bar{x}_0, \tag{11.30}$$

$$u(t) = h(t, \bar{x}(t), y(t-1)), \tag{11.31}$$

where $\bar{x} : \Omega \times T \to \mathbb{R}^{n_{\bar{x}}}$, $f : T \times \mathbb{R}^{n_{\bar{x}}} \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_{\bar{x}}}$, $h : T \times \mathbb{R}^{n_{\bar{x}}} \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_u}$. Call the stochastic system (11.30,11.31) a *dynamic controller* for the stochastic control system (11.27,11.28). Because $\bar{x}$ takes values in $\mathbb{R}^{n_{\bar{x}}}$ it may be called a *finite-dimensional dynamic controller.* This term is not fully appropriate here, but for the moment it suffices. The closed-loop stochastic control system then takes the form,

$$\begin{aligned}
x^g(t+1) &= Ax^g(t) + Bh(t, \bar{x}^g(t), y^g(t)) + Mv(t), \ x^g(t_0) = x_0, \\
\bar{x}^g(t+1) &= f(t, \bar{x}^g(t), y^g(t)), \ \bar{x}^g(0) = z_0, \\
y^g(t) &= Cx^g(t) + Dh(t, \bar{x}^g(t), y^g(t-1)) + Nv(t).
\end{aligned}$$

A controller as presented above is easy to implement for control of an engineering system.

The controller presented above puts the problem in a different light. The example motivates that one should search for controllers of the form (11.30,11.31). Control synthesis of such controllers may be achieved by two different approaches.

1. The first approach is to select the combination of stochastic control system and optimization criterion in such a way that the optimal stochastic control problem admits a solution in the form of a finite-dimensional controller. The optimal stochastic control problem mentioned in Example 11.5.3 will turn out to have a solution with this property.
2. The second approach is to limit attention to control laws of the form of a finite-dimensional controller. The disadvantage of this approach is that it leads to an optimization problem over function spaces that is in general difficult to solve analytically.

The first approach mentioned above may be reformulated as a stochastic realization problem. The problem is to determine quadruples of stochastic control systems, information structures, sets of control laws, and optimization criteria such that the corresponding optimal stochastic control problem for this quadruple results in a finite-dimensional controller.

Control synthesis approach of establishing structural properties of control laws. Consider a stochastic control system, an information structure, and a set of control laws G_1. Either the problem formulation or engineering design suggests that a useful set of control laws is G_2 with $G_2 \subsetneq G_1$. The problem is then to show that the optimal control law over the set G_1 is an element of the set G_2. In mathematical terms, prove that,

$$J^* = J(g_1^*) = \inf_{g_1 \in G_1} J(g_1) = \inf_{g_2 \in G_2} J(g_2) = J(g_2^*), \;\; g_2^* \in G_2.$$

Because $G_2 \subsetneq G_1$, the inequality $\inf_{g_1 \in G_1} \; J(g_1) \leq \inf_{g_2 \in G_2} \; J(g_2)$ always holds. To achieve equality, the converse inequality has to hold. This can be achieved if $J(g_1^*) = \inf_{g_1 \in G_1} \; J(g_1)$ implies that $g_1 \in G_2$. This can be proven if one can solve the optimal control problem. A second more difficult question is to find the smallest subset of control laws $G_2 \subsetneq G_1$ for which the above equality holds.

11.6 Statistical Decision Problems

In this section the reader is introduced to problems and concepts from statistical decision theory. Statistical decision problems may be considered as elementary stochastic control problems in which there is no dynamics but only one time step. These problems have the advantage that the difficulties in the problem formulation are easily pointed out, that the problems are, relatively speaking, easily solved, and that properties of the solution are easily illustrated.

The following two examples illustrate that statistical decision problems are different from deterministic decision problems.

Example 11.6.1. *Portfolio selection* An investor has to select one of two investment opportunities. These opportunities are:

- Investment opportunity A, promising to yield a return of 4.9%.
- Investment opportunity B, promising to yield a return of 40% with probability 0.75 and a return of -100% with probability 0.25.

Suppose that the objective of the investor is to maximize the average return. These average returns are:

- Investment opportunity A 4.9%
- Investment opportunity B $40 \times 0.75 - 100 \times 0.25 = 5\%$

The objective of maximization of the average returns leads to the conclusion to invest exclusively in Opportunity B. Would you accept this advice?

Most investors would not accept this advice! They are scared away by the chance of loosing their money, in the case above this chance has probability 0.25. Apparently the objective of maximization of average returns does not take account of the risk the investor takes.

Example 11.6.2. The St. Petersburg paradox You are invited to participate in a game. The rules are as follows. You pay EUR x to participate in the game. The pay-off is based on repeated coin tossing. If k heads show up before a tail comes up for the first time then you receive EUR $2^k x$. How much money, EUR x, would you want to pay in advance of the game to participate in it?

Note that the probability that k heads show up before a tail does is $2^{-(k+1)}$. The average return if you pay initially EUR x is thus

$$\sum_{k=0}^{\infty}(2^{-(k+1)}2^k)x - x = +\infty.$$

The conclusion is that you should bet as large an amount as possible to play the game. Yet, not many people would like to follow this advice. Again the risk of losing the initial payment is not taken into account in the problem formulation. For a history of this problem see [44].

Below the terminology and the notation of control theory is used rather than the terminology and the notation of the statistical decision literature. This eases the strain on the readers of this book who earlier in the book have become accustomed to the terminology and the notation of control theory.

Definition 11.6.3. *Statistical decision model.* In a statistical decision model a decision maker has to choose a control law which minimizes a cost function. Specifically, the decision maker receives information of the state of the model, chooses a control law from the available set, calculates the value of the control law when applied to the state, calculates the controlled variable and from this the value of the cost criterion.

The problem formulated later on is to determine a control law from the available set which minimizes the cost criterion.

In terms of mathematical notation,

$$\begin{aligned} &x : \Omega \rightarrow X \subseteq \mathbb{R}^{n_x},\ u : \Omega \rightarrow U \subseteq \mathbb{R}^{n_u},\ z : \Omega \rightarrow Z \subseteq \mathbb{R}^{n_x}, \\ &z = h(x,u),\ h : X \times U \rightarrow Z,\ \text{a measureable function.} \end{aligned}$$

Call x the state of the model, u the input of the model, and z the controlled output of the model.

Denote the set of *control laws* by

$$\begin{aligned} G &= \{g : X \rightarrow U \,|\, g \text{ measurable function}\}, \\ u^g &= g(x),\ \text{input as the control law function applied to the state,} \\ z^g &= h(x,u^g) = h(x,g(x)),\ \text{the controlled output as function of the state.} \end{aligned}$$

Though by definition the variable u is a random variable, only through the function g does it become dependent on the state of the model. Similarly, then the controlled output z becomes dependent on the state of the model.

Define the *cost function* of the model by the expression,

$$J(g) = E[b(z^g)] = E[b(h(x, g(x)))], \ \ b : Z \to \mathbb{R}_+, \text{ a measureable function.}$$

In the literature on the statistical decision problem one maximizes the cost function which is in contrast with optimal control theory in which one minimizes the cost function. In this section the literature of the statistical decision literature is followed hence the cost function is maximized.

Problem 11.6.4. The *statistical decision problem*. Consider the model as defined in 11.6.3

Determine a control law $g^* \in G$ such that,

$$J(g^*) = \sup_{g \in G} J(g).$$

This amounts to determine the value $J^* = \sup_{g \in G} J(g)$, to prove existence of an element $g^* \in G$ such that $J^* = \sup_{g \in G} J(g) = J(g^*)$ or, if no such $g^* \in G$ exists, to show that for any $\varepsilon \in (0, \infty)$ there exists a control law $g^*_\varepsilon \in G$ such that
$J^* - \varepsilon < J(G^*_\varepsilon) < J^*$.

Call then $g^* \in G$ the *optimal control law*, $g^*_\varepsilon \in G$ the *ε-optimal control law*, and $J^* \in \mathbb{R}_+$ the *value* of the problem.

Example 11.6.5. Example. Investment opportunity Consider the question which part of an available amount to invest in investment opportunity A and which part in investment opportunity B? The data of the problem are:

- Investment opportunity A: return EUR 1.50 for EUR 1.00 with probability 1.0
- Investment opportunity B: return EUR 1.00 for EUR 1.00 with probability 0.5, and return EUR 3.00 for EUR 1.00 with probability 0.5.

The construction of the model for this problem follows. Denote the input set $U = [0, 1]$ and let $u \in U$ represent the fraction of capital invested in A, hence $(1 - u) \in U$ represents the fraction of capital invested in B. Let $\Omega = \{\omega_1, \omega_2\}$ where ω_1 represents that the return on B is EUR 1.00, and ω_2 represents that the return on B is EUR 3.00. Let F be the σ-algebra of all subsets of Ω. Let $Z = [1, 3]$, $f : U \times \Omega \to Y$

$$f(u, \omega) = \begin{cases} 1.5u + 1 \times (1 - u), & \text{if } \omega = \omega_1, \\ 1.5u + 3 \times (1 - u), & \text{if } \omega = \omega_2. \end{cases} \tag{11.32}$$

Let $\{P_u, \ u \in U\}$ be defined by

$$P_u(\{z = 1.5u + (1 - u)\}) = \frac{1}{2} = P_u(\{ \text{ z = 1.5 u + 3 (1-i)} \}). \tag{11.33}$$

$b : Z \to \mathbb{R}$, $b(u) = cu - u^2$, $c \in \mathbb{R}_+$, $c > 6$. For $c > 6$ the function b is increasing on the interval $[0,3]$. $J : U \to \mathbb{R}_+$, $J(u) = E_u[b(f(u,\omega))]$. The problem is then to solve $\sup_{u \in U} J(u)$. The optimal decision is

$$u^* = \begin{cases} (8-c)/5, & \text{if } 6 < c \leq 8, \\ 0, & \text{if } 8 \leq c. \end{cases}$$

Proof. Note that,

$$\begin{aligned} J(u) &= E[b(f(u,\omega))] = \frac{1}{2}b(1.5u + (1-u)) + \frac{1}{2}b(1.5u + 3(1-u)) \\ &= -\frac{10}{8}u^2 + (4 - 0.5c)u + (2c - 5). \end{aligned}$$

Hence the solution of the unconstrained optimization problem is $u^* = (8-c)/5$. Then $u^* \in [0,1]$ if and only if $c \in [3,8]$. Because by assumption $c \in (6,\infty)$, this solution holds for $c \in (6,8]$. □

Note that for $c \in (6,8)$, $u^* > 0$. Thus even though the average return of investment opportunity A is strictly smaller than that of B, the investor allocates a strictly positive fraction of his capital to investment opportunity A. Hence the risk the investor runs is represented in the problem formulation, here in the utility function U.

Concepts of Statistical Decision Theory

The reader may learn below of the concept of risk, of risk-sensitive criteria, and of certainty equivalence as part of statistical decision theory.

The concept of risk plays an important role in statistical decision problems. K.J. Arrow, [3, 4], has written about risk to which the reader is referred. Risks are basically an evaluation of all effects of a decision which are the responsibility of the decisionmaker. This requires of the decisionmaker a complete specification of the decision and the quantitative description of all financial consequences of the decision. This task is far from simple.

In an investment opportunity the risk is often limited to the loss of the entire investment, for example in a share of a company. In a decision which results in public actions, the financial consequences are often not easy to describe. Most companies enter into a contract with an insurance firm which firm then bears the risk for a financial reward. The insurance firm has hopefully developed expertise on the quantitative modeling of risks.

In this book the discussion of risk is limited to a specification and discussion of a cost function. There follow below the formulation of two concepts, a *risk-sensitive cost function* and the concept of a em certainty-equivalence control law.

Definition 11.6.6. *Risk sensitive statistical decision problems.* Consider Statistical Decision Model 11.6.3 with $Z = \mathbb{R}$ and Problem 11.6.4. In these definitions the expectations are assumed to be finite.

The decisionmaker associated with this problem is called:

- *Risk averse* if $E[b(z)] \leq b(E[z])$.
- *Risk preferring* if $E[b(z)] \geq b(E[z])$.
- *Risk neutral* if $E[b(z)] = b(E[z])$.

An interpretation of these conditions follows. Consider a statistical decision problem in which the pay-off is to be maximized. The interpretation of a risk averse decisionmaker $E[b(z)] \leq b(E[z])$, is that he or she prefers the pay-off $b(E[z])$ associated with the mean value $E[z]$ above the uncertain pay-off $E[b(z)]$. Conversely, the interpretation of a risk preferring decisionmaker is that he or she prefers the uncertain pay-off $E[b(z)]$ above $b(E[z])$. The latter attitude may be associated with that of a dedicated gambler.

Definition 11.6.7. Consider Statistical Decision Model 11.6.3 and Problem 11.6.4. Assume that $Z = \mathbb{R}$,

1. $b : Z \to \mathbb{R}$ is twice-continuously differentiable
2. $b' : Z \to \mathbb{R}$, $b'(y) = db(z)/dz$, is nonzero, or, for all $z \in Z$, $b'(z) \neq 0$.

Define the *index of absolute risk aversion* as the function $r : Z \to \mathbb{R}$

$$r(z) = -b''(y)/b'(z). \tag{11.34}$$

The index of absolute risk aversion measures locally, at $z \in Z$, the risk aversion of the decisionmaker as will be explained below.

An interpretation of the index of absolute risk aversion follows. Consider the Statistical Decision Problem 11.6.4 with $Z = \mathbb{R}$. Assume that the random variable z has finite expectation and finite variance. Denote $\bar{z} = E[z]$, $q = E[(z-\bar{z})^2]$, and define a to be the amount the decisionmaker is willing to pay in order to avoid the decision problem and receive the outcome $\bar{z}$. The following definitions and calculations then lead to the formula for the index of absolute risk aversion.

$$b(\bar{z}-a) = E[b(z)]; \tag{11.35}$$

$$b(\bar{z}-a) = b(\bar{z}) - ab'(\bar{z}) + o(a), \text{by the Taylor expansion,} \tag{11.36}$$

$$o : \mathbb{R} \to \mathbb{R},\ \lim_{s \to 0} o(s)/s = 0,$$

$$\begin{aligned} E[b(z)] &= E[b(\bar{z}) + (z-\bar{z})b'(\bar{z}) + \frac{1}{2}(z-\bar{z})^2 b''(\bar{z}) + o((z-\bar{z})^2)] \\ &= b(\bar{z}) + \frac{1}{2}qb''(\bar{z}) + E[o((z-\bar{z})^2)]. \end{aligned} \tag{11.37}$$

From (11.35,11.36,11.37) then follows that,

$$a = -\frac{1}{2}q\frac{b''(\bar{z})}{b'(\bar{z})} + \frac{o(a)}{b'(\bar{z})} - \frac{E[o((z-\bar{z})^2)]}{b'(\bar{z})} = \frac{1}{2}qr(\bar{z}) + \frac{o(a)}{b'(\bar{z})} - \frac{E[o((z-\bar{z})^2)]}{b'(\bar{z})}.$$

Thus the amount the decisionmaker is willing to pay in order to avoid the decision problem is in its first-order term proportional to $r(\bar{z})$. This is the interpretation of the index of absolute risk aversion.

A realistic model is to take the function $r : Z \to \mathbb{R}$ decreasing. Then the associated decisionmaker will become less risk averse when his capital increases, or he or she will more readily accept risks when their capital increases. This type of behavior has been observed of humans who have to make daily decisions on investment opportunities. For example, if $b(z) = \ln(z)$ then $r(y) = 1/z$.

The reader is referred to the book of D.P. Bertsekas, [9, Subsection 1.3.4, Section 3.3], in which the concept of absolute risk-aversion is illustrated for the problem of allocation between a secure and a risky asset.

Definition 11.6.8. The solution to a statistical decision problem is said to have the *certainty equivalence property* if the solution to this problem is identical to the solution of a corresponding deterministic decision problem. The latter problem is obtained from the first by replacing all random variables by their expectation. The solution of the first problem is then said to be *certainty equivalent* to the solution of the second problem.

Certainty equivalence is a curious property. It will be shown below that the solution to the statistical decision problem for a quadratic utility function and with an outcome with a Gaussian distribution has the certainty equivalence property. The term was introduced specifically for this case. Certainty equivalence of this solution should be seen as a remarkable coincidence in the problem specification that is not likely to occur in general.

Example 11.6.9. Example. A statistical decision problem with a quadratic utility function Consider a statistical decision problem with $U = \mathbb{R}^{n_u}$, $X = \mathbb{R}^{n_x}$, for $u \in U$ $z : \Omega \to Z$ with $z \in G(m_z + u, Q_z)$ in which $m \in \mathbb{R}^{n_u}$, $Q_z \in \mathbb{R}^{n_z \times n_z}_{spd}$. Let $b : Z \to \mathbb{R}_+$ be a quadratic form

$$\begin{aligned} b(z) &= \frac{1}{2}\begin{pmatrix} z \\ u \end{pmatrix}^T L_1 \begin{pmatrix} z \\ u \end{pmatrix}, \; L_1 = \begin{pmatrix} Q & S \\ S^T & R \end{pmatrix} \in \mathbb{R}^{(n_z+n_u)\times(n_z+n_u)}_{sspd}, \\ J(u) &= E[b(z)]. \end{aligned}$$

The problem is then to solve

$$\inf_{u \in U} J(u).$$

Proposition 11.6.10. *Consider the deterministic decision problem associated with Example 11.6.9 that is obtained from Example 11.6.9 by replacing the random variable y by its expectation $m + u$. This problem is to solve*

$$\begin{aligned} & \mathit{inf}_{u \in D} J_1(u), \; \textit{where}, \\ y &= m + u, \\ J_1(u) &= \frac{1}{2}\begin{pmatrix} y \\ u \end{pmatrix}^T L_1 \begin{pmatrix} y \\ u \end{pmatrix} = \frac{1}{2}\begin{pmatrix} m+u \\ u \end{pmatrix}^T L_1 \begin{pmatrix} m+u \\ u \end{pmatrix}, \\ L_1 &= \begin{pmatrix} Q & S \\ S^T & R \end{pmatrix} = L_1^T \in \mathbb{R}^{2p \times 2p}. \; \textit{Denote}, \\ W &= Q - [Q+S][Q+S+S^T+R]^{-1}[Q+S]^T. \end{aligned}$$

(a)If $L_1 > 0$ then $[Q+S+S^T+R] > 0$. If $L_1 \geq 0$ and $[Q+S+S^T+R] > 0$ then $W = W^T \geq 0$.

(b)Assume that $L_1 \geq 0$ and $[Q+S+S^T+R] > 0$. The optimal decision and the optimal value are respectively

$$u^* = -[Q+S+S^T+R]^{-1}[Q+S^T]m, \tag{11.38}$$

$$J_1(u^*) = \frac{1}{2}m^T Wm, \text{ with } W = W^T \geq 0. \tag{11.39}$$

Proof. Note the calculations,

$$T_1 = \begin{pmatrix} I & I \\ 0 & I \end{pmatrix} \in R^{2p \times 2p},$$

$$L_2 = T_1^T L_1 T_1 = \begin{pmatrix} Q & Q+S \\ Q+S^T & Q+S+S^T+R \end{pmatrix}; \tag{11.40}$$

$$0 < [Q+S+S^T+R] \text{ by assumption}, \tag{11.41}$$

$$T_2 = \begin{pmatrix} I & 0 \\ -[Q+S+S^T+R]^{-1}[Q+S^T] & I \end{pmatrix},$$

$$L_3 = T_2^T L_2 T_2 = \begin{pmatrix} W & 0 \\ 0 & Q+S+S^T+R \end{pmatrix}. \tag{11.42}$$

(a) The first claim below follows from (11.40), from the non-singularity of the matrix T_1 and the first formula below. The second claim below follows from (11.42) and the assumption that $L_1 \geq 0$.

$$\begin{pmatrix} Q & Q+S \\ Q+S^T & Q+S+S^T+R \end{pmatrix} = L_2 = T_1^T L_1 T_1,$$

$$\begin{pmatrix} W & 0 \\ 0 & Q+S+S^T+R \end{pmatrix} = L_3 = T_2^T L_2 T_2 = T_2^T T_1^T L_1 T_1 T_2 \geq 0,$$

(b) Note that

$$\begin{aligned} J_1(u) &= \frac{1}{2}\begin{pmatrix} y \\ u \end{pmatrix}^T L_1 \begin{pmatrix} y \\ u \end{pmatrix} = \frac{1}{2}\begin{pmatrix} m+u \\ u \end{pmatrix}^T L_1 \begin{pmatrix} m+u \\ u \end{pmatrix} \\ &= \frac{1}{2}\begin{pmatrix} m \\ u \end{pmatrix}^T L_2 \begin{pmatrix} m \\ u \end{pmatrix}, \text{ by (1)}, \\ &= \frac{1}{2}\begin{pmatrix} m \\ u-u^* \end{pmatrix}^T \begin{pmatrix} W & 0 \\ 0 & Q+S+S^T+R \end{pmatrix} \begin{pmatrix} m \\ u-u^* \end{pmatrix}, \text{ by (2)}. \end{aligned}$$

Because by assumption $[Q+S+S^T+R] > 0$, $u-u^* = 0$ achieves the minimum in $\inf_{u \in U} J_1(u)$. Thus $u = u^*$ is the optimal decision. From (a) follows that $W = W^T \geq 0$.

Proposition 11.6.11. *Consider Example 11.6.9 of the statistical decision problem, different from the deterministic optimization problem of Proposition 11.6.10. Assume that $[Q+S+S^T+R] > 0$ and that $L_1 \geq 0$.*

(a)The optimal solution is

$$u^* = -[Q+S+S^T+R]^{-1}[Q+S^T]m, \tag{11.43}$$

$$J^* = J(u^*) = \frac{1}{2}m^T Wm + \frac{1}{2}tr(QQ_v), \tag{11.44}$$

$$W = Q - [Q+S][Q+S+S^T+R]^{-1}[Q+S^T]. \tag{11.45}$$

(b)The solution to this problem has the certainty equivalence property.

Proof. (a) Note that

$$\begin{aligned} J(u) &= E_u[\frac{1}{2}\begin{pmatrix} y \\ u \end{pmatrix}^T L_1 \begin{pmatrix} y \\ u \end{pmatrix}] \\ &= \frac{1}{2}\begin{pmatrix} m+u \\ u \end{pmatrix}^T L_1 \begin{pmatrix} m+u \\ u \end{pmatrix} + \frac{1}{2}tr(\begin{pmatrix} Q & S \\ S^T & R \end{pmatrix}\begin{pmatrix} Q_v & 0 \\ 0 & 0 \end{pmatrix}), \\ &\quad \text{by Proposition 2.7.6,} \\ &= \frac{1}{2}\begin{pmatrix} m+u \\ u \end{pmatrix}^T L_1 \begin{pmatrix} m+u \\ u \end{pmatrix} + \frac{1}{2}tr(QQ_v). \end{aligned}$$

The result then follows from Proposition 11.6.10.
(b) This follows from (a) and Proposition 11.6.10.

Note that the solutions presented in (11.38) and (11.43) are identical, but that the values presented in (11.39) and (11.44) differ.

Example 11.6.12. Example. Exponential-Quadratic cost Consider the statistical decision problem specified by $U = \mathbb{R}^{n_u}$, $Y = \mathbb{R}^{n_u}$, for $n_u \in \mathbb{Z}_+$ and $u \in U$, $z : \Omega \to Z$ $z \in G(m+u, Q_v)$ in which $m \in \mathbb{R}^{n_u}$, $Q_v \in \mathbb{R}^{n_u \times n_u}$, $Q_v = Q_v^T > 0$. Let $c \in R$. Let $b : Y \to \mathbb{R}$ be the exponential-of-a-quadratic utility function

$$\begin{aligned} b(y) &= c\exp\left(\frac{1}{2}c\begin{pmatrix} y \\ u \end{pmatrix}^T \begin{pmatrix} Q & S \\ S^T & R \end{pmatrix}\begin{pmatrix} y \\ u \end{pmatrix}\right), \\ L_1 &= \begin{pmatrix} Q & S \\ S^T & R \end{pmatrix} = L_1^T; \text{ denote } J : U \to \mathbb{R}; \\ J(u) &= E[b(y)] = E\left[c\exp\left(\frac{1}{2}c\begin{pmatrix} y \\ u \end{pmatrix}^T \begin{pmatrix} Q & S \\ S^T & R \end{pmatrix}\begin{pmatrix} y \\ u \end{pmatrix}\right)\right]. \end{aligned}$$

Consider the problem,

$$\inf_{u \in U} J(u).$$

Note that if $f : \mathbb{R} \to \mathbb{R}$ is given by $f(x) = c\ exp(cx)$, with $c \in \mathbb{R}$, then for both $c > 0$ and $c < 0$ the function f is strictly increasing on $\mathbb{R}$. Hence minimization of f has the same effect as minimization of the variable x for all $c \in R$, $c \neq 0$. If $c = 0$ then $J(u) = 0$.

Proposition 11.6.13. *Consider the statistical decision problem of Example 11.6.12. Assume that*

$$\begin{aligned}
&0 \leq L_1 = L_1^T,\ 0 < [Q_v^{-1} - cQ],\ 0 \leq \frac{1}{c} M \geq 0,\\
&0 < \frac{1}{c}[M_{11} + M_{12} + M_{12}^T + M_{22}];\\
&M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix},\\
&M_{11} = Q_v^{-1}(Q_v^{-1} - cQ)^{-1}V^{-1} - V^{-1}, \quad M_{12} = cQ_v^{-1}(Q_v^{-1} - cQ)^{-1}S,\\
&M_{21} = cS^T(Q_v^{-1} - cQ)^{-1}Q_v^{-1}, \quad M_{22} = cR + c^2 S^T(Q_v^{-1} - cQ)^{-1}S.
\end{aligned}$$

The optimal decision and the optimal value are respectively,

$$u^* = -[M_{11} + M_{12} + M_{12}^T + M_{22}]^{-1}[M_{11} + M_{12}^T]m, \tag{11.46}$$

$$J^* = J(u^*) = c\left(\frac{det((Q_v^{-1} - cQ)^{-1})}{det(Q_v)}\right)^{\frac{1}{2}} \exp(\frac{1}{2}m^T W m), \textit{ where,} \tag{11.47}$$

$$W = M_{11} - [M_{11} + M_{12}][M_{11} + M_{12} + M_{12}^T + M_{22}]^{-1}[M_{11} + M_{12}]^T, \tag{11.48}$$

$$W = W^T \succeq 0. \tag{11.49}$$

Proof.

$$\begin{aligned}
&E_u[b(y)]\\
&= c\left(\frac{det((Q_v^{-1} - cQ)^{-1})}{det(Q_v)}\right)^{\frac{1}{2}} \exp(\frac{1}{2}c\begin{pmatrix} m+u \\ u \end{pmatrix}^T (M/c)\begin{pmatrix} m+u \\ u \end{pmatrix}.
\end{aligned} \tag{11.50}$$

by Proposition 19.4.8; note that the mean is $m+u$. By the remark above Proposition 11.6.13 minimization of (11.50) is equivalent to minimization of

$$\frac{1}{2c}\begin{pmatrix} m+u \\ u \end{pmatrix}^T M \begin{pmatrix} m+u \\ u \end{pmatrix}. \tag{11.51}$$

From Proposition 11.6.10 it follows that the solution is given by

$$u^* = -[M_{11} + M_{12} + M_{12}^T + M_{22}]^{-1}[M_{11} + M_{12}^T]m,$$

with value given in (11.47).

The above example will be extended to a stochastic control problem in the Chapters 12, 14, and 15.

11.7 Exercises

Problem 11.7.1. Consider the state equation of a Gaussian stochastic control system with $x : \Omega \times T \to X$, $T = \mathbb{Z}$, $X = U = \mathbb{R}$,

$$x(t+1) = ax(t) + bu(t) + mv(t),$$

$v(t) \in G(0,w)$. Assume that $b \neq 0$, which condition implies stochastic controllability. Consider the Markov information structure and the stationary Markov control law $g : X \to U$, $g(x) = fx$, with $f \in \mathbb{R}$.

(a) Determine the representation of the closed-loop stochastic control system.
(b) Determine a set $F \subset \mathbb{R}$ such that for any $f \in F$ the closed-loop stochastic control system is exponentially stable.
(c) Assume that $f \in F$. Determine the variance q of the state process if the system is started in the equilibrium distribution.
(d) Solve

$$\inf_{f \in F} J(f),$$

where $J(f) = q$. Thus determine $f^* \in F$ and J^* such that

$$J^* = J(f^*) = \inf_{f \in F} J(f).$$

Problem 11.7.2. Consider the following stochastic team problem. There are two agents, labelled (1) and (2), that are able to control a system. They have the same control objective, namely to minimize the variance of the state process. Assume that the model may be represented by a time-invariant Gaussian stochastic control system. Decompose the state and input into two components that correspond to the two agents, say

$$x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}, \; u(t) = \begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix}.$$

The system representation may then be written as

$$x(t+1) = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} x(t) + \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} u(t) + Mv(t).$$

(a) The two agents are located at different places. Agent 1 observes state vector $x_1(t)$ at time $t \in T$. At that moment he transmits this information also to Agent 2 who receives this information one time step later, thus at $t+1$. Agent 2 observes $x_2(t)$ at time $t \in T$ and transmits this information to Agent 1 who receives this at $t+1$. Specify the information structures of both agents if at any time they are able to recall all information received so far.
(b) Suppose that both agents use a time-invariant linear control law such that at any time this control law uses only the most recently received information. Formulate this control law and derive the representation of the closed-loop stochastic control system.
(c) What do you think is the state process of the closed-loop stochastic control system derived in b? Derive the stochastic system representation in terms of this state vector.

Problem 11.7.3. An investor has to decide which fraction of his capital he should invest in the following two investment opportunities:

- investment in a sure asset with return r_1;
- investment in a risky asset with return r_2 with probability 0.5, or with return r_3 with probability 0.5.

Suppose that $0 < r_2 < r_1 < \frac{1}{2}(r_2 + r_3) < r_3$. The initial capital of the investor is $y_0 \in (0,\infty)$. Let $u \in [0,1]$ denote the fraction of the capital invested in the risky asset. The return on the sure asset is then $r_1(1-u)y_0$. Let y denote the total return. Suppose that the utility function of this investor is $U : (0,\infty) \to \mathbb{R}$, $U(y) = \ln(y)$.

(a) Calculate the index of absolute risk aversion of this investor.
(b) Solve the problem

$$\sup_{u \in [0,1]} E[U(y)].$$

(c) Compute the value of the optimal decision in case $r_1 = 1.05$, $r_2 = 0.80$, and $r_3 = 1.32$.

11.8 Further Reading

History. *Control of stochastic systems* Control theory of stochastic systems has a basis in economic decision theory and statistical decision theory. Classical sources for these subjects are the books of J. von Neumann and O. Morgenstern, [38] for economic decision theory and of [10, 30, 52] for statistical decision theory. A recent book on statistical decision is [40]. In economics, the subject of microeconomics is related to control theory, see [32, 51].

For a history of feedback control see the book of O. Mayr, [34].

Books on stochastic control problems. Books on stochastic control at the level of this book are [5, 14, 26, 29, 33]. Books that discuss stochastic control problems from an economics viewpoint are [1, 12, 22, 48]. General books on *continuous-time stochastic control problems* are [7, 8, 15, 25, 27, 28]. Stochastic control problems of economics are described in [2].

Motivation. The example of course keeping of a ship may be found in [50].

Control laws. The presentation of control laws and closed-loop stochastic control systems is based on the book [26]. A standard form for sequential stochastic control is presented in [54]. In that paper a detailed formulation is provided for the analysis of the sequential stochastic control problem.

Closed-loop stochastic control systems. Example 11.3.7 is adjusted from [26, Section 2.4].

Statistical decision theory. Books that present statistical decision theory are, the older books, [10, 11, 30, 38, 52], and the more recent books, [9, 32]. See the paper by C.M. Harvey for modeling of statistical decision problems with quite small risks, [17]. The concept of absolute risk aversion has been introduced by J.W. Pratt, [39].

Examples of particular stochastic control problems. [6, 21, 41]. *Residence time control*. [23, 35, 37, 36, 42]. *Continuous-time optimal stopping problems* are treated in [7, Ch. VII] and [20, Ap. D]. [49]. *Mathematical physics approach to modeling of stochastic control systems*. [24].

Numerical methods for stochastic control. [13, 18, 43].

References

1. M. Aoki. *State space modeling of time series*. Springer-Verlag, Berlin, 1987. 120, 410
2. V.I. Arkin and I.V. Evstigneev. *Stochastic models of control and economic dynamics*. Academic Press, New York, 1987. 410
3. K.J. Arrow. *Aspects of the theory of risk bearing*. Yrjo Jahnsson Lecture Series. U. of Helsinki, Helsinki, Finland, 1965. 403, 468
4. K.J. Arrow. The theory of risk aversion. In *Essays in the theory of risk-bearing*. North-Holland, Amsterdam, 1971. 403
5. K.J. Aström. *Introduction to stochastic control*. Academic Press, New York, 1970. 376, 410, 467, 522, 575, 596
6. V.E. Benes. Full "bang" to reduce predicted miss is optimal. *SIAM J. Control & Opt.*, 14:62–84, 1976. 411, 575
7. A. Bensoussan. *Stochastic control by functional analysis methods*. North-Holland Publ. Co., Amsterdam, 1982. 410, 411
8. A. Bensoussan and J.L. Lions. *Applications of variational inequalities in stochastic control*. North-Holland, Amsterdam, 1982. 410
9. D.P. Bertsekas. *Dynamic programming and stochastic control*. Academic Press, New York, 1976. 376, 405, 410, 439, 468, 502, 525, 526, 575, 595
10. D. Blackwell and M.A. Girshick. *Theory of games and statistical decisions*. Wiley, New York, 1954. 353, 410, 467
11. D. Blackwell and M.A. Girshick. *Theory of games and statistical decisions*. Dover Publications Inc., New York, 1979. 410
12. G. Chow. *Econometric analysis by control methods*. John Wiley, New York, 1981. 120, 410
13. P. Colleter, F. Delebecque, F. Falgarone, and J.P. Quadrat. Application of stochastic control methods to the management of energy production in new caledonia. In A. Bensoussan, P. Kleindorfer, and C.S. Tapiero, editors, *Applied stochastic control in econometrics and management science*, pages 203–232. North-Holland Publ. Co., Amsterdam, 1980. 411
14. M.H.A. Davis and R.B. Vinter. *Stochastic modelling and control*. Chapman and Hall, London, 1985. 120, 376, 410, 468, 575, 595
15. W.H. Fleming and R.W. Rishel. *Deterministic and stochastic optimal control*. Springer-Verlag, Berlin, 1975. 410
16. L.J. Forys. Performance analysis of a new overload strategy. In *10th International Teletraffic Congres*, 1983. 78, 379
17. C.M. Harvey. Preference functions for catastrophe and risk inequity. *Large Scale Systems*, 8:131–146, 1985. 410
18. R.H.W. Hoppe. Multi-grid methods for hamilton-jacobi-bellman equations. *Numerische Mathematik*, 49:239–254, 1986. 411
19. S. Stidham Jr. Optimal control of admission to a queueing system. *IEEE Trans. Automatic Control*, 30:705–713, 1985. 78, 379
20. I. Karatzas and S.E. Shreve. *Methods of mathematical finance*. Number 39 in Applications of Mathematics. Springer, Berlin, 1998. 411, 605, 742
21. Ioannes Karatzas. Probabilistic aspects of finite-fuel stochastic control. *Proc. Natl. Acad. Sci. USA*, 82:5579–5581, 1985. 411

22. D. Kendrick. *Stochastic control for economic models*. McGraw-Hill Book Co., New York, 1981. 120, 410, 575
23. S. Kim, S.M. Meerkov, and T. Runolfsson. Residence probability control. *Computers Math. Applic.*, 19:121–125, 1990. 411
24. Peter Kosmol and Michele Pavon. Lagrange approach to the optimal control of diffusions. *Acta Appl. Math.*, 32:101–122, 1993. 411
25. N.V. Krylov. *Controlled diffusion processes*. Springer-Verlag, Berlin, 1980. 410
26. P.R. Kumar and P. Varaiya. *Stochastic systems: Estimation, identification, and adaptive control*. Prentice Hall Inc., Englewood Cliffs, NJ, 1986. 376, 410, 468, 525, 575, 595, 596
27. H.J. Kushner. *Stochastic stability and control*. Academic Press, New York, 1967. 121, 376, 410, 467
28. H.J. Kushner. *Introduction to stochastic control*. Holt, Rinehart and Winston Inc., New York, 1971. 121, 376, 410, 467, 525
29. H. Kwakernaak and R. Sivan. *Linear optimal control systems*. Wiley-Interscience, New York, 1972. 120, 376, 410, 467, 489, 593, 822, 823
30. R.D. Luce and H. Raiffa. *Games and decisions*. John Wiley & Sons, New York, 1957. 376, 410, 467
31. J. Marschak and R. Radner. *Economic theory of teams*. Yale University Press, New Haven, 1972. 384
32. A. Mas-Colell, M.D. Whinston, and J.R. Green. *Microeconomic theory*. Oxford University Press, Oxford, 1995. 410
33. P.S. Maybeck. *Stochastic models, estimation and control: Volume 1,2 and 3*. Academic Press, New York, 1979. 120, 376, 410
34. O. Mayr. *The origins of feedback control*. MIT Press, Cambridge, MA, 1970. 410
35. S.M. Meerkov and T. Runolfsson. Residence time control. *IEEE Trans. Automatic Control*, 33:323–332, 1988. 411
36. S.M. Meerkov and T. Runolfsson. Theory of residence-time control by output feedback. *Dynamics and Control*, 1:63–81, 1991. 411
37. S.M. Merkov and T. Runolfsson. Output residence time control. *IEEE Trans. Automatic Control*, 34:1171–1176, 1989. 411
38. J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton University Press, Princeton, NJ, 1947. 410, 467
39. J.W. Pratt. Risk aversion in the small and in the large. *Econometrica*, 32:1964, 1964. 410
40. J.W. Pratt, H. Raiffa, and R. Schlaifer. *Introduction to statistical decision theory*. MIT Press, Cambridge, MA, 2008. 319, 353, 410
41. R. Rishel. Controlled wear process: Modeling-optimal control. In *Proceedings 28th Conference on Decision and Control*, pages 724–726, New York, 1989. IEEE Press. 411
42. T. Runolfsson. Stationary risk-sensitive LQG control and its relation to LQG and H-infinity control. In *Proceedings 29th IEEE Conference on Decision and Control*, pages 1018–1023, New York, 1990. IEEE Press. 411, 526
43. N. Saldi, S. Yüksel, and T. Linder. Finite model approximations and asymptotic optimality of quantized policies in decentralized stochastic control. *IEEE Trans. Automatic Control*, 62:2360–2373, 2017. 411
44. P.A. Samuelson. St. Petersburg paradoxes: Defanged, dissected, and historically described. *J. Economic Literature*, 7:24–55, 1969. 401
45. F.C. Schoute. Optimal control and call acceptance in a SPC exchange. In *9th International Teletraffic Congres*, 1981. 78, 379
46. F.C. Schoute. Adaptive overload control of an SPC exchange. In *10th International Teletraffic Congres*, 1983. 78, 379
47. F.C. Schoute. Overload control in spc processors. *Philips Telecommunication Review*, 41:300–310, 1983. 78, 379
48. J.K. Sengupta. *Stochastic optimization and economic models*. D. Reidel Publ. Co., Dordrecht, 1986. 410
49. L.A. Shepp. Explicit solutions to some problems of optimal stopping. *Ann. Math. Statist.*, 40:993–1010, 1969. 411

50. J. van Amerongen. Adaptive steering of ships - A model reference approach. *Automatica*, 20:3–14, 1984. 410, 468
51. H.R. Varian. *Intermediate microeconomics - A modern approach*. W.W. Norton & Conpany, New York, 1987. 410
52. A. Wald. *Statistical decision functions*. John Wiley, New York, 1950. 315, 353, 410
53. Hans S. Witsenhausen. Separation of estimation and control for discrete time systems. *Proc. IEEE*, 59:1557–1566, 1971. 384, 575
54. H.S. Witsenhausen. A standard form for sequential stochastic control. *Math. Systems Theory*, 7:5–11, 1973. 410, 575

Chapter 12
Stochastic Control with Complete Observations on a Finite Horizon

Abstract Optimal stochastic control problems are formulated for a stochastic control system with complete observations on a finite horizon. Dynamic programming yields necessary and sufficient conditions for optimality rather than local optimality conditions as provided by methods based on the calculus of variations or on the maximum principle. Sufficient conditions are formulated for a subset of value functions to be invariant with respect to the dynamic programming operator. Reduction in complexity of a stochastic control system with the controlled output signal is proven using dynamic programming. Examples include: the linear-quadratic-Gaussian optimal control problem, a gambling problem with an exponential value function, and a finite stochastic control system.

Key words: Optimal stochastic control. Dynamic programming.

Readers who first learn about this control procedure may want to focus attention on the Sections 12.2, 12.3, 12.6, and 12.7. Subsequently it will be useful if they read the Sections 12.9 and 12.10.

In this chapter uses is made of concepts and results of the theory of optimization. The reader finds a summary of those concepts and results in Section 17.7.

The reader may be interesting in noticing that the stochastic control system in closed-loop with the optimal control law is a stochastic realization of the stochastic control system with a finite-dimensional or finite state set satisfying the optimality criterion.

12.1 Control Problems

Example 12.1.1. *Course keeping of a ship.* Automatic steering of ships was introduced early in the 20th century. With the development of technology throughout the 20th century, autopilots have changed from mechanical devices via electronic systems to computers on chips.

The problem of designing an autopilot for a ship has over the years continuously received attention. A ship operates in a changing environment. Environmental factors that affect a ship are water depth, water currents, and wind. The ships conditions as loading and speed also affect the dynamics. Therefore the autopilot should be able to adjust to the actual conditions. An adaptive control approach to the design of an autopilot is therefore quite appropriate. In this example only a stochastic control problem is formulated.

Two steering modes are distinguished: course keeping and course changing. Control problems for both modes may be considered. Below attention is restricted to course keeping.

Modeling A simple model of the dynamics of a ship is a mechanical model described by the differential equation,

$$\tau_1 \frac{d^2\phi(t)}{dt^2} + \frac{d\phi(t)}{dt} = K_1 u(t), \tag{12.1}$$

where $u : T \to \mathbb{R}$, $u(t)$ denotes the angle of the rudder with respect to a reference direction, $\phi : T \to \mathbb{R}$ denotes the ship's heading with respect to the same reference direction, and $K_1, \tau_1 \in \mathbb{R}$. The constants K_1, τ_1 depend on the ship's speed and length.

The rudder is actuated by means of a hydraulic machine which has nonlinear dynamics. The rudder angle and the rudder speed are limited in magnitude. Compared with the limitation of the rudder speed, other time constants of the steering machine may be disregarded for controller design.

Disturbances that affect ship steering are caused by wind, waves, and water current. The effect of the wind and the waves on the ship may be modelled by stochastic processes. Practitioners suggest as a model of a fully developed sea, a stationary Gaussian process with a particular covariance function or spectral density. The effect of the waves on the ship depends also on the angle between the direction of the waves and that of the ship's heading, and on the speed of the ship. In the following these effects are modelled simply by a disturbance process.

The model may be converted to the standard state-space form as follows. Then,

$$\begin{aligned}
x(t) &= \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix},\ x_1(t) = \phi(t),\ x_2(t) = \frac{d\phi(t)}{dt}, \\
\dot{x}_1(t) &= \dot{\phi}(t) = x_2(t),\ \ \dot{x}_2(t) = \ddot{\phi}(t) = \frac{-1}{\tau_1}\dot{\phi}(t) + \frac{K_1}{\tau_1}u(t); \\
\dot{x}(t) &= Ax(t) + Bu(t),\ x(0) = x_0, \\
A &= \begin{pmatrix} 0 & 1 \\ 0 & a_{22} \end{pmatrix},\ \ B = \begin{pmatrix} 0 \\ b_2 \end{pmatrix},\ \ a_{22} = -\frac{1}{\tau_1},\ b_2 = \frac{K_1}{\tau_1}.
\end{aligned}$$

The disturbance for a continuous-time stochastic system is modelled by a Brownian motion process resulting in the stochastic differential equation,

$$dx(t) = Ax(t)dt + Bu(t)dt + Mdv(t), \tag{12.2}$$

where $v : \Omega \times T \to \mathbb{R}^2$ is a Brownian motion process that represents the effect of the wind and wave forces on the ship's heading, and $M \in \mathbb{R}^{2\times 2}$.

Control The control problem is to synthesize a control law such that the stochastic control system combined with the control law satisfies the control objectives. Control objectives for course keeping are: (1) to minimize course deviations; (2) to minimize the costs, mainly fuel costs. The weights attached to these objectives depend on the shipping environment. In confined waters with dense traffic the closed-loop control system has to follow the course very accurately. On a full sea, Control Objective 2, the cost minimization, dominates the objectives. Because the course deviation and the control effort are dynamically coupled, it seems appropriate to consider a quadratic cost criterion, say

$$E[\frac{1}{t_1}\int_0^{t_1}(e(t)^2 + cu(t)^2)dt], \tag{12.3}$$

where ϕ_0 denotes the heading direction of the control objective, $e(t) = \phi(t) - \phi_0$ denotes the heading error, $u(t)$ denotes the rudder angle, and $c \in (0,\infty)$ is a constant of the cost function.

The following state feedback controller for this problem has been proposed,

$$u(t) = K_p e(t) - K_d\dot{\phi}(t) + K_i(t) = K_p(x_1(t) - \phi_0) - K_d x_2(t) + K_i(t), \tag{12.4}$$

where $K_i : T \to R$ has been added to correct for the slowly varying moment of the wind. It is suggested that for the above mentioned criterion the parameters of the controller should be chosen as

$$K_p = \lambda^{-\frac{1}{2}}, K_d = \frac{1}{K_1}[\left(1 + 2\frac{K_1\tau_1}{\sqrt{\lambda}}\right)^{\frac{1}{2}} - 1], K_i(t) = \frac{1}{t_2}\int_{t-t_2}^{t} u(s)ds, \tag{12.5}$$

where $t_2 \in (0, t_1)$, such that K_i compensates for the average deviation of the rudder angle.

The frequencies of the ship's motions caused by waves are so high that it makes no sense to compensate for them by rudder movements. The approach suggested is to apply a noise reduction filter to the measurement of the heading and to apply the controller (12.4) not to the state but to the filtered estimate of the state. Thus,

$$u(t) = K_p[\hat{x}_1(t) - \phi_0] - K_d\hat{x}_2(t) + K_i(t),$$
$$\text{where } \hat{x} = (\hat{x}_1, \hat{x}_2) \text{ represents the filtered estimate of the state.}$$

Such a controller was shown to operate reasonably well and to perform better than an autopilot from a preceding generation and an experienced professional.

Another example is the control of a shock absorber, Example 1.3.1.

12.2 Problem Formulation

Problem 12.2.1. The *optimal stochastic control problem.* Define the *finite horizons*,

$$T = T(0:t_1) = \{0,1,\ldots,t_1\},\ \ T(0:t_1-1) = \{0,1,\ldots,t_1-1\},\ \ t_1 \in \mathbb{Z}_+.$$

Consider a recursive state-observed stochastic control system,

$$x(t+1) = f(t,x(t),u(t),v(t)),\ x(0) = x_0, \tag{12.6}$$

$$X \subseteq \mathbb{R}^{n_x},\ U \subseteq \mathbb{R}^{n_u}, \tag{12.7}$$

the past-state information structure, and the set of past-state control laws G with Borel measurable functions. For any $g \in G$ with $g = \{g_0, g_1, g_2, \ldots\}$ define the closed-loop stochastic control system and the input by,

$$\begin{aligned} x^g(t+1) &= f(t,x^g(t),g_t(x^g(0:t)),v(t)),\ x^g(0) = x_0, \\ x^g(0:t) &= (x^g(0),x^g(1),\ldots,x^g(t)), \\ u^g(t) &= g_t(x^g(0:t)) = g_t(x^g(0),\ldots,x^g(t)). \end{aligned}$$

Consider the positive cost function,

$$\begin{aligned} &b : T_1 \times X \times U \to \mathbb{R}_+,\ b_1 : X \to \mathbb{R}_+,\ \text{Borel measurable functions}, \\ J(g) &= E\left[\sum_{s=0}^{t_1-1} b(s,x^g(s),u^g(s)) + b_1(x^g(t_1))\right],\ J : G \to \mathbb{R}_+. \end{aligned}$$

Assume that for any $g \in G$,

$$\begin{aligned} &E|b_1(x^g(t_1))| < \infty, \\ &E|b(t,x^g(t),u^g(t))| < \infty,\ \ \forall t = 0,\ldots,t_1-1. \end{aligned}$$

If the assumption does not hold then restrict attention to a subset of control laws $G_1 \subseteq G$ for which it holds. If $G_1 = \emptyset$, then for every control law at least one of the above expectations is infinite, then comparison of control laws of G is not possible because the cost of each control law is infinite hence there is no distinction between the costs of control laws.

The optimal stochastic control problem is then to solve,

$$\inf_{g \in G} J(g).$$

This problem should be understood as: (1) to determine the *value* $J^* \in \mathbb{R}$ which always exists because, $\forall\ g \in G$, $J(g) \geq 0$, and (2) to determine an *optimal control law* $g^* \in G$, if it exists, such that,

$$J^* = \inf_{g \in G} J(g) = J(g^*).$$

In general $g^* \in G$ is not unique. The problem also asks for a procedure to determine an optimal control law.

If $g_s \in G$ is a particular control law then call the *regret* of that control law the variable $J(g_s) - J^*$; which equals the extra cost of that control law above the infimal cost. This term is due to Tze Leung Lai, [47].

It may be the case that there does not exist a $g^* \in G$ such that $J^* = J(g^*)$. In that case the problem asks for the determination of, for any $\varepsilon \in (0,\infty)$, an ε*-optimal control law* $g^*_\varepsilon \in G$, which satisfies,

$$J^* < J(g^*_\varepsilon) < J^* + \varepsilon.$$

In the formulation of Problem 12.2.1 attention is restricted to stochastic control systems as defined in Equation (12.6) or Definition 10.1.2 with the past-state information structure. This is called the *complete observations case.* Optimal stochastic control problems with the past-output information structure, called *stochastic control problem with partial observations*, receive attention in Chapter 14.

In addition, attention is first given to a cost function with an additive cost structure followed by those with a multiplicative cost structure. For an approach that includes general cost functions see Chapter 16.

A discussion of the conditions of the problem follows. In the literature one distinguishes the following cases for the state set X and the input set U:

- the *finite case*: the finite state set X and the finite input set U;
- the *countable case*: the countable state set X and the countable input set U, where for example $X = \mathbb{N} = \{0,1,2,\ldots\}$ is a countable set;
- the *continuous case*: the state set X and the input set U are intervals or noncountable subsets of tupes of the real numbers; a special case is when both X and U are Borel spaces.

A measurable space is a *Borel space*, Def. 2.4.8, if it is bijectively related (isomorphic) to a measurable subset of a complete separable metric space. Borel spaces were defined by D. Blackwell, [13, 15]. In this book the focus is on the third case described above though occasionally examples are used which fit in the first and the second case.

In the general case, it will be assumed that, for all $(t,x) \in \times X$, there exists a compact and convex subset $U(t,x) \subseteq U$. This formulation allows the set of inputs to depend on the time and state which is useful for examples. On the cost rate b an assumption of continuity and strict convexity with respect to $u \in U$ may be imposed.

As stated above, in particular case an optimal control law does not exist in the set of admissable control laws. In that case one may investigate the existence of an $\varepsilon \in (0,\infty)$ optimal control law as described for optimization problem in Section 17.7.

Are their guarantees on the performance of the optimal control law? Define the *zero control law* as the function $g_{uz} \in G$ such that $g_{uz}(.) = 0$ for all its arguments. It is assumed that such a control law exists in the set G of admissable control laws. Can one guarantee that the optimal control law $g^* \in G$, if its exists, satisfies $J(g^*) < J(g_{uz})$, or even by a margin $J_{marg} \in (0,\infty)$ such that $J(g^*) + J_{marg} < J(g_{uz})$?

From the view point of control theory, one expects that the existence of an optimal control law and the improvement of performance has the necessary and sufficient condition of stochastic controllability of the stochastic control system. Such a condition is not used much in the literature. The dynamic procedure, described below in this chapter, to solve an optimal stochastic control problem could be applied even if the system is not stochastically controllable. Stochastic controllability is required if one wants to guarantee that the value of the cost function strictly goes down compared to the case without the use of a control law. Therefore in this section, the assumption of stochastic controllability and of stochastic observability will be formulated and illustrated. Chapter 13 provides more explicit results in this regard. The system theoretic interpretation of dynamic programming using stochastic controllability is discussed in Section 12.10.

The reader is provided in this chapter the approach of dynamic programming to solve an optimal stochastic control problem.

12.3 Explanation of Dynamic Programming

In this section the dynamic programming method is informally introduced.

In this chapter the dynamic programming method is used to solve optimal stochastic control problems. This method has the advantage that it yields global conditions for optimality. This is in contrast with the variational approach to stochastic control that yields only local optimality conditions.

Dynamic programming is a solution method for optimal stochastic control problems. It applies to problems of deterministic and of stochastic control. Most readers have probably used dynamic programming even though they were not aware of it. For example, the reader uses dynamic programming if he/she wants to determine the departure time of a trip if he/she wants to arrive at the destination on or before a specified time. Dynamic programming is best introduced with an example.

Example 12.3.1. *Introduction to dynamic programming.* Consider the network of a deterministic finite-state control system as displayed in Figure 12.1. The time set is $T = \{0,1,2\}$. The state space has only three elements $X = \{1,2,3\}$. The possible state transitions are indicated by arrows between the states. The input determines which transition is taken if there is more than one transition. The cost rate and the terminal cost are described in the Table 12.1. The expression $b(t,x,u)$ denotes the cost rate at time $t \in T$, at state $x \in X$, and with input value $u \in U$. The values of the cost rate are indicated in Fig. 12.1 near the corresponding arrows. The input value u denotes the state to which the system will move from the tuple $(t,x) \in T \times X$.

The problem is to compute the control law that specifies how to go from a state x_0 at $t = 0$ to a state at $t = 2$ with minimal cost.

The dynamic programming procedure will be explained next. The dynamic programming procedure computes:

1. the minimal cost-to-go $V(t,x)$ from any tuple (t,x) at a time $t \in T$ and from state $x \in X$ to a tuple (t_1,x_1) at the terminal time t_1 and to the reachable state x_1;
2. the optimal control law $g^*(t,x)$ which specifies at any time and state (t,x) which input to choose so as to achieve the lowest cost.

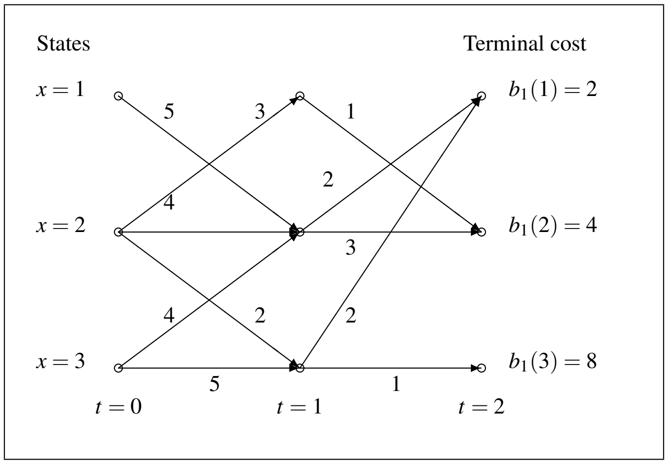


Fig. 12.1 Network of a finite deterministic control system.

(t,x,u)	x^+	$b(t,x,u)$	(t,x,u)	x^+	$b(t,x,u)$
(0,1,2)	2	5	(1,1,2)	2	1
(0,2,1)	1	3	(1,2,1)	1	2
(0,2,2)	2	4	(1,2,2)	2	3
(0,2,3)	2	2	(1,3,1)	1	2
(0,3,2)	2	4	(1,3,3)	3	1
(0,3,3)	3	5			
x		$b_1(x)$			
1		2			
2		4			
3		8			

Table 12.1 Table of cost rate and of terminal cost for example.

Denote the *minimal cost-to-go* for the problem from state $(t,x) \in T \times X$ to the terminal time and a state $(t_1,x_1) \in T \times X$ by $V(t,x) \in \mathbb{R}$. Then, at the terminal time, by definition of the terminal cost,

$$V(2,1) = b_1(1) = 2,\ V(2,2) = b_1(2) = 4,\ V(2,3) = b_1(3) = 8.$$

Compute next the value function at time $t = 1$ with the formula,

$$V(1,x) = \min_{u \in U(1,x)} \{b(1,x,u) + V(2,f(1,x,u))\}.$$

Compute the minimal cost from $(t,x) = (1,1)$ to the terminal time as,

$$\begin{aligned} V(1,1) &= \min_{u \in U(1)=\{2\}} (b(1,1,u) + V(2,f(t,x,u)) = \\ &= b(1,1,2) + V(2,2) = 1 + 4 = 5, \end{aligned}$$

which is the sum of the transition cost $b(1,1,2)$ and of the terminal cost $V(2,2)$. In this case there was no optimization because only one transition is possible in the system. Similarly,

$$\begin{aligned}
V(1,2) &= \min_{u \in U(1,2)=\{1,2\}} (b(1,2,u)+V(2,f(1,2,u))) = \min\{2+2,3+4\} = 4, \\
V(1,3) &= \min_{u \in U(1,3)=\{2,3\}} (b(1,3,u)+V(2,f(1,3,u))) = \min\{2+2,1+8\} = 4.
\end{aligned}$$

The optimal control law at time $t = 1$ is thus determined by the arguments of the minimizations as, where $g^*(1,x)$ denotes the state to which the system then moves,

$$g^*(1,x) = \begin{cases} 2, & \text{if } x = 1, \\ 1, & \text{if } x = 2, \\ 1, & \text{if } x = 3. \end{cases}$$

Next the value function is determined at time $t = 0$ by the recursion,

$$V(0,x) = \min_{u \in U(0,x)} \{b(0,x,u)+V(1,f(t,x,u))\}.$$

The computations are,

$$\begin{aligned}
V(0,1) &= \min_{u \in U(0,1)=\{2\}} (b(0,1,u)+V(1,f(0,1,u)) = 5+4 = 9, \\
V(0,2) &= \min_{u \in U(0,2)=\{1,2,3\}} (b(0,2,u)+V(1,f(0,2,u)) \\
&= \min\{3+5,4+4,2+4\} = 6, \\
V(0,3) &= \min_{u \in U(0,3)=\{2,3\}} (b(0,3,u)+V(1,f(0,3,u)) = \min\{4+4,5+4\} = 8.
\end{aligned}$$

The optimal control law at time $t = 0$ is thus,

$$g^*(0,x) = \begin{cases} 2, & \text{if } x = 1, \\ 3, & \text{if } x = 2, \\ 2, & \text{if } x = 3, \end{cases}$$

Note that the dynamic programming procedure computes the minimal cost $V(t,x)$ for every state $x \in X$ and every time $t \in T$, not just for the particular initial state $x_0 \in X$.

The dynamic programming procedure consists thus of a backward recursion and (1) at each step one computes the value function $V(t,.)$ for the time considered, and (2) at each step one determines the optimal control law $g^*(t,.) : X \to U$.

12.4 Digression on Optimization

The formulation of the dynamic programming procedure requires a discussion of integration and of optimization.

In the dynamic programming procedure there appears an expression of conditional expectation. It is explained how to calculate this conditional expectation. Consider Problem 12.2.1. Let $X = \mathbb{R}^{n_x}$ and $U \subseteq \mathbb{R}^{n_u}$. Below, the variable x_V denotes an argument of the function V while $x^g(t)$ denotes the state of the closed-loop system as defined in Problem 12.2.1.

Dynamic programming uses a backward recursion for the minimal cost-to-go. Let $V : T \times X \to \mathbb{R}$ and let $V(t, x_V)$ represent the minimal cost on the future time horizon $\{t, t+1, \ldots, t_1\}$ if the control system starts in state $x^g(t) = x_V$ at time $t \in T$. The following relation may then be conjectured to hold,

$$V(t, x_V) \leq b(t, x_V, u_V) + E[V(t+1, f(t, x_V, u_V, v(t)))|F^{x_V, u_V}], \tag{12.8}$$

for all $u_V \in U(x_V)$, while for an optimal control law g^*,

$$\begin{aligned} V(t, x^{g^*}(t)) = {} & b(t, x^{g^*}(t), g^*(x^{g^*}(t))) + \\ & + E[V(t+1, f(t, x^{g^*}(t), u^{g^*}(t), v(t)))|F^{x^{g^*}, u^{g^*}}]. \end{aligned} \tag{12.9}$$

The interpretation of Equation (12.8) is that the infimal cost on the horizon $\{t, t+1, \ldots, t_1\}$ is less than or equal to the sum of the cost rate $b(t, x(t), u(t))$ at time $t \in T$ and of the infimal cost on the horizon $\{t+1, t+2, \ldots, t_1\}$ projected on the σ-algebra F^{x_V, u_V}. The interpretation of Equation (12.9) is that for an optimal control law one achieves equality in (12.8). The Equations (12.8) and (12.9) together describe what is called the *principle of optimality*. This principle has been proposed and popularised by R.E. Bellman. The principle may actually be deduced from other properties.

In discrete time and with discrete spaces, the value function can be computed numerically. In discrete time and with continuous state space and input space, one has to calculate the analytic form of the value function backward recursively in time. In continuous time and with continuous spaces one has to solve a partial differential equation for the value function. In practice one computes a numerical approximation of the value function and of the control law.

Consider Problem 12.2.1. For any past-state control law g, any measurable function $V : T \times X \to \mathbb{R}$, and any time $t \in T$, consider the conditional expectation,

$$E[V(t+1, f(t, x^g(t), g_t(x^g(0:t), v(t)))|F_t^{x^g, u^g}]. \tag{12.10}$$

It follows from Theorem 2.8.6 that, because for all $t \in T$, $v(t)$ is independent of $F_t^{x^g}$; moreover, the current state $x^g(t)$ and the past states $x^g(0), \ldots, x^g(t-1)$ are measurable with respect to $F_t^{x^g}$. Hence the calculation of the conditional expectation (12.10) may be performed by integrating over the probability distribution function $f_{v(t)}$ of $v(t)$, while treating $x^g(0), \ldots, x^g(t)$ and u as known values, according to,

$$\int V(t+1, f(t, x^g(t), u, v)) f_{v(t)}(dv). \tag{12.11}$$

This calculation may also be performed if $x^g(t)$ and $g_t(x^g(0), \ldots, x^g(t))$ are replaced by the indeterminates x_V and u_V respectively, in which case (12.11) is denoted by,

$$E[V(t+1, f(t, x_V, u_V, v(t)))|F^{x_V, u_V}],$$

where the conditioning on F^{x_V, u_V} should be understood as in equation (12.10). The latter displayed expression will be used in this chapter.

Next optimization is discussed. The dynamic programming procedure requires for all time $t \in T$ the solution of an infimization problem,

$$\inf_{u \in U(t,x)} h(t,x,u), \text{ where } h : T \times X \times U \rightarrow \mathbb{R}_+.$$

Issues to be investigated include: the existence of a minimal element $u^* \in U(t,x)$, the uniqueness of a minimal element, and a procedure to determine a minimal element. Next a sufficient condition is stated after which conditions are determined which imply the sufficient conditions. The reader is expected to know several concepts of analysis and convex functions, see partly Chapter 17. Needed is also the concept of a compact subset of the vector space $\mathbb{R}^{n_u}$ for an integer $n_u \in \mathbb{Z}_+$. A closed and bounded subset of $\mathbb{R}^{n_u}$ is a compact set.

Theorem 12.4.1. *Consider a finite time index set $T = \{0,1,2,\ldots,t_1\}$ for $t_1 \in \mathbb{Z}_+$, and subsets of tuples of the real numbers, $X \subseteq \mathbb{R}^{n_x}$ and $U \subseteq \mathbb{R}^{n_u}$. Consider a measurable function, $h : T \times X \times U \rightarrow \mathbb{R}_+$. Assume that for all $(t,x) \in T \times X$, there exists a subset $U(t,x) \subseteq U$.*

(a)If (1) $\forall\ (t,x) \in T \times X$ the subset $U(t,x) \subseteq U$ is a compact set; (2) $\forall\ (t,x) \in T \times X$ the function $h(t,x,.) : U(t,x) \rightarrow \mathbb{R}_+$ is a continuous function; then,

$$\forall\, x \in X,\ \exists\, u^* \in U(t,x),\ \forall\, u \in U(t,x),\ h(t,x,u^*) \leq h(t,x,u).$$

In general a minimizer u^* is not unique.

(b)If (1) $\forall\ (t,x) \in T \times X$ the subset $U(t,x) \subseteq U \subseteq \mathbb{R}^{n_u}$ is a compact and convex set; (2) $\forall\ (t,x) \in T \times X$ the function $h(t,x,.) : U(t,x) \rightarrow \mathbb{R}_+$ is a continuous function; and (3) $\forall\ (t,x) \in T \times X$ the function $h(t,x,.) : U(t,x) \rightarrow \mathbb{R}_+$ is strictly convex; then,

$$\forall\, x \in X,\ \exists\, u^* \in U(t,x),\ \forall\, u \in U(t,x),\ h(t,x,u^*) \leq h(t,x,u).$$

In general, a minimizer u^* is unique.

(c)If (1) $\forall\ (t,x) \in T \times X$ the subset $U(t,x) \subseteq U \subseteq \mathbb{R}^{n_u}$ is a convex set; (2) $\forall\ (t,x) \in T \times X$ the function $h(t,x,.) : U(t,x) \rightarrow \mathbb{R}_+$ is a continuous function; and (3) $\forall\ (t,x) \in T \times X$ the function $h(t,x,.) : U(t,x) \rightarrow \mathbb{R}_+$ is strictly convex; (4) $\forall\ (t,x) \in T \times X$, the following limit holds true $\lim_{\|u\| \rightarrow \infty} h(t,x,u) = +\infty$; then,

$$\forall\, x \in X,\ \exists\, u^* \in U(t,x),\ \forall\, u \in U(t,x),\ h(t,x,u^*) \leq h(t,x,u).$$

Proof. (a) This follows from [19, Thm. 3.15, Cor. 6.57].
(b) The existence of $u^* \in U(t,x)$ follows from (a) while the uniqueness follows from the facts that $U(t,x)$ is a convex set and that $h(t,x,.)$ is strictly convex. Existence of two different minimizers directly contradicts the strict convexity assumption.
(c) Because of condition (4), there exists a $u_0 \in U(t,x)$ such that $h(t,x,u_0) \in \mathbb{R}_+$. It is best not to take as u_0 the minimizer of the function. Consider the sublevel set of $h(t,x,u_0)$, $U_s(t,x) = \{u \in U(t,x) |\ h(t,x,u) \leq h(t,x,u_0)\}$. This set is not empty because $u_0 \in U_s(t,x)$, it is closed, and it is bounded because of condition (4). Hence $U_s(t,x)$ is a compact subset of $U(t,x)$. By condition (1), the subset $U(t,x)$ is convex, and from the definition of $U_s(t,x)$ then follows that it is also a convex set. By condition (3), the function $h(t,x,.)$ is strict convex on $U(t,x)$ hence strictly convex

on $U_s(t,x)$. It then follows from (b) that there exists a unique $u^* \in U(t,x)$ achieving the minimal value. □

Example 12.4.2. Consider the function for $T = T(0:t_1)$, $X = \mathbb{R}$, and $U(t,x) = U = (0,\infty)$, $H(t,x,u) = 1/u$. Then for all $(t,x) \in T \times X$, $U(t,x)$ is a convex set, $h(t,x,.) : U \rightarrow \mathbb{R}_+$ is a continuous function on U, and $h(t,x,.) : U \rightarrow \mathbb{R}_+$ is a strictly convex function, due to $dH(t,x,u)/du = -1/u^2 < 0$ and $d^2H(t,x,u)/du^2 = 2/u^3 > 0$ on the set U. Yet, condition (4) of Theorem 12.4.1.(c) does not hold. It is directly clear from the function that there does not exist an element $u^* \in U(t,x)$ which achieves the infimum which is zero.

Another function on the entire real line is $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = \exp(-x)$. This function is strictly convex, $\inf_{x\in\mathbb{R}} f(x) = 0$, but no minimizer exists on the real line.

The conclusion is that there exists a convex set on which is defined a continuous and strictly convex function which nevertheless does not admit a minimal value on the domain of the function. This example shows that condition (4) is necessary in Theorem 12.4.1.(c).

If attention is restricted to the set $U(t,x) = (0,1000]$ which is a compact set then it follows from part (b) of the above theorem that there exists an unique minimum $u^* = 1000 \in U(t,x)$ for all $(t,x) \in T \times X$.

Of the above sufficient conditions, the continuity cannot be deleted due to difficulties with calculating a minimal value. The convexity is preferably also not dropped due to again difficulties in determining a minimal value. Convexity of the function requires convexity of the subset $U(t,x) \subseteq U$.

Next sufficient conditions for Theorem 12.4.1 are investigated.

Lemma 12.4.3. *(a)Consider the sets and functions,*

$$\begin{aligned}
T &= \{0,1,2,\ldots,t_1\},\ t_1 \in \mathbb{Z}_+,\ X \subseteq \mathbb{R}^{n_x},\ U \subseteq \mathbb{R}^{n_u},\ W \subseteq \mathbb{R}^{n_w}, \\
&\quad V : T \times X \rightarrow \mathbb{R}_+,\ \forall\, t \in T,\ V(t,.) : X \rightarrow \mathbb{R}_+ \text{ is continuous,} \\
&\quad f_s : T \times X \times U \times W \rightarrow X, \\
&\quad \forall\, t \in T,\ f_s(t,.,.,.) : X \times U \times W \rightarrow X \text{ is continuous,} \\
&\quad p_v : W \rightarrow \mathbb{R}_+ \text{ is continuous},\ \int_W p_w(w_p)dw_p = 1; \\
&\quad \forall\, (t,x,u) \in T \times X \times U,\ \int_W V(f_s(t,x,u,w_p))\ p_w(w_p)dw_p < \infty; \\
&\quad h_V(t,x,u) = \int_W V(f_s(t,x,u,w_p))p_w(w_p)dw_p,\ h_V : T \times X \times U \rightarrow \mathbb{R}_+.
\end{aligned}$$

Then, for all $t \in T$, the function $h(t,.,.) : X \times U \rightarrow \mathbb{R}_+$ is a continuous function.

(b)If in addition, (1) for all $(t,x) \in T \times X$, $U(t,x) \subseteq U$ is a convex set, (2) for all $(t,x) \in T \times X$, $f_s(t,x,.)$ is an affine function of the variable u, and (3) $V(t,.) : X \rightarrow \mathbb{R}_+$ is a convex function; then, for all $t \in T$, the function $h_V(t,x,.) : U(t,x) \rightarrow \mathbb{R}_+$ is a convex function.

Proof. (a) By assumption, for all (t,x,u), the integral defining $h(t,x,u)$ is finite or real valued. Because f_s and V are continuous functions, so is the map

$(x,u,w_p) \mapsto V(t, f_s(t,x,u,w_p))$ for all $t \in T$. It follows from the theory of integration of continuous functions over a product space and from the continuity of p_v that the integral $h(t,x,u)$ is a continuous function of (x,u) for all $t \in T$.
(b) It follows from [17, p.79, Subsec. 3.2.2], from V being a convex function, and from f_s being an affine function of u, that the function, for all $t \in T$, $(x,u,w_p) \mapsto V(f_s(t,x,u,w_p))$ is a convex function. It follows from the previous conclusion and from [17, p.79] that, for all $t \in T$, the function $(x,u) \mapsto h(t,x,u)$ is a convex function. □

The convexity of a composition of two convex functions is discussed in [17, p. 84, Subsec. 3.2.4]. That the integral over a measurable function of two variables integrated over one of the variables, is again a measurable function is proven in [19, Th. 10.5.1].

Lemma 12.4.4. *Consider the sets and functions,*

$$\begin{aligned}
T &= \{0,1,2,\ldots,t_1\},\ t_1 \in \mathbb{Z}_+,\ X \subseteq \mathbb{R}^{n_x},\ U \in \mathbb{R}^{n_u},\ W \in \mathbb{R}^{n_w},\\
&\quad V : T \times X \to \mathbb{R}_+,\ \forall\, t \in T,\ V(t,.) : X \to \mathbb{R}_+ \textit{ is continuous,}\\
&\quad f_s : T \times X \times U \times W \to X,\\
&\quad \forall\, t \in T,\ f_s(t,.,.,.) : X \times U \times W \to X \textit{ is continuous,}\\
&\quad p_w : W \to \mathbb{R}_+ \textit{ is continuous},\\
&\quad \forall\, (t,x,u) \in T \times X \times U,\ \int_V V(t, f_s(t,x,u,w_p))\ p_w(w_p) dw_p < \infty;\\
&\quad h_V(t,x,u) = \int_V V(t, f_s(t,x,u,w_p)) p_w(w_p) dw_p,\\
&\quad h_V : T \times X \times U \to \mathbb{R}_+,\ \forall\, (t,x) \in T \times X,\ h_V(t,x,.) : U(t,x) \to \mathbb{R}_+\\
&\quad h_V(t,x,.) : U \to \mathbb{R}_+ \textit{ assumed to be strictly convex,}\\
&\quad b : T \times X \times U \to \mathbb{R}_+,\ \forall\, (t,x) \in T \times X,\ b(t,x,.) : U(t,x) \to \mathbb{R}_+,\\
&\quad b(t,x,.) \textit{ assumed to be continuous and strictly convex,}\\
h(t,x,u) &= b(t,x,u) + E[V(t, f_s(t,x,u,w)) |\, F^{x,u}]\\
&= b(t,x,u) + h_V(t,x,u),\ h : T \times X \times U \to \mathbb{R}_+.
\end{aligned}$$

Then for all $(t,x) \in T \times X$, the function $h(t,x,.) : U(t,x) \to \mathbb{R}_+$ is continuous and strictly convex.

Proof. From [17, p. 84] follows that the sum of two convex functions defined on the same domain, is a convex function. The assumptions on V and f_s being continuous imply that h_V is continuous. The sum of two continuous functions on the same domain is a continuous function. □

How to calculate the minimum of a function?

Procedure 12.4.5 *Consider a function $h : U \to \mathbb{R}_+$ for a convex subset $U \subseteq \mathbb{R}^{n_u}$. Assume that h is twice continuously differentiable.*

1. *Calculate the following second derivative. If the indicated relation holds then the function is strictly convex.*

$$\frac{\partial^2 h(u)}{\partial u^2} \succ 0, \; \forall \, u \in U.$$

2. *Calculate the first derivative and solve the next equation for $u^* \in U$. From strict convexity follows that $u^* \in U$ is the unique minimizer.*

$$0 = \frac{\partial h(u)}{\partial u}|_{u=u^*},$$

Example 12.4.6. Consider the quadratic function,

$$X = \mathbb{R}^{n_x}, \; U = \mathbb{R}^{n_u}, \; h(x,u) = \begin{pmatrix} x \\ u \end{pmatrix}^T Q \begin{pmatrix} x \\ u \end{pmatrix}, \;\; h : X \times U \to \mathbb{R}_+,$$
$$Q = \begin{pmatrix} Q_{xx} & Q_{xu} \\ Q_{xu}^T & Q_{uu} \end{pmatrix} \in \mathbb{R}^{(n_x+n_u)\times(n_x+n_u)}, \; 0 \prec Q_{uu} \in \mathbb{R}^{n_u \times n_u}.$$

Note that,

$$\lim_{\|u\|\to\infty} h(x,u) = +\infty, \text{ because } 0 \prec Q_{uu};$$
$$\frac{\partial h(x,u)}{\partial u} = 2u^T Q_{uu} + 2x^T Q_{xu} = 0, \;\; \frac{\partial^2 h(x,u)}{\partial u^2} = 2Q_{uu} \succ 0;$$
$$\Rightarrow \; 0 = Q_{uu}u^* + Q_{xu}^T x \; \Rightarrow \; u^* = -Q_{uu}^{-1} Q_{xu}^T x.$$

Thus the function h is for all $x \in X$ a continuous and strictly convex function, and the minimizer is u^* as specified above.

12.5 Digression on Measurable Control Laws

The reader has to be informed about optimal control laws which are required to be measurable. In the dynamic programming procedure stated in the next section, a condition is imposed that a candidate optimal control law and the value function are actually measurable functions. This condition is necessary and the explanation is mathematically quite technical. A reader interested in the formulation of dynamic programming may skip the section initially and return to it later after having learned about dynamic programming.

In the dynamic programming equation, the conditions are imposed that, for all times $t \in T$, the functions $g_t^* : X \to U$ and $V(t,.) : X \to \mathbb{R}$ are Borel or Lebesgue measurable functions. This is a simple sufficient condition which is satisfactory for this book.

The theoretical basis for this condition requires an understanding of mathematics which is beyond the scope of this book. The research issue was formulated and investigated first by D. Blackwell and later by S.E. Shreve and his research advisor

D.P. Bertsekas. The reader is referred to their book, [12, Ch. 1], for an introduction to the problem.

A more detailed explanation follows. In the dynamic programming procedure one constructs, by a backward recursion, the value function V and the optimal control law g^*. The steps of the procedure are well defined only if for all times $t \in T$, the functions $V(t,.) : X \to \mathbb{R}$ and $g_t^* : X \to U$ are measurable functions. If the state set X and the input set U are countable sets, then measureability is always satisfied. But when X and U are not countable, then measureability may not hold. The first issue for this is that the orthogonal projection of a Borel set in $\mathbb{R}^2$ on an axis need not be a Borel measurable subset of $\mathbb{R}^1$. A second issue is to prove that for any real number $\varepsilon \in (0,\infty) \subset \mathbb{R}$, there exists a Borel measurable control law $g_\varepsilon^* \in G$. In the literature one terms the construction of a measurable value function and that of a measurable control law, a *measurable selection*, [66].

Three approaches have been developed for sufficient conditions for these measureability issues. Approach (1) of C. Striebel is a general model in which the value stochastic process is well defined. However, it leaves open how to construct an ε-optimal control law. Approach (2) based on analysis imposes sufficient conditions in the form of semi-continuity, compactness, and convexity of sets and of functions. Approach (3) restricts attention to a universal-measurable cost rate and a universal-measurable control laws. This approach is due to D. Blackwell, [14], further developed by R.E. Strauch, A.A. Yushkevic (also written as A.A. Juskevič), K. Hinderer, [61, 72, 33], and in particular by S.E. Shreve. The reader is referred for references and theory to the book of S.E. Shreve and D.P. Bertsekas, [12].

For relatively simple optimal control problems the control law is a linear, affine, polynomial, exponential, or logarithmic function which all satisfy the sufficient condition of measurability. For the general case, the reader has to investigate whether the value function and the calculated optimal control law satisfy the measurability condition.

12.6 Dynamic Programming for Additive Cost Functions

The reader finds in this section the solution procedure of dynamic programming for an optimal stochastic control problem with an additive cost function. The case of a multiplicative cost function requires a different treatment which is provided in Section 12.11.

Definition 12.6.1. Consider Problem 12.2.1. Define the *conditional cost-to-go* as the function,

$$\begin{aligned}
& J : G \times \Omega \times T \rightarrow \mathbb{R}, \\
J(g,t) &= E[\sum_{s=t}^{t_1-1} b(s,x^g(s),u^g(s)) + b_1(x^g(t_1))|F_t^{x^g}] \qquad (12.12)\\
&= E[\sum_{s=t}^{t_1-1} b(s,x^g(s),g(t,x^g(0:t))) + b_1(x^g(t_1))|F_t^{x^g}];\ \text{where,}\\
x^g(t+1) &= f(t,x^g(t),g(t,x^g(0:t)),v(t)),\ x^g(0) = x_0,\\
u^g(t) &= g(t,x^g(0:t)),\ \ J(g) = E[J(g,0)].
\end{aligned}$$

Note that, at time $t \in T$,

$$\sum_{s=t}^{t_1-1} b(s,x^g(s),u^g(s)) + b_1(x^g(t_1))$$

is the cost to be incurred on the horizon $\{t,t+1,\ldots,t_1\}$ when control law g is used. Here $J(g,t)$ depends on g only through its values at the times $t,t+1,\ldots,t_1-1$, thus on $g(t,.),g(t+1,.),\ldots,g(t_1-1,.)$.

The reader best distinguishes two approaches to optimal stochastic control used in the literature:

1. *Value function approach*. *Define* the *value function* as the infimal cost-to-go over the remaining horizon at any time.

$$\begin{aligned}
& V : T \times X \rightarrow \mathbb{R}_+,\ \forall\, x_V \in X,\ x^g(t) = x_V,\\
V(t,x_V) &= \mathrm{P} - essinf_{(g(t,.),\ g(t+1,.),\ \ldots,\ g(t_1-1,.))\in G(t:t_1-1)}\\
& \quad E[\sum_{s=t}^{t_1-1} b(s,x^g(s),u^g(s)) + b_1(x^g(t_1))|F^{x_V}].\\
G(t:t_1-1) &= \{(g(t,.),\ldots,g(t_1-1,.))\}.
\end{aligned}$$

   This requires the concept of $\mathrm{P} - essinf$ because in general the set of control laws $G(t:t_1-1)$ is neither finite nor countable, and because the cost-to-go is a random variable due to the conditioning of the cost-to-go on the current state. Then one proves that this value function satisfies the dynamic programming equation. See Chapter 16 for this approach.
2. *Dynamic programming recursion approach*. Define the value function by the dynamic program recursion stated below. Then prove that this function equals the infimal cost-to-go at any time as expressed in item (1).

In this chapter the second approach is used.

The reader has to distinguish (1) the value function V evaluated at an argument x_V as in $V(t,x_V)$, (2) from the evaluation of that function at a trajectory of the closed-loop system, as in $V(t,x^g(t))$. The input space is allowed to depend per time on the current state of the system, hence the input space is denoted by $U(t,x_V)$ for
$t \in T(0:t_1-1)$ and for $x_V \in X$. This generality is useful for particular examples.

Procedure 12.6.2 *The* dynamic programming procedure (for additive cost) *(DP-procedure). Consider Problem 12.2.1.*

1. *Initialization. Let* $V(t_1,.) : X \to \mathbb{R}_+$, $V(t_1,x_V) = b_1(x_V)$.
2. *For* $t = t_1 - 1, t_1 - 2, \ldots, 0$ *do: Determine,* $V(t,.) : X \to \mathbb{R}_+$,

$$\begin{aligned} & V(t,x_V) && (12.13) \\ & = \inf_{u_V \in U(t,x_V)} \{b(t,x_V,u_V) + E[V(t+1, f(t,x_V,u_V,v(t)))|F^{x_V,u_V}]\}. \end{aligned}$$

$$\begin{aligned} h(t,x,u_V) &= b(t,x_V,u_V) + E[V(t+1, f(t,x_V,u_V,v(t)))|F^{x_V,u_V}], \quad (12.14) \\ & \forall\ (t,x_V) \in T \times X, h(t,x_V,.) : U \to \mathbb{R}_+. \end{aligned}$$

This includes: (1) for all $(x_V,u_V) \in X \times U$*, the calculation of the conditional expectation used in (12.13); and (2) for all* $(x_V,u_V) \in X \times U$*, the solution of the infimization problem. The reader should note that the infimization of the sum equals the infimization of (1) the instantaneous cost* $b(t,x_V,u_V)$ *and (2) the conditional expectation of the cost-to-go, from time* $t+1$ *to the end of the horizon, conditioned on the current state* x_V *and the current input* u_V *considered. The optimization strikes a balance between both of these terms.*

3. *If for all* $x_V \in X$ *the infimum in (12.13) is attained, or, equivalently, if for all* $x_V \in X$ *there exists a* $u^* \in U(t,x_V)$ *such that the following equality holds, then define,*

$$\begin{aligned} & b(t,x_V,u^*) + E[V(t+1, f(t,x_V,u^*,v(t)))|F^{x_V}] && (12.15) \\ & = \inf_{u \in U(t,x_V)} \{b(t,x_V,u_V) + E[V(t+1, f(t,x_V,u_V,v(t)))|F^{x_V,u_V}]\}, \end{aligned}$$

$$g^*(t,x_V) = u^*, \ \ g^*(t,.) : X \to U(t,x_V),$$

where (t,x_V) *and* u^* *are related by (12.15). Note that by definition of the control law* $g^*(t,x_V)$*, it depends only on the current state* x_V *and not on the past states* $x^g(0:t-1)$*, hence* $g^* \in G_M$ *is a Markov control law.*

If there does not exist an input value in the input space which achieves the infimum then, for any $\varepsilon \in (0,\infty)$*, determine an input* u^*_ε *which achieves a value which is within an* $\varepsilon \in (0,\infty)$*-approximation of the infimal value, and similarly* $g_\varepsilon(t,x_V)$*; Proposition 17.7.7.*

4. *Determine whether for all* $t \in T$*, the function* $V(t,.) : X \to \mathbb{R}_+$ *and the function* $g^*(t,.) : X \to U$ *are Borel measurable functions. Stop if the condition is not satisfied, see for an explanation Section 12.5.*
5. *If the conditions of Steps (2)–(4) are met then output the value function* V *and the optimal control law* g^**. The control law* $g^* \in G$ *is optimal in the set of nonlinear Borel measurable control laws.*

The word *procedure* in the dynamic programming procedure is used to avoid misunderstanding by computer scientists for whom the word *algorithm* has a well defined meaning in the theory of computation. Because the procedure involves the determination of an infimum over the possibly continuous input set U, the procedure is in general not an algorithm in the sense of the theory of computation.

The dynamic programming procedure prescribes a stepwise backward recursion, with, at each step, to perform an infimization of a function over the input space U. This is a significant reduction in computational complexity with respect to the optimal stochastic control problem which demands an infimization over the set of control laws G.

The uniqueness of the input vector u^* is often obtained from a strict convexity assumption on the function to be infinimized; see for the conditions, Section 12.4.

Condition (4) of the Dynamic Programming Procedure is necessary because there exists an example where the condition is not satisfied. For a more-satisfactory measure theoretic formulation of dynamic programming see Chapter 16 and the references [12, 62].

Definition 12.6.3. Define the time-varying *dynamic programming operator of additive cost* of Procedure 12.6.2 by the formula,

$$\begin{aligned} DP(t,V)(x_V) &= \inf_{u_V \in U(t,x_V)} \{b(t,x_V,u_V) + E[V(t,f(t,x_V,u_V,v(t)))|F^{x_V,u_V}]\}, \\ &DP : T \times F(T \times X,\ \mathbb{R}_+) \to \mathbb{R}_+. \end{aligned}$$

The expression in curly brackets is a measurable function. It will be assume that the expression $DP(.,.)$ is a measurable function. See the discussion in Section 12.5.

Theorem 12.6.4. Sufficient and necessary conditions for optimality of an optimal control law. *Consider Problem 12.2.1.*

(a)A condition on the value function. *Let $V : T \times X \to \mathbb{R}_+$ be produced by the dynamic programming procedure. Then,*

$$\begin{aligned} &\forall\, g \in G, t \in T, \\ V(t,x^g(t)) &\leq J(g,t) = E\left[\sum_{s=t}^{t_1-1} b(s,x^g(s),u^g(s)) + b_1(x^g(t_1))|F_t^{x^g}\right],\ a.s. \\ E[V(0,x_0)] &\leq E[J(g,0)] = J(g)\ \Rightarrow\ E[V(0,x_0)] \leq J^* = \inf_{g \in G} J(g). \end{aligned}$$

Note that in the above first inequality the value function is evaluated at $x^g(t)$ and not at all values $x_V \in X$.

Note also that the inequality $V(t,x^g(t)) \leq J(g,t)$ states that for any control law $g \in G$, the value function evaluated at the current state $x^g(t)$ is lower than then conditional cost to go at the same time for the same state. Because this holds for any time $t \in T$, the value $V(t,x^g(t))$ is a lower bound for the future cost over the remaining horizon $\{t+1,t+2,\ldots,t_1\}$.

*(b)*Sufficient conditions for existence of a minimizer. *If the conditions of Theorem 12.4.1(b) or (c) hold for the function h defined in equation (12.14) of Procedure 12.6.2 then, for all $(t,x) \in T \times X$, there exists a unique minimizer $u^* \in U(t,x)$ such that,*

$$b(t,x_V,u^*) + E[V(t+1,f(t,x_V,u^*,v(t)))|F^{x_V}] \tag{12.16}$$

$$= \inf_{u \in U(t,x_V)} \{b(t,x_V,u_V) + E[V(t+1,f(t,x_V,u_V,v(t)))|F^{x_V,u_V}]\}. \tag{12.17}$$

Weaker conditions can be imposed to guarantee the existence of a minimal element without the uniqueness.

(c)A sufficient condition for optimality. *If the infima in Equation (12.13) are attained or, equivalently, if, for all $(t,x) \in T \times X$, there exists a $u^*(t,x_V) \in U(t,x_V)$ such that (12.15) holds, then*
$g^* : T \times X \to U$ produced by Procedure 12.6.2 is well defined, and for all $t \in T$,

$$\begin{aligned} &V(t,x^{g^*}(t)) = J(g^*,t), \\ &E[V(0,x_0)] = E[J(g^*,0)] = J(g^*) = J^*; \end{aligned}$$

hence $g^* \in G$ is an optimal control law and a Markov control law.
Note that by the problem formulation any control law $g \in G$ can for any $t \in T$ depend on all past states, $g(t,x^g(0),\ldots,x^g(t))$, but that the optimal control law $g^*(t,x^g(t))$ depends only on the current state $x^g(t)$.

(d)A necessary condition for optimality. *Assume that there exists an optimal Markov control law $g^* : T \times X \to U$, $g^* \in G$. Let $V : T \times X \to \mathbb{R}_+$ be produced by the dynamic programming Procedure 12.6.2. Assume that the conditions of Theorem 12.4.1.(c) hold. Denote the closed-loop system of the optimal control law by*

$$x^{g^*}(t+1) = f(t,x^{g^*}(t),g^*(t,x^{g^*}(t),v(t)),\ x^{g^*}(0) = x_0.$$

Denote for all $t \in T$ the support set of $x^{g^*}(t)$ by $X_{supp}(x^{g^*}(t)) \subseteq X$ which, by definition of a support set, is an open subset of $X = \mathbb{R}^{n_x}$.
Then the following equality holds in the dynamic programming procedure,

$$\begin{aligned} &\forall\, t \in T\backslash\{t_1\},\ \ \forall\, x_s \in X_{supp}(x^{g^*}(t)), \\ &b(t,x_s,g^*(t,x_s)) + E[V(t+1,f(t,x_s,g^*(t,x_s),v(t))|\ F^{x_s,g^*(t,x_s)}] \\ = &\inf_{u_V \in U(t,x_s)} b(t,x_s,u_V) + E[V(t+1,f(t,x_s,u_V,v(t))|\ F^{x_s,u_V}]. \end{aligned}$$

In words, the infimum in Equation (12.15) is attained for all $t \in T$, for all $x_s \in X_{supp}(x^{g^*}(t))$ at $u_V = g^*(t,x^{g^*}(t))$. Note that these equations hold for the trajectory x^{g^*} corresponding to the optimal control law.

The proof of Theorem 12.6.4 is based on the following intermediary results.

Lemma 12.6.5. *The* comparison principle. *Consider Problem 12.2.1. Let the function V be such that,*

$$\begin{aligned} &V : T \times X \to \mathbb{R}_+, \\ V(t_1,x_V) &\leq b_1(x_V), \forall x_V \in X, and, \\ V(t,x_V) &\leq b(t,x_V,u_V) + E[V(t+1,f(t,x_V,u_V,v(t)))|F_t^{x_V,u_V}], \qquad (12.18) \\ &\forall\, x_V \in X,\ u_V \in U(t,x_v),\ t = 0,1,\ldots,t_1-1.\ Then, \\ V(t,x^g(t)) &\leq J(g,t)\ a.s. \forall\, t \in T,\ \forall\, g \in G. \qquad (12.19) \end{aligned}$$

The interpretation of this result is that $V(t,x^g(t))$ is a lower bound of the conditional cost-to-go $J(g,t)$ for any control law $g \in G$.

Proof. Let $g \in G$. Recall the definition of the closed-loop system of Problem 12.2.1 with the state process x^g and the input process u^g. Note that,

$$\begin{aligned}
V(t_1, x^g(t_1)) &\leq b_1(x^g(t_1)), \text{ by assumption,} \\
&= E[b_1(x^g(t_1))|F_{t_1}^{x^g}], \text{ by Theorem 2.8.2.(c),} \\
&= J(g, t_1), \text{ by definition of } J(g, .).
\end{aligned}$$

Suppose that Equation (12.19) holds for $s = t+1,\ t+2,\ \ldots,\ t_1$. Then it is proven for $s = t$,

$$\begin{aligned}
&V(t, x^g(t)) \\
&\leq b(t, x^g(t), u^g(t)) + E[V(t+1, f(t, x^g(t), u^g(t), v(t)))|F_t^{x^g}], \\
&\quad \text{by Equation (12.18),} \\
&= E[b(t, x^g(t), u^g(t)) + V(t+1, x^g(t+1))|F_t^{x^g}], \text{ by Theorem 2.8.2.(c),} \\
&\leq E[b(t, x^g(t), u^g(t)) + J(g, t+1)|F_t^{x^g}], \text{ by the induction hypothesis (12.19)} \\
&\quad \text{and by monotonicity of conditional expectation,} \\
&= E[b(t, x^g(t), u^g(t)) + E[\sum_{s=t+1}^{t_1-1} b(s, x^g(s), u^g(s)) + b_1(x^g(t_1))|F_{t+1}^{x^g}]|F_t^{x^g}], \\
&\quad \text{by definition of } J(g, t+1), \\
&= E[\sum_{s=t}^{t_1-1} b(s, x^g(s), u^g(s)) + b_1(x^g(t_1))|F_t^{x^g}], \text{ by Theorem 2.8.2.(d),} \\
&= J(g, t), \text{ by definition of } J(g, t).
\end{aligned}$$

The result follows by the induction theorem. □

Lemma 12.6.6. The value function associated with a Markov control law. *Let $g \in G_M$ be a Markov control law. Define $V^g : T \times X \to \mathbb{R}_+$ by,*

$$V^g(t_1, x_V) = b_1(x_V), \tag{12.20}$$

$$V^g(t, x_V) = b(t, x_V, g(t, x_V)) + E[V^g(t+1, f(t, x_V, g(t, x_V), v(t)))|F^{x_V}]. \tag{12.21}$$

Then the following equalities hold,

$$\forall\, t \in T,\ V^g(t, x^g(t)) = J(g, t),\ a.s. \tag{12.22}$$

Proof. Let $g \in G_M$. From Proposition 11.3.3 follows that the state process x^g of the closed-loop stochastic control system is a Markov process. Then,

$$\begin{aligned}
V^g(t_1, x^g(t_1)) &= b_1(x^g(t_1)), \text{ by definition of } V^g(t_1, .), \\
&= E[b_1(x^g(t_1))|F_{t_1}^{x^g}], \text{ by Proposition 2.8.2.(c),} \\
&= J(g, t_1), \text{ by Def. 12.6.1 of } J(g, .), \\
&= E[b_1(x^g(t_1))|F^{x^g(t_1)}],
\end{aligned}$$

because x^g is a Markov process. Suppose that (12.22) holds for $s = t+1,\ t+2,\ \ldots,\ t_1$. Then it is proven for $s = t$,

$$\begin{aligned}
&V^g(t,x^g(t))\\
&= b(t,x^g(t),g(t,x^g(t))) + E[V^g(t+1,f(t,x^g(t),g(t,x^g(t)),v(t)))|F^{x^g(t)}]\\
&\quad \text{by definition of } V^g(t,x^g(t)),\\
&= b(t,x^g(t),g(t,x^g(t))) + E[V^g(t+1,x^g(t+1))|F_t^{x^g}]\\
&\quad \text{because } g \in G_M \text{ is a Markov control law hence } x^g \text{ is a Markov process},\\
&= b(t,x^g(t),g(t,x^g(t))) +\\
&\quad +E[E[\sum_{s=t+1}^{t_1-1} b(s,x^g(s),g(s,x^g(s))) + b_1(x^g(t_1))|F_{t+1}^{x^g}]|F_t^{x^g}],\\
&\quad \text{by the induction hypothesis},\\
&= E[\sum_{s=t}^{t_1-1} b(s,x^g(s),g(s,x^g(s))) + b_1(x^g(t_1))|F_t^{x^g}],\\
&\quad \text{by } F_t^{x^g} \subseteq F_{t+1}^{x^g} \text{ and by Theorem 2.8.2.(d)},\\
&= J(g,t), \text{ by Definition 12.6.1 },\\
&= E[\sum_{s=t}^{t_1-1} b(s,x^g(s),g(s,x^g(s))) + b_1(x^g(t_1))|F^{x^g(t)}],
\end{aligned}$$

because x^g is a Markov process. □

Proof. Proof of Theorem 12.6.4.
(a) Let $V : T \times X \to \mathbb{R}_+$ be constructed by the dynamic programming procedure. Then V satisfies the conditions of Lemma 12.6.5. From Lemma 12.6.5 then follows that for any $g \in G$,

$$\begin{aligned}
V(t,x^g(t)) &\leq J(g,t), \text{ a.s.} \forall\, t \in T;\\
E[V(0,x_0)] &\leq E[J(g,0)] = J(g), \text{ by Def. 12.6.1;}\\
E[V(0,x_0)] &\leq \inf_{g\in G} J(g) = J^*, \text{ since } g \in G \text{ was arbitrary.} \qquad (12.23)
\end{aligned}$$

(b) The claims follow directly from the indicated theorem.
(c) Let $g^* \in G$ be a control law determined by the dynamic programming procedure. Hence it achieves the infimum in Equation (12.15). Note that then V satisfies the conditions for V^{g^*} presented in the Equations (12.20) and (12.21). From Lemma 12.6.6 then follows that,

$$\begin{aligned}
V(t,x^{g^*}(t)) &= J(g^*,t) \text{ a.s., } \forall t \in T. \text{ Then,} \qquad (12.24)\\
J(g^*) &= E[J(g^*,0)], \text{ by Equation (12.12),}\\
&= E[V(0,x_0)], \text{ by (12.24),}\\
&\leq J^*, \text{ by (12.23).}\\
&\leq J(g^*), \text{ because } g^* \in G \text{ and by definition of } J^*, \qquad (12.25)\\
J(g^*) &= J^* = \inf_{g\in G} J(g),
\end{aligned}$$

hence that $g^* \in G$ is an optimal control law.
(d) Assume that the control law $g^* \in G_M$ is optimal. It will be proven by contradiction that, for all $t \in T$, $g^*(t,x^{g^*}(t))$ achieves the infimum in Eqn. (12.15).

Recall the definition of the support set $X_{supp}(x^{g^*}(t))$. Recall also the definition of the function

$$h(t,x_h,u_h) = b(t,x_h,u_h) + E[V(t,f(t,x_h,u_h,v(t))|F^{x_h,u_h}].$$

Suppose that the infimum of Eqn. (12.15) is not attained for time $t_a = t_1 - 1$ at the value $u_h = g^*(t_a,x^{g^*}(t_a))$. Then there exists a state $x_s \in X_{supp}(x^{g^*}(t_a))$ such that

$$\begin{aligned}
& h(t_a,x_s,g^*(t_a,x_s)) \\
&= b(t_a,x_s,g^*(t_a,x_s)) + E[V(t_a+1,f(t_a,x_s,g^*(t_a,x_s),v(t_a))|F^{x_s}] \\
&> \inf_{u_V \in U(t_a,x_s)} \{\, b(t_a,x_s,u_V) + E[V(t_a+1,f(t_a,x_s,u_V,v(t_a))|F^{x_s}]\} \qquad (12.26) \\
&= h^*(t_a,x_s), \quad \text{where } t_a = t_1 - 1 \ \Rightarrow\ t_a + 1 = t_1,\ V(t_a+1,x) = b_1(x).
\end{aligned}$$

Because $X_{supp}(x^{g^*}(t_a)) \subset X = \mathbb{R}^{n_x}$ is an open set there exists a Borel measurable open subset $X_s \subseteq X_{supp}(x^{g^*}(t_a))$ such that (1) $P_X(X_s) > 0$ and (2) for all $x_s \in X_s$, Eqn. (12.26) holds. Here $P_X : B(X) \to \mathbb{R}_+$ is the probability measure induced on $(X,B(X))$ by the random variable $x^{g^*}(t_a) : \Omega \to X$.

By assumption on the function h, for all $(t_a,x) \in T \times X$, $U(t_a,x_s)$ is a convex set, and the function $h(t_a,x,.) : U(t_a,x) \to \mathbb{R}_+$ is continuous, strictly convex, and satisfies condition (4) of Theorem 12.4.1(c). From that theorem follows that there exists a minimizer $u^*(t_a,x_s) \in U(t_a,x_s)$ of Eqn. (12.26). From Proposition 17.7.7 and, if necessary, by further restricting the subset $X_s \subset X_{supp}(x^{g^*}(t_a))$ to be a Borel measurable open set, follows that

$$\begin{aligned}
& \exists\, \varepsilon \in (0,1),\ \{\forall\, x_r \in X_s\ \exists\, u_\varepsilon \in U(t_a,x_r)\} \text{ such that } \forall\, x_s \in X_s \\
& h^*(t_a,x_s) = h(t_a,x_s,u^*(t_a,x_s)) \le h(t_a,x_s,u_\varepsilon) < h(t_a,x_s,u^*(t_a,x_s)) + \varepsilon \\
&< h(t_a,x_s,g^*(t_a,x_s)) \ \Rightarrow\ h(t_a,x_s,g^*(t_a,x_s)) > h(t_a,x_s,u_\varepsilon).
\end{aligned}$$

Define the function $g : T \times X \to U$

$$g(t,x) = \begin{cases} u_\varepsilon, & \text{if } (t,x) = (t_a,x_s) \in T \times X_s, \\ g^*(t_a,x), & \text{if } (t_a,x) \in T \times X \backslash X_s, \\ g^*(t,x), & \text{otherwise.} \end{cases}$$

By the choice of X_s and of u_ε, the function g is a measurable function though not necessarily a continuous function. From the above and the inequality that $P_X(X_s) > 0$ follows that

$$\begin{aligned}
& E[h(t_a,x^{g^*}(t_a),g^*(t_a,x^{g^*}(t_a)))] \\
&= E[b(t_a,x^{g^*(t_a)},g^*(t_a,x^{g^*}(t_a)) + b_1(f(t_a,x^{g^*}(t_a),g^*(t,x^{g^*}(t_a)),v(t_a)))] \\
&> E[b(t_a,x^{g^*(t_a)},g(t,x^{g^*}(t_a)) + b_1(f(t_a,x^{g^*}(t_a),g(t,x^{g^*}(t_a)),v(t_a))] \\
&= E[h(t_a,x^{g^*}(t_a),g(x^{g^*}(t_a)))].
\end{aligned}$$

Then $x^g(s) = x^{g^*}(s)$ for all $s = 0,\ldots,t_1 - 1$, hence $u^g(s) = u^{g^*}(s)$, for all $s = 0\ldots,t_1 - 1$, and,

$$\begin{aligned}
u^g(t_1 - 1) = g(t_1 - 1,x^g(t_1 - 1)) = g^*(t_1 - 1),x^{g^*}(t_1 - 1)). \text{ Hence,} \\
E[b(s,x^g(s),u^g(s))] = E[b(s,x^{g^*}(s),u^{g^*}(s))],\ \forall s = 0,\ldots,t_1 - 2.
\end{aligned}$$

Adding the latter equalities to the strict inequality above yields $J(g^*) > J(g)$. But this inequality contradicts that the control law $g^* \in G_M$ is optimal. Thus $g^*(t_1 - 1, x^{g^*}(t_1 - 1))$ achieves the infimum in the Dynamic Programming Equation at time $t_1 - 1$. From Lemma 12.6.6 then follows that, $V^{g^*}(t_1 - 1, x^{g^*}(t_1 - 1)) = J(g^*, t_1 - 1)$.

Next the induction hypothesis is that $g^*(s, x^{g^*}(s))$ achieves the infimum in Equation (12.15) and that $J(g^*, s) = V(s, x^{g^*}(s))$ for $s = t+1, t+2, \ldots, t_1 - 1$ and for a $t \in T$. Suppose that $g^*(t, x^{g^*}(t))$ does not achieve the infimum in Equation (12.15). Using the corresponding arguments as described above for the time $t_a = t_1 - 1 \in T$, it follows that there exists a function $g(t,.) : X \to U$ such that

$$\begin{aligned} & b(t, x^{g^*}(t), g^*(t, x^{g^*}(t))) + E[V(t+1, f(t, x^{g^*}(t), g^*(t, x^{g^*}(t)), v(t)))|F^{x^{g^*}(t)}] \\ > \; & b(t, x^{g^*}(t), g(t, x^{g^*}(t))) + E[V(t+1, f(t, x^{g^*}(t), g(t, x^{g^*}(t)), v(t)))|F^{x^{g^*}(t)}], \end{aligned}$$

and this inequality is strict with positive probability. Consequently,

$$\begin{aligned} & E[b(t, x^{g^*}(t)), g^*(t, x^{g^*}(t))) + V(t+1, f(t, x^{g^*}(t), g^*(t, x^{g^*}(t)), v(t)))] \\ > \; & E[b(t, x^{g^*}(t), g(t, x^{g^*}(t))) + V(t+1, f(t, x^{g^*}(t), g(t, x^{g^*}(t)), v(t)))]. \end{aligned}$$

Consider the Markov control law $g : T \times X \to U$,

$$g(s,x) = \begin{cases} g^*(s,x), \; \forall x \in X, s = 0, \ldots, t-1, t+1, \ldots, t_1, \\ g(t,x), \;\; \forall x \in X, \text{ where } g(t,x) \text{ is defined above.} \end{cases}$$

Then, $E[b(s, x^{g^*}(s), u^{g^*}(s))] = E[b(s, x^g(s), u^g(s))]$, $\forall s = 0, \ldots, t-1$. By the induction hypothesis, $g^*(s, x^{g^*}(s))$ for all $s = t+1, \ldots, t_1 - 1$. achieves the infimum in Equation (12.15). From Lemma 12.6.6 then follows that,

$$\begin{aligned} & E[V(t+1, x^{g^*}(t+1))] = E[J(g^*, t)] \\ = \; & E \sum_{s=t+1}^{t_1 - 1} b(x^{g^*}(s), g^*(s, x^{g^*}(s)) + b_1(x^{g^*}(t_1)], \\ & E[V(t+1, x^g(t+1))] = E[J(g,t)] \\ = \; & E \sum_{s=t+1}^{t_1 - 1} b(x^g(s), g^*(s, x^g(s)) + b_1(x^g(t_1)]. \end{aligned}$$

Using the inequality and the expressions stated above one obtains

$$\begin{aligned} J(g^*) &= E[\sum_{s=0}^{t_1 - 1} b(s, x^{g^*}(s), u^{g^*}(s)) + b_1(x^{g^*}(t_1)] \\ &= E[\sum_{s=0}^{t-1} b(s, x^{g^*}(s), u^{g^*}(s)) + b(t, x^{g^*}(t), u^{g^*}(t)) + \\ &\quad + \sum_{r=t+1}^{t_1 - 1} b(r, x^{g^*}(s), u^{g^*}(s)) + b_1(x^{g^*}(t_1)] \\ &= E[\sum_{s=0}^{t-1} b(s, x^{g^*}(s), u^{g^*}(s)) + b(t, x^{g^*}(t), u^{g^*}(t)) + V(t+1, x^{g^*}(t+1))] \\ &> E[\sum_{s=0}^{t-1} b(s, x^g(s), u^g(s))s + b(t, x^g(t), u^g(t)) + V(t+1, x^g(t+1))] = J(g), \end{aligned}$$

and the optimality of $g^* \in G_M$ is contradicted. Hence $g^*(t, x^{g^*}(t))$ achieves the infimum and from Lemma 12.6.6 follows that $J(g^*, t) = V(t, x^{g^*}(t))$. □

The dynamic programming procedure may be used to calculate or compute optimal control laws. If the input space U and state space X are finite sets, then elementary numerical computations lead to the control law as shown below for an example. If the input and state space are countably infinite or uncountable then attention must be directed to analytic solutions. There are several cases in which formula's can be derived for the value function and the control law. Several of these cases are presented in the subsequent sections.

Derivation of Explicit Optimal Control Laws

In the following sections the reader is provided a set of special cases and of examples of optimal stochastic control problems for which explicit optimal control laws can be obtained. The purpose of this section is clarify for the reader how these special cases and examples were selected.

The problem which motivates this section is: For which combination of a stochastic control system including the dynamics and the probability distributions, and of which cost rates on a finite horizon, does either (1) there exists an analytic solution of the value function and of the optimal control law; or (2) there exists a procedure to compute the optimal control law in a finite number of steps? An answer to problem (2) is almost clear and this is stated below. The answer to problem (1) is basically a realization problem of which the solution is not known to the author.

The following special cases and examples are discussed in this chapter.

1. The case of a Gaussian stochastic control system with a cost rate which is a quadratic form of the controlled output. In this case the optimal control law is a linear function of the current state, hence a Markov control law, and the value function is a quadratic function of the state. See Section 12.7.
2. The case of a finite stochastic control system with an arbitrary cost rate. In this case the optimal control law can be computed by solving a finite set of linear equations at every time. This is an answer to problem (2) formulated above. See Section 12.8.
3. The case of an optimal stochastic control problem in which the input set is a finite set. The optimal control law has then the from of an index control law where in practice the user has to compute the values of a finite set of indices and then to choose that index that minimizes a function. This is an answer to problem (2) formulated above.
4. The portfolio selection problem over a finite horizon. The optimal control law can be determined partly based on the strucuture of the control system and of the cost rate. In case the cost rate has a specific form, a more explicit result can be obtained. See Section 12.12.

5. A simple gambling problem with a logarithmic terminal cost. See Problem 12.9.12 and its solution.

It appears that problem (1) formulated above can be solved if the following steps can be carried out. These steps are related to a realization problem of which the solution is not yet known.

Consider a subset of functions of the state to be regarded as a set of specific value functions $V_s \subseteq X$. To help the reader, he/she may think of the set of V_s quadratic functions or of logarithmic functions. If it can be proven that for $t = t_1$, $V(t_1,.) \in V_s$; and if $V(t+1,.) \in V_s$ then $V(t,.) \in V_s$, then the analytic form of the value function at all times is known. From this analytic form it is then possible to derive properties of the optimal control law. In the case of a Gaussian stochastic control system and a quadratic cost rate in terms of the controlled output, the optimal control law will be a linear function of the state.

The above procedure is described in Section 12.9 with more details than provided here.

Which combination of stochastic control systems, terminal costs, and cost rates produce a value function for which there exists a subset V_s as formulated above? One can formulate sufficient conditions for this. One such condition is the conjugate optimization property described elsewhere.

Specific cases of the existence of a subset of value functions on the state set are:

- A Gaussian stochastic control system, with a quadratic terminal cost, and a quadratic cost rate in the state and the input, see Section 12.7.
- The gambling problem of Problem 12.9.12. with a logarithmic terminal cost and a zero cost rate.

More examples are likely to exist.

In the literature several properties of optimal control laws are defined. Here the concept of certainty equivalence is defined.

Definition 12.6.7. An optimal stochastic control problem is said to have the *certainty-equivalence property* if the optimal control law is identical to that of an associated deterministic control problem. The latter problem is deduced from the first by replacing all the random variables by their expectation.

Note that certainty equivalence requires that the control laws of the two problems to be the same. But the values of the associated control problems may be different. The certainty equivalence property defined above for stochastic control problems, corresponds to the concept with the same name for statistical decision problems as defined in Definition 11.6.8.

The solution to Problem 12.7.1 has the certainty equivalence property. The solution of the associated deterministic optimal control problem with a linear control system and quadratic cost function is presented in, for example, [18].

Definition 12.6.8. An optimal stochastic control problem on a finite horizon is said to have a *myopic control law* if at any time the control law is identical to the control law of an associated statistical decision problem for a single period. The word myopic is from the Greek language and in English it means *short sighted*.

For a examples of a myopic control law see [9, Sec. 3.3].

12.7 Control of a Gaussian Control System

The reader finds in this section the problem of optimal stochastic control for a Gaussian stochastic control system and a cost function which is a quadratic form of the controlled output. The optimal control law in the set of nonlinear Borel measurable control laws is a linear control with a time-varying feedback matrix. The optimal control law and its extension to time-invariant Gaussian stystems, are much used in engineering.

Problem 12.7.1. The *Problem of optimal stochastic control of a time-varying Gaussian stochastic control system with complete observations and with a quadratic cost function (acronym LQG-CO-FH)*. Consider the state equation of a Gaussian stochastic control system,

$$\begin{aligned}
x(t+1) &= A(t)x(t)+B(t)u(t)+M(t)v(t),\ x(0)=x_0,\\
z(t) &= C_z(t)x(t)+D_z(t)u(t),\ \ \forall\, t\in T(0:t_1-1),\\
z(t_1) &= C_z(t_1)x(t_1);\\
&\quad T(0:t_1),\ X=\mathbb{R}^{n_x},\ U=\mathbb{R}^{n_u},\ Z=\mathbb{R}^{n_z},\ n_x,\ n_u,n_z\in\mathbb{Z}_+,\ n_u\le n_z,\\
&\quad x_0\in G(0,Q_0),\ v(t)\in G(0,I),\ F^{x_0},\ F^v_{t_1}\text{ independent},\\
&\quad \forall\, t\in T(0:t_1-1),\ \operatorname{rank}(D_z(t))=n_u\ \Rightarrow\ D_z(t)^TD_z(t)\succ 0;
\end{aligned}$$

where v is standard Gaussian white noise. Consider the past-state information structure and the corresponding set G of control laws. For any $g\in G$ the closed-loop stochastic control system representation is given by,

$$\begin{aligned}
x^g(t+1) &= A(t)x^g(t)+B(t)g(t,x^g(0:t))+M(t)v(t),\ x^g(0)=x_0,\\
z^g(t) &= C_z(t)x^g(t)+D_z(t)g(t,x^g(0:t)),\ \forall\, t\in T(0:t_1-1),\\
z^g(t_1) &= C_z(t_1)x^g(t_1),\ \ u^g(t)=g(t,x^g(0:t)).
\end{aligned}$$

Define the cost rate and the terminal cost by the expressions,

$$\begin{aligned}
Q_{cr}(t) &= \begin{pmatrix} C_z^T(t)C_z(t) & C_z(t)^TD_z(t)\\ D_z^T(t)C_z(t) & D_z(t)^TD_z(t)\end{pmatrix}\succeq 0,\ D_z^T(t)D_z(t)\succ 0;\\
b(t,x_V,u_V) &= \begin{pmatrix} x_V\\ u_V\end{pmatrix}^T Q_{cr}(t)\begin{pmatrix} x_V\\ u_V\end{pmatrix}\succeq 0,\ \forall\, t\in T(0:t_1-1),\\
b_1(x_V) &= z_V^Tz_V=x_V^TC_z^T(t_1)C_z(t_1)x_V,\ C_z^T(t_1)C_z(t_1)\succeq 0.
\end{aligned}$$

The case in which,

$$\operatorname{rank}(D_z(t))<n_u \text{ hence } D_z(t)^TD_z(t)\succeq 0,\ \ \not\succ 0,$$

is called the *singular optimal control problem*. There is a solution procedure for this case, which the reader may find for the case of control of a deterministic linear system with a singular quadratic cost function in [27, 59, 71].

Define the cost function in terms of the cost rate and the terminal cost, and the optimal stochastic control problem by the following equations,

$$J(g) = E\left[\left(\sum_{s=0}^{t_1-1} z^g(s)^T z^g(s)\right) + z^g(t_1)^T z^g(t_1)\right], \; J : G \to \mathbb{R}_+,$$

$$\inf_{g \in G} J(g).$$

In the literature in several other research areas than control theory, no conditions are imposed on the stochastic control system for the existence of an optimal control law. The reader is expected to compute a solution and a solution should exist. It is clear that, if the control system does not depend on the input signal at all, for example of for all times $t \in T$, $B(t) = 0$, and if the controlled output does not depend on the input signal at all, $\forall\, t \in T$, $D_z(t) = 0$, then the optimal control problem is useless because the cost cannot be decreased at all. To make the optimal control approach effective, one has to impose an assumption of stochastic controllability.

J.C. Doyle has argued that for LQG regulators with partial observations, there do not exist guaranteed performance guarantees, see [25]. This is correct as will be proven in subsequent chapters. For the understanding of the reader, it is necessary to discuss the performance of an optimally controlled system in more detail for .

Based on the control objectives for optimal stochastic control problem for a Gaussian stochastic control system with complete observations one expects that the optimally controlled stochastic system meets the following particular control objectives:

1. the closed-loop system is exponentially stable;
2. in case of a time-invariant stochastic control system the eigenvalues of the closed-loop system are bounded away from the instability boundary by a prespecified margin; and
3. the value of the cost function is lower than in the case of no control.

It will be argued in Chapter 22, see below Theorem 22.2.4, that in general the eigenvalues of the closed-loop system are not bounded away from the instability boundary. The condition of stochastic controllability or stabilizability is required for the satisfaction of the first and the third particular control objectives mentioned above.

A condition of stochastic controllability on the stochastic control system will be imposed. This is done because it is necessary for the first and the third conditions mentioned above. In the literature of control theory little attention has been paid to stochastic controllability of time-varying stochastic control systems. In the literature of operations research, the concept of controllability of a stochastic system is not treated at all. There one applies the dynamic programming procedure even if the system is fully uncontrollable. Controllability is necessary to attain the control objectives of stability and of strict decrease of the cost function.

Stochastic controllability of a time-varying Gaussian system has been formulated in Def. 4.6.6. A characterization of the stochastic controllability is provided in Theorem 4.6.8. The expressions below are based on the characterization of controllability of a time-varying linear control system, see Theorem 21.2.15.

Assumption 12.7.2 *Consider the time-varying Gaussian stochastic control system of Problem 12.7.1. Define the conditions:*

1. *the system is supportable which in this case requires that,*

$$\begin{aligned} n_x &= \mathrm{rank}(\sum_{s=0}^{t_1-1} \Phi(s,0)M(s)M(s)^T\Phi(s,0)^T); \\ \Phi(t,s) &= A(t-1)A(t-2)A(t-3)\ldots A(s+1)A(s),\ t>s, \\ &\quad \Phi(t,t)=I,\ \Phi(t,s)=0 \text{ } if \, t<s; \Phi : T\times T \to \mathbb{R}^{n_x\times n_x}; \end{aligned}$$

2. *the system is controllable which in this case requires that,*

$$n_x = \mathrm{rank}(\sum_{s=0}^{t_1-1} \Phi(s,0)B(s)B(s)^T\Phi(s,0)^T);$$

3. *the system with the controlled output z is stochastically observable, which in this case is equivalent to,*

$$n_x = \mathrm{rank}(\sum_{s=0}^{t_1-1} \Phi(s,0)^T C(s)^T C(s)\Phi(s,0)).$$

Example 12.7.3. *Example of a non-controllable Gaussian stochastic control system.* Consider a time-varying Gaussian stochastic control system with the following system representation,

$$\begin{aligned} x(t+1) &= \begin{pmatrix} A_{11}(t) & 0 \\ A_{21}(t) & A_{22}(t) \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} + \begin{pmatrix} 0 \\ B_2(t) \end{pmatrix} + \begin{pmatrix} M_1(t) \\ M_2(t) \end{pmatrix} v(t),\ x(0)=x_0, \\ z(t) &= \begin{pmatrix} C_{z,1}(t) & C_{z,2}(t) \end{pmatrix} x(t) + D_z(t)u(t). \end{aligned}$$

Such a system representation arises if a noise-shaping system determines the noise process of a Gaussian stochastic control system.

Note that the input u does not directly influence the first component x_1 of the state vector x, Neither does the second state component x_2 influence the first state component x_1. Therefore the above Gaussian stochastic control system is not stochastically controllable. One cannot achieve a particular measure on the first state component at a future time $t_1 \in T$.

This stochastic control system does not satisfy condition (2) of Assumption 12.7.2. In Chapter 13 a condition of stabilizability will be defined that is relevant for a system with the above representation.

Definition 12.7.4. Define the *optimal control law* for the optimal stochastic control problem with a time-varying Gaussian stochastic control system with complete observations on a finite horizon (LQG-CO-FH) by the sets and functions,

$$\begin{aligned}
&g^*_{LQG,CO,FH} : T \times X \rightarrow U, \\
&Q_c : T \rightarrow \mathbb{R}^{n_x \times n_x}_{pds},\ F : T \times \mathbb{R}^{n_x \times n_x}_{pds} \rightarrow \mathbb{R}^{n_u \times n_x}, \\
&g^*_{LQG,CO,FH}(t, x_V) = F(t, Q_c(t+1))x_V, \text{ where,} \qquad (12.27) \\
Q_c(t_1) &= C_z^T(t_1)C_z(t_1), \\
Q_c(t) &= A(t)^T Q_c(t+1)A(t) + C_z^T(t)C_z(t) + \\
&\quad -[A(t)^T Q_c(t+1)B(t) + C_z^T(t)D_z(t)] \times \\
&\quad \times [B(t)^T Q_c(t+1)B(t) + D_z^T(t)D_z(t)]^{-1} \times \\
&\quad \times [A(t)^T Q_c(t+1)B(t) + C_z^T(t)D_z(t)]^T,\ \forall\, t \in T(0:t_1-1); \qquad (12.28) \\
&\quad \text{then, } \forall\, t \in T(0:t_1-1),\ [B(t)^T Q_c(t+1)B(t) + D_z^T(t)D_z(t)] \succ 0; \\
&F(t, Q_c(t+1)) \qquad (12.29) \\
&= -[B(t)^T Q_c(t+1)B(t) + D_z^T(t)D_z(t)]^{-1} \times \\
&\quad \times [A(t)^T Q_c(t+1)B(t) + C_z^T(t)D_z(t)]^T.
\end{aligned}$$

Call equation (12.28) for the matrix-valued function Q_c with the terminal state $Q_c(t_1)$, the (backward) *control Riccati recursion*, the control law of equation (12.27) the *optimal LQG-CO-FH* control law, and the matrix function F of equation (12.29) the *optimal LQG-CO-FH feedback matrix*.

Define the function *control Riccati recursion* by the formula,

$$\begin{aligned}
f_{CR} &: T \times \mathbb{R}^{n_x \times n_x}_{pds} \rightarrow \mathbb{R}^{n_x \times n_x}_{pds}, \\
f_{CR}(t, Q) &= A(t)^T Q A(t) + C_z^T(t)C_z(t) + \\
&\quad -[A(t)^T Q B(t) + C_z^T(t)D_z(t)][B(t)^T Q B(t) + D_z^T(t)D_z(t)]^{-1} \times \\
&\quad \times [A(t)^T Q B(t) + C_z^T(t)D_z(t)]^T.
\end{aligned}$$

Theorem 12.7.5. *Consider Problem 12.7.1. All the conditions of Assumption 12.7.2 are expected to hold. Assume that for all $t \in T$ and for all $x_V \in X$ the $U(t, x_V) \subseteq U$ is a nonempty convex set.*

(a)The control law $g^*_{LQG-CO-FH} \in G$ and the backward Riccati recursion for the matrix-valued function Q_c as specified in Def. 12.7.4 are well defined.

(b)The closed-loop system associated with the Gaussian stochastic control system and the control law $g^*_{LQG-CO-FH}$ is a Gaussian system represented by,

$$\begin{aligned}
x^{g^*}(t+1) &= [A(t) + B(t)F(t, Q_c(t+1))]x^{g^*}(t) + M(t)v(t),\ x^{g^*}(0) = x_0, \\
z^{g^*}(t) &= [C(t) + D(t)F(t, Q_c(t+1))]x^{g^*}(t).
\end{aligned}$$

The stochastic processes x^{g^*}, z^{g^*}, v and the random variable x_0 are jointly Gaussian, hence the above is a Gaussian system representation. The mean and the variance functions of the state and of the controlled output of the above closed-loop system can then be calculated, of which the variance equations are represented by,

$$\begin{aligned}
Q_{x^{g^*}}(t+1) &= [A(t)+B(t)F(t,Q_c(t+1))]Q_{x^{g^*}}(t)\times \\
&\quad \times[A(t)+B(t)F(t,Q_c(t+1))]^T+M(t)M(t)^T, \\
&\quad Q_{x^{g^*}}(0)=Q_{x_0}, \\
Q_{z^{g^*}}(t) &= [C(t)+D(t)F(t,Q_c(t+1))]Q_{x^{g^*}}(t)\times \\
&\quad \times[C(t)+D(t)F(t,Q_c(t+1))]^T.
\end{aligned}$$

(c)The value function and the value are given by,

$$\begin{aligned}
&Q_c : T \rightarrow \mathbb{R}_{pds}^{n_x\times n_x} \text{ as defined in (a),} \\
V(t,x_V) &= x_V^T Q_c(t)x_V + r(t),\ r : T \rightarrow R, \\
r(t_1) &= 0,\ r(t) = r(t+1) + \text{tr}\left(Q_c(t+1)M(t)M(t)^T\right) \\
&= \sum_{s=t}^{t_1-1} \text{tr}(Q_c(s+1)M(s)M(s)^T); \qquad (12.30) \\
J^*_{LQG,CO,FH} &= J(g^*_{LQG,CO,FH}) = E[V(0,x_0)] = E[x_0^T Q_c(0)x_0] + r(0) \\
&= \text{tr}(Q_c(0)\ Q_{x_0}) + r(0). \qquad (12.31)
\end{aligned}$$

The optimal control law is a linear function of the state and it is optimal over the set of all measurable nonlinear control laws.

(d)Note that the optimal control law $g^*_{LQG,co,fh}$ does depend on the system matrix functions of the system (A, B), and on the matrix functions of the controlled output (C_z,D_z) but does not *depend on the matrix function M of the noise model. However, the value $J^*_{LQG,co,fh}$ depends on the matrix function M of the noise model, see equation (12.30).*

Proof. (a) The backward recursion of the control Riccati recursion is well defined. The formula of the control law $g^*_{LQG-CO-FH}$ then follows.

(b) The formulas of the closed-loop system are directly calculated using the control law $g_{LQG-CO-FH}(t,x) = F(t,Q_c(t+1))x$. The formulas of the variances then follow from Theorem 4.3.5.

(c) The dynamic programming equation will be applied using induction. Define the functions,

$$\begin{aligned}
H_{xx}(t) &= A(t)^T Q_c(t+1)A(t) + C_z^T(t)C_z(t), \\
H_{xu}(t) &= A(t)^T Q_c(t+1)B(t) + C_z^T(t)D_z(t), \\
H_{uu}(t) &= B(t)^T Q_c(t+1)B(t) + D_z^T(t)D_z(t), \\
H(t) &= \begin{pmatrix} H_{xx}(t) & H_{xu}(t) \\ H_{xu}(t)^T & H_{uu}(t) \end{pmatrix}.
\end{aligned}$$

By the first step of Procedure 12.6.2, $V(t_1,x_V) = b_1(x_V) = x_V^T C_z(t_1)^T C_z(t_1)x_V$. Define,

$$\begin{aligned}
Q_c(t_1) &= C_z(t_1)^T C_z(t_1),\ r(t_1) = 0, \text{ then, } Q_c(t_1) = Q_c(t_1)^T \succeq 0, \\
V(t_1,x_V) &= x_V^T Q_c(t_1)x_V + r(t_1).
\end{aligned}$$

Suppose that for $s = t+1,\ldots,t_1$,

$$Q_c(s) = Q_c(s)^T \succeq 0,\ V(s,x_V) = x_V^T Q_c(s) x_V + r(s).$$

It will be shown that then this formula also holds for $s = t$. The dynamic programming procedure prescribes to solve,

$$\inf_{u_V \in U(x_V)} \left\{ z^g(t)^T z^g(t) + E[V(t+1, f(t,x_V,u_V,v(t)))|F^{x_V,u_V}] \right\}.$$

The conditional expectation may be calculated as follows.

$$\begin{aligned}
&E[V(t+1,f(t,x_V,u_V,v(t)))|F^{x_V,u_V}] \\
&= E\big[\big(A(t)x_V + B(t)u_V + M(t)v(t)\big)^T Q_c(t+1)\big(A(t)x_V + B(t)u_V + M(t)v(t)\big) + \\
&\quad + r(t+1)|F^{x_V,u_V}\big],\ \text{ by the induction hypothesis for } s = t+1, \\
&= \begin{pmatrix} x_V \\ u_V \end{pmatrix}^T \left(\begin{array}{c|c} A(t)^T Q_c(t+1) A(t) & A(t)^T Q_c(t+1) B(t) \\ \hline B(t)^T Q_c(t+1) A(t) & B(t)^T Q_c(t+1) B(t) \end{array}\right) \begin{pmatrix} x_V \\ u_V \end{pmatrix} \\
&\quad + \text{tr}\left(Q_c(t+1) M(t) M(t)^T\right) + r(t+1), \\
&\quad \text{because of } E[v(t)|F^{x_v}] = 0,\ v(t) \in G(0,I),\ \text{ and of Proposition 2.7.6.}
\end{aligned}$$

Note that $H(t) = H(t)^T \succeq 0$ and that from $Q_c(t+1) \succeq 0$ follows that,

$$H_{uu}(t) = H_{uu}(t)^T = B(t)^T Q_c(t+1) B(t) + D_z(t) D_z(t) \succeq D_z(t)^T D_z(t) \succ 0,$$

by the induction hypothesis and the assumption of Problem 12.7.1 on D_z. Hence,

$$\begin{aligned}
&V(t,x_V)\\
&= \inf_{u_V \in U(t,x_V)} \left\{ z^g(t)^T z^g(t) + E[V(t+1, f(t,x_V,u_V,v(t)))|F^{x_V}] \right\}\\
&= \inf_{u \in U(t,x_V)} \left\{ \begin{pmatrix} x_V \\ u_V \end{pmatrix}^T \begin{pmatrix} C_z(t) \\ D_z(t) \end{pmatrix}^T \begin{pmatrix} C_z(t) \\ D_z(t) \end{pmatrix} \begin{pmatrix} x_V \\ u_V \end{pmatrix} + \right.\\
&\quad \left. + E[V(t+1, f(t,x_V,u_V,v(t)))|F^{x_V}] \right\}\\
&= \inf_{u_V \in U(t,x_V)} \left\{ \begin{pmatrix} x_V \\ u_V \end{pmatrix}^T H(t) \begin{pmatrix} x_V \\ u_V \end{pmatrix} + \operatorname{tr}\left(Q_c(t+1)M(t)M(t)^T\right) + r(t+1) \right\}\\
&= \inf_{u_V \in U(t,x_V)} \left\{ \begin{pmatrix} x_V \\ u_V + H_{22}(t)^{-1}H_{12}(t)^T x_V \end{pmatrix}^T \times \right.\\
&\quad \times \begin{pmatrix} H_{11}(t) - H_{12}(t)H_{22}^{-1}(t)H_{12}^T(t) & 0 \\ 0 & H_{22}(t) \end{pmatrix} \times\\
&\quad \left. \times \begin{pmatrix} x_V \\ u_V + H_{22}(t)^{-1}H_{12}(t)^T x_V \end{pmatrix} + \operatorname{tr}\left(Q_c(t+1)M(t)M(t)^T\right) + r(t+1) \right\}\\
&\quad \text{using the technique of completion of squares,}\\
&= x_V^T [H_{xx}(t) - H_{xu}(t)H_{uu}(t)^{-1}H_{xu}(t)^T] x_V +\\
&\quad + \operatorname{tr}\left(Q_c(t+1)M(t)M(t)^T\right) + r(t+1),\\
&\quad \text{because } H_{22}(t) \succ 0 \text{ and for } u^* = u_V = -H_{uu}(t)^{-1}H_{xu}(t)^T x_V,\\
&= x_V^T Q_c(t) x_V + r(t), \text{ if,}\\
&\quad Q_c(t) = H_{xx}(t) - H_{xu}(t)H_{uu}(t)^{-1}H_{xu}(t)^T,\\
&\quad r(t) = r(t+1) + \operatorname{tr}\left(Q_c(t+1)M(t)M(t)^T\right).
\end{aligned}$$

Note that the backward recursions for Q_c and r are as given in Equations (12.28) and (12.30). Define,

$$\begin{aligned}
&g^*_{LQG,co,fh} : T \times X \to U,\\
g^*(t,x_V) &= u^* = -H_{22}(t)^{-1}H_{12}(t)^T x_V\\
&= -[B(t)^T Q_c(t+1)B(t) + D_z(t)^T D_z(t)]^{-1} \times\\
&\quad \times [A(t)^T Q_c(t+1)B(t) + C_z(t)^T D_z(t)]^T x_V.
\end{aligned}$$

Note that $g^*_{LQG,co,fh}(t,x_V)$ is a linear function in x_V hence a measurable function. Thus g^* is a solution of the dynamic programming procedure. From Theorem 12.6.4 then follows that $g^*_{LQG,co,fh}$ is an optimal control law. Because $H_{uu}(t) \succ 0$ for all $t \in T$, the argument of the optimization $u^* = g^*(t,x_V)$ is unique, hence the control law $g^*_{LQG,co,fh}$ is unique. This proves (a) and (b).
(c) This follows directly from (a) and (b). □

In control engineering there are many applications of the optimal control law for an optimal control problem with a Gaussian stochastic control system and with a quadratic cost function on a finite horizon. Most of these applications are for time-invariant Gaussian control systems where attention is restricted to a time-invariant control law.

Example 12.1.1 for shipsteering is an example where the LQG optimal control law is used. The example is modified to account for slow various of the wind. Such modifications occur often in control engineering.

Example 4.1.1 with control of a paper machine is a singular optimal stochastic control problem. That the problem has a singular costs is because the amount of material flow from the pulp was not taken into consideration. That particular control requires a different solution procedure discussed in Chapter 15. However, if one formulates a controlled output satisfying the condition that $\text{rank}(D_z) = n_u$ then one satisfies the conditions of Theorem 12.7.5 and the LQG optimal control law can be applied. In this case the amount of material flow enters into the cost function which contributes both to a lower cost and to sustainability.

There exists a minor generalization of the LQG-optimal control law which establishes a robustness property of this result, see the following corollary.

Corollary 12.7.6. *Consider the LQG-CO-FH optimal control Problem 12.7.1 except that the noise process v is not Gaussian white noise. Assume that the stochastic process $v : \Omega \times T \to \mathbb{R}^{n_v}$ is a sequence of independent random variables, square integrable with, for all $t \in T$, $E[v(t)] = 0$ and $Q_v(t) = E[v(t)v(t)^T] \succ 0$. The probability distribution is not specified for this stochastic process. Assume also that $F^v_{t_1}$ and F^{x_0} are independent.*

The optimal control for this modified LQG optimal stochastic control problem is equal to $g^*_{LQG-CO-FH}$ except that the value is different; both are specified by,

$$g_{LQG-CO-FH}(x) = F(t, Q_c(t+1))x,$$

$$J^* = \text{tr}(Q_c(0)Q_{x_0}) + \sum_{s=0}^{t_1-1} \text{tr}(Q_c(s+1)M(s)Q_v(s)M(s)^T).$$

Proof. The probability distribution function of the noise enters only at one point in the proof which is described below,

$$\begin{aligned}
&E[V(t+1, f(t, x_V, u_V, v(t)))|\, F^{x,u}] \\
&= E[(A(t)x_V + B(t)u_V + M(t)v(t))^T Q_c(t+1)(\ldots) + r(t+1)] \\
&= (A(t)x_V + B(t)u_V)^T Q_c(t+1)(A(t)x_V + B(t)u_V) + \\
&\quad +2(A(t)x_V + B(t)u_v)^T Q_c(t+1)ME[v(t)] + \\
&\quad +\text{tr}(M(t)^T Q_c(t+1)M(t)Q_v(t)) + r(t+1) \\
&= (A(t)x_V + B(t)u_V)^T Q_c(t+1)(A(t)x_V + B(t)u_V) \\
&\quad +\text{tr}(Q_c(t+1)M(t)Q_v(t+1)M(t)^T) + r(t+1).
\end{aligned}$$

The remainder of the proof is identical to that of Theorem 12.7.5. □

12.8 Control of a State-Finite Stochastic Control System

The general framework for optimal stochastic control systems also applies to an optimal stochastic control problem for a state-finite stochastic control system with any cost function. Thus the results of Section 12.6 apply.

For an optimal stochastic control problem with a state-finite stochastic system the computations are elementary and can be carried out by a computer program. The value function can be obtained by computation. The computation is possible because for this stochastic control system the state set is a finite set. The input set may be a finite set in which case the computation is simple or it may be a continuous-subset of the real numbers in which case there is a calculation required. How to carry out the computations is illustrated by the following example of inventory control.

Example 12.8.1. *Inventory control.* Consider Example 11.1.3 of a shop selling radios. The problem is to determine an optimal control law for ordering radios of a shop such that the economic cost of the shop is minimized. The notation here is,

$x(t)$	Stock available at beginning of the t-th period
$u(t)$	Stock ordered and immediately delivered at beginning of the t-th period
$v(t)$	The random variable denoting the demand in the t-th period
h	Holding cost per item
k	Ordering cost per item
b	Shortage cost per item for customer demand that cannot be met

Let $T = \{0,1,2,3\}$, $X = \{0,1,2\}$, $U = \{0,1,2\}$, $x, v : \Omega \times T \rightarrow \mathbb{N}$. Suppose that $\{x_0, v(0), v(1), v(2), v(3)\}$ are independent random variables. Moreover, let v be an independent identically-distributed sequence with for any $t \in T$,

$$P(\{v(t) = 0\}) = 0.2,\ P(\{v(t) = 1\}) = 0.6,\ P(\{v(t) = 2\}) = 0.2.$$

To simplify the computations, the condition is imposed that $x(t) + u(t) \leq 2$ for all $t \in T$. Hence $U(x) = \{u \in U|\ x + u \leq 2\}$, thus $U(0) = \{0,\ 1,\ 2\}$, $U(1) = \{0,\ 1\}$, and $U(2) = \{0\}$. Suppose further that customer demand in excess of the available supply, occuring if $v(t) - x(t) - u(t) > 0$, is lost to the shop. For each unsatisfied customer, a cost is to be paid. Take the numerical values equal to $h = k = 1$, $b = 2$. The state dynamics are,

$$x(t+1) = \max\{0, x(t) + u(t) - v(t)\}. \tag{12.32}$$

The control objective is to minimize the cost rather than to maximize profit as in Example 11.1.3. The cost function is then,

$$J(g) = E\left[\sum_{s=0}^{2} \begin{pmatrix} k \times u(s) + b\max\{0, v(s) - x(s) - u(s)\} + \\ +h\max\{0, x(s) + u(s) - v(s)\} \end{pmatrix} + b_1(x(3))\right].$$

The control objectives confict between disappointing customers if there is no stock left and the holding costs. Note that if the stock is zero at a particular day then all customers requesting products that day are disappointed. Moreover, if the shop is fully stocked at a day but less customers request products then a holding cost has to be paid at the end of the day. The optimal control law strikes a balance between these two conflicting control objectives.

Consider the past-state information structure and the corresponding set of control laws $g = \{g_0, g_1, g_2, \ldots, g_{t_1-1}\}$ with $g_t : X^{t+1} \to U$. The problem is then to solve the optimal stochastic control problem $\inf_{g\in G} J(g)$.

The solution is determined by application of the dynamic programming procedure. Due to the formulation of the cost rate, in the dynamic programming procedure the cost rate is included in the conditional expectation operator.

First, $\forall\ x \in X$, define the value function at the terminal time by $V(3,x) = b_1(x)$ or $V(3,0) = 3$, $V(3,1) = 2$, and $V(3,2) = 3$. Next,

$$\begin{aligned}
&V(2,x(2)) \\
&= \min_{u(2)\in U(x(2))} E[ku(2) + b\max\{0, v(2) - x(2) - u(2)\} + \\
&\quad +h\max\{0, x(2) + u(2) - v(2)\} + V(3, f(x(2), u(2), v(2))|F^{x(2),u(2)}], \\
&V(2,0) \\
&= \min_{u(2)\in U(0)=\{0,1,2\}} \{E[ku(2) + b\max\{0, v(2) - u(2)\} + \\
&\quad +h\max\{0, u(2) - v(2)\} + V(3, \max\{0, u(2) - v(2)\})|F^{u(2)}]\Big\} \\
&\quad (u(2)=0) \quad P(\{v(2)=0\})V(3,0) + \\
&\qquad +P(\{v(2)=1\})[b + V(3,0)] + P(\{v(2)=2\})[2b + V(3,0)] \\
&\qquad = b + V(3,0) = 2 + 3 = 5; \\
&\quad (u(2)=1)\ P(\{v(2)=0\})[k + h + V(3,1)] + \\
&\qquad +P(\{v(2)=1\})[k + V(3,0)] + P(\{v(2)=2\})[k + b + V(3,0)] \\
&\qquad = k + 0.2b + 0.2h + 0.8V(3,0) + 0.2V(3,1) = 4.4; \\
&\quad (u(2)=2)\ P(\{v(2)=0\})[2k + 2h + V(3,2)] + \\
&\qquad +P(\{v(2)=1\})[2k + h + V(3,1)] + P(\{v(2)=2\})[2k + V(3,0)] \\
&\qquad = 2k + h + [0.2V(3,2) + 0.6V(3,1) + 0.2V(3,0)] = 5.4; \\
&= \min_{u(2)\in\{0,1,2\}} \{5.0,\ 4.4,\ 5.4\} = 4.4,\ \text{thus},\ g^*(2,0) = 1.
\end{aligned}$$

$$\begin{aligned}
V(2,1) &= \min_{u(2)\in U(1)=\{0,1\}} \{E[ku(2) + b\max\{0, v(2) - 1 - u(2)\} + \\
&\quad +h\max\{0, 1 + u(2) - v(2)\} + V(3, \max\{0, 1 + u(2) - v(2)\})|\ F^x]\} \\
&= \min_{u(2)\in\{0,1\}} \{3.4,\ 4.4\} = 3.4,\ g^*(2,1) = 0.
\end{aligned}$$

$$\begin{aligned}
&V(2,2) = E[ku + b\max\{0, v(2) - 2\} + h\max\{0, 2 - v(2)\}|\ F^x] = 3.4, \\
&g^*(2,2) = 0.
\end{aligned}$$

$$\begin{aligned}
V(1,0) = & \min_{u(1)\in U(0)=\{0,1,2\}} \{E[ku(1)+b\max\{0,v(1)-0-u(1)\}+\\
& +h\max\{0,0+u(1)-v(1)\}+V(2,\max\{0,0+u(1)-v(1)\})|\, F^x]\},\\
& (u(1)=0) \quad P(\{v(2)=0\})[0+V(2,0)]+\\
& \quad +P(\{v(2)=1\})[b+V(2,0)]+P(\{v(2)=2\})[2b+V(2,0)]\\
& \quad b+V(2,0)=2+4.4=6.4,\\
& (u(1)=1) \quad P(\{v(2)=0\})[k+h+V(2,1)]+\\
& \quad +P(\{v(2)=1\})[k+V(2,0)]+\\
& \quad +P(\{v(2)=2\})[k+b+V(2,0)]\\
& \quad =k+0.2b+0.2h+[0.2V(2,1)+0.8V(2,0)]\\
& \quad =1.2+0.4+4.2=5.8,\\
& (u(1)=2) \quad P(\{v(2)=0\})[2k+2h+V(2,2)]+\\
& \quad +P(\{v(2)=1\})[2k+h+V(2,1)]+\\
& \quad +P(\{v(2)=2\})[2k+V(2,0)]\\
& \quad =2k+h+[0.2V(2,2)+0.6V(2,1)+0.2V(2,0)]\\
& \quad =3+3.6=6.6,\\
= & \min\{6.4,\ 5.8,\ 6.6\}=5.8,\ \ g^*(1,0)=1;
\end{aligned}$$

$$\begin{aligned}
V(1,1) = & \min_{u(1)\in U(1)=\{0,1\}} \{E[ku(1)+b\max\{0,v(1)-1-u(1)\}+\\
& +h\max\{0,1+u(1)-v(1)\}+V(2,\max\{0,1+u(1)-v(1)\})|\, F^x]\},\\
= & \min\{4.8,\ 5.6\}=4.8,\ g^*(1,1)=0,;
\end{aligned}$$

$$V(1,2) = \min_{u(1)\in U(0)=\{0\}} \{4.6\}=4.6,\ g^*(1,2)=0.$$

$$\begin{aligned}
V(0,0) = & \min_{u(0)\in U(x(0))=\{0,1,2\}} \{E[ku(0)+b\max\{0,v(0)-0-u(0)\}\\
& +h\max\{0,0+u(0)-v(0)\}+\\
& +V(1,\max\{\ 0,0+u(0)-v(0)\})|\, F^x]\}\\
= & \min\{7.8,7.2,7.96\}=7.2,\ \ g^*(0,0)=1;
\end{aligned}$$

$$V(0,1) = \min_{u\in U(1)=\{0,1\}} \{6.2,\ 6.96\}=6.2,\ \ g^*(0,1)=0.$$

$$V(0,2) = \min_{u\in U(2)=\{0\}} 5.96=5.96,\ \ g^*(0,0)=0.$$

The optimal control law is then,

$$g^*(0,x(1)) = \begin{cases} 1, & \text{if } x(1)=0,\\ 0, & \text{if } x(1)=1,\\ 0, & \text{if } x(1)=2, \end{cases} \quad g^*(1,x(1)) = \begin{cases} 1, & \text{if } x(1)=0,\\ 0, & \text{if } x(1)=1,\\ 0, & \text{if } x(1)=2, \end{cases}$$

$$g^*(2,x(2)) = \begin{cases} 1, & \text{if } x(2)=0,\\ 0, & \text{if } x(2)=1,\\ 0, & \text{if } x(2)=2. \end{cases}$$

The optimal control law is such that for all $t \in \{0,1,2\}$ if the current state $x(t) = 0$ then one radio is ordered, $u(t) = 1$, while, if the current state is $x(t) > 0$, then no radio is ordered, $u(t) = 0$.

12.9 Invariance of a Subset of Value Functions

The reader may wonder whether optimal stochastic control problems admit an analytic solution for the value function of a broader set of stochastic control systems than that of a Gaussian stochastic control system with a quadratic cost rate.

For the Gaussian stochastic control problem it was proven that the value function consists of the sum of a quadratic function of the state and of a deterministic term not depending on the state,

$$V(t,x_V) = x_V^T Q_c(t) x_V + r(t). \tag{12.33}$$

If the value function has a particular analytic form then the control law can also be derived in analytic form. This is useful for control theory because the analytic form of a control law reveals the way in which the control law depends on the state of the system and on the parameters of the problem.

Below is discussed the research issue: Do there exist subsets of value functions which have a particular invariance property? There exist algebraic conditions on a stochastic system and on the terminal cost which imply the existence of in invariant value function. These conditions are investigated in this section.

The problem formulation is preceeded by several preliminaries. The stochastic system in the recursive formulation has the representation $x(t+1) = f(t,x(t),u(t),v(t))$. In terms of the conditional transition probability, assuming time-invariance of this function, the recursive stochastic control system is described by the conditional probability function $f(.|.,.;\ x(t+1)|x(t),u(t))$. The latter conditional probability distribution function denotes the cpdf of the random variable $x(t+1)$ conditioned in $(x(t),\ u(t))$. Note that for a value function V_1, the conditional expectation can thus be written as,

$$E[V_1(f(x(0),u(0),\ v))|\ F^{x(0),u(0)}] = \int V_1(w_{x_1}) f(dw_{x_1}|.,.;\ x(0),u(0)).$$

Attention is restricted to one step of the dynamic programming procedure starting at time $t = 1$ with the value function V_1, moving backwards to time $t = 0$, and obtaining the value function V_0. Below the zeros of (x_0,u_0) are no longer written hence one writes $f(.|.,.;\ x(1)|\ x,u)$. The expression of interest is thus,

$$\begin{aligned} V_0(x) &= \inf_{u \in U(x)} \{b(x,u) + E[V_1(f(x,u,v))|\ F^{x,u}]\} \\ &= \inf \{b(x,u) + \int V_1(w_{x_1})\ f(dw_{x_1}|.,.;\ x,u)\} \\ &= \inf_{u \in U(x)} \{\int \ [b(x,u) +\ V_1(w_{x_1})]\ f(dw_{x_1}|.,.;\ x,u)\}. \end{aligned}$$

Problem 12.9.1. Consider an optimal stochastic control problem for a stochastic control system described by,

$$\begin{aligned}
& X = \mathbb{R},\ U = \mathbb{R},\ \forall\, x \in X,\ U(x) = U,\ T = \{0,1\}, \\
& b : X \times U \to \mathbb{R}_+,\ \text{a Borel measurable function;} \\
& f(.|.,.;\ x(1)|\ x,u) \text{ the conditional pdf of } x(1) \text{ conditioned on } F^{x,u}, \\
V_s \subseteq & \{V : X \to \mathbb{R}_+|\ V \text{ Borel measurable function}\},\ V_1 \in V_s, \\
V_0(x) = & \inf_{u \in U} \left\{ b(x,u) + \int V_1(w_{x(1)})\, f(dw_{x(1)}|.,.;\ x,u) \right\}.
\end{aligned}$$

In case of a multiplicative cost function the operator has to be modified to,

$$V_0(x) = \inf_{u \in U} \left\{ b(x,u) \times \int V_1(w_{x(1)})\, f(dw_{x_1}|.,.;\ x,u). \right\}.$$

The operations needed to calculate V_0 are: (1) integration over the conditional pdf, (2) addition or multiplication of the integral with the cost rate $b(x,u)$; and (3) infimization of a function over $u \in U$.

For which subsets V_s of value functions is it true that $V_1 \in V_s$ implies that $V_0 \in V_s$? What is the smallest such subset?

Definition 12.9.2. Consider the setting of Problem 12.9.1 Call the triple $(V_s,\ f(.|.,.;\ x(1)|\ x,u),\ b(x,u))$ a *control-conjugate triple* if the condition holds that in the quoted problem $V_0 \in V_s$ implies that $V_1 \in V_s$.

How to determine control-conjugate triples of functions?

In regard to the conditional probability functions one may consider the sets of conditional probability distributions: Bernoulli, Poisson, Beta, Gamma, and Gaussian, etc. where, in case there the cpdf has two or more parameters, one has to choose which parameter depends on the state and the input variables.

In regard to the subsets of value functions V_s one can consider the subsets consisting of one or more of the following algebraic forms,

$$x,\ x^2,\ x^k \text{ for } k \in \mathbb{Z}_+,\ \exp(-ax) \text{ with } a \in (0,\infty),\ x^k \exp(-ax),\ \ln(x),\ \text{etc.}$$

In regard to the cost rate, this is best chosen based on the analytic form of the integral.

The reader finds below the solution to Problem 12.9.1 for a Gaussian stochastic control problem which is similar to the result derived in Section 12.7. There are also shown calculations for other examples which so far did not lead to other examples of control-conjugate functions.

Proposition 12.9.3. Invariance of a subset of value functions in case of a Gaussian conditional probability distribution function. *Consider Problem 12.9.1 for the case in which,*

$$\begin{aligned}
& f(.|.,.;\ x(1)|x,u) \textit{ is Gaussian with } CG(.;\ x(1)|\ ax + bu, q_v), \\
& x \in X = \mathbb{R},\ u \in U = \mathbb{R},\ \forall\, x \in X,\ U(x) = U,\ a,\ b \in \mathbb{R},\ b \neq 0,\ q_v \in (0,\infty), \\
& b(x,u) = q_{xx}x^2 + q_{uu}u^2 + 2q_{xu}xu,\ q_{uu} \in (0,\infty),\ q_{xx},\ q_{xu} \in \mathbb{R}_+, \\
V_s = & \left\{ V : x \to \mathbb{R}_+|\ V(x) = q_x\, x^2 + c,\ q_x \in (0,\infty),\ c \in \mathbb{R}_+ \right\}.
\end{aligned}$$

Then,

$$\begin{aligned}
&V_0(x) = q_x x^2 \in V_s \;\Rightarrow\; V_1(x) = q_1 x^2 + c_1 \in V_s,\\
q_1 &= a^2 q_x + q_{xx} - \frac{(abq_x + q_{xu})^2}{b^2 q_x + q_{uu}},\; c_1 = q_v q_x,\;\; u^* = -\frac{abq_x + q_{xu}}{b^2 q_x + q_{uu}}\, x.
\end{aligned}$$

Thus $(V_s,\, CG(.;\, x_1|\, ax+bu, q_v), b(x,u))$ *is a control-conjugate tuple.*

Proof. A calculation shows that,

$$\begin{aligned}
V_0(x) &= \inf_{u \in U(x)} \left\{ b(x,u) + \int V_1(w_{x_1})\, f_{x_1|\,x,u}(dw_{x_1}; ax+bu, q_v) \right\}\\
&= \inf\, b(x,u) + \int q_x w_{x_1}^2\, \exp(-(w_{x_1} - [ax+bu])^2/2q_v) dw_{x_1}\\
&= \inf\, [b(x,u) + q_x(ax+bu)^2 + q_v q_x], \text{ by Proposition 2.7.6,}\\
&= \inf\, [q_{xx}x^2 + q_{uu}u^2 + 2q_{xu}xu + q_x a^2 x^2 + q_x b^2 u^2 + 2abq_x xu + q_v q_x]\\
&= [q_{xx}x^2 + q_x a^2 x^2 + q_v q_x] + \inf_{u \in U}\, [(b^2 q_x + q_{uu})u^2 + 2ux(abq_x + q_{xu})]\\
&= x^2[q_x a^2 + q_{xx} - \frac{(abq_x + q_{xu})^2}{b^2 q_x + q_{uu}}] + q_v q_x +\\
&\quad + \inf\, (b^2 q_x + q_{uu})[u + \frac{(abq_x + q_{xu})x}{b^2 q_x + q_{uu}}]^2\\
&= \left[q_x a^2 + q_{xx} - \frac{(abq_x + q_{xu})^2}{(b^2 q_x + q_{uu})} \right] x^2 + q_v q_x = q_1 x^2 + c_1 = V_0(x),
\end{aligned}$$

$$u^* = -\frac{(abq_x + q_{xu})}{b^2 q_x + q_{uu}}\, x.$$

□

The case of a conditional Gamma pdf is investigated.

Proposition 12.9.4. *Consider the sets and maps,*

$$\begin{aligned}
&X = (0,\infty),\; U = (0,\infty),\; T = \{0,1\},\\
&f(.|.,.;\, x(1)|x,u) \text{ a Gamma cpdf with parameters } (\gamma_1(x,u), \gamma_2(x,u)),\\
V_s &= \{V : X \to \mathbb{R}_+ |\, V(x) = \exp(-ax) - 1 + bx,\; a,\; b \in (0,\infty),\; b < a\},\\
V_1(x) &= \exp(-ax) - 1 + bx.
\end{aligned}$$

Then,

$$E[\exp(-ax_1)|\, F^{x,u}] = (1 + a\gamma_2(x,u))^{-\gamma_1(x,u)}.$$

Proof.

$$
\begin{aligned}
& E[\exp(-ax_1)|\ F^{x,u}] \\
&= \int_0^\infty \exp(-aw_1)\ w_1^{\gamma_1(x,u)-1} \exp(-w_1/\gamma_2(x,u)) dw_1\ \gamma_2(x,u)^{-\gamma_1(x,u)}/\Gamma(\gamma_1(x,u)) \\
&= \int w_1^{\gamma_1(x,u)-1} \exp(-w_1[a+1/\gamma_2(x,u)]) dw_1\ \gamma_2(x,u)^{-\gamma_1(x,u)}/\Gamma(\gamma_1(x,u)) \\
&= (a+1/\gamma_2(x,u))^{-\gamma_1(x,u)}\ \gamma_2(x,u)^{-\gamma_1(x,u)} \\
&= (1+a\gamma_2(x,u)))^{-\gamma_1(x,u)}.
\end{aligned}
$$

□

Consider the optimization problem,

$$
\begin{aligned}
& \inf_{u \in (0,\infty)} h(u), \\
& h(u) = \exp(-au) - 1 + bu,\ U = (0,\infty),\ a,\ b \in (0,\infty),\ b < a,\ h : U \rightarrow \mathbb{R}_+.
\end{aligned}
$$

The function h is strictly convex on its domain of definition. The solution of this optimization problem is,

$$
u^* = \frac{-1}{a} \ln(\frac{b}{a}) \in (0,\infty),\ \ h(u^*) = \frac{b}{a}[1 - \ln(\frac{b}{a})] - 1.
$$

It is not clear yet which functions form a control-conjugate tuple for a Gamma state conditional probability distribution function.

Next the case of a Beta cpdf.

Proposition 12.9.5. *Consider the sets and maps,*

$$
\begin{aligned}
& X = (0,1),\ U = (0,1),\ T = \{0,1\}, \\
& f(.|.,.;\ x(1)|x,u) \text{ a Beta cpdf with parameters } (\beta_1(x,u), \beta_2(x,u)), \\
& V_s = \{V : X \rightarrow \mathbb{R}_+|\ V(x) = x^{\alpha_1}(1-x)^{\alpha_2},\ \ \alpha_1,\ \alpha_2 \in (0,\infty)\}; \\
& V_1(x) = x^{\alpha_1}(1-x)^{\alpha_2} \in V_s,\ \alpha_1,\ \alpha_2 \in (0,\infty).
\end{aligned}
$$

From Def. 2.1.7 and Def. 2.1.8 then follows that,

$$
\begin{aligned}
E[V_1(x_1)|\ F^{x,u}] &= B(\alpha_1 + \beta_1(x,u), \alpha_2 + \beta_2(x,u))/B(\beta_1(x,u), \beta_2(x,u)); \\
B(\beta_1(x,u), \beta_2(x,u)) &= \frac{\Gamma(\beta_1(x,u))\Gamma(\beta_2(x,u))}{\Gamma(\beta_1(x,u) + \beta_2(x,u))}, \\
\Gamma(\beta_1(x,u)) &= \int_{(0,\infty)} w^{\beta_1(x,u)-1}\ \exp(-w)\ dw.
\end{aligned}
$$

Proof. The proof is an integration,

$$
\begin{aligned}
& E[V_1(x_1)|\ F^{x,u}] = \int w^{\alpha_1}\ (1-w)^{\alpha_2}\ w^{\beta_1(x,u)-1}(1-w)^{\beta_2(x,u)-1} dw \\
& /\ B(\beta_1(x,u), \beta_2(x,u)) \\
&= \int w^{\alpha_1+\beta_1(x,u)-1}(1-w)^{\alpha_2+\beta_2(x,u)-1} dw/\ B(\beta_1(x,u), \beta_2(x,u)) \\
&= B(\alpha_1 + \beta_1(x,u), \alpha_2 + \beta_2(x,u))/B(\beta_1(x,u), \beta_2(x,u)).
\end{aligned}
$$

□

From the above proposition it is clear that the choice of the subset of value functions leads to a complicated optimization problem. The function V_0 is a function of the input variable u which does not lead to a simple analytic expression.

Proposition 12.9.6. *Consider the function*

$$f(u) = u^a(1-u)^b, \ f : (0,\infty) \to \mathbb{R}_+, \ a, \ b \in (2,\infty).$$

The function f is convex on its domain and its infimization leads to,

$$u^* = \frac{a}{a+b}, \ f(u^*) = \left(\frac{a}{a+b}\right)^a \times \left(1 - \frac{a}{a+b}\right)^b.$$

Proof. A simple calculation yields the result,

$$\begin{aligned} df(u)/du &= au^{a-1}(1-u)^b + bu^a(1-u)^{b-1}, \\ d^2f(u)/du^2 &= a(a-1)u^{a-2}(1-u)^b + 2abu^{a-1}(1-u)^{b-1} + \\ &\quad +b(b-1)u^a(1-u)^{b-2} \\ &= u^{a-2}(1-u)^{b-2} \times \\ &\quad \times [a(a-1)(1-u)^2 + 2abu(1-u) + b(b-1)(1-u)^2] > 0, \\ &\quad \forall \, u \in U = (0,\infty), \\ 0 = df(u)/du \ &\Rightarrow \ a(1-u) - bu = a - u(a+b) \ \Rightarrow \ u^* = \frac{a}{a+b}. \end{aligned}$$

□

Invariance over a Horizon

If a set of value functions has been determined which has the required invariance property then this formulation can be generalized to the finite-horizon case as described below.

Problem 12.9.7. For which stochastic control systems in combination with which cost functions, does the value function have, for all time moments, the same analytic form?

Definition 12.9.8. *Assumptions on the invariance of an additive value function set.* Consider the recursive control system and the cost function of Problem 12.2.1,

$$\begin{aligned} x(t+1) &= f(t,x(t),u(t),v(t)), \ x(0) = x_0, \\ u(t) &= g(t,x(t)), \\ J(g) &= E\left[\sum_{s=0}^{t_1} b(x(s),u(s)) + b_1(x(t_1))\right]. \end{aligned}$$

Define the triple of function spaces, where the symbol s stands for the word *special*,

$$\begin{aligned}
&(V_s, H_s, G_s),\\
&V_s \subseteq \{V : X \to \mathbb{R}_+ | V \text{ measurable function}\},\\
&H_s \subseteq \{H : X \times U \to \mathbb{R}_+ | H \text{ measurable function}\},\\
&G_s \subseteq \{g : X \to U | g \text{ a measurable function}\}.
\end{aligned}$$

Note that G_s is a set of time-invariant Markov control laws.

These function sets are said to satisfy the *additive invariance conditions* if the following conditions all hold:

1. if $V \in V_s$ and if $t \in T \backslash \{t_1\}$ then
$$E[V(f(t, x_v, u, v(t))|F^{x,u}] \in H_s; \tag{12.34}$$
2. the function set H_s is closed with respect to addition: if $H_1,\ H_2 \in H_s$ then $H_1 + H_2 \in H_s$; (linearity is not imposed)
3. the infimization is attained: if $H \in H_s$ and for all $x_v \in X$ there exists a unique $u^* \in U$ such that,
$$H(x_v, u^*) = \inf_{u \in U} H(x_v, u); \tag{12.35}$$
if one defines $g^* : X \to U$ by $g^*(x_v) = u^*$, then $g^* \in G_s$;
4. if $H \in H_s$ and if one defines $V(x_v) = H(x_v, u^*) = \inf_{u \in U} H(x_v, u)$ then $V \in V_s$;
5. $b \in H_s$; and
6. $b_1 \in V_s$.

Theorem 12.9.9. *Consider the setting of Def. 12.9.8 and assume that the additive invariance conditions of the value function hold. Let the function V be the value function and the control law g^* be the optimal control law as determined by the Dynamic Programming Procedure 12.6.2. Then,*

$$V(t, .) : X \to \mathbb{R}_+,\ \ g^*(t, .) : X \to U, \tag{12.36}$$
$$\forall\, t \in T,\ V(t, .) \in V_s; \quad \forall\, r \in T \backslash \{t_1\},\ g^*(r, .) \in G_s. \tag{12.37}$$

Proof. The dynamic programming procedure is used. By condition 12.9.8.(6), $V(t_1, .) = b_1(.) \in V_s$. Suppose that $V(s, .) \in V_s$ for $s = t+1,\ t+2, \ldots, t_1$. It will be proven that then $V(t, .) \in V_s$. By the dynamic programming procedure,

$$V(t, x_V) = \inf_{u \in U}\ \{b(x_V, u) + E[V(t+1, f(t, x_V, u, v(t)))|F^{x_V, u}]\}.$$

By the induction assumption, $V(t+1, .) \in V_s$. By condition 12.9.8.(1),

$$E[V(t+1, f(t, x_V, u, v(t)))|F^{x_V, u}] \in H_s.$$

This, condition 12.9.8.(5) that $b \in H_s$, and condition 12.9.8.(2) on the additive closure of H_s, imply that,

$$b(x_v, u) + E[V(t+1, f(t, x_v, u, v(t)))|F^{x_v, u}] \in H_s.$$

Then the conditions 12.9.8.(3+4) imply that,

$$\begin{aligned} V(t,x_V) &= H(x_V,u^*) = \min_{u\in U} H(x_V,u) \\ &= \inf_{u\in U} \, [b(x_V,u) + E[V(t+1,f(t,x_V,u,v(t)))|F^{x_V,u}] \in V_s, \\ g^*(t,x_V) &= \text{argmin}_{u\in U} \, H(x_V,u) = u^* = \\ &= \text{argmin}_{u\in U} \, [b(x_V,u) + E[V(t+1,f(t,x_V,u,v(t)))|F^{x_V,u}] \in G_s. \end{aligned}$$

The proof then completes by induction. □

Proposition 12.9.10. *Consider the Gaussian optimal stochastic control problem of Problem 12.7.1. Define the function sets,*

$$\begin{aligned} & (V_s, H_s, G_s), \\ V_s &\subseteq \left\{ V : X \to \mathbb{R}_+ |\ \exists\ Q_V \in \mathbb{R}^{n\times n}_{pds},\ V(x_V) = x_V^T Q_V x_V + r, \quad r \in \mathbb{R}_+ \right\}, \\ H_s &\subseteq \left\{ \begin{array}{l} H : X \times U \to \mathbb{R}_+ |\ \exists\ Q_H \in \mathbb{R}^{n\times n}_{pds}, \\ H(x_v) = \begin{pmatrix} x_v \\ u \end{pmatrix}^T Q_H \begin{pmatrix} x_v \\ u \end{pmatrix} + r_H, \quad r_H \in \mathbb{R} \end{array} \right\}, \\ G_s &\subseteq \{ g : X \to \mathbb{R}_+ | \exists\ F \in \mathbb{R}^{m\times n},\ g(x_v) = F x_v \}. \end{aligned}$$

Assume that, in the infimization condition of Def. 12.9.8.(3) of the function in H_s to be minimized, it is true that $Q_H|_U \succ 0$. Then the triple $(V_s,\ H_s,\ G_s)$ satisfies the additive invariance conditions. From Theorem 12.9.9 then follows that the value function of this problem satisfies that, for all $t \in T$, $V(t,.) \in V_s$ and that the optimal control law satisfies $g^*(t,.) \in G_s$.

Proof. That the conditions 12.9.8.(5) and (6) hold follows from Problem 12.7.1. That condition 12.9.8.(2) holds follows directly from the definition of the function set H_s. That the conditions 12.9.8.(1, 3, 4) hold follows from the proof of Theorem 12.7.5. □

The reader is faced with the problem of how to find an appropriate set of value functions. The following procedure may help with this problem.

Procedure 12.9.11 *Consider the optimal stochastic control problem of Def. 12.2.1. Construct a subset of value functions which is invariant by application of the following steps.*

1. *Choose the initial set of value functions $V_s(0)$ such that $V(t_1,x_v) = b_1(x_v) \in V_s(0)$.*
2. *Using the dynamic programming procedure 12.6.2 construct the value function $V(t_1-1,.) : X \to \mathbb{R}_+$. Choose a subset of value functions such that $V_s(t_1,.),\ V_s(t_1-1,.) \in V_s(1)$.*
3. *Construct a sequence of subsets of value functions such that,*

$$\begin{aligned} & \forall\ k \in T,\ V_s(k+1) \subseteq V_s(k+2) \subseteq \ldots \subseteq V_s(t_1-1), \\ & V_s(k) \subseteq V_s(k+1),\ V(k,.) \in V_s(k), \end{aligned}$$

similarly to the previous step. Stop construction of the sequence if there exists a $k^* \in \mathbb{Z}_+$ such that $V_s(k^*+1) = V_s(k^*)$. Then output $V_s(k^*)$.

A Gambling Problem with a Logarithmic Reward

A gambler at a casino is faced with the following control problem. The gambler has a current capital. At each time moment he or she can offer a bid which is less than the current capital. There is then a lottery procedure which determines whether the bid is successful or not. If the bid is successful then the casino returns an amount which equals twice the bid. If the bid is not successful then the entire bid is lost. What is the optimal control law for a bid if the gambler has as control objective to maximize her or his profit at the terminal time?

The solution of this example shows the invariance of a particular subset of value functions with a finite-valued conditional probability distribution function.

Problem 12.9.12. *Gambling problem with a logarithmic reward function.* Consider the gambling problem informally described above. A mathematical model is described by: (Ω, F, P) a probability space, $T = \{0, 1, \ldots, t_1\}$ the time index set, $x : \Omega \times T \to \mathbb{R}_+$, $x(t)$ denotes the capital of the gambler at time $t \in T$, $u : \Omega \times T \to \mathbb{R}_+$, $u(t)$ denotes the bid of the gambler made at time $t \in T$, with the constraint that $u(t) \in (0, x(t))$, $v : \Omega \times T \to \{0, 1\}$, $v(t)$ denotes whether the bid is successful denoted by $v(t) = 1$ with probability p and not successful denoted by $v(t) = 0$ with probability $(1-p)$. Assume that a bid is successful with probability $p \in (1/2, 1)$. This assumption is not realistic in general. The control system is then described by,

$$\begin{aligned} x(t+1) &= [x(t) - u(t) + 2u(t)]I_{\{v(t)=1\}} + [x(t) - u(t)]I_{\{v(t)=0\}}, \\ x(0) &= x_0 \in (0, \infty). \end{aligned}$$

The reward to be maximized is the expectation of the logarithm of the capital at the terminal time,

$$J(g) = E[\ln(x^g(t_1))], \;\; J^* = \sup_{g \in G} J(g).$$

Note that the cost rate is zero, $b(t, x_V, u_V) = 0$ for all $(x_V, u_V) \in \mathbb{R}_+^2$.

Proposition 12.9.13. *Consider Problem 12.9.12. The optimal control law $g^* \in G$ for the bid and the value function are,*

$$\begin{aligned} g^*(x) &= 2(p - 1/2)x, \\ V(t, x_V) &= \ln(x_V) + (t_1 - t)c(p), \\ c(p) &= p\ln(p) + (1-p)\ln(1-p) + \ln(2). \end{aligned}$$

The optimal control law is a linear function of the state and it is optimal over the set of all Borel measurable nonlinear control laws. In addition, it is a myopic control law, see Def. 12.6.8.

Proof. The dynamic programming procedure is carried out. The value function at the terminal time equals the terminal cost, $V(t_1, x_V) = \ln(x_V)$. Then,

$$\begin{aligned}
V(t_1-1,x_V) &= \sup_{u_V \in (0,x_V)} E[V(t_1,x(t_1))|F^{x_V,u_V}] \\
&= \sup E\left[\ln((x_V-u_V)+2u_V)I_{\{v(t)=1\}}+\right. \\
&\quad \left. +\ln(x_V-u_V)I_{\{v(t)=0\}}|F^{x_V,u_V}\right] \\
&= \sup\left[p\ln(x_V+u_V)+(1-p)\ln(x_V-u_V)\right], \\
H(x_V,u_V) &= p\ln(x_V+u_V)+(1-p)\ln(x_V-u_V), \\
0 &= \frac{\partial H(x_V,u_V)}{\partial u_V} = \frac{p}{x_V+u_V}-\frac{1-p}{x_V-u_V}, \\
&\Rightarrow p(x_V-u_V)=(1-p)(x_V+u_V),\ u^*=2(p-1/2)x_V \in (0,x_V), \\
H(x_V,u^*) &= \ln(x_V)+c(p), \\
\frac{\partial^2 H(x_V,u_V)}{\partial u_V^2} &= \frac{-p}{(x_V+u_V)^2}+\frac{-(1-p)}{(x_V-u_V)^2}<0, \\
&\quad \forall\, x_V \in (0,\infty),\ \forall\, u_V \in U(t,x_V)=(0,x_V), \\
V(t_1-1,x_V) &= \ln(x_V)+c(t,p).
\end{aligned}$$

Note that $\partial^2 H(x_V,u_V)/\partial x_v^2 < 0$ implies that the function $H(x_V,u_V)$ is strictly concave in u_V. Note further that the term $\ln(x_V)$ occurs both in $V(t_1,x_V)$ and in $\ln(V(t_1-1,x_V))$. The result then follows by using an induction argument. □

12.10 Relation of Optimal Control Law and State

A researcher of control theory is expected to be interested in the structure of the control laws. In particular, on which components of the state does the optimal control law depend? This question can be answered using results from realization theory. This section establishes a relation between optimal control and realization theory.

The principle of the proposed approach is to consider the function from the input to the controlled output via the stochastic control system. If the stochastic control system with that input and that controlled output is a minimal realization then the optimal control law may depend on all state components. However, if the stochastic control system with that input and that output is not a minimal stochastic realization then there can be made a transformation of the state set corresponding to the analogon of the Kalman decomposition of linear systems. Then it can be proven that the optimal control law will depend only on the observable and on the controllable part of the control system.

Theorem 12.10.1. *Consider a recursive stochastic control system representation with the particular decomposition,*

$$
\begin{aligned}
x(t+1) &= \begin{pmatrix} f_1((x_1(t),0),(u_1(t),0),(v_1(t),0)) \\ f_2((x_1(t),x_2(t)),(u_1(t),u_2(t)),(v_1(t),v_2(t))) \end{pmatrix}, \\
& \quad x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} \in X_1 \times X_2,\ u(t) = \begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix} \in U_1 \times U_2, \\
J(g) &= E\left[\sum_{s=0}^{t_1-1} b((x_1^g(s),0),(u_1^g(s),0)) + b_1((x_1(t_1),0))\right].
\end{aligned}
$$

It is assumed that the stochastic processes v_1 and v_2 each are sequences of independent random variables. In addition, the processes v_1 and v_2 are independent. Assume in addition that the optimal control law exists.

Then the optimal control law and the value function satisfy,

$$
\begin{aligned}
g^*(t,(x_{V,1},x_{V,2})) &= g^*(t,(x_{V,1},0)),\ \forall\, t \in T\backslash\{t_1\},\ \forall\, (x_{V,1},x_{V,2}) \in X_1 \times X_2, \\
V(t,(x_{V,1},x_{V,2})) &= V(t,(x_{V,1},0)),\ \forall\, t \in T\backslash\{t_1\},\ \forall\, (x_{V,1},x_{V,2}) \in X_1 \times X_2.
\end{aligned}
$$

Thus, both the value function and the optimal control law depend only on the first state component!

Proof. The dynamic programing procedure is applied. Note that at the terminal time, $V(t_1,x_v) = b_1((x_{v,1},0))$. Assume that for all $s = t+1, t+2, \ldots, t_1$, the induction assumption holds that $V(s,(x_{v,1},x_{v,2})) = V(s,(x_{v,1},0))$ for all $(x_{v,1},x_{v,2}) \in X_1 \times X_2$. The dynamic programming procedure for time $t \in T$ then becomes,

$$
\begin{aligned}
V(t,(x_{V,1},x_{V,2})) &= \inf_{(u_1,u_2)\in U_1\times U_2} [b(t,x_V,u) + E[V(t+1,x(t+1))|F^{x_V,u}]] \\
&= \inf[b(t,(x_{V,1},0),(u_1,0)) + \\
&\quad +E[V(t+1,(f_1(t,(x_{V,1},0),(u_1,0),(v_1(t),0), \\
&\quad\quad f_2(t,(x_{V,1},x_{V,2}),(u_1,u_2)),(v_1(t),0))|F^{x_V,u}] \\
&= \inf[b(s,(x_{V,1},0),(u_1,0)) + \\
&\quad\quad +E[V(t+1,(f_1(t,(x_{V,1},0),(u_1,0),(v_1(t),0),0)))|F^{x_V,u}] \\
&= \inf_{(u_1,u_2)\in U_1\times U_2} H(s,(x_{V,1},0),(u_1,0)) = V(t,(x_{V,1},0)),\ \text{where,} \\
g^*(t,(x_{V,1},x_{V,2})) &= (u_1^*,u_2^*) = (u_1^*,0) = g^*(t,(x_{V,1},0)).
\end{aligned}
$$

The result then follows by induction. □

Corollary 12.10.2. *Consider a Gaussian stochastic control system with a controlled output. Suppose that the linear system from the input u to the controlled output z is not a minimal realization of its response function. Construct then the Kalman decomposition of this linear system from the input to the controlled output of the form, again denoted by the state x and the input u,*

$$
\begin{aligned}
x(t+1) &= \begin{pmatrix} A_{11} & 0 & A_{13} & 0 \\ A_{21} & A_{22} & A_{23} & A_{24} \\ 0 & 0 & A_{33} & 0 \\ 0 & 0 & A_{43} & A_{44} \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{pmatrix} + \begin{pmatrix} B_1 \\ B_2 \\ 0 \\ 0 \end{pmatrix} u(t) + \begin{pmatrix} M_1 \\ M_2 \\ M_3 \\ M_4 \end{pmatrix} v(t), \\
z(t) &= \begin{pmatrix} C_{z,1} & 0 & C_{z,3} & 0 \end{pmatrix} x(t) + D_z u(t).
\end{aligned}
$$

(a)The following reduced-order Gaussian stochastic control system describes the same impulse response function from the input u to the controlled output z and from the noise process v to the controlled output z as the system considered above.

$$\begin{aligned}
\begin{pmatrix} x_1(t+1) \\ x_3(t+1) \end{pmatrix} &= \begin{pmatrix} A_{11} & A_{13} \\ 0 & A_{33} \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_3(t) \end{pmatrix} + \begin{pmatrix} B_1 \\ 0 \end{pmatrix} u(t) + \begin{pmatrix} M_1 \\ M_3 \end{pmatrix} v(t) \\
&= A_{(1,3)} x_{(1,3)}(t) + B_{(1,3)} u(t) + M_{(1,3)} v(t), \\
z(t) &= \begin{pmatrix} C_{z,1} & C_{z,3} \end{pmatrix} x_{1,3}(t) + D_z u(t) = C_{z,(1,3)} x_{(1,3)}(t) + D_z u(t), \\
x_{(1,3)}(t) &= \begin{pmatrix} x_1(t) \\ x_3(t) \end{pmatrix}.
\end{aligned}$$

The state component x_1 is affected by the input u and affects the controlled variable z. However, the state component x_3 influences the controlled output z, is affected by the noise process v, but is not influenced by the input.

(b)Then this stochastic control problem satisfies the conditions of Theorem 12.10.1. From that theorem then follows that the optimal control law and the value function satisfy,

$$\begin{aligned}
g^*(t,(x_{v,1},x_{v,2},x_{v,3},x_{v,4})) &= g^*(t,(x_{v,1},0,x_{v,3},0)), \\
V(t,(x_{v,1},x_{v,2},x_{v,3},x_{v,4})) &= V(t,(x_{v,1},0,x_{v,3},0)), \\
&\forall\, t \in T,\ \forall\ (x_{v,1},x_{v,2},x_{v,3},x_{v,4}) \in X_1 \times X_2 \times X_3 \times X_4.
\end{aligned}$$

The state component x_3 is not influenced by the input but it affects the controlled output. Hence it shows up in the value function and in the cost function. Therefore this component cannot be deleted from the system representation according to the formulation considered.

Proof. (a) It will be shown that the impulse response function from u to z of the considered system and of the reduced-order system are identical. It is to be shown that $C_z A^k B = C_{z,1} A_{11}^k B_1$ for all $k \in \mathbb{Z}_+$.

Note that,

$$CB = C_{z,1}B_1,\ C_z AB = \begin{pmatrix} C_{z,1} & 0 & C_{z,3} & 0 \end{pmatrix} \begin{pmatrix} A_{11}B_1 \\ A_{21}B_1 + A_{22}B_2 \\ 0 \\ 0 \end{pmatrix} = C_{z,1}A_{11}B_1,$$

$$C_z A^k B = C \begin{pmatrix} A_{11}^k B_1 \\ * \\ 0 \\ 0 \end{pmatrix} = C_{z,1} A_{11}^k B_1,\ \forall\, k \in \mathbb{Z}_+.$$

Similarly one proves that $C_z A^k M = C_{z,(1,3)} A_{(1,3)}^k M_{(1,3)}$.

(b) This follows directly from the indicated theorem. □

12.11 Dynamic Programming for Multiplicative Cost Functions

An explicit expression of both the value function and of the optimal control law can also be obtained in case the system is a Gaussian control system and the cost function is the expectation of an exponential cost function with a quadratic cost rate. The fact that explicit functions can be obtained was rather surprising initially. It can be explained by analogy with a result in physics.

In case the cost function is of the form of an expectation of an exponential cost, the cost function is no longer additive but multiplicative, hence a different dynamic programming procedure is needed. The approach below can be generalized to many multiplicative cost function. Below attention is restricted to an exponential cost function.

Problem 12.11.1. *The optimal stochastic control problem with an exponential cost function.* Consider the recursive state-observed stochastic control system,

$$x(t+1) = f(t,x(t),u(t),v(t)), x(0) = x_0,$$

on $T = \{0,1,\ldots,t_1\}$ as defined in Problem 12.2.1, the past-state information structure, and the set G of past-state control laws. Consider the exponential cost function,

$$J(g) = E\left[\exp\left(b_1(x^g(t_1)) + \sum_{s=0}^{t_1-1} b(s,x^g(s),u^g(s))\right)\right], \; J : G \to \mathbb{R}.$$

The optimal stochastic control problem is to solve $\inf_{g\in G} J(g)$.

Below the dynamic programming method for exponential cost functions is presented to solve the above problem.

Procedure 12.11.2 The dynamic programming procedure for an exponential cost function. *Consider the optimal stochastic control problem of 12.11.1.*

1. *Initialization. Let $V(t_1,.) : X \to \mathbb{R}$ be defined by $V(t_1,x_V) = \exp(b_1(x_V))$.*
2. *For $t = t_1-1, t_1-2, \ldots, 0$ do:*
   - *Determine $V(t,.) : X \to \mathbb{R}$*

     $$V(t,x_V) = \inf_{u\in U(t,x_V)} \; \{ \; \exp(b(t,x_V,u))E[V(t+1,f(t,x_V,u,v(t)))|F^{x_V,u}] \; \}.$$

   - *If, for all $x_V \in X$, the infimum in the above infimization is attained, or, equivalently, if for all $x_V \in X$ there exists a $u^* \in U(t,x_V)$ such that,*

     $$\begin{aligned}&\exp(b(t,x_V,u^*))E[V(t+1,f(t,x_V,u^*,v(t)))|F^{x_V,u}] \\ &= \inf_{u\in U(t,x_V)} \{\exp(b(t,x_V,u))E[V(t+1,f(t,x_V,u,v(t)))|F^{x_V,u}]\},\end{aligned}$$

     then define $g^*(t,.) : X \to U$ as $g^*(t,x_V) = u^*$.
3. *Check if for all $t \in T\backslash\{t_1\}$ the function $g^*(t,.)$ is a measurable function. If so proceed else stop.*

4. *Output (V, g^*) with the value function V and the optimal control law g^*.*

It can be proven that Procedure 12.11.2 produces the value function and the optimal control law.

Linear-Exponential-Quadratic-Gaussian Stochastic Control

Problem 12.11.3. The *linear-exponential-quadratic-Gaussian optimal stochastic control problem (LEQG-CO-FH)*. Consider the state equation of a Gaussian stochastic control system representation,

$$\begin{aligned}
x(t+1) &= A(t)x(t) + B(t)u(t) + M(t)v(t),\ x(0) = x_0,\ v(t) \in G(0,I),\\
z(t) &= C_z(t)x(t) + D_z(t)u(t),\ \forall\, t \in T(0:t_1-1),\\
z(t_1) &= C_z(t_1)x(t_1),\\
&\quad n_z \geq n_u,\ \forall\, t \in T,\ \text{rank}(D_z(t)) = n_u \ \Rightarrow\ D_z(t)^T D_z(t) \succ 0.
\end{aligned}$$

Consider the past-state information structure and the set G of past-state control laws. Define the exponential-quadratic cost function $J : G \rightarrow \mathbb{R}$,

$$\begin{aligned}
J(g) &= E\left[c\exp\left(\frac{1}{2}c\left[x^g(t_1)^T C_z(t_1)^T C_z(t_1)x^g(t_1) + \sum_{s=0}^{t_1-1}\begin{pmatrix} x^g(s)\\ u^g(s)\end{pmatrix}^T L(s)\begin{pmatrix} x^g(s)\\ u^g(s)\end{pmatrix}\right]\right)\right],\\
Q_{cr}(t) &= \begin{pmatrix} C_z(t)^T C_z(t) & C_z(t)^T D_z(t)\\ D_z(t)^T C_z(t) & D_z(t)^T D_z(t)\end{pmatrix} \in \mathbb{R}_{pds}^{(n_x+n_u)\times(n_x+n_u)},\\
&\quad D_z(t)^T D_z(t) \succ 0,\ \forall t \in T,\ Q_{cr} : T \rightarrow \mathbb{R}^{(n_x+n_u)\times(n_x+n_u)},
\end{aligned}$$

for a value $c \in \mathbb{R}$, $c \neq 0$. Both the case $c > 0$ and $c < 0$ are of interest and the closed-loop systems show different properties for those two cases. The optimal stochastic control problem is then to solve,

$$\inf_{g \in G} J(g).$$

This problem is known in the literature as the complete observations case of the *linear-exponential-quadratic-Gaussian* (LEQG) stochastic control problem.

Theorem 12.11.4. *Consider the optimal stochastic control problem of 12.11.3. Assume that:*

1. for all $t \in T$, $[I - cM(t)^T Q_c(t+1)M(t)] \succ 0$;
2. for all $t \in T$, $D_z(t)^T D_z(t) + B(t)^T Q_c(t)B(t) \succ 0$.

Define the backward recursions,

$$\begin{aligned}
&Q_c : T \rightarrow \mathbb{R}_{pds}^{n_x\times n_x},\ Q_r : T_1 \rightarrow \mathbb{R}_{pds}^{n_x\times n_x},\\
Q_c(t_1) &= C_z(t_1)^T C_z(t_1),\\
Q_r(t) &= Q_c(t+1) + cQ_c(t+1)M(t)\left[I - cM(t)^T Q_c(t+1)M(t)\right]^{-1} M(t)^T Q_c(t+1),\\
Q_c(t) &= A(t)^T Q_r(t)A(t) + C_z(t)^T C_z(t) - [C_z(t)^T D_z(t) + A(t)^T Q_r(t)B(t)] \times\\
&\quad \times [D_z(t)^T D_z(t) + B(t)^T Q_r(t)B(t)]^{-1}[C_z(t)^T D_z(t) + A(t)^T Q_r(t)B(t)]^T;\\
\overline{Q_v}(t)^{-1} &= I - cM(t)^T Q_c(t+1)M(t).
\end{aligned}$$

(a)The optimal control law is given by, $g^*_{LEQG,co,fh} : T \times X \rightarrow U$,

$$\begin{aligned} & g^*_{LEQG,co,fh}(t,x_V) \\ & = -[D_z(t)^T D_z(t) + B(t)^T Q_r(t) B(t)]^{-1} [C_z(t)^T D_z(t) + A(t)^T Q_r(t) B(t)]^T x_V. \end{aligned}$$

(b)The value function and the value are given by,

$$\begin{aligned} V(t,x_v) &= r(t)c \ \exp(\frac{1}{2} c x_V^T Q_c(t) x_V), \ r : T \rightarrow \mathbb{R}, \\ r(t) &= r(t+1) \left(\frac{\det(\overline{Q_v}(t))}{\det(Q_v(t))} \right)^{\frac{1}{2}}, \ r(t_1) = 1, \\ J^* &= E[V(0,x_0)]. \end{aligned}$$

(c)$Q_r(t)^{-1} = Q_c(t+1)^{-1} - cM(t) Q_c(t+1) M(t)^T$.

Proof. The dynamic programming procedure is applied. Note that,

$$\begin{aligned} Q_c(t_1) = C_z(t_1)^T C_z(t_1), \ r(t_1) = 1, \ \Rightarrow \\ V(t_1,x_V) = c\exp(\frac{1}{2} c x_V^T C_z(t_1)^T C_z(t_1) x_V) = c r(t_1) \exp(\frac{1}{2} c x_V^T Q_c(t_1) x_V). \end{aligned}$$

Suppose that for $s = t+1, \ldots, t_1$, $Q_c(s) = Q_c(s)^T$, $r(s) > 0$, and,

$$V(s,x_V) = c r(s) \exp(\frac{1}{2} c x_V^T Q_c(s) x_V).$$

It will be proven that then $Q_c(t) = Q_c(t)^T$, $r(t) > 0$, and,

$$\forall \, x_V \in X, \ V(t,x_V) = c r(t) \exp(\frac{1}{2} c x_V^T Q_c(t) x_V).$$

Procedure 12.11.2 prescribes to solve,

$$\begin{aligned} & \inf_{u \in U(t,x_V)} \left\{ c \exp\left(\frac{1}{2} c \begin{pmatrix} x_V \\ u \end{pmatrix}^T Q_{cr}(t) \begin{pmatrix} x_V \\ u \end{pmatrix} \right) \times \right. \\ & \left. \times E[V(t+1, f(t,x_V,u,v(t))) | F^{x_V,u}] \right\}. \end{aligned}$$

The conditional expectation may be calculated as follows. Note that $x(t+1)$ is conditionally Gaussian given $F^{x_V,u}$. Then

$$
\begin{aligned}
& E[V(t+1, f(t,x_V,u,v(t)))|F^{x_v,u}] \\
& = E[cr(t+1)\exp(\frac{1}{2}c(A(t)x_V + B(t)u + M(t)v(t))^T Q_c(t+1) \times \\
& \quad (A(t)x_V + B(t)u + M(t)v(t)))|F^{x_V,u}] \\
& = E[cr(t+1)\exp(\frac{1}{2}c \begin{pmatrix} v(t) \\ A(t)x_V + B(t)u \end{pmatrix}^T \times \\
& \quad \times \begin{pmatrix} M(t)^T Q_c(t+1) M(t) & M(t)^T Q_c(t+1) \\ Q_c(t+1) M(t) & Q_c(t+1) \end{pmatrix} \begin{pmatrix} v(t) \\ A(t)x_V + B(t)u \end{pmatrix})|F^{x_v,u}] \\
& = cr(t+1) \left(\frac{\det(\overline{Q_v}(t))}{\det(Q_v(t))} \right)^{\frac{1}{2}} \times \\
& \quad \times \exp(\frac{1}{2}c\ (A(t)x_V + B(t)u)^T Q_r(t)(A(t)x_V + B(t)u)), \\
& \quad \text{by Proposition 19.4.8 and the assumptions (1) and (2),} \\
& = cr(t)\exp\left(\frac{1}{2}c \begin{pmatrix} x_V \\ u \end{pmatrix}^T \begin{pmatrix} A(t)^T Q_r(t) A(t) & A(t)^T Q_r(t) B(t) \\ B(t)^T Q_r(t) A(t) & B(t)^T Q_r(t) B(t) \end{pmatrix} \begin{pmatrix} x_V \\ u \end{pmatrix} \right).
\end{aligned}
$$

Denote,

$$
\begin{aligned}
H(t) &= \begin{pmatrix} H_{11}(t) & H_{12}(t) \\ H_{12}(t)^T & H_{22}(t) \end{pmatrix} \\
&= \begin{pmatrix} A(t)^T Q_r(t) A(t) + C_z(t)^T C_z(t) & A(t)^T Q_r(t) B(t) + C_z(t)^T D_z(t) \\ B(t)^T Q_r(t) A(t) + D_z(t)^T C_z(t) & B(t)^T Q_r(t) B(t) + D_z(t)^T D_z(t) \end{pmatrix}.
\end{aligned}
$$

Note that by Assumption (2),

$$
\begin{aligned}
H_{22}(t) &= H_{22}(t)^T = D_z(t)^T D_z(t) + B(t)^T Q_r(t) B(t) \succeq D_z(t)^T D_z(t) \succ 0; \text{ hence,} \\
&\inf_{u \in U(t,x_v)} \left\{ cr(t) \exp(\frac{1}{2} c \begin{pmatrix} x_V \\ u \end{pmatrix}^T Q_{cr}(t) \begin{pmatrix} x_V \\ u \end{pmatrix} \times \right. \\
&\quad \left. \exp(\left(\frac{1}{2} c \begin{pmatrix} x_V \\ u \end{pmatrix}^T \begin{pmatrix} A(t)^T Q_r(t) A(t) & B(t)^T Q_r(t) B(t) \\ B(t)^T Q_r(t) A(t) & B(t)^T Q_r(t) B(t) \end{pmatrix} \begin{pmatrix} x_V \\ u \end{pmatrix} \right)) \right\} \\
&= \inf_{u \in U(t,x_v)} \left\{ cr(t) \exp(\frac{1}{2} c \begin{pmatrix} x_V \\ u \end{pmatrix}^T H(t) \begin{pmatrix} x_V \\ u \end{pmatrix}) \right\} \\
&= \inf_{u \in U(t,x_V)} \left\{ cr(t) \exp \left(\frac{1}{2} c \begin{pmatrix} x_V \\ u + H_{22}(t)^{-1} H_{12}(t)^T x_V \end{pmatrix}^T \times \right. \right. \\
&\quad \times \begin{pmatrix} H_{11}(t) - H_{12}(t) H_{22}(t)^{-1} H_{12}(t)^T & 0 \\ 0 & H_{22}(t) \end{pmatrix} \times \\
&\quad \left. \left. \times \begin{pmatrix} x_V \\ u + H_{22}(t)^{-1} H_{12}(t)^T x_V \end{pmatrix} \right) \right\} \\
&= cr(t) \exp(\frac{1}{2} c x_V^T \left[H_{11}(t) - H_{12}(t) H_{22}(t)^{-1} H_{12}(t)^T \right] x_V) \\
&= cr(t) \exp(\frac{1}{2} c \, x_V^T Q_c(t) x_V), \text{ if,} \\
g^*(t,x_v) &= u^* = -H_{22}(t)^{-1} H_{12}^T(t) x_V, \\
Q_c(t) &= H_{11}(t) - H_{12}(t) H_{22}(t)^{-1} H_{12}(t)^T.
\end{aligned}
$$

The formula for $Q_r(t)^{-1}$ follows from that of $Q_r(t)$ by the matrix inversion lemma 17.4.28. □

One can formulate *multiplicative invariance conditions* for the value function analogous to those of the additive invariance conditions formulated in Def. 12.9.8.

12.12 Stochastic Control Problems of Economics and of Finance

The research areas of economics and of finance have several problems of stochastic control. It was decided not to include in the book several of the well known examples of such stochastic control problems except for a few comments formulated below.

Portfolio selection is an optimal stochastic control problem motivated by a problem of investors. For the continuous-time case one obtains directly from optimal control theory a control law for the optimal stochastic control problem of portfolio selection. However, for discrete-time portfolio selection problem the optimal control law has to be computed numerically. This requires a formulation and examples with numerical computations. There is insufficient space in this book for such computations.

12.13 Control via System Approximation

The reader finds in this section a brief description on how to obtain an approximate control law for general optimal stochastic control problems.

Consider the stochastic control system with an additive cost function of Problem 12.2.1 with the equations,

$$\begin{aligned} x(t+1) &= f(t,x(t),u(t),v(t)),\ x(0)=x_0, \\ J(g) &= E\left[\sum_{s=0}^{t_1-1} b(s,x^g(s),u^g(s)) + b_1(x^g(t_1))\right], \\ &\inf_{g\in G} J(g). \end{aligned}$$

In general it may not be possible to analytically determine the value function either because the calculation of the conditional expectation in Procedure 12.6.2 does not yield a well known expression or because the infimization can not be carried out analytically. The reader is then faced what the problem how to determine an optimal control law.

As in many problems of mathematics, one searches for an approximation procedure. Research issues to be investigated are then: How to discrete the stochastic control system? How to compute the optimal control law for the discretized stochastic control system and the corresponding cost? How to transform the optimal control law for the discretized optimal stochastic control problem to a control law for the original control system? How to bound the performance of the computed approximate control law from the value of the optimal control problem for the original system? Can a theorem be proven for the convergence of the optimal discretized control law to the optimal control law of the original optimal control problem? The theory required for answering these questions is beyond the scope of this book. The reader is referred to the books of H.J. Kushner [43, 44, 45] and to the book of R.Z. Hasminkski [30] for the theory.

12.14 Exercises

Problem 12.14.1. Preventive maintenance. Consider a machine that over a finite time horizon may be in an operating state, $x(t)=1$, or in a failed state, $x(t)=0$. The inputs represent: $u=0$ that no maintenance is performed, $u=1$ that preventive maintenance is performed, and $u=2$ that the machine is repaired. The input space is state dependent, if $x=1$ then $U(1)=\{0,1\}$ and if $x=0$ then $U(2)=\{2\}$. Depending on the state and the control action, the state transition is as follows:

- if $x(t)=1$ and $u(t)=0$ then $x(t+1)=0$ with probability p_1;
- if $x(t)=1$ and $u(t)=1$ then $x(t+1)=0$ with probability p_2;
- if $x(t)=0$ and $u(t)=2$ then $x(t+1)=1$ almost surely.

The cost function is $c(1,0) = c_1$, $c(1,1) = c_2$ and $c(0,2) = c_3$. Suppose that the terminal cost is,

$$V(t_1,x) = \begin{cases} 5, & \text{if } x = 1, \\ 10, & \text{if } x = 0, \end{cases}$$

and that $0 < p_2 < p_1$ and $c_1 < c_2 < c_3$.

(a)Let $c_1 = 0$, $c_2 = 1$, $c_3 = 10$, $p_1 = 0.3$ and $p_2 = 0.1$. Compute the value function and the optimal control over the finite horizon $T = \{0,1,2\}$.
(b)Determine a relation between $V(t,1)$ and $V(t,0)$ for arbitrary fixed $t \in \mathbb{Z}_+$ such that preventive maintenance is better than no maintenance.

Problem 12.14.2. Formulate and prove the comparison principle,
analogous to 12.6.5. for the optimal stochastic control system with exponential
cost 12.11.1.

12.15 Further Reading

History. Optimal stochastic control with complete observations on a finite horizon by dynamic programming was much stimulated by the publications of R.E. Bellman [5, 6]. It is not known to the author whether there are substantial earlier sources for dynamic programming than the books and papers of Bellman which have appeared after 1945. The expression of *dynamic programming* was motivated by the term *linear programming* which had become in use since the early part of the 20th century. The calculations of linear programming were stimulated by the availability of electronic computers in the period after 1940.

Researchers of economics including those active in the research areas of optimization, decision theory, games, and dynamics games, stimulated and interacted with researchers of control theory. The book of J. von Neumann and O. Morgenstern on games has stimulated much interest, [53]. Decision theory and games are also to be found in the books by D. Blackwell and M.A. Girshick, [16], and that of R.D. Luce and H. Raiffa, [49].

In the operations research literature, dynamic programming for stochastic control systems was further developed. An early book is that of R. Howard, [34].

History of LQG optimal stochastic control with complete observations. The original references are papers of P.D. Joseph and J.T. Tou, [36], and of T.L. Gunkel III and G.F. Franklin, [28]. The special issue of the journal *IEEE Transactions on Automatic Control* of December 1971 has many papers on this problem, including for stochastic control with partial observations.

Books on optimal stochastic control with complete observations. In control theory, the early books are those of H. Kushner, [41, 42], of J.S. Meditch, [50], and that of K.J. Aström, [2]. See for deterministic and stochastic control of linear systems the book of H. Kwakernaak and R. Sivan, [46].

At the level of this publication are the books [39], with the new publication [40], and [9], with its later editions [10, 11].

Other books on optimal stochastic control of discrete-time stochastic control systems are [12, 22, 31, 32, 62]. From the Russian school the book of E.B. Dynkin and A.A. Yushkevich is mentioned [26].

For a treatment of the measure theoretic properties of dynamic programming see the book [12], the Ph.D. thesis of S.E. Shreve, [58], and the references provided there. A survey of measurable selection theorems is [66].

A rather general model formulated using the concept of P-essential infimum was developed by R.Rishel, M.H.A. Davis and P. Varaiya, and C. Striebel [12, 21, 62]. The approach of [62] is applicable to additive and to multiplicative cost functions.

Books on optimal stochastic control with dynamic programming written from the viewpoint of operations research are [24, 54, 56, 55, 70]. A reference on dynamic programming for Markov chains is [3].

Applications of stochastic control to control of communication networks are treated in [23, 37, 68, 69]. For applications to manufacturing see [7]. Optimal stopping problems are treated in [57].

Control problems. The references of the examples used: Example 12.1.1 on course keeping of a ship, is based on [65]. The example of control of an automobile suspension is a brief summary of the paper [4]. An example of an optimal stochastic control problem inspired by interplanetary guidance is presented in [64].

Invariance of a subset of value functions. This issue is formulated in the Ph.D. thesis of P.R. Kumar, see [38]. This issue is discussed in the book [39, p. 153]. See also the papers [29, 48, 67]. The gambling problem 12.9.12 with a logarithmic reward function was investigated by T.M. Cover, [20].

The relation of an optimal control and the state. This approach has probably been known for a long time but the author does not know any reference.

Optimal stochastic control with a multiplicative cost function and complete observations. The linear-exponential-quadratic-Gaussian stochastic control problem with complete observations was solved by D.H. Jacobson in [35]. According to a report, the problem was suggested by L. Zadeh. Risk-sensitive cost functions are treated in [52].

Control problems of economics and of mathematical finance. Portfolio selection on a finite horizon is treated in the book of D.P. Bertsekas [9, Sections 1.3 and 1.4]. Also see the paper [8]. In the economics literature, see K.J. Arrow, [1], and J. Mossin, [51]. The concept of certainty equivalence was proposed by H.A. Simon [60], see also H. Theil [63].

References

1. K.J. Arrow. *Aspects of the theory of risk bearing*. Yrjo Jahnsson Lecture Series. U. of Helsinki, Helsinki, Finland, 1965. 403, 468
2. K.J. Aström. *Introduction to stochastic control.* Academic Press, New York, 1970. 376, 410, 467, 522, 575, 596

3. J. Bather. Optimal decision procedures for finite markov chains. *Adv. Applied Probab.*, 5:328–339, 521–540, 541–553, 1973. 468, 525
4. S. Bellizzi, R. Bouc, F. Campillo, and E. Pardoux. Contrôle optimal semi-actif de suspension de véhicule. In A. Bensoussan and J.L. Lions, editors, *Analysis and optimization of systems*, volume 111 of *Lecture Notes in Control and Information Sciences*, pages 689–699. Springer-Verlag, Berlin, 1988. 9, 377, 468
5. R.E. Bellman. *Dynamic programming*. Princeton University Press, Princeton, 1957. 467
6. R.E. Bellman and S.E. Dreyfus. *Applied dynamic programming*. Princeton University Press, Princeton, 1962. 467
7. A. Bensoussan. Stochastic control in discrete time and applications to the theory of production. *Math. Programm. Study*, 18:43–60, 1982. 468
8. D.P. Bertsekas. Necessary and sufficient conditions for the existence of an optimal portfolio. *J. Econom. Theory*, 8:235–247, 1974. 468
9. D.P. Bertsekas. *Dynamic programming and stochastic control*. Academic Press, New York, 1976. 376, 405, 410, 439, 468, 502, 525, 526, 575, 595
10. D.P. Bertsekas. *Dynamic programming and optimal control, Volume I*. Athena Scientific, Belmont, MA, 1995. 468
11. D.P. Bertsekas. *Dynamic programming and optimal control, Volume II*. Athena Scientific, Belmont, MA, 1995. 468
12. D.P. Bertsekas and S.E. Shreve. *Stochastic optimal control: The discrete time case*. Academic Press, New York, 1978. 49, 428, 431, 468, 575, 595
13. D. Blackwell. On a class of probability spaces. In *Proc. Third Berkeley Symp. Math. Statist. Prob.*, volume 2, pages 1–6, Berkeley, CA, 1956. University of California. 49, 419
14. D. Blackwell. Discounted dynamic programming. *Ann. Math. Statist.*, 36:226–235, 1965. 428, 525, 526
15. D. Blackwell. The stochastic processes of Borel gambling and dynamic programming. *Ann. Statist.*, 4:370–374, 1976. 49, 376, 419
16. D. Blackwell and M.A. Girshick. *Theory of games and statistical decisions*. Wiley, New York, 1954. 353, 410, 467
17. Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, U.K., (with corrections) edition, 2007. 426, 636
18. R.W. Brockett. *Finite dimensional linear systems*. Wiley, New York, 1970. 217, 438
19. A. Browder. *Mathematical analysis - An introduction*. Undergraduate texts in mathematics. Springer-Verlag, New York, 1996. 30, 49, 424, 426, 475, 526, 635, 636, 677, 815
20. T.M. Cover. An algorithm for maximizing expected log investment return. *IEEE Trans. Inform. Theory*, 30:369–373, 1984. 468
21. M.H.A. Davis and P. Varaiya. Dynamic programming conditions for partially observable stochastic systems. *SIAM J. Control*, 11:226–261, 1973. 468, 575, 605
22. M.H.A. Davis and R.B. Vinter. *Stochastic modelling and control*. Chapman and Hall, London, 1985. 120, 376, 410, 468, 575, 595
23. P.R. de Waal. *Overload control of telephone exchanges*. PhD thesis, Catholic University Brabant, Tilburg, 1990. 468
24. C. Derman. *Finite state Markov decision processes*. Academic Press, New York, 1970. 468, 525
25. John C. Doyle. Guaranteed margins for LQG regulators. *IEEE Trans. Automatic Control*, 23:756–757, 1978. 440, 592, 596, 822, 824
26. E.B. Dynkin and A.A. Yushkevich. *Controlled Markov Processes*. Springer, New York, 1979. 468
27. A. Ferrante, G. Picci, and S. Pinzoni. Silverman algorithm and the structure of discrete-time stochastic systems. *Linear Algebra & its Applications*, 351–352:219–242, 2002. 439
28. T.L. Gunckel and G.F. Franklin. A general solution for linear, sampled-data control. *Trans. ASME, J. Basic Eng., Ser. D*, 85:197–203, 1963 (June). 467
29. B. Hajek. Optimal control of two interacting service stations. *IEEE Trans. Automatic Control*, 29:491–499, 1984. 468

30. R.Z. Hasminski. *Stochastic stability of differential equations*. Sijthoff & Noordhoff, Alphen aan de Rijn, 1980. 121, 466
31. O. Hernandez-Lerma and Lasserre J.-B. *Discrete-time Markov control processes*. Springer, New York, 1995. 468, 525
32. O. Hernandez-Lerma and J.B. Lasserre. *Further topics on discrete-time Markov control processes*. Springer-Verlag, New York, 1999. 468
33. K. Hinderer. *Foundations of non-stationary dynamic programming with discrete time-parameter*. Number 33 in Lecture Notes in Operations Research and Mathematical Systems. Springer-Verlag, Berlin, 1970. 428
34. R. Howard. *Dynamic programming and Markov processes*. M.I.T. Press, Cambridge, 1960. 376, 467, 525
35. D.H. Jacobson. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Trans. Automatic Control*, 18:124–131, 1973. 468, 605
36. P.D. Joseph and J.T. Tou. On linear control theory. *AIEE Trans. (Appl. Ind.)*, 80:193–196, 1961 (Sep.). 467, 574
37. F.P. Kelly. *Reversibility and stochastic networks*. John Wiley & Sons, Chichester, 1979. 73, 169, 468
38. P.R. Kumar. *Stochastic optimal control and stochastic differential games*. PhD thesis, Sever Institute of Technology, Washington University, St. Louis, 1977. 468
39. P.R. Kumar and P. Varaiya. *Stochastic systems: Estimation, identification, and adaptive control*. Prentice Hall Inc., Englewood Cliffs, NJ, 1986. 376, 410, 468, 525, 575, 595, 596
40. P.R. Kumar and P. Varaiya. *Stochastic systems: Estimation, identification, and adaptive control*. Number 75 in Classics in Applied Mathematics. SIAM, Philadelphia, 2015. 468, 575, 595
41. H.J. Kushner. *Stochastic stability and control*. Academic Press, New York, 1967. 121, 376, 410, 467
42. H.J. Kushner. *Introduction to stochastic control*. Holt, Rinehart and Winston Inc., New York, 1971. 121, 376, 410, 467, 525
43. H.J. Kushner. *Approximation and weak convergence methods for random processes, with applications to stochastic system theory*. M.I.T. Press, Cambridge, 1984. 466
44. H.J. Kushner and P.G. Dupuis. *Numerical methods for stochastic control problems in continuous time*. Springer-Verlag, Berlin, 1992. 466
45. H.J. Kushner and P.G. Dupuis. *Numerical methods for stochastic control problems in continuous time (2nd Ed.)*. Number 24 in Applications of Mathematics. Springer, New York, 2001. 9, 376, 377, 466
46. H. Kwakernaak and R. Sivan. *Linear optimal control systems*. Wiley-Interscience, New York, 1972. 120, 376, 410, 467, 489, 593, 822, 823
47. T.L. Lai and C.Z. Wei. Extended least squares and their application to adaptive control and prediction in linear systems. *IEEE Trans. Automatic Control*, 31:898–906, 1986. 419
48. Woei Lin and P.R. Kumar. Optimal control of a queueing system with two heterogeneous servers. *IEEE Trans. Automatic Control*, 29:696–703, 1984. 468
49. R.D. Luce and H. Raiffa. *Games and decisions*. John Wiley & Sons, New York, 1957. 376, 410, 467
50. J.S. Meditch. *Stochastic optimal estimation and control*. McGrawHill, New York, 1969. 310, 467
51. J. Mossin. Optimal multi-period portfolio policies. *J. Business*, 41:215–229, 1968. 468
52. H. Nagai. Bellman equations of risk-sensitive control. *SIAM J. Control & Opt.*, 34:74–101, 1996. 468
53. J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton University Press, Princeton, NJ, 1947. 410, 467
54. M.L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, Inc., New York, 1994. 468
55. S. Ross. *Introduction to stochastic dynamic programming*. Academic Press, New York, 1983. 468

56. S.M. Ross. *Applied probability models with optimization applications*. Holden-Day, San Francisco, 1970. 468, 525
57. A. Shiryayev. *Optimal stopping rules*. Springer, New York, 1978. 468
58. S. Shreve. *Dynamic programming in complete seperable spaces*. PhD thesis, University of Illinois, Department of Mathematics, Urbana, 1977. 468
59. L.M. Silverman. Discrete Riccati equations: Alternative algorithms, asymptotic properties, and system theory interpretations. In C.T. Leondes, editor, *Control and Dynamic Systems*, pages 313–386. Academic Publishers, New York, 1976. 439, 786, 808
60. H.A. Simon. Dynamic programming under uncertainty with a quadratic criterion function. *Econometrica*, 24:74–81, 1956. 468
61. R. Strauch. Negative dynamic programming. *Ann. Math. Statist.*, 37:871–890, 1966. 428, 525
62. C. Striebel. *Optimal control of discrete time stochastic systems*, volume 110 of *Lecture Notes in Economic and Mathematical Systems*. Springer-Verlag, Berlin, 1975. 431, 468, 575, 595, 603, 605, 742
63. H. Theil. A note on certainty equivalence in dynamic planning. *Econometrica*, 25:346–349, 1957. 468
64. F. Tung and C. Striebel. A stochastic optimal control problem and its applications. *J. Math. Anal. Appl.*, 12:350–359, 1965. 468, 575
65. J. van Amerongen. Adaptive steering of ships - A model reference approach. *Automatica*, 20:3–14, 1984. 410, 468
66. D.H. Wagner. Survey of measureable selection theorems. *SIAM J. Control & Opt.*, 15:859–903, 1977. 428, 468
67. J. Walrand. A note on 'optimal control of a queueing system with two heterogenous servers'. *Syst. Control Lett.*, 4:131–134, 1984. 468
68. J. Walrand. *An introduction to queueing networks*. Prentice Hall, Englewood Cliffs, NJ, 1988. 468
69. J. Walrand and P.P. Varaiya. *High-performance communication networks*. Kaufmann, San Francisco, 1996. 9, 468
70. P. Whittle. *Optimization over time*. John Wiley, New York, 1982. 468
71. J.C. Willems, A. Kitapci, and L.M. Silverman. Singular optimal control: A geometric approach. *SIAM J. Control Optim.*, 24:323–337, 1986. 439
72. A.A. Yuskhkevich. Reduction of a controlled Markov model with incomplete data to a problem with complete information in the case of a Borel state and control spaces. *Theory Probab. Appl.*, 21:153–158, 1976. 428

Chapter 13
Stochastic Control with Complete Observations on Infinite Horizon

Abstract Optimal stochastic control problems with complete observations and on an infinite-horizon are considered. Control theory for both the average cost and the discounted cost function is treated. The dynamic programming approach is formulated as a procedure to determine the value and the value function; from the value function one can derive the optimal control law. Stochastic controllability is in general needed to prove that there exists a control law with finite average cost in case of positive cost. Special cases treated in depth are: the case of a Gaussian stochastic control system and of a finite stochastic control system.

Key words: Stochastic control. Complete observations. Infinite-horizon.

The reader may be interesting in noticing that the stochastic control system in closed-loop with the optimal control law is a stochastic realization of the stochastic control system with a finite-dimensional or finite state set satisfying the optimality criterion in terms of the cost criterion. See Theorem 13.2.15(d) for this property.

13.1 Introduction to Control on an Infinite-Horizon

The stochastic control problem on the time index set $T = \mathbb{N} = \{0, 1, \ldots\}$ is referred to as the *infinite-horizon case.* For many engineering control problems the control is expected to be operating for relatively long periods. Moreover, the effects near the end point of the time horizon are of minor interest. In such a situation, synthesis of a control law may be taken care of by solving an infinite-horizon control problem. A time-invariant control law can often be implemented for online control of an engineering system in a relatively simple way.

Consider then a time-invariant recursive stochastic control system of the form

$$x(t+1) = f(x(t), u(t), v(t)), x(0) = x_0,$$

as in Problem 12.2.1, in which the dynamics of the stochastic system f does not depending on time explicitly. Consider a set of control laws G, and construct, for any $g \in G$, the closed-loop system according to,

$$\begin{aligned} g &= (g_0, g_1(.), \ldots, g_t(.), \ldots, g_{t_1-1}(.)) \in G, \\ x^g(t+1) &= f(x^g(t), g_t(x^g(0:t)), v(t)), x^g(0) = x_0, \\ u^g(t) &= g_t(x^g(0:t)),\ x^g(0:t) = (x^g(0), x^g(1), \ldots, x^g(t)). \end{aligned}$$

In contrast with the finite-horizon case, on an infinite horizon there are two cost criteria of interest.

Definition 13.1.1. The *discounted cost function* or *discounted cost criterion* with *discount factor* $r \in (0,1)$ for the above formulated stochastic control system is defined as,

$$J_{dc}(g) = \limsup_{t_1 \to \infty} E\left[\sum_{s=0}^{t_1-1} r^s b(x^g(s), u^g(s))\right],\ J_{dc} : G \to \mathbb{R}_+ \cup \{\infty\}. \tag{13.1}$$

Note that the cost rate $b : X \times U \to \mathbb{R}_+$ is assumed not to depend explicitly on the time variable.

Definition 13.1.2. The *average cost function* or *average cost criterion* for the above formulated stochastic control system is defined by,

$$J_{ac}(g) = \limsup_{t \to \infty} \frac{1}{t} E\left[\sum_{s=0}^{t-1} b(x^g(s), u^g(s))\right],\ J_{ac} : G \to \mathbb{R}_+ \cup \{\infty\}. \tag{13.2}$$

$$\tag{13.3}$$

In the average cost case limit superior is used rather than limit because the limit may not exist. Note that in general the average cost differs from the cost function,

$$J(g) = E\left[\lim_{t \to \infty} \frac{1}{t} \sum_{s=0}^{t-1} b(x^g(s), u^g(s))\right],$$

Because the interchange of the limit operation and of the expectation operation requires a condition.

An explanation of the two cost functions on an infinite-horizon follows. The direct extension of the cost function from a finite-horizon to an infinite-horizon is the total cost,

$$J_{tc}(g) = E\left[\sum_{s=0}^{\infty} b(x^g(s), u^g(s))\right],\ J_{tc} : G \to \mathbb{R}_+.$$

Note that assumptions are required to make the infinite sum finite. If for all control laws $g \in G$ the total cost $J_{tc}(g) = \infty$ then there is no distinction between control laws in regard to their cost values hence all control laws are optimal. Therefore an assumption is required such that there exists at least one control law which has a finite cost. Note that, due to $J_{tc} : G \to \mathbb{R}_+$,

$$J_{tc}(g) < \infty \;\Rightarrow\; \sum_{s=0}^{\infty} b(x^g(s), u^g(s)) < \infty \text{ a.s.}$$

The finiteness of the infinite sum implies by [13, Thm. 2.27]
that $\lim_{s\to\infty} \; b(x^g(s), u^g(s)) = 0 \; a.s.$ If X, U are finite subsets, if $b : X \times U \to \mathbb{R}_+$, and if there exists a control $g \in G$ such that $J_{tc}(g) < \infty$ then there exists at least one tuple $(x_b, u_b) \in X \times U$ such that $b(x_b, u_b) = 0$. The system will in the long run be mostly at tuples satisfying that condition. Thus J_{tc} is in general not so useful as a cost criterion.

Which cost criterion, the discounted or average cost one, is suitable for which optimal stochastic control problem? The answer depends on the problem considered. In the discounted cost criterion, present costs are given more weight than future cost. This cost criterion is often used for control problems of economics where an euro at time $t = 0$ is more valuable than an euro at time $t = 1$. In the average cost criterion, present and future costs receive equal weight. In engineering the average cost function is often used.

Mathematical arguments often lead to consideration of the optimal stochastic control problem with the discounted cost criterion because that problem is easier to solve than that with the average cost criterion. However, the analysis of the average cost criterion is more interesting for control theory and yields more results about the properties of the control laws.

13.2 Average Cost

13.2.1 Problem Formulation

The problem formulation is phrased in terms of a recursive stochastic control system. In subsequent sections several special cases are treated.

Problem 13.2.1. *Optimal stochastic control problem for a recursive stochastic control system with complete observations on an infinite-horizon with an average cost function.*

Consider the time-invariant recursive stochastic control system with a controlled output process,

$$\begin{aligned} x(t+1) &= f(x(t), u(t), v(t)),\; x(0) = x_0, \\ z(t) &= h(x(t), u(t)),\; \forall\, t \in T\backslash\{t_1\},\; z(t_1) = h(x(t_1)), \\ & \quad T = \mathbb{N},\; X = \mathbb{R}^{n_x},\; U = \mathbb{R}^{n_u},\; Z = \mathbb{R}^{n_z},\; n_x,\; n_u,\; n_v,\; n_z \in \mathbb{Z}_+, \\ & \quad x_0 : \Omega \to \mathbb{R}^{n_x},\; v : \Omega \times T \to \mathbb{R}^{n_v}\; \{v(t),\; t \in T\} \text{ independent sequence}, \\ & \quad F^{x_0},\;\; F^v_\infty \text{ independent.} \end{aligned}$$

Define the set of time-varying control laws G_{tv}, that of time-varying Markov control laws $G_{tv,M}$, and the set of time-invariant control laws G_{ti} according to,

$$G_{tv} = \left\{ \begin{array}{l} g = (g_0, g_1(.),\ g_2(.) \ldots | \\ \forall\, t \in T,\ g_t : X^{t+1} \to U, \text{ Borel measurable} \end{array} \right\},\ g_t(x(0:t)),$$

$$G_{tv,M} = \left\{ \begin{array}{l} g = (g_0, g_1(.),\ g_2(.) \ldots | \\ \forall\, t \in T,\ g_t : X \to U, \text{ Borel measurable} \end{array} \right\},\ g_t(x(t)),$$

$$G_{ti} = \left\{ g : X \to U \text{ Borel measurable} \right\},\ g(x(t)),$$
$$\forall\, g \in G_{ti} \text{ define, } g_T = \left\{ (g, g, \ldots) \right\} \in G_{tv,M}.$$

Construct, for any $g \in G_{tv}$, the closed-loop system according to,

$$x^g(t+1) = f(x^g(t), g_t(x^g(0:t)), v(t)), x^g(0) = x_0,$$
$$u^g(t) = g_t(x^g(0:t)) = g_t((x^g(0), x^g(1), \ldots, x^g(t))).$$

Define the *average cost* on the infinite horizon as,

$$J_{ac} : G_{tv} \to \mathbb{R}_+ \cup \{\infty\},\ b : X \times U \to \mathbb{R}_+,$$
$$J_{ac}(g) = \limsup_{t_1 \to \infty} \frac{1}{t_1} E\left[\sum_{s=0}^{t_1 - 1} b(x^g(s), u^g(s))\right].$$

Denote by $G_{acf} \subseteq G$ the subset of control laws achieving a finite cost and define the optimal stochastic control problem as,

$$\inf_{g \in G_{tv,acf}} J_{ac}(g);$$
$$G_{tv,acf} = \{g \in G_{tv} |\ J_{ac}(g) < \infty\},\ G_{ti,acf} = \{g \in G_{ti} |\ J_{ac}(g) < \infty\}.$$

The reader should notice that the set of control laws is such that the control law g_t used at time $t \in T$ can depend on the entire past of the state process x^g hence g_t depends on time and is thus time-varying. See below for a discussion of this issue.

Uniqueness of the optimal control law is implied by a strict convexity condition of the cost rate.

Define the *sample-path average-cost function* and the corresponding *sample-path optimal control law* $g^* \in G_{tv}$ by the equations,

$$J_{sp,ac}(g, x_0) = \limsup_{t_1 \to \infty} \frac{1}{t_1} \sum_{s=0}^{t_1} b(x^g(s), u^g(s)) : \Omega \to \mathbb{R}_+ \cup \{\infty\},$$
$$J^*_{sp,ac}(x_0) = \mathrm{P} - essinf_{g \in G_{tv}}\ J_{sp,ac}(g, x_0) = J_{sp,ac}(g^*, x_0),\ \forall\, x_0.$$

This problem requires the concept of ergodicity and is therefore not further treated in this book. See the section *Further Reading* for references. The sample-path optimal control law is used in information theory, where one also uses the corresponding ε-optimal sample-path optimal control law.

An extension of the average cost problem is to ask for a sequence of finite valued control laws $\{g_k \in G_{ti},\ k \in \mathbb{Z}_+\}$ such that one or both of the following limit exist,

$$J^*_{ac} < \lim_{k \to \infty}\ \limsup_{t \to \infty} \frac{1}{t} E\left[\sum_{s=0}^{t-1} b(x^{g_k}(s), u^{g_k}(s))\right] < J^*_{ac} + \varepsilon,\ \ \forall\, \varepsilon \in (0, \infty),$$

$$J^*_{ac} < \lim_{k \to \infty}\ \limsup_{t \to \infty} \frac{1}{t} \sum_{s=0}^{t-1} b(x^{g_k}(s), u^{g_k}(s)) < J^*_{ac} + \varepsilon,\ a.s.\ \ \forall\, \varepsilon \in (0, \infty).$$

This particular problem is used in information theory for source coding and channel coding. The control law has to apply to a finite set of message and has to produce an input in a finit set of messages. For any ε one then asks for a suboptimal control law $g_{\varepsilon}^{*} \in G$ which achieves the second of the above inequalities.

Note that for a time-varying Markov control law $g \in G_{tv,M}$ with input $u^{g}(t) = g(t,x^{g}(t))$ the state process x^{g} of the closed-loop system is in general is a Markov process. For an arbitrary time-varying control law $g \in G_{tv}$ the state process of the closed-loop system is in general not a Markov process.

Control theory has to establish whether an optimal control law is time-varying or time-invariant. There exists an example of an optimal stochastic control problem with average cost where a time-varying control law achieves a lower cost then a time-invariant control law, see Example 13.2.19. In case of time-varying control laws, control theory has to establish that an optimal control law is a time-varying Markov control law $\overline{g} \in G_{tv,M}$, such that for all $t \in T$, $\overline{g}_t : X \rightarrow U$, $\overline{g}(x^{\overline{g}}(t))$. Control synthesis with the set of time-varying control laws requires an investigation not explored in this book.

In the remainder of this chapter attention is restricted to time-invariant control laws. Time-invariant control laws are relatively simple to implement in engineering practice.

It will be illustrated later in this chapter that the conditions of stochastic controllability and of stochastic observability are useful and necessary to establish that, for an arbitray system, there exists a control law which achieves a finite cost on the infinite horizon. The following conditions will be discussed in the special cases of optimal stochastic control problems of this chapter.

Assumption 13.2.2 *Consider the stochastic control system of Problem 13.2.1. Define the conditions on this system:*

1. *stochastic controllability of the stochastic control system, for the relation from the input to the state;*
2. *stochastic observability of the stochastic control system with the controlled output, thus for the relation from the state to the controlled output z.*

The theory of optimal stochastic control on an infinite horizon with an average cost criterion is distinguished into the cases:

1. a finite state and a finite input space;
2. a countable state and a countable input space;
3. an uncountable state and an uncountable input space with positive cost function;
4. an uncountable state and an uncountable input space with negative cost function.

This distinction is made because the theory and the proofs differ by case. The reader may want to use this information when searching for books or papers for a particular optimal stochastic control problem.

In this chapter the results and proofs will be presented for optimal stochastic control on an infinite horizon for the cases: (1) a state and input set which are measurable subsets of tuples of the real numbers with the Borel σ-algebra, and with positive cost rates; and (2) finite-state and finite-input spaces. For the other cases the reader has to search in the literature, see the Section *Further Reading* at the end of this chapter.

The solution of the optimal stochastic control problem will be based on the following procedure.

Procedure 13.2.3 Solution procedure for average cost. *The solution procedure for an optimal stochastic control problem with complete observations on an infinite-horizon for an average cost, consists of the following steps,*

1. *Show existence of a nonempty subset of control laws $G_f \subseteq G$ such that for every $g \in G_f$, the infinite horizon costs is finite; or, equivalently, such that $J(g) < \infty$. If no such subset of control law exists then the control problem is not interesting because no control laws can be compared. The existence requires a condition of stochastic controllability or of stochastic stabilizability.*
2. *Formulate the* dynamic programming equation *for the value function.*
3. *Prove existence of a value function and of a value which tuple is a solution of the dynamic programming equation. Uniqueness of a solution holds if specific assumptions hold.*
4. *Calculate or compute a tuple of a value and a value function which tuple is a solution of the dynamic programming equation.*
5. *Derive the optimal control law from the value function.*

There exists an example, Example 13.2.19, for which there exists a time-varying control law which is optimal while there does not exist a time-invariant control law achieving the infimal cost. But many other optimal stochastic control problems have an optimal control law which is time-invariant. The author does not know of a control synthesis procedure for the construction of a time-varying optimal control law. In this book attention is therefore restricted to the subset of time-invariant control laws.

13.2.2 Positive Cost

The reader finds in this section the dynamic procedure to determine the optimal control law for the average cost optimal stochastic control problem. The stochastic control system is with complete observations and with state and input spaces which are measurable subsets of tuples of the real numbers. The cost function has a positive cost rate.

Below the following expression is evaluated as first described in Section 12.6.

$$E[V(f(x_V,u_V,v(t))|F^{x_V,u_V}] = \int V(f(x_V,u_V,w)f_{v(t)}(dw),$$

where $f_{v(t)}$ denotes the probability distribution function of the random variable $v(t)$ and the variables $(x_V,\ u_V) \in X \times U$ are fixed. Recall from the definition of a recursive stochastic control system, Def. 10.1.2, that $\{v(t) \in \mathbb{R}^{n_v},\ t \in T\}$ is a sequence of independent random variables. The integral displayed above is thus an ordinary expectation over the probability distribution function $f_{v(t)}(.)$ while the variables x_V, u_V are treated as indeterminates.

The reader is expected to be aware of the digressions on optimization and on measurable control laws of the Sections 12.4 and 12.5.

Definition 13.2.4. The *dynamic programming equation of average cost.* Consider Problem 13.2.1 of optimal stochastic control on an infinite-horizon with an average cost and a positive cost rate.

Define the *dynamic programming equation of average cost* as the equation for a tuple (J^*, V),

$$\begin{aligned} & (J^*, V),\ J^* \in \mathbb{R}_+,\ V : X \to \mathbb{R}_+,\ \forall\, x_V \in X,\ \exists\, U(x_V) \subseteq U; \\ J^* + V(x_V) = & \inf_{u \in U(x_V)} \{b(x_V, u_V) + E[V(f(x_V, u_V, v(t)))|F^{x_V, u_V}]\},\ \ \forall\, x_V \in X. \end{aligned}$$

Call J^* the *value* and V the *value function* of this dynamic programming equation.

Define the time-invariant *dynamic programming operator* as the function,

$$\begin{aligned} DP(V)(x_V) = & \inf_{u_V \in U(x_V)} \{b(x_V, u_V) + E[V(f(x_V, u_V, v(t)))|F^{x_V, u_V}]\}, \\ & DP : F(X,\ \mathbb{R}_+) \to F(X,\ \mathbb{R}_+), \end{aligned}$$

where the expression in curly brackets is a measurable function. It will be *assumed* that the expression $DP(V)$ is also a measurable function.

Proposition 13.2.5. *Consider the dynamic programming equation of average cost.*

(a)If (J_1^*, V) and (J_2^*, V) are two solutions of the dynamic programming equation of Def. 13.2.4 then $J_1^* = J_2^*$.
(b)If (J^*, V) is a solution of the dynamic programming equation and $c \in \mathbb{R}_+$ then $(J^*, V + c)$ is another solution of the dynamic programming equation of average cost. Thus the dynamic programming equation does not have a unique solution for the function V.

Proof. (a). From the dynamic programming equation and the assumption that (J_1^*, V) and (J_2^*, V) are both solutions, follows by substraction of the two equations that $J_1^* = J_2^*$.
(b) If $(J^*, V + c)$ is a solution of the dynamic programming equation then the constant c occurs on both the left-hand side and the right-hand side of the equation and thus cancels. The resulting equation is then for (J^*, V). Hence (J^*, V) and $(J^*, V + c)$ are solutions of the same equation. □

How to solve the dynamic programming equation of average cost for the value and the value function?

Note that if one is provided a candidate solution for the value function V then it can be calculated whether it is a solution. Note that in the dynamic programming equation of average cost one is provided the dynamics of the stochastic control system in the form of the function $f(x, u, v(t))$, the probability distribution function of the noise variable $v(t)$, and of the average cost function, the cost rate $b(x,u)$. With a candidate solution V one can calculate the right-handside of the dynamic programming equation which involves an infimization problem over a finite, a countable, or a finite-dimensional space. One can then check whether there exists a real number J^* such that the equality of the left-hand side and the right-hand side holds for all states. If equality holds then the tuple (J^*,V) is a solution of the dynamic programming equation.

In case of a stochastic control system with a continuous state space X the procedure to solve the dynamic programming equation of average costs is as follows. Formulate a conjecture for the analytic form of the value function as a function of the state. In the case of a Gaussian stochastic control system one selects a quadratic function $V(x) = x^T Q_c x$. Calculate then the right-hand side of the dynamic programming equation and check whether it equals the left-hand side. In the case of a Gaussian stochastic control system with the candidate function $V(x) = x^T Q_c x$ one then discovers that the matrix Q_c has to satisfy an algebraic Riccati equation with side conditions. A separate analysis of that equation is then necessary. One then also finds the value J^*. In case of other stochastic control system it is best to calculate a few steps of the dynamic programming operator to formulate a conjecture for the analytic form of the value function.

In case of a finite stochastic control system, Section 13.2.4 provides a control law iteration procedure to compute the value and the value function.

Procedure 13.2.6 *The* procedure for dynamic programming with average cost.

Consider the stochastic control problem on an infinite horizon with the average cost criterion and a positive cost rate, see Problem 13.2.1. Assume that the subset of time-invariant control laws achieving a finite cost is nonempty, equivalently, $G_{ti,acf} \neq \emptyset$.

1. *Determine a tuple (J^*,V) which is a solution of the dynamic programming equation of average cost, of which existence is assumed,*

$$\begin{aligned} & J^* \in \mathbb{R}_+, \; V : X \to \mathbb{R}_+, \\ J^* + V(x_V) = & \inf_{u \in U(x_V)} \{ b(x_V,u_V) + E[V(f(x_V,u_V,v(t)))|F^{x_V,u_V}] \}. \end{aligned} \tag{13.4}$$

2. *If,*

$$\begin{aligned} & \forall \, x_V \in X, \; \exists \, u^* \in U(x_V), \text{ such that,} \\ & b(x_V,u^*) + E[V(f(x_V,u^*,v(t))|F^{x_V,u^*}] \\ = & \inf_{u_V \in U(x_V)} \{ b(x_V,u_V) + E[V(f(x_V,u_V,v(t))|F^{x_V,u_V}] \}, \\ & \text{then define } g^*(x_V) = u^*, \;\; g^* : X \to U. \end{aligned} \tag{13.5}$$

   In examples one may impose a condition such that the infimization has a unique solution.

3. *Check whether or not the functions g^* and V are measurable function and such that $g^* \in G_{acf}$. Stop if the condition is not met.*

4. *Output the triple (J^*, V, g^*), of the value J^*, of the value function V, and of the optimal control law g^*.*

Theorem 13.2.7. Sufficient condition for an optimal control law of average cost. *Consider the stochastic control problem on an infinite horizon with the average cost criterion and with positive cost rate, see Problem 13.2.1. Assume that the subset of control laws with finite average cost is nonempty, equivalently, $\emptyset \neq G_{acf} \subseteq G$. Assume that there exist a tuple (V, J^*) of the dynamic programming equation for average cost, equation (13.4). Assume further that the following two conditions both hold,*

$$\forall\, g \in G_{acf},\ \limsup_{t_1 \to \infty} \frac{1}{t_1}\, E[V(x^g(t_1)] = 0, \text{ and } E[V(x_0)] < \infty.$$

(a)Then,

$$J^* \leq J(g),\ \forall\, g \in G.$$

(b)If, in addition to (a), if it is assumed that, for all $x_V \in X$, there exist a minimizer $u^* \in U(x_V)$ in equation (13.5), if the functions g^* and V are measurable functions, and if $g^* \in G_{acf}$, then the function $g^* \in G_{acf}$ is an optimal control law satisfying,

$$J^* = J(g^*) = \inf_{g \in G_{acf}} J(g).$$

Proof. (a) Consider an arbitrary control law $g \in G_{ac}$. Let $t_1 \in T$. Note the calculations,

$$\begin{aligned}
& b(x^g(s), u^g(s)) + E[V(f(x^g(t), u^g(t), v(t)))|F^{x^g(s), u^g(s)}] \\
\geq & \inf_{u \in U(x^g(s))} \left\{ b(x^g(s), u) + E[V(f(x^g(s), u, v(s)))|F^{x^g(s), u}] \right\} \\
= & J^* + V(x^g(s)), \text{ by equation (13.4)}, \\
\Rightarrow & \sum_{s=0}^{t_1 - 1} \left(b(x^g(s), u^g(s)) + E[V(x^g(s+1))|F^{x^g(s), u^g(s)}] \right) \\
\geq & \sum_{s=0}^{t_1 - 1} [J^* + V(x^g(s))] = t_1 J^* + \sum_{s=0}^{t_1 - 1} V(x^g(s)); \ \Rightarrow \\
J^* \leq & \frac{1}{t_1} \sum_{s=0}^{t_1 - 1} E[b(x^g(s), u^g(s))] + \\
& + \frac{1}{t_1} \sum_{s=0}^{t_1 - 1} E[E[V(x^g(s+1))|F^{x^g(s), u^g(s)}] - \frac{1}{t_1} \sum_{s=0}^{t_1 - 1} E[E[V(x^g(s))]] \\
= & \frac{1}{t_1} \sum_{s=0}^{t_1 - 1} E[b(x^g(s), u^g(s))] + \frac{1}{t_1}\, E[V(x^g(t_1))] - \frac{1}{t_1}\, E[V(x(0))]
\end{aligned}$$

$$\begin{aligned}
J^* &\leq \limsup_{t_1 \to \infty} \frac{1}{t_1} \sum_{s=0}^{t_1-1} E[b(x^g(s), u^g(s))] + \\
&\quad + \limsup_{t_1 \to \infty} \frac{1}{t_1} E[V(x^g(t_1))] - \liminf \frac{1}{t_1} E[V(x(0))] \\
&= J(g), \text{ because by assumption, } E[V(x_0)] < \infty, \ \limsup_{t_1 \to \infty} \frac{1}{t_1} E[V(x^g(t_1))] = 0, \\
J^* &\leq \inf_{g \in G_{acf}} J(g), \text{ because the inequality holds for all } g \in G_{acf}.
\end{aligned}$$

This proves that J^* is a lower bound of the cost function.
(b) In case the infima in equation (13.5) are all attatained, then equality holds in the first inequality in the above calculations. Hence equality holds also in the other inequalities. From the assumption that there exists a nonempty subset of control laws achieving a finite cost, it follows that $g^* \in G_{acf}$. Hence,

$$J^* = J(g^*), \text{ and } J^* = \inf_{g \in G_{acf}} J(g).$$

Thus J^* is the value and $g^* \in G_{acf}$ is an optimal control law. □

13.2.3 Control of a Gaussian Control System

For control engineering, the optimal control law of the optimal stochastic control problem for a Gaussian stochastic control system with complete observations on an infinite-horizon with a quadratic cost rate, is much used. The reader finds in this section a derivation.

Problem 13.2.8. *The optimal stochastic control problem with complete observations and average cost for a time-invariant Gaussian system with quadratic cost rate (LQG-CO-AC).*

Consider a time-invariant Gaussian stochastic control system representation,

$$\begin{aligned}
x(t+1) &= Ax(t) + Bu(t) + Mv(t), x(0) = x_0, \\
z(t) &= C_z x(t) + D_z u(t), \\
&\quad n_u \leq n_z, \ D_z \in \mathbb{R}^{n_z \times n_u}, \ \text{rank}(D_z) = n_u \ \Rightarrow \\
&\quad \text{rank}(D_z^T D_z) = n_u \ \Rightarrow \ D_z^T D_z \succ 0.
\end{aligned}$$

In general it is recommended to take the exact-actuated case with $n_u = n_z$, see Def.10.2.3.

Consider the past-state information pattern and the corresponding set G of control laws. Then the closed-loop system has the representation,

$$\begin{aligned}
x^g(t+1) &= Ax^g(t) + Bg_t(x^g(0:t)) + Mv(t), \ x^g(0) = x_0, \\
u^g(t) &= g_t(x^g(0:t)) = g_t((x^g(0), \ldots, x^g(t))), \\
z^g(t) &= C_z x^g(t) + D_z u^g(t).
\end{aligned}$$

Define the average-cost function, and the problem to be solved,

$$J_{ac}(g) = \limsup_{t_1 \to \infty} \frac{1}{t_1} E\left[\sum_{s=0}^{t_1-1} z^g(s)^T z^g(s)\right], \; J_{ac} : G \to \mathbb{R}_+ \cup \{+\infty\},$$
$$\inf_{g \in G} J_{ac}(g).$$

To be more specific, the closed-loop control system should achieve:

1. the control objective that the closed-loop system is exponentially stable;
2. the control objective that the eigenvalues of the closed-loop system are bounded away from the instability boundary by a margin $\delta \in (0,1)$ hence in the subset $\{c \in \mathbb{C} |\; |c| < 1-\delta\}$;
3. the control objective that the average cost of the optimal control law $g^* \in G_{acf}$ is strictly less than the average cost of a *zero control law*, $J_{ac}(g^*) < J_{ac}(g_{zo})$, where $g_{zo} \in G$ is defined as $g_{zo}(.) = 0$ for all its arguments.

For the control objectives specified above, controllability and observability of the stochastic control system are necessary. It will be argued below that the control objective of a stability margin of the eigenvalues of the closed-loop system does not hold in general.

Below an assumption is stated which will be imposed to make the optimal control problem well defined. It will be proven later that the sufficient condition implies that there exists a subset of control laws achieving a finite cost. In addition, one wants to prove that the sufficient conditions are also necessary.

Assumption 13.2.9 *Consider the Gaussian stochastic control system of Problem 13.2.8. Define the matrices,*

$$A_c = A - B(D_z^T D_z)^{-1} D_z^T C_z \in \mathbb{R}^{n_x \times n_x},$$
$$C_c^T C_c = C_z^T C_z - C_z^T D_z (D_z^T D_z)^{-1} D_z^T C_z \in \mathbb{R}_{pds}^{n_x \times n_x}, \; C_c \in \mathbb{R}^{n_x \times n_x}.$$

It follows from Proposition 17.4.33 for the Schur complement that $C_c^T C_c \succeq 0$ hence a factorization for C_c exists.

(a)The controllability and the observability conditions *hold if: (1) (A,B) is a controllable pair; and (2) (A_c,C_c) is an observable pair. If $D_z^T C_z = 0$ then $A_c = A$ and C_c can be chosen such that $C_c = C_z$, hence $(A_c,C_c) = (A,C_z)$ and condition (2) is adjusted correspondingly.*

(b)The stabilizability and detectability conditions *holds if: (1) (A,B) is a stabilizable pair; and (2) (A_c,C_c) is a detectable pair. Again, if $D_z^T C_z = 0$ then a similar implication as in case (a) holds.*

An explanation of the above assumptions follows. If the time-invariant Gaussian stochastic control system is neither controllable nor observable then it follows from the Kalman decomposition of a linear system, Def. 21.3.7, that the Gaussian control system representation may be transformed by a state-space transformation to the following form,

$$\begin{aligned} x(t+1) &= \begin{pmatrix} A_{11} & A_{12} & 0 & 0 \\ 0 & A_{22} & 0 & 0 \\ A_{31} & A_{32} & A_{33} & A_{34} \\ 0 & A_{42} & 0 & A_{44} \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{pmatrix} + \begin{pmatrix} B_1 \\ 0 \\ B_3 \\ 0 \end{pmatrix} u(t) + Mv(t), \\ z(t) &= \begin{pmatrix} C_{z,1} & C_{z,2} & 0 & 0 \end{pmatrix} x(t) + D_z u(t), \\ & \left(\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \begin{pmatrix} C_{z,1} & C_{z,2} \end{pmatrix} \right), \text{ an observable pair,} \\ & \left(\begin{pmatrix} A_{11} & 0 \\ A_{31} & A_{33} \end{pmatrix}, \begin{pmatrix} B_1 \\ B_3 \end{pmatrix} \right), \text{ a controllable pair.} \end{aligned}$$

The Kalman decomposition has been carried out only for the linear subsystem relating the input u to the state x and to the controlled output z.

The subsystem with state component x_1 is both controllable and observable; that with state component x_2 is not controllable but observable; that with state component x_3 is controllable but not observable; and that with state component x_4 is neither controllable nor observable.

Consider the conditions of Assumption 13.2.9(b). Stabilizability of the tuple (A,B) implies that the following spectral conditions both hold $\text{spec}(A_{22}) \subset \mathrm{D}_o$ and $\text{spec}(A_{44}) \subset \mathrm{D}_o$. Detectability of the tuple (A,C_z) implies that the following spectral conditions both hold $\text{spec}(A_{33}) \subset \mathrm{D}_o$ and $\text{spec}(A_{44}) \subset \mathrm{D}_o$.

If the average cost is to be finite and if the tuple $(A,\ M)$ is supportable then it will be proven below that the controlled output z should not contain any dynamics with unstable eigenvalues. If the controlled output z is not to contain dynamics based on unstable eigenvalues, those on or strictly outside the unit disc, then this implies the condition that $\text{spec}(A_{22}) \subset \mathrm{D}_o$. Note that the modes of the subsystem of the state components $x_3,\ x_4$ will never show up in the controlled output z because those modes are unobservable. The modes of the subsystem with state component x_1 are controllable and observable, hence can be made stable by feedback. Thus the only necessary condition for a finite cost is the condition $A_{22} \subset \mathrm{D}_o$. Note that the state component x_2 affects in the system dynamics the state component x_1 via the matrix A_{12}.

A comparison of the conditions follows. Assumption 13.2.9(b) implies that $\text{spec}(A_{22}) \subset \mathrm{D}_o$, $\text{spec}(A_{33}) \subset \mathrm{D}_o$, and $\text{spec}(A_{44}) \subset \mathrm{D}_o$. But finiteness of the cost requires only that $\text{spec}(A_{22}) \subset \mathrm{D}_o$.

In general it seems advisable to require stabilizability of all subsystems with the state components $x_2,\ x_3,\ x_4$. Then the entire closed-loop system can be made asymptotic stability. The latter control objective is not included in the average cost function.

Definition 13.2.10. Consider Problem 13.2.8. Assume that (A,B) is a controllable pair and that (A,C) is an observable pair. Define the set of *linear Markov control laws* as,

$$G_{ML} = \left\{ g : X \to U \,|\, \exists\, F \in \mathbb{R}^{n_u \times n_x},\ g(x) = Fx,\ \text{spec}(A+BF) \subset \mathrm{D}_o \right\}.$$

For a Markov control law with $g(x) = Fx$, the closed-loop system has the representation,

$$\begin{aligned} x^g(t+1) &= Ax^g(t)+BFx^g(t)+Mv(t) = (A+BF)x^g(t)+Mv(t),\ x^g(t_0)=x_0, \\ z^g(t) &= (C_z+D_zF)x^g(t). \end{aligned}$$

Is, for any feedback matrix F, the matrix tuple $(A+BF,M)$ a supportable pair and is the tuple $(A+BF,C+DF)$ an observable pair? If $(A+BF,M)$ is a supportable pair then it follows from Theorem 4.4.5.(d) that the state variance of the state x^g has support on the full state space, equivalently $0 \prec Q_{x^g}$. If $(A+BF,C+DF)$ is in addition an observable pair then the variance of the controlled output z^g depends fully on the variance of the state x^g of the closed-loop system. Answers to these questions of supportability and of observability follow directly from the theory of linear control systems. The concept of strong observability implies that $(A+BF,C+DF)$ is an observable pair for all F, which is established for continuous-time linear control systems in [38, Thm. 7.16]. For the supportability of $(A+BF,M)$ there is not yet an appropriate concept in the literature, note that M and B are in general different matrices.

Is the average cost function for this closed-loop system finite? If so, what is the value $J_{ac}(g)$ for any control law $g \in G_{ML}$. These questions are treated below.

Example 13.2.11. *Nonstabilizable system and infinite cost.* Below is provided a stochastic control system which is not stabilizable and it is proven that for any control law $g \in G_{ML}$, the cost is infinity.

Consider the single-state Gaussian control system,

$$\begin{aligned} x(t+1) &= ax(t)+bu(t)+mv(t),\ x(0)=x_0,\ x_0 \in G(0,q_{x_0}),\ q_{x_0}>0, \\ z(t) &= c_zx(t)+d_zu(t), \\ &\quad n_x=1,\ n_v=1,\ |a|>2,\ m\neq 0,\ q_v\in(0,\infty),\ b=0,\ c_z\neq 0. \end{aligned}$$

Thus the $B=b=0$ matrix is zero and the system matrix is exponentially unstable. For any control law $g \in G_{ML}$, the state process x^g is the same as that of x for control law $g=0$ because the B matrix is zero. The variance of the state process and the average cost satisfy,

$$\begin{aligned} q_x(t+1) &= a^2q_x(t)+m^2q_v,\ q_x(0)=q_{x_0}>0,\ \lim_{t\to\infty} q_x(t)=+\infty, \\ z^g(t) &= (c_z+d_zf)x^g(t),\ E[x^g(t)]=0, \\ J_{ac}(g) &= \lim_{t_1\to\infty}\frac{1}{t_1}\sum_{s=0}^{t_1-1} E[z^g(s))^2] = \lim_{t_1\to\infty}\frac{1}{t_1}\sum_{s=0}^{t_1-1} E[x^g(s)^2](c_z+d_zf)^2 \\ &= (c_z+d_zf)^2 \lim_{t_1\to\infty}\frac{1}{t_1}\sum_{s=0}^{t_1-1} q_x(s) = +\infty. \end{aligned}$$

Proposition 13.2.12. *Consider Problem 13.2.8. Consider a linear Markov control law $g \in G_{ML}$ with representation $g(x)=Fx$. The closed-loop system is then,*

$$\begin{aligned} x^g(t+1) &= (A+BF)x^g(t)+Mv(t),\ x^g(0)=0, \\ z^g(t) &= C_zx^g(t)+D_zg(x(t)) = (C_z+D_zF)x^g(t). \end{aligned}$$

(a)Consider Problem 13.2.8 and assume that the stabilizability and detectability conditions of Assumption 13.2.9 hold. Thus the stochastic control system is stabilizable. From Theorem 21.2.11.(c) follows that there exists a matrix $F \in \mathbb{R}^{n_u \times n_x}$ such that $\mathrm{spec}(A+BF) \subset \mathrm{D}_o$. *Hence the class G_{ML} is not empty.*

(b)For any control law $g \in G_{ML}$ with $g(x) = Fx$ there exists a state variance matrix $Q_x^g \in \mathbb{R}^{n_x \times n_x}$ which is the unique solution of the following Lyapunov equation, and the average cost of that control law is then,

$$\begin{aligned} Q_x^g &= (A+BF)Q_x^g(A+BF)^T + MM^T, \\ J_{ac}(g) &= \mathrm{tr}((C_z+D_zF)Q_x^g(C_z+D_zF)^T) < \infty. \end{aligned}$$

In addition, there exists a cost matrix $Q_c^g \in \mathbb{R}^{n_x \times n_x}$ which is the unique solution of the Lyapunov equation,

$$Q_c^g = (A+BF)^T Q_c^g (A+BF) + (C_z+D_zF)^T(C_z+D_zF) \in \mathbb{R}_{pds}^{n_x \times n_x}.$$

Then,

$$J_{ac}(g) = \mathrm{tr}((C_z+D_zF)Q_x^g(C_z+D_zF)^T) = \mathrm{tr}(M^T Q_c^g M) < \infty.$$

Proof. (a) This follows from Proposition 21.2.11.(c).

(b) The closed-loop system satisfies $\mathrm{spec}(A+BF) \subset \mathrm{D}_o$. Then it follows from Theorem 22.1.2 that there exists an invariant measure for the state process, $x(t) \in G(0, Q_x^g)$, with Q_x^g the unique solution of a Lyapunov equation such that,

$$Q_x^g = (A+BF)Q_x^g(A+BF)^T + MM^T.$$

Because $x_0 = 0$, for all $t \in T$, $E[x(t)] = 0$. Then,

$$\begin{aligned} Q_x^g(t+1) &= (A+BF)Q_x^g(t)(A+BF) + MM^T, \; Q_x^g(0) = Q_{x_0}, \\ &\lim_{t\to\infty} Q_x^g(t) = Q_x^g = (A+BF)Q_x^g(A+BF) + MM^T, \\ &\lim_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} Q_x^g(s) = Q_x^g. \end{aligned}$$

where the convergence of the last sequence follows from Theorem 22.1.2.(e). Then,

$$\begin{aligned} z^g(t) &= (C_z+D_zF)x^g(t), \\ E[z^g(t)^T z^g(t)] &= \mathrm{tr}((C_z+D_zF)Q_x^g(t)(C_z+D_ZF)^T), \\ J(g) &= \lim_{t_1\to\infty} \frac{1}{t_1} \sum_{s=0}^{t_1-1} E[z^g(s)^T z^g(s)] \\ &= \lim_{t_1\to\infty} \frac{1}{t_1} \sum_{s=0}^{t_1-1} \mathrm{tr}((C_z+D_zF)Q_x^g(s)(C_z+D_zF)^T) \\ &= \mathrm{tr}((C_z+D_zF)Q_x^g(C_z+D_zF)^T) < \infty. \end{aligned}$$

It follows from the assumption that $\mathrm{spec}(A+BF) \subset \mathrm{D}_o$ and from Theorem 22.1.2 that the Lyapunov equation for the matrix Q_c has a unique solution. It then follows from Theorem 22.1.2.(f) that

$$\mathrm{tr}((C_z+D_zF)Q_x^g(C_z+D_zF)^T) = \mathrm{tr}(M^T Q_c^g M).$$

□

Theorem 13.2.13. *Consider the optimal stochastic control problem 13.2.8.*

(a)There exists a time-invariant control law of the form $g(x) = Fx$ such that the eigenvalues of the closed-loop system matrix satisfy $\text{spec}(A+BF) \subset \text{D}_o$ *if and only if the stabilizability condition of Assumption 13.2.9(b) holds. With such a control law the average cost function is finite.*

(b)For any set $\Lambda \subset \text{D}_o$ of n_x complex-conjugate eigenvalues there exists a control law of the form $g(x) = Fx$ such that the eigenvalues of the closed-loop system matrix equal the specified set of eigenvalues, $\text{spec}(A+BF) = \Lambda$*, if and only if the matrix tuple (A,B) is a controllable pair, which condition belongs to the controllability condition of Assumption 13.2.9(a). In this case the average cost is also finite.*

(c)Consider the time-invariant control law $g(x) = Fx$ with $g \in G_{ML}$ and the resulting closed-loop system. Assume that the matrix tuple $(A+BF,\ M)$ is a supportable pair.
If the average cost $J(g)$ is finite then the subsystem which is both observable and noncontrollable, is exponentially stable. In general this condition is strictly weaker than stabilizability.

Proof. (a) ($\Leftarrow$) From Proposition 13.2.12 follows that stabilizability implies that that there exists a linear Markov control law $g = Fx \in G_{ML}$ such that the cost $J(g) = \text{tr}((C_z + D_zF)Q_x^g(C_z + D_zF)^t) < \infty$.
($\Rightarrow$) The proof is by contradiction. Suppose that the system is not stabilizable, or, equivalently, that (A,B) is not a stabilizable pair. It has to be shown that, for all control laws $g \in G$, the cost is infinite. Example 13.2.11 shows that there exists a nonstabilizable system of state-space dimension one, which is by assumption observable by the controlled output and supportable, such that for all control laws, the cost is infinity. A corresponding example of a time-invariant Gaussian system such that the average cost is infinite, can then be easily constructed by the reader.
(b) See Theorem 21.2.7.(b) which establishes that there exists a feedback matrix $F \in \mathbb{R}^{n_u \times n_x}$ such that $\text{spec}(A+BF) = \Lambda$, the set of conjugate eigenvalues specfied in the problem statement, if and only if the tuple (A,B) is a controllable pair.
(c) Due to the assumption of supportability and of $g \in G_{ML}$, the variance matrix Q_x^g of the invariant distribution of the state is strictly positive definite by Theorem 22.1.2(d). Note the expression of the average cost obtained in Proposition 13.2.12, $J(g) = \text{tr}((C_z + D_zF)Q_x^g(C_z + D_zF)^T)$. Consider then the Kalman decomposition of the closed-loop Gaussian system considered with input u and output z. The conclusion then follows directly be a contradiction argument. □

The preceeding theorem establishes necessary and sufficient conditions for either stabilizability of the closed-loop system or assignment of the state dynamics of the closed-loop system in terms of the spectrum of the closed-loop system. If either of these conditions hold then one can do more and optimize the performance of the average cost.

The reader is expected to have read Theorem 22.2.4 which deals with the existence of a solution of the control algebraic Riccati equation.

Definition 13.2.14. *Average-cost optimal stochastic control law for a time-invariant Gaussian stochastic control system.* Consider Problem 13.2.8 and assume that the stabilizability and the detectability conditions of Assumption 13.2.9 hold.

It follows from Theorem 22.2.4 that there exists a matrix $Q_c^* \in \mathbb{R}_{pds}^{n_x \times n_x}$ which is a solution of the control algebraic Riccati equation with side conditions,

$$\begin{aligned} Q_c^* &= A^T Q_c^* A + C_z^T C_z + \\ &\quad -[A^T Q_c^* B + C_z^T D_z][B^T Q_c^* B + D_z^T D_z]^{-1}[A^T Q_c^* B + C_z^T D_z]^T, \\ &\quad Q_c^* \in \mathbb{R}_{pds}^{n_x \times n_x},\ \text{spec}(A + BF(Q_c^*)) \subset \text{D}_o,\ \text{where}, \\ F(Q_c^*) &= -[B^T Q_c^* B + D_z^T D_z]^{-1}[A^T Q_c^* B + C_z^T D_z]^T. \end{aligned}$$

Define then the *optimal control law* of Problem 13.2.8 by the equations,

$$\begin{aligned} g_{LQG-CO-AC}^*(x) &= g^*(x) = F(Q_c^*)x, \\ F(Q_c^*) &= -[B^T Q_c^* B + D_z^T D_z]^{-1}[A^T Q_c^* B + C_z^T D_z]^T. \end{aligned}$$

It will be proven in Theorem 13.2.15 that the defined control law is indeed optimal.

In the above control algebraic Riccati equation there is a relation between the matrix Q_c^* being positive-definite and the condition on the spectrum of $\text{spec}(A + BF(Q_c^*)) \subset \text{D}_o$, in fact, either one of these conditions implies the other. In the literature the statements in regard to this issue are not always clear. Therefore the two conditions are mentioned above together. The reader is referred to Theorem 22.2.4 for the details of the control algebraic Riccati equation and on the above implications.

Theorem 13.2.15. *The* optimal control law *(LQG-CO-AC). Consider the optimal stochastic control problem with complete observations and with average cost as defined in Problem 13.2.8. Assume that the stabilizability and detectability conditions of Assumption 13.2.9 hold.*

(a)The optimal control law $g_{LQG-CO-AC}^*$ of Def. 13.2.14 is well defined and there exists a matrix $Q_c^* \in \mathbb{R}_{pds}^{n_x \times n_x}$ which is a solution of the control algebraic Riccati equation with side conditions of Theorem 22.2.4.

(b)The closed-loop system with the optimal control law is exponentially stable,

$$\begin{aligned} x^{g^*}(t+1) &= (A + BF(Q_c^*)x^{g^*}(t) + Mv(t),\ x^{g^*}(0) = x_0, \\ z^{g^*}(t) &= (C + DF(Q_c^*))x^{g^*}(t), \\ &\quad \text{spec}(A + BF(Q_c^*)) \subset \text{D}_o. \end{aligned}$$

The state process x^{g^*} of the closed-loop system and the controlled output process z^{g^*} are jointly Gaussian processes of which the mean and variance can be calculated according to Theorem 4.3.5.

Moreover, the average cost of this control law is finite, $J_{ac}(g^*) < \infty$ and $g^* \in G_{acf}$.

(c)The value function, the value, and the minimal average cost of the optimal control law are specified by the equations,

$$V(x_V) = x_V^T Q_c^* x_V, \tag{13.6}$$

$$J_{ac}^* = \text{tr}(M^T Q_c^* M), \tag{13.7}$$

$$J_{ac}^* = J_{ac}(g^*), \tag{13.8}$$

The optimal control law of Def. 13.2.14 is optimal over the set of measurable nonlinear control laws.

(d)*Denote for any feedback matrix* $F \in \mathbb{R}^{n_u \times n_x}$ *such that* $\text{spec}(A+BF) \subset \text{D}_o$ *the corresponding control law by* $g_s(x) = Fx$. *The difference in cost of this suboptimal control law and the LQG-optimal control law is provided by the expression,*

$$\begin{aligned} J(g^*) - J(g_s) &= \text{tr}(M^T Q_c^* M) - \text{tr}(M^T Q_c M) = \text{tr}(M^T (Q_c^* - Q_c) M) \leq 0, \\ Q_c &= (A+BF)^T Q_c (A+BF) + (C_z + D_z F)^T (C_z + D_z F). \end{aligned}$$

The above inequality shows explicitly that the cost of a linear stabilizing control law *has a higher or equal cost than the LQG optimal control law, and it provides an expression for the regret.*

Note that the optimal control law depends on the system matrices (A,B,C_z,D_z) but not on the noise matrix M. However, the value J^* depends on all five system matrices (A,B,M,C_z,D_z).

The analysis of the behavior of the eigenvalues of the closed-loop LQG-controlled linear system is due to H. Kwakernaak and R. Sivan, see the references of Chapter 22. An example of a *continuous-time* linear system have a linear system zero is the system which is a model of an inverted pendulum mounted on a cart, see the book of H. Kwakernaak and R. Sivan, [**?**, Ex.]. The reader is referred for this issue of this item to Section 22.2 for a description of the asymptotic behavior of the eigenvalues of the deterministic system.

There follows a longer discussion of the analysis of the behavior of the eigenvalues of the closed-loop LQG control system. Consider the deterministic linear system

$$\begin{aligned} x_s(t+1) &= Ax_s(t) + Bu(t),\ x_s(0) = 0, \\ z_s(t) &= C_z x_s(t) + r\, D_z\, u(t),\ \ r \in (0,\infty). \end{aligned}$$

Determine whether or not there exist linear system zeroes of this linear system from the input u to the special output z_s inside the open unit disc. See Section 21.5 for the concept of a zero of a linear system.

Distinguish the cases:

(a)There do not exist such linear system zeroes in the open unit disc. For the parameter value r converging towards zero but without reaching the value zero, the eigenvalues of the closed-loop Gaussian system either are zero or converge to zero with a Butterworth pattern consisting of several radii. Consequently the ratio $J(g^*)/J(g_0)$ can be made arbitrarily small if the parameter r converges to zero.

(b)There exists one or more linear system zeroes in the open unit disc. If the parameter r converges to zero but is not zero yet then the eigenvalues of the closed-loop stochastic control system have the following properties:

(b.1)There may exist one or more eigenvalues in zero which do not change with the parameter r.

(b.2) There may exist eigenvalues of the closed-loop system, as many as there are linear system zeroes, such that, if the parameter value r converges to zero but is not zero yet, then such an eigenvalue converges to the location of a particular linear system zero. In case the linear system zero is close to the instability boundary, for example in the set

$$\{c \in \mathbb{C} |\ 1-\delta < |c| < 1\}, \quad \text{for a } \delta \in (0,1),$$

then also the corresponding eigenvalue will be near that zero and thus also be close to the instability boundary. However, due to the definition of a system zero, the system trajectory corresponding to that eigenvalue has no or little mathematical effect on the controlled output and on the average cost criterion. In practice and due to round-off errors, there may be a nonnegligible effect on the average cost. A user may consider to restrict the value of r or to modify the matrices C_z and D_z such that no system zeros exist.

(b.3) The remaining eigenvalues of the closed-loop system matrix, if any, converge to the zero of the complex plane in a Butterworth pattern with several radii when the parameter r converges to zero.

Proof. Proof of Theorem 13.2.15. (a) This follows directly from the stabilizability and detectability conditions of Assumption 13.2.9 and Theorem 22.2.4.
(b) This follows from the side condition of the control algebraic Riccati equation, see Def. 13.2.14.
(c) From the stabilizability and detectability conditions Assumption 13.2.9 follows that (A,B) is a stabilizable pair and from Theorem 13.2.13 follows that there exists a linear Markov control law with finite average cost. Hence the set $G_{acf} \subset G$ of control laws with finite average cost is not empty.

It will be shown below that the pair (V, J^*_{ac}) given by the equations (13.6,13.7) is a solution of the dynamic programming equation,

$$\begin{aligned} & J^* + V(x_V) \\ & = \inf_{u_V \in U(x_V)} \left((C_z x_V + D_z u_V)^T (C_z x_V + D_z u_V) + E[V(f(x_V, u_V, v(t)))|F^{x_V,u_V}] \right). \end{aligned}$$

A calculation shows that,

$$\begin{aligned} & E[V(Ax_V + Bu_V + Mv(t))|F^{x_V,u_V}] \\ & = E[(Ax_V + Bu_V + Mv(t))^T Q_c^* (Ax_V + Bu_V + Mv(t))|F^{x_V,u_V}] \\ & = \begin{pmatrix} x_V \\ u_V \end{pmatrix}^T \begin{pmatrix} A^T Q_c^* A & A^T Q_c^* B \\ (A^T Q_c^* B)^T & B^T Q_c^* B \end{pmatrix} \begin{pmatrix} x_V \\ u_V \end{pmatrix} + tr(M^T Q_c^* M); \text{ define,} \\ H & = \begin{pmatrix} H_{11} & H_{12} \\ H_{12} & H_{22} \end{pmatrix} = \begin{pmatrix} (A^T Q_c^* A + C_z^T C_z) & (A^T Q_c^* B + C_z^T D_z) \\ (A^T Q_c^* B + C_z^T D_z)^T & (B^T Q_c^* B + D_z^T D_z) \end{pmatrix}. \end{aligned}$$

Note that by $Q_c^* \succeq 0$ and by assumption, $H_{22} = B^T Q_c^* B + D_z^T D_z \succeq D_z^T D_z \succ 0$. By definition the matrix H is both symmetric and positive definite. This implies by the Schur complement of the matrix H, see Proposition 17.4.33, and the algebraic Riccati equation, that $Q_c = H_{11} - H_{12} H_{11}^{-1} H_{12}^T \succeq 0$. Then,

$$
\begin{aligned}
& \inf_{u_V \in U(x_V)} \{(C_z x_V + D_z u_V)^T (C_z x_V + D_z u_V) + E[V(f(x_V, u_V, v(t))))|F^{x_V, u_V}]\} \\
& = \inf_{u_V \in U(x_V)} \begin{pmatrix} x_V \\ u_V \end{pmatrix}^T H \begin{pmatrix} x_V \\ u_V \end{pmatrix} + \mathrm{tr}(M^T Q_c^* \, M) \\
& = \inf_{u_V \in U(x_V)} \begin{pmatrix} x_V \\ u_V + H_{22}^{-1} H_{12}^T x_V \end{pmatrix}^T \begin{pmatrix} Q_c^* & 0 \\ 0 & H_{22} \end{pmatrix} \begin{pmatrix} x_V \\ u_V + H_{22}^{-1} H_{12}^T x_V \end{pmatrix} + \\
& \quad + \mathrm{tr}(M^T Q_c^* \, M) \\
& = x_V^T Q_c^* x_V + \mathrm{tr}(M^T Q_c^* \, M), \text{ by } H_{22} \succ 0 \text{ and for } u_V^* = -H_{22}^{-1} H_{12}^T x_V, \\
& = V(x_V) + J^*.
\end{aligned}
$$

Thus (V, J^*) is a solution of the dynamic programming equation of average cost.

Note further that from Theorem 22.2.4 follows that,

$$
\begin{aligned}
& \lim_{t \to \infty} Q_{x^g}(t) = Q_{x^g}(\infty) < \infty, \\
& E[V(x^g(t))] = E[x^g(t)^T Q_c^* x^g(t)] = \mathrm{tr}(Q_c^* Q_{x^g}(t)] \\
\Rightarrow & \lim_{t \to \infty} \frac{1}{t} E[V(x^g(t))]] = \lim_{t \to \infty} \frac{1}{t} \mathrm{tr}(Q_c^* Q_{x^g}(t)) = \lim \ \mathrm{tr}(Q_c^* \lim Q_{x^g}(t)/t) < \infty; \\
& E[V(x_0)] = E[(x_0^g)^T Q_c^* x_0^g] = \mathrm{tr}(Q_c^* Q_{x_0}) < \infty.
\end{aligned}
$$

Hence the assumptions of Theorem 13.2.7 are satisfied.

From Theorem 13.2.7 follows that then J^* is the value of the problem.

From the calculation above follows that,

$$
\begin{aligned}
g^*(x_V) & = \mathrm{argmin}_{u_V \in U} \\
& \quad \left\{ (C_z x_V + D_z u_V)^T (C_z x_V + D_z u_V) + E[V(f(x_V, u_V, v(t))|F^{x_V, u_V}] \right\} \\
& = -H_{22}^{-1} H_{12}^T x_V = -[B^T Q_c^* B + D_z^T D_z]^{-1} [A^T Q_c^* B + C_z^T D_z]^T x_V \\
& = F(Q_c) x_V.
\end{aligned}
$$

Moreover, because $g^* \in G_{ML}$ is a linear function, it is a measurable function. Hence g^* is the optimal control law.

(d) It follows from Proposition 13.2.12 that for any feedback matrix F satisfying $\mathrm{spec}(A + BF) \subset \mathrm{D}_o$ the average cost is finite and equals $J_{ac}(g_s) = \mathrm{tr}(M^T Q_c M)$ where Q_c is the unique solution of the displayed Lyapunov equation. Because the optimal control law is a linear control and $\mathrm{spec}(A + BF(Q_c^*)) \subset \mathrm{D}_o$ it follows from the same proposition as quoted above that the average cost satisfies $J_{ac}(g^*) = \mathrm{tr}(M^T Q_c^* M)$. It follows from Theorem 22.2.7 that the solution Q_c^* of the control algebraic Riccati equation equals the state variance matrix associated with Proposition 13.2.12.

(e) See Section 22.2 in particular below Def. 22.2.5, and the references quoted there.

□

How to compute a LQG-optimal control law?

Procedure 13.2.16 Computation of an optimal control law for an optimal stochastic control problem with a Gaussian stochastic control system, with complete observations, on an infinite-horizon, and with an average cost.

Data: The system parameters $(n_x,\ n_u,\ n_v,\ n_z,\ A,\ B,\ M,\ C_z,\ D_z)$.

1. *Check the stabilizability and detectability conditions of Assumption 13.2.9. If the conditions do not hold then stop else proceed.*
2. *Compute the matrix $Q_c^* \in \mathbb{R}_{pds}^{n_x \times n_x}$ which is the solution of the Control Algebraic Riccati Equation with side conditions, see Theorem 13.2.15(a).*
3. *Compute the optimal control law and its performance according to,*

$$\begin{aligned} F(Q_c^*) &= -[C_z^T Q_c^* C_z + D_z^T D_z]^{-1} [A^T Q_c^* B^T + C_z^T D_z]^T, \\ g^*_{LQG,CO,AC}(x) &= F(Q_c^*) x, \\ J^*_{ac} &= \mathrm{tr}(M^T Q_c^* \, M). \end{aligned}$$

4. *Output $(g^*_{LQG,CO,AC},\ J^*_{ac},\ Q_c^*)$.*

13.2.4 Control of a State-Finite Stochastic Control System

In this section the optimal stochastic control problem with average cost is considered for a finite stochastic control system with complete observations.

Definition 13.2.17. A *time-invariant finite stochastic control system with complete observations*.

Consider the system of Def. 13.2.1 with $T = N = \{0, 1, 2, \ldots\}$, finite state set X, finite input set U, and with representation,

$$x(t+1) = f(x(t), u(t), v(t)), x(0) = x_0.$$

Assume that the information pattern for the input is the *past-state information pattern* $\{F_t^x,\ t \in T\}$. Assume that the set G of control laws is the set of past-state control laws. Thus any $g \in G$ is such that for all $t \in T \backslash \{t_1\}$, $g_t : X^{t+1} \to U$.

The closed-loop stochastic control system is then given by the representation,

$$\begin{aligned} x^g(t+1) &= f(x^g(t), g_t(x^g(0:t)), v(t)),\ x^g(0) = x_0, \\ u^g(t) &= g_t(x^g(0:t)). \end{aligned}$$

It is assumed that, for all control laws $g \in G$, the closed-loop system is irreducible, nonperiodic, and stochastically controllable.

The *cost rate* is a map $b : X \times U \to \mathbb{R}_+$. Measurability is not required on finite sets. Because both the state set X and the input set U are finite, there exists a real number $b_{max} \in \mathbb{R}_+$ such that $b_{max} = \max_{(x,u) \in X \times U}\ b(x,u)$. Then,

$$\lim_{t \to \infty} \frac{1}{t}\ E[\sum_{s=0}^{t-1} b(x^g(s), u^g(s))] \leq \lim_{t \to \infty} \frac{1}{t}\ t b_{max} = b_{max} < +\infty.$$

Thus the average cost is finite for every control law $g \in G$.

Denote the average cost by,

$$J(g) = \lim_{t \to \infty} \frac{1}{t}\ E[\sum_{s=0}^{t-1} b(x^g(s), u^g(s))],\ J : G \to \mathbb{R}_+.$$

Problem 13.2.18. *Optimal stochastic control problem with average cost for a finite stochastic control system with complete observations.*

Consider the stochastic control system of Def. 13.2.17. The optimal stochastic control problem with average cost is to determine a control law $g^* \in G$ such that,

$$J^*_{ac} = \inf_{g \in G} J_{ac}(g) = J_{ac}(g^*).$$

Is an optimal control law time-invariant or time-varying? An example was constructed by S. Ross.

Example 13.2.19. *Optimality of a time-varying control law*. There exist an example of a time-varying control law which achieves the infimal value while no time-invarant control law achieves that infimal value.

The control system is deterministic hence a special case of a finite stochastic control system, specified by,

$$\begin{aligned}
&X = \mathbb{Z}_+ = \{1,2,\ldots\},\ U = \{u_1,\ u_2\},\ T = \mathbb{N} = \{0,1,2,\ldots\},\ x_0 = 1,\\
&P(\{x(t+1) = k+1\}|\ \{x(t) = k,\ u(t) = u_1\}) = 1,\\
&P(\{x(t+1) = k\}|\ \{x(t) = k,\ u(t) = u_2\}) = 1,\ \forall\, k \in X,\ \forall\, t \in T.
\end{aligned}$$

The state trajectory is thus such that, if $u(t) = u_2$ then the system stays at the current state, while if $u(t) = u_1$ then the system makes a transition to the neighboring state with a higher state number.

The cost rate function is defined as,

$$b : X \times U \to \mathbb{R}_+,\ b(x,u_1) = 1,\ \forall\, x \in X,\ b(x,u_2) = 1/x,\ \forall\, x \in X.$$

Consider the sets of time-varying and time-invariant control laws as specified in Def. 13.2.1. Define the average cost in the usual form,

$$J_{ac}(g) = \limsup_{t\to\infty} \frac{1}{t} E[\sum_{s=1}^{t-1} b(x^g(s),u^g(s))],\ J_{ac} : G_{tv} \to \mathbb{R}_+ \text{ and } J_{ac,ti} : G_{ti} \to \mathbb{R}_+.$$

The set G_{ti} of time-invariant control laws contains only the following elements,

$$\begin{aligned}
&g_a \in G_{ti},\ g_a(x) = u_1,\ \forall\, x \in X,\\
&\forall\, k \in \mathbb{Z}_+,\ g_{b,k} \in G_{ti},\\
g_{b,k}(x) = &\begin{cases} u_1, & \forall\, x \in \{1,2,\ldots,k-1\},\\ u_2, & x = k,\\ \text{arbitrary}, & x \in X\backslash\{1,2,\ldots,k-1\}\end{cases}
\end{aligned}$$

The behavior of the closed-loop system is described next. In case the control law is g_a then the behavior of the closed-loop system is such that with $g_a(x) = u_1$ for all states, hence, at any state, the system makes transitions to the state with an index number one higher than the current state. In case the control law is $g_{b,k}$ then the behavior of the closed-loop system is such that it moves from the initial state $x_0 = 1$ to the subsequent states $\{2,\ 3,\ \ldots,\ k\}$ If $x(t) = k$, hence $t = k$, then $g_{b,k}(k) = u_2$ and by definition of the system $x(t+1) = k$, and the system stays at state k for every after. The values of the control law $g_{b,k}(x)$ for $x > k$ are never used hence are not specified.

The average cost of the above time-invariant control laws is calculated.

$$\begin{aligned}
J_{ac}(g_a) &= \limsup_{t\to\infty} \frac{1}{t} E[\sum_{s=1}^{t-1} b(x^{g_a}(s), u^{g_a}(s))] = \limsup \frac{1}{t} \sum_{s=0}^{t} 1 \\
&= \limsup \frac{t}{t} = 1; \\
& \quad E[\sum_{s=0}^{t_k} b(x^{g_{b,k}}(s), u^{g_{b,k}}(s))] = [1+1+1+\ldots+1] + \sum_{s=k+1}^{t} 1/k \\
&= (k-1) + \frac{t-k}{k}, \\
J_{ac}(g_{b,k}) &= \lim_{t\to\infty} [\frac{k-1}{t} + \frac{t-k}{tk}] = \frac{1}{k},\ 0 < J_{ac}(g_{b,k}) = \frac{1}{k}, \\
J^*_{ac,ti} &= \inf_{g\in G_{ti}} J(g) = \inf_{g_{b,k}\in G_{ti},\ k\in\mathbb{Z}_+} J(g_{b,k}) = \inf_{k\in\mathbb{Z}_+} \frac{1}{k} = 0.
\end{aligned}$$

Define the time-varying control law,

$$\begin{aligned}
& g \in G_{tv}, \\
& \forall\, k \in \mathbb{Z}_+,\ t_k = \frac{1}{2}k(k+1)+1,\ \ \{t_1, t_2, t_3, \ldots\} = \{1,\ 2,\ 4,\ 7,\ 11, \ldots\}, \\
g_{tv}(t,x) &= \begin{cases} u_1 \in U \text{ if } t = t_k, \\ u_2 \in U \text{ if } t = t_k+1, \ldots, t_{k+1}-1, \end{cases} \\
& \{g_{tv}(1,x),\ g_{tv}(2,x),\ \ldots,\} \\
&= \{u_1,\ u_1,\ u_2, u_1,\ u_2,\ u_2,\ u_1,\ u_2,\ u_2, u_2,\ u_1, \ldots\}.
\end{aligned}$$

The control law is in essence an open-loop control law in that it does not depend on the state of the system. The average cost of this time-varying control law is calculated,

$$\begin{aligned}
& \forall\, k \in \mathbb{Z}_+,\ t_k = \frac{1}{2}k(k+1)+1,\ T(0:t_k) = \{1, 2, \ldots, t_k\}, \\
& E[\sum_{s=0}^{t_k} [1+1+1/2+1/2+1+1/3+1/3+1/3+1+\ldots+1] \\
&= \sum_{s=0}^{k} (1+1) = 2(k+1), \\
J_{ac}(g_{tv}) &= \limsup_{t\to\infty} \frac{1}{t} E[\sum_{s=1}^{t-1} b(x^g(s), u^g(s))] = \lim_{k\to\infty} \frac{2(k+1)}{t_k} \\
&= \lim_{k\to\infty} \frac{2(k+1)}{\frac{1}{2}k(k+1)+1} = 0.
\end{aligned}$$

The conclusion is thus that,

$$J^*_{ac,ti} = \inf_{g \in G_{ti}} J_{ac,ti}(g) = 0, \ \ \forall \, g \in G_{ti}, \ 0 < J_{ac,ti}(g) < \infty,$$
$$J_{ac}(g_{tv}) = 0 = J^*_{ac,ti}.$$

The conclusion of this example is that any time-invariant control law has a strictly-positive average cost and is strictly larger than the infimal value, while there exists a time-varying control law achieving the infimal average cost of value zero.

Below attention is restricted to time-invariant control laws for average cost optimal stochastic control problems for a finite stochastic control system.

Within the set of time-invariant control laws, there are questions: (1) Does there exists a time-invariant control law which achieves the infimal value? (2) Is an optimal control law unique? In regard to uniqueness, an assumption of strict convexity of the cost rate is often assumed. For the case of a finite stochastic control system, in general the control law is not unique.

The reader may also think of a finite stochastic control system where there are at least two discrete states. A transition is possible from the first considered discrete state to the second one, and conversely. If the costs at both discrete states are the same and the associated transition costs are zero, then an optimal control law can include arbitrary state transitions from the first discrete state to the second and back. For example, the optimal control law may be time-varying meaning that it switches between the two states depending on the value of a specified time function. This is an example with a set of a time-varying optimal control laws all achieving the same cost.

The reader is expected to have read Section 18.6 on similarity of positive matrices to be able to appreciated the following results.

Control of a finite stochastic control system requires knowledge of the structure of such systems and how control affects this structure. One distinguishes the control synthesis into the following subsets of finite stochastic control systems:

1. the stochastic control system is such that for any value of the input, the state transition matrix is an irreducible and nonperiodic stochastic matrix;
2. the stochastic control system is such that there exists an input such that the closed-loop control system has a state set which has a decomposition into two or more terminal subsets each of which has an irreducible and nonperiodic stochastic matrix.

Consider first the special case of a finite stochastic control system of which the state transition matrix is irreducible and nonperiodic for all inputs.

Derivation of Dynamic Programming Equation

There follows an informal derivation of the dynamic programming equation of average cost for a finite stochastic control system. This discussion is to motivate the dynamic programming equation, afterwards there is presented a formal proof.

For the finite-horizon $T(0:t_1)$ optimal stochastic control problem with cost function, $E[\sum_{s=0}^{t_1} b(x(s),u(s))]$, the value function satisfies the dynamic programming equation,

$$V(t,x) = \min_{u\in U} \left\{ b(x,u) + \sum_{x_1\in X} P(x_1,x,u)V(t+1,x_1) \right\}.$$

Reverse the time axis by defining, $\overline{V} : T \times X \to \mathbb{R}$, $\overline{V}(t,x) = V(t_1 - t,x)$. Then $\overline{V}$ satisfies the forward equation,

$$\begin{aligned}
\overline{V}(t,x) &= \min_{u\in U} \left\{ b(x,u) + \sum_{x_1\in X} P(x_1,x,u)\overline{V}(t-1,x_1) \right\};\\
&\overline{V}(t,x) + \frac{1}{t}\overline{V}(t,x) = \min_{u\in U} \qquad (13.9)\\
&= \{b(x,u)(1+1/t) + \sum_{x_1\in X} P(x_1,x,u) \times (\overline{V}(t-1,x_1) + \frac{1}{t}\overline{V}(t-1,x))\}.
\end{aligned}$$

Assume that there exists a $V : X \to R$ and $J^*_{ac} \in R$ such that for all $x \in X$, $\lim_{t\to\infty}[\overline{V}(t,x) - tJ^*_{ac}] = V(x)$. Then,

$$\exists\, V : X \to \mathbb{R},\ J^*_{ac} \in \mathbb{R}_+,\ \forall\, x \in X,\ \lim_{t\to\infty} \frac{1}{t}\overline{V}(t,x) = J^*_{ac}, \qquad (13.10)$$

which limit does not depend on $x \in X$. Further,

$$\begin{aligned}
J^*_{ac} + V(x) &= \lim_{t\to\infty} [\frac{1}{t}\overline{V}(t,x) + (\overline{V}(t,x) - tJ^*_{ac})],\\
&\quad \text{by definition of } V \text{ and Equation (13.10)},\\
&= \lim_{t\to\infty} \min_{u\in U}\ \{b(x,u)(1+1/t)+\\
&\quad + \sum_{x_1\in X} P(x_1,x,u)(\overline{V}(t-1,x_1) - (t-1)J^*_{ac} + \frac{1}{t}\overline{V}(t,x) - J^*_{ac})) \Bigg\},\\
&\quad \text{by Equation (13.9)},\\
&= \min_{u\in U}\ \{b(x,u) + \sum_{x_1\in X} P(x_1,x,u)V(x_1)\},
\end{aligned}$$

because interchange of the limit and the minimization operation is allowed by the finiteness of the input space U and by definition of V. Hence one obtains the dynamic programming equation for (J^*,V).

To show that the dynamic programming approach is useful the following questions must be answered:

- Does there exist a pair (V,J^*_{ac}) satisfying the dynamic programming equation? In general, the dynamic programming equation of average cost does not have a unique solution.
- Can the optimal control law $g^* \in G$ be determined from the value function?
- How to compute the solution (V,J^*_{ac}) of the dynamic programming equation?

These questions are answered below.

The dynamic programming equation of average cost in case of a finite stochastic control system, can be transformed into an equation for a vector. Introduce the following notation for the case of a finite state set. Relate the function $V : X \rightarrow \mathbb{R}^{n_x}$ to a vector $V \in \mathbb{R}_+^{n_x}$, and similarly relate the function $g \in G_M$ to the vector g, and relate the cost rate to the vector $b(g)$ according to,

$$V = \begin{pmatrix} V(x_1) \\ V(x_2) \\ \vdots \\ V(x_{n_x}) \end{pmatrix}, \; g = \begin{pmatrix} g(x_1) \\ g(x_2) \\ \vdots \\ g(x_{n_x}) \end{pmatrix}, \; b(g) = \begin{pmatrix} b(x_1, g(x_1)) \\ b(x_2, g(x_2)) \\ \vdots \\ b(x_{n_x}, g(x_{n_x})) \end{pmatrix},$$

$$A(g) \in \mathbb{R}_+^{n_x \times n_x}, \; A(g)_{i,j} = P(\{x^+ = i\} | \{x = j, \; u = g(j)\}),$$
$$p_{x_0} \in \mathbb{R}_+^{n_x}, \; p_{x_0}(i) = P(\{x_0 = i\}); \text{ consequently,}$$
$$p_{x(t)} = A(g)^t p_{x_0} \in \mathbb{R}_{st}^{n_x}.$$

The expected cost on a finite horizon is then,

$$E[\sum_{s=0}^{t-1} b(x(s), g(x(s)))] = \sum_{s=0}^{t-1} b(g)^T A(g)^s p_{x_0} = b(g)^T (\sum_{s=0}^{t-1} A(g)^s) p_{x_0}.$$

The average cost is then finite and, due to the definition of b_{max} in Def.13.2.17,

$$\frac{1}{t} \, E[\sum_{s=0}^{t-1} b(x(s), g(x(s)))] \leq b_{max}, \; \forall \, t \in \mathbb{Z}_+,$$
$$J(g) = \lim_{t \rightarrow \infty} \frac{1}{t} \, E[\sum_{s=0}^{t-1} b(x(s), g(x(s)))] = \lim_{t \rightarrow \infty} b(g)^T (\frac{1}{t} \sum_{s=0}^{t-1} A(g)^s) p_{x_0}$$
$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{s=0}^{t-1} A(g)^s = Q^* s(g); \text{ by Theorem 18.8.17,}$$
$$J(g) = b(g)^T Q^*(g) p_{x_0}.$$

See Section 18.8 for the matrix $P_s(g)$.

Dynamic Programming

Definition 13.2.20. Consider the finite stochastic control system problem of Def.13.2.17 and the optimal stochastic control problem, Problem 13.2.18.

Define the dynamic programming equation of a finite stochastic control system and of average cost for the value $J^* \in \mathbb{R}$ and for the value function $V : X \rightarrow \mathbb{R}$, by the formulas,

$$J^* + V(x) = \min_{u \in U(x)} \{b(x,u) + \sum_{x_1 \in X} P(x_1, x, u) V(x_1)\}, \forall \, (x,u) \in X \times U, \quad (13.11)$$
$$P(x_1, x, u) = P(\{x(t+1) = x_1\} | \{x(t) = x, \; u(t) = u\}) \in \mathbb{R}_+.$$

Procedure 13.2.21 *The* dynamic programming procedure for average cost and for a finite stochastic control system.
Input: $n_x \in \mathbb{Z}_+$*,* $b : X \times U \to \mathbb{R}_+$ *and* $\{P(x_1, x, u) \in \mathbb{R}_+,\ \forall\, x_1, x \in X,\ u \in U\}$.

1. *Solve the* dynamic programming equation of average cost

$$J^* + V(x) = \min_{u \in U(x)} \{b(x,u) + \sum_{x_1 \in X} P(x_1, x, u) V(x_1)\},$$

for the value function $V : X \to \mathbb{R}$ *and for the value* $J^* \in \mathbb{R}$*. Procedure 13.2.26 provides a policy iteration procedure to produce a solution of this equation.*

2. *Define the optimal control law according to,*

$$g^*(x) = \arg \min_{u \in U(x)} \{b(x,u) + \sum_{x_1 \in X} P(x_1, x, u) V(x_1)\},\ g^* : X \to U. \tag{13.12}$$

3. *Output the triple* $(J^*,\ V,\ g^*)$.

Theorem 13.2.22. A sufficient condition for optimality.
Consider Problem 13.2.18. If

1. *there exists a solution* (J^*, V) *of the Dynamic Programming Equation (13.11);*
2. *the control law* $g^* \in G$ *is produced by Procedure 13.2.21;*

then $g^* \in G$ *is an optimal control law and* $J^* \in \mathbb{R}_+$ *is the value of the problem.*

Proof. The result follows from Theorem 13.2.7 if the conditions of that theorem are satisfied. Because the value function V is defined on a finite set, there exists a bound $b_V \in \mathbb{R}_+$ such that, for all $x \in X$, $|V(x)| \le b_V$. Moreover, $\lim_{t \to \infty} E[V(x(t))]/t = 0$ and $E[V(x_0)] \le b_V < \infty$. Because the input set and the state set are finite sets the infima are actually minima hence the values are attained. The conditions of the theorem are then satisfied and the conclusions of the theorem hold. □

Theorem 13.2.23. *Consider Problem 13.2.18. Assume that, for all stationary Markov control laws* $g \in G_{SM}$*, the transition matrix* $P(g)$ *is irreducible.*

(a)*There exists a tuple* (J^*, V) *which is a solution of the dynamic programming equation,*

$$J^* + V(x) = \min_{u \in U} \{b(x,u) + \sum_{x_1 \in X} P(x_1, x, u) V(x_1)\},\ \forall\, x \in X.$$

(b)*The solution* V *of the dynamic programming equation is unique up to an additive constant. Thus, if* V *is a solution and* $a \in \mathbb{R}$*, then* $V + a$ *is also a solution, while if* V_1*,* V_2 *are solutions then* $V_1 - V_2 = a$ *for an element* $a \in \mathbb{R}$*. However,* $J^* \in \mathbb{R}_+$ *is unique.*

(c)Sufficient condition for optimality. *Let* g^* *be as constructed in Step (2) of Procedure 13.2.21. Then* $g^* \in G_{SM}$ *is an optimal control law and* J^* *is the value.*

*(d)*Necessary condition for optimality. *If $g_1 \in G_{SM}$ is such that ,*

$$J(g_1) = \inf_{g \in G} J(g),$$

then the dynamic programming equation holds with $V(g_1)$, defined in Proposition 13.2.24, and g_1 attains the minimum in Step (2) of Procedure 13.2.21.

The proof of Theorem 13.2.23 is based on the following results.

It follows from the assumption that for every Markov control law $g \in G_{SM}$ the state transition matrix $A(g)$ is irreducible, and from Theorem 18.8.2 that there exists a unique stochastic vector $p(g) \in \mathbb{R}^{n_x}_{st}$ such that $A(g)p(g) = p(g)$. Then the average cost of that stationary Markov control law satisfies the formula, see Theorem 18.8.17,

$$J(g) = b(g)^T Q^*(A(g))\ p_{x_0} = b(g)^T p(g), \text{ where,} \tag{13.13}$$

$$A(g)p(g) = p(g), \ \ p(g) \in \mathbb{R}^{n_x}_{st}, \tag{13.14}$$

where $p(g) \in \mathbb{R}^{n_x}_{st}$ is the unique solution of equation (13.14). From Equation (13.13) follows that the average cost does not depend on the distribution p_0 of x_0. The results of this discussion are summarized below.

Proposition 13.2.24. *Let $g \in G_{SM}$ be a stationary Markov control law. Assume that $A(g) \in \mathrm{R}^{n \times n}_{+}$ is irreducible.*

(a)There exists an unique $p(g)$ such that,

$$\begin{aligned}
&\exists\ p(g) \in \mathbb{R}^{n_x}_{s+,st}\ A(g)p(g) = p(g); \\
&\exists\ Q^*(A(g)) \in \mathbb{R}^{n_x \times n_x}_{st} \text{ such that} \\
Q^*(A(g)) = &\lim_{t \to \infty} \frac{1}{t} \sum_{s=0}^{t-1} A(g)^s, \\
&\exists\ p^*(g) \in \mathbb{R}^{n)x}_{s+,st}\ Q^*(A(g))\ p^*(g) = p^*(g).
\end{aligned}$$

(b)The average cost associated with the control law $g \in G$ is, $J(g) = b(g)^T p^*(g)$.
(c)There exists a $V(g) \in \mathbb{R}^{n_x}_{+}$ such that $J(g)1_{n_x} + V(g) = b(g) + A(g)V(g)$.

Proof. (a) The existence of $p(g)$ follows from Theorem 18.8.2. The existence of the limit of the averaged powers of $A(g)$ follows from Theorem 18.8.17. The existence of the vector $p^*(g)$ follows from Theorem 18.8.17.(c).
(b) This follows from the above discussion. (c) By (b)

$$J(g) = p(g)^T b(g) \Leftrightarrow [J(g)1_{n_x} - b(g)]^T p(g) = 0, \tag{13.15}$$

or $p(g)$ is orthogonal to $[J(g)1 - b(g)]$. By (a), $p(g)$ is the unique solution of,

$$\begin{aligned}
A(g)p(g) \ = \ p(g), 1^T p(g) = 1 \ &\Leftrightarrow \ [A(g) - I]p(g) = 0, \ 1^T p(g) = 1; \\
&\left\{\ [J(g)1_{n_x} - b(g)]^T p(g) = 0, \ [A(g) - I]p(g) = 0\ \right\} \\
&\Rightarrow \exists\ V(g) \in \mathbb{R}^{n_x} \text{ such that, } J(g)1_{n_x} - b(g) = [A(g) - I]V(g).
\end{aligned}$$

The latter equation is the dynamic programming equation. The implication ⇒ used above holds because equation (13.15) has a unique solution $p(g)$ and because $[A(g)-I]p(g)=0$. □

Lemma 13.2.25. *Assume that:*

1. *For any stationary Markov control law $g \in G_{SM}$, $A(g) \in \mathbb{R}_{st}^{n_x \times n_x}$ is irreducible.*
2. *There exists a $g^* \in G_{SM}$ such that,*

$$J(g^*) = \inf_{g \in G_{SM}} J(g).$$

Then there exists a tuple $(J(g^*), V(g^*,.))$ which is a solution to the dynamic programming equation,

$$\begin{aligned} J(g^*) + V(g^*,x) &= \min_{u \in U}\{b(x,u) + \sum_{x_1 \in X} P(x_1,x,u)V(g^*,x_1), \ \forall\, x \in X; \\ J(g^*) + V(g^*,x) &= b(x,g^*(x)) + \sum_{x_1 \in X} P(x_1,x,g^*(x))V(g^*,x_1), \ \forall\, x \in X. \end{aligned}$$

Proof. From Proposition 13.2.24.(c) there exists a $V(g^*,.) : X \to \mathbb{R}$ and $J(g^*)$ such that for all $x \in X$,

$$\begin{aligned} J(g^*) + V(g^*,x) &= b(x,g^*(x)) + \sum_{x_1 \in X} P(x_1,x,g^*(x))V(g^*,x_1); \\ J(g^*) + V(g^*,u) &\geq \min_{u \in U} \ \{b(x,u) + \sum_{x_1 \in X} P(x_1,x,u)V(g^*,x_1)\}, \ \forall\, x \in X. \end{aligned} \tag{13.16}$$

Define,

$$g_1 : X \to U, \ \ g_1(x) = \arg\min_{u \in U} \ \{b(x,u) + \sum_{x_1 \in X} P(x_1,x,u)V(g^*,x_1)\}.$$

Note that by Theorem 18.8.2, $p(g_1)(x) > 0$ for all $x \in X$. Suppose that for some $x_2 \in X$

$$J(g*) + V(g*,x_2) > \min_{u \in U}\{b(x_2,u) + \sum_{x_1 \in X} P(x_1,x_2,u)V(g*,x_1)\}. \tag{13.17}$$

Multiplying Equation (13.17) by $p(g_1)(x_0) > 0$ and summing over $x_0 \in X$ yields,

$$
\begin{aligned}
& J(g^*) + \sum_{x_0 \in X} p(g_1) V(g^*, x_0) \\
& > \min_{u \in U} \sum_{x_0 \in X} \{ p(g_1)(x_0) b(x_0, u) + \sum_{x_0 \in X} p(g_1)(x_0) \sum_{x_1 \in X} P(x_1, x_0, u) V(g^*, x_1) \} \\
& = \sum_{x_0 \in X} p(g_1)(x_0) b(x_0, g_1(x_0)) + \\
& \quad + \sum_{x_0 \in X} p(g_1)(x_0) \sum_{x_1 \in X} P(x_1, x_0, g_1(x_0)) V(g^*, x_1), \\
& \quad \text{by definition of } g, \\
& = J(g_1) + \sum_{x_1 \in X} p(g_1)(x_1) V(g^*, x_1), \text{ by Proposition 13.2.24.(a),} \\
J(g_1) & < J(g^*),
\end{aligned}
$$

contradicting Assumption (2). Hence equality holds in equation (13.16). □

Proof. Proof of Theorem 13.2.23. (a & c) Because the state space X and input space U are finite sets, the set of stationary Markov control laws G_{SM} is a finite set. Therefore there exists a $g^* \in G_{SM}$ such that, $J(g^*) = \min_{g \in G_{SM}} J(g)$. From Lemma 13.2.25 follows that there exists a pair $V(g^*, .) : X \to R$ and $J(g^*)$ that is a solution of the dynamic programming equation. Moreover, g^* attains the minimum in Equation (13.12).
(b) This follows from Proposition 13.2.5.
(d) This follows from Lemma 13.2.25. □

Computation of an Optimal Control Law

There exists both a *control law iteration procedure*, also called a policy iteration procedure, and a *value iteration procedure* to determine the value function and the optimal control law from the dynamic programming equation of average cost for a finite stochastic control system. Only the first procedure is discussed in this section.

The computations for an example are preceded by an introduction to the notation. Note that

$$
\begin{aligned}
& E[V(t+1, f(x, u, v(t))) | F^{x,u}] = E[\sum_{k=1}^{n_x} V(x_k) I_{\{x(t+1) = x_k\}} | F^{x,u}] \\
& = \sum_{k=1}^{n_x} V(x_k)\, P(\{x(t+1) = x_k\} | \{x(t) = x,\ u(t) = u\}) \\
& = \sum V(x_k) (A(u)x)_k = V^T\, A(u) x,\ \ V = \begin{pmatrix} V(x_1) \\ V(x_2) \\ \vdots \\ V(x_{n_x} \end{pmatrix} \in \mathbb{R}^{n_x}.
\end{aligned}
$$

Procedure 13.2.26 Control law iteration for average-cost dynamic programming of a finite stochastic control system.
(In the literature also called the policy iteration procedure.)
Data: the cost rate $b : X \times U \to \mathbb{R}_+$, the state transition function $A : U \to \mathbb{R}_{st}^{n_x \times n_x}$, and the initial control law $g_0 : X \to U$ which is a stationary Markov control law.

1. *Initialization. Let $m = 0$. Solve the equation,*
   $$J(g_0) + V_{g_0}(x) = b(x, g_0(x)) + V_{g_0}^T A(g_0(x))x, \ \forall\, x \in X,$$
   for $(J(g_0),\ V_{g_0}) \in \mathbb{R} \times \mathbb{R}^{n_x}$.
2. *For $m := m + 1$ while,*
   $$\exists\, x \in X, \text{ such that}$$
   $$J(g_{m-1}) + V_{g_{m-1}}(x) > \min_{u \in U} \{b(x,u) + V_{g_{m-1}}^T A(u)x\} \text{ compute}$$
   (2.1) $g_m(x) = \arg\min_{u \in U} \{b(x,u) + V_{g_{m-1}}^T A(u)x\}, \ \forall\, x \in X;$

   (2.2) *solve the following equation for* $(J(g_m), V_{g_m}) \in \mathbb{R} \times \mathbb{R}^{n_x}$,
   $$J(g_m) + V_{g_m}(x) = b(x, g_m(x)) + V_{g_m}^T A(g_m(x))x, \ \forall\, x \in X.$$
3. *Define the optimal control law as $g^* = g_m$.*

In Step 1 and Step (2.2), the solution may be computed by first subtracting the row of x_{n_x} from all other rows. The effect of this is that the variable $J(g_0)$ is eliminated from the first $n_x - 1$ equations. Next set $V(x_x) = 0$ which can be done because the solution is up to an additive constant. If the stochastic control system is irreducible then the set of $n_x - 1$ equations has a unique solution which can be computed by a linear algebra program. If you like you may add the smallest component of the vector $V(g_0)$ to all components such that $V(g_0) \in \mathbb{R}_+^{n_x}$. Finally compute $J(g_0) \in \mathbb{R}$ from the equation of $x_1 \in X$. It is not known to the author whether there exists a standard software package for this procedure.

Proposition 13.2.27. *[7, Proposition 8.8, p. 352]. Assume that for all input values $u \in U$ the transition matrix $A(g) \in \mathbb{R}_{st}^{n_x \times n_x}$ is irreducible. Then the policy iteration procedure yields an optimal control law in a finite number of steps.*

Example 13.2.28. *Computation of an optimal control law by control law iteration.* Consider a finite stochastic control system with the representation

$$X = \{x_1, x_2, x_3\}, \ U = \{u_1, \ u_2\},$$
$$E[x(t+1)|F_t] = A(u(t))x(t),$$
$$A(u_1) = \begin{pmatrix} 0.2\ 0.2\ 0.4 \\ 0.6\ 0.8\ 0.0 \\ 0.2\ 0.0\ 0.6 \end{pmatrix}, \ A(u_2) = \begin{pmatrix} 0.2\ 0.2\ 0.4 \\ 0.2\ 0.8\ 0.0 \\ 0.6\ 0.0\ 0.6 \end{pmatrix} \in \mathbb{R}_{st}^{3 \times 3}.$$

The control system has three states. From state x_1 one proceeds either to state x_2 or to state x_3 or one stays at state x_1. From state x_2 and from state x_3 one returns to state x_1 with a probability, the probabilities differ by state. The latter property makes the system asymmetric. The user can influence the stochastic control system by the choice of the input at state x_1, with u_1 there is a higher probability to go to state x_2 than to state x_3, with the input u_2 the probabilities are reversed. The inputs have no effect at the states x_2 and x_3 to keep the problem simple. Note that for all $u \in U$, the stochastic matrix $A(u)$ is irreducible.

The average cost function is defined as

$$J(g) = \limsup_{t_1 \to \infty} \frac{1}{t_1} E[\sum_{s=0}^{t_1-1} b(x(s),u(s))],\ b(x,u) = \begin{cases} 1, \text{ if } x = x_1,\ \forall\, u \in U, \\ 2, \text{ if } x = x_2,\ \forall\, u \in U, \\ 5, \text{ if } x = x_3,\ \forall\, u \in U. \end{cases}$$

The policy iteration procedure is applied.

Initialization. Let $g_a : X \to U$ be $g_a(x) = u_2$ for all $x \in X$.

Solve the equations

$$J(g_0) + V_{g_a}(x_1) = b(x, g_a(x)) + V_{g_a}^T A(g_a(x))x,\ \ \forall\, x \in X,\ V_{g_a}(x_3) = 0,$$

for $(J(g_a), V_{g_a})$. The condition $V_{g_a}(x_3) = 0$ is imposed because the solution is up to an additive constant.

Substract the equation of x_3 from that of x_1 so that one obtains

$$\begin{aligned} & V_{g_a}^T \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} = V_{g_a}(x_1) - V_{g_a}(x_3) \\ &= b(x_1, g_a(x_1)) - b(x_3, g_a(x_3)) + V_{g_a}^T A(g_a(x))x_1 - V_{g_a}^T A(g_a(x))x_3 \\ &= 1 - 5 + V_{g_a}^T \left[\begin{pmatrix} 0.2 \\ 0.2 \\ 0.6 \end{pmatrix} - \begin{pmatrix} 0.4 \\ 0 \\ 0.6 \end{pmatrix} \right] \ \Leftrightarrow\ \begin{pmatrix} 1.2 & -0.2 & -1 \end{pmatrix}^T V_{g_a} = -4. \end{aligned}$$

Similarly for the second equation,

$$\begin{aligned} & V_{g_a}^T \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} = V_{g_a}(x_2) - V_{g_a}(x_3) \\ &= b(x_1, g_a(x_2)) - b(x_3, g_a(x_3)) + V_{g_a}^T A(g_a(x))x_2 - V_{g_a}^T A(g_a(x))x_3 \\ &= 2 - 5 + V_{g_a}^T \left[\begin{pmatrix} 0.2 \\ 0.8 \\ 0 \end{pmatrix} - \begin{pmatrix} 0.4 \\ 0 \\ 0.6 \end{pmatrix} \right] \ \Leftrightarrow\ \begin{pmatrix} 0.2 & 0.2 & -0.4 \end{pmatrix}^T V_{g_a} = -3. \end{aligned}$$

The combined set of equations then has the solution

$$\begin{pmatrix} 1.2 & -0.2 \\ 0.2 & 0.2 \end{pmatrix} \begin{pmatrix} V_{g_a}(x_1) \\ V_{g_a}(x_2) \end{pmatrix} = \begin{pmatrix} -4 \\ -3 \end{pmatrix},\ V_{g_a} = \begin{pmatrix} -5 \\ -10 \\ 0 \end{pmatrix} \ \Rightarrow V_{g_a} = \begin{pmatrix} 5 \\ 0 \\ 10 \end{pmatrix};$$

where the latter vector is also a solution of the equation because V_{g_a} is unique upto an additive constant. Hence

$$\begin{aligned} J(g_a) &= -V_{g_a}(x_1) + b(x_1, g_a(x_1)) - A(g_a(x_1))x_1 \\ &= -5 + 1 + V_{g_a}^T \begin{pmatrix} 0.2 \\ 0.2 \\ 0.6 \end{pmatrix} = -5 + 1 + 7 = 3. \end{aligned}$$

The reader can verify that the solution $(J(g_a),\ V_{g_a})$ satisfies the three equations.

For V_{g_a} fixed, one determines a new control law g_b according to

$$\begin{aligned} h^*(x) &= \min_{\{u \in U = \{u_1, u_2\}} [b(x,u) + V_{g_a}^T A(u)x] \\ &= \min \left\{ b(x,u_1) + \begin{pmatrix} 5 \\ 0 \\ 10 \end{pmatrix}^T \begin{pmatrix} 0.2 & 0.2 & 0.4 \\ 0.6 & 0.8 & 0.0 \\ 0.2 & 0.0 & 0.6 \end{pmatrix}, \right. \\ &\qquad \left. b(x,u_2) + \begin{pmatrix} 5 \\ 0 \\ 10 \end{pmatrix}^T \begin{pmatrix} 0.2 & 0.2 & 0.4 \\ 0.2 & 0.8 & 0.0 \\ 0.6 & 0.0 & 0.6 \end{pmatrix} \right\}, \\ h^*(x_1) &= \min\{1+3,\ 2+7\} = 4,\ g_b^*(x_1) = 1, \\ h^*(x_2) &= \min\{2+1,\ 1+2\} = 3,\ g_b(x_2) = 1, \\ h^*(x_3) &= \min\{5+8,\ 5+8\} = 13,\ g_b(x_3) = 1, \end{aligned}$$

where it is noted that the control law at the states x_2 and x_3 can be either u_1 or u_2 though that has no effect on the dynamics as is clear from the stochastic control system.

Next the value function V_{g_b} has to be computed. Note that,

$$\begin{aligned} & V_{g_b}^T \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} = b(x_1, g_b(x_1)) - b(x_3, g_b(x_3)) + V_{g_b}^T \left[\begin{pmatrix} 0.2 \\ 0.6 \\ 0.2 \end{pmatrix} - \begin{pmatrix} 0.4 \\ 0.0 \\ 0.6 \end{pmatrix} \right] \\ \Leftrightarrow\ & \begin{pmatrix} 1.2 & -0.6 & -0.6 \end{pmatrix} V_{g_b} = -4; \\ & V_{g_b}^T \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} = b(x_2, g_b(x_2)) - b(x_3, g_b(x_3)) + V_{g_b}^T \left[\begin{pmatrix} 0.2 \\ 0.8 \\ 0.0 \end{pmatrix} - \begin{pmatrix} 0.4 \\ 0.0 \\ 0.6 \end{pmatrix} \right] \\ \Leftrightarrow\ & \begin{pmatrix} 0.2 & 0.2 & -0.4 \end{pmatrix} V_{g_b} = -3; \end{aligned}$$

$$\begin{aligned} & \begin{pmatrix} 1.2 & -0.6 \\ 0.2 & 0.2 \end{pmatrix} \begin{pmatrix} V_{g_b}(x_1) \\ V_{g_b}(x_2) \end{pmatrix} = \begin{pmatrix} -4 \\ -3 \end{pmatrix}, \\ V_{g_b} &= \begin{pmatrix} -65/9 \\ -70/9 \\ 0 \end{pmatrix} \Rightarrow V_{g_b} = \begin{pmatrix} 5/9 \\ 0 \\ 70/9 \end{pmatrix}; \\ J(g_b) &= -V_{g_b}(x_1) + b(x_1, g_b(x_1)) + V_{g_b}^T A(g_b(x_1))x_1 \\ &= -5/9 + 1 + 15/9 = 2 + 1/9 < 3 = J(g_a). \end{aligned}$$

A further computation shows that g_b is the optimal control law with $(J^*, V) = (J(g_b), V_{g_b})$.

13.3 Discounted Cost

13.3.1 Positive Cost

The reader finds in this section a brief summary of optimal stochastic control on an infinite horizon with discounted cost and with a positive cost-rate. The reader may find references with the corresponding proofs in Sect. 13.6.

Problem 13.3.1. *Optimal stochastic control problem for a recursive stochastic control system with complete observations on an infinite-horizon with a discounted cost function.* Consider the time-invariant recursive stochastic control system with a controlled output process,

$$\begin{aligned} x(t+1) &= f(x(t),u(t),v(t)),\ x(0)=x_0, \\ z(t) &= h(x(t),u(t)),\ \forall\, t \in T\backslash\{t_1\},\ z(t_1)=h(x(t_1)), \\ & T=\mathbb{N},\ X=\mathbb{R}^{n_x},\ U=\mathbb{R}^{n_u},\ Z=\mathbb{R}^{n_z},\ n_x,\ n_u,\ n_v,\ n_z \in \mathbb{Z}_+, \\ & v:\Omega\times T \to \mathbb{R}^{n_v},\ \{v(t),\ t\in T\} \text{ independent sequence}, \\ & x_0:\Omega \to \mathbb{R}^{n_x},\ F^{x_0},\ \ F^{v}_{\infty} \text{ independent}. \end{aligned}$$

Define the set of time-varying control laws G_{tv}, that of time-varying Markov control laws $G_{tv,M}$, and the set of time-invariant control laws G_{ti} according to,

$$\begin{aligned} G_{tv} &= \left\{ \begin{array}{l} g=(g_0,g_1(.),\ g_2(.)\ldots | \\ \forall\, t\in T,\ g_t: X^{t+1}\to U,\ \text{Borel measurable} \end{array} \right\},\ g_t(x(0:t)), \\ G_{tv,M} &= \left\{ \begin{array}{l} g=(g_0,g_1(.),\ g_2(.)\ldots | \\ \forall\, t\in T,\ g_t: X\to U,\ \text{Borel measurable} \end{array} \right\},\ g_t(x(t)), \\ G_{ti} &= \left\{ g: X\to U \text{ Borel measurable} \right\},\ g(x(t)), \\ & \forall\, g\in G_{ti} \text{ define},\ \ g_T = \left\{ (g,g,\ldots) \right\} \in G_{tv,M}. \end{aligned}$$

Construct, for any $g \in G_{tv}$, the closed-loop system according to,

$$\begin{aligned} x^g(t+1) &= f(x^g(t),g_t(x^g(0:t)),v(t)),x^g(0)=x_0, \\ u^g(t) &= g_t(x^g(0:t)) = g_t((x^g(0),x^g(1),\ldots,x^g(t))). \end{aligned}$$

Define the *discounted cost* on the infinite horizon as,

$$\begin{aligned} & J_{dc}: G_{tv} \to \mathbb{R}_+ \cup \{\infty\},\ b: X\times U \to \mathbb{R}_+, \\ J_{dc}(g) &= \limsup_{t_1\to\infty}\ E\left[\sum_{s=0}^{t_1-1} r^s\ b(x^g(s),u^g(s)) \right]. \end{aligned}$$

Denote by $G_{dcf} \subseteq G$ the subset of control laws achieving a finite cost and define the optimal stochastic control problem as,

$$\begin{aligned} & \inf_{g\in G_{tv,dcf}} J_{ac}(g); \\ & G_{tv,dcf} = \{g\in G_{tv} |\ J_{dc}(g)<\infty\},\ G_{ti,dcf} = \{g\in G_{ti} |\ J_{dc}(g)<\infty\}. \end{aligned}$$

The reader should notice that the set of control laws G_{tv} is such that the control law g_t used at time $t \in T$ can depend on the entire past of the state process x^g hence g_t depends on time and is thus time-varying.

Definition 13.3.2. Consider Problem 13.3.1. Define the *discounted cost dynamic programming equation* for a function V by the equation,

$$V(x_V) = \inf_{u_V \in U(x_V)} \{b(x_V, u_V) + r\, E[V(f(x_V, u_V, v))|\, F^{x_V, u_V}]\},\ V : X \to \mathbb{R}_+.$$

In general the set of solutions of the discounted cost dynamic programming equation has two or many solutions. If an infimal solution exists in the set of postively-valued measurable value functions then denote it by V^*, hence $V^*(x_V) \leq V(x_V)$ for all $x_V \in X$ and for all measurable solutions V of the discounted cost dynamic programming equation.

Denote the *discounted cost dynamic programming operator* and the *dynamic programming operator associated with the particular control law* $g \in G_{ti}$, respectively by,

$$DP_{dc}(V)(x_V) = \inf_{u_V \in U(x_V)} \{b(x_V, u_V) + r\, E[V(f(x_V, u_V, v))|\, F^{x_V, u_V}]\},$$
$$DP_{dc,g}(V)(x_V) = b(x_V, g(x)) + r\, E[V(f(x_V, g(x_V), v))|\, F^{x_V, u_V}].$$

Theorem 13.3.3. *Consider the problem of optimal stochastic control with a discounted cost, Problem 13.3.1.*

(a)If there exists a function V^* which is an infimal solution of the discounted cost dynamic programming equation, then,

$$E[V^*(x_0)] \leq J_{dc}(g),\ \forall\, g \in G_{tv}.$$

(b)If, in addition to (a), there exists a measurable control law $g^* \in G_{tv}$ which attains the infima in,

$$\begin{aligned} & b(x_v, g^*(x_V)) + E[V^*(f(x_V, g^*(x_V), v)|\, F^{x_V, g^*(x_v)}] \\ = & \inf_{u_V \in U(x_V)} \{b(x_V, u_V) + r\, E[V^*(f(x_V, u_V, v))|\, F^{x_V, u_V}]\},\ g^*(x_V) \in U(x_V), \\ & \text{then } E[V^*(x_0)] = J_{dc}(g^*). \end{aligned}$$

Theorem 12.4.1 provides sufficient conditions for existence of a minimizer $g^*(x_V) \in U(x_V)$.

Procedure 13.3.4 Construction of the infimal value function. *Consider the problem of optimal stochastic control with a discounted cost, Problem 13.3.1.*

Define the sequence of value functions,

$$V_0(x_V) = 0,\ \forall\, x_v \in V;$$
$$V_{k+1}(x_V) = DP_{dc}(V_k)(x_V),\ V_k : X \to \mathbb{R}_+,\ \{V_k,\ k \in \mathbb{Z}_+\}.$$

This sequence is only well defined if V_0 equals zero for all its arguments.

Proposition 13.3.5. *Consider the problem of optimal stochastic control with a discounted cost, Problem 13.3.1. A condition of stochastic controllability is assumed to hold. This condition has to be formulated along the lines for the stochastic controllability of a Gaussian stochastic control system.*

Then there exists a measurable value function $V_\infty : X \rightarrow \mathbb{R}_+$ which is a solution of the dynamic programming equation and satisfies,

$$\begin{aligned} V_\infty(x_V) &= \lim_{k\rightarrow\infty} V_k(x_V),\ \forall\, x_V \in X; \\ V_\infty &= DP_{dc}(V_\infty);\ V^* = V_\infty. \end{aligned}$$

Moreover, V_∞ is the infimal solution of the discounted cost dynamic programming equation.

The solution procedure for the optimal stochastic control problem with discounted cost is thus: (1) First check whether a condition of stochastic controllability holds. (2) Second construct a solution of the discounted cost dynamic programming equation according to the above procedure. (3) Thirdly, construct the optimal discounted-cost optimal control law according to Theorem 13.3.3 The above steps are carried out below for control of a Gaussian stochastic control system.

13.3.2 Control of a Gaussian Stochastic Control System

Problem 13.3.6. *Optimal stochastic control problem of a time-invariant Gaussian stochstic control system with complete observations on an infinite-horizon with the discounted cost function*.

Consider a time-invariant Gaussian stochastic control system representation,

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) + Mv(t) = f(x(t),u(t),v(t)),\ x(0) = x_0, \\ z(t) &= C_z x(t) + D_z u(t), \\ &\quad \text{rank}(D_z) = n_u \ \Rightarrow\ 0 \prec D_z^T D_z. \end{aligned}$$

Consider the past-state information pattern and the corresponding set G of control laws. Consider the discounted cost criterion,

$$\begin{aligned} J_{dc}(g,x_0) &= \overset{t_1\rightarrow\infty}{\lim\sup}\ E\left[\sum_{s=0}^{t_1} r^s z^g(t)^T z_g(t)|F^{x_0}\right],\ J_{dc} : G \times X \rightarrow \mathbb{R}_+,\ r \in (0,1), \\ J_{dc}(g) &= E[J_{dc}(g,x_0)],\ \ J_{dc} : G \rightarrow \mathbb{R}_+, \\ b(x,u) &= z^T z = \begin{pmatrix} x \\ u \end{pmatrix}^T \begin{pmatrix} C_z^T C_z & C_z^T D_z \\ D_z^T C_z & D_z^T D_z \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix} = \begin{pmatrix} x \\ u \end{pmatrix}^T Q_{cr} \begin{pmatrix} x \\ u \end{pmatrix}, \\ &\quad Q_{cr} \in \mathbb{R}_{pds}^{(n_x+n_u)\times(n_x+n_u)}. \end{aligned}$$

The problem is to solve the optimal stochastic control problem,

$$\inf_{g\in G} J_{dc}(g).$$

Below the control algebraic Riccati equation of discounted cost is needed which differs from that of the control algebraic Riccati equation of average cost. It is stated explicitly to avoid confusion.

Theorem 13.3.7. *Consider Problem 13.3.6. Define the matrices*

$$\begin{aligned} A_c &= A - B(D_z^T D_z)^{-1} D_z^T C_z \in \mathbb{R}^{n_x \times n_x}, \\ C_c^T C_c &= C_z^T C_z - C_z^T D_z (D_z^T D_z)^{-1} D_z^T C_z, \; C_c \in \mathbb{R} \end{aligned}$$

(a)If (A,B) is a stabilizable pair and (A_c,C_c) is a detectable pair, then there exists a solution Q_c^* of the Control Algebraic Riccati Equation with Discounted Cost *(13.18) with the side condition (13.19), written for the matrix Q_c*

$$\begin{aligned} Q_c &= f_{CR,DC}(Q_c) = rA^T Q_c A + C_z^T C_z + \qquad (13.18) \\ &\quad -[rA^T Q_c B + C_z^T D_z][rB^T Q_c B + D_z^T D_z]^{-1}[rA^T Q_c B + C_z^T D_z], \\ &\quad \text{spec}(A + BF(Q_c)) \subset \text{D}_o, \; Q_c \in \mathbb{R}_{pds}^{n_x \times n_x}; \qquad (13.19) \\ F(Q_c) &= -[rB^T Q_c B + D_z^T D_z]^{-1}[rA^T Q_c B + C_z^T D_z]^T \in \mathbb{R}^{n_u \times n_x}. \qquad (13.20) \end{aligned}$$

(b)If in addition (A_c,C_c) is an observable pair then $0 \prec Q_c^*$.
(c)If $Q_c \in \mathbb{R}^{n_x \times n_x}$ is any symmetric solution of the discounted-cost control Riccati equation $Q_c = f_{CR}(Q_c)$ then $Q_c \preceq Q_c^*$.

Proof. (a) This follows from Theorem 22.2.4.(e) and .(f).

The control algebraic Riccati equation of discounted cost is transformed to the following form,

$$\begin{aligned} & A \mapsto (r^{1/2}A), \; B \mapsto (r^{1/2}B); \\ Q_c &= (r^{1/2}A)^T Q_c (r^{1/2}A) + C_z^T C_z + \\ &\quad -[(r^{1/2}A)^T Q_c (r^{1/2}B) + C_z^T D_z][(r^{1/2}B)^T Q_c (r^{1/2}B) + D_z^T D_z]^{-1} \times \\ &\quad \times [(r^{1/2}A)^T Q_c (r^{1/2}B) + C_z^T D_z], \\ &\quad \text{spec}((r^{1/2}A) + (r^{1/2}B)F(Q_c)) \subset \text{D}_o, \; Q_c \in \mathbb{R}_{pds}^{n_x \times n_x}; \\ F(Q_c) &= -[(r^{1/2}B)^T Q_c (r^{1/2}B) + D_z^T D_z]^{-1}[(r^{1/2}A)^T Q_c (r^{1/2}B) + C_z^T D_z]^T \\ &\quad \in \mathbb{R}^{n_u \times n_x}. \end{aligned}$$

Because (A,B) is assumed to be a stabilizable pair and $r \in (0,1)$, so is
$((r^{1/2}A),(r^{1/2}B))$. Note that $(r^{1/2}A_c) = (r^{1/2}A) + (r^{1/2}B)(D_z^T D_z) D_z^T C_z$.
Thus (A_c,C_c) a detectable pair implies that $((r^{1/2}A_c),C_c)$ is a detectable pair.

It follows from Theorem 22.2.4.(e) and .(f) and the above discussion that there exists a unique solution Q_c^*. Then $Q_c^* = f_{CARE}(Q_c^*)$ and $Q_c^* \in \mathbb{R}_{pds}^{n_x \times n_x}$. It follows from Theorem 22.2.4.(d) that then $\text{spec}(A + BF(Q_c^*)) \subset \text{D}_o$.
(b) This follows from Theorem 22.2.4.(f).
(c) This follows from Theorem 22.2.4.(g). □

Theorem 13.3.8. *Consider Problem 13.3.6. Assume that the conditions of Theorem 13.3.7 hold. Let $Q_c^* \in \mathbb{R}_{pds}^{n_x \times n_x}$ be as defined in Theorem 13.3.7.*

(a)There exists a nonempty subset of control laws $G_{dcf} \subseteq G$ such that for every control law $g \in G_{fc}$, $J_{dc}(g) < \infty$.

(b)The dynamic programming equation for the value function is specified by,

$$\begin{aligned} V(x) &= \inf_{u \in U(x)} \{b(x,u) + rE[V(f(x,u,v))|F^{x,u}]\}, \;\; V : X \to \mathbb{R}_+, \\ f(x,u,v) &= Ax + Bu + Mv. \end{aligned}$$

(c)The unique solution of the dynamic programming equation is the function specified by,

$$\begin{aligned} V(x_V) &= x_V^T Q_c^* x_V + \frac{r}{(1-r)} \operatorname{tr}(M^T Q_c M), \\ E[V(x_0)] &\leq E[J_{dc}(g, x_0)] = J_{dc}(g), \; \forall\, g \in G, \\ E[V(x_0)] &= E[J_{dc}(g^*, x_0)] = J_{dc}(g^*), \\ J_{dc}^* &= E[J_{dc}(g^*, x_0)] = E[V(x_0)] = J_{dc}(g^*). \end{aligned}$$

(d)The optimal control law (LQG-CO-DC) is a Markov control law specified by,

$$g_{LQG,co,dc}^*(x_g) = -[rB^T Q_c^* B + D_z^T D_z]^{-1} \; [rA^T Q_c^* B + C_z^T D_z]^T x_g.$$

The optimal control law g^* is optimal over the set of measurable nonlinear control laws.

Proof. (a) Because the tuple (A, B) is a stabilizable pair, it follows from Theorem 21.2.11.(c) that there exists a feedback matrix $F \in \mathbb{R}^{n_u \times n_x}$ such that the closed-loop system matrix $A + BF$ is exponentially stable, $\operatorname{spec}(A + BF) \subset \mathrm{D}_o$. Consider the control law g defined below and the variance of the state of the closed-loop system,

$$\begin{aligned} g(x) &= Fx, \\ x^g(t+1) &= Ax^g(t) + Bu^g(t) + Mv(t) = (A + BF)x^g(t) + Mv(t), \\ u^g(t) &= Fx^g(t), \\ Q_{x^g}(t+1) &= (A+BF)Q_{x^g}(t)(A+BF)^T + MM^T, \; Q_{x^g}(0) = Q_{x_0}. \end{aligned}$$

It follows from the exponential stability of the closed-loop system and from Theorem 22.1.2 that,

$$\begin{aligned} \lim_{t \to \infty} Q_x(t) &= Q_x(\infty) \in \mathbb{R}^{n_x \times n_x}, \\ Q_x(\infty) &= (A+BF)Q_x(\infty)(A+BF)^T + MM^T, \end{aligned}$$

and that $Q_x(\infty)$ is the unique solution of the above discrete-time Lyapunov equation.

Note the calculations,

$$
\begin{aligned}
& E[\sum_{s=0}^{t-1} r^s \begin{pmatrix} x^g(s) \\ u^g(s) \end{pmatrix}^T Q_{cr} \begin{pmatrix} x^g(s) \\ u^g(s) \end{pmatrix}] \\
= & E[\sum_{s=0}^{t-1} r^s (x^g(s)^T Q_{cc} x^g(s))] = \sum_{s=0}^{t-1} r^s (\text{trace}(Q_{x^g}(s) Q_{cc}) \\
& \text{where } Q_{cc} = C_z^T C_z + C_z^T D_z F + F^T D_z^T D_z + F^T D_z^T D_z F, \\
= & \frac{1-r^t}{1-r} \ (\text{trace}(Q_x(\infty) Q_{cc})) + \sum r^s \ \text{trace}((Q_x(s) - Q_x(\infty)) Q_{cc}) \\
\leq & \frac{1-r^t}{1-r} (\text{trace}(Q_x(\infty) Q_{cc}) + \\
& + (\sum (r^2)^s)^{1/2} \ (\sum \text{trace}((Q_x(s) - Q_x(\infty)) Q_{cc})^2)^{1/2}, \\
& J_{ac}(g) = \lim_{t \rightarrow \infty} \frac{1}{t} \ E[\ldots] \\
= & \lim_{t \rightarrow \infty} \frac{1-r^t}{1-r} \text{trace}(Q_x(\infty) Q_{cc}) + \\
& + \lim \ (\sum_{s=0}^{t-1} (r^2)^s)^{1/2} \ \lim_{t \rightarrow \infty} \ (\sum_{s=0}^{t-1} \text{trace}((Q_x(s) - Q_x(\infty)) Q_{cc})^{1/2}) < \infty.
\end{aligned}
$$

(b) See Theorem 13.3.3.

(c) It is necessary to calculate,

$$\begin{aligned}
V(x_V) &= x_V^T Q_c^* x_V + \frac{r}{1-r}\text{trace}(M^T Q_c^* M), \\
&\quad rE[V(f(x,u,v))|F^{x,u}] \\
&= rE[(Ax+Bu+Mv)^T Q_c^* (Ax+Bu+Mv) + \frac{r^2}{1-r}\text{trace}(M^T Q_c^* M)|F^{x,u}] \\
&= \begin{pmatrix} x \\ u \end{pmatrix}^T \begin{pmatrix} rA^T Q_c^* A & rA^T Q_c^* B \\ (rA^T Q_c^* B)^T & rB^T Q_c^* B \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix} + (\frac{r^2}{1-r} + r)\text{trace}(M^T Q_c^* M), \\
&\quad b(x,u) + rE[V(f(x,u,v)|F^{x,u}] \\
&= \begin{pmatrix} x \\ u \end{pmatrix}^T \begin{pmatrix} (rA^T Q_c^* A + C_z^T C_z) & (rA^T Q_c^* B + C_z^T D_z) \\ (rA^T Q_c^* B + C_z^T D_z)^T & (rB^T Q_c^* B + D_z^T D_z) \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix} + \\
&\quad + \frac{r}{1-r}\text{trace}(M^T Q_c^* M); \\
&\quad \inf_{u \in U(x)} \; (b(x,u) + rE[V(f(x,u,v)|F^{x,u}]) \\
&= \inf_{u \in U(x)} \begin{pmatrix} x \\ u \end{pmatrix}^T \begin{pmatrix} (rA^T Q_c^* A + C_z^T C_z) & (rA^T Q_c^* B + C_z^T D_z) \\ (rA^T Q_c^* B + C_z^T D_z)^T & (rB^T Q_c^* B + D_z^T D_z) \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix} + \\
&\quad + \frac{r}{1-r}\text{trace}(M^T Q_c^* M) \\
&= \inf_{u \in U(x)} \begin{pmatrix} x \\ u - Fx \end{pmatrix}^T \begin{pmatrix} Q_c & 0 \\ 0 & rB^T Q_v B + D_z^T D_z \end{pmatrix} \begin{pmatrix} x \\ u - Fx \end{pmatrix} + \\
&\quad + \frac{r}{1-r}\text{trace}(M^T Q_c^* M) \\
&= x_V^T Q_c^* x_V + \frac{r}{1-r}\text{trace}(M^T Q_c^* M) = V(x_V); \\
F(Q_c^*) &= -[rB^T Q_c^* B + D_z^T D_z]^{-1} \; [rA^T Q_c^* B + C_z^T D_z]^T.
\end{aligned}$$

Thus the candidate expression $V(x)$ is a solution of the dynamic programming equation.
(d) It follows from the proof of (c) above that,

$$\begin{aligned}
g_{LQG,co,dc}^*(x_g) &= F(Q_c^*) x_g, \\
F(Q_c^*) &= -[rB^T Q_c^* B + D_z^T D_z]^{-1} \; [rA^T Q_c^* B + C_z^T D_z]^T.
\end{aligned}$$

□

13.3.3 Control of a State-Finite Stochastic Control System

The discounted cost criterion was introduced in Section 13.1. The problem formulation follows.

Definition 13.3.9. Consider a recursive time-invariant state-observed finite stochastic control system as defined Definition 10.1.2 with $T = N = \{0, 1, 2, \ldots\}$, with finite state set X and finite input set U, and with representation,

$$x(t+1) = f(x(t),u(t),v(t)), x(0) = x_0. \tag{13.21}$$

The function $f : X \times U \times W \to X$ does not explicitly depend on time. Here $x_0 : \Omega \to X$ is a random variable representing the initial state.

Assume that the information pattern is the past state information pattern $\{F_t^x,\ t \in T\}$. Assume that the set G of control laws is the set of past-state control laws. The closed-loop stochastic control system is then given by the representation,

$$\begin{aligned} x^g(t+1) &= f(x^g(t), g_t(x^g(0:t)), v(t)), x^g(0) = x_0, \\ u^g(t) &= g_t(x^g(0:t)) = g_t((x^g(0), x^g(1), \ldots, x^g(t))). \end{aligned}$$

Denote the upper bound on the cost-rate function $b : X \times U \to \mathbb{R}_+$ by, forall $(x,u) \in X \times U$, $b(x,u) \leq b_2 \in \mathbb{R}_+$,

Define the discounted-cost criterion with discount factor $r \in (0,1)$ as,

$$J(g) = E[\sum_{s=0}^{\infty} r^s b(x^g(s), u^g(s))],\ J_{ac} : G \to \mathbb{R}_+.$$

Because there exists a bound on the cost rate, it is directly proven that the discounted cost is a real number for every control law in G.

Problem 13.3.10. *Optimal stochastic control problem with complete observations and discounted cost for a finite stochastic control system.* Consider the finite stochastic control system of 13.3.9 with the assumptions stated there. The *optimal stochastic control problem with complete observations and with discounted cost* is to determine a control law $g^* \in G$ and a value $J^* \in \mathbb{R}_+$ such that,

$$J^* = J(g^*) = \inf_{g \in G} J(g).$$

In case of a finite stochastic control system special notation is used which is introduced below. The description of a stochastic control system in the form of,

$$x(t+1) = f(x(t),u(t),v(t)),$$

is equivalent to the specification of a *transition measure* $P_1 : X \times U \to \mathbf{P}(X)$ where $\mathbf{P}(X)$ denotes the set of probability measures on the state space X. Starting from Equation (13.21) one may define $P_1 \in \mathbf{P}(X)$ by

$$\begin{aligned} P_1(i,j,u) &= P(\{x(t+1) = i\} | \{x(t) = j, u(t) = u\}) \\ &= P(\{f(j,u,v(t)) = i\} | \{x(t) = j, u(t) = u\}). \end{aligned} \tag{13.22}$$

In the following the transition measure P_1 is denoted by the map $P : X \times X \times U \to \mathbb{R}_+$. With a transition measure, conditional expectation may be represented as follows. Let $V : T \times X \to \mathbb{R}_+$ be a measurable function. Then,

$$E[V(t+1, x^g(t+1)) | F^{x^g(t), u^g(t)}] = \sum_{x_1 \in X} V(t+1, x_1) P(x_1, x^g(t), u^g(t)). \tag{13.23}$$

Note that because the state space X is finite, the sum in Equation (13.23) is finite. Below Equation (13.23) will also be written as,

$$E[V(t+1,x^g(t+1))|F^{x_V,u_V}] = \sum_{x_1 \in X} V(t+1,x_1)P(x_1,x_V,u_V).$$

Control Theory

The reader finds in this section the solution procedure for optimal control with the discounted cost for a finite stochastic control system.

The solution procedure is preceded by an informal derivation of the dynamic programming equation.

Consider Problem 13.3.10 but with the finite-horizon $T(0:t_1-1) = \{0,1,\ldots,t_1\}$ and the cost function,

$$J_1(g) = E[\sum_{s=0}^{t_1-1} r^s b(x(s),u(s))].$$

The value function for this problem as produced by the backward recursion of Procedure 12.6.2 is determined by,

$$\begin{aligned} W(t_1,x) &= 0, \quad W : T \times X \to \mathbb{R}_+, \; t = t_1-1, \; t_1-2, \; \ldots, \; 0, \\ W(t,x) &= \min_{u \in U} \left\{ r^t b(x,u) + \sum_{x_1 \in X} W(t+1,x_1)P(x_1,x,u) \right\}. \end{aligned}$$

Note that by Theorem 12.6.4.(a), $W(t,x_W)$ is the minimal cost-to-go on the horizon $\{t, \; t+1, \; \ldots, \; t_1\}$ when starting at time $t \in T$ in state $x_W \in X$. Transform the backward recursion into a forward recursion.

$$\begin{aligned} V(t,x) &= r^{t-t_1} W(t_1-t,x), \; V : T(0:t_1) \times X \to \mathbb{R}_+. \text{ Then,} \\ V(0,x) &= 0, \; \forall x \in X, \\ V(t,x) &= \min_{u \in U} \left\{ b(x,u) + r \sum_{x_1 \in X} V(t-1,x_1)P(x_1,x,u) \right\}. \end{aligned}$$

This may be proven as follows.

$$\begin{aligned} V(0,x) &= r^{-t_1} W(t_1,x) = 0, \\ V(t,x) &= r^{t-t_1} W(t_1-t,x) \\ &= \min_{\{u \in U\}} \left\{ r^{t-t_1} r^{t_1-t} b(x,u) + r^{t-t_1} \sum_{x_1 \in X} W(t_1-t+1,x_1)P(x_1,x,u) \right\} \\ &= \min_{\{u \in U\}} \left\{ b(x,u) + r \sum_{x_1 \in X} V(t-1,x_1)P(x_1,x,u) \right\}. \qquad (13.24) \end{aligned}$$

Note that the forward transition measure P is used in connection with $V(t-1,.)$. It may then be deduced from Theorem 12.6.4.(a) that,

$$
\begin{aligned}
V(t,x_0) &= \inf_{g \in G} E[\sum_{s=0}^{t-1} r^s b(x(s),u(s))|F^{x_0}], \text{ hence that,} \\
V(\infty,x_0) &= \inf_{g \in G} E[\sum_{s=0}^{\infty} r^s b(x(s),u(s))|F^{x_0}],
\end{aligned}
$$

and from this expression and from Equation (13.24) it can be conjectured that the expression $V(\infty,x_0)$ satisfies the equation,

$$
V(\infty,x) = \min_{u \in U} \left\{ b(x,u) + r \sum_{x_1 \in X} V(\infty,x_1)P(x_1,x,u) \right\}, \forall\, x \in X. \tag{13.25}
$$

The purpose of this derivation was to motivate that Equation (13.25) is the dynamic programming equation for the discounted cost case. In Theorem 13.3.16 it is proven that (13.25) is the proper dynamic programming equation for discounted cost.

The formal construction proceeds with concepts of functional analysis.

Definition 13.3.11. Consider Problem 13.3.10. Define the sets and the *dynamic programming operator DP* as,

$$
\begin{aligned}
& \text{For } z : X \to \mathbb{R} \text{ define} \\
\|z\| &:= \max_{x \in X} |z(x)|, \text{ which can be proven to be a norm on the space, } X, \\
L &= L(X,\mathbb{R}) = \left\{ z : X \to \mathbb{R} | \|z\| < \infty \right\}, \; DP : L \to L, \\
DP(z)(x) &= \min_{u \in U} \left\{ b(x,u) + r \sum_{x_1 \in X} z(x_1)P(x_1,x,u) \right\}. \\
& \forall\, g \in G_M, \text{ define, } DP_g : L \to L, \\
DP_g(z)(x) &= \left\{ b(x,g(x)) + r \sum_{x_1 \in X} z(x_1)P(x_1,x,g(x)) \right\}.
\end{aligned}
$$

Definition 13.3.12. Consider Problem 13.3.10 with Assumption 13.3.9. The *dynamic programming equation for the discounted cost function* $V : X \to \mathbb{R}_+$ is defined as the equation,

$$
\begin{aligned}
& V = DP(V); \text{ equivalently, } \forall x_V \in X, \\
V(x_V) &= DP(V)(x_V) = \min_{u_V \in U(x_V)} \left\{ b(x_V,u_V) + r \sum_{x_1 \in X} V(x_1)\, P(x_1,x_V,u_V) \right\}.
\end{aligned}
$$

The dynamic programming equation is an equation for the value function V in terms of the dynamic programming operator.

Dynamic programming for the optimal stochastic control problem with discounted cost may make use of the concepts of a normed space and of a contraction mapping.

Definition 13.3.13. Let $(L, \|.\|)$ be a normed linear space consisting of a non-empty linear space L and a norm $\|.\|$. A *contraction mapping* is a map $K : L \rightarrow L$, where Ky denotes $K(y)$, such that there exists a real number $r \in (0,1)$ satisfying,

$$\|Ky - Kz\| \leq r\|y - z\|, \quad \forall\, y,\, z \in L.$$

Example 13.3.14. Example. Contraction mapping.
Consider the normed linear space $(\mathbb{R}, |.|)$ and the affine function $k : \mathbb{R} \rightarrow \mathbb{R}$, $k(x) = ax + b$ for an $a \in (0,1)$ and a $b \in \mathbb{R}$. Then k is a contraction mapping with $|k(x) - k(y)| \leq a|x - y|$. Note that if $k(x^*) = ax^* + b = x^*$ then $x^* = b/(1-a)$.

Theorem 13.3.15. *Let $(L, \|.\|)$ be a complete normed space and let $K : L \rightarrow L$ be a contraction mapping.*

(a)There exists a unique $w^* \in L$ such that $w^* = Kw^*$. The element $w^* \in L$ is called the fixpoint *of the map K.*
(b)Define the sequence $\{K^n z,\ \forall\, n \in Z_+\}$ for $z \in L$,

$$K^1(z) = z,$$
$$K^{n+1}(z) = K(K^n(z)),\ \forall\, n \in \mathbb{Z}_+. \text{ Then, for } w^* \text{ defined in (a),}$$
$$0 = \lim_{n\rightarrow\infty} \|K^n(z) - w^*\|.$$

Theorem 13.3.16. *Consider Problem 13.3.10 with discounted cost.*

(a)The dynamic programming equation $V = DP(V)$, or equivalently,

$$V(x) = \min_{u\in U} \left\{ b(x,u) + r \sum_{x_1 \in X} V(x_1) P(x_1, x, u) \right\},\ \forall\, x \in X,$$

for the function $V : X \rightarrow \mathbb{R}_+$ has a unique solution.
(b)The function V is such that for any $x^g(0) = x_0 \in X$,

$$V(x_0) = \inf_{g\in G} E[\sum_{s=0}^{\infty} r^s b(x^g(s), u^g(s)) | F^{x_0}]. \tag{13.26}$$

Hence the function V satisfies the interpretation of the value function.

The interpretation of Equation (13.26) is that, if the initial state is x_0, then $V(x_0)$ is the conditional discounted cost when the initial state of the finite stochastic control system equals x_0. The proof of Theorem 13.3.16 is based on the following lemma.

Lemma 13.3.17. *Consider Problem 13.3.10.*

(a)The map DP is well defined and a contraction mapping.
(b)There exists a unique function $V : X \rightarrow \mathbb{R}_+$ such that $DP(V) = V$.
(c)For any $z \in L(X, \mathbb{R})$

$$\lim_{n\rightarrow\infty} \|(DP)^n z - V\| = 0 \Leftrightarrow \lim_{n\rightarrow\infty} ((DP)^n z)(x) = V(x),\ \forall\, x \in X.$$

(d)If $z \in L(X,\mathbb{R})$ satisfies $z = 0$, or, equivalently, $z(x) = 0$ for all $x \in X$, then,

$$((DP)^k z)(x_0) = \inf_{g \in G} E[\sum_{s=0}^{k-1} r^s b(x^g(s), u^g(s))|F^{x_0}], \ \forall\, k \in \mathbb{Z}_+,\ x^g(0) = x_0 \in X.$$

Corollary 13.3.18. *Let $g : X \to U$ be a stationary Markov control law. Define*

$$(DP_g z)(x) = b(x, g(x)) + r \sum_{x_1 \in X} z(x) P(x_1, x, g(x)),\ DP_g : L(X,\mathbb{R}) \to L(X,\mathbb{R}).$$

(a)The map DP_g is well defined and a contraction mapping.
(b)There exists a unique function $V(g,.) : X \to \mathbb{R}$ such that $DP_g V(g,.) = V(g,.)$.
(c)For any $z \in L(X,\mathbb{R})$ and $x \in X$

$$\lim_{k \to \infty} ((DP_g)^k z)(x) = V(g,x).$$

(d)If $z \in L(X,\mathbb{R})$ satisfies $z = 0$ then for all $k \in Z_+$ and for all $x^g(0) = x_0 \in X$,

$$((DP_g)^k z)(x_0) = E[\sum_{s=0}^{k-1} r^s b(x^g(s), u^g(s))|F^{x_0}].$$

Proof. This follows from Lemma 13.3.17 by restricting G to the set $\{g\}$ and for any $x \in X$, restricting U to the set $U_x = \{g(x)\}$. □

Proof. Of Lemma 13.3.17.
(a) That DP is well defined follows from the boundedness of b, see Assumption 13.3.9, and from $\sum_{x_1 \in X} P(x_1, x, u) = 1$. It will be proven that DP is a contraction mapping. Let $y, z \in L(X,\mathbb{R})$, $x \in X$, and,

$$u_1 = \arg\min_{u \in U} \left\{ b(x,u) + r \sum_{x_1 \in X} y(x_1) P(x_1, x, u) \right\}.$$

Then,

$$\begin{aligned}
&(DP(z))(x) - (DP(y))(x) \\
&= \inf_{u \in U} \left\{ b(x,u) + r \sum_{x_1 \in X} z(x_1) P(x_1, x, u) \right\} - [b(x,u_1) + r \sum_{x_1 \in X} y(x_1) P(x_1, x, u_1)] \\
&\leq b(x,u_1) + r \sum_{x_1 \in X} z(x_1) P(x_1, x, u_1) - b(x,u_1) - r \sum_{x_1 \in X} y(x_1) P(x_1, x, u_1) \\
&= r \sum_{x_1 \in X} [z(x_1) - y(x_1)] P(x_1, x, u_1) \leq r\|z - y\| \sum_{x_1 \in X} P(x_1, x, u_1) = r\|z - y\|.
\end{aligned}$$

By reversing the roles of z and y one obtains

$$\begin{aligned}
&(DP(y))(x) - (DP(z))(x) \leq r\|y - z\|; \text{ hence } \forall x \in X, \\
&|(DP(z))(x) - (DP(y))(x)| \leq r\|y - z\|, \text{ or, } \|DP(z) - DP(y)\| \leq r\|y - z\|.
\end{aligned}$$

(b) Then Theorem 13.3.15 implies that the equation $DP(V) = V$ has an unique solution V.

(c) This follows from Theorem 13.3.15.(b).
(d) This follows (c) and the formula of $DP(z)$. □

Proof. Of Theorem 13.3.16.
(a) This follows from Lemma 13.3.17.(b).
(b) Let $z \in L(X, \mathbb{R})$, $z = 0$. By Lemma 13.3.17.(d)

$$(DP^k z)(x) = \inf_{g \in G} E[\sum_{s=0}^{k-1} r^s b(x^g(s), u^g(s))|F^x].$$

Fix $x \in X$. Then,

$$\inf_{g \in G} E[\sum_{s=0}^{\infty} r^s b(x^g(s), u^g(s))|F^x] \geq \inf E[\sum_{s=0}^{t_1-1} r^s b(x^g(s), u^g(s))|F^x],$$

because by Assumption 13.3.9, $0 \leq b(x,u)$,

$$= (DP^{t_1} z)(x), \text{ by Lemma 13.3.17.(d), hence,}$$

$$\inf_{g \in G} E[\sum_{s=0}^{\infty} r^s b(x^g(s), u^g(s))|F^x]$$

$$\geq \lim_{t_1 \to \infty} ((DP)^{t_1} z)(x) = V(x) \quad \text{by Lemma 13.3.17.(c); let } t_1 \in T; \text{ then,}$$

$$\inf_{g \in G} E[\sum_{s=0}^{\infty} r^s b(x^g(s), u^g(s))|F^x]$$

$$\leq \inf_{g \in G} E[\sum_{s=0}^{t_1-1} r^s b(x^g(s), u^g(s))|F^x] + \sum_{s=t_1}^{\infty} r^s b_2,$$

since by Assumption 13.3.9, $b(x,u) \leq b_2$,

$$= ((DP)^{t_1} z)(x) + \frac{r^{t_1} b_2}{1-r}, \text{ by } r \in (0,1) \text{ and Lemma 13.3.17.(d); thus,}$$

$$\inf_{g \in G} E[\sum_{s=0}^{\infty} r^s b(x^g(s), u^g(s))|F^x] \leq \lim_{t_1 \to \infty} [((DP)^{t_1} z)(x) + \frac{r^{t_1} b_2}{1-r}]$$

$$= V(x), \text{ by Lemma 13.3.17.(c) and } r \in (0,1).$$

□

From the solution of the discounted-cost dynamic programming equation, one can derive by a limit argument on the discount factor, the dynamic programming equation of average cost. This result is referred to other sources, see the section *Further Reading*.

Procedures

How to construct an optimal control law for the discounted cost criterion? Recall from Definition 11.2.3.(e) that a control law of the form $g : X \to U$ is called time-invariant. Denote by G_{ML} the set of time-invariant Markov control laws.

Procedure 13.3.19 Calculation of optimal control law for discounted-cost optimal stochastic control.
Data: $r \in (0,1)$, $P : X \times X \times U \to \mathbb{R}$, $b : X \times U \to \mathbb{R}_+$, $V : X \to \mathbb{R}_+$ where V is the value function assumed to be available. Define,

$$g^*(x) = \arg\min_{u \in U} \left\{ b(x,u) + r \sum_{x_1 \in X} P(x_1,x,u)V(x_1) \right\}, \; g^* : X \to U. \tag{13.27}$$

Theorem 13.3.20. *Let $V : X \to \mathbb{R}$ be the value function of the discounted cost optimal stochastic control problem as constructed in Theorem 13.3.16. Let $g^* \in G_{ML}$ be the control law constructed according to Procedure 13.3.19. Then g^* is an optimal control law.*

Proof. By definition of g^*,

$$V(x) = b(x,g^*(x)) + r \sum_{x_1 \in X} P(x_1,x,g^*(x))V(x_1) = DP_{g^*}(V)(x),$$

for all $x \in X$. But by Corollary 13.3.18.(b) this equation has the unique solution $V(g^*,.)$, hence $V(g^*,x) = V(x)$, for all $x \in X$. Then,

$$E_{g^*}[\sum_{s=0}^{\infty} r^s b(x(s),g^*(x(s)))|F^x] = V(g^*,x) \text{ by Theorem 13.3.16.(b)},$$
$$= V(x), \text{ by the above discussion},$$
$$= \inf_{g \in G} E[\sum_{s=0}^{\infty} r^s b(x(s),u(s))|F^x], \text{ by Equation (13.26)},$$

hence $g^* \in G$ is an optimal control law. □

Procedure 13.3.19 shows how to construct an optimal control law from the value function. But how to determine the value function? There are two approaches: value iteration and policy iteration.

Procedure 13.3.21 Value iteration.
Data: $P : X \times X \times U \to \mathbb{R}_+$, $b : X \times U \to \mathbb{R}_+$, $r \in (0,1)$.

1. *Initialization. Let $h(0,.) : X \to \mathbb{R}_+$, $h(0,x) = 0$ for all $x \in X$.*
2. *For $m = 0,1,2,\ldots$ do,*

$$h(m+1,x) = \inf_{u \in U} \left\{ b(x,u) + r \sum_{x_1 \in X} P(x_1,x,u)h(m,x_1) \right\}. \tag{13.28}$$

3. *Let $V : X \to \mathbb{R}_+$, $V(x) = \lim_{m \to \infty} h(m,x)$.*
4. *Let $g^* : X \to U$,*

$$g^*(x) = \arg\min_{u \in U} \left\{ b(x,u) + r \sum_{x_1 \in X} P(x_1,x,u)V(x_1) \right\}. \tag{13.29}$$

5. *Output g^*.*

It follows from the Theorems 13.3.16 and 13.3.20 that the function V constructed by the value iteration procedure is the value function and g^* an optimal control law.

The value iteration procedure contains a limit procedure. In general convergence does not take place in a finite number of steps. In practice one stops if the changes in the candidate value function and the candidate optimal control law change little from step to step.

Because the state space X and the input space U are finite sets one expects a procedure that converges in a finite number of steps. The policy iteration procedure presented below has this property.

Notation is introduced first. By Assumption 13.3.9.(1) the state space X is a finite set, say $X = \{x_1, x_2, \ldots, x_{n_x}\}$ with n_x elements. A function $V : X \rightarrow \mathbb{R}_+$ may therefore be represented as a vector in $\mathbb{R}_+^{n_x}$ and, correspondingly, any control law $g \in G_{ML}$, $g : X \rightarrow U$, and the cost rate $b(g)$ for a control law $g \in G_{ML}$,

$$V = \begin{pmatrix} V(x_1) \\ V(x_2) \\ \vdots \\ V(x_{n_x}) \end{pmatrix} \in \mathbb{R}_+^{n_x}; \; g = \begin{pmatrix} g(x_1) \\ g(x_2) \\ \vdots \\ g(x_{n_x}) \end{pmatrix} \in \mathbb{R}_+^{n_u},$$

$$b(g) = \begin{pmatrix} b(x_1, g(x_1)) \\ b(x_2, g(x_2)) \\ \vdots \\ b(x_n, g(x_{n_x})) \end{pmatrix} \in \mathbb{R}_+^{n_x};$$

$$P(g) \in \mathbb{R}_+^{n_x \times n_x}, \; P(g)_{ij} = P(x_j, x_i, g(x_i)); \text{ then,}$$

$$DP(g)z = b(g) + rP(g)z, \;\; DP(g) : \mathbb{R}^{n_x} \rightarrow \mathbb{R}_+^{n_x},$$

Define the order relation $\leq$ on $\mathbb{R}^{n_x}$ by,

$$y \leq x, \text{ if } \forall \; i \in \mathbb{Z}_{n_x}, \; y_i \leq z_i; \;\; y < z, \text{ if } y \leq z \text{ and } y \neq z.$$

Furthermore, define $DP : \mathbb{R}^n \rightarrow \mathbb{R}^n$,

$$DP(z) = \min_{g \in \mathbb{R}^n} [b(g) + rP(g)z], \; DP : \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_x}.$$

where the minimization is defined componentwise. Thus $g(x_1)$ is the minimum argument of the first component of $b(g) + rP(g)z$, etc. Recall from Lemma 13.3.17 that for any $g \in G_{SM}$, $DP(g)$ is a contraction. From Lemma 13.3.17 follows that the equation,

$$V(g) = DP_g(V) = b(g) + rP(g)V(g), \; V(g) \in \mathbb{R}_+^{n_x},$$

has a unique solution $V(g)$.

Procedure 13.3.22 Policy iteration procedure *or* policy improvement procedure. *Data: $n_x \in \mathbb{Z}_+$, $b : X \times U \rightarrow \mathbb{R}_+^{n_x}$, $P : G \rightarrow \mathbb{R}_+^{n_x \times n_x}$, $g_0 \in \mathbb{R}^{n_x}$ representing a time-invariant control law.*

1. *Initialization. Let $m = 0$. Solve the following equation for $V(g_0) \in \mathbb{R}_+^{n_x}$,*

$$b(g_0) + rP(g_0)V(g_0) = V(g_0).$$

2. *For $m := m+1$ while,*

$$\min_{g \in \mathbb{R}^n} \{b(g) + rP(g)V(g_{m-1})\} < V(g_{m-1}),\ do, \tag{13.30}$$

(2.1) determine $g_m \in \mathbb{R}^{n_u}$ such that,

$$b(g_m) + rP(g_m)V(g_{m-1}) = \min_{g \in \mathbb{R}^{n_u}} \{b(g) + rP(g)V(g_{m-1})\};$$

(2.2) solve for $V(g_m) \in \mathbb{R}_+^{n_x}$, the equation,

$$b(g_m) + rP(g_m)V(g_m) = V(g_m). \tag{13.31}$$

3. *Output $g^* = g_m$.*

It follows from Proposition 13.3.25 that the above procedure stops after a finite number of steps.

Example 13.3.23. Illustration of policy iteration for discounted cost. Consider the stochastic control system and the cost function,

$$\begin{aligned} X &= \{1,2\},\ U = \{u_1, u_2\}, \\ P(u_1) &= \frac{1}{4}\begin{pmatrix} 3 & 1 \\ 3 & 1 \end{pmatrix},\ P(u_2) = \frac{1}{4}\begin{pmatrix} 1 & 3 \\ 1 & 3 \end{pmatrix}, \\ b(1,u_1) &= 2,\ b(1,u_2) = 0.5,\ b(2,u_1) = 1,\ b(2,u_2) = 3,\ r = 0.9. \end{aligned}$$

The policy iteration procedure is applied. Let $g_0 : X \rightarrow U$
$g_0(1) = u_1$, $g_0(2) = u_2$, and $m = 0$.

1. Initialization. Solve the equation,

$$\begin{aligned} b(g_0) + rP(g_0)V(g_0) &= V(g_0) \in \mathbb{R}^2,\ \Leftrightarrow \\ \begin{pmatrix} 2 \\ 3 \end{pmatrix} + \frac{0.9}{4}\begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix} V(g_0) &= V(g_0);\quad V(g_0) = \begin{pmatrix} 24.12 \\ 25.96 \end{pmatrix}. \end{aligned}$$

2. Let $m = 1$.

   a. Determine $g_1 \in \mathbb{R}^2$ such that

$$\begin{aligned} b(g_1) + rP(g_1)V(g_1) &= \min_{g \in R^2}\{b(g) + rP(g)V(g_0)\} \\ &= \min \begin{pmatrix} \{2 + 0.9 \times (\frac{3}{4} \times 24.12 + \frac{1}{4} \times 25.96), \\ 0.5 + 0.9 \times (\frac{1}{4} \times 24.12 + \frac{3}{4} \times 25.96)\} \\ \{1 + 0.9 \times (\frac{3}{4} \times 24.12 + \frac{1}{4} \times 25.96), \\ 3 + 0.9 \times (\frac{1}{4} \times 24.12 + \frac{3}{4} \times 25.96)\} \end{pmatrix} \\ &= \begin{pmatrix} \min\{24.12, 23.45\} \\ \min\{23.12, 25.95\} \end{pmatrix} = \begin{pmatrix} 23.45 \\ 23.12 \end{pmatrix}, \end{aligned}$$

   hence $g_1(1) = u_2$ and $g_1(2) = u_1$. Note that,

$$\min_{g \in R^2} b(g) + rP(g)V(g_0) = \begin{pmatrix} 23.45 \\ 23.12 \end{pmatrix} < \begin{pmatrix} 24.12 \\ 25.96 \end{pmatrix}.$$

b. Solve the equation

$$b(g_1) + rP(g_1)V(g_1) = V(g_1) \in \mathbb{R}^2; \;\; V(g_1) = \begin{pmatrix} 7.33 \\ 7.67 \end{pmatrix}.$$

3. Let $m = 2$.

a. Determine $g_2 \in R^2$ such that,

$$\begin{aligned} b(g_2) = rP(g_2)V(g_1) &= \min_{g \in R^2} \{b(g) + rP(g)V(g_1)\} \\ = \begin{pmatrix} \min\{8.67, 7.33\} \\ \min\{7.67, 9.83\} \end{pmatrix} &= \begin{pmatrix} 7.33 \\ 7.67 \end{pmatrix}, \end{aligned}$$

$g_2(1) = u_2$ and $g_2(2) = u_1$. Note that $g_2 = g_1$.

b. Note that the solution of (13.31) is such that $V(g_2) = V(g_1)$.

4. Let $m = 3$. Note that

$$V(g_2) = \min_{g \in R^2} b(g) + rP(g)V(g_1),$$

hence condition (13.30) is satisfied.

5. Then the optimal control law and the value are,

$$g^* = g_2 = \begin{pmatrix} u_2 \\ u_1 \end{pmatrix}, \; V(g^*) = \begin{pmatrix} 7.33 \\ 7.67 \end{pmatrix}.$$

Proposition 13.3.24. *Let $g \in G$. Then DP, $DP(g) : \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ are monotone operators: if $z,\ y \in \mathbb{R}^{n_x}$ with $z \leq y$ then $DP(z) \leq DP(y)$ and $DP(g)z \leq DP(g)y$.*

Proof. Let $z, y \in \mathbb{R}^{n_x}$ with $z \leq y$. Then

$$DP(g)z - DP(g)y = b(g) + rP(g)z - [b(g) + rP(g)y] = rP(g)(z - y) \leq 0,$$

because $(z - y) \leq 0$, $r \in (0,1)$, and $P(g) \in \mathbb{R}_+^{n_x \times n_x}$. Then also,

$$\begin{aligned} DP(z) = \min_{g \in \mathbb{R}^{n_u}} &\{b(g) + rP(g)z\} \leq b(g_1) + rP(g_1)z, \text{ for any } g_1 \in \mathbb{R}^{n_u}, \\ &\leq b(g_1) + rP(g_1)y, \text{ by } DP(g)z \leq DP(g)y, \text{ hence,} \\ DP(z) \leq \min_{g_1 \in \mathbb{R}^{n_u}} &\{b(g_1) + rP(g_1)y\} = DP(y). \end{aligned}$$

□

Proposition 13.3.25. *Procedure 13.3.22, the policy iteration procedure, converges in a finite number of steps to an optimal control law $g^* \in G_{ML}$.*

Proof. Suppose that for some $m \in \mathbb{Z}_+$, $DP(V)(g_m) \leq V(g_m)$. According to Step (2.a), $g_{m+1} \in \mathbb{R}^{n_u}$ is such that

$$DP(g_{m+1})V(g_m) = DP(V)(g_m) \leq V(g_m). \tag{13.32}$$

Then

$$\begin{aligned}
V(g_m) &\geq DP(g_{m+1})V(g_m), \text{ by (13.32)}, \\
&\geq DP(g_{m+1})^2 V(g_m), \\
&\quad \text{by the monotonicity of Proposition 13.3.24}, \\
&\geq \ldots \geq (DP(g_{m+1}))^k V(g_m), \text{ and} && (13.33) \\
V(g_{m+1}) &= \lim_{k\to\infty} (DP(g_{m+1}))^k V(g_m), \text{ by Corollary 13.3.18.(a)}, \\
&\leq V(g_m) \text{ by Equation (13.33)}. && (13.34)
\end{aligned}$$

If $DP(V)V(g_m) = V(g_m)$ then, by Lemma 13.3.17.(b) $V(g_m) = V$ and by Theorem 13.3.20 $g^* = g_m$ is an optimal control law.

$$DP(V)(g_m) < V(g_m) \;\Rightarrow\; V(g_{m+1}) = \lim_{k\to\infty} (DP(g_{m+1}))^k V(g_m) < V(g_m),$$

hence g_m is not optimal. Because X and U are finite sets, the set of time-invariant Markov control laws is finite. Hence the procedure converges in a finite number of steps. □

13.4 Minimum-Variance Control with Complete Observations

A special case of a stochastic control problem is that in which the cost rate depends only on the state process but not at all on the input process. This problem was first considered by K.J. Aström for the variance of the output process only. But it can also be considered for the variance of the state process.

The minimun-variance control problem with complete observations is to determine a control law based on complete observations such that the variance of the state process is minimal when an infinite horizon control problem is considered. Below this problem is discussed for a Gaussian stochastic control system.

The solution for the partial-observations single-input discrete-time case was presented by K.J. Aström, [3, 4]. The solution technique uses polynomials. The multivariable version in discrete-time, thus with an input process with two or more input components, was solved by U. Shaked and P.R. Kumar in [36]. The control theory for minimum-variance control in case the stochastic control system is formulated in terms of polynomial matrices and was developed by V. Kučera, see for example the reference [26]. There does not exist a proof of the minimum-variance control problem using state-space representations exclusively. A first approach to provide such as proof is in [2]. The reader finds below a discussion of the minimum-variance control problem with complete observations for a Gaussian stochastic control system.

Problem 13.4.1. *Minimum-variance control of a Gaussian stochastic control system.* Consider a time-invariant Gaussian stochastic control system with representation,

$$\begin{aligned} x(t+1) &= Ax(t)+Bu(t)+Mv(t),\ x(0)=x_0,\ v(t)\in G(0,Q_v),\\ z(t) &= C_z x(t). \end{aligned}$$

The condition is imposed that the linear system specified by the system matrices $(A,B,C_z,0)$ has no system zeroes, Def. 21.5.2, on the unit circle or strictly outside the unit disc.

Define the class of linear Markov control laws of the form,

$$G_L = \left\{ g : X \rightarrow U | \exists F \in \mathbb{R}^{n_u \times n_x},\ g(x) = Fx,\ \mathrm{spec}(A+BF) \subset \mathrm{D}_o \right\}.$$

For a control law in this class, the closed-loop system has the representation,

$$\begin{aligned} x^g(t+1) &= Ax^g(t)+Bg(x^g(t))+Mv(t) = (A+BF)x^g(t)+Mv(t),\ x^g(0)=x_0,\\ z^g(t) &= C_z x^g(t),\\ u^g(t) &= Fx^g(t). \end{aligned}$$

The minimun-variance cost function is a special case of the average-cost function according to,

$$\begin{aligned} b(x^g(s),u^g(s)) &= z^g(s)^T z^g(s) = x^g(s)^T C_z^T C_z x^g(s),\\ J_{ac}(g) &= \limsup_{t_1\rightarrow\infty} \frac{1}{t_1} E\left[\sum_{s=0}^{t_1-1} x^g(s)^T C_z^T C_z x^g(s)\right],\ J_{ac} : G_L \rightarrow \mathbb{R}_+ \cup \{\infty\}. \end{aligned}$$

The problem is then to solve the following problem for the optimal control law $g^* \in G_L$ and the value J^*_{ac},

$$J^*_{ac} = \inf_{g\in G_L} J_{ac}(g) = J_{ac}(g^*).$$

Proposition 13.4.2. *Consider Problem 13.4.1. For any control law $g \in G_L$ with $g(x) = Fx$, the cost criterion has the value,*

$$J_{ac}(g) = \mathrm{tr}(C_z Q_x^g(\infty) C_z^T) < \infty, \quad \textit{where } Q_x^g(\infty) \in \mathbb{R}^{n_x\times n_x}_{pds}, \tag{13.35}$$

$$Q_x^g(\infty) = (A+BF)Q_x^g(\infty)(A+BF)^T + MM^T. \tag{13.36}$$

Proof. Denote the recursion of the state variance process by,

$$\begin{aligned} Q_x^g(t+1) &= (A+BF)Q_x^g(t)(A+BF)^T + MM^T,\ Q_x^g(0)=Q_0,\\ & Q_x^g : T \rightarrow \mathbb{R}^{n_x\times n_x}, \text{ and the limit value by,}\\ Q_x^g(\infty) &= \lim_{t\rightarrow\infty} Q_x^g(t),\\ Q_x^g(\infty) &= (A+BF)Q_x^g(\infty)(A+BF)^T + MM^T. \end{aligned}$$

That the limit exists and that $Q_x^g(\infty)$ is the unique solution of the Lyapunov equation (13.36) follows from $g \in G_L$, $g(x) = Fx$, hence $\mathrm{spec}(A+BF) \subset \mathrm{D}_o$, and Theorem

22.1.2. Then the following calculations, using a result on the convergence of a sum of a series, yield the result,

$$\begin{aligned} J_{ac}(g) &= \lim_{t_1 \to \infty} \frac{1}{t_1} \sum_{s=0}^{t_1-1} \text{tr}(E[x^g(s)x^g(s)^T]C_z^T C_z) \\ &= \lim_{t_1 \to \infty} \frac{1}{t_1} \sum_{s=0}^{t_1-1} \text{tr}(Q_x^g(s)C_z^T C_z) = \text{tr}(C_z Q_x^g(\infty) C_z^T). \end{aligned}$$

□

Proposition 13.4.3. *Consider the minimum variance control Problem 13.4.1. Consider the special case with $n_u = n_x$, hence $B \in \mathbb{R}^{n_u \times n_x}$, and impose the strong condition* $\text{rank}(B) = n_x$ *which implies that the system is stochastically controllable.*

Then an optimal control law and the value are,

$$g^*(x) = F^* x = -B^{-1}Ax, \;\; J^* = \text{tr}(C_z M (C_z M)^T) = \text{tr}(C_z M M^T C_z^T).$$

Proof. Because $\text{rank}(B) = n_x$, the matrix $F^* = -B^{-1}A$ is well defined and $g^*(x) = F^* x$ results in $(A + BF^*) = 0$, hence $g^* \in G_L$. The Lyapunov equation for the steady-state variance of the invariant distribution of the system equals,

$$Q_x^{g^*}(\infty) = (A + BF^*) Q_x^{g^*}(\infty)(A + BF^*)^T + MM^T = MM^T.$$

If $g \in G_L$ is any control law then $g(x) = Fx$ and $\text{spec}(A + BF) \subset \text{D}_o$. The associated steady-state variance is then the solution of the Lyapunov equation,

$$\begin{aligned} Q_x^g(\infty) &= (A + BF) Q_x^g(\infty)(A + BF)^T + MM^T \geq MM^T, \\ & \quad Q_x^g(\infty) \geq 0 \; \Rightarrow Q_x^g(\infty) \geq MM^T = Q_x^{g^*}(\infty), \\ J(g^*) &= \text{tr}(C_z {Q_x}^{g^*}(\infty) C_z^T) = \text{tr}(C_z M M^T C_z^T) \leq \text{tr}(C_z Q_x^g(\infty) C_z^T) = J(g). \end{aligned}$$

Thus the control law $g^* \in G_L$ is optimal. □

The approach of the above proposition seems extendable to the case where $1 \leq n_u < n_x$, $(A, B, C_z, 0)$ are the system matrices of a minimal linear system, and that there do not exist linear system zeroes.

13.5 Exercises

Problem 13.5.1. Preventive maintenance revisited. Consider a machine that over a infinite time horizon may be in an operating state, $x(t) = 1$, or in a failed state, $x(t) = 0$. The set of controls is $U = \{u_1, u_2, u_3\}$, where

- u_1 no maintenance is performed;
- u_2 light maintenance is performed;
- u_3 the machine is repaired.

The set of control policies is:

- g_1: if $x(t) = 1$ then no maintenance is performed, while if $x(t) = 0$ then the machine is repaired;
- g_2: if $x(t) = 1$ then light maintenance is performed, while if $x(t) = 0$ then the machine is repaired.

The possible transitions are then described by:

$$P_{g_1} = \begin{pmatrix} 0 \; p_1 \\ 1 \; 1-p_1 \end{pmatrix}, P_{g_2} = \begin{pmatrix} 0 \; p_2 \\ 1 \; 1-p_2 \end{pmatrix},$$

The cost function is $c : X \times U \to \mathbb{R}$

$$c(0,u_1) = c_3, c(1,u_1) = c_1, c(0,u_2) = c_3, c(1,u_2) = c_2,$$

while an average cost criterion is used. Assume that $0 < p_2 < p_1 < 1$ and $0 < c_1 < c_2 < c_3$. Determine a relation between c_1, c_2, c_3, p_1, p_2 such that in state $x = 1$ the control of no maintenance is optimal.
(Hint. Let $g_0 = g_1$ and apply the policy iteration procedure for the average cost criterion.)

13.6 Further Reading

History of infinite-horizon optimal stochastic control. The early approaches are due to D. Blackwell, [8, 9, 10]. The book by R. Howard [22] is also an early source.

Books on the subject. The average cost and the discounted cost optimal stochastic control problem are discussed in the book of D.P. Bertsekas, [7, Ch. 8], the book of P.R. Kumar and P. Varaiya [24, 4.5, 8.5, 8.6], and that of H. Kushner [25, Chapter 6, Section 9.4].

Average cost. The theoretical framework is based on the publications, [8, 10, 35, 37]. See also the paper [11].

Later weaker conditions were formulated for the existence of the value function and for existence of the optimal control law. See [6, 12]. See also publications [21, 15]. A survey paper on optimal stochastic control with average cost function is [1]

Average cost – Control of a Gaussian stochastic control system.. The books mentioned above discuss this case. The relation between the optimal stochastic control problem for a time-invariant Gaussian stochastic control system with average cost and a quadratic cost rate with the algebraic Riccati equation, is told by J.C. Willems in [41].

Average cost – Control of a state-finite stochastic control system. The example of S. Ross may be found in [30, Ex. 2, p. 143], the example in this book has a more lengthy explanation. Further research has been done on determining time-varying optimal control laws, see [18, 19] to which the reader is referred.

Optimal stochastic control with complete observations on an infinite horizon for a finite stochastic control system is treated in [5, 16, 17, 22, 27, 28, 29]. A useful bounding technique for the value function was proposed by P. Varaiya, [40]. See for the relation of the average cost between (1) a closed-loop system with a time-invariant control law and (2) a closed-loop system with a time-varying control law, [18].

Average cost – Control of a state-countable stochastic control system. See [16]. A counter example is provided in [14].

Average cost – Stochastic control systems. For the criterion of total variance distance ambiguity, see [39].

Average cost with a risk-sensitive cost function is treated by R. Howard and J. Matheson, [23]. See also the report of D.Hernández-Hernández and S.I. Marcus, for countable-state space systems, [20]. A framework for risk-sensitive control problems and its relation with stochastic differential games was developed by T. Runolfsson, [32, 33, 34].

Discounted cost. The discounted cost optimal stochastic control problem is discussed in [7]. An early reference is [9]. The positive case is treated in [10].

Discounted cost for a state-finite stochastic control system. Theorem 13.3.15 on the existence of a fixpoint of a contraction mapping is standard in analysis. Its proof is based on [13, Th. 6.44] and on the convergence of sequences, [31, Prop. 7.14].

References

1. A. Arapostathis, V.S. Borkar, E. Fernández-Gaucherand, M.K. Ghosh, and S.I. Marcus. Discrete-time controlled Markov processes with average cost criterion: A survey. *SIAM J. Control & Opt.*, 31:282–344, 1993. 525
2. C. Arnoldi and R.H. Kwong. A state space approach to minimum variance control of multivariable ARMAX systems. In *Proceedings of the 28th Conference on Decision and Control*, pages 2125–2126, New York, 1989. IEEE Press. 522, 596
3. K.J. Aström. Computer control of a paper machine - An application of linear stochastic control theory. *IBM J. Res. & Developm.*, 11:389–405, 1967. 9, 78, 120, 522, 575, 596
4. K.J. Aström. *Introduction to stochastic control*. Academic Press, New York, 1970. 376, 410, 467, 522, 575, 596
5. J. Bather. Optimal decision procedures for finite markov chains. *Adv. Applied Probab.*, 5:328–339, 521–540, 541–553, 1973. 468, 525
6. A. Bensoussan and M. Robin. On the convergence of the discrete time dynamic programming equation. *SIAM J. Control & Opt.*, 20:722–746, 1982. 525
7. D.P. Bertsekas. *Dynamic programming and stochastic control*. Academic Press, New York, 1976. 376, 405, 410, 439, 468, 502, 525, 526, 575, 595
8. D. Blackwell. Discrete dynamic programming. *Ann. Math. Statist.*, 33:719–726, 1962. 525
9. D. Blackwell. Discounted dynamic programming. *Ann. Math. Statist.*, 36:226–235, 1965. 428, 525, 526
10. D. Blackwell. Positive dynamic programming. In *Proc. 5th Berkeley Symp. Math. Statist. and Probability*, pages 415–418. University of California Press, Berkeley, 1965. 525, 526
11. D. Blackwell. On stationary policies. *J. Roy. Statist. Soc., Ser. A*, 133:33–38, 1970. 525
12. Vivek S. Borkar. Control of Markov chains with long-run average cost criterion: The dynamic programming equations. *SIAM J. Control & Opt.*, 27:642–657, 1989. 525
13. A. Browder. *Mathematical analysis - An introduction*. Undergraduate texts in mathematics. Springer-Verlag, New York, 1996. 30, 49, 424, 426, 475, 526, 635, 636, 677, 815

14. Rolando Cavazos-Cadena. A counterexample on the optimality equation in Markov decision chains with the average cost criterion. *Systems & Control Lett.*, 16:387–392, 1991. 526
15. Raúl Montes de Oca. The average cost optimality equation for Markov control processes on Borel spaces. *Systems & Control Lett.*, 22:351–357, 1994. 525
16. C. Derman. Denumerable state Markovian decision processes – Average cost criterion. *Ann. Math. Statist.*, 37:1545–1554, 1966. 525, 526
17. C. Derman. *Finite state Markov decision processes*. Academic Press, New York, 1970. 468, 525
18. C. Derman and R. Strauch. A note on memoryless rules for controlling sequential control processes. *Ann. Math. Statist.*, 37:276–279, 1966. 525, 526
19. L. Fisher and S. Ross. An example in denumerable decision processes. *Ann. Math. Statist.*, 39:674–675, 1968. 525
20. D. Hernández-Hernández and S.I. Marcus. Risk sensitive control of Markov processes in countable state space. Report, U. Maryland, Dept. of Electrical Engineering, College Park, MD, 1996. 526
21. O. Hernandez-Lerma and Lasserre J.-B. *Discrete-time Markov control processes*. Springer, New York, 1995. 468, 525
22. R. Howard. *Dynamic programming and Markov processes*. M.I.T. Press, Cambridge, 1960. 376, 467, 525
23. R.A. Howard and J.A. Matheson. Risk-sensitive Markov decision processes. *Management Science*, 18:356–369, 1972. 526, 605
24. P.R. Kumar and P. Varaiya. *Stochastic systems: Estimation, identification, and adaptive control*. Prentice Hall Inc., Englewood Cliffs, NJ, 1986. 376, 410, 468, 525, 575, 595, 596
25. H.J. Kushner. *Introduction to stochastic control*. Holt, Rinehart and Winston Inc., New York, 1971. 121, 376, 410, 467, 525
26. V. Kučera. *Discrete linear control - The polynomial equation approach*. Czechoslovak Academy of Sciences, Prague, 1979. 522, 596
27. P. Mandl. Estimation and control in Markov chains. *Adv. Appl. Probab.*, 6:40–60, 1974. 525
28. S. Ross. Non-discounted denumerable Markovian decision models. *Ann. Math. Statist.*, 39:412–423, 1968. 525
29. S.M. Ross. Arbitrary state Markov decision processes. *Ann. Math. Statist.*, 39:2118–2122, 1968. 525
30. S.M. Ross. *Applied probability models with optimization applications*. Holden-Day, San Francisco, 1970. 468, 525
31. H.L. Royden. *Real analysis, 2nd edition*. MacMillan Co., New York, 1968. 49, 526, 636
32. T. Runolfsson. Stationary risk-sensitive LQG control and its relation to LQG and H-infinity control. In *Proceedings 29th IEEE Conference on Decision and Control*, pages 1018–1023, New York, 1990. IEEE Press. 411, 526
33. T. Runolfsson. On the stationary control of a controlled diffusion with an exponential-of-integral performance criterion. In *Proceedings of the 30th Conference on Decision and Control*, pages 935–936, New York, 1991. IEEE Press. 526
34. T. Runolfsson. The equivalence between infinite horizon control of stochastic systems with exponential-of-integral performance index and stochastic differential games. *IEEE Trans. Automatic Control*, 39:1551–1563, 1994. 526
35. M. Schäl. Dynamic programming under continuity and compactness assumptions. *Adv. Appl. Probab.*, 5:24–25, 1973. 525
36. U. Shaked and P.R. Kumar. Minimum variance control of discrete time multivariable systems. *SIAM J. Control & Opt.*, 24:396–411, 1986. 522, 596
37. R. Strauch. Negative dynamic programming. *Ann. Math. Statist.*, 37:871–890, 1966. 428, 525
38. H.L. Trentelman, A.A. Stoorvogel, and M. Hautus. *Control theory for linear systems*. Springer, United Kingdom, 2001. 485, 781, 784, 785, 786, 808
39. Ioannis Tzortzis, Charalambos D. Charalambous, and Themistokles Charalambous. Infinite horizon average cost dynamic programming subject to total variation distance ambiguity. *SIAM J. Control & Opt.*, 57:2843–2872, 2019. 526, 742

40. P. Varaiya. Optimal and suboptimal stationary controls for Markov chains. *IEEE Trans. Automatic Control*, 23:388–394, 1978. 526

41. J.C. Willems. Least squares stationary optimal control and the algebraic Riccati equation. *IEEE Trans. Automatic Control*, 16:621–634, 1971. 525, 867, 885

Chapter 14
Stochastic Control with Partial Observations on a Finite Horizon

Abstract In stochastic control with partial observations, the control law at any time can depend only on the past outputs and the past inputs of the stochastic control system. Neither is available to the control law the current state nor the past states. Control theory for stochastic systems with partial observations is poorly developed and poorly understood. The approach to this control problem is to first determine a stochastic realization of the stochastic control system with respect to the past outputs and the past inputs and with a finite-dimensional state set. Secondly, one constructs a control law or an optimal control law such that the closed-loop control system is another stochastic realization of the output process. The approach is illustrated for a Gaussian stochastic control system and for an output-finite-state-finite stochastic control system. The tracking problem is treated.

Key words: Stochastic control. Partial observations. Finite horizon.

The approach to the stochastic control problem with partial observations is to first determine a stochastic realization of the stochastic control system with respect to the past outputs and the past inputs and with a finite-dimensional state set. Secondly, one constructs a control law or an optimal control law such that the closed-loop control system is another stochastic realization of the output process with the needed properties.

A reader who first learns about stochastic control for partially-observed systems is advised to focus attention on the Sections 14.2, 14.3, 14.4.1, 14.4.2.

14.1 Motivation

Control problems with partial observations include the control of a mooring tanker and control of a paper machine. The latter example is briefly described for future reference.

Example 14.1.1. *Control of a paper machine*. This example is a continuation of that of Example 4.1.1. Recall that a model for paper basis weight has been derived consisting of the ARMAX representation

$$y(t) = \sum_{s=0}^{p} a(s)y(t-s) + \sum_{s=0}^{p} b(s)u(t-s) + \sum_{s=0}^{p} c(s)v(t-s), \tag{14.1}$$

where y represents the measurement of basis weight of paper, u the input process, and v is a Gaussian white noise process. Note that the state components $v(t),\ v(t+1),\ldots,v(t+p)$ are not observed hence this is a stochastic control problem with partial observations.

The control objective is to minimize the variance in basis weight. Optimal stochastic control theory for this problem is called minimum variance control.

14.2 Problem Formulation

Problem 14.2.1. The *optimal stochastic control problem with partial observations on a finite horizon*.

Consider a recursive stochastic control system with partial observations on a finite-horizon with the system representation,

$$\begin{aligned} x(t+1) &= f(t,x(t),u(t),v(t)), x(0) = x_0, \\ y(t) &= h(t,x(t),u(t),v(t)), \\ z(t) &= h_z(t,x(t),u(t)),\ \forall\, t \in T\backslash\{t_1\},\ \ z(t_1) = h_{z,1}(x(t_1)), \\ &\quad f : T \times X \times U \times V \to X,\ \ h : T \times X \times U \times V \to Y, \\ &\quad h_z : T \times X \times U \to Z,\ h_{z,1} : X \to Z, \end{aligned}$$

where (Ω,F,P) denotes a probability space, $T(0:t_1) = \{0,1,\ldots,t_1\}$ a finite horizon for $t_1 \in \mathbb{Z}_+$, $(X,B(X))$, $(Y,B(Y))$, $(U,B(U))$, $(V,B(V))$, and $(Z,B(Z))$ are measurable spaces with X, Y, U, Z measurable subsets of tuples of the real numbers, $x_0 : \Omega \to X$ is a random variable, $v : \Omega \times T \to V$ is a stochastic process consisting of an independent sequence, $F^{v}_{t_1-1}$ and F^{x_0} are independent σ-algebras, and the Borel measurable maps f, h, h_z, and $h_{z,1}$.

Consider the past-output and past-input information structure,

$$\{F^{y}_{t-1} \vee F^{u}_{t-1}, t \in T\}.$$

Consider the set G of control laws compatible with this information structure. For any $g \in G$, $g = \{g_0,\ldots,g_{t_1-1}\}$, $g_0 \in U$, and, for $t \in T(1:t_1-1)$, $g_t : Y^t \times U^t \to U$.

For any $g \in G$ the closed-loop stochastic control system has the representation,

$$\begin{aligned}
x^g(t+1) &= f(t,x^g(t),g_t(y^g(0:t-1),u^g(0:t-1)),v(t)), \;\; x^g(0)=x_0,\\
y^g(t) &= h(t,x^g(t),g_t(y^g(0:t-1),u^g(0:t-1)),v(t)), \qquad (14.2)\\
z^g(t) &= h_z(t,x^g(t),g_t(y^g(0:t-1),u^g(0:t-1))), \; \forall\, t \in T\backslash\{t_1\},\\
z^g(t_1) &= h_{z,1}(x^g(t_1)),\\
u^g(t) &= g_t(y^g(0:t-1),u^g(0:t-1)),\\
y^g(0:t-1) &= (y^g(0),y^g(1),\ldots,y^g(t-1))^T,\\
u^g(0:t-1) &= (u^g(0),u^g(1),\ldots,u^g(t-1))^T.
\end{aligned}$$

The index g is used on x^g,y^g,u^g to indicate that these processes are defined only if the control law $g \in G$ is specified and than these processes depend on the control law g. Note that for any time $t \in T$, only after $g_0,g_1,\ldots,g_t$ are specified then the state $x^g(t)$ and the output $y^g(t)$ are random variables hence the measure on $F_t^x \vee F_{t-1}^y$ is specified.

Note that g_t depends on $y^g(0),\ldots,y^g(t-1)$, and on $u^g(0),\ldots,u^g(t-1)$, but not on $y^g(t)$. Dependence of g_t on $y^g(t)$ would make equation (14.2) an equation for $y^g(t)$ and then conditions on the existence of $y^g(t)$ are necessary. Therefore the results of this chapter cannot be compared directly with references that do not satisfy this convention.

For the understanding of the reader, there follows the ordered sequence of definitions of the successive variables for every time $t \in T$,

$$\begin{aligned}
u^g(t) &= g_t(y^g(0:t-1),u^g(0:t-1)),\\
x^g(t+1) &= f(t,x^g(t),u^g(t),v(t)),\\
y^g(t) &= h(t,x^g(t),u^g(t),v(t)),\\
y^g(0:t) &= (y^g(0),y^g(1),\ldots,y^g(t))^T, \; u^g(0:t) = (u^g(0),u^g(1),\ldots,u^g(t))^T.
\end{aligned}$$

Consider the positive cost function,

$$J(g) = E\left[\sum_{s=0}^{t_1-1} b(s,z^g(s)) + b_1(z^g(t_1))\right],$$
$$b : T \times Z \to \mathbb{R}_+, \; b_1 : Z \to \mathbb{R}_+, \; J : G \to \mathbb{R}_+.$$

The functions b and b_1 are assumed to be Borel measurable.

The problem is to solve the optimal stochastic control problem,

$$J^* = \inf_{g \in G} J(g) = J(g^*).$$

The problem implies to determine the *value* J^* and an optimal control law $g^* \in G$ so that the above formulas hold, if an optimal control law exists. If no optimal control law exists then the problem asks, for any $\varepsilon \in (0,\infty)$, for an ε-optimal control law $g_\varepsilon \in G$ such that,

$$J^* < J(g_\varepsilon^*) < J^* + \varepsilon.$$

Assumption 14.2.2 *Consider the stochastic control system of Problem 14.2.1. Define the conditions:*

1. *stochastic controllability for the relation between the input u and the state x;*
2. *stochastic observability for the relation between the state x and the observed output y;*
3. *stochastic observability for the relation between the state x and the observed output z;*
4. *supportability of the state process for the relation from the noise process v to the state process x.*

It is possible to relax several of the above conditions from controllability to stabilizability or from observability to detectability. This will not be specified in this chapter.

In each specific optimal stochastic control problem, those of the above conditions which are assumed to hold, will be specified.

The conditions of Assumption 14.2.2 seem sufficient for the optimal stochastic control problem.

In a stochastic control problem for a partially-observed stochastic system, the control law has two tasks:

1. To regulate the state so as to miminize the cost function.
2. To reduce uncertainty about the current state of the stochastic control system so as to better meet task 1.

These two tasks have to be combined according to the cost function. From the solution one can deduce how the two tasks are related. It will be shown how these tasks are performed in the special cases of stochastic control problems for which a solution has been obtained.

In several publications, a transformation of the above problem is proposed. For a set of control laws, and for any control law in this set, the input equals $u^g(t) = g_t(y^g(0:t-1), u^g(0:t-1))$. Therefore this input is a function of past observations and of past inputs. Therefore, the information structure at time $t \in T$, $F_t^{y(0:t-1)} \vee F_t^{u(0:t-1)}$, can be transformed to $F_t^{y(0:t-1)}$ by substitution of the inputs and writing the resulting closed-loop equations as functions of past outputs only. This is a correct substitution.

The advice of the author is not to use substitution described above. There are control problems for networked stochastic systems where the optimal control law depends on the last-available input or on a short sequence of recent inputs. By keeping the generality of the problem formulation, no restrictions are imposed on the problem. In stochastic control problems of networked stochastic systems one has to work with the proposed general model because there exist control laws depending on the past inputs.

In the subsequent sections the following procedure will be used for the solution of an optimal stochastic control problem with partial observations. The proposal to approach the problem in this way is due to C. Striebel in her paper published in 1965, [47].

Procedure 14.2.3 Control synthesis procedure for stochastic control with partial observations. *Consider the stochastic control problem with partial observations, Problem 14.2.1. This control synthesis procedure is defined by the steps:*

1. *If not already so specified, construct a stochastic realization of the stochastic control system which is measurable on the information structure with a finite or finite-dimensional state set. Refer to this system theoretic object as the* observation-based stochastic control system. *A way to construct such a stochastic realization is to construct the conditional filter system or, in a special case, the filter system of the Kalman filter. In the literature the term 'information system' is occasionally used for an observation-based stochastic control system.*
2. *Project the cost rate and the terminal cost on the information structure using the stochastic realization of the previous step. These variables will become functions of the state of the observation-based stochastic control system.*
3. *Solve the optimal stochastic control problem for the observation-based stochastic control system with the projected cost rates by a dynamic programming procedure, slightly different from that of the complete observations case. Prove that the value function is a lower bound on the conditional-cost-to-go for any control law and, if a candidate optimal control law exists, prove its optimality.*

This synthesis procedure is used in the Sections 14.4.1 and 14.5. The example of LEQG discussed in Section 14.4.4 shows that the above synthesis procedure may lead to an approach that differs significantly from that for a Gaussian stochastic control system with a quadratic cost function.

In the literature there is the concept of separability of a stochastic control system with partial observations. With the increased understanding of control theory and with the availability of other examples than that of a Gaussian stochastic control system with a quadratic cost function, it seems that this property does not always hold and is not useful in control theory. Therefore the concept of separability is not used in this book.

14.3 Stochastic Realization of a Stochastic Control System

Problem 14.2.1 of optimal stochastic control for a stochastic control system with partial observations cannot be solved directly by the dynamic programming method of Chapter 12.

As suggested by C. Striebel [47] in 1965, one must search for a new state process that is measurable with respect to the information structure. This then leads to the observation-based stochastic control system defined below.

Suppose that the past-output and past-input information structure is available to the controller. At any time $t \in T$ one may define the vector of the past-outputs and past-inputs as an information state, say

$$x_{is}(t) = (y(t-1), \ldots, y(0), u(t-1), \ldots, u(0))^T.$$

The information state could be the state of the new stochastic control system. Such a formulation meets several of the conditions of the information system defined below. But note that the memory needed to store this information state grows with time $t \in T$ without any bound. If the time variable goes to infinity then the memory needed grows beyond any bound and thus cannot be stored in any computer which necessarily has finite memory. Therefore this suggestion is not practical.

The conclusion from the previous paragraph is that it is best to impose a condition that the observation-based stochastic control system has a finite or a finite-dimensional state set. This condition is imposed in realization theory for almost all sets of control systems, for example for a linear control system and for a Gaussian system. The following definition is now motivated.

Definition 14.3.1. *Observation-based stochastic control system*. Consider the optimal stochastic control problem of Problem 14.2.1. Recall that the information structure is,

$$\{F^y_{t-1} \vee F^u_{t-1},\ t \in T\} \ \Rightarrow\ F^{u(t)} \subseteq F^y_{t-1} \vee F^u_{t-1},\ \forall\, t \in T.$$

The *observation* refers to the combined output and input process, (y,u).

Define an *observation-based stochastic control system* for the stochastic control system of this problem by the conditions,

$$\begin{aligned} &\forall\, t \in T,\ (x_o(t),u(t)) \mapsto \text{cpdf}((x_o(t+1),y(t),z(t))|\ F^y_{t-1} \vee F^u_{t-1}) \\ &= \text{cpdf}((x_o(t+1),y(t),z(t))|\ F^{x_o(t),u(t)}),\ \text{ and } F^{x_o(t)} \subseteq F^y_{t-1} \vee F^u_{t-1},\ \forall\, t \in T. \end{aligned}$$

Note that in the above formulation the measurability of the state process is restricted as described above.

A *observation-based stochastic control system representation* is defined as the recursive stochastic control system,

$$\begin{aligned} x_o(t+1) &= f_o(t,x_o(t),u(t),y(t),v_o(t)),\ x_o(0) = x_{o,0} \in X_o, \\ y(t) &= h_o(t,x_o(t),u(t),v_o(t)), \\ z(t) &= h_{z,o}(t,x_o(t),u(t)), \\ &\quad v_o : \Omega \times T \to V_o \subseteq \mathbb{R}^{n_{v_o}},\ \forall\, t \in T,\ F^{v0(t)} \subseteq F^y_t \vee F^u_{t-1}, \\ &\quad v_o \text{ is a sequence of independent random variables.} \end{aligned}$$

It has then to be proven that the above defined control system representation is an observation-based stochastic control system.

The set of control laws is defined in the problem statement as those of the form, for $t \in T$, $u(t) = g_t(y(0:t-1),u(0:t-1))$. It remains to prove, see the subsequent sections, that the optimal control law can be formulated in terms of the state of the observation-based stochastic control system, $u(t) = g^*_t(x_o(t))$.

The reader is alerted about the formulation of an ARMAX representation used in the literature, Example 4.1.1, which in general is not an observation-based stochastic system representation. That representation of a stochastic system in the single-input-single-output case has the form,

$$
\begin{aligned}
y(t) &= \sum_{i=1}^{n_y} a_i y(t-1) + \sum_{j=0}^{n_u} b_i u(t-i) + \sum_{k=0}^{n_w} c_i w(t-i), \\
& \quad y : \Omega \times T \rightarrow \mathbb{R},\ u : \Omega \times T \rightarrow \mathbb{R},\ w : \Omega \times T \rightarrow \mathbb{R},
\end{aligned}
$$

where y represents an output process, u represents in input process, and w represents a noise process, meaning a sequence of independent random variables. The initial condition at time zero is almost never specified. The above system representation can be transformed to a state-space system representation by the steps,

$$
\begin{aligned}
x(t) &= \begin{pmatrix} y(t-1) \\ y(t-2) \\ \vdots \\ y(t-n_y) \end{pmatrix},\ x_u(t) = \begin{pmatrix} u(t-1) \\ u(t-2) \\ \vdots \\ u(t-n_u) \end{pmatrix},\ x_w(t) = \begin{pmatrix} w(t-1) \\ w(t-2) \\ \vdots \\ w(t-n_w) \end{pmatrix}, \\
x(t) &= \begin{pmatrix} x_y(t) \\ x_u(t) \\ x_w(t) \end{pmatrix}, \\
x(t+1) &= Ax(t) + Bu(t) + Mw(t), \\
y(t) &= Cx(t) + Du(t) + Nw(t).
\end{aligned}
$$

Note that then,

$$
F^{x(t)} \subseteq F^y_{t-1} \vee F^u_{t-1} \vee F^w_{t-1},\ \forall\, t \in T;\ \text{but}\ F^{x(t)} \not\subseteq F^y_{t-1} \vee F^u_{t-1}.
$$

Therefore the above system representation is not an observation-based stochastic system representation. Only if the noise process w is measurable with respect to the observation filtration, $F^{w(t)} \subseteq F^y_{t-1} \vee F^u_{t-1}$ for all $t \in T$, then it is an observation-based stochastic system representation.

There are several ways to establish the existence of an observation-based stochastic control system:

1. *The approach of LQG.* The approach of W.M. Wonham for the case of a Gaussian stochastic control system, [58], which is described in Section 14.4.1.
2. *The approach of deriving the observation-based stochastic control system via filtering.* The literature on this approach is not always clear. The approach can be used only if the filtration of the observation process does not depend on the control law used. In the case of output-finite-state-finite stochastic system, this issue is not discussed at all, one assumes that the filter equations take the same form as in case of no dependence on the control law. A second issue is whether the conditional distribution of the next state based on past observations has an analytic form which depends on a finite-number of parameters or on a variable taking values in a measurable subset of a finite-dimensional space. This requires conditions on the stochastic control system.
3. *The approach of assuming one starts with an observation-based stochastic control system.* One can assume that the optimal stochastic control problem is formulated for a stochastic control system which is already an observation-based stochastic control system. This approach is used in information theory where it

is assumed that the state of the channel can be expressed as a function of future outputs without noise. There is a modeling issue with many problems of information theory that will not be discussed further in this book. The author advises readers to use this approach if it is not possible to proceed otherwise.

4. *The approach of a conditional filter*. The use of a conditional filter, see the conditional-Kalman filter of Theorem 8.9.2 for the case of a Gaussian stochastic system. This approach is described in more detail at the end of Section 14.4.1.

In the literature one may find the concept of a separated control law. Those definitions are too vague for a solid control theory for stochastic control systems with partial observations. Hence that concept will not be used in this book.

14.4 Stochastic Control of a Gaussian Stochastic Control System

14.4.1 Stochastic Realization by Filtering

The stochastic control problem for a Gaussian stochastic control system with partial observations and additive cost function has been solved as described in this chapter. Needed is first the existence of an observation-based stochastic control system. There are two approaches to the existence of such a stochastic realization both discussed in this section:

1. the derivation of the stochastic realization via Kalman filtering, formulated by W.M. Wonham in [58] for a continuous-time stochastic control system; this approach seems limited to the case of a stochastic control system with a linear dynamics and with a linear output equation;
2. the approach of the conditional Kalman filter, see below in this section.

Problem 14.4.1. The *optimal stochastic control problem for a Gaussian stochastic control system with partial observations on a finite horizon with a quadratic cost function*. Consider a Gaussian stochastic control system representation,

$$\begin{aligned}
x(t+1) &= A(t)x(t)+B(t)u(t)+M(t)v(t), x(0)=x_0,\\
y(t) &= C(t)x(t)+D(t)u(t)+N(t)v(t),\\
z(t) &= C_z(t)x(t)+D_z(t)u(t),\ \forall\, t \in T(0:t_1-1),\\
z(t_1) &= C_z(t_1)x(t_1);\\
&\quad x_0 \in G(m_{x_0},Q_{x_0}),\ \ v(t) \in G(0,I),\ T(0:t_1)=\{0,1,\ldots,t_1\};\\
&\quad n_y \leq n_v,\ \forall\, t \in T,\ \text{rank}(N(t))=n_y,\ \text{ hence } N(t)N(t)^T \succ 0,\\
&\quad n_z \geq n_u,\ \forall\, t \in T,\ \text{rank}(D_z(t))=n_u,\ \text{ hence } D_z(t)^T D_z(t) \succ 0.
\end{aligned}$$

Consider further the past-output and past-input information structure,

$$\{F^y_{t-1} \vee F^u_{t-1},\ t \in T\}$$

and the corresponding set G of control laws. For any $g \in G$, $g = \{g_0, g_1, \ldots, g_{t_1-1}\}$, $g_0 \in U$, and for $t \in \mathbb{Z}_+$, $g_t : Y^t \times U^t \to U$, the closed-loop stochastic control system is described by,

$$\begin{aligned}
x^g(t+1) &= A(t)x^g(t) + B(t)g_t(y^g(0:t-1), u^g(0:t-1)) + M(t)v(t),\\
&\quad x^g(0) = x_0,\\
y^g(t) &= C(t)x^g(t) + D(t)g_t(y^g(0:t-1), u^g(0:t-1)) + N(t)v(t),\\
u^g(t) &= g_t(y^g(0:t-1), u^g(0:t-1));\\
z^g(t) &= C_z(t)x^g(t) + D_z(t)g_t(y^g(0:t-1), u^g(0:t-1)),\\
&\quad \forall\, t \in T(0:t_1-1);\\
z^g(t_1) &= C_z(t_1)x^g(t_1);\ \text{where},\\
y^g(0:t-1) &= (y^g(0), y^g(1), \ldots, y^g(t-1))^T,\\
u^g(0:t-1) &= (u^g(0), u^g(1), \ldots, y^g(t-1))^T.
\end{aligned}$$

Define the cost function,

$$J(g) = E\left[\sum_{s=0}^{t_1-1} (z^g(s))^T z^g(s) + z^g(t_1)^T z^g(t_1)\right],\ \ J : G \to \mathbb{R}_+.$$

The problem is then to solve the following optimal stochastic control problem,

$$J^* = \inf_{g \in G} J(g) = J(g^*).$$

If one starts the dynamic programming approach for Problem 14.4.1 according to Procedure 14.2.3 than the first step is to determine the conditional distribution of $x^g(t)$ conditioned on $F^{y,g}_{t-1} \vee F^{u,g}_{t-1}$. Recall that,

$$F^{y,g}_{t-1} \vee F^{u,g}_{t-1} = \sigma(\{y^g(0), \ldots, y^g(t-1)\}) \vee \sigma(\{u^g(0), \ldots, u^g(t-1)\}) \tag{14.3}$$

In equation (14.3) the index g is attached to the σ-algebra at the same level as y for notational reasons that will become clear later in the discussion. The above question is thus equivalent to determining the conditional characteristic function,

$$E[\exp(iw_x^T x^g(t))|F^{y,g}_{t-1} \vee F^{u,g}_{t-1}].$$

Note that the distribution depends on the control law g, in particular on the functions $g_0, g_1(.), \ldots, g_{t-1}(.)$.

If the control law g is nonlinear then the processes x^g and y^g need not be Gaussian. Therefore filter theory for Gaussian systems cannot be applied directly. Another approach is needed. This fundamental issue is a source of confusion in the literature.

Theorem 14.4.2. *Consider Problem 14.4.1. Fix a control law $g \in G$.*

(a)For any time $t \in T$,

$$F^{y,g}_t \vee F^{u,g}_t = F^{y,0}_t \vee F^{u,0}_t, \tag{14.4}$$

where $\{F_t^{y,0},\ t \in T\}$ *represents the* σ*-algebra family obtained if in the Gaussian stochastic control system representation the control law is used that makes the input process identically zero for all times; thus for the control law* $g = 0$ *such that for all* $t \in T$, $u^g(t) = g_t(.) = 0$. *This rather remarkable property shows that Problem 14.4.1 is very special.*

(b)The requested conditional characteristic function is Gaussian and specified by,

$$\begin{aligned}&E[\exp(iw_x^T x^g(t))|F_{t-1}^{y,g} \vee F_{t-1}^{u,g}] && (14.5)\\ &= \exp(iw_x^T \hat{x}^g(t) - w_x^T Q_f(t) w_x/2)\ a.s.,\ \ \forall w_x \in \mathbb{R}^{n_x},\ \forall t \in T(1:t_1),\end{aligned}$$

where $\hat{x}^g : \Omega \times T \to \mathbb{R}^{n_x}$ *and* $Q_f : T \to \mathbb{R}_{spd}^{n_x \times n_x}$ *are determined by the recursions,*

$$\begin{aligned}&\hat{x}^g(t+1) && (14.6)\\ &= A(t)\hat{x}^g(t) + B(t)u^g(t) + K(t, Q_f(t))[y^g(t) - C(t)\hat{x}^g(t) - D(t)u^g(t)]\\ &= f_{KF}(t, \hat{x}^g(t), Q_f(t), y^g(t), u^g(t)),\ \hat{x}^g(0) = m_{x_0}, && (14.7)\\ &Q_f(t+1) && (14.8)\\ &= A(t)Q_f(t)A(t)^T + M(t)M(t)^T +\\ &\quad -[A(t)Q_f(t)C(t)^T + M(t)N(t)^T][C(t)Q_f(t)C(t)^T + N(t)N(t)^T]^{-1} \times\\ &\quad \times[A(t)Q_f(t)C(t)^T + M(t)N(t)^T]^T = f_{FR}(t, Q_f(t)),\ \ Q_f(0) = Q_{x_0}, && (14.9)\\ &K(t, Q_f(t))\\ &= [A(t)Q_f(t)C(t)^T + M(t)N(t)^T][C(t)Q_f(t)C(t)^T + N(t)N(t)^T]^{-1} && (14.10)\end{aligned}$$

The recursion of equation (14.7) is called the Kalman filter of stochastic control *and the recursion of equation (14.9) is called the* Filter Riccati Recursion *for this control problem. The recursion of the Kalman filter for the conditional mean depends on the output and on the input of the system while that of the Filter Riccati Recursion depends only on the conditional variance but neither on the output nor on the input.*

For later use, define the functions,

$$\begin{aligned}&f_{KF} : T \times \mathbb{R}^{n_x} \times \mathbb{R}_{pds}^{n_x \times n_x} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_x}, && (14.11)\\ &f_{FR} : T \times \mathbb{R}_{pds}^{n_x \times n_x} \to \mathbb{R}_{pds}^{n_x \times n_x},\\ &f_{KF}(t, x_v, Q_v, y_v, u_v)\\ &= A(t)x_v + B(t)u_v + K(t, Q_v)[y_v - C(t)x_v - D(t)u_v], && (14.12)\\ &f_{FR}(t, Q_v)\\ &= A(t)Q_vA(t)^T + M(t)M(t)^T +\\ &\quad -[A(t)Q_vC(t)^T + M(t)N(t)^T][C(t)Q_vC(t)^T + N(t)N(t)^T]^{-1} \times\\ &\quad \times[A(t)Q_vC(t)^T + M(t)N(t)^T]^T. && (14.13)\end{aligned}$$

(c)For all $t \in T$ *and all* $w_x \in \mathbb{R}^{n_x}$,

$$E[\exp(iw_x^T(x^g(t) - \hat{x}^g(t)))|F_{t-1}^{y,g} \vee F_{t-1}^{u,g}] = \exp(-1/2 w_x^T Q_f(t) w_x), \qquad (14.14)$$

which, because Q_f is a deterministic function, implies that $(x^g(t) - \hat{x}^g(t))$ is independent of $F^{y,g}_{t-1} \vee F^{u,g}_{t-1}$ for all $t \in T$.

(d) *Define the* innovation process,

$$\begin{aligned} &\overline{v}^g : \Omega \times T(0:t_1-1) \to \mathbb{R}^{n_y}, \\ \overline{v}^g(t) &= y^g(t) - C(t)\hat{x}^g(t) - D(t)u^g(t). \end{aligned}$$

Then $\overline{v}^g$ is a Gaussian white noise process with,

$$\begin{aligned} &\overline{v}^g(t) \in G(0, Q_{\overline{v}}(t)),\ \overline{v}^g(t) \text{ is } F^{y^g}_t \vee F^{u^g}_t \text{ measurable}, \\ Q_{\overline{v}}(t) &= C(t)Q_f(t)C(t)^T + N(t)N(t)^T, \qquad (14.15) \\ &F^{\overline{v}^g(t)},\ \ F^{y,g}_{t-1} \vee F^{u,g}_{t-1}, \text{ are independent}, \forall\, t \in T(0:t_1-1). \end{aligned}$$

By definition, the variance $Q_{\overline{v}}$ of the innovations process does not depend on the control law!

(e) *Define the* observation-based Gaussian stochastic control system *associated with the stochastic control system considered and with respect to the information structure, as the Gaussian stochastic control system,*

$$\begin{aligned} \hat{x}^g(t+1) &= A(t)\hat{x}^g(t) + B(t)u^g(t) + K(t, Q_f(t))\overline{v}^g(t),\ \hat{x}^g(0) = m_{x_0}, \\ y^g(t) &= C(t)\hat{x}^g(t) + D(t)u^g(t) + \overline{v}^g(t), \\ &\forall\, t \in T(0:t_1-1),\ \ \overline{v}^g(t) \in G(0, Q_{\overline{v}}(t)), \\ Q_{\overline{v}}(t) &= C(t)Q_f(t)C(t)^T + N(t)N(t)^T, \\ &\{\overline{v}^g(t),\ F^{y^g}_t \vee F^u_t,\ \forall\, t \in T(0:t_1-1)\}. \end{aligned}$$

The innovation process $\overline{v}$ is a Gaussian white noise process with the indicated mean and variance functions, and it is adapted to the indicated filtration.

This system representation is an observation-based stochastic control system as defined in Def. 14.3.1

(f) *The formulas (14.16) and (14.19) are needed in the dynamic programming procedure that follows.*

$$\begin{aligned} &\forall\, t = 1, \ldots, t_1 - 1 \in T,\ \ w_x \in \mathbb{R}^{n_x}, \\ &E[\exp(iw_x^T x(t+1))|F^{y,g}_{t-1} \vee F^{u,g}_{t-1}] \qquad (14.16) \\ &= \exp(iw_x^T[A(t)\hat{x}^g(t) + B(t)u^g(t)] - \frac{1}{2}w_x^T Q_{f_+}(t)w_x), \text{ where}, \\ Q_{f_+}(t) &= Q_f(t+1) + K(t, Q_f(t))Q_{\overline{v}}(t)K(t, Q_f(t))^T; \qquad (14.17) \\ &E[\exp(iw_x^T \hat{x}^g(t+1))|F^{y,g}_{t-1} \vee F^{u,g}_{t-1}] \\ &= \exp(iw_x^T[A(t)\hat{x}^g(t) + B(t)u^g(t)] - \frac{1}{2}w_x^T Q_{K\overline{v}}(t, Q_f(t)w_x)), \qquad (14.18) \\ &\forall w_x \in \mathbb{R}^{n_x}, \\ Q_{K\overline{v}}(t) &= K(t, Q_f(t))Q_{\overline{v}}(t)K(t, Q_f(t))^T, \\ &= [A(t)Q_f(t)C(t)^T + M(t)N(t)^T][C(t)Q_f(t)C(t)^T + N(t)N(t)^T]^{-1} \times \\ &\quad \times [A(t)Q_f(t)C(t)^T + M(t)N(t)^T]^T. \qquad (14.19) \end{aligned}$$

Note that the conditional distribution of $x(t)$ conditioned on $F_{t-1}^{y,g} \vee F_{t-1}^{u,g}$ displayed in equation (14.5) is such that the input influences only the conditional mean $\hat{x}$ and does not affect the conditional variance Q_f. Therefore the control cannot be used to reduce the uncertainty about the state of the system, meaning that the input cannot reduce the conditional error variance. This property has been termed *neutrality* in [39]. The Gaussian stochastic control system is also in this regard a rather special case. This fact also simplifies the solution of the stochastic control problem.

Proof. (a) Fix $g \in G$. Define $x_1^g, x_2 : \Omega \times T \to \mathbb{R}^{n_x}$ and $y_1^g, y_2 : \Omega \times T \to \mathbb{R}^{n_y}$ by,

$$\begin{aligned} x_1^g(t+1) &= A(t)x_1^g(t) + B(t)u^g(t), x_1^g(0) = 0, \\ y_1^g(t) &= C(t)x_1^g(t) + D(t)u^g(t), \\ x_2(t+1) &= A(t)x_2(t) + M(t)v(t), x_2(0) = x_0, \\ y_2(t) &= C(t)x_2(t) + N(t)v(t). \end{aligned}$$

Because of the linearity of the above equations it can be proven by induction that for all $t \in T$ both the left-hand side and the right-hand side of the following expressions satisfy the same recursions or equations respectively.

$$x^g(t) = x_1^g(t) + x_2(t), \;\; y^g(t) = y_1^g(t) + y_2(t), \tag{14.20}$$

(a.1). Note that by definition of $F_t^{y,0}$

$$F_t^{y,0} = F_t^{y_2} \tag{14.21}$$

for all $t \in T$. Note also that x_2 and y_2 do not depend on the control law $g \in G$.

Claim 1. It will be proven by induction that,

$$F_0^{y_1} = \{\Omega, \emptyset\}, \;\; F_s^{x_1^g} \subset F_{s-1}^{y,g}, \;\; F_s^{y_1^g} \subset F_{s-1}^{y,g}, \forall\, s = 1,2,\ldots,t_1. \tag{14.22}$$

Note that $u^g(0) = g_0 \in U$ is deterministic by definition of a past-output control law and hence measurable with respect to any σ-algebra. Then,

$$\begin{aligned} u^g(t) &= g_t(y^g(0),\ldots,y^g(t-1),u^g(0),\ldots,u^g(t-1)) && (14.23) \\ & \qquad \forall\, t = 1,\ldots,t_1, \;\; \text{so } F_t^{u^g} \subset F_{t-1}^{y,g} \;\; \forall t \in T. && (14.24) \\ y_1^g(0) &= C(0)x_1^g(0) + D(0)u^g(0) = D(0)g_0 \in \mathbb{R}^{n_y}, && (14.25) \\ x_1^g(1) &= A(0)x_1^g(0) + B(0)u^g(0) = B(0)g_0 \in \mathbb{R}^{n_x}, && (14.26) \\ u^g(1) &= g_1(y^g(0),u^g(0)) \in \mathbb{R}^{n_u}, && (14.27) \\ y_1^g(1) &= C(1)x_1^g(1) + D(1)u^g(1) \in \mathbb{R}^{n_y}, && (14.28) \\ & \Rightarrow F_1^{x_1^g} \subset F_0^{y,g}, \;\; F_1^{y_1^g} \subset F_0^{y,g}. && (14.29) \end{aligned}$$

Suppose that equation (14.22) holds for $s = 1,2,\ldots,t-1$ for $t \in T$. It will be proven that equation (14.22) holds for t. Note that,

$$\begin{aligned}
&\quad u^g(t-1) \text{ is } F^{y,g}_{t-2} \text{ measurable by (14.23),}\\
&\quad x^g_1(t-1) \text{ is } F^{y,g}_{t-2}, \text{ measureable by the induction hypothesis,}\\
x^g_1(t) &= A(t-1)x^g_1(t-1)+B(t-1)u^g(t-1)\text{is } F^{y,g}_{t-1}, \text{ measureable}\\
&\quad F^{x^g_1}_{t-1} \subseteq F^{y,g}_{t-2} \subseteq F^{y,g}_{t-1}, \text{ by the induction hypothesis,}\\
F^{x^g_1}_t &= (F^{x^g_1(t)} \vee F^{x^g_1}_{t-1}) \subset F^{y,g}_{t-1};\\
y^g_1(t) &= C(t)x^g_1(t)+D(t)u^g(t), \text{ is } F^{y,g}_{t-1}, \text{ measurable}\\
&\quad F^{y^g_1}_{t-1} \subseteq F^{y,g}_{t-2} \subseteq F^{y,g}_{t-1}, \text{ by the induction hypothesis,}\\
F^{y^g_1}_t &= (F^{y^g_1(t)} \vee F^{y^g_1}_{t-1}) \subset F^{y,g}_{t-1}.
\end{aligned}$$

Induction then proves the claim.
(a.2) It will be proven by induction that,

$$F^{y_2}_s = F^{y,0}_s \subseteq F^{y,g}_s \ \forall s \in T. \tag{14.30}$$

Note that $g_0 \in U$,

$$\begin{aligned}
y_2(0) &= y^g(0)-y^g_1(0) = y^g(0)-D(0)g_0, \text{ is } F^{y,g}_0 \text{ measureable,}\\
F^{y,0}_0 &= F^{y_2}_0 \subset F^{y,g}_0.
\end{aligned}$$

Suppose that (14.30) holds for $s = 0,1,\ldots,t-1$.

$$\begin{aligned}
y_2(t) &= y^g(t)-y^g_1(t)\\
&\quad F^{y^g_1}_t \subset F^{y,g}_{t-1} \subset F^{y,g}_t, \text{ by (14.22) and increasingness,}\\
&\Rightarrow F^{y(t),0} = F^{y_2(t)} \subset F^{y,g}_t, \text{ by } y_2(t) = y^g(t)-y^g_1(t),\\
&\Rightarrow F^{y,0}_t = F^{y_2}_t = (F^{y_2(t)} \vee F^{y_2}_{t-1}) \subset F^{y,g}_t, \text{ by (14.30) for } s = t-1.
\end{aligned}$$

(a.3) It will be proven by induction that,

$$F^{y,g}_s \subset F^{y,0}_s = F^{y_2}_s, \forall\ s \in T. \tag{14.31}$$

Note that, by $g_0 \in U$,

$$y^g(0) = y^g_1(0)+y_2(0) = D(0)g_0+y_2(0),\ F^{y,g}_0 \subset F^{y,0}_0 = F^{y_2}_0.$$

Suppose that Equation (14.31) holds for $s = 0,1,\ldots,t-1$ for $t \in T$. Then,

$$\begin{aligned}
y^g(t) &= y^g_1(t)+y_2(t),\\
&\quad F^{y^g_1}_t \subset F^{y,g}_{t-1} \subset F^{y,g}_t, \text{ by (14.22) and increasingness,}\\
&\quad y^g_1(t) \text{ is by (14.20) measurable with respect to } F^{y,g}_{t-1} \subseteq F^{y_2}_{t-1} \subseteq F^{y_2}_t,\\
&\quad \text{hence } y^g(t) \text{ is measurable with respect to } F^{y_2}_t,\ F^{y^g(t)} \subseteq F^{y_2}_t,\\
&\quad F^{y,g}_t = (F^{y^g(t)} \vee F^{y,g}_{t-1}) \subset (F^{y_2}_t \vee F^{y_2}_{t-1}) = F^{y_2}_t = F^{y,0}_t.
\end{aligned}$$

From the Equations (14.30) and (14.31) follows that,

$$F^{y,g}_t \vee F^{u,g}_t = F^{y,g}_t = F^{y,0}_t = F^{y_2}_t,\ \forall\ t \in T. \tag{14.32}$$

(b) Consider the stochastic filtering problem for the Gaussian system

$$x_2(t+1) = A(t)x_2(t) + M(t)v(t), x_2(0) = x_0, \tag{14.33}$$

$$y_2(t) = C(t)x_2(t) + N(t)v(t), \tag{14.34}$$

with $x_0 \in G(m_{x_0}, Q_{x_0})$, $v(t) \in G(0,I)$. It follows from Theorem 8.3.2 that,

$$E[\exp(iw_x^T x_2(t)|F_{t-1}^{y_2}] = \exp(iw_x^T \hat{x}_2(t) - \frac{1}{2} w_x^T Q_f(t) w_x), \ \forall \, w_x \in \mathbb{R}^{n_x}, \ t \in T, \tag{14.35}$$

where $\hat{x}_2 : \Omega \times T \to \mathbb{R}^{n_x}$ is determined by the recursion,

$$\hat{x}_2(t+1) = A(t)\hat{x}_2(t) + K(t, Q_f(t))[y_2(t) - C(t)\hat{x}_2(t)], \ \hat{x}_2(0) = m_{x_0}, \tag{14.36}$$

$$\hat{x}_2(t) = E[x_2(t)|F_{t-1}^{y_2}], \tag{14.37}$$

and $Q_f : T \to \mathbb{R}^{n_x \times n_x}$ is determined by the recursion of Equation (14.9). Define,

$$\hat{x}(t) = x_1^g(t) + \hat{x}_2(t). \ \hat{x} : \Omega \times T \to \mathbb{R}^{n_x}. \text{ Then } \forall t \in T, \tag{14.38}$$

$$E[\exp(iw_x^T x(t))|F_{t-1}^{y,g}] \tag{14.39}$$

$$= E[\exp(iw_x^T x_2(t))|F_{t-1}^{y,g}] \exp(iw_x^T x_1^g(t)),$$

by the equations (14.20), (14.38), and (14.22),

$$= E[\exp(iw_x^T x_2(t))|F_{t-1}^{y_2}] \exp(iw_x^T x_1^g(t)), \text{ by Equation (14.32)},$$

$$= \exp(iw_x^T \hat{x}_2(t) - \frac{1}{2} w_x^T Q_f(t) w_x + iw_x^T x_1^g(t))), \tag{14.40}$$

$$= \exp(iw_x^T \hat{x}(t) - \frac{1}{2} w_x^T Q_f(t) w_x), \text{ by Equation (14.38)},$$

$$\begin{aligned} \hat{x}(t+1) &= x_1^g(t+1) + \hat{x}_2(t+1) \\ &= A(t)x_1^g(t) + B(t)u^g(t) + A(t)\hat{x}_2(t) + K(t, Q_f(t))[y_2^g(t) - C(t)\hat{x}_2(t)] \\ &= A(t)\hat{x}(t) + B(t)u^g(t) + K(t, Q_f(t))[y^g(t) - C(t)\hat{x}(t) - D(t)u^g(t)], \\ &\quad \hat{x}(0) = m_{x_0}. \end{aligned}$$

(c) This follows directly from (b) because $\hat{x}^g(t)$ is $F_{t-1}^{y,g} \vee F_{t-1}^{u,g}$ measureable for all $t \in T$.
(d) By Equation (14.20),

$$\bar{v}(t) = y(t) - C(t)\hat{x}^g(t) - D(t)u^g(t) = y_2(t) - C(t)\hat{x}_2(t).$$

The result then follows from the theorem for the Kalman filter and from (b).
(e) This follows directly from (d).
(f) Note that by (b) and (c)

$$\hat{x}^g(t+1) = A(t)\hat{x}^g(t) + B(t)u^g(t) + K(t)\bar{v}(t),$$

and by (c) that $\bar{v}(t) \in G(0, Q_{\bar{v}})$ is independent of $F_{t-1}^{y,g} \vee F_{t-1}^{u,g}$. Then,

$$
\begin{aligned}
&E[\exp(iw_x^T \hat{x}^g(t+1))|\, F_{t-1}^{y,g} \vee F_{t-1}^{u,g}] \\
&\quad = \exp(iw_x^T[A(t)\hat{x}^g(t)+B(t)u^g(t)])E[\exp(iw_x^T K(t,Q_f(t)))\overline{v}(t))|F_{t-1}^{y,g} \vee F_{t-1}^{u,g}], \\
&\qquad \text{because } \hat{x}^g(t) \text{ and } u^g(t) \text{ are } F_{t-1}^{y,g} \vee F_{t-1}^{u,g} \text{ measurable,} \\
&\quad = \exp(iw_x^T[A(t)\hat{x}^g(t)+B(t)u^g(t)])\exp(-\frac{1}{2}w_x^T Q_{K\overline{v}}(t)w_x),
\end{aligned}
$$

because $\overline{v}(t)$ is independent of the σ-algebra $F_{t-1}^{y,g} \vee F_{t-1}^{u,g}$ and $\overline{v}(t) \in G(0,Q_{\overline{v}}(t))$. Similarly,

$$
\begin{aligned}
&E[\exp(iw_x^T x(t+1))|F_{t-1}^{y,g} \vee F_{t-1}^{u,g}] \\
&\quad = E[E[\exp(iw_x^T x(t+1))|F_t^{y,g} \vee F_t^{u,g}]|F_{t-1}^{y,g} \vee F_{t-1}^{u,g}], \text{ by reconditioning,} \\
&\quad = E[\exp(iw_x^T \hat{x}^g(t+1) - \frac{1}{2}w_x^T Q_f(t+1)w_x)|F_{t-1}^{y,g} \vee F_{t-1}^{u,g}], \text{by equation (14.5),} \\
&\quad = \exp(iw_x^T[A(t)\hat{x}^g(t)+B(t)u^g(t)]) \times \\
&\qquad \times \exp(-\frac{1}{2}w_x^T[Q_f(t+1)+K(t,Q_f(t))Q_{\overline{v}}(t)K(t,Q_f(t)))^T]w_x), \\
&\quad = \exp(iw_x^T[A(t)\hat{x}^g(t)+B(t)u^g(t)]) - w_x^T Q_{f+}(t+1)w_x),
\end{aligned}
$$

by the above calculation. The alternate expression for Q_{f_+} follows from the Equations (14.15) and (14.9). □

Stochastic Realization by the Conditional Kalman Filter

The reader finds below the second approach to the construction of an observation-based stochastic control system for a Gaussian stochastic control system. It is based on the conditional Kalman filter, Theorem 8.9.2, which statement the reader is expected to have read.

Consider the optimal control problem for a Gaussian stochastic control system with partial observations, Problem 14.4.1. Next consider any input process u adapted to the information structure and hence satisfying the following conditions,

$$
\begin{aligned}
&\{u(t),\ F_{t-1}^y \vee F_{t-1}^u,\ t \in T\},\ u : \Omega \times T \to U = \mathbb{R}^{n_u}, \\
&\text{adapted } (\Leftrightarrow\ \forall\, t \in T,\ F^{u(t)} \subseteq F_{t-1}^y \vee F_{t-1}^u,\ u(0) \in U).
\end{aligned}
$$

This is a generalization compared to the approach earlier in this section. No control laws are used, only the measurability of the input process is restricted.

Theorem 14.4.3. *Consider Problem 14.4.1. Assume that the input process u is measurable with respect to the information structure as described directly above the theorem.*

The filtering problem of Problem 14.4.1 has the solution,

$$\begin{aligned}
& E[\exp(iw_x^T\, x(t))|\, F_{t-1}^y \vee F_{t-1}^u] = \exp(iw_x^T \hat{x}(t) - w_x^T Q_f(t) w_x/2), \\
& \forall\, w_x \in \mathbb{R}^{n_x},\ t \in T; \\
\hat{x}(t+1) &= A(t)\hat{x}(t) + B(t)u(t) + K(t, Q_f(t))[y(t) - C(t)\hat{x}(t) - D(t)u(t)], \\
& x(0) = m_{x_0}, \\
Q_f(t+1) &= f_{FR}(t, Q_f(t)),\ Q_f(0) = Q_{x_0}, \\
K(t, Q_f(t)) &= [A(t)Q_f(t)C(t)^T + M(t)N(t)^T][C(t)Q_f(t)C(t)^T + N(t)N(t)^T]^{-1}, \\
\bar{v}(t) &= y(t) - C(t)\hat{x}(t) - D(t)u(t),\ \forall\, w_y \in \mathbb{R}^{n_y},\ \forall\, t \in T, \\
& E[\exp(iw_y^T\, \bar{v}(t))|\, F_{t-1}^y \vee F_{t-1}^u] = \exp(-w_y^T Q_f(t) w_y/2).
\end{aligned}$$

Proof. The result follows from the conditional Kalman filter, Theorem 8.9.2, with the following comments. Define the processes,

$b : \Omega \times T \to \mathbb{R}^{n_x},\ \forall\, t \in T,\ b(t) = B(t)u(t),$

hence $b(0) = B(0)u(0) \in \mathbb{R}^{n_x}$ is not random;

$d : \Omega \times T \to \mathbb{R}^{n_y},\ \forall\, t \in T,\ d(t) = D(t)u(t),$

hence $d(0) = D(0)u(0) \in \mathbb{R}^{n_y}$ is not random;

$\{b(t),\ F_{t-1}^y \vee F_{t-1}^u,\ t \in T\}$, is adapted, $\{d(t),\ F_{t-1}^y \vee F_{t-1}^u,\ t \in T\}$, is adapted.

Then the assumptions of Theorem 8.9.2 are satisfied and the theorem statements follow. The result for the innovations process $\bar{v}$ follows from the results of the theorem similarly as in the proof of Theorem 14.4.2. □

The above result can be generalized to the case where the input enters the system in a nonlinear way, like in,

$$x(t+1) = A(t)x(t) + b(t, u(t)) + M(t)v(t),\ x(0) = x_0.$$

The approach of this subsection can be extended to other stochastic systems.

14.4.2 Dynamic Programming

Next consider the optimal stochastic control Problem 14.4.1. The solution method has been sketched in Section 14.2. According to Procedure 14.2.3, the projection of the cost rate and of the terminal cost are calculated next.

Proposition 14.4.4. *Consider the optimal stochastic control Problem 14.4.1 for a Gaussian stochastic control system. Recall the Kalman filter of stochastic control formulated in Theorem 14.4.2. Define the projection of the cost rate on the information structure of the control law, and the projection of the terminal cost on the information structure by the formulas,*

$$\overline{b} : T \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x \times n_x}_{pds} \times \mathbb{R}^{n_u} \to \mathbb{R}_+, \; \overline{b}_1 : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x \times n_x}_{pds} \to \mathbb{R}_+,$$

$$\begin{aligned}
\overline{b}(t, (\hat{x}_b, Q_b), u_b) &= \int b(t, w_x, u_b) \; f_G(dw_x; (\hat{x}_b, Q_b)) \\
\overline{b}_1((\hat{x}_b, Q_b)) &= \int b_1(w_x) f_G(dw_x; (\hat{x}_b, Q_b)), \\
&\quad f_G(.) = f(., (\hat{x}^g(t), Q_f(t)); \; x(t)|F^{y^g}_{t-1} \vee F^{u^g}_{t-1}).
\end{aligned}$$

Then

$$\begin{aligned}
&\forall \, g \in G, \; t \in T, \\
&\overline{b}(t, (\hat{x}^g(t), Q_f(t)), u^g(t)) = E[b(t, x^g(t), u^g(t))|F^{y^g}_{t-1} \vee F^{y^g}_{t-1}], \\
&\overline{b}_1((\hat{x}^g(t_1), Q_f(t_1))) = E[b_1(x^g(t_1))|F^{y^g}_{t_1-1} \vee F^{u^g}_{t-1}], \\
J(g) &= E\left[\sum_{s=0}^{t_1-1} \overline{b}(s, (\hat{x}^g(s), Q_f(s)), u^g(s)) + \overline{b}_1((\hat{x}^g(t_1), Q_f(t_1)))\right].
\end{aligned}$$

Proof. Note that, by definition of the information structure and of the control law g, the input $u^g(t)$ is measurable with respect to $F^{y^g}_{t-1} \vee F^{u^g}_{t-1}$. Then,

$$\begin{aligned}
&\forall \, g \in G, \; t \in T, \\
&E[b(t, x^g(t), u^g(t))|F^{y^g}_{t-1} \vee F^{y^g}_{t-1}] \\
= &\int b(t, w_x, u^g(t)) \; f_G(dw_x; (\hat{x}^g(t), Q_f(t))), \text{ by Theorem 14.4.2,} \\
= &\; \overline{b}(t, (\hat{x}^g(t), Q_f(t)), u^g(t)), \text{ by definition of } \overline{b}(t, ., .); \\
&E[b_1(x^g(t_1))|F^{y^g}_{t_1-1} \vee F^{u^g}_{t-1}] \\
= &\int b_1(w_x) f_G(dw_x; (\hat{x}^g(t_1), Q_f(t_1))) = \overline{b}_1(\hat{x}^g(t_1), Q_f(t_1)).
\end{aligned}$$

Note that,

$$\begin{aligned}
J(g) &= E[\sum_{s=0}^{t_1-1} b(s, x^g(s), u^g(s)) + b_1(x^g(t_1))] \\
&= \sum_{s=0}^{t_1} E[E[b(s, x^g(s), u^g(s))|F^{y^g}_{s-1} \vee F^{u^g}_{s-1}]] + E[E[b_1(x^g(t_1))|F^{y^g}_{t_1-1} \vee F^{u^g}_{t-1}]] \\
&= E[\sum_{s=0}^{t_1} \overline{b}(s, (\hat{x}^g(s), Q_f(s)), u^g(s)) + \overline{b}_1((\hat{x}^g(t_1), Q_f(t_1)))].
\end{aligned}$$

□

After the transformation one obtains a new stochastic control problem with as stochastic control system the filter system and as cost function the expression of the above proposition. This is formally stated as follows.

Problem 14.4.5. Consider the stochastic control system consisting of the filter system,

$$\begin{aligned}
&\forall\, g \in G,\\
\hat{x}^g(t+1) &= A(t)\hat{x}^g(t) + B(t)g_t(y^g(0:t-1), u^g(0:t-1)) + K(t, Q_f(t))\overline{v}^g(t)\\
&= f_{KF}(t, \hat{x}^g(t), Q_f(t), g_t(.,.), \overline{v}^g(t)),\ \ \hat{x}^g(0) = m_{x_0},\\
y^g(t) &= C(t)\hat{x}^g(t) + D(t)g_t(y^g(0:t-1), u^g(0:t-1)) + \overline{v}^g(t),\\
Q_f(t+1) &= f_{FR}(t, Q_f(t)),\ Q_f(0) = Q_{x_0},\\
&\overline{v}^g \text{ a Gaussian white noise process, } \overline{v}^g(t) \in G(0, Q_{\overline{v}}(t)),\\
Q_{\overline{v}}(t) &= C(t)Q_f(t)C(t)^T + N(t)N(t)^T.
\end{aligned}$$

The filter system can be regarded as a stochastic realization of the stochastic control system with respect to the information structure.

Consider the cost function projected on the information structure, see Proposition 14.4.4.(b),

$$J(g) = E\left[\sum_{s=0}^{t_1-1} \overline{b}(t, (\hat{x}^g(t), Q_f(t)), u^g(t)) + \overline{b}_1((\hat{x}^g(t_1), Q_f(t_1)))\right].$$

The optimal stochastic control problem is to solve,

$$\inf_{g \in G} J(g).$$

Note that the above problem has now almost become a stochastic control problem with complete observations because the state of the filter system $(\hat{x}^g(t), Q_f(t))$ is measurable with respect to the information structure and the cost rate is a function only of the state of the filter system and of the input. However, in the above problem the control law can only depend on the information structure while in the complete observations case the control law depends on the state $(\hat{x}^g, Q_f)$ of the filter system only.

Notation is recalled and introduced for a Gaussian probability distribution function f_G on $\mathbb{R}^n$ with mean $m \in \mathbb{R}^n$ and variance $Q \in \mathbb{R}^{n\times n}_{pds}$,

$$\begin{aligned}
&f_G : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{n\times n}_{pds} \to \mathbb{R}_+,\ \ p_G(w;\ m, Q);\\
&Q \in \mathbb{R}^{n\times n}_{spds}\ \Rightarrow\\
p_G(dw;\ m, Q) &= (2\pi \det(Q))^{-1/2} \exp(-(w-m)^T Q^{-1}(w-m)/2),\\
f_G(dw;\ m, Q) &= p_G(w;\ m, Q)dw.
\end{aligned}$$

Below use is made of the Gaussian conditional probability distribution $f_G(.;\hat{x}, Q_f)$ in terms of $(\hat{x}, Q_f)$ as defined as in Theorem 14.4.2.

Denote the *conditional cost-to-go* and its projection on the information structure by respectively,

$$\begin{aligned}
&J : G \times T \to \mathbb{R}_+,\\
J(g,t) &= E\left[\sum_{s=t}^{t_1-1} \overline{b}(s, (\hat{x}^g(s), Q_f(s)), u^g(s)) + \overline{b}_1((\hat{x}^g(t_1), Q_f(t_1)))|F^{y,g}_{t-1} \vee F^{u,g}_{t-1}\right].
\end{aligned}$$

That the conditional cost-to-go of the original problem indeed equals the latter expression can be proven as in Proposition 14.4.4.

Below the reader has to distinguish between the arguments (x_V, Q_V) of the value function V and the particular values of the stochastic processes $(\hat{x}^g, Q_f)$ produced by the Kalman filter and the Riccati recursion, see the equations (14.7, 14.9).

Procedure 14.4.6 *The* dynamic programming procedure for a partially observed Gaussian stochastic control system in terms of the associated filter system representation.

Consider Problem 14.4.1 and Theorem 14.4.2 with the conditional distribution and the equations of the Kalman filter and the filter Riccati recursion, (14.7, 14.9) Consider the projection of the cost function of Proposition 14.4.4.

1. *Initialization. Define,*

$$V(t_1,.) : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x \times n_x}_{pds} \rightarrow \mathbb{R}_+,$$

$$V(t_1,(\hat{x}_V,Q_V)) = \overline{b}_1((\hat{x}_V,Q_V)).$$

2. *For* $t = t_1 - 1,\ t_1 - 2,\ \ldots,\ 0$ *do: Determine the function* V *at the indicated times,*

$$\begin{aligned}
& V(t,.) : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x \times n_x}_{pds} \rightarrow \mathbb{R}_+, \\
& V(t,\ (\hat{x}_V,\ Q_V)) \qquad (14.41) \\
& = \inf_{u_V \in U(t,(\hat{x}_V,Q_V))} \left\{ \overline{b}(t,(\hat{x}_V,Q_V),u_V) + \right. \\
& \quad \left. + E[V(t+1,\ (\hat{x}(t+1),\ Q_f(t+1)))|\ F^y_{t-1} \vee F^u_{t-1}] \right\}. \\
& = \inf_{u_V \in U(t,(\hat{x}_V,Q_V))} \left\{ \int b(t,w,u_V)\ f_G(dw;x(t)|\ (\hat{x}(t),\ Q_f(t))) + \right. \\
& \quad \left. + \int V(t+1,(\overline{w}_x, f_{FR}(t,Q_V)))\ f_G(d\overline{w}_x;[A(t)\hat{x}_V + B(t)u_V],Q_{K\overline{v}}(t,Q_V)) \right\}.
\end{aligned}$$

The last equality follows from Theorem 14.4.2(b) and (g) and equation (14.16). If, for all $(t,(x_V,Q_V))$*, the infimum in Equation (14.41) is attained, say there exists an element* $u^*_V \in U(t,(x_V,Q_V))$ *according to,*

$$\begin{aligned}
& \overline{b}(t,(\hat{x}_V,Q_V),u^*_V) + \\
& + \int V(t+1,(\overline{w}_x, f_{FR}(t,Q_V)))\ f_G(d\overline{w}_x;[A(t)\hat{x}_V + B(t)u^*_V],Q_{K\overline{v}}(t)) \\
& = \inf_{u_V \in U(t,(\hat{x}_V,Q_V))} \left\{ \overline{b}(t,(\hat{x}_V,Q_V),u_V) + \right. \qquad (14.42) \\
& \quad \left. + \int V(t+1,(\overline{w}_x, f_{FR}(t,Q_V)))\ f_G(d\overline{w}_x;[A(t)\hat{x}_V + B(t)u_V],Q_{K\overline{v}}(t)) \right\};
\end{aligned}$$

then define, $g^*_{V,t}(\hat{x}_V,Q_V) = u^*_V,\ \ g^*_{V,t}(.) : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x \times n_x}_{pds} \rightarrow \mathbb{R}^{n_u}.$

3. *If the functions*

$$\{g^*_{V,t} : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x \times n_x}_{spd} \rightarrow U,\ \forall\, t \in T(1:t-1)\}$$

are all measurable functions then proceed else stop.

4. *Define and output the optimal dynamic control law and the information system as,*

$$\begin{aligned}
& g_t^* : \mathbb{R}^{n_y(t+1)} \times \mathbb{R}^{n_u(t+1)} \to \mathbb{R}^{n_u}, \\
& g_t^*(y^{g^*}(0:t-1), u^{g^*}(0:t-1)) = g^*_{V,t}(\hat{x}^{g^*}(t), Q_f(t)), \\
Q_f(t+1) &= f_{FR}(t, Q_f(t)),\ Q_f(0) = Q_{x_0}, \\
\hat{x}^{g^*}(t+1) &= f_{KF}(t, \hat{x}^{g^*}(t), Q_f(t), g^*_{V,t}(t, (\hat{x}^{g^*}(t), Q_f(t)), \overline{v}(t)) \\
&= A(t)\hat{x}^{g^*}(t) + B(t) g^*_{V,t}((\hat{x}^{g^*}(t), Q_f(t))) + \\
&\quad + K(t, Q_f(t))[y(t) - C(t)\hat{x}^{g^*}(t) - D(t) g^*_{V,t}(t, (\hat{x}^{g^*}(t), Q_f(t)))], \\
&\quad \hat{x}(0) = m_{x_0}, \\
u^{g^*}(t) &= g^*_{V,t}(\hat{x}^{g^*}(t), Q_f(t)).
\end{aligned}$$

Note that the expression in brackets in the right-hand side of Equation (14.41) is a function of (t, u_V, x_V, Q_V) only. The infimization of that expression strikes a balance between the instantenous conditional cost rate and the cost on the remaining future horizon conditioned on the current information structure. The optimal control law $g^*_{V,t}$ will in general for every $t \in T(0:t_1-1)$, be a function of the tuple (x_V, Q_V). Note further that the function $g^*_{V,t}$, neither depends explicitly on the past outputs strictly before time $t \in T$ nor depends on the past inputs strictly before time $t \in T$.

Theorem 14.4.7. *Consider Problem 14.4.5. Consider the Kalman filter of stochastic control specified in Theorem 14.4.2. Let $V : T \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x \times n_x}_{pds} \to \mathbb{R}_+$ be produced by the Dynamic Programming Procedure 14.4.6.*

(a)Then, $\forall\ g \in G$ and $\forall\ t \in T$,

$$V(t, (\hat{x}^g(t), Q_f(t))) \leq J(g,t) \ a.s.\ , \tag{14.43}$$

$$E[V(0, (\hat{x}^g(0), Q_f(0)))] \leq E[J(g,0)] = J(g). \tag{14.44}$$

Note that the above statements only hold for the value function V at the specified values $(\hat{x}^g, Q_f)$.

(b)Assume that for any time $t \in T$ there exists a measurable function $g^*_{V,t}$ which attains the infima in Equation (14.42). Then $g^* \in G$ is an optimal control law, and,

$$V(t, (\hat{x}^{g^*}(t), Q_f(t))) = J(g^*, t) \ a.s.\ \ \forall\ t \in T, \tag{14.45}$$

$$J^* = E[V(0, (\hat{x}(0), Q_f(0)))] = J(g^*),$$

$$\inf_{g \in G} J(g) = \inf_{g \in G_D} J(g). \tag{14.46}$$

The proof of the theorem is based on the following two lemmas. The first lemma is for all $u_V \in U$ while the second is for a Markov control law $g \in G_M$.

Lemma 14.4.8. The comparison principle. *Let,*

$$\begin{aligned}
& V : T \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x \times n_x}_{pds} \to \mathbb{R}, \textit{ be such that,} \\
V(t_1,(\hat{x}_V,Q_V)) \le & \int b_1(w_x) f_G(dw_x;(\hat{x}_V,Q_V)), \ \forall\ (\hat{x}_V,Q_V) \in X \times \mathbb{R}^{n_x \times n_x}_{pds}; \\
V(t,(\hat{x}_V,Q_V))) \le & \int\ b(t,w_x,u_V) f_G(dw_x;(\hat{x}_V,Q_V)) + \\
& + \int\ V(t+1,(\overline{w}_x, f_{FR}(Q_V))) \times \\
& \times f_G(d\overline{w}_x;([A(t)\hat{x}_V + B(t)u_V], Q_{K\overline{v}}(t))), \\
& \forall\, t = t_1-1,\ t_1-2,\ \ldots, 0,\ (x_V,Q_V),\ \textit{and } u_V \in U;\ \textit{then} \\
V(t,(\hat{x}^g(t),Q_f(t))) \le & \ J(g,t) \textit{ a.s. } \forall g \in G,\ t \in T.
\end{aligned} \tag{14.47}$$

Proof. Analogous to Lemma 12.6.5 and to Lemma12.6.6. □

Lemma 14.4.9. *Let $g \in G_M$ be a Markov control law in terms of the state of the filter system $(\hat{x}^g, Q_f)$, hence $g_t : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x \times n_x}_{pds} \to U$. Define recursively the function,*

$$\begin{aligned}
& V^g : T \times X \times \mathbb{R}^{n_x \times n_x}_{pds} \to \mathbb{R}_+, \\
V^g(t_1,(\hat{x}_V,Q_V)) = & \int b_1(w_x) f_G(dw_x;(\hat{x}_V,Q_V)), \\
& \textit{and for } t = t_1-1,\ldots,0, \\
V^g(t,(\hat{x}_V,Q_V)) = & \int\ b(t,w_x,g_t(\hat{x}_V,Q_V))\ f_G(dw_x;(\hat{x}_V,Q_V))) \\
& + \int V^g(t+1,(\overline{w}_x, f_{FR}(Q_V))) \times \\
& \times f_G(d\overline{w}_x;[A(t)\hat{x}_V + B(t)g(t,(x_V,Q_V)))], Q_{K\overline{v}}(t));
\end{aligned} \tag{14.48}$$

then,

$$\begin{aligned}
& V^g(t,(\hat{x}^g(t),Q_f(t))) = J(g,t) \qquad (14.49) \\
& = E\left[\sum_{s=t}^{t_1-1} \overline{b}(s,(\hat{x}^g(s),Q_f(s)), g(s,(\hat{x}(s),Q_f(s)))) + \right. \\
& \qquad \left. + \overline{b}_1((\hat{x}^g(t_1),Q_f(t_1)))|\ F^{y,g}_{t-1} \vee F^{u,g}_{t-1} \right],\ \forall\, t \in T.
\end{aligned}$$

Proof. The proof is analogous to that of Lemma 12.6.6. From 14.4.2 follows that

$$f_G(.;\hat{x}^g(t_1)|F^{y,g}_{t_1-1} \vee F^{u,g}_{t_1-1}) = f_G(.,(\hat{x}^g(t_1),Q_f(t_1));\ \hat{x}^g(t_1)|F^{y,g}_{t_1-1} \vee F^{u,g}_{t_1-1}).$$

Then

$$\begin{aligned}
V^g(t_1,(\hat{x}^g(t_1),Q_f(t_1)) = & \int \overline{b}_1(w_x) f_G(dw_x;(\hat{x}^g(t_1),Q_f(t_1))) \\
= & \ E[\overline{b}_1(\hat{x}(t_1))|F^{y,g}_{t_1-1} \vee F^{u,g}_{t_1-1}] = J(g,t_1).
\end{aligned}$$

Suppose that Equation (14.49) holds for $t_1, t_1-1, \ldots, t+1$. Then

$$\begin{aligned}
& V^g(t,(\hat{x}^g(t),Q_f(t))) \\
& = \int \overline{b}(t,(w_x,Q_f(t)),g(t,(\hat{x}^g(t),Q_f(t))))f_G(dw_x;\hat{x}^g(t),Q_f(t)) \\
& \quad + \int V^g(t+1,(\overline{w}_x,f_{FR}(t,Q_f(t)))) \times \\
& \quad \times f_G(d\overline{w}_x;A(t)\hat{x}^g(t)+B(t)g(t,(\hat{x}^g(t),Q_f(t))),Q_{K\overline{v}}(t,Q_f(t)))) \\
& \quad \text{by equation (14.48),} \\
& = E[\overline{b}(t,(\hat{x}^g(t),Q_f(t)),g(t,(\hat{x}^g(t),Q_f(t)))|F^{y,g}_{t-1} \vee F^{u,g}_{t-1}]+ \\
& \quad +E[V^g(t+1,(\hat{x}^g(t+1),Q_f(t+1)))|F^{y,g}_{t-1} \vee F^{u,g}_{t-1}], \\
& \quad \text{by Theorem 14.4.2.(b) and (d), and, } Q_f(t+1)=f_{FR}(t,Q_f(t)), \\
& = E[\overline{b}(t,(\hat{x}^g(t),Q_f(t)),u^g(t))|\, F^{y,g}_{t-1} \vee F^{u,g}_{t-1}] \\
& \quad +E\left[E\left[\sum_{s=t+1}^{t_1-1} \overline{b}(s,(\hat{x}^g(s),Q_f(s)),u^g(s))+ \right.\right. \\
& \qquad \left.\left. +\overline{b}_1((\hat{x}^g(t_1),Q_f(t_1)))|\, F^{y,g}_t \vee F^{u,g}_t\right] |F^{y,g}_{t-1} \vee F^{u,g}_{t-1}\right], \\
& \quad \text{by the induction hypothesis for } s=t+1, \\
& = E[\sum_{s=t}^{t_1-1} \overline{b}(s,(x^g(s),Q_f(s)),u^g(s))+b_1((x(t_1),Q_f(t_1)))|F^{y,g}_{t-1} \vee F^{u,g}_{t-1}] = J(g,t),
\end{aligned}$$

by reconditioning and addition of conditional expectation. □

Proof. Of Theorem 14.4.7. This follows from Lemma 14.4.9 and Lemma 14.4.8 as Theorem 12.6.4 follows from Lemma 12.6.6 and Lemma 12.6.5. □

14.4.3 Quadratic Cost Rate

Problem 14.4.10. The *optimal stochastic control problem for a partially-observed Gaussian stochastic control system and with a quadratic cost function* (LQG-PO-FH). Consider a Gaussian stochastic control system representation,

$$\begin{aligned}
x(t+1) &= A(t)x(t)+B(t)u(t)+M(t)v(t), x(0)=x_0, \\
y(t) &= C(t)x(t)+D(t)u(t)+N(t)v(t), \\
z(t) &= C_z(t)x(t)+D_z(t)u(t),\ \forall\, t \in T(0:t_1-1), \\
z(t_1) &= C_z(t_1)x(t_1), \\
& \quad v(t) \in G(0,I),\ x_0 \in G(m_{x_0},Q_{x_0}),\ \ T(0:t_1)=\{0,1,\ldots,t_1\}, \\
& \quad n_y \leq n_v,\ \forall\, t \in T,\ \ \text{rank}(N(t))=n_y \ \Rightarrow\ N(t)N(t)^T \succ 0, \\
& \quad n_z \geq n_u,\ \forall\, t \in T,\ \ \text{rank}(D_z(t))=n_u \ \Rightarrow\ D_z(t)^T D_z(t) \succ 0.
\end{aligned}$$

Consider the information structure and the set of control laws,

$$\begin{aligned}
& \{F^y_{t-1} \vee F^u_{t-1},\ t \in T\}, \\
& \forall\, g \in G,\ g=\{g_0,g_1,\ldots,g_{t_1-1}\},\ g_0 \in U,\ g_t: Y^t \times U^t \to U.
\end{aligned}$$

Define the quadratic cost function,

$$J(g) = E\left[\left(\sum_{s=0}^{t_1-1} z^g(s)^T z^g(s)\right) + z^g(t_1)^T z^g(t_1)\right],\ J : G \to \mathbb{R}_+.$$

The problem is to solve the optimal stochastic control problem,

$$\inf_{g \in G} J(g).$$

Definition 14.4.11. Define the *LQG-PO-FH optimal control law* or the *optimal control law for a Gaussian stochastic control system with partial observations on a finite horizon with a quadratic cost rate* by the sets and functions,

$$\begin{aligned}
&Q_f : T \to \mathbb{R}^{n_x \times n_x}_{pds},\ Q_c : T \to \mathbb{R}^{n_x \times n_x}_{pds},\\
&K : T(0:t_1-1) \times \mathbb{R}^{n_x \times n_x}_{pds} \to \mathbb{R}^{n_x \times n_y},\\
&F : T(0:t_1-1) \times \mathbb{R}^{n_x \times n_x}_{pds} \to \mathbb{R}^{n_u \times n_x},
\end{aligned}$$

$$Q_f(t+1) = f_{FR}(t, Q_f(t)),\ \ Q_f(0) = Q_{x_0},\ \text{see equation (14.7)}, \tag{14.50}$$

$$\begin{aligned}
&K(t, Q_f(t)) \\
&= [A(t)Q_f(t)C(t)^T + M(t)N(t)^T][C(t)Q_f(t)C(t)^T + N(t)N(t)^T]^{-1},\\
&\quad \text{see equation (14.10)},
\end{aligned} \tag{14.51}$$

$$Q_c(t_1) = C_z(t_1)^T C_z(t_1), \tag{14.52}$$

$$\begin{aligned}
&Q_c(t) \\
&= A(t)^T Q_c(t+1)A(t) + C_z(t)^T C_z(t) + \\
&\quad -[A(t)^T Q_c(t+1)B(t) + C_z(t)^T D_z(t)] \times \\
&\quad \times [B(t)^T Q_c(t+1)B(t) + D_z(t)^T D_z(t)]^{-1} \times \\
&\quad \times [A(t)^T Q_c(t+1)B(t) + C_z(t)^T D_z(t)]^T,\ \forall\, t \in T(0:t_1-1);
\end{aligned} \tag{14.53}$$

$$\begin{aligned}
&F(t, Q_c(t+1)) \\
&= -[B(t)^T Q_c(t+1)B(t) + D_z(t)^T D_z(t)]^{-1} \times \\
&\quad \times [A(t)^T Q_c(t+1)B(t) + C_z(t)^T D_z(t)]^T,
\end{aligned} \tag{14.54}$$

$$g^*_t(\hat{x}_V, Q_V) = F(t, Q_V)\hat{x}_V,\ g^*_t : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x \times n_x}_{pds} \to \mathbb{R}^{n_u}, \tag{14.55}$$

$$\begin{aligned}
&\hat{x}^{g^*}(t+1) = A(t)\hat{x}^{g^*}(t) + B(t)F(t, Q_c(t+1))\hat{x}^{g^*}(t) + \\
&\quad + K(t, Q_f(t))[y^{g^*}(t) - C(t)\hat{x}^{g^*}(t) - D(t)F(t, Q_c(t+1))\hat{x}^{g^*}(t)],
\end{aligned} \tag{14.56}$$

$$\begin{aligned}
&= [A(t) + B(t)F(t, Q_c(t+1)) - K(t, Q_f(t))C(t) + \\
&\quad -K(t, Q_f(t))D(t)F(t, Q_c(t+1))]\hat{x}^{g^*}(t) + K(t, Q_f(t)))y^{g^*}(t),
\end{aligned} \tag{14.57}$$

$$\hat{x}^{g^*}(0) = m_{x_0},$$

$$u^{g^*}(t) = F(t, Q_c(t+1))\, \hat{x}^{g^*}(t), \tag{14.58}$$

$$g^*_{LQG,PO,FH}(t, \hat{x}_V, Q_V) = F(t, Q_V)\, \hat{x}_V. \tag{14.59}$$

Define the function of the *Control Riccati Recursion* as,

$$
\begin{aligned}
f_{CR} &: T \times \mathbb{R}_{pds}^{n_x \times n_x} \rightarrow \mathbb{R}^{n_x \times n_x}, \\
f_{CR}(t,Q) &= A(t)^T Q A(t) + C_z(t)^T C_z(t) + \\
&\quad -[A(t)^T Q B(t) + C_z(t)^T D_z(t)][B(t)^T Q B(t) + D_z(t)^T D_z(t)]^{-1} \times \\
&\quad \times [A(t)^T Q B(t) + C_z(t)^T D_z(t)]^T; \qquad (14.60) \\
Q_c(t) &= f_{CR}(t, Q_c(t+1)),\ Q_c(t_1) = C_z(t_1)^T C_z(t_1). \qquad (14.61)
\end{aligned}
$$

The optimal control law is a linear function of the state of the observation-based Gaussian stochastic control system representation.

Theorem 14.4.12. *Consider the optimal stochastic control Problem 14.4.10.*

(a)The value function and the value are given by,

$$
\begin{aligned}
& V(t,(\hat{x}_V, Q_V)) = \hat{x}_V^T Q_c(t) \hat{x}_V + r(t, Q_V, Q_c(t)), \qquad (14.62) \\
& \quad r : T \times \mathbb{R}_{pds}^{n_x \times n_x} \times \mathbb{R}_{pds}^{n_x \times n_x} \rightarrow \mathbb{R}, \\
& r(t_1, Q_V, Q_c(t_1)) = \mathrm{tr}(Q_v\ C_z(t_1)^T C_z(t_1)), \qquad (14.63) \\
& r(t, Q_V, Q_c(t)) = r(t+1, f_{FR}(t, Q_V), Q_c(t+1)) + \qquad (14.64) \\
& \quad + \mathrm{tr}\left(Q_c(t+1)\ Q_{K\overline{v}}(t, Q_f(t))\right) + \mathrm{tr}(Q_V\ C_z(t)^T C_z(t)), \\
& Q_{K\overline{v}}(t, Q_f(t)) \\
& = K(t, Q_f(t))\ [C(t) Q_f(t) C(t)^T + N(t) N(t)^T]\ K(t, Q_f(t))^T, \qquad (14.65) \\
J^* & = E[V(0,(x_0, Q_{x_0}))] = E[x_0^T Q_c(0) x_0] + r(0, Q_f(0), Q_c(1)) \qquad (14.66) \\
& = m_{x_0}^T Q_c(0) m_{x_0} + \mathrm{tr}(Q_c(0)\ Q_{x_0}) + \\
& \quad + \sum_{s=0}^{t_1-1} \mathrm{tr}(Q_c(s+1) Q_{f^+}(Q_f(s))) + \sum_{s=0}^{t_1} \mathrm{tr}(Q_f(s) C_z(s)^T C_z(s)).
\end{aligned}
$$

(b)The optimal control law is a linear function of the available observations. The optimal control law is optimal over the set of nonlinear measurable control laws.

(c)The optimal control law (LQG-PO-FH) consists of: the information system with the equations (14.56,14.50), the control Riccati recursion (14.53), and the feedback law (14.55). It is a dynamic control law.

Proof. The dynamic programming procedure is applied. It will be proven that the function V of equation (14.62) is a solution of the dynamic programming equation and that J^* of equation (14.66) is the associated value.

Define the function,

$$
\begin{aligned}
H &: T \rightarrow \mathbb{R}_{pds}^{(n_x+n_u) \times (n_x+n_u)}, \\
H_{xx}(t) &= A(t)^T Q_c(t+1) A(t) + C_z(t)^T C_z(t) \in \mathbb{R}^{n_x \times n_x}, \\
H_{xu}(t) &= A(t)^T Q_c(t+1) B(t) + C_z(t)^T D_z(t) \in \mathbb{R}^{n_x \times n_u}, \\
H_{uu}(t) &= B(t)^T Q_c(t+1) B(t) + D_z(t)^T D_z(t) \in \mathbb{R}^{n_u \times n_u}, \\
H(t) &= \begin{pmatrix} H_{xx}(t) & H_{xu}(t) \\ H_{xu}(t)^T & H_{uu}(t) \end{pmatrix}.
\end{aligned}
$$

From Theorem 14.4.2 follows that, for any control law $g \in G$, the conditional distribution of $x(t_1)$ given $F^{y,g}_{t_1-1} \vee F^{u,g}_{t_1-1}$ is Gaussian and denoted by $f_G(.;\hat{x}^g(t_1), Q_f(t_1))$. Hence,

$$\begin{aligned} V(t_1,(\hat{x}_V,Q_V)) &= E[x(t_1)^T C_z(t_1)^T C_z(t_1) x(t_1) | F^y_{t_1-1}] \\ &= \int w_x^T C_z(t_1)^T C_z(t_1) w_x \; f_G(dw_x;(x_V,Q_V)) \\ &= \hat{x}_V^T C_z(t_1)^T C_z(t_1) \hat{x}_V + \text{tr}(C_z(t_1)^T C_z(t_1) \; Q_V), \end{aligned}$$

by Proposition 2.7.6. Define,

$$\begin{aligned} Q_c(t_1) &= C_z(t_1)^T C_z(t_1), \; r(t_1,Q_V,Q_c(t_1)) = \text{tr}(Q_c(t_1) \; Q_V); \text{ then,} \\ Q_c(t_1) &= Q_c(t_1)^T \succeq 0, \text{ and,} \\ V(t_1,(\hat{x}_V,Q_V))) &= \hat{x}_V^T Q_c(t_1) \hat{x}_V + r(t_1,Q_V,Q_c(t_1)). \end{aligned}$$

Suppose that for $s = t_1, \; t_1-1, \; t_1-2, \; \ldots, \; t+1$,

$$Q_c(s) = Q_c(s)^T \succeq 0, \text{ and,} \tag{14.67}$$
$$V(s,\hat{x}_V,Q_V)) = \hat{x}_V^T Q_c(s) \hat{x}_V + r(s,Q_V,Q_c(s)). \tag{14.68}$$

It will be proven that the equations (14.67,14.68) hold for $s = t$. By Theorem 14.4.2 the conditional distribution of $x(t)$ conditioned on $F^{y,g}_{t-1} \vee F^{u,g}_{t-1}$ is Gaussian and denoted by $f_G(.;\hat{x}^g(t), Q_f(t))$. Denote,

$$\begin{aligned} Q_{cr}(t) &= \begin{pmatrix} C_z(t)^T C_z(t) & C_z(t)^T D_z(t) \\ D_z(t)^T C_z(t) & D_z(t)^T D_z(t) \end{pmatrix}, \\ Q_{f_+}(t) &= A(t) Q_f(t) A(t)^T + M(t) M(t)^T. \end{aligned}$$

Recall from Theorem 14.4.2.(f) that

$$\begin{aligned} &\text{cpdf}(x(t) | \; F^{y^g}_{t-1} \vee F^{u^g}_{t-1}) = G(\hat{x}(t), Q_f(t)), \\ &\text{cpdf}(x(t+1) | \; F^{y^g}_{t-1} \vee F^{u^g}_{t-1}) = G([A(t)\hat{x}^g(t) + B(t) u^g(t)], Q_{K\bar{v}}(t, Q_f(t))). \end{aligned}$$

Then,

$$
\begin{aligned}
&\int \begin{pmatrix} w_x \\ u_V \end{pmatrix}^T Q_{cr}(t) \begin{pmatrix} w_x \\ u_V \end{pmatrix} f_G(dw_x; \hat{x}_V, Q_V) \\
&\quad + \int V(t+1, (\overline{w}_x, f_{FR}(t, Q_V)))) \, f_G(d\overline{w}_x; [A(t)\hat{x}_V + B(t)u_V], Q_{K\overline{v}}(Q_f(t)))) \\
&= \int \begin{pmatrix} w_x \\ u_V \end{pmatrix}^T Q_{cr}(t) \begin{pmatrix} w_x \\ u_V \end{pmatrix} f_G(dw_x; \hat{x}_V, Q_V) \\
&\quad + \int [\overline{w}_x^T Q_c(t+1)\overline{w}_x + r(t+1, f_{FR}(t, Q_V), Q_c(t+1))] \times \\
&\quad \times f_G(d\overline{w}_x; [A(t)\hat{x}_V + B(t)u_V], Q_{K\overline{v}}(t)), \text{ by the induction hypothesis,} \\
&= \begin{pmatrix} \hat{x}_V \\ u_V \end{pmatrix}^T Q_{cr}(t) \begin{pmatrix} \hat{x}_V \\ u_V \end{pmatrix} + \mathrm{tr}(C_z(t)^T C_z(t)\ Q_V) + \\
&\quad + [A(t)\hat{x}_V + B(t)u_V]^T Q_c(t+1)[A(t)\hat{x}_V + B(t)u_V] \\
&\quad + \mathrm{tr}(Q_c(t+1)Q_{K\overline{v}}(Q_f(t))) + r(t+1, f_{FR}(t, Q_V), Q_c(t+1)) \\
&= \begin{pmatrix} \hat{x}_V \\ u_V \end{pmatrix}^T H(t) \begin{pmatrix} \hat{x}_V \\ u_V \end{pmatrix} + \mathrm{tr}(C_z(t)^T C_z(t)\ Q_V) \\
&\quad + \mathrm{tr}(Q_c(t+1)Q_{K\overline{v}}(Q_f(t))) + r(t+1, f_{FR}(t, Q_V), Q_c(t+1)).
\end{aligned}
$$

Then also,

$$
\begin{aligned}
&\inf_{u_V \in U(\hat{x}_V, Q_V)} \left\{ \int \begin{pmatrix} w_x \\ u_V \end{pmatrix}^T Q_{cr}(t) \begin{pmatrix} w_x \\ u_V \end{pmatrix} f_G(dw_x; \hat{x}_V, Q_V) \right. \\
&\quad \left. + \int V(t+1, (\overline{w}_x, f_{FR}(t, Q_V))) f_G(d\overline{w}_x; [A(t)\hat{x}_V + B(t)u_V], Q_{K\overline{v}}(Q_f(t)))) \right\}
\end{aligned}
$$

$$
\begin{aligned}
&= \inf_{u_V \in U(\hat{x}_V, Q_V)} \begin{pmatrix} \hat{x}_V \\ u_V \end{pmatrix}^T H(t) \begin{pmatrix} \hat{x}_V \\ u_V \end{pmatrix} + \mathrm{tr}(C_z(t)^T C_z(t)\ Q_V) + \\
&\quad + \mathrm{tr}(Q_c(t+1)Q_{K\overline{v}}(Q_f(t))) + r(t+1, f_{FR}(t, Q_V), Q_c(t+1)), \\
&= \hat{x}_V^T Q_c(t)\hat{x}_V + r(t, Q_V, Q_c(t)), \text{ by completion of squares,} \\
u_V^* &= -H_{uu}(t)^{-1} H_{xu}(t)^T \hat{x}_V, \\
Q_c(t) &= H_{xx}(t) - H_{xu}(t) H_{uu}^{-1}(t) H_{xu}^T(t) = f_{CR}(t, Q_c(t+1)), \\
r(t, Q_V, Q_c(t)) &= r(t+1, f_{FR}(t, Q_V), Q_c(t+1)) + \\
&\quad + \mathrm{tr}(C_z(t)C_z(t)Q_V) + \mathrm{tr}(Q_c(t+1)Q_{K\overline{v}}(Q_f(t))), \\
H_{uu}(t) &= B(t)^T Q_c(t+1)B(t) + D_z(t)^T D_z(t) \succeq D_z(t)^T D_z(t) \succ 0;
\end{aligned}
$$

and where the infimum is attained for $u = u^*$. It follows from the above calculation that $Q_c(t) = Q_c(t)^T \succeq 0$. The result then follows from Theorem 14.4.7. □

14.4.4 Examples

Example 14.4.13. *Control of a mooring tanker*. Consider again the control problem of a mooring tanker, described in Example 1.1.1. There is a tanker moored at a single loading station at sea. The tanker is attached to the loading station by a stretcheable tube through which oil can be pumped from the station to the tanker. The tanker is to manouevre so as to keep the distance between the loading station and the point on the ship to which the hawser is attached, within specified limits. It should be clear that if there is locally a storm then the loading operation is temporarily stopped till quiet times have returned.

A model in the form of a stochastic control system was formulated. The system models both the ship dynamics and the dynamic behavior of the sea for the period in which the control is active. The control problem is to determine a control law which in closed-loop with the control system will meet the control objectives. The control objectives are (1) to guarantee exponential stability of the closed-loop system; and (2) to minimize a performance criterion which includes the distance between the loading station and the ship and the cost of fuel. The LQG optimal control law has been computed and the performance of the closed-loop system was extensively tested. The results are satisfactory to the investigators.

The original stochastic control system is in continuous-time. Using a well-known procedure, a discrete-time time-invariant Gaussian stochastic control system can be formulated whose behavior is close to that of the continuous-time system. Then the optimal control law of Theorem 14.4.12 can be applied, which is stated here again for the following discussion

$$\begin{aligned}
\hat{x}^{g^*}(t+1) &= [A(t)+B(t)F(t,Q_c(t+1)) - K(t,Q_f(t))C(t) \\
&\quad -K(t,Q_f(t))D(t)F(t,Q_c(t+1))]\hat{x}^{g^*}(t) + K(t,Q_f(t)))y^{g^*}(t), \\
&\quad \hat{x}^{g^*}(0) = m_{x_0}, \\
u^{g^*}(t) &= F(t,Q_c(t+1))\,\hat{x}^{g^*}(t).
\end{aligned}$$

Note that the above theorem deals with a time-varying Gaussian stochastic control system while the model of the mooring tanker is described by a time-invariant Gaussian stochastic control system. The optimal control law for a time-invariant Gaussian stochastic control system is derived in Chapter 15.

The control design requires attention for several design issues. The system matrices (A,B,C,D,M,N) are determined by the model chosen. The matrices of the controlled output $C_z,\ D_z$ are to be choosen by the control engineer. The choice of the authors of the paper is to select for C_z a matrix so that all state components of the ship appear in the controlled output but the state components of the sea model do not appear. The latter components are not controllable in the system. It makes sense to include in the controlled output all components of the input vector, so the matrix D_z can be chosen as a partly diagonal matrix. The relative contribution between the matrix C_z and D_z is a design issue that is best solved by evaluation of the performance of the closed-loop system in simulation for a range of values. A larger C_z while keeping D_z fixed results in better performance at a higher cost for the input.

A smaller C_z while keeping D_z fixed results in a poorer performance at a lower cost for the input. The trade-off between these choices is an engineering design issue.

The performance of the closed-loop system was checked by simulation. The investigators were satisfied with the performance. It is then to the company requesting the controller to select the actual matrices C_z and D_z depending on its preferences.

Example 14.4.14. *Control of a paper machine by LQG control law*. Consider again the example of control of a paper machine as described in Example 14.1.1.

The control objective is to reduce the variance of the dry paper weight. To this we now add also the control objective to reduce the variance of the input to the material flow. The latter control objective transforms the control problem which meets the conditions of the LQG control Problem 14.4.10. The condition that $\text{rank}(D_z) = n_u$ is then met. The control problem is therefore different from the minimum variance control problem considered in Example 14.1.1.

An optimal control law can then be computed along the lines of Theorem 14.4.12. Note that the optimal control law is then time-varying. The associated time-invariant optimal control law is derived in Chapter 15.

The closed-loop system is then asymptotically stable and minimizes the selected performance index.

The primary control design issue is to select the relative values of the matrices of the controlled output, C_z and D_z. It seems best to compute the performance of the closed-loop system and to evaluate simulations of the closed-loop system for various initial conditions and pseudo-random noise, before selecting values for those matrices and hence the optimal control law.

14.4.5 Predicted Miss

The following example shows that the information system need not be the conditional distribution and may depend on the control objective.

Problem 14.4.15. *Predicted miss as information state.* Consider a time-invariant Gaussian stochastic control system as defined in Problem 14.4.10,

$$\begin{aligned}
x(t+1) &= Ax(t) + Bu(t) + Mv(t), x(0) = x_0,\\
y(t) &= Cx(t) + Nv(t),\\
z &= Hx(t_1),\\
&\quad U = [-1,+1] \subset \mathbb{R}, n_u = 1,\ z : \Omega \to \mathbb{R}^{n_z},\ H \in \mathbb{R}_+^{n_z \times n_x},\\
&\quad v(t) \in G(0,I),\ x_0 \in G(m_{x_0}, Q_{x_0}),\ NN^T \succ 0.
\end{aligned}$$

Consider the past-output information structure, the corresponding set of control laws, and the closed-loop system. Define the cost function as,

$$J(g) = E[z^T z],\ \ J : G \to \mathbb{R}_+.$$

There is no cost rate, only a terminal cost.

The optimal stochastic control problem is to determine a control law g^* that is optimal for the defined cost function.

Proposition 14.4.16. *Consider Problem 14.4.15. As in Theorem 14.4.2 the Kalman filter for this problem is given, for $g \in G$, by*

$$\begin{aligned} \hat{x}^g(t+1) &= A\hat{x}^g(t) + Bu^g(t) + K(t)[y(t) - C\hat{x}^g(t)],\ \hat{x}^g(0) = m_{x_0}, \\ \overline{v}(t) &= y(t) - C\hat{x}^g(t). \end{aligned}$$

Define the predicted miss *process by,*

$$\begin{aligned} \hat{z}^g &: \Omega \times T \to \mathbb{R}^{n_z},\ H : T \to \mathbb{R}^{n_z \times n_x}, \\ \hat{z}^g(t) &= HA^{t_1 - t}\hat{x}(t) = H(t)\hat{x}(t); \\ H(t) &= HA^{t_1 - t},\ Q_{e_z}(t) = H(t)K(t)Q_{\overline{v}}(t)K(t)^T H(t)^T, \end{aligned}$$

and where $Q_{\overline{v}} : T \to \mathbb{R}_+$ is as defined in Equation (14.15).

Then the information system for the predicted miss is given by,

$$\hat{x}(t+1) = A\hat{x}(t) + Bu(t) + K(t)[y(t) - C\hat{x}(t)],\ \hat{x}(0) = m_{x_0}, \tag{14.69}$$

$$\hat{z}(t) = HA^{t_1 - t}\hat{x}(t), \tag{14.70}$$

$$p_g(.;z|F^{y,g}_{t_1-1}) = G(.;\hat{z}(t_1), HQ_f(t_1)H^T), \tag{14.71}$$

$$\begin{aligned} & f_G(.;\hat{z}^g(t+1)|F^{y,g}_{t-1} \vee F^{u,g}_{t-1}) \\ & = f_G(.;\hat{z}^g(t) + H(t)B(t)u^g(t), Q_{e_z}(t)). \end{aligned} \tag{14.72}$$

Proof. This follows from Theorem 14.4.2. □

Note that $\hat{x}$ is the information state of the information system of Equation (14.69). One may consider $\hat{z}$, the predicted miss, as a sufficient information state for the cost function.

A control law $g \in G$ for Problem 14.4.15 is a dynamic control law based on the information system described by the Equations (14.69) and (14.70), and on the explicit control law in terms of the state of the information system, $u(t) = h(t, \hat{z}(t))$, for a function $h : T \times \mathbb{R} \to U$. Denote by G_{IS} the class of separate control laws. Note that a separated control law does not directly depend on the information state but on $\hat{z}$, which in turn depends on the cost function.

Proposition 14.4.17. *Define the function $V : T \times \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$*

$$\begin{aligned} V(t_1, (x_V, q_V)) &= \int w^2 G(dw; x_V, HQ_f(t_1)H^T), \\ V(t, (x_V, q_V)) &= \inf_{u_V \in U} \int V(t+1, (w_V, q_V))G(dw; x_V + H(t)Bu_V, Q_{e_z}(t)) \end{aligned} \tag{14.73}$$

For $g \in G$ and define,

$$J(g,t) = E[z^2|F^{y,g}_{t-1} \vee F^{u,g}_{t-1}].$$

(a)Then for any $g \in G$ and $t \in T$

$$V(t_1, (x_V, q_V)) = x_V^2 + Hq_V H^T = x_V^2 + h^2 q_V, \tag{14.74}$$

$$V(t, (\hat{z}(t), q_f(t))) \leq J(g,t) \text{ a.s.} \tag{14.75}$$

(b)Assume that there exists a function $h^* : T \times \mathbb{R} \to U$ that attains the infima in Equation (14.73), or such that

$$V(t,(x,q_f)) = \int V(t+1,(w,q_f))\; f_G(dw; x_V + H(t)Bu^*, Q_{e_z}(t)), \tag{14.76}$$
$$u^* = h^*(t, x_V, q_V). \tag{14.77}$$

The optimal control law g^* then consists of,

$$\begin{aligned}
\hat{x}(t+1) &= A\hat{x}(t) + Bu(t) + K(t)[y(t) - C\hat{x}(t)], \hat{x}(0) = m_{x_0},\\
\hat{z}(t) &= HA^{t_1-t}\hat{x}(t),\\
u(t) &= h^*(t, \hat{z}(t), q_f(t)).\\
V(t,(\hat{z}(t),q_f(t)))) &= J(g^*,t) \text{ a.s.}\\
J^* &= J(g^*) = \inf_{g \in G} J(g),
\end{aligned}$$

and g^* is an optimal control law. Hence the optimal control is a dynamic control law based on the filter system.

Note that the value function (14.73) and the optimal control law (14.77) depend on the predicted miss, not on the information state.

Proof. This follows analogously to that of Theorem 14.4.7. □

Proposition 14.4.18. *Consider Problem 14.4.15. Let $r = 1$, $c = HB \in \mathbb{R}$, and assume that $c > 0$. Recall that $U = [-1,+1]$. Then*

$$\begin{aligned}
&g^*(t_1-1,x)\\
&= \begin{cases} 1, & x < -c,\\ -x/c, & -c \le x \le c,\\ -1, & c < x, \end{cases} \tag{14.78}\\
&V(t_1-1,\hat{z}(t_1-1))\\
&= \begin{cases} HQ_f(t_1)H^T + S(t_1-1) + (\hat{z}(t_1-1)+c)^2, & \text{if } \hat{z}(t_1-1) < -c,\\ HQ_f(t_1)H^T + S(t_1-1), & \text{if } -c \le \hat{z}(t_1-1) \le c,\\ HQ_f(t_1)H^T + S(t_1-1) + (\hat{z}(t_1-1)-c)^2, & \text{if } c < \hat{z}(t_1-1), \end{cases}
\end{aligned}$$

The mathematical formula of the optimal control law of equation (14.78) is summarized by the expression full bang to reduce expected miss. *The expression* full bang *denotes that the input takes either the value $+1$ or -1, the extreme values of the input set.*

Proof. A direct calculation. □

14.4.6 Cost Function is Exponential-Quadratic

Problem 14.4.19. Consider a Gaussian stochastic control system representation

$$\begin{aligned}
x(t+1) &= A(t)x(t)+B(t)u(t)+M(t)v(t), x(0)=x_0, \\
y(t) &= C(t)x(t)+D(t)u(t)+N(t)v(t), \\
z(t) &= C_z(t)x(t)+D_z(t)u(t), \ \forall\, t \in T(0:t_1-1), \\
z(t_1) &= C_z(t_1)x(t_1), \\
& \quad v(t) \in G(0,I), \ x_0 \in G(m_{x_0}, Q_{x_0}), \ T(0:t_1) = \{0,1,\ldots,t_1\}, \\
& \quad n_y \leq n_v, \ \forall\, t \in T, \ \text{rank}(N(t)) = n_y \ \Rightarrow N(t)N(t)^T] \succ 0, \\
& \quad n_z \leq n_u, \ \forall\, t \in T, \ \text{rank}(D_z(t)) = n_u \ \Rightarrow \ D_z(t)^T D_z(t)] \succ 0.
\end{aligned}$$

Consider further the past-output and past-input information structure,

$$\{F^y_{t-1} \vee F^u_{t-1}, \ t \in T\}$$

and the corresponding set G of control laws. For any $g \in G$, let $g = \{g_0, g_1, \ldots, g_{t_1-1}\}$, $g_0 \in U$, and $g_t : Y^t \times U^t \to U$. Define the exponential-of-quadratic cost function $J : G \to \mathbb{R}$ for $c \in \mathbb{R} \backslash \{0\}$,

$$J(g) = E\left[c \exp\left(\frac{1}{2}c \sum_{s=0}^{t_1-1} z^g(s)^T z^g(s) + z^g(t_1)^T z^g(t_1)\right)\right]. \tag{14.79}$$

The problem is to solve the optimal stochastic control problem,

$$\inf_{g \in G} J(g).$$

This problem is called the *partial observations linear-exponential-quadratic-Gaussian (LEQG) optimal stochastic control problem.* It is mainly known by its acronym of LEQG or by the name *risk-sensitive* stochastic control problem.

There is a relation between the solution of the above formulated problem and the solution of a particular robust control problem. The relation is due to that of the exponential-of-quadratic cost function and an entropy criterion in robust control.

The solution to this problem will not be presented here. See the section *Further Reading*, 14.7, for references.

14.4.7 A Tracking Problem

In control engineering a subset of problems deals with tracking. In a *tracking problem* a control systems has to be controlled such that the controlled output of the control system tracks or follows a prespecified reference signal. The prespecified reference signal is often assumed to be generated by a *reference system*. That the reference system is used rather than that the control system includes a model of the reference system, is due to the option that the control system may be used for several different reference systems. Therefore a separation between the control system and the reference system is useful.

An example of a tracking problem is when a ship has to follow a particular trajectory specified by the position and the speed of another ship, where the other ship may be the leader of a convoy where the leader escorts the follower ships. This can then also be applied to a set of airplanes. In chemical engineering, a tracking problem can be that the system of a chemical reactor which tracks the reference system which models the recipe for the production of the end product. In an information system a tracking problem can be that the control system follows a particular procedure which describes a sequence of operations.

The tracking problem for a time-invariant linear system was solved by B.A. Francis and W.M. Wonham, see [26]. See also the book, [25]. Later generalization to a control system in a differential geometric structure was developed.

In this section the tracking problem is formulated for a Gaussian stochastic control system. The reference system is assumed to be a Gaussian system, hence it is a stochastic system. The fact that the reference system is a stochastic system requires comments. In a deterministic reference system it is often assumed that the system is asymptotically stable hence the state converges to a set point which may be the zero state. In case the reference system is deterministic but nonlinear then the reference system can exhibit other behavior for example a periodic trajectory as the Van der Pol oscillator. If there is in control engineering a tracking problem where the reference signal is fluctuating around a nominal trajectory, like the trajectory of the leader ship, then it is realistic that the reference system is modelled as a stochastic system. Hence our choice to model the reference system as a stochastic system.

As first stated by B.A. Francis and W.M. Wonham, the control law for a tracking problem of a linear control system is a dynamic control law, where the dynamic control law best includes a copy of the reference system. This result also holds for the result of this section. That the dynamic control law includes a copy of the reference system follows directly from engineering modeling of the tracking problem.

Problem 14.4.20. *Tracking problem of a Gaussian control system.* Consider the tracking problem for a time-varying Gaussian stochastic control system. Assume that the reference system is a Gaussian stochastic system and that the associated control system is a Gaussian stochastic control system, with representations

$$\begin{aligned}
T &= \{0,1,\ldots,t_1\},\ (\Omega,F,P),\ n_{x_r},\ n_{v_r},\ n_{y_r},\ n_{z_r} \in \mathbb{Z}_+,\\
x_r(t+1) &= A_r(t)x_r(t) + M_r(t)v_r(t),\ x_r(0) = x_{r,0} \in G(0,Q_{x_{r,0}}),\\
y_r(t) &= C_r(t)x_r(t) + N_r(t)v_r(t),\ \ v_r(t) \in G(0,Q_{v_r}),\\
z_r(t) &= C_{z_r}(t)x_r(t),\ \forall\, t \in T,\ N_r(t)N_r(t)^T \succ 0;\\
&\qquad n_{x_s},\ n_{v_s},\ n_{y_s},\ n_{z_s} \in \mathbb{Z}_+,\ n_{z_s} = n_{z_r},\\
x_s(t+1) &= A_s(t)x_s(t) + B_s(t)u(t) + M_s(t)v_s(t),\ x_s(0) = x_{s,0} \in G(0,Q_{x_{s,0}}),\\
y_s(t) &= C_s(t)x_s(t) + D_s(t)u(t) + N_s(t)v_s(t),\ \ v_s(t) \in G(0,Q_{v_s}),\\
z_s(t) &= C_{z_s}(t)x_s(t) + D_{z_s}(t)u(t),\\
&\qquad \forall\, t \in T,\ N_s(t)N_s(t)^T \succ 0,\ D_z(t)^T D_z(t) \succ 0;\\
&\qquad F^{x_{r,0}},\ F^{x_{s,0}},\ F_{t_1}^{v_r},\ F_{t_1}^{v_s} \text{ are independent.}
\end{aligned}$$

Alternatively, the reference system could be described as a Gaussian system in the Kalman realization where the state is measureable with respect to the past outputs.

The *tracking error* is defined by the equation,

$$z(t) = z_r(t) - z_s(t),\ z : \Omega \times T \to \mathbb{R}^{n_z}.$$

The combined system is defined as the product of the reference system and the stochastic control system and it has the representation,

$$\begin{aligned}
x(t) &= \begin{pmatrix} x_r(t) \\ x_s(t) \end{pmatrix},\ y(t) = \begin{pmatrix} y_r(t) \\ y_s(t) \end{pmatrix},\ v(t) = \begin{pmatrix} v_r(t) \\ v_s(t) \end{pmatrix}, \\
x(t+1) &= A(t)x(t) + B(t)u(t) + M(t)v(t),\ x(0) = x_0, \\
y(t) &= C(t)x(t) + D(t)u(t) + N(t)v(t), \\
z(t) &= C_z(t)x(t) + D_z(t)u(t), \\
&\quad A(t) = \begin{pmatrix} A_r(t) & 0 \\ 0 & A_s(t) \end{pmatrix},\ B(t) = \begin{pmatrix} 0 \\ B_s(t) \end{pmatrix},\ M(t) = \begin{pmatrix} M_r(t) & 0 \\ 0 & M_s(t) \end{pmatrix}, \\
&\quad C(t) = \begin{pmatrix} C_r(t) & 0 \\ 0 & C_s(t) \end{pmatrix},\ D(t) = \begin{pmatrix} 0 \\ D_s(t) \end{pmatrix},\ N(t) = \begin{pmatrix} N_r(t) & 0 \\ 0 & N_s(t) \end{pmatrix}, \\
&\quad C_z = \begin{pmatrix} C_{z,r}(t) & -C_{z,s}(t)(t) \end{pmatrix},\ D_z(t) = -D_{z,s}(t), \\
&\quad \forall\, t \in T,\ N(t)N(t)^T \succ 0,\ D_z(t)^T D_z(t) = D_{z,s}(t)^T D_{z,s}(t) \succ 0.
\end{aligned}$$

Define the set of partially-observed control laws as in Problem 14.2.1 and the corresponding closed-loop system as also stated in that problem. The cost function is then

$$\begin{aligned}
J(g) &= E\left[\sum_{s=0}^{t_1-1} z^g(s)^T z^g(s) + z^g(t_1)^T z^g(t_1)\right],\ J : G \to \mathbb{R}_+, \\
&\inf_{g \in G} J(g).
\end{aligned}$$

Note that the combined system contains both the reference system and the stochastic control system.

The optimal control law follows directly from Theorem 14.4.12. However, note that for the tracking problem the combined system is not at all controllable though it can be required to be stabilizable. This is due to the fact that the reference system is not at all affected by the input signal. The solution to the tracking problem is summarized in the next theorem.

Theorem 14.4.21. *Consider the tracking problem of Problem 14.4.20.*

(a)Of the time-varying Kalman filter, both the filter error variance function and the Kalman gain function are block-diagonal

$$Q_f(t) = \begin{pmatrix} Q_{f,r}(t) & 0 \\ 0 & Q_{f,s}(t) \end{pmatrix},\ K(t, Q_f(t)) = \begin{pmatrix} K_r(t, Q_f(t)) & 0 \\ 0 & K_s(t, Q_f(t)) \end{pmatrix}.$$

Consequently the Kalman filter of the combined system decomposes as two unrelated time-varying Kalman filters for the reference system and for the stochastic control system

$$\begin{aligned}
\hat{x}(t) &= \begin{pmatrix} \hat{x}_r(t) \\ \hat{x}_s(t) \end{pmatrix}, \\
\hat{x}_r(t+1) &= A_r(t)\hat{x}_r(t) + K_r(t, Q_f(t))[y_r(t) - C_r(t)\hat{x}_r(t)],\ \hat{x}_r(0) = E[x_{r,0}], \\
\hat{x}_s(t+1) &= A_s(t)\hat{x}_s(t) + B_s(t)u(t) + \\
&\quad + K_s(t, Q_f(t))[y_s(t) - C_s(t)\hat{x}_s(t) - D_s(t)u(t)],\ \hat{x}_s(0) = E[x_{s,0}].
\end{aligned}$$

(b)The solution of the control Riccati recursion Q_c is a matrix function with in general four nonzero blocks. The optimal control law exists. In general, the feedback matrix function $F(Q_c(t+1))$ is a matrix with two nonzero blocks.

$$\begin{aligned}
Q_f(t+1) &= f_{FR}(t, Q_f(t)),\ Q_f(0) = Q_{x_0}, \\
Q_c(t) &= f_{CR}(t, Q_c(t+1)),\ Q_c(t_1) = C_z^T C_z, \\
u(t) &= F(t, Q_c(t+1))\hat{x}(t) \\
&= F_r(t, Q_c(t+1))\hat{x}_r(t) + F_s(t, Q_c(t+1))\hat{x}_s(t), \\
\hat{x}(t+1) &= [A(t) + B(t)F(t, Q_c(t+1)) - K(t, Q_f(t))C(t) + \\
&\quad - K(t, Q_f(t))D(t)F(t, Q_c(t+1))]\hat{x}(t) + K(t, Q_f(t))y(t), \\
u(t) &= F(t, Q_c(t+1))\hat{x}(t), \\
\hat{z}(t) &= C_z(t)\hat{x}(t) - D_z(t)F(t, Q_c(t+1))\hat{x}(t) \\
&= C_{z,r}(t)\hat{x}_r(t) - C_{z,s}\hat{x}_s(t) + \\
&\quad - D_{z,s}(t)F_r(t, Q_c(t+1))\hat{x}_r(t) - D_{z,s}(t)F_s(t, Q_c(t+1))\hat{x}_s(t).
\end{aligned}$$

Proof. (a) The formulas for the Kalman filter of a Gaussian stochastic control system follow from Theorem 14.4.2. The statements that the matrix functions $Q_c(.)$ and $K(Q_f(.))$ are both block-diagonal follows directly from the block-diagonal structure of the matrix functions A, C, M, N, and Q_{x_0}.
(b) This follows directly from Theorem 14.4.12. □

14.5 Control of a State-Finite Stochastic Control System

The optimal control problem for an output-finite-state-finite stochastic control system with partial observations is quite different from that of the case of a Gaussian stochastic control problem with partial observations.

The optimal control problem for a state-finite stochastic control system with *complete observations* has been solved in Chapter 12. There are many applications of optimal control of such a system in control engineering and in operations research. The computations are relatively simple, for dynamic programming at each time and each state of a finite set of states, one has to minimize a function over the input set, which input set can be either a finite set or a subset of tuples of the real numbers.

Based on the result for the optimal control problem with complete observations, researchers have formulated the optimal control problem for an output-finite-state-finite stochastic control system with *partial observations*. Possibly the expectation was that also in this case one can obtain a simple dynamic programming procedure to compute the optimal control law.

This expectation turned out to be not correct. Due to the partial observations, one has to first construct a stochastic realization of the stochastic control system with respect to the filtration generated by the output process and the input process. The stochastic system of this stochastic realization is the filter system and the state process of that stochastic system takes values in a subset of the probability simplex. Therefore, dynamic programming involves, for each time and each state in the probability simplex, the infimization of a real-valued function over the input set. Even if the input set is finite, one has to carry out the infimization for each state in a subset of the probability simplex. This infimization in the partial observations case is if a much higher complexity than that in the complete observations case. So far there is no theoretical framework for this case. An object to expect is an analytic form of the value function at each time.

There is however an alternative as developed below, see Problem 14.5.10.

Below the reader finds the optimal control theory for an output-finite-state-finite stochastic control system with partial observations. There are few examples which can be solved with this approach. That theory is followed by a another way to look at the problem.

14.5.1 Problem Formulation

Problem 14.5.1. Consider an output-finite-state-finite stochastic control system with partial observations as formulated in Def. 10.4.1 with the system in the indicator representation,

$$E\left[\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix} \mid F_t^x \vee F_{t-1}^y\right] = \begin{pmatrix} A(u(t)) \\ C(u(t)) \end{pmatrix} x(t),\ x(0) = x_0 \in X_e,$$

where (Ω, F, P) is a probability space, $T = \{0, 1, \ldots, t_1\}$ is a finite horizon, $n_x,\ n_y,\ n_u \in \mathbb{Z}_+$, $X_e = \{e_1,\ e_2, \ldots, e_{n_x}\} \subseteq \mathbb{R}_+^{n_x}$ with e_i the i-the unit vector of the space $\mathbb{R}_+^{n_x}$, $Y_e = \{v_1,\ v_2, \ldots, v_{n_x}\} \subseteq \mathbb{R}_+^{n_y}$, with v_i the i-the unit vector of the space $\mathbb{R}_+^{n_y}$, $U \subseteq \mathbb{R}_+^{n_u}$ either a finite set or a subset, $A : U \rightarrow \mathbb{R}_{st}^{n_x \times n_x}$ and $C : U \rightarrow \mathbb{R}_{st}^{n_y \times n_x}$ are measureable maps, $u : \Omega \times T \rightarrow U$ is a stochastic process, $x_0 : \Omega \rightarrow X_e$ is a random variable; and the system representation determines with the right-hand side a conditional probability measure on the tuple of vectors $(x(t+1),\ y(t))$ which is then probabilistically realized by a choice of the values of these vectors according to the determined conditional probability measures such that $x(t+1) \in X_e$ and $y(t) \in Y_e$.

Consider the past-output and past-input information structure,

$$\{F_{t-1}^y \vee F_{t-1}^u, t \in T\},\ F_t^y = \sigma(\{y(s),\ s \leq t\}),\ F_t^u = \sigma(\{u(s),\ s \leq t\}).$$

Consider the set G of control laws compatible with the information structure. For any $g \in G$, denote $g = \{g_0, \ldots, g_{t_1}\}$, such that $g_0 \in U$ and, for all $t \in T$, $g_t : Y^t \times U^t \rightarrow U$.

For any $g \in G$ the closed-loop stochastic control system has the system representation,

$$\begin{aligned} & E\left[\begin{pmatrix} x^g(t+1) \\ y^g(t) \end{pmatrix} \mid F_t^x \vee F_{t-1}^y\right] \\ & = \begin{pmatrix} A(g_t(y^g(0:t-1),u^g(0:t-1))) \\ C(g_t(y^g(0:t-1),u^g(0:t-1))) \end{pmatrix} x^g(t),\ x^g(0) = x_0 \in X_e, \\ u^g(t) & = g_t(y^g(0:t-1),u^g(0:t-1)) \\ & = g_t(y^g(0),\ldots,y^g(t-1),u^g(0),\ldots,u^g(t-1)). \end{aligned}$$

Consider the cost function,

$$J(g) = E[\sum_{s=0}^{t_1-1} b(s,x^g(s),u^g(s)) + b_1(x^g(t_1))],\ \ J : G \rightarrow \mathbb{R}_+.$$

Define the *conditional cost-to-go* as $J : G \times \Omega \times T \rightarrow \mathbb{R}_+$,

$$J(g,t) = E[\sum_{s=t}^{t_1-1} b(s,x^g(s),u^g(s)) + b_1(x^g(t_1))|\ F_{t-1}^{y,g} \vee F_{t-1}^{u,g}]. \tag{14.80}$$

The problem is to solve the optimal stochastic control problem $\inf_{g\in G}\ J(g)$.

14.5.2 Filtering

According to the control synthesis Procedure 14.2.3, the next step is to derive the stochastic realization of the state and the output process with respect to the information structure. This amounts to solve the filter problem.

Problem 14.5.2. Consider the finite stochastic control system of Problem 14.5.1. Determine for any control law $g \in G$ the conditional probability measure and the associated filter system specified by,

$$E\left[\begin{pmatrix} x^g(t+1) \\ y^g(t) \end{pmatrix} \mid F_{t-1}^{y^g} \vee F_{t-1}^{u^g}\right],\ \forall\, t \in T.$$

.

As in the case of a Gaussian stochastic control system, there is the issue that the information structure depends in principle on the control law considered.

Theorem 14.5.3. *The conditional probability measure requested in Problem 14.5.2 is determined by the equations,*

$$\begin{aligned}
&\forall\, g \in G,\ \forall\, t \in T, \\
\hat{x}^g_{un}(t+1) &= A(g_t(y^g(0:t-1), u^g(0:t-1))) \times \\
&\quad \times \text{Diag}(C(g_t(y^g(0:t-1), u^g(0:t-1)))^T y(t))\ \times \hat{x}^g_{un}(t), \\
&\quad \hat{x}^g_{un}(0) = p_{x_0}, \\
\hat{x}^g(t+1) &= \frac{\hat{x}^g_{un}(t+1)}{1^T_{n_x}\hat{x}^g_{un}(t+1)}, \\
E_1[y(t)|\, F^{y^g}_{t-1}] &= C(g_t(y^g(0:t-1), u^g(0:t-1)))\, \hat{x}^g(t), \\
&\quad \hat{x}^g_{un} : \Omega \times T \to \mathbb{R}^{n_x}_+,\ \hat{x}^g : \Omega \times T \to \mathbb{R}^{n_x}_+.
\end{aligned}$$

Call $\hat{x}^g_{un}$ *the* unnormalized conditional probability *of* $x^g(t+1)$ *conditioned on* $F^{y^g}_{t-1}$ *and* $\hat{x}^g$ *the* conditional probability *of the same variables.*

Note that the term $1^T_{n_x}\hat{x}^g_{un}(t+1)$ *in the denominator of the equation of* $\hat{x}^g(t+1)$ *is a* normalization factor *only.*

Proof. The measure transformation approach is used for this problem. The approach is explained for the Gaussian case in Section 16.3.

By the measure transformation approach, one constructs on the probability space (Ω, F, P) a Radon-Nikodym derivative and then a probaility measure P_0 such that: the probability measures P_0 and P are equivalent probability measures, with respect to the new measure P_0 the output process is a sequence of independent random variables each having the same probability distribution, and such that the σ-algebra $F^{y(t+1)}$ is independent of $F^y_t \vee F^x_{t+1}$ for all $t \in T\backslash\{t_1\}$.

Because with respect to the probability measure P_0 the output process y is a sequence of independent and identically distributed random variables, the filtration $\{F^{y^g}_t,\ t \in T\}$ cannot depend in the control law $g \in G$, hence $F^{y_g}_t = F^y_t$ for all $t \in T$. This conclusion could also have been achieved by starting the construction of the probability measure of the problem statement with the probability measure P_0 as described above and then constructing the new measure P such that the properties of the problem statement hold.

After this change, the proof is identical to that of Theorem 9.7.2 and the result follows from that theorem. □

The cost function can be projected on the filtration of the information structure.

Proposition 14.5.4. *Consider Problem 14.5.1 for a state-finite stochastic control system. Then the projections of the terminal cost function and of the cost rate on the information structure are provided by the equations,*

$$E[b_1(x(t_1))|\, F^y_{t_1-1} \vee F^u_{t_1-1}] = \sum_{i=1}^{n_x} b_1(e_i)\hat{x}_i(t_1);$$

$$E[b(t, x(t), u(t))|\, F^y_{t-1} \vee F^u_{t-1}] = \sum_{i=1}^{n_x} b(t, e_i, u(t))\, \hat{x}_i(t), \forall\, t \in T\backslash\{t_1\}.$$

Both terms are linear functions of the estimated state $\hat{x}$.

Proof. Note that,

$$
\begin{aligned}
& E[b_1(x(t_1))|\, F^y_{t_1-1} \vee F^u_{t_1-1}] \\
& = \sum_{i=1}^{n_x} E[b_1(e_i) I_{\{x(t_1)=e_i\}} | F^y_{t_1-1} \vee F^u_{t_1-1}] = \sum_{i=1}^{n_x} b_1(e_i) \hat{x}_i(t_1); \\
& E[b(t,x(t),u(t))|\, F^y_{t-1} \vee F^u_{t-1}] \\
& = \sum_{i=1}^{n_x} E[b(t,e_i,u(t))\; I_{\{x(t)=e_i\}}|\, F^y_{t-1} \vee F^u_{t-1}] = \sum_{i=1}^{n_x} b(t,e_i,u(t))\; \hat{x}_i(t).
\end{aligned}
$$

Note that by assumption, $u(t)$ is $F^y_{t-1} \vee F^u_{t-1}$ measurable hence comes out of the conditional expectation in the second relation. □

14.5.3 Dynamic Programming

Below Problem 14.5.1 is solved by dynamic programming according to the approach sketched in Section 14.2.

Problem 14.5.5. Consider Problem 14.5.1. Consider the stochastic realization of the stochastic control system on the information structure provided by the filter realization as stated in Theorem 14.5.3. The problem is specified by the sets and the equations for the stochastic realization,

$$
\begin{aligned}
& E_1 \left[\begin{pmatrix} \hat{x}_{un}(t+1) \\ y(t) \end{pmatrix} |\, F^y_{t-1} \vee F^u_{t-1} \right] \\
& = \begin{pmatrix} A(u(t))\ \mathrm{Diag}\left(C(u(t))^T\ E_1[y(t))|\, F^y_{t-1} \vee F^u_{t-1}]\ \hat{x}_{un}(t)\right) \\ C(u(t))\ \hat{x}_{un}(t) \end{pmatrix}, \\
& \hat{x}_{un}(0) = p_{x_0}, \\
\hat{x}^g(t+1) = {} & \frac{\hat{x}^g_{un}(t+1)}{1^T_{n_x} \hat{x}^g_{un}(t+1)},\ \ \hat{x}^g_{un} : \Omega \times T \to \mathbb{R}^{n_x}_+,\ \hat{x}^g : \Omega \times T \to \mathbb{R}^{n_x}_{st}.
\end{aligned}
$$

For any control law one obtains the closed-loop system of the form,

$$
\begin{aligned}
& E_1 \left[\begin{pmatrix} \hat{x}^g_{un}(t+1) \\ y^g(t) \end{pmatrix} |\, F^{y^g}_{t-1} \vee F^{u^g}_{t-1} \right] \\
& = \begin{pmatrix} A(g_t(y^g(0:t-1), u^g(0:t-1)))\times \\ \times \mathrm{Diag}\left(C(g_t(y^g,u^g))^T\ E_{[}y^g(t))|\, F^{y^g}_{t-1} \vee F^{u^g}_{t-1}]\right)\ \hat{x}^g_{un}(t) \\ C(g_t(y^g,u^g))\ \hat{x}^g_{un}(t) \end{pmatrix}, \\
& \hat{x}_{un}(0) = p_{x_0},\ \ \forall\, g \in G,\ \forall\, t \in T, \\
& g_t(y^g,u^g) = g_t(y^g(0:t-1), u^g(0:t-1)))^T, \\
\hat{x}^g(t+1) = {} & \frac{\hat{x}^g_{un}(t+1)}{1^T_{n_x} \hat{x}^g_{un}(t+1)},\ \hat{x}^g_{un} : \Omega \times T \to \mathbb{R}^{n_x}_+,\ \hat{x}^g : \Omega \times T \to \mathbb{R}^{n_x}_{st}.
\end{aligned}
$$

This problem statement requires much explanation. The problem statement describes the stochastic realization of the original stochastic control system with respect to the observation filtration $\{F^y_{t-1} \vee F^u_{t-1},\ t \in T\}$. The state process of this stochastic system takes values in $\hat{x}_{un} : \Omega \times T \to \mathbb{R}^{n_x}_+$. Note that $\{\hat{x}_{un}(t),\ F^{yg}_{t-1} \vee F^{ug}_{t-1},\ t \in T\}$ is an adapted process. The output process is a finite valued process and takes values in $Y_e = \{e_1,\ e_2, \ldots,\ e_{n_y}\} \subset \mathbb{R}^{n_y}_{st}$.

Note that, due to the stochastic realization chosen, any function $g(\hat{x}_{un}(t))$ of the state process $\hat{x}_{un}$ is adapted to the observation filtration. Therefore, one can apply the dynamic programming procedure stated below.

For the solution of the optimal stochastic control problem one needs conditions of controllability and of observability to hold. Controllability requires the calculation of the attainable subset of probability vectors in terms of $\hat{x}_{un}(t) \in \mathbb{R}^{n_x}_+$. This requires an understanding how the stochastic system generates the next state $\hat{x}_{un}(t+1)$ based on the past state, on the output $y(t)$, and on the input $u(t)$. A characterization of the attainable subset of probability measures as a subset of the space $\mathbb{R}^{n_x}_+$, is not determined yet. A solution approach can be that similar to the characterization of observability of an output-finite-state-finite stochastic system, see Theorem 5.7.27.

Also one needs a condition of observability. But note that in the stochastic system of Problem 14.5.5 the observation matrix $C(u(t))$ depends on the value of the input at that time. Therefore, the input affects the observability. There may exist inputs for which observability does not hold. There may exist other inputs for which observability holds and the state estimation is really good. In the optimal control problem then a trade-off is made between the tasks of state estimation and of performance minimization. For this stochastic system, the observability and the controllability are thus thightly intertwined. No satisfactory characterization is known of observability for this filter realization of the stochastic control system.

The problem of characterization of observability of the stochastic realization of the stochastic control system does not occur in control of a Gaussian stochastic control problem with partial observations as discussed earlier in this chapter. This is primarily due to the linearity of the system equation for that stochastic control system.

Below the dynamic programming procedure is specified and the theorem for the value function is stated.

Procedure 14.5.6 *The* dynamic programming procedure *for a partially-observed state-finite stochastic control system.*

1. *Initialization. Define,*

$$V(t_1,\hat{x}_{un,V}) = \sum_{i=1}^{n_x} b_1(e_i)\, \hat{x}_{un,V,i} = r(t_1)^T\, \hat{x}_{un,V},\ \ V(t_1,.) : T \times \mathbb{R}^{n_x}_+ \to \mathbb{R}_+,$$

$$r(t_1)^T = \big(b_1(e_1)\ \ldots\ b_1(e_{n_x})\big)/[1^T_{n_x}\, \hat{x}_{un}(t_1)] \in \mathbb{R}^{1\times n_x}.$$

2. *For* $t = t_1 - 1,\ t_1 - 2,\ \ldots,\ 0$ *do:*

- *Determine,*

$$\begin{aligned}
&V(t,\hat{x}_{un,V}) \\
&= \inf_{u_V \in U(t,\hat{x}_{un,V})} \left\{ \begin{array}{l} E[b(t,x(t),u(t))|F^y_{t-1} \vee F^u_{t-1}] + \\ +E[V(t+1,\hat{x}_{un}(t+1))|F^y_{t-1} \vee F^u_{t-1}] \end{array} \right\} \\
&= \inf_{u_V \in U(t,\hat{x}_{un,V})} \left\{ \sum_{i=1}^{n_x} s_i(t,u_V)\, \hat{x}_{un,V,i} \right\}, \;\; V(t,.) : \mathbb{R}^{n_x}_+ \rightarrow \mathbb{R}_+; \\
&s_i(t,u_V) \\
&= b(t,e_i,u_V) + \\
&\quad +E[r(t+1)^T\, A(u_V)\, \text{Diag}(C(u_V)^T y(t))|\, F^y_{t-1} \vee F^u_{t-1}]_i.
\end{aligned}$$

- *If in the dynamic programming procedure the infima are all attained,*

$$\begin{aligned}
&\forall\, \hat{x}_{un,V},\; \exists\, u^* \in U(t,\hat{x}_{un,V}) \text{ such that,} \\
&\left\{ \sum_{i=1}^{n_x} s_i(t,u^*)\, \hat{x}_{un,V,i} \right\} = \inf_{u_V \in U(t,\hat{x}_V)} \left\{ \sum_{i=1}^{n_x} s_i(t,u_V)\, \hat{x}_{un,V,i} \right\}, \quad (14.81) \\
&\text{then define,} \quad (14.82) \\
&g^*(t,\hat{x}_{un,V}) = u^*, \;\; g^* : T \times \mathbb{R}^{n_x}_+ \rightarrow U; \;\; \forall\, i \in \mathbb{Z}_{n_x}, \\
r_i(t) &= b(t,e_i,u^*) + \quad (14.83) \\
&+E[r(t+1)^T\, A(u^*)\, \text{Diag}(C(u^*)^T y(t))|\, F^y_{t-1} \vee F^u_{t-1}]_i.
\end{aligned}$$

If the set $U(t,x_V)$ is finite then the minimum is of course attained.

3. *Check whether V and g^* are measurable functions. Stop if not else proceed.*
4. *Output the value function, the optimal control law (V,g^*).*

The infimization in the dynamic programming procedure formulated above has to be performed over the input set U for any element $\hat{x}_{un,V} \in \mathbb{R}^{n_x}_+$. If the input set U is a finite set then there exists a partition of the state set $\hat{X}_{un}$ in the form of $\{\hat{X}_{un,i} \subseteq \hat{X}_{un},\; i \in \mathbb{Z}_{n_u}\}$ such that, if $\hat{x}_{un} \in X_{un,i}$, then $u^* = g^*(t,\hat{x}_{un}) = u_i$. It wil be nice if the subsets of the partition have a special geometric structure, like being a polytope or a polyhedral cone. If U is not a finite set then one has to determine the analytic, algebraic, or geometric form of the optimal control law. The complexity of such a description can be quite large.

Note further that $\hat{x}_{un,V} \in \mathbb{R}^{n_x}_+$ belongs to the positive orthant. The optimal control law will therefore be a function of $\hat{x}_{un}(t)$ of which the analytic or algebraic form has to be determined. The calculation of the value function is therefore of a much higher complexity than in the case of complete observations where the state set is a finite set.

These comments show that dynamic programming for the partially-observed stochastic control problem is quite different from that of the complete observations case.

Proposition 14.5.7. *Procedure 14.5.6 is well defined for Problem 14.5.1.*

Proof. The initialization of the value function. Note that by Proposition 14.5.4,

$$\begin{aligned}
V(t_1,\hat{x}_{un,V}) &= E[b_1(x(t_1))|F^{y^g}_{t_1-1}] = \sum_{i=1}^{n_x} b_1(e_i)E[I_{\{x(t_1)=e_i)\}}||\, F^{y^g}_{t_1-1}] \\
&= \sum_{i=1}^{n_x} b_1(e_i)\,\hat{x}_i(t_1) = \sum b_1(e_i)\,\hat{x}_{un,i}(t_1)/[1^T_{n_x}\hat{x}_{un}(t_1)] \\
&= \sum_{i=1}^{n_x} r_i(t_1)\,\hat{x}_{un,V,i}.
\end{aligned}$$

By induction it is proven that the value function is a linear function of the state $\hat{x}_{un}$. This is true for t_1 as proven above. Suppose the linearity holds for $s = t_1,\ t_1 - 1, \ldots,\ t+1$. It will then be proven for $s = t$.

$$\begin{aligned}
& V(t,\hat{x}_{un,V}) = V(t,\hat{x}_{un}(t)) \\
& = \inf_{u_V \in U(t,\hat{x}_{un,V})} \left\{ \begin{array}{l} E[b(t,x(t),u(t))|F^y_{t-1} \vee F^u_{t-1}]+ \\ +E[V(t+1,\hat{x}_{un}(t+1))|F^y_{t-1} \vee F^u_{t-1}] \end{array} \right\}; \\
& = \inf \left\{ \begin{array}{l} \sum_{i=1}^{n_x} b(t,e_i,u_V)\,\hat{x}_{un,i}(t)+ \\ +E[V(t+1,\hat{x}_{un}(t+1))|F^y_{t-1} \vee F^u_{t-1}] \end{array} \right\}; \\
& E[V(t+1,\hat{x}_{un}(t+1))|F^y_{t-1} \vee F^u_{t-1}] \\
& = E[r(t+1)^T\ \hat{x}_{un}(t+1)|\ F^y_{t-1} \vee F^u_{t-1}] \\
& = E[E[r(t+1)^T\ \hat{x}_{un}(t+1))|\ F^y_t \vee F^u_t]|\ F^y_{t-1} \vee F^u_{t-1}] \\
& = E[r(t+1)^T\ A(u(t))\ \text{Diag}(C(u(t))^T y(t))\ \hat{x}_{un}(t)|\ F^y_{t-1} \vee F^u_{t-1}] \\
& = E[r(t+1))^T\ A(u(t))\ \text{Diag}(C(u(t))^T y(t))|\ F^y_{t-1} \vee F^u_{t-1}]\ \hat{x}_{un}(t) \\
& = E[r(t+1)^T\ A(u_V)\ \text{Diag}(C(u_V)^T y(t))|\ F^y_{t-1} \vee F^u_{t-1}]\ \hat{x}_{un}(t); \\
& s_i(t) \\
& = b(t,e_i,u_V) + E[r(t+1)^T\ A(u_V)\ \text{Diag}(C(u_V)^T y(t))|\ F^y_{t-1} \vee F^u_{t-1}]_i,
\end{aligned}$$

$$V(t,\hat{x}_{un,V}) = \inf_{u_V \in U(t,x_V)} \left\{ \sum_{i=1}^{n_x} s_i(t,u_V)\,\hat{x}_{un,V,i} \right\};$$

if $\forall\ (t,\hat{x}_V) \in T \times \hat{X}_{un}\ \exists\ u^* \in U(t,\hat{x}_{un})$ such that

$$V(t,\hat{x}_{un,V}) = \left\{ \sum_{i=1}^{n_x} s_i(t,u^*)\,\hat{x}_{un,V,i} \right\} = \inf_{u_V \in U(t,x_V)} \left\{ \sum_{i=1}^{n_x} s_i(t,u_V)\,\hat{x}_{un,V,i} \right\},$$

then define $r_i(t) = s_i(t,u^*(t,x_V))$; consequently $\forall\ i \in \mathbb{Z}_{n_x}$,

$$\begin{aligned}
r_i(t) &= b(t,e_i,u^*) + \\
&\quad +E[r(t+1)^T\ A(u^*)\ \text{Diag}(C(u^*)^T y(t))|\ F^y_{t-1} \vee F^u_{t-1}]_i, \\
V(t,\hat{x}_{un,V}) &= r(t)^T \hat{x}_{un,V}.
\end{aligned}$$

□

Theorem 14.5.8. *Consider Problem 14.5.1. Let the function $V : T \times \mathbb{R}^{n_x}_{st} \rightarrow \mathbb{R}_+$ be defined by the Dynamic Programming Procedure 14.5.6.*

(a)For any $g \in G$ and $t \in T$

$$\begin{aligned} V(t,\hat{x}(t)) &\leq J(g,t) \text{ a.s.} \quad (14.84)\\ E[V(0,\hat{x}(0)] &\leq E[J(g,0)] = J(g),\\ E[V(0,\hat{x}(0)] &\leq \inf_{g \in G} J(g). \end{aligned}$$

(b)Let $g^* \in G$, $g^* : T \times \hat{X}_{un} \to U$, be such that it attains the infima in the dynamic programming procedure stated above. Then,

$$\begin{aligned} V(t,\hat{x}(t)) &= J(g^*,t) \text{ a.s.}, \forall\, t \in T,\\ E[V(0,x_0)] &= E[J(g^*,0)] = J(g^*) = \inf_{g \in G} J(g) = J^*, \quad (14.85) \end{aligned}$$

hence $g^* \in G$ is an optimal control law.

Proof. (a)

$$\begin{aligned} J(g,t_1) &= E[b_1(x(t_1))|F^y_{t_1-1} \vee F^u_{t_1-1}] \quad (14.86)\\ &= \sum_{i \in X} b_1(i)p(i;x(t_1)|y^{t_1-1},u^{t_1-1}) \quad (14.87) \end{aligned}$$

hence Equation (14.84) holds with equality for $t = t_1$. Suppose that Equation (14.84) holds for $s = t+1, t+2, \ldots, t_1$. Then

$$\begin{aligned} J(g,t) &= E\begin{bmatrix} b(t,x(t),u(t))+ \\ +E[\sum_{s=t+1}^{t_1-1} b(s,x(s),u(s)) + \; b_1(x(t_1))|F^y_t \vee F^u_t]|F^y_{t-1} \vee F^u_{t-1} \end{bmatrix}\\ &= E\left[b(t,x(t),u(t)) + J(g,t+1)|F^y_{t-1} \vee F^u_{t-1}\right]\\ &\quad \text{by definition of } J(g,t+1),\\ &\geq E[b(t,x(t),u(t)) + V(t+1,\hat{x}_{un}(t+1))|F^y_{t-1} \vee F^u_{t-1}]\\ &\quad \text{by the induction assumption,}\\ &= E[E[b(t,x(t),u(t)) + V(t+1,\hat{x}_{un}(t+1))|F^y_t \vee F^u_t]|F^y_{t-1} \vee F^u_{t-1}] \end{aligned}$$

$$\begin{aligned} &= E[E[b(t,x(t),u(t))+\\ &\quad +V(t+1,f_1(t,\hat{x}_{un}(t)),y(t),u(t)))|\hat{x}_{un}(t),u(t)]|F^y_{t-1} \vee F^u_{t-1}]\\ &\geq E[V(t,\hat{x}_{un}(t))|F^y_{t-1} \vee F^u_{t-1}] \text{ by Equation (14.81),}\\ &= V(t,\hat{x}_{un}(t)), \end{aligned}$$

hence Equation (14.84) holds for t and the proof completes by induction.

(b)From Equation (14.86) follows that Equation (14.85) holds with equality for $s = t_1$. Suppose that Equation (14.85) holds for $s = t+1, t+2, \ldots, t_1$. In the proof of (a) both inequalities become equalities, the first inequality by the induction assumption and the second because $u(t) = g^*(t,\hat{x}_{un}(t))$ achieves the minimum in Equation (14.81). Note further that for any $g \in G$

$$\begin{aligned} J(g^*) &= E[J(g^*,0)], \text{ by (14.80),}\\ &= E[V(0,\hat{x}_{un}(0))], \text{ by (14.85),}\\ &\leq E[J(g,0)] = J(g) \end{aligned}$$

by Equations (14.84) and (14.80). This and Equation (14.84) imply that

$$J^* = J(g^*) = \inf_{g \in G} J(g),$$

and $g^* \in G$ is an optimal control law.

□

14.5.4 Example

Example 14.5.9. *Optimal control of an output-finite-state-finite stochastic control system.*

Consider the special case of Problem 14.5.1 for an output-finite-state-finite stochastic control system in the indicator representation,

$$E\left[\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix} \mid F_t^x \vee F_{t-1}^y \vee F_t^u\right] = \begin{pmatrix} A(u(t)) \\ C(u(t)) \end{pmatrix} x(t),\ x(0) = x_0,$$
$$n_x = 2,\ n_u = 1,\ n_y = 2,\ U = [0, 0.3] \subset \mathbb{R},$$
$$x : \Omega \times T \to X_e = \{e_1,\ e_2 \in \mathbb{R}^2\},\ y : \Omega \times T \to Y_e = \{e_1,\ e_2 \in \mathbb{R}^2\},$$
$$A(u) = \left(\begin{array}{c|c} 0.8-u & 0.5+u \\ \hline 0.2+u & 0.5-u \end{array}\right),\ C(u) = \left(\begin{array}{c|c} 0.6+u & 0.2 \\ \hline 0.4-u & 0.8 \end{array}\right).$$

The dynamics of the system is such that the state is either e_1 or $e_2 \in \mathbb{R}_+^2$. For $u = 0$ the probability that the state is e_1 is higher than that it is at state e_2. For $u = 0.3$ the state is at is at either state e_1 or e_2 with probabilities which both are not small. The cost rate is set such that the cost rate is lowest if the state is e_1 and $u = 0$.

The output equation with $C(u)$ is such that for $u = 0.3$ one obtains the output $y(t) = e_1$ with the high probability of 0.9, while if $u = 0$ the probability of the output $y(t) = e_1$ is only 0.6 and that it is $y(t) = e_2$ is 0.4 which are not far apart. Therefore, the output is most expressive if u has a high value.

Note that there is thus a conflict in the determination of the input $u \in U$ between (1) infimization of the cost and (2) the probability of obtaining a reliable output. The optimal control problem is to find a balance between these conflicting tasks.

For any input value $u_r \in U$, the state-transition matrix $A(u_r)$ has only nonzero elements hence is an irreducible matrix. Therefore there exists a steady state vector $p_s(u_r)$ satisfying $A(u_r)p_s(u_r) = p_s(u_r)$, which is easily calculated to be,

$$p_s(u_r) = \begin{pmatrix} 0.5+u_r \\ 0.2+u_r \end{pmatrix} \frac{1}{0.7+2u_r} \in \mathbb{R}_{st}^{n_x}.$$

The filter for the estimation of the state based on the past observations, based on Theorem 14.5.3, is described by the formulas,

$$\hat{x}_{un}(t+1) = A(u(t))\ \mathrm{Diag}(C(u(t))^T y(t))\ \hat{x}_{un}(t),\ \hat{x}_{un}(0) = p_{x_0},$$
$$\hat{x}(t+1) = \hat{x}_{un}(t+1)/[1_{n_x}^T \hat{x}_{un}(t+1)].$$

The stochastic realization of the stochastic control system with respect to the filtration of the output and the input is then,

$$\begin{aligned}
&\{F^y_{t-1} \vee F^u_{t-1},\ t \in T\},\ \text{the observation filtration,}\\
&\{u(t),\ F^y_{t-1} \vee F^u_{t-1},\ t \in T\},\ \text{the adaptedness of the input process,}\\
&E\left[\begin{pmatrix} \hat{x}_{un}(t+1) \\ y(t) \end{pmatrix} \mid F^y_{t-1} \vee F^u_{t-1}\right]\\
= &\begin{pmatrix} A(u(t))\ \mathrm{Diag}(C(u(t))^T C(u(t))\ \hat{x}_{un}(t))\ \frac{1}{1^T \hat{x}_{un}(t)} \\ C(u(t)) \end{pmatrix} \hat{x}_{un}(t)/[1^T \hat{x}_{un}(t)],\\
&\hat{x}_{un}(0) = p_{x_0};\ \text{ where use is made of the calculation,}\\
&E[y(t)|\ F^y_{t-1} \vee F^u_{t-1}] = E[E[y(t)|\ F^x_t \vee F^y_{t-1} \vee F^u_{t-1}]|\ F^y_{t-1} \vee F^u_{t-1}]\\
= &E[C(u(t))x(t)|\ F^y_{t-1} \vee F^u_{t-1}] = C(u(t))E[x(t)|\ F^y_{t-1} \vee F^u_{t-1}]\\
= &C(u(t))\hat{x}_{un}(t)/[1^T \hat{x}_{un}(t)].
\end{aligned}$$

Define the cost function by the expressions,

$$\begin{aligned}
J(g) &= E\left[\sum_{s=0}^{t_1} b(s,x(s),u(s))\ b_1(x(t_1))\right],\\
b_1(x) &= I_{\{x=e_1\}} + 2I_{\{x=e_2\}},\\
b(s,x,u) &= (I_{\{x=e_1\}} + 2I_{\{x=e_2\}})(0.3-u),\\
&\quad E[b_1(x(t_1))|F^y_{t_1-1} \vee F^u_{t_1-1}]\\
&= 1\hat{x}_{un,1}(t_1)/[1^T \hat{x}_{un}(t_1)] + 2\hat{x}_{un,2}(t_1)/[1^T \hat{x}_{un}(t_1)],\\
&\quad E[b(t,x(t),u)|F^y_{t-1} \vee F^u_{t-1}]\\
&= [\hat{x}_{un,1} + 2\hat{x}_{un,2}](0.3-u)/[1^T \hat{x}_{un}(t)].
\end{aligned}$$

The dynamic programming procedure is written explicitly for the first two steps.

$$\begin{aligned}
&V(t_1,\hat{x}_{un,V})\\
&= E[b_1(x(t_1))|F^y_{t_1-1} \vee F^u_{t_1-1}] = [1\hat{x}_{un,V,1} + 2\hat{x}_{un,V,2}]/[1^T \hat{x}_{un,V}],\\
&E[\hat{x}_{un}(t_1)|\ F^y_{t_1-1} \vee F^u_{t_1-1}]\\
&= A(u(t_1-1))\ \mathrm{Diag}(C(u(t_1-1))^T y(t_1-1))\ \hat{x}_{un}(t_1-1),\\
&\quad 1^T \hat{x}_{un}(t_1) = y(t_1-1)^T C(u(t_1-1))\hat{x}_{un}(t_1-1),\\
&V(t_1-1,\hat{x}_{un,V})\\
&= \inf_{u_V \in U} \left\{E[b(t,x(t),u)|F^y_{t-1} \vee F^u_{t-1}] + E[V(t_1,\hat{X}(t_1))|\ F^y_{t_1-1} \vee F^u_{t_1-1}]\right\}\\
&= \inf \left\{\begin{array}{l} [\hat{x}_{un,1}(t_1-1) + 2\hat{x}_{un,2}(t_1-1)](0.3-u(t_1-1))/[1^T \hat{x}_{un}(t_1-1)]+ \\ +(1\ 2)\ A(u(t_1-1))\ \mathrm{Diag}(C(u(t_1-1))^T y(t_1-1))\ \hat{x}_{un}(t_1-1)\times \\ /[y(t_1-1)^T C(u(t_1-1))\hat{x}_{un}(t_1-1)] \end{array}\right\}\\
&= \inf_{u_V \in U} \left\{\begin{array}{l} [\hat{x}_{un,V,1} + 2\hat{x}_{un,V,2}](0.3-u_V)/[1^T \hat{x}_{un,V}]+ \\ +(1\ 2)\ A(u_V)\ \mathrm{Diag}(C(u_V)^T y(t_1-1))\ \hat{x}_{un,V} \\ \quad /[y(t_1-1)^T C(u_V)\hat{x}_{un,V}] \end{array}\right\}.
\end{aligned}$$

The calculation of $V(t_1-1,\hat{x}_{un,V})$ now requires the infimization of the last expression above, for any $\hat{x}_{un,V}\in\mathbb{R}^2_{st}$. The function to be infimized is a rational function of the variable $u_V\in U=[0,0.3]$. Infimization is therefore possible but depends on the value of $\hat{x}_{un,V}$ chosen.

The complexity of the required calculations do not make one enthusiastic about the solution of this optimal control problem.

14.5.5 Alternative Control Problem

For the stochastic control problem for an output-finite-state-finite stochastic control system with partial observations it is possible to reformulate the problem statement so that the problem can be solved in a simple way. The problem formulation stated below is an approximation of Problem 14.5.1. In general, the two problems are not identical.

Problem 14.5.10. *Stochastic control of an output-finite-state-finite stochastic control system formulated as an observation-based realization.* Consider an output-finite-state-finite stochastic system in the observation-based stochastic realization described by the system,

$$\begin{aligned}
x(t+1) &= A(t,u(t),y(t))\,x(t),\ x(0)=x_0,\\
E[y(t)|F^y_{t-1}] &= C(t,u(t),y(t))\,x(t),\\
&\qquad X_e=\{e_1,\,e_2,\,\ldots,\,e_{n_x}\in\mathbb{R}^{n_x}\},\ Y_e=\{e_1,\,e_2,\,\ldots,\,e_{n_y}\in\mathbb{R}^{n_y}\},\\
&\qquad U\subseteq\mathbb{R}^{n_u}\text{ either a finite set or a measurable subset},\\
&\qquad x:\Omega\times T\to X_e,\ y:\Omega\times T\to Y_e,\ u:\Omega\times T\to U,\ x_0\in X_e,\\
&\qquad A:T\times U\times Y\to\mathbb{R}^{n_x\times n_x}_{st},\ C:T\times U\times Y\to\mathbb{R}^{n_y\times n_x}_{st}.
\end{aligned}$$

Define the set of control laws as,

$$G=\left\{(g_0,g_1,\ldots,g_{t_1-1})|\ g_0\in U,\ \forall\,t\in T,\ g_t:U^t\times Y^t\to U\right\}.$$

The closed-loop system is then defined by the formulas,

$$\begin{aligned}
x^g(t+1) &= A(t,g(t,u^g(0:t-1),y^g(0:t-1)),y^g(t))\,x^g(t),\\
&\quad x^g(0)=x_0\in X_e,\\
E[y^g(t)|F^y_{t-1}] &= C(t,g(t,u^g(0:t-1),y^g(0:t-1)),y^g(t))\,x^g(t).
\end{aligned}$$

Define the cost function,

$$\begin{aligned}
J(g) &= E\left[\sum_{s=0}^{t_1-1} b(s,x^g(x),u^g(s),y^g(s))+b_1(x^g(t_1))\right],\\
&\quad J:G\to\mathbb{R}_+,\ b:T\times X_e\times U\times Y_e\to\mathbb{R}_+,\ b_1:X\to\mathbb{R}_+.
\end{aligned}$$

The optimal control problem is then to solve the problem $\inf_{g\in G} J(g)$ for the optimal control law.

The above stochastic control problem is such that the state set X_e is a finite set and such that $\{x(t), F^y_{t-1}, t \in T\}$ is an adapted process. There is no need to obtain a stochastic realization of this stochastic control with respect to the observation filtration because the state process is by definition of the stochastic control system already adapted to the observation filtration.

The solution of the above formulated optimal stochastic control problem follows directly from the dynamic programming Procedure 12.6.2 and from Theorem 12.6.4. Note that the state set X_e is a finite set so that in the dynamic programming procedure one has to determine the value function at every time for a finite set of states. If the input set U is a finite set then the infimization of the dynamic procedure is actually a minimization over a finite set which can be computed by a computer program. If the input set U is a subset of the real numbers then a computation with a higher complexity has to be carried out.

A point requiring attention is to select the parameters of the above stochastic control system such that it is a realistic model for the considered phenomenon.

14.6 Exercises

Problem 14.6.1. Provide the proof of Proposition 14.4.17.

Problem 14.6.2. Provide the proof of Proposition 14.4.18.

14.7 Further Reading

History. The solution of the finite-horizon LQG optimal stochastic control problem for a Gaussian stochastic control system with partial observations is due to T.L. Gunckel, P.D. Joseph, and J.T. Tou, [27, 29]. For continuous-time optimal stochastic control problems the construction of the observation-based stochastic realization of the stochastic control system with respect to the σ-algebra family generated by the past observations and the past inputs, and the optimal control theory, was proven by W.M. Wonham [58]. Novel in that paper is the proof that the filtration does not depend on the control law and that consequently the filter system of the stochastic control system can be derived by the existing results for the Kalman filter. A second paper of W.M. Wonham on the existence of a solution of the control Riccati differential equation is also essential for the solution, [57]. A general discussion of the relevance of optimal stochastic control to control engineering is presented by W.M. Wonham in [59].

The paper of C. Striebel, [47], has clarified in a general setting the concept of an observation-based stochastic realization and its relation with the literature on sufficient statistics. Related publications of the same period are by M. Aoki, [3, 4], K.J. Aström, [31]. From the Russian school there are contributions by E.B. Dynkin [22] and A. Sirjaev [44]. Y. Sawaragi and T. Yoshikawa formulated the framework of control with partial observations in terms of Borel spaces and formulated a stochastic realization in terms of the conditional distribution which is then the new stochastic control system for which the problem can be solved, [43].

A framework to investigate decentralized and distributed stochastic control systems was developed by H.V. Witsenhausen, [55]. The paper [56] provides a formulation for sequential stochastic control in a standard form.

Other papers on the theoretical framework for control of stochastic systems with partial observations include [1, 18, 17]. The predicted miss example is treated in [47, 49].

For the case that the cost function is the expected value of the natural exponent with in the exponent an additive cost function, see [13, 51, 54, 52, 53]. This is also called the risk-sensitive control optimal control problem with partial observations.

The synthesis approach for continuous-time stochastic control systems based on filtering, is formulated by V.E. Beneš and I. Karatzas in [11].

Optimal stochastic control with partial observations on a finite horizon of a output-finite-state-finite stochastic system. In the research area of operations research there are early papers on control of partially-observed output-finite-state-finite stochastic systems, [8, 9, 40, 42, 43]. See also the paper of A.A. Yuskhevich, [60]. A survey dated 1982 is [38] which includes references to the early literature and several examples. A proposal for a geometric approach is provided by Hao Zhang in [61].

Books. At the level of these notes, books on the subject of this chapter are [14, Ch. 4], [16], [19, 6.3], [34, 6.4 - 6.7] and [35]. Books at an advanced level are [15, 48]. Older books are [6, 4]. Not mentioned here are several books exclusively focused on these problems for continuoust-time stochastic systems.

Books with text on control of partially-observed output-finite-state-finite stochastic control systems include, beside the books quoted above, also that by V. Krishnamurthy [33].

Motivation. The example of a moored tanker is presented in [36]. The optimal stochastic control problem for a paper machine is treated in [5]. For applications of optimal stochastic control to economics see [30, 50].

Control synthesis. The concepts of information state, information system, and sufficient information state were introduced by C. Striebel [47]. A survey of concepts and theorems for stochastic control problems, stochastic game problems, and team problems has been presented by H. Witsenhausen in [55]. The concept of neutrality, and the interaction of the control tasks of caution and probing are discussed in [28, 39]. An interpretation of the control tasks is also presented in [7].

For the construction of ε-optimal control laws for optimal stochastic control problems of nonlinear stochastic control systems with partial observations, see [12, 37, 41].

For more on the example of a predicted miss see [10].

Stochastic control of partially-observed output-finite-state-finite stochastic systems. Books which include text about the subject include [34, 6.4-6.7], [14, Ch. 4], and [33].

A very early publication for this subject is by A. Drake, [20] which is a Sc.D. thesis of MIT published in 1962. Of the Russian school of probability there are publications by E. Dynkin, [21], and by A. Sirjaev (most likely to be identical to the mathematican whose name is now written as A. Shiryayev), [44]. E. Sondik has published about the computational issues of control of a output-finite-state-finite stochastic system, [23, 46, 45]. A survey of this research topic with many references is [38]. Later papers are [24, 32]. For optimization of controllers, see [2].

References

1. C. Aldrich and D.W. Peterson. Quadraticity and neutrality in discrete time stochastic linear quadratic control. *Automatica*, 13:307–312, 1977. 575
2. C. Amato, D.S. Bernstein, and S.Zilberstein. Optimizing fixed-size stochastic controllers for POMDP and decentralized POMPD. *Autonomous Agents and Multi-Agent Systems*, 21:293–320, 2010. 576
3. M. Aoki. Optimal control of partially observable Markov systems. *J. Franklin Institute*, 280:367–386, 1965. 574
4. M. Aoki. *Optimization of stochastic systems*. Academic Press, New York, 1967. 574, 575
5. K.J. Aström. Computer control of a paper machine - An application of linear stochastic control theory. *IBM J. Res. & Developm.*, 11:389–405, 1967. 9, 78, 120, 522, 575, 596
6. K.J. Aström. *Introduction to stochastic control*. Academic Press, New York, 1970. 376, 410, 467, 522, 575, 596
7. Y. Bar-Shalom and E. Tse. Dual effect, certainty equivalence, and separation in stochastic control. *IEEE Trans. Automatic Control*, 19:494–500, 1974. 575
8. L.E. Baum and T. Petrie. Statistical inference for probabilistic functions of Markov chains. *Ann. Math. Statist.*, 37:1554–1563, 1966. 575
9. L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, 41:164–171, 1970. 575
10. V.E. Benes. Full "bang" to reduce predicted miss is optimal. *SIAM J. Control & Opt.*, 14:62–84, 1976. 411, 575
11. V.E. Benes and I. Karatzas. Filtering of diffusions controlled through their conditional measures. *Stochastics*, 13:1–23, 1984. 575
12. A. Bensoussan and W. Runggaldier. An approxmation method for stochastic control problems with partial observations of the state – A method for construction ε-optimal controls. *Acta Appl. Mathem.*, 10:145–170, 1987. 575
13. A. Bensoussan and J.H. van Schuppen. Optimal control of partially observable stochastic systems with an exponential-of-integral performance index. *SIAM J. Control Optim.*, 23:599–613, 1985. 575
14. D.P. Bertsekas. *Dynamic programming and stochastic control*. Academic Press, New York, 1976. 376, 405, 410, 439, 468, 502, 525, 526, 575, 595
15. D.P. Bertsekas and S.E. Shreve. *Stochastic optimal control: The discrete time case*. Academic Press, New York, 1978. 49, 428, 431, 468, 575, 595
16. P.E. Caines. *Linear stochastic systems*. John Wiley & Sons, New York, 1988. 120, 276, 302, 310, 575
17. M.H.A. Davis. The separation principle in stochastic control via Girsanov solutions. *SIAM J. Control & Optim.*, 14:176–188, 1976. 575
18. M.H.A. Davis and P. Varaiya. Dynamic programming conditions for partially observable stochastic systems. *SIAM J. Control*, 11:226–261, 1973. 468, 575, 605
19. M.H.A. Davis and R.B. Vinter. *Stochastic modelling and control*. Chapman and Hall, London, 1985. 120, 376, 410, 468, 575, 595

20. A. Drake. *Observation of a Markov process through a noisy channel*. PhD thesis, MIT, Dept. of Electrical Engineering, Cambridge, MA, USA, 1962. 576
21. E.B. Dynkin. Controlled random sequences. *Th. Probab. & Appl.*, 10:1–14, 1965. 576
22. E.B. Dynkin. *Markov processes, volume 1, volume 2*. Academic Press Inc., Publishers, New York, 1965. 73, 574
23. J.E. Eckles. Optimum maintenance with incomplete information. *Oper. Res.*, 16:1058–1067, 1968. 576
24. E. Fernandez-Gaucherand and A. Arapostathis. On partially observable Markov decision processes with an average cost criterion. In *Proceedings of the 28th IEEE Conference on Decision and Control (CDC.1989)*, pages 1267–1272, New York, 1989. IEEE, IEEE Press. 576
25. B.A. Francis. *A course in H_∞ control theory*. Number 88 in Lecture Notes in Control and Information Sciences. Springer-Verlag, New York, 1987. 560, 850
26. B.A. Francis and W.M. Wonham. The internal model principle for linear multivariable regulators. *Appl. Math & Optim.*, 2:170–194, 1975. 560
27. T.L. Gunckel. Optimum design of sampled-data systems with random parameters. Technical Report SEL TR 2102-2, Stanford Electron. Lab., Stanford, 1961. 574
28. O.L.R. Jacobs and J.W. Patchell. Caution and probing in stochastic control. *Int. J. Control*, 16:189–199, 1972. 575
29. P.D. Joseph and J.T. Tou. On linear control theory. *AIEE Trans. (Appl. Ind.)*, 80:193–196, 1961 (Sep.). 467, 574
30. D. Kendrick. *Stochastic control for economic models*. McGraw-Hill Book Co., New York, 1981. 120, 410, 575
31. K.J. Aström. Optimal control of Markov processes with incomplete state information. *J. Math. Anal. Appl.*, 10:174–205, 1965. 574
32. G. Koole. A transformation method for stochastic control problems with partial observations. *Systems & Control Lett.*, 35:301–308, 1998. 576
33. V. Krishnamurthy. *Partially observed Markov decision processes*. Cambridge University Press, Cambridge, 2016. 169, 277, 353, 575
34. P.R. Kumar and P. Varaiya. *Stochastic systems: Estimation, identification, and adaptive control*. Prentice Hall Inc., Englewood Cliffs, NJ, 1986. 376, 410, 468, 525, 575, 595, 596
35. P.R. Kumar and P. Varaiya. *Stochastic systems: Estimation, identification, and adaptive control*. Number 75 in Classics in Applied Mathematics. SIAM, Philadelphia, 2015. 468, 575, 595
36. G.B. Di Masi, L. Finesso, and G. Picci. Design of LQG controller for single point moored large tankers. *Automatica J.-IFAC*, 22:155–169, 1986. 8, 78, 575
37. Giovanni B. Di Masi and Wolfgang J. Runggaldier. An approach to discrete-time stochastic control problems under partial observations. *SIAM J. Control & Opt.*, 25:38–48, 1987. 575
38. G.E. Monahan. A survey of partially observable Markov decision processes: theory, models, and algorithms. *Manag. Sci.*, 28:1–16, 1982. 575, 576
39. J.W. Patchell and O.L.R. Jacobs. Separability, neutrality, and certainty equivalence. *Int. J. Control*, 13:337–342, 1971. 540, 575
40. D. Rhenius. Incomplete information in Markovian decision models. *Ann. Statist.*, 2:1327–1334, 1974. 575
41. W. Runggaldier. On the construction of ε-optimal strategies in partially observed MDSs. *Ann. Oper. Res.*, 28:81–96, 1991. 575
42. J. Satia and R. Lave. Markovian decision processes with probabilistic observation of states. *Manag. Sci.*, 20:1–13, 1973. 575
43. Y. Sawaragi and T. Yoshikawa. Discrete-time Markovian decision processes with incomplete state information. *Ann. Math. Statist.*, 41:78–86, 1970. 575
44. A. Sirjaev. On the theory of decision functions and control of a process of observations based on incomplete observation. *Selected Translations in Math., Stat., and Probability*, 6:162–188, 1966. 574, 576
45. R.D. Smallwood and E.J. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Oper. Research*, 21:1071–1088, 1973. 576

46. E.J. Sondik. *The optimal control of partially observable Markov processes*. PhD thesis, Stanford University, Stanford, 1971. 576
47. C. Striebel. Sufficient statistics in the optimum control of stochastic systems. *J. Math. Anal. Appl.*, 12:576–592, 1965. 532, 533, 574, 575
48. C. Striebel. *Optimal control of discrete time stochastic systems*, volume 110 of *Lecture Notes in Economic and Mathematical Systems*. Springer-Verlag, Berlin, 1975. 431, 468, 575, 595, 603, 605, 742
49. F. Tung and C. Striebel. A stochastic optimal control problem and its applications. *J. Math. Anal. Appl.*, 12:350–359, 1965. 468, 575
50. S.J. Turnovsky. Optimal control of linear systems with stochastic coefficients and additive disturbances. In S.J. Turnovsky J.D. Pitchford, editor, *Applications of control theory to economic analysis*, pages 293–335. North-Holland Publishing Company, Amsterdam, 1977. 575
51. P. Whittle. Risk-sensitive Linear/Quadratic/Gaussian control. *Adv. Appl. Prob.*, 13:764–777, 1981. 575
52. P. Whittle. Entropy-minimising and risk-sensitive control rules. *Systems & Control Lett.*, 13:1–7, 1989. 575
53. P. Whittle. *Risk-sensitive optimal control*. Wiley, Chichester, 1990. 575
54. P. Whittle and J. Kuhn. A Hamiltonian formulation of risk-sensitive Linear/Quadratic/Gaussian control. *Int. J. Control*, 43:1–12, 1986. 575
55. Hans S. Witsenhausen. Separation of estimation and control for discrete time systems. *Proc. IEEE*, 59:1557–1566, 1971. 384, 575
56. H.S. Witsenhausen. A standard form for sequential stochastic control. *Math. Systems Theory*, 7:5–11, 1973. 410, 575
57. W.M. Wonham. On a matrix Riccati equation of stochastic control. *SIAM J. Control*, 6:681–697, 1968. 574, 824, 849
58. W.M. Wonham. On the separation theorem of stochastic control. *SIAM J. Control*, 6:312–326, 1968. 535, 536, 574
59. W.M. Wonham. Optimal stochastic control. *Automatica J. IFAC*, 5:113–118, 1969. 574
60. A.A. Yuskhevich. Reduction of a controlled Markov model with incomplete data to a problem with complete information in the case of Borel state and control spaces. *Theory Probab. Appl.*, 21:153–158, 1976. 575
61. Hao Zhang. Partially observable Markov decision processes: A geometric technique and analysis. *Oper. Research*, 58:214 – 228, 2010. 575

Chapter 15
Stochastic Control with Partial Observations on an Infinite Horizon

Abstract Optimal stochastic control problems are considered for a time-invariant stochastic control system with partial observations on an infinite-horizon. Such problems can be solved by a dynamic programming method for partial observations. Both the average cost and the discounted cost functions are considered. Treated as special cases are optimal control problems for a Gaussian stochastic control system and for a finite stochastic control system.

Key words: Stochastic control. Partial observations. Infinite Horizon.

15.1 Problem Issues

Consider a stochastic control problem for a time-invariant stochastic control system with partial observations. The control law is expected to operate on a relatively long horizon. For a control engineering problem, a partial observations optimal stochastic control problem on an infinite-horizon may be suitable for the derivation of a control law. Both the case of an average cost function and of a discounted cost function are discussed in this chapter.

Control theory for the optimal stochastic control problem with partial observations and on an infinite horizon is underdeveloped. For the case of a optimal control of a Gaussian stochastic control system the number of satisfactory proofs is limited. The case of control of partially-observed finite stochastic systems is theoretically analysed but even for that case a deeper analysis is needed. The existing theoretical framework is not fully satisfactory.

The theoretical framework for optimal stochastic control problems with partial observations on an infinite-horizon with the average cost function, requires attention for several research issues. The first research issue is that an assumption is needed which guarantees that the average cost is *finite*. The appropriate condition for this is that the stochastic control system is stochastically controllable and stochastically observable by the controlled output used in the cost function. Weaker conditions of stochastic stabilizability and of stochastic detectability may be considered also. Informally stated, a stochastic control system is stochastically controllable if any measure on state set within a subset determined by the stochastic control system, can be achieved by a sequence of inputs. If the stochastic control system is stochastically controllable and stochastically observable then it has to be proven that the closed-loop system is exponentially stable and that hence the average cost is finite.

A second research issue is the time-invariance of the control law. A control law can be either time-varying or time-invariant. As discussed in Chapter 13, there exists an example, Example 13.2.19, of an optimal stochastic control problem with complete observations on an infinite-horizon for which the optimal control law is time-varying and not time-invariant. It is likely that the same conclusion holds for partially observed stochastic control problems on an infinite-horizon. There is no satisfactory control synthesis method for time-varying control laws on an infinite-horizon.

Therefore attention is in this chapter restricted to time-invariant control laws for the optimal stochastic control problem on an infinite-horizon for a time-invariant stochastic control system with partial observations.

A third research issue is whether one can obtain structural properties of the optimal control law from an analysis of the value function. Sufficient conditions for the optimal control law to have particular structural properties, can be formulated. This issue has also been discussed in previous chapters

15.2 Control of a Gaussian Stochastic Control System

15.2.1 Problem Formulation

In the first part of this chapter attention is restricted to a time-invariant Gaussian stochastic control system though with an arbitrary cost rate, not necessarily a cost rate which is a quadratic function of the state and of the input. This restriction to a Gaussian stochastic control system is due to the need for the conditional Kalman filter for the development of control theory for this case.

Problem 15.2.1. The *average-cost infinite-horizon optimal stochastic control problem for a time-invariant partially-observed Gaussian stochastic control system.* Consider a time-invariant partially-observed Gaussian stochastic control system representation,

$$\begin{aligned}
x(t+1) &= Ax(t)+Bu(t)+Mv(t), x(0)=x_0 \in G(m_{x_0},Q_{x_0}),\\
y(t) &= Cx(t)+Du(t)+Nv(t),\ \ v(t)\in G(0,I),\\
z(t) &= C_z x(t)+D_z u(t);\ \ T=\mathbb{N}=\{0,1,\ldots\},\ F^{x_0},\ F^{v}_{\infty} \text{ are independent},\\
&\quad n_y \le n_v,\ \text{rank}(N)=n_y\ \Rightarrow\ NN^T \succ 0;\\
&\quad n_z \ge n_u,\ \text{rank}(D_z)=n_u\ \Rightarrow\ D_z^T D_z \succ 0.
\end{aligned}$$

Consider further the past-output and past-input information pattern, and the set of time-invariant control laws,

$$\begin{aligned}
&\{F^y_{t-1} \vee F^u_{t-1},\ t \in T\};\\
G_{tv} &= \left\{ \begin{array}{l} g = (g_0, g_1, g_2, \ldots, g_t, \ldots |\ g_t : Y^t \times U^t \to U \\ \text{a measurable function} \end{array} \right\},\\
G_{ti} &= \{g : Y \times U \to U|\ \text{a measurable function}\}.
\end{aligned}$$

The set of control laws G_{tv} is that of time-varying control laws while that of G_{ti} is of time-invariant control laws using only output and input feedback. Later another set of control laws is introduced which is a subset of the set G_{tv}. The set of time-invariant control laws as defined above has a limited usefulness though there are examples where it can be effective.

Note that the arguments of g_t are $y^g(0:t-1)$ and $u^g(0:t-1)$ which depend on time. Note also that g_t depends on $y^g(t-1)$ but not on $y^g(t)$ because dependence on $y^g(t)$ would make the relation for $y^g(t)$ an equation.

The closed-loop system for a time-varying control law $g \in G_{tv}$ is then denoted by,

$$\begin{aligned}
x^g(t+1) &= Ax^g(t) + Bg_t(y^g(0:t-1), u^g(0:t-1)) + Mv(t),\ x^g(0) = x_0,\\
y^g(t) &= Cx^g(t) + Dg_t(y^g(0:t-1), u^g(0:t-1)) + Nv(t),\\
z^g(t) &= C_z x^g(t) + D_z g_t(y^g(0:t-1), u^g(0:t-1)),\\
u^g(t) &= g_t(y^g(0:t-1), u^g(0:t-1)).
\end{aligned}$$

In part of the chapter, a general cost function is used which is later replaced by the usual quadratic cost in terms of the state and the input process. Both of these cost rates are defined below. Define thus the infinite-horizon average cost function,

$$\begin{aligned}
&J_{ac} : G_{tv} \to \mathbb{R}_+ \cup \{+\infty\},\\
&b : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}_+,\ \text{a measurable function},\\
J_{ac}(g) &= \limsup_{t_1 \to \infty} \frac{1}{t_1} E\left[\sum_{s=0}^{t_1-1} b(x^g(s), u^g(s))\right];\ \text{define the quadratic cost rate as,}\\
b(x_b, u_b) &= (z^g)^T z^g = \begin{pmatrix} x_b \\ u_b \end{pmatrix}^T Q_{cr} \begin{pmatrix} x_b \\ u_b \end{pmatrix},\\
Q_{cr} &= \begin{pmatrix} C_z^T C_z & C_z^T D_z \\ D_z^T C_z & D_z^T D_z \end{pmatrix} \in \mathbb{R}^{(n_x+n_u) \times (n_x+n_u)};\ \text{recall } D_z^T D_z \succ 0.
\end{aligned}$$

The problem is: (a) to show there exists a control law $\overline{g} \in G_{tv}$ such that the cost is finite, $J_{ac}(\overline{g}) < +\infty$; denote the subset of control laws with finite average cost by,

$$G_{tv,fc} = \{g \in G_{tv}|\ J_{ac}(g) < \infty\};$$

(b) to show existence of a value $J^*_{ac} \in \mathbb{R}_+$ and of an optimal control law $g^*_{ac} \in G_{tv,fc}$, if it exists, such that,

$$J^*_{ac} = \inf_{g \in G_{tv,fc}} J_{ac}(g) = J_{ac}(g^*_{ac});$$

(c) if an optimal control law does not exist then determine for any $\varepsilon \in (0,\infty)$ an ε-optimal control law $g_\varepsilon \in G_{tv,f}$ such that $J_{ac}^* < J_{ac}(g_\varepsilon) < J_{ac}^* + \varepsilon$.

Definition 15.2.2. Consider the optimal stochastic control problem of Problem 15.2.1. Recall that $NN^T \succ 0$ and $D_z^T D_z \succ 0$. Define the matrices,

$$\begin{aligned} &A_c,\ C_c,\ A_f,\ M_f \in \mathbb{R}^{n_x \times n_x}, \\ &A_c = A - B(D_z^T D_z)^{-1} D_z^T C_z,\ \ C_c^T C_c = C_z^T C_z - C_z^T D_z (D_z^T D_z)^{-1} D_z^T C_z, \\ &A_f = A - MN^T (NN^T)^{-1} C,\ \ M_f M_f^T = MM^T - MN^T (NN^T)^{-1} NM^T. \end{aligned}$$

(a)The *controllability and the observability conditions* of Problem 15.2.1 are said to hold if: (1) (A,B) is a controllable pair; (2) (A_c,C_c) is an observable pair; (3) (A,C) is an observable pair; and (4) (A_f,M_f) is a supportable pair.
If $MN^T = 0$ then $A_f = A$ and M_f can be chosen as $M_f = M$ hence $(A_f,M_f) = (A,M)$ and condition (4) is then equivalent to (A,M) being a supportable pair. Similary, if $D_z^T C_z = 0$ then $A_c = A$ and C_c can be chosen as $C_c = C_z$ hence $(A_c,C_c) = (A,C_z)$.

(b)The *stabilizability and the detectability conditions* of Problem 15.2.1 are said to hold if: (1) (A,B) is a stabilizable pair; (2) (A_c,C_c) is a detectable pair; (3) (A,C) is a detectable pair; and (4) (A_f,M_f) is a supportable-stable pair.
If either $MN^T = 0$ or $D_z^T C_z = 0$ or both then implications similar to case (a) hold.

It follows by a result for linear control systems that the controllability and the observability conditions of Def. 15.2.2.(a) imply that the stabilizability and the detectability conditions (b) hold.

See the text below Assumption 13.2.9 for an interpretation of the above assumption.

15.2.2 Stochastic Realization

The procedure for optimal stochastic control problems with partial observations outlined in Section 14.2 will be followed. First the filtering problem is treated. As described in Theorem 14.4.2 there exists a time-varying Kalman filter for the closed-loop system and for any control law. If the assumptions of Def. 15.2.2 hold, the solution of the Riccati recursion for the variance of the state estimate converges to the solution of the Filtering Algebraic Riccati Equation (FARE), see Theorem 8.5.3. Consequently, the time-varying Kalman gain function converges to the Kalman gain matrix which is a function of the solution of Filter Algebraic Riccati Equation (FARE), see Def. 8.5.4.

Because in the average-cost function the average cost depends on the long term behavior of the closed-loop system, the average cost does not depend on the initial transient behavior of the filter system during which period the convergence of the solution of the Riccati recursion to the solution of FARE takes place. Therefore the average cost depends only on the time-invariant Kalman gain.

Below the assumption is used that the variance of the initial state x_0 equals the solution of the algebraic Filter Riccati equation (FARE), see Proposition 8.5.5.(c). It then follows that the Filter Riccati recursion equals for all time the solution of the algebraic Filter Riccati equation and the Kalman gain function equals the Kalman gain of the time-invariant Kalman filter. In terms of formulas,

$$\begin{aligned} Q_f^* &= f_{FARE}(Q_f^*) \in \mathbb{R}_{pds}^{n_x \times n_x},\ \text{spec}(A + K(Q_f^*)C) \subset \mathrm{D}_o, \\ & \quad x_0 \in G(0, Q_{x_0}),\ Q_{x_0} = Q_f^*, \\ Q_f(t+1) &= f_{FARE}(Q_f(t)),\ Q_f(0) = Q_{x_0} = Q_f^* \\ & \quad \Rightarrow Q_f(t) = Q_f^*,\ K(Q_f(t)) = K(Q_f^*),\ \forall\, t \in T. \end{aligned}$$

The transient effect of the convergence of the solution of the Riccati recursion to the solution of FARE can be proven to have no effect on the average cost though this will not be proven here.

The considered stochastic control system is time-invariant. From Theorem 14.4.2 follows the result on the time-varying Kalman filter in the case of a time-invariant stochastic control system.

Theorem 15.2.3. *Consider the optimal stochastic control Problem 15.2.1. Assume that $Q_{x_0} = Q_f^*$ as discussed above. Assume that the stabilizability and the detectability conditions hold of Def. 15.2.2.(b). Consider a control law $g \in G_{tv,fc}$.*

(a)Then the conditional distribution of the state process conditioned in the past outputs and past inputs is Gaussian with,

$$\begin{aligned} & E[\exp(i w_x^T x^g(t)) | F_{t-1}^{y^g} \vee F_{t-1}^{u^g}] \\ &= \exp(i w_x^T \hat{x}^g(t) - \frac{1}{2} w_x^T Q_f^* w_x),\ \forall\, w_x \in \mathbb{R}^{n_x}, \end{aligned}$$

of which the conditional density function is denoted by,

$$p_G(.;\ \hat{x}^g(t), Q_f^*);$$

$$\begin{aligned} \hat{x}^g(t+1) &= A\hat{x}^g(t) + Bu^g(t) + K(Q_f^*)[y^g(t) - C\hat{x}^g(t) - Du^g(t)],\ \hat{x}^g(0) = m_{x_0}, \\ Q_f^* &= f_{FARE}(Q_f^*) \in \mathbb{R}_{pds}^{n_x \times n_x},\ \text{spec}(A + K(Q_f^*)C) \subset \mathrm{D}_o; \\ K(Q_f^*) &= [AQ_f^*C^T + MN^T][CQ_f^*C^T + NN^T]^{-1}. \end{aligned}$$

(b)Define the innovation process *by the formulas,*

$$\begin{aligned} \overline{v}^g(t) &= y^g(t) - C\hat{x}^g(t) - Du^g(t),\ \overline{v}^g : \Omega \times T \to \mathbb{R}^{n_y}, \\ & \quad \overline{v}^g(t) \in G(0, Q_{\overline{v}}),\ Q_{\overline{v}} = CQ_f^*C^T + NN^T; \\ & \quad \overline{v}^g(t) \text{ is independent of } F_{t-1}^{y^g} \vee F_{t-1}^{u^g}. \end{aligned}$$

The innovation process is a Gaussian white noise process.

(c)The stochastic realization of the Gaussian stochastic control system with respect to the information pattern, or, equivalently, the filter system driven by the innovations process, used subsequently in stochastic control, is specified by,

$$\begin{aligned} \hat{x}^g(t+1) &= A\hat{x}^g(t) + Bu^g(t) + K(Q_f^*)\overline{v}^g(t),\ \hat{x}^g(0) = m_{x_0}; \\ & \quad x^g(t) \text{ is } F_{t-1}^{y^g} \vee F_{t-1}^{u^g} \text{ measurable} \end{aligned}$$

(d) *The following conditional characteristic function is needed in the dynamic programming procedure.*

$$\begin{aligned}
&E[\exp(iw_x^T \hat{x}^g(t+1)) |\, F_{t-1}^{y^g} \vee F_{t-1}^{u^g}] \\
&= \exp(iw_x^T [A\hat{x}^g(t) + Bu^g(t)] - \frac{1}{2} w_x^T K(Q_f^*) Q_{\overline{v}}(Q_f^*) K(Q_f^*)^T w_x), \\
&\forall\, w_x \in \mathbb{R}^{n_x}, \text{ of which the conditional density is denoted by,} \\
&p_G(.;\ [A\hat{x}^g(t) + Bu^g(t)], K(Q_f^*) Q_{\overline{v}}(Q_f^*) K(Q_f^*)^T), \\
&Q_{K(Q_f^*)\overline{v}} = K(Q_f^*) Q_{\overline{v}}(Q_f^*) K(Q_f^*)^T.
\end{aligned}$$

Proof. This result follows from Theorem 14.4.2 and Proposition 8.5.5.(c) using the time-invariance of the Kalman filter. □

15.2.3 Dynamic Programming

According to Procedure 14.2.3 the next step is to solve the optimal stochastic control problem now for the stochastic realization with respect to the information pattern. First the projection of the cost function is calculated.

Below the variance of the Kalman filter of stochastic control is denoted by $Q_f \in \mathbb{R}_{pds}^{n_x \times n_x}$. If $Q_{x_0} = Q_f^*$ where Q_f^* is the solution of the filter algebraic Riccati equation, then this is stated explicitly. The notation of Q_f allows for a certain generality.

Proposition 15.2.4. *Consider the optimal stochastic control Problem 15.2.1. Assume that $Q_{x_0} = Q_f$ as discussed above. Assume that the stabilizability and the detectability conditions of Def. 15.2.2.(b) hold. Define the projection of the cost rate as the function,*

$$\overline{b}(x_b, Q_f, u_b) = \int b(w_x, u_b)\ f_G(dw_x; (x_b, Q_f)).$$

Note that the functions b and $\overline{b}$ have different arguments!

Then the average cost can be written as,

$$\begin{aligned}
J_{ac}(g) &= \limsup_{t_1 \rightarrow} \frac{1}{t_1} E\left[\sum_{s=0}^{t_1-1} \overline{b}(\hat{x}^g(s), Q_f, u^g(s))\right], \\
\overline{b}(\hat{x}^g(s), Q_f, u^g(s)) &= E[b(x^g(s), u^g(s)) | F_{s-1}^{y^g} \vee F_{s-1}^{u^g}],\ \forall\, s \in T.
\end{aligned}$$

Proof. Note that for all $s \in T$,

$$\begin{aligned}
&E[b(x^g(s), u^g(s))] = E[E[b(x^g(s), u^g(s)) | F_{s-1}^{y^g} \vee F_{s-1}^{u^g}]] \\
&= E[\int b(x_b, u^g(s)) f_G(dx_b; \hat{x}^g(s), Q_f)] = E[\overline{b}(\hat{x}^g(s), Q_f(s), u^g(s))]; \\
&E\left[\sum_{s=0}^{t_1-1} b(x^g(s), u^g(s))\right] = E\left[\sum_{s=0}^{t_1-1} \overline{b}(\hat{x}^g(s), Q_f(s), u^g(s))\right].
\end{aligned}$$

Problem 15.2.5. Consider the Gaussian stochastic control system in terms of the stochastic realization with respect to the information pattern, see Theorem 15.2.3.(c), and the projected cost rate of Proposition 15.2.4,

$$\hat{x}^g(t+1) = A\hat{x}^g(t) + Bu^g(t) + K(Q_f)\overline{v}^g(t), \; \hat{x}^g(0) = 0,$$
$$J_{ac}(g) = \lim_{t_1 \to \infty} \frac{1}{t_1} \, E\left[\sum_{s=0}^{t_1-1} \overline{b}(\hat{x}^g(s), Q_f, u^g(s))\right].$$

Solve the optimal control problem,

$$\inf_{g \in G_{ti}} \, J_{ac}(g).$$

The above formulated problem is now almost a stochastic control problem for a stochastic control system with complete observations because the input at any time is measurable with respect to the information pattern at the time and because the state of the filter system, $(\hat{x}^g(t), Q_f)$, is measurable with respect to the information pattern at the time. However, note that the problem is not precisely one with complete observations because the set of control laws G_{tv} is defined in Problem 15.2.1 hence may depend on the past outputs and the past inputs of the system of that problem statement. Therefore the dynamic programming procedure for average-cost optimal stochastic control with partial observations differs slightly from that with complete observations.

Procedure 15.2.6 Procedure of dynamic programming with partial observations on an infinite horizon and with average cost. *Consider the optimal stochastic control Problem 15.2.5.*

Calculate according to the following steps.

1. *Determine the value and a value function satisfying the following dynamic programming equation of average cost,*

$$(J_{ac}^*, V), \; J_{ac}^* \in \mathbb{R}_+, \; V : \mathbb{R}^{n_x} \times \mathbb{R}_{pds}^{n_x \times n_x} \to \mathbb{R}_+,$$
$$\begin{aligned} & J_{ac}^* + V(\hat{x}_V, Q_f) \\ & = \inf_{u_V \in U(\hat{x}_V, Q_f)} \left\{ \overline{b}(\hat{x}_V, Q_f, u_V) + E[V(\hat{x}(t+1), Q_f(t+1)) | \, F_{t-1}^y \vee F_{t-1}^u] \right\} \\ & = \inf_{u_V \in U(\hat{x}_V, Q_f)} \left\{ \int b(w,u) \; p_G(w; \; \hat{x}_V, \; Q_f) dw + \right. \\ & \quad \left. + \int V(w, f_{FARE}(Q_f)) \; p_G(w; \; [A\hat{x}_V + Bu_V], Q_{K(Q_f)\overline{v}}(Q_f)) dw \right\}, \\ & \quad \forall \; (\hat{x}_V, Q_f) \in \mathbb{R}^{n_x} \times \mathbb{R}_{pds}^{n_x \times n_x}. \end{aligned}$$

2. *If for every $(\hat{x}_V, Q_V)$ there exists an input $u_V^* \in U(\hat{x}_V, Q_V)$ which attains the infimum in the infimization of Step (1),*

$$\begin{aligned}
&\forall\ (\hat{x}_V, Q_f) \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_x \times n_x}_{pds},\ \exists\ u^* \in U(\hat{x}_V, Q_f),\ \textit{such that,} \\
&\int b(w, u^*)\ p_G(w;\ \hat{x}_V,\ Q_f) dw + \\
&+ \int V(w, f_{FARE}(Q_f))\ p_G(w;\ [A\hat{x}_V + Bu^*], Q_{K(Q_f)\overline{v}}(Q_f)) dw, \\
= &\inf_{u_V \in U(\hat{x}_V, Q_f)} \left\{ \int b(w, u_V)\ p_G(w;\ \hat{x}_V,\ Q_f) dw + \right. \\
&\left. + \int V(w, f_{FARE}(Q_f))\ p_G(w;\ [A\hat{x}_V + Bu_V], Q_{K(Q_f)\overline{v}}(Q_f)) dw \right\}, \\
&\textit{then define } \overline{g}^*(\hat{x}_V, Q_f) = u^*_V,\ \ \overline{g}^* : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x \times n_x}_{spd} \to \mathbb{R}^{n_u}.
\end{aligned}$$

3. *Check whether $\overline{g}^*$ is a measurable function. Proceed if it is, stop otherwise.*
4. *Output $(J^*_{ac},\ V,\ \overline{g}^*)$.*

Theorem 15.2.7. *Consider Problem 15.2.1. Consider the transformed Problem 15.2.5 in terms of the filter system and the projected cost rate.*

Assume that: (1) the stabilizability and the detectability assumptions of Def. 15.2.2.(b) hold; (2) the set of control laws achieving a finite average cost on the infinite horizon is nonempty; (3) there exists a tuple $(J^*_{ac}, V, \overline{g}^*)$ which is a solution of the dynamic programming equation of Procedure 15.2.6;

$$(4)\ \ \forall\ g \in G_{ac,fc},\ \forall\ t \in T,\ E|V(x^g(t), Q^*_f)| < \infty;$$

$$(5)\ \ \forall\ g \in G_{ti,f},\ \lim_{t_1 \to \infty} \frac{1}{t_1}\ E[V(\hat{x}^g(t_1), Q^*_f)] = 0.$$

(a)Then J^*_{ac}, determined by the dynamic programming procedure, is a lower bound of the average cost for all control laws achieving a finite cost

$$J^*_{ac} \leq J_{ac}(g),\ \forall\ g \in G_{ti,fc}.$$

(b)Assume, in addition to (a), that for any $(x_V, Q_f) \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_x \times n_x}_{spd}$ the infimum in Procedure 15.2.6.Step2 is attained, there exist an element $u^*_V \in U(x_V, Q_f) \subseteq \mathbb{R}^{n_u}$ which is a minimizer. Assume that $\overline{g}^*$ is a measurable function. Then the particular control law $\overline{g}^*$ is an optimal control law of Problem 15.2.5 and satisfies,

$$J^*_{ac} = J_{ac}(\overline{g}^*) = \inf_{g \in G_{ac,fc}} J_{ac}(g).$$

The optimal control law is specified by the equations,

$$\begin{aligned}
\hat{x}(t+1) &= A\hat{x}(t) + B\overline{g}^*(\hat{x}(t), Q_f) + K(Q_f)[y(t) - C\hat{x}(t) - D\overline{g}^*(\hat{x}(t), Q_f)], \\
&\quad \hat{x}(0) = m_{x_0}, \\
u^*(t) &= \overline{g}^*(\hat{x}(t), Q_f).
\end{aligned}$$

Proof. The proof is similar to the proof of Theorem 13.2.7. □

Note that the optimal control law $g^*(x_V, Q^*_f)$ depends on the state $\hat{x}^g$ of the filter system and on the solution of the algebraic filter Riccati equation. Moreover, the control law g^* is time invariant. Due the definition of $\hat{x}$, the control law is adapted to the filtration $\{F^y_{t-1} \vee F^u_{t-1},\ t \in T\}$. Note that Q^*_f does not depend explicitly on either the output process nor the input process though it depends on the system matrices.

15.2.4 Quadratic Cost Rate

The reader finds in this section the LQG optimal control law and its proof of optimality.

Proposition 15.2.8. *Consider the optimal stochastic control Problem 15.2.1. Assume that $Q_{x_0} = Q_f$ as discussed above. Assume that the stabilizability and the detectability conditions of Def. 15.2.2.(b) hold.*

In case the cost function is quadratic, one obtains the following projection of the quadratic cost rate, using the definition of Q_{cr} from Problem 15.2.1,

$$\begin{aligned}
J_{ac}(g) &= \limsup_{t_1 \to \infty} \frac{1}{t_1} E\left[\sum_{s=0}^{t_1-1} \hat{z}^g(s)^T \hat{z}^g(s) + \mathrm{tr}(C_z\, Q_f^* C_z^T)\right] \\
&= \limsup_{t_1 \to} \frac{1}{t_1} E\left[\sum_{s=0}^{t_1-1} \begin{pmatrix} \hat{x}^g(s) \\ u^g(s) \end{pmatrix}^T \begin{pmatrix} C_z^T C_z & C_z^T D_z \\ D_z^T C_z & D_z^T D_z \end{pmatrix} \begin{pmatrix} \hat{x}^g(s) \\ u^g(s) \end{pmatrix} + \mathrm{tr}(C_z\, Q_f^* C_z^T)\right].
\end{aligned}$$

Proof. Define,

$$\begin{aligned}
\hat{z}^g(t) &= E[z^g(t)|F_{t-1}^{y^g} \vee F_{t-1}^{u^g}] = C_z \hat{x}^g(t) + D_z u^g(t); \\
& \quad z^g(t) - \hat{z}^g(t) = C_z(x^g(t) - \hat{x}^g(t)); \\
& \quad E[z^g(t)^T z^g(t)] \\
&= E[(\hat{z}^g(t) + (z^g(t) - \hat{z}^g(t))^T (\hat{z}^g(t) + (z^g(t) - \hat{z}^g(t))] \\
&= E[\hat{z}^g(t)^T \hat{z}^g(t)] + E[(z^g(t) - \hat{z}^g(t))^T (z(t) - \hat{z}^g(t))] + \\
& \quad +2E[\hat{z}^g(t)^T (z^g(t) - \hat{z}^g(t))] \\
&= E[\hat{z}^g(t)^T \hat{z}^g(t)] + E[(C_z(x^g(t) - \hat{x}^g(t)))^T (C_z(x^g(t) - \hat{x}^g(t)))] + \\
& \quad +2E[\hat{z}^g(t)^T E[(z^g(t) - \hat{z}^g(t))|F_{t-1}^{y^g} \vee F_{t-1}^{u^g}]] \\
&= E[\hat{z}^g(t)^T \hat{z}^g(t)] + E[(x^g(t) - \hat{x}^g(t))^T C_z^T C_z (x^g(t) - \hat{x}^g(t))] \\
&= E[\hat{z}^g(t)^T \hat{z}^g(t)] + \mathrm{tr}(E[(x^g(t) - \hat{x}^g(t))(x^g(t) - \hat{x}^g(t))^T] C_z^T C_z) \\
&= E[\hat{z}^g(t)^T \hat{z}^g(t)] + \mathrm{tr}(Q_f(t) C_z^T C_z) = E[\hat{z}^g(t)^T \hat{z}^g(t)] + \mathrm{tr}(C_z Q_f(t) C_z^T); \\
\hat{z}^g(t)^T \hat{z}^g(t) &= \begin{bmatrix} \hat{x}^g(t) \\ u^g(t) \end{bmatrix} \begin{bmatrix} C_z^T C_z & C_z^T D_z \\ D_z^T C_z & D_z^T D_z \end{bmatrix} \begin{bmatrix} \hat{x}^g(t) \\ u^g(t) \end{bmatrix} = \begin{bmatrix} \hat{x}^g(t) \\ u^g(t) \end{bmatrix} Q_{cr} \begin{bmatrix} \hat{x}^g(t) \\ u^g(t) \end{bmatrix}^T.
\end{aligned}$$

Then,

$$\begin{aligned}
J_{ac}(g) &= \limsup_{t_1 \to \infty} \frac{1}{t_1} E\left[\sum_{s=0}^{t_1-1} \left(\hat{z}^g(s)^T \hat{z}^g(s) + \mathrm{tr}(C_z Q_f(s) C_z^T)\right)\right] \\
&= \limsup_{t_1 \to \infty} \frac{1}{t_1} E\left[\sum_{s=0}^{t_1-1} \hat{z}^g(s)^T \hat{z}^g(s)\right] + \limsup_{t_1 \to \infty} \frac{1}{t_1} E\left[\sum_{s=0}^{t_1-1} \mathrm{tr}(C_z Q_f(s) C_z^T)\right] \\
&= \limsup_{t_1 \to \infty} \frac{1}{t_1} E\left[\sum_{s=0}^{t_1-1} \left(\hat{z}^g(s)^T \hat{z}^g(s)\right)\right] + \mathrm{tr}(C_z Q_f^* C_z^T) \\
&= \limsup_{t_1 \to \infty} \frac{1}{t_1} E\left[\sum_{s=0}^{t_1-1} \left(\hat{z}^g(s)^T \hat{z}^g(s) + \mathrm{tr}(C_z Q_f^* C_z^T)\right)\right].
\end{aligned}$$

The last equality follows from Theorem 22.2.2 and from Proposition 17.5.9. Here the limit $Q_f(\infty) = \lim_{t\to\infty} Q_f(t)$ exists and it equals the unique solution of the Filter Algebraic Riccati Equation. From now on the limit is denoted as the matrix Q_f^*. □

Definition 15.2.9. *Time-invariant LQG optimal control law*.
Consider Problem 15.2.1. Assume that the stabilizability and detectability conditions of Def. 15.2.2.(b) hold. Then it follows from Theorem 22.2.2 and from Theorem 22.2.4 that there exist matrices Q_f^* and Q_c^* satisfying the following algebraic Riccati equations with side conditions such that,

$$\begin{aligned}
Q_f^* &= AQ_f^*A^T + MM^T + \\
&\quad -[AQ_f^*C^T + MN^T][CQ_f^*C^T + NN^T]^{-1}[AQ_f^*C^T + MN^T]^T, \\
&\quad Q_f^* \in \mathbb{R}_{pds}^{n_x \times n_x},\ \mathrm{spec}(A - K(Q_f^*)C) \subset \mathrm{D}_o; \\
K(Q_f^*) &= [AQ_f^*C^T + MN^T][CQ_f^*C^T + NN^T]^{-1} \in \mathbb{R}^{n_x \times n_y}, \\
Q_{\bar{v}}(Q_f^*) &= CQ_f^*C^T + NN^T; \\
Q_c^* &= A^T Q_c^* A + C_z^T C_z + \\
&\quad -[A^T Q_c^* B + C_z^T D_z][B^T Q_c^* B + D_z^T D_z]^{-1}[A^T Q_c^* B + C_z^T D_z]^T, \\
&\quad Q_c^* \in \mathbb{R}_{pds}^{n_x \times n_x},\ \mathrm{spec}(A + BF(Q_c^*)) \subset \mathrm{D}_o; \\
F(Q_c^*) &= -[B^T Q_c^* B + D_z^T D_z]^{-1}[A^T Q_c^* B + C_z^T D_z]^T \in \mathbb{R}^{n_u \times n_x}.
\end{aligned}$$

Define the *time-invariant LQG optimal control law*, denoted by $g_{LQG,PO,AC}^* \in G$, of Problem 15.2.1 with a quadratic cost rate, as a dynamic control law defined by the equations,

$$\begin{aligned}
&\quad \bar{x} : \Omega \times T \to \mathbb{R}^{n_x}, \\
\bar{x}(t+1) &= A\bar{x}(t) + Bu(t) + K(Q_f^*)[y(t) - C\bar{x}(t) - Du(t)],\ \bar{x}(0) = m_{x_0}, \\
u(t) &= F(Q_f^*)\bar{x}(t); \\
&\quad \text{equivalently, after substitution of the control law,} \\
\bar{x}(t+1) &= [A + BF(Q_c^*) - K(Q_f^*)C - K(Q_f^*)DF(Q_c^*)]\bar{x}(t) + K(Q_f^*)y(t), \\
u(t) &= F(Q_c^*)\bar{x}(t); \\
z(t) &= (C_z + D_z F(Q_c^*))\bar{x}(t), \\
&\quad g_{LQG,PO,AC}^*(\hat{x}_V, Q_f^*) = F(Q_c^*)\hat{x}_V.
\end{aligned}$$

The associated closed-loop control system consisting of the Gaussian stochastic control system of Problem 15.2.1 and of the LQG optimal control law is described by the system representation,

$$\begin{aligned}
e(t) &= x(t) - \bar{x}(t), \; e : \Omega \times T \to \mathbb{R}^{n_x}, \\
x_{\bar{x},e}(t) &= \begin{pmatrix} \bar{x}(t) \\ e(t) \end{pmatrix}, \; x_{\bar{x},e} : \Omega \times T \to \mathbb{R}^{2n_x}, \\
x_{\bar{x},e}(t+1) &= \begin{pmatrix} A + BF(Q_c^*) & K(Q_f^*)C \\ 0 & A - K(Q_f^*)C \end{pmatrix} x_{\bar{x},e}(t) + \begin{pmatrix} KN \\ M - KN \end{pmatrix} v(t) \\
&= A_{\bar{x},e} x_{\bar{x},e}(t) + M_{\bar{x},e} v(t), \;\; x_{\bar{x},e}(0) = \begin{pmatrix} m_{x_0} \\ x_0 - m_{x_0} \end{pmatrix}.
\end{aligned}$$

Note that the notation $\bar{x}$ is used in the above definition for a process that satisfies a recursion similar to the Kalman filter in the time-varying control law. If the variance of the initial state Q_{x_0} is chosen equal to Q_f^* then the time-varying Kalman filter is such that $Q_f(t) = f_{FARE}(Q_f) = Q_f^*$ for all $t \in T$ hence $K(t, Q_f(t)) = K(t, Q_f^*) = K(Q_f^*)$. In that case $\bar{x}(t) = \hat{x}^g(t)$ for all $t \in T$. But those formulas do not hold if the initial variance Q_{x_0} is not equal to Q_f^*.

Theorem 15.2.10. *Consider the optimal stochastic control problem of Problem 15.2.1 with the quadratic cost rate. Assume that the stabilizability and the detectability conditions of Definition 15.2.2.(b) hold. Consider the average-cost LQG optimal control law of Def. 15.2.9.*

(a)*There exist solutions of the filter algebraic Riccati equation with side conditions of Def. 15.2.9 and of the control algebraic Riccati equation with side conditions of the same definition. Then the LQG optimal control law of that definition is well defined.*

(b)*The closed-loop system associated with the stochastic control system and the LQG optimal control law of Def. 15.2.9 is exponentially stable with*

$$\mathrm{spec}(A_{\bar{x},e}) = \mathrm{spec}(A + BF(Q_c^*)) \cup \mathrm{spec}(A - K(Q_f^*)C) \subset \mathrm{D}_o.$$

(c)*The average cost of the candidate LQG optimal control law equals the expression,*

$$J_{ac}(g^*_{LQG,po,ac}) = \mathrm{tr}(C_z Q_f^* \, C_z^T) + \mathrm{tr}(Q_c^* \; K(Q_f^*) \; Q_{\bar{v}}(Q_f^*) \; K(Q_f^*)^T).$$

Hence the average cost of this control law is finite and the set of control laws, which achieve a finite cost, is nonempty.

(d)*The following expressions for the value and for the value function are a solution of the dynamic programming equation of average cost projected on the information structure,*

$$\begin{aligned}
J_{ac}^* &= \mathrm{tr}(C_z Q_f^* \, C_z^T) + \mathrm{tr}(Q_c^* \; K(Q_f^*) \; Q_{\bar{v}}(Q_f^*) \; K(Q_f^*)^T), \\
V(x_V, Q_V) &= x_V^T Q_c^* x_V.
\end{aligned}$$

(e)*The following relations both hold, with J_{ac}^* determined by the dynamic programing procedure,*

$$J_{ac}^* \leq J(g), \;\; \forall \, g \in G_{tv,fc}, \tag{15.1}$$

$$J_{ac}^* = J_{ac}(g^*). \tag{15.2}$$

Thus g^* is an optimal control law.
The average-cost optimal control LQG law $g^*_{LQG,PO,AC}$ is a linear function of the state of the observation-based realization of the stochastic control system and it is optimal over the set of all measurable nonlinear control laws achieving finite cost.

The value is determined by two cost rates. The term $\mathrm{tr}(C_z\, Q_f^*\, C_z^T)$ accounts for the cost of filtering determined by the variance of the state error Q_f^* as measured in terms of the controlled output z by the matrix C_z. The second term, $\mathrm{tr}(Q_c^*\, KQ_{\bar{v}}K^T)$, is the cost of control which depends on the matrix Q_c^* which is the solution of the control algebraic Riccati equation and on the variance of the innovation process affecting the observation-based realization of the stochastic control system. Note the relation of the above two terms with the value of the infinite-horizon optimal stochastic control problem with complete observations, see Theorem 13.2.15. The value of the latter theorem is related but slightly different, but the structure of the terms is similar. Only the variances of the noise components differ.

Proof. Proof of Theorem 15.2.10.
(a) The existence of the solutions of the algebraic Riccati equations follows from Theorem 22.2.2 and from Theorem 22.2.4. Then the LQG optimal control law is well defined.
(b) The formulas of the closed-loop system follow from Def. 15.2.9. Because the closed-loop system is linear with a structured system matrix, the spectrum is the union of the two spectra. Because of Def. 15.2.9, the eigenvalues of $A+BF(Q_c^*)$ and of $A-K(Q_f^*)C$ are inside the open unit disc of the complex plane.
(c) It follows from Theorem 22.2.2 and from Theorem 22.2.4 that the limits of the filter Riccati recursion and of the control Riccati recursion exist and are solutions of the corresponding algebraic Riccati equations,

$$\begin{aligned}
Q_f(t+1) &= (A-KC)Q_f(t)(A-KC)^T + (M-KN)(M-KN)^T,\ Q_f(0) = Q_{x_0},\\
Q_c(t-1) &= (A+BF)Q_c(t)(A+BF)^T + (C_z+D_zF)(C_z+D_zF)^T,\\
&\quad Q_c(0) = C_z^T C_z,\\
Q_f^* &= Q_f(\infty) = \lim_{t\to\infty} Q_f(t),\ Q_c^* = Q_c(-\infty) = \lim_{t\to-\infty} Q_c(t),\\
K(Q_f^*) &= K(Q_f(\infty)) = \lim_{t\to\infty} K(Q_f(t)),\\
F(Q_c^*) &= F(Q_c(-\infty)) = \lim_{t\to-\infty} F(Q_c(t)),\\
Q_{\bar{v}}(Q_f^*) &= \lim_{t\to\infty} Q_{\bar{v}}(Q_f(t)) = \lim CQ_f(t)C^T + NN^T = CQ_f^*C^T + NN^T.
\end{aligned}$$

It follows from Theorem 14.4.12 that the cost on the finite horizon $T(0:t-1) = \{0,1,\ldots,t-1\}$ is,

$$\begin{aligned}
J_t(g) &= \sum_{s=0}^{t-1} \mathrm{tr}(C_z\, Q_f(s)\, C_z^T) + m_{x_0}^T Q_{x_0} m_{x_0} + \mathrm{tr}(Q_c(0)Q_{x_0}) + \\
&\quad + \sum_{s=0}^{t-2} \mathrm{tr}(Q_c(s+1)\, K(Q_f(s))Q_{\bar{v}}(Q_f(s))K(Q_f(s))^T).
\end{aligned}$$

The cost on the infinite horizon of this dynamic control law is thus,

$$\begin{aligned}
J_{ac}(g) &= \lim_{t\to\infty} \frac{1}{t}\, J_t(g) \\
&= \lim_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} \mathrm{tr}(Q_f(s)\, C_z^T C_z) + \lim_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-2} \mathrm{tr}(Q_c(s+1)\, K Q_{\overline{v}}(Q_f(s)) K^T) + \\
&\quad + \lim\, [m_{x_0}^T Q_c(0) m_{x_0} + \mathrm{tr}(Q_c(0) Q_{x_0})]/t \\
&= \mathrm{tr}(Q_f(\infty)\, C_z^T C_z) + \mathrm{tr}(Q_c(-\infty)\, K Q_{\overline{v}}(Q_f(\infty)) K^T),\ \text{by Proposition 17.5.9 and} \\
&\quad \text{using the exponential-asymptotic convergence rate of the Riccati sequences,} \\
&= \mathrm{tr}(C_z Q_f^*\, C_z^T) + \mathrm{tr}(Q_c^*\, K(Q_f^*) Q_{\overline{v}}(Q_f^*) K(Q_f^*)^T).
\end{aligned}$$

(d) It is proven that the value and the value function, both defined in the theorem statement are a solution of the dynamic programming equation of Theorem 15.2.10. Note that,

$$\begin{aligned}
& \begin{pmatrix} x_V \\ u_V \end{pmatrix}^T Q_{cr} \begin{pmatrix} x_V \\ u_V \end{pmatrix} + \mathrm{tr}(C_z Q_V C_z^T) + \\
& + E\left[V(f_{FS}(x_V, Q_V, u_V), f_{FR}(Q_V)) |\, F_{t-1}^{y^g} \vee F_{t-1}^{u} \right], \\
= & \begin{pmatrix} x_V \\ u_V \end{pmatrix}^T Q_{cr} \begin{pmatrix} x_V \\ u_V \end{pmatrix} + \mathrm{tr}(C_z Q_V\ C_z^T) + \\
& + E[(Ax_V + Bu_V + K\overline{v}(t))^T Q_c^* (Ax_V + Bu_V + K\overline{v}(t)) | F_{t-1}^{y^g} \vee F_{t-1}^{u^g}] \\
& \text{by Theorem 15.2.3, using that } f_{FR}(Q_f^*) = Q_f^*, \\
& \text{and noting that } \overline{v}(t) \in G(0, Q_{\overline{v}}(Q_f^*)), \\
= & \begin{pmatrix} x_V \\ u_V \end{pmatrix}^T H \begin{pmatrix} x_V \\ u_V \end{pmatrix} + \mathrm{tr}(C_z Q_V\ C_z^T) + \mathrm{tr}(Q_c^*\ K Q_{\overline{v}} K^T) \\
& H = \begin{pmatrix} H_{xx} & H_{xu} \\ H_{xu}^T & H_{uu} \end{pmatrix}, \\
& H_{xx} = A^T Q_c^* A + C_z^T C_z,\ H_{xu} = A^T Q_c^* B + C_z^T D_z,\ H_{uu} = B^T Q_c^* B + D_z^T D_z, \\
= & \begin{pmatrix} x_V \\ u_V - u_V^* \end{pmatrix}^T \overline{H} \begin{pmatrix} x_V \\ u_V - u_V^* \end{pmatrix} + \mathrm{tr}(C_z Q_V\ C_z^T) + \mathrm{tr}(Q_c^*\ K Q_{\overline{v}} K^T) \\
& \overline{H} = \begin{pmatrix} H_{xx} - H_{xu} H_{uu}^{-1} H_{xu}^T & 0 \\ 0 & H_{uu} \end{pmatrix}, \text{ by completion of squares,} \\
& \inf_{u_V \in U(x_V, Q_V)} E[\begin{pmatrix} x_V \\ u_V \end{pmatrix}^T Q_{cr} \begin{pmatrix} x_V \\ u_V \end{pmatrix} + \\
& + V(f_{FS}(x_V, Q_V, u_V, \overline{v}(t)), f_{FR}(Q_V)) | F_{t-1}^{y^g} \vee F_{t-1}^{u^g}] \\
= & \inf_{u_V \in U(x_V, Q_V)} \begin{pmatrix} x_V \\ u_V - u_V^* \end{pmatrix}^T \begin{pmatrix} Q_c^* & 0 \\ 0 & D_z^T D_z \end{pmatrix} \begin{pmatrix} x_V \\ u_V - u_V^* \end{pmatrix} + \\
& + \mathrm{tr}(C_z Q_V\ C_z^T) + \mathrm{tr}(Q_c^*\ K Q_{\overline{v}} K^T) \\
= & x_V^T Q_c^* x_V^T + \mathrm{tr}(Q_V\ C_z^T C_z) + \mathrm{tr}(Q_c^*\ K Q_{\overline{v}} K^T) = V(x_V, Q_V) + J_{ac}^*, \\
u^* = & g^*(x_V, Q_V) = F(Q_c^*) x_V.
\end{aligned}$$

Thus J_{ac}^* and V satisfy the dynamic programming equation of Procedure 15.2.6. It then follows from Theorem 15.2.7 that the inequality (15.1) and the equality (15.2) both hold.
(e) This follows from Theorem 15.2.7. □

The performance of closed-loop system controlled by an LQG-PO-AC control law is discussed. Recall that J.C. Doyle, [8], has published that there are no guaranteed margins for LQG regulators. This was mentioned in Chapter 12 after Problem 12.7.1. It is described in more detail in Section 22.2.

The various control objectives are below discussed one by one. Consider a time-invariant Gaussian stochastic control system and the optimal stochastic control problem with partial observations on an infinite horizon. Assume that the stabilizability and the detectability conditions of Assumption 15.2.2 all hold.

The first control objective is the stability of the closed-loop system. It follows from Theorem 15.2.10.(a) that the closed-loop system is asymptotically stable.

The second control objective is whether the eigenvalues of the closed-loop system can be bounded away from the instability boundary, which for a time-invariant discrete-time Gaussian system is the unit circle. No such bound can exist in general! See the discussion in Section 22.2 after Def. 22.2.5 on the exponential-asymptotic convergence rate of the filter algebraic Riccati equation. See also Theorem 13.2.15 of Chapter 13 and the discussion following it.

In case the system matrix representing the noise affecting the output has the form $r\,N$ where $N \in \mathbb{R}^{n_y \times n_v}$ satisfies $0 \prec NN^T$ and $r \in (0,\infty)$ satisfies $\lim\; r \downarrow 0$, then it is known that the eigenvalues of the closed-loop filter system, thus of $A-KC$, move partly to the zeros of the system relating (x_0,u) to (x,z) and move partly to zero in a discrete-time Butterworth pattern. The reader can find this result in [16, 15]. If the system described above has a system zero, see Def. 21.5.2 close to the instability boundary or on the instability boundary, then there will be an eigenvalue of the closed-loop system which approaches the value of that zero when the parameter r approaches zero. Therefore, there cannot be a performance guarantee in general for the eigenvalues of the closed-loop system. The same conclusion holds for the eigenvalues of the the system matrix $A+BF(Q_c^*)$.

The third control objective is considered. It can be proven using the assumptions of stabilizability and of detectability and if the open-loop system is asymptotically stable that then the value of the optimal control law satisfies, $J_{ac}(g^*) < J_{ac}(g_{u0})$, where $g_{u0} \in G_{tv,fc}$ denotes the zero control law.

A special case of the tracking problem of a time-invariant Gaussian stochastic control system with a time-invariant Gaussian system as reference system, see Problem 14.4.20, is such that the combined system is not controllable but stabilizable. Hence the stabilizability conditions are necessary for this example.

15.3 Minimum-Variance Control with Partial-Observations

The problem of minimum-variance control has been motivated in Example 4.1.1. The optimal stochastic control theory for minimum-variance control has been developed primarily for polynomial descriptions of Gaussian stochastic control systems. This is unfortunate because the problem has unexplored connections with the optimal control problem for a linear system and a singular quadratic cost function. In this section the problem of minimum-variance control is briefly mentioned and the result described.

Problem 15.3.1. Consider a time-invariant Gaussian stochastic control system representation with conditions,

$$\begin{aligned} x(t+1) &= Ax(t)+Bu(t)+Mv(t),\; x(0)=x_0 \in G(m_{x_0},Q_{x_0}),\\ y(t) &= Cx(t)+Du(t)+Nv(t),\\ &\quad v(t) \in G(0,I),\; T=\mathbb{N},\; \text{rank}(N)=n_y \;\Rightarrow\; NN^T \succ 0. \end{aligned}$$

Consider further the past-output and past-input information pattern,

$$\{F_{t-1}^{y} \vee F_{t-1}^{u},\ t \in T\}$$

and the corresponding set G of time-invariant control laws. For any $g \in G$, let $g = \{g_0, g_1, \ldots, g_{t_1-1}\}$, $g_0 \in U$, and for $t \in T$, $g_t : Y^t \times U^t \rightarrow U$. Consider the average-cost quadratic cost function,

$$J_{mv}(g) = \limsup_{t\to\infty} \frac{1}{t} E[\sum_{s=0}^{t-1} x^g(s)^T Q_c x^g(s)],\ J_{mv} : G \rightarrow \mathbb{R}_+ \cup \{\infty\}, \tag{15.3}$$

where $Q_c \in \mathbb{R}_{pds}^{n_x \times n_x}$ hence satisfies $Q_c = Q_c^T \succeq 0$. The problem is then to solve the optimal stochastic control problem for the value and the optimal control law (J_{mv}^*, G^*),

$$J_{mv}^* = \inf_{g\in G} J_{mv}(g) = J_{mv}(g_{mv}^*).$$

This problem is called the *optimal stochastic control problem for the variance of the controlled output.* The case where $Q_c = C^T C$ such that $x^T Q_c x = y^T y$, is in the literature often called the *minimum-variance control problem*, where it is understood that the variance of the output process is referred to. That in this case the controlled output equals the observed output process is rather special.

A further special is to restrict attention to the set of control laws g such that y^g is a stationary process and to infimize $\inf_{g\in G_{mv}} \operatorname{tr}(Q_c Q_{y^g})$ where $Q_{y^g} = E[y^g(t) y^g(t)^T]$.

The minimum-variance control problem will be treated in this book only for the single-input-single-output case, thus with $y : \Omega \times T \rightarrow \mathbb{R}$ and $u : \Omega \times T \rightarrow \mathbb{R}$, and in the ARMAX representation

$$y(t) = -\sum_{i=1}^{n_y} -a_i y(t-i) + \sum_{i=0}^{n_u} b_i u(t-i-d) + \sum_{i=0}^{n_w} c_i w(t-i)$$

where $n_y,\ n_u,\ n_w \in \mathbb{N}$, $w : \Omega \times T \rightarrow \mathbb{R}$ is a sequence of independent random variables with $E[w(t)] = 0$ and $E[w(t)^2] = q_w \in (0,\infty)$. It is assumed that $b_0 \neq 0$ and that $c_0 = 1$. In case that the values of $n_y,\ n_u,\ n_v$ are different, one can determine their maximum and then add zero coefficient so that in a new representation all sums have the same upper limit. With the operator theory notation $zx(t) = x(t-1)$, the representation may be rewritten as,

$$\begin{aligned} a(z) &= 1 + \sum_{i=1}^{n_a} a_i z^{i+1},\ b(z) = z^d \sum_{i=0}^{n_b} b_i z^i, c(z) = \sum_{i=0}^{n_c} c_i z^i, \\ a(z)y(t) &= b(z)u(t) + c(z)w(t). \end{aligned}$$

Consider the above defined ARMAX description of a single-input-single-output ($n_u = n_y = 1$) Gaussian stochastic control system and the optimal stochastic control Problem 15.3.1 with as cost function,

$$J_{mv}(g) = \limsup_{t\to\infty} \frac{1}{t} E[\sum_{s=0}^{t-1} y(s)^2].$$

Restrict attention to those control laws $G_1 \subseteq G$ for which $E|y(t)|^2 < \infty$ for all $t \in T$ and assume that this set is not empty. Assume that the roots of the polynomials $b(.)$ and $c(.)$ are strictly outside the unit circle. The condition on the polynomial b is called a *minimum-phase condition*.

The polynomial equation for the polynomials (f, g)

$$c(z) = a(z)f(z) + z^d b(z)g(z),$$
$$f(z) = \sum_{i=0}^{n_f} f_i z^i, \;\; g(z) = \sum_{i=0}^{n_g} g_i z^i,$$

has a unique solution.

Consider the control law g^* specified by the controller

$$0 = b(z)f(z)u(t) + g(z)y(t), \;\; \text{with } b(z)f(z) = \sum_{i=1}^{n_{bf}} k_i z^i, \; g(z) = \sum_{s=0}^{n_g} g_i z^i,$$
$$u(t) = \sum_{i=1}^{r} k_i u(t-i) + \sum_{i=0}^{s} g_i y(t-i).$$

The resulting closed-loop stochastic control system is asymptotically stable.

The output process of the closed-loop system is, asymptotically, given by

$$y(t) = f(z)w(t) = \sum_{i=1}^{n_f} f_i w(t-i).$$

The control law g^* defined in (b) minimizes the average cost, or,

$$J_{mv}(g^*_{mv}) = \inf_{g\in G} J_{mv}(g) = q_w \sum_{i=0}^{d-1} f_i^2.$$

15.4 Further Reading

The topic of this chapter is optimal stochastic control with partial observations on an infinite-horizon with either average cost or with discounted cost. There are few books which treat this research topic in depth. The number of papers on the relevant theoretical framework is quite limited.

Books which contain chapters on the subject of optimal stochastic control for either a Gaussian stochastic control system with partial observations on an infinite horizon or for a output-finite-state-polytopic stochastic system with partial observations at the level of this book include: [4, Ch. 4], [7, 6.3], [12, 6.4 - 6.7], and [13]. Books addressing control theory for a broader set than Gaussian stochastic control systems include [6, 20].

Average cost. The problem of infinite-horizon optimal stochastic control for a Gaussian stochastic control system representation with the discounted and average cost function is covered poorly in the literature.

Discounted cost. Optimal stochastic control with partial observations of a finite stochastic control system on an infinite horizon with discounted cost is treated in [19]. See for references also the section *Further Reading* of the previous chapter. Papers by S.I. Marcus and co-workers, see [9, 10, 11].

Minimum variance stochastic control. The minimum variance control problem was formulated and solved in [2], and extended in [17]. The multi-input-multi-output case is treated in by U. Shaked and P.R. Kumar, [18]. The polynomial description of stochastic systems and algorithms for polynomial equations are mentioned in [14]. A state space approach to the minimum variance control problem is presented in [1] but the associated full paper has not been published. Books on stochastic control which discuss minimum variance control are [3, 5, 12]. The author conjectures that a treatment of minimum-variance control formulated in terms of a state-space stochastic system is possible if a better understanding is reached on the invertibility of a discrete-time linear system.

Performance guarantees. J.C. Doyle has published about the performance of the LQG optimal stochastic control problems. He has concluded that no substantial performance guarantees exist. See for a publication [8].

References

1. C. Arnoldi and R.H. Kwong. A state space approach to minimum variance control of multivariable ARMAX systems. In *Proceedings of the 28th Conference on Decision and Control*, pages 2125–2126, New York, 1989. IEEE Press. 522, 596
2. K.J. Aström. Computer control of a paper machine - An application of linear stochastic control theory. *IBM J. Res. & Developm.*, 11:389–405, 1967. 9, 78, 120, 522, 575, 596
3. K.J. Aström. *Introduction to stochastic control*. Academic Press, New York, 1970. 376, 410, 467, 522, 575, 596
4. D.P. Bertsekas. *Dynamic programming and stochastic control*. Academic Press, New York, 1976. 376, 405, 410, 439, 468, 502, 525, 526, 575, 595
5. D.P. Bertsekas. *Dynamic programming: Deterministic and stochastic models*. Prentice-Hall, Englewood Cliffs, 1987. 596
6. D.P. Bertsekas and S.E. Shreve. *Stochastic optimal control: The discrete time case*. Academic Press, New York, 1978. 49, 428, 431, 468, 575, 595
7. M.H.A. Davis and R.B. Vinter. *Stochastic modelling and control*. Chapman and Hall, London, 1985. 120, 376, 410, 468, 575, 595
8. John C. Doyle. Guaranteed margins for LQG regulators. *IEEE Trans. Automatic Control*, 23:756–757, 1978. 440, 592, 596, 822, 824
9. E. Fernandez-Gaucherand, A. Arapostathis, and S.I. Marcus. On the average cost optimality equation and the structure of optimal policies for partially observable markov decision processes. *Ann. Oper. Res.*, 91:439–470, 1991. 596
10. E. Fernández-Gaucherand, A. Arapostathis, and S.I. Marcus. On the average cost optimality equation and the structure of optimal policies for partially observed Markov decision processes. *Ann. Oper. Res.*, 29:439–470, 1991. 596
11. E. Fernandez-Gaucherand and S.I. Marcus. Risk-sensitive optimal control of hidden markov models: Structural results. *IEEE Trans. Automatic Control*, 42:1418–1422, 1997. 596

12. P.R. Kumar and P. Varaiya. *Stochastic systems: Estimation, identification, and adaptive control*. Prentice Hall Inc., Englewood Cliffs, NJ, 1986. 376, 410, 468, 525, 575, 595, 596
13. P.R. Kumar and P. Varaiya. *Stochastic systems: Estimation, identification, and adaptive control*. Number 75 in Classics in Applied Mathematics. SIAM, Philadelphia, 2015. 468, 575, 595
14. V. Kučera. *Discrete linear control - The polynomial equation approach*. Czechoslovak Academy of Sciences, Prague, 1979. 522, 596
15. H. Kwakernaak. Asymptotic root loci of multivariable linear optimal regulators. *IEEE Trans. Automatic Control*, 21:378–382, 1976. 593, 822, 823
16. H. Kwakernaak and R. Sivan. *Linear optimal control systems*. Wiley-Interscience, New York, 1972. 120, 376, 410, 467, 489, 593, 822, 823
17. V. Peterka. On steady state minimum variance control strategy. *Kybernetica*, 8:219–232, 1972. 596
18. U. Shaked and P.R. Kumar. Minimum variance control of discrete time multivariable systems. *SIAM J. Control & Opt.*, 24:396–411, 1986. 522, 596
19. E.J. Sondik. The optimal control of partially observable Markov processes over the infinite horizon: Discounted cost. *Oper. Res.*, 26:282–304, 1978. 596
20. C. Striebel. *Optimal control of discrete time stochastic systems*, volume 110 of *Lecture Notes in Economic and Mathematical Systems*. Springer-Verlag, Berlin, 1975. 431, 468, 575, 595, 603, 605, 742

Chapter 16
Stochastic Control Theory

Abstract Stochastic control issues of a general character are presented. Problems of control theory are mentioned which require research interest the coming years. A general method for sufficient and necessary conditions for the existence of an optimal control law is discussed. The framework covers arbitrary cost functions, including additive and multiplicative functions. In addition, the approach of a measure transformation is formulated and illustrated for stochastic control of a partially-observed stochastic control system.

Key words: Problems. Optimality conditions. Measure transformation approach.

This chapter contains several research issues which do not fit the structure of the past four chapters.

16.1 Research Problems of Control of Stochastic Systems

There follow a set of several problems of control theory which are motivated by control engineering in a broad sense. The order of the problems is arbitrary.

Control for the Effective Interaction of Control and Observation

How to solve problems of stochastic control with partial observations where the input and the observations are tightly intertwined?

The motivation of this problem comes from control engineering. The input signal to a stochastic control system can in several systems be used to obtain observations of a higher quality than without the use of the input for the observation task. A Gaussian stochastic control system does not exhibit this form of interaction hence in control theory the problem has received little attention so far. The interaction shows up in stochastic control of a state-finite stochastic control system as described in Section 14.5. In the literature for control of this system, little attention has been spend on this research issue.

Needed are new concepts. It is expected that new concepts of observability and of controllability are needed for this problem. It is not yet clear to the author what these concepts are. The interaction of two control tasks for this problem may help in the analysis: (1) to minimize the control performance; and (2) to optimize the state observation by improving the quality of the observations and exciting the state process so as to better control the state. Of course, in the end the optimal control law will handle the trade-off between these tasks. A better understanding of the performances of these tasks will be useful, also for synthesis of suboptimal control laws.

Control of Partially-Observed Stochastic Control Systems

How to advance the understanding of stochastic realization of stochastic control systems on the information structure?

This research issue is motivated by the approach to control of a stochastic control system with partial observations as formulated in Chapter 14 and in Chapter 15. The stochastic realization viewpoint seems of interest, though it is close to the early approach to stochastic control with partial observations. Does the stochastic realization with respect to the original stochastic realization still play a role? Possibly not. But if it plays a role, how should one think of the original stochastic realization.

It is clear from the case of stochastic control of a state-finite stochastic control system that a new characterization is needed of stochastic controllability of the stochastic realization with respect to the information structure. Because the stochastic realization is different, the algebraic and geometric structure of the system is different and therefore other mathematical tools are needed.

An exploration will be useful of examples and models of control engineering problems in areas different from classical control engineering. Most of the developments in optimal stochastic control in the past decades have taken place in the area of control of communication and computer networks. The many structural properties of control laws derived in this development may point to a yet undetermined coherent theoretical framework. Other application areas that stimulate stochastic control theory are control of urban and freeway traffic, financial mathematics, mathematical economics, and modeling of the operations of banks and insurance companies.

Control of Communication Systems

How to formulate and solve stochastic control systems for the communication of information?

The motivation of this area comes from the communication of information. Claude Shannon has formulated the model of the communication system and solved several of the elementary models. A communication system consists of a source, a channel, and a receiver. The engineer has to synthesize and to design an encoder and a decoder. The information objectives are that the specified information is communicated from the source to the receiver with a minimal loss of information. Such problems are problems of stochastic control. The research area of information theory and of communication theory is dedicated to these problems. Yet, there is a serious difference between the approaches of those research areas and that of control theory. With the emphasis on stochastic control of networked systems, the communication of one subsystem to another in a networked system is of critical importance.

It is suggested that attention is spend on modeling of stochastic control problems of the communication of information. In particular the formulation of sources and channels is best changed compared with what is customary in other research areas. The stochastic control problems can first be formulated as is custormary in control theory. In a second phase it has to be proven that the optimal control laws also achieve path-wise optimality conditions. The path-wise optimality is requested by researchers of information theory but it restricts the interest of researchers of control theory for this research area. However, the extra research required is not that complicated.

Of primary interest is the communication of information from one subsystem to another in a networked stochastic control system. This problem also arises in decentralized control, where it is currently the most important obstacle to further research progress.

Control of a Networked Stochastic Control System

How to control a network of stochastic systems so that it together achieves the control objectives set for the entire networked system?

The motivation for this research issue comes from control engineering. Current control systems are almost exclusively a network of control systems, not all stochastic control systems. But it seems best to think of a network of only stochastic control systems. The research of this issue is in a very early stage. The current research issues deal with modeling and simple forms of interaction.

One research issue is how to coordinate the activities of a set of networked control systems such that they together achieve control objectives? This requires an understanding of the capacities of each subsystem, the interaction of these subsystems, the information exchange of these subsystems, and the control law for achieving the control objectives. The author with co-workers has investigated a form of coordination control and of a multilevel structure for control. The concept of a multilevel structure is again defined in terms of the conditional independence relation. But that approach is in a very early stage. More research is required.

Multiple Conditional Independence

How to relate the various state σ-algebras of the multiple conditional independence relation? How to estimate a subset of unobserved outputs from other observed outputs? How to relate subsets of outputs such that conditional independence is stronger than the general case? Characterize minimal state σ-algebras. See Section 5.9 for the multiple conditional-independence relation.

The motivation for this research problem is the modeling and analysis of databases. Each element of a data base has a finite set of items. How to relate the various database elements to each other? How to determine the minimal information which makes subsets of items conditionally independent?

16.2 General Optimality Conditions

The reader finds a generalization of optimal control of stochastic systems in this section. This approach was developed in the early 1970's by M.H.A. Davis, R. Rishel, C. Striebel, and P. Varaiya at more or less the same time, possibly independently. The formulations of the individual researchers differ slightly. The reader finds the references of those authors in the *Further Reading* of this chapter.

The theory presented in the Chapters 12 to Chapter 15 is not fully satisfactory in its formulation of optimality conditions. Moreover, the proof of necessity is not so clear.

The theory of this section uses the concept of a P-essential infimum of which a definition and results may be found in Section 19.12.

An explanation of the unsatisfactory character of the optimality conditions follows. Consider Problem 12.2.1. Recall the variables and stochastic processes, with a slight abuse of notation,

$$\begin{aligned}
& b : T_1 \times X \times U \rightarrow \mathbb{R}_+,\ b_1 : X \rightarrow \mathbb{R}_+, \text{ Borel measurable functions,} \\
J(g) &= E\left[\sum_{s=0}^{t_1-1} b(s, x^g(s), u^g(s)) + b_1(x^g(t_1))\right],\ J : G \rightarrow \mathbb{R}_+; \\
& J : G \times \Omega \times T \rightarrow \mathbb{R}, \\
J(g, x^g(t)) &= E[\sum_{s=t}^{t_1-1} b(s, x^g(s), u^g(s)) + b_1(x^g(t_1)) | F_t^{x^g}] \\
J(g, x_0) &= E[\sum_{s=0}^{t_1-1} b(s, x^g(s), u^g(s)) + b_1(x^g(t_1)) | F^{x_0}] \\
J(g) &= E[J(g, x_0)].
\end{aligned}$$

The value function is constructed by the dynamic programming procedure. Subsequently the results are then that,

$$\begin{aligned}
V(x_0) &\le J(g,x_0),\ \forall\, g \in G,\ \Rightarrow \\
E[V(x_0)] &\le E[J(g,x_0)] = J(g),\ \forall\, g \in G,\ \Rightarrow \\
E[V(x_0)] &\le J^* = \inf_{g \in G} J(g); \\
E[V(x_0)] &= J(g^*) = J^* = \inf_{g \in G} J(g).
\end{aligned}$$

But there is no use of the expression $V(x_0) \le \inf_{g\in G}\ J(g,x_0)$. The difficulty here is how to define the infimum in the last expression. Note that $J(g,x_0) : \Omega \to \mathbb{R}_+$ is a random variable, the initial state x_0 is fixed, and that $g \in G$. If the set of control laws G is finite then $\inf_{g\in G}\ J(g,x_0)$ is the minimum of a finite number of random variables. But if G is neither finite nor countable, then one has to properly define the infimum.

The concept needed is that of the P-essential infimum of a set of random variables, in general uncountable, denoted by $\mathrm{P}-essinf$, see Section 19.12. The existence of such an infimum is proven using the Radon-Nikodym theorem. It is also possible to derive theorems on the interchange of conditional expectation and the P-essential infimum.

C. Striebel has formulated general optimality conditions which do not only cover the case of additive and of multiplicative cost functions, but arbitrary cost functions. This approach then also include optimal stopping problems.

Using the concept of P-essential infimum, one can then make sense of the expression, $\mathrm{P}-essinf_{g\in G}\, J(g,x_0)$ and prove that $V(x_0) = \mathrm{P}-essinf_{g\in G}\, J(g,x_0)$. The optimality conditions are formulated in terms of martingales. The value function $\{V(x^{g^*}(t)), F_t,\ t \in T\}$ is a martingale, while for any nonoptimal control law $g \in G$, $\{J(g,t,x(t)),\ F_t,\ t \in\}$ is a submartingale. From the general optimality conditions one then directly derives the corresponding dynamic programming equations for the special case of an additive cost function and of a multiplicative cost function.

The optimality conditions are not stated in this book so as to avoid an overlap. The best source for these conditions is the book of C. Striebel, [21, Ch. 4].

It has to be explicitly stated that the optimality conditions using the $\mathrm{P}-essinf$ does not solve the existence of a measurable optimal control law, see Section 12.5. For that research issue another approach is needed.

16.3 Stochastic Control via a Measure Transformation

The reader finds in this section the measure transformation approach to optimal stochastic control problems with partial observations.

The approach was first published by V.E. Beneš, [1] for the case of continuous-time stochastic control systems with partial observations. In this section the discrete-time case is formulated.

The reader may be now know of the Radon-Nikodym theorem relating one probability measure to another one in case of absolute continuity of the measares involved.

A generalization to the case of Brownian motion was published by I.V. Girsanov in 1960, [13].

Beneš used the measure transformation approach in the following way. One starts with a probability measure P_0 and takes a Brownian motion process y with respect to P_0. At this point there is no control law stated yet. One defines an input process u which is adapted to the observation process y, $\{u(t),\ F^y_{t-1} \vee F^u_{t-1},\ t \in T\}$. Then one defines a measure transformation yielding a probability measure P_1 which is equivalent to P_0 such that with respect to P_1, the observation process has a representation exactly equal to that of the stochastic control system. In an optimal stochastic control problem the σ-algebra of the observations depends on the control law. However, if one transforms the problem from probability measure P_1 to P_0 then one obtains a new problem in which the filtration $\{F^y_t,\ t \in T\}$ does not depend on the control law. Then it is possible to prove the existence of an observation-based stochastic realization. of the consider stochastic control system and the optimal control problem with respect to the probability measure P_1.

The approach is sketched below for a discrete-time stochastic system.

Definition 16.3.1. Consider a complete probability space (Ω, F, P) and a finite time set $T = T(0:t_1) = \{0,1,2,\ldots,t_1\} \subset \mathbb{Z}_+$ for $t_1 \in \mathbb{Z}_+$. Define a stochastic process $y : \Omega \times T \to \mathbb{R}^{n_y}$ which is a Gaussian white noise process with for all $t \in T$, $y(t) \in G(0, Q_v)$ and $0 \prec Q_v$. Note the subindex v of Q_v which will became clear later.

Define a state process by the random variables and processes,

$$\begin{aligned} & x(t+1) = Ax(t) + Bu(t) + Mv(t),\ x(0) = x_0, \\ & \qquad \text{as in a Gaussian stochastic control system,} \\ & \qquad u : \Omega \times T \to \mathbb{R}^{n_u},\ \{u(t),\ F^y_{t-1} \vee F^u_{t_1},\ t \in\}, \text{adapted process.} \end{aligned}$$

Define the Radon-Nikodym process as,

$$\begin{aligned} & r_{1|0} : \Omega \times T \to \mathbb{R}_+,\ r_{1|0}(0) = 1, \\ r_{1|0}(t+1) = {} & r_{1|0}(t) \times \exp\Big((Q_v^{-1/2} Cx(t+1))^T y(t+1) \\ & - (Q_v^{-1/2} Cx(t+1))^T Q_v (Q_v^{-1/2} Cx(t+1))/2 \Big). \end{aligned}$$

Theorem 16.3.2. *Consider the objects and relations of Def. 16.3.1.*

(a)$\{r_{1|0}(t),\ F^y_t \vee F^u_{t-1},\ t \in T\} \in M_1(P_0)$ *and for all* $t \in T$, $r_{1|0}(t) > 0$ *a.s.*(P_0).

(b)The formula,

$$\frac{dP_1}{dP_0} = r(t_1),$$

defines a probability measure P_1 *on* (Ω, F) *and* P_1 *and* P_0 *are equivalent probability measures.*

(c)Define the stochastic process $r_{0|1} : \Omega \times T \to \mathbb{R}_+$ *by* $r_{0|1}(t) = r_{1|0}(t)^{-1}$. *Then* $\{r_{0|1}(t),\ F^y_t \vee F^u_t,\ t \in T\} \in M_1(P_1)$.

(d)With respect to the probability measure P_1 the stochastic processes (x,y) have the representation,

$$\begin{aligned} x(t+1) &= Ax(y)+Bu(t)+Mv(t),\ x(0)=x_0,\\ y(t) &= Cx(t)+Q_v^{1/2}w(t),\\ &\quad w:\Omega\times T\rightarrow\mathbb{R}^{n_w},\ \forall\, t\in T,\ w(t)\in G(0,I_{n_y}),\end{aligned}$$

and w is a Gaussian white noise process. Moreover, P_1 and P_0 are identical on the σ-algebra $F^x_{t_1}$. Thus the measure P_1 equals the probability measure of Problem 14.4.1 except for the independence of the v and w processes.

The proof is simple using the knowledge of a measure transformation, Section 19.9.

16.4 Further Reading

Problems of control theory Interaction of control and observation, [2, 3]. Control of the communication of information and of communication systems, [11, 15, 23]. Research of C.D. Charalambous and his research group is very promising, see the following paper and the references provided in that paper, [4].

General optimality conditions for optimal control. An approach for additive cost functions has been formulated by R. Rishel, [20]. The paper by M.H.A. Davis and P. Varaiya develops this for the continuous-time case, [6]. C. Striebel has published this, [21, Ch. 4 and 5], for the discrete-time case and in [22] for the continuous-time case. Text on the P-essential infimum may be found at [18, Ap. A] and [21, Sec. A.2].

LEQG. The optimal stochastic control problem with the exponential quadratic cost function and with complete observations, is treated in [5, 14, 16, 19, 12]. Infinite-horizon LEQG with complete observations, [9, 10].

Stochastic control via the measure transformation approach was proposed by V.E. Benes for continuous-time stochastic control systems and the relevant references may be found in [1, 6, 7, 8] and in the book [17, Ch. 5].

References

1. V.E. Beneš. Existence of optimal control laws. *SIAM J. Control*, 3:446–475, 1971. 603, 605
2. René Boel and Jan H. van Schuppen. Optimal control and optimal sensor activation for Markov decision problems with costly observations. In *Proc. IEEE Multi-Conference on Systems and Control (MSC.2015)*, page to appear, New York, 2015. IEEE, IEEE Press. 605
3. René K. Boel and Jan H. van Schuppen. Control of sensors of a gaussian stochastic control system. In *Proc. IEEE Conference on Decision and Control (CDC.2015)*, New York, 2015. IEEE Press. 605
4. C.D. Charalambous, K. Kourtellaris, and I. Tzortzis. Information transfer of control strategies: Dualities of stochastic optimal control theory and feedback capacity of information theory. *IEEE Trans. Automatic Control*, 62:5010–5025, 2017. 605

5. Charalambos D. Charalambous and Robert J. Elliott. Classes of nonlinear partially observable stochastic optimal control problems with explicit optimal control laws. *SIAM J. Control & Opt.*, 36:542–578, 1998. 605
6. M.H.A. Davis and P. Varaiya. Dynamic programming conditions for partially observable stochastic systems. *SIAM J. Control*, 11:226–261, 1973. 468, 575, 605
7. Tyrone Duncan and Pravin Varaiya. On the solutions of a stochastic control system. *SIAM J. Control*, 9:354–371, 1971. 605
8. Tyrone Duncan and Pravin Varaiya. On the solutions of a stochastic control system - ii. *SIAM J. Control*, 13:1077–1092, 1975. 605
9. W.H. Fleming and W.M. McEneaney. Risk sensitive control with ergodic cost criteria. In *Proceedings of the 31st IEEE Conference on Decision and Control*, pages 2048–2052, New York, 1992. IEEE Press. 605
10. W.H. Fleming and W.M. McEneaney. Risk-sensitive control on an infinite time horizon. *SIAM J. Control & Opt.*, 33:1881–1915, 1995. 605
11. R.G. Gallager. *Information theory and reliable communication*. John Wiley & Sons, New York, 1968. 605
12. Zo gang Pan and Tamer Basar. Model simplification and optimal control of stochastic singularly perturbed systems under exponentiated quadratic cost. Report UILU-ENG-93-2249, DC-157, Coordinated Science Laboratory, University of Illinois, Urbana, 1993. 605
13. I.V. Girsanov. On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. *Th. Probab. Appl.*, 5:285–301, 1960. 352, 604, 742
14. R.A. Howard and J.A. Matheson. Risk-sensitive Markov decision processes. *Management Science*, 18:356–369, 1972. 526, 605
15. S. Ihara. *Information theory for continuous systems*. World Scientific, Singapore, 1993. 605
16. D.H. Jacobson. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Trans. Automatic Control*, 18:124–131, 1973. 468, 605
17. Yuri Kabanov and Sergei Pergamenshchikov. *Two-scale stochastic systems*. Number 49 in Applications of Mathematics. Springer, Berlin, 2003. 605
18. I. Karatzas and S.E. Shreve. *Methods of mathematical finance*. Number 39 in Applications of Mathematics. Springer, Berlin, 1998. 411, 605, 742
19. P.R. Kumar and J.H. van Schuppen. On the optimal control of stochastic systems with an exponential-of-integral performance index. *J. Math. Anal. Appl.*, 80:312–332, 1981. 605
20. R. Rishel. Necessary and sufficient dynamic programming conditions for continuous time stochastic optimal control. *SIAM J. Control & Opt.*, 8:559–571, 1970. 605
21. C. Striebel. *Optimal control of discrete time stochastic systems*, volume 110 of *Lecture Notes in Economic and Mathematical Systems*. Springer-Verlag, Berlin, 1975. 431, 468, 575, 595, 603, 605, 742
22. C. Striebel. Martingale conditions for the optimal control of continuous time stochastic systems. *Stoc. Proc. Appl.*, 18:329–347, 1984. 605
23. J. Walrand and P.P. Varaiya. *High-performance communication networks (2nd ed.)*. Morgan Kaufmann, San Francisco, 2000. 169, 605

Chapter 17
Appendix A Mathematics

abstractThe reader finds in this short appendix concepts and results of various topics of mathematics. These topics are used in the body of the book but are not part of control theory. Topics covered are: algebra of set theory, a canonical form, algebraic structures including monoids, groups, and rings; linear algebra and matrices; analysis; geometry, convex sets, affine sets; and optimization.

Key words: Algebra, linear algebra, analysis, and geometry.

The reader finds in this chapter concepts and results of mathematics used in the body of the book. The chapter is best considered to be a partial encyclopedia rather than a study chapter. In case a reader needs other or more detailed theory, she or he is referred to the Section *Further Reading* where references are provided.

17.1 Algebra of Sets

The concepts of this section belong to the algebra of sets and are also described as either *elementary algebra* or *universal algebra*.

The reader is expected to know the following subsets of the real numbers for which notation is introduced below. The *natural numbers* $\mathbb{N} = \{0,\ 1,\ 2, \ldots\}$; the *integers* $\mathbb{Z}$ and the *positive integers* $\mathbb{Z}_+$; the *rational numbers* p/q for $p,\ q \in \mathbb{Z}$ with $q \neq 0$; the *real numbers* denoted by $\mathbb{R}$ the *positive real numbers* $\mathbb{R}_+$, and the *strictly positive real numbers* $(0, \infty) \subset \mathbb{R}$; the *complex numbers* $\mathbb{C}$; the *n-tuples of the real numbers* $\mathbb{R}^n$ for an integer $n \in \mathbb{Z}_+$. A subset of unordered integers is denoted for example by $\{1,\ 2,\ 3\} \subset \mathbb{Z}_+$ while a subset of ordered integers is denoted for example by $(1,\ 2,\ 3) \subset \mathbb{Z}_+$.

Definition 17.1.1. A set A is called a *finite set* if there exists an integer $n \in \mathbb{Z}_+$ such that A is bijectively related to $\mathbb{Z}_n = \{1,\ 2,\ \ldots,\ n\}$. Hence there exists a bijection $b : \mathbb{Z}_n \to A$.

The set A is called a *countable set* or a *denumerable set* if either it is a finite set or if it is bijectively related to the natural numbers $\mathbb{N}$.

Relations and Canonical Forms

Consider a set X. A *relation* E on the set X is a subset of the product set, $E \subseteq X \times X$. Several relations are explicitly defined.

Definition 17.1.2. Consider a set X. The relation $E \subseteq X \times X$ is called an *equivalence relation* if the following three conditions all hold:

1. *reflexivity*: $\forall\, x \in X$, $(x,\ x) \in E$;
2. *symmetry* : $\forall\, x,\ y \in X$, $(x,\ y) \in E$ implies that $(y,\ x) \in E$;
3. *transitivity*: for all $x,\ y,\ z \in X$, $(x,\ y) \in E$ and $(y,\ z) \in E$ imply that $(x,\ z) \in E$.

Example 17.1.3. *Orthogonal equivalence of symmetric positive-definite matrices.* In this example use is made of several concepts of matrices which are defined in Section 17.4.

Consider the set of positive-definite symmetric matrices denoted by $\mathbb{R}^{n\times n}_{pds}$ for an integer $n \in \mathbb{Z}_+$ and the set of orthogonal matrices denoted by $\mathbb{R}^{n\times n}_{ortg}$ of the same size. A matrix U is an orthogonal matrix if $UU^T = I = U^T U$ which is denoted by $U \in \mathbb{R}^{n\times n}_{ortg}$

Define the relation,

$$E_{ortg} = \left\{ (Q_1,\ Q_2) \in \mathbb{R}^{n\times n}_{pds} \times \mathbb{R}^{n\times n}_{pds} |\ \exists\, U \in \mathbb{R}^{n\times n}_{ortg} \text{ such that } Q_1 = UQ_2U^T \right\}.$$

It will be proven that E_{ortg} is an equivalence relation. For any matrix $Q \in \mathbb{R}^{n\times n}_{spd}$, $(Q,\ Q) \in E_{ortg}$ if one takes the orthogonal matrix $U = I$. Because an orthogonal matrix is invertible with $U^{-1} = U^T$, if $(Q_1,\ Q_2) \in E_{ortg}$ with $Q_1 = UQ_2U^T$ then $Q_2 = U^T Q_1 U$ hence $(Q_2,\ Q_1) \in E_{ortg}$. If $Q_1,\ Q_2,\ Q_3 \in \mathbb{R}^{n\times n}_{pds}$ with $(Q_1,\ Q_2)$, $(Q_2,\ Q_3) \in E_{ortg}$ hence there exist $U_{12},\ U_{23} \in \mathbb{R}^{n\times n}_{ortg}$ such that $Q_1 = U_{12}Q_2U_{12}^T$ and $Q_2 = U_{23}Q_3U_{23}^T$ then, by substitution, $Q_1 = U_{12}Q_2U_{12}^T = U_{12}U_{23}Q_3U_{23}^TU_{12}^T$ where $U_{12}U_{23} \in \mathbb{R}^{n\times n}_{ortg}$, hence $(Q_1,\ Q_3) \in E_{ortg}$. Thus E_{ortg} is an equivalence relation.

Universal algebra provides the concept of a canonical form.

Definition 17.1.4. Consider a set X and an equivalence relation $E \subseteq X \times X$.

Define a *canonical form* $(X_c,\ f_c)$ or a *quotient set* of (X,E) for the set X with the equivalence relation $E \subseteq X \times X$, as a subset $X_c \subseteq X$ such that, for any $x \in X$, there exists a unique element $x_c \in X_c$ such that $(x,x_c) \in E$. Define then the *transformation to the canonical form* as the map $f_c : X \to X_c$ by the formula $f_c(x) = x_c$ if $(x,x_c) \in E$. That map transforms any element $x \in X$ into the unique element $x_c \in X_c$ of the canonical form to which it is equivalent.

The construction of a canonical form for a set with an equivalence relation is arbitrary because a canonical form is not unique. From examples one learns how to formulate canonical forms. Arguments of algebra and of convenience play a role in the selection of the canonical form. Once a candidate canonical form $(X_c,\ f_c)$ has been selected then the researcher has to prove that:

1. for any element $x \in X$ of the set there exists a member $x_c \in X_c$ of the canonical form set which is equivalent to the considered member of the original set, $(x,\ x_c) \in E$; in particular cases, the existence has to be proven;
2. if any two members of the canonical form are equivalent then they are identical; equivalently, if $x,\ y \in X_c$ and if $(x,\ y) \in E$ then $x = y$;
3. from (1) and (2) it follows that there exists a function $f_c : X \to X_c$ which maps any member $x \in X$ of the original set to a member $x_c = f_c(x) \in X_c$ of the canonical form set to which it is equivalent.

If the properties (1-3) hold then $(X_c,\ f_c)$ is a canonical form of $(X,\ E)$. The second property establishes the required uniqueness of the canonical form.

Example 17.1.5. There follows a very elementary example of a canonical form in terms of sets. Consider the set X and a relation on that set,

$$X = [0,1]^2,\ \ E = \{(x,\ y) \in X \times X |\ x = (x_1,\ x_2),\ y = (y_1,\ y_2),\ x_1 = y_1\}.$$

It follows directly from the above definition that the set E is an equivalence relation.

Define the candidate canonical form,

$$X_c = \{x \in X |\ x_2 = 0.8\},$$
$$f_c : X \to X_c,\ f_c(x) = (x_1,\ 0.8),\ \text{if}\ x = (x_1,\ x_2);\ (X_c,\ f_c).$$

Note that: (1) For all $x = (x_1,\ x_2) \in X$, there exists the element $x_c = f_c(x) = (x_1, 0.8) \in X_c$ such that $(x,\ x_c) \in E$. (2) If $x = (x_1,\ 0.8) \in X_c$ and $y = (y_1,\ 0.8) \in X_c$ are such that $(x,\ y) \in E$ then $x_1 = y_1$ hence $x = y$. Thus $(X_c,\ f_c)$ is a canonical form for the above equivalence relation.

The above defined canonical form is not unique. The following alternative sets can also be the basis of a canonical form,

$$X_{c,a} = \{x \in X |\ x_2 = 0.3\},\ \ X_{c,b} = \{x \in X |\ x_2 = 0.3 + 0.5x_1\}.$$

The above example is made such that the set X is decomposed with the equivalence relation into the product of the independent subset and the dependent subset.

Order Relation

Definition 17.1.6. A *partial-order relation* on a set X is denoted by the symbol $\leq$ and defined by the conditions: (1) *reflexivity*: for all $x \in X$, $x \leq x$; and (2) *transitivity*: for all $x,\ y,\ z \in X$, $x \leq y$ and $y \leq z$ imply that $x \leq z$; (3) *anti-symmetric*: for all $x,\ y \in X$, $x \leq y,\ y \leq x\ \Rightarrow\ x = y$. Call then $(X, \leq)$ a *partially-ordered set* or a *poset*. If $x \leq y$ then one also writes $y \geq x$.

Define the *strict-partial-order relation* denoted by the symbol $<$ and defined by the conditions (1) transitivity as above; and (2) for all $x \in X$, $x \not< x$.

A *total-order relation* is defined as a partial-order relation with the additional condition that (4) *trichotomy* holds: for all $x, y \in X$, either $x < y$, or $x > y$, or $x = y$.

Example 17.1.7. Consider a set X and the inclusion relation defined on subsets of that set, $\subseteq$ is a subset of $\text{Pwrset}(X) \times \text{Pwrset}(X)$, denoted by $X_2 \subseteq X_1$. Then $\subseteq$ is a partial-order relation. It is not a total-order relation; an example consists of two sets which are neither included one in the other and which have a nonempty intersection.

Example 17.1.8. Consider the set of tuples of integers, $\mathbb{Z}_+^2$. Define the relation $\leq \subseteq (\mathbb{Z}_+^2 \times \mathbb{Z}_+^2)$ by the conditions that $(x_1, x_2) \leq (y_1, y_2)$ if $\{ x_1 \leq y_1$ and $x_2 \leq y_2\}$.

It can then be proven that the relation $\leq$ is a partial order on the considered set. Note there are elements of $\mathbb{Z}_+^2$ which are not ordered, for example neither $(1,2) \not\leq (2,1)$ nor $(2,1) \not\leq (1,2)$. Thus the relation $\leq$ on $\mathbb{Z}_+^2$ is not a total-order relation.

Definition 17.1.9. Consider a partially-ordered set $(X, \leq)$ and a subset $X_s \subseteq X$. The element:

- $x^- \in X_s$ is called a *minimum* of X_s or a *minimal element* of the set X_s if for all $x \in X_s$, $x \leq x^-$ implies that $x = x^-$;
- $x^+ \in X_s$ is called a *maximum* of X_s or a *maximal element* of the set X_s if for all $x \in X_s$, $x^+ \leq x$ implies that $x = x^+$;
- the element $x_l \in X$ is called a *lower bound* of the subset X_s if for all $x_s \in X_s$, $x_l \leq x_s$; $\inf_{x_t \in X_s} x_t \in X$ is called the *infimum* or the *greatest lower bound* of X_s if $\inf_{x_t \in X_s} x_t$ is a lower bound and if for any lower bound $x_l \in X$, $x_l \leq \inf_{x_t \in X_s} x_t$;
- the element $x_u \in X$ is called an *upper bound* of the subset X_s if, for all $x_t \in X_s$, $x_t \leq x_u$; $\sup_{x_t \in X_s} x_t \in X$ is called the *supremum* or the *lowest upper bound* of X_s if it is an upper bound and if for any upper bound $x_u \in X$, $\sup_{x_t \in X_s} x_t \leq x_u$.

Note that by definition, the minimum and the maximum belong to the subset $X_s \subseteq X$. The infimum and the supremum belong to the set X but may not belong to the subset $X_s \subseteq X$. The literature is not always clear on the set membership of the minimum, maximum, infimum, or the supremum.

Example 17.1.10. *Infimum and supremum of an interval of the real numbers.* Consider the set of the real numbers. and the open interval $(0,1) \subset \mathbb{R}$. Then $\inf_{x \in (0,1)} x = 0 \notin (0,1)$ and $\sup_{x \in (0,1)} x = 1 \notin (0,1)$.

Functions

Definition 17.1.11. Consider two sets X and Y. A relation $f \subseteq X \times Y$ of these sets is called a *function* if for every $x \in X$ there exists a *unique* $y \in Y$ such that $(x, y) \in f$. Denote a function then as $f : X \to Y$ with $y = f(x)$ rather than $f \subseteq X \times Y$. Call f also a *function* or a *map* from the set X to the set Y.

Call of a function $f : X \to Y$, X the *domain*, Y the *range*, and the set $f(X) = \{y \in Y | \exists\, x \in X, \text{ such that } y = f(x)\}$, the *image* of f from X.

The function $f : X \to Y$ is called *surjective* with respect to Y if for all $y \in Y$ there exists a $x \in X$ such that $y = f(x)$, hence $Y = f(X)$. It is called *injective* if for all $x_1, x_2 \in X$, $f(x_1) = f(x_2)$ implies that $x_1 = x_2$.

Consider two functions $g : U \to X$ and $h : X \to Y$. Define the *composition* of g followed by h as the function $f : U \to Y$, $f(u) = h(g(u))$. Call then f a *composite function*. It then follows that the composition operation is *associative*, $k \circ (h \circ g) = (k \circ h) \circ g$ for all functions $k,\ h,\ g$ between the appropriate spaces.

Definition 17.1.12. Consider a function $f : X \to Y$. Define a *left-inverse* of f as a function $g_L : Y \to X$ such that for all $x \in X$, $g_L(f(x)) = x$. Define a *right-inverse* of f as a function $g_R : Y \to X$ such that for all $y \in X$, $f(g_R(y)) = y$. Define an *inverse function* of f as a function $g : Y \to X$ which is both a left-inverse and a right-inverse of f. If an inverse function exists then call the function f *invertible*. Denote an *inverse function* of f by $f^{-1} : Y \to X$.

Proposition 17.1.13. *Consider a function $f : X \to Y$. There exists a left-inverse of f if and only if f is injective. There exists a right-inverse of f if and only if f is surjective. There exists an inverse of f if and only if f is bijective.*

Example 17.1.14. Consider the bilinear function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = ax + b$, with $a,\ b \in \mathbb{R}$ and $a \neq 0$. Then f is injective as is proven according to $f(x_1) = f(x_2)$ implies that $ax_1 + b = ax_2 + b$, $a(x_1 - x_2) = 0 \Rightarrow\ x_1 = x_2$ because $a \neq 0$. The inverse function is then, $y = f(x) = ax + b$, $ax = y - b$, $f^{-1}(y) = x = (y - b)/a$.

17.2 Algebraic Structures

Several algebraic structures are defined. These structures are also used in Chapter 18 for positive matrices.

Definition 17.2.1. An *additive monoid* is a nonempty set X with an operation written as addition, $+ : X \times X \to X$, such that all of the following conditions hold: (1) *associativity*: for all $x,\ y,\ z \in X$, $(x+y)+z = x+(y+z)$; (2) *existence of an additive identity element*: there exists a particular element called *zero* denoted by $0 \in X$ such that, for all $x \in X$, $x + 0 = x$ and $0 + x = x$. Denote an additive monoid by $(X, +, 0)$.

A *multiplicative monoid* is a nonempty set X and a function called the *product* denoted by $\times : X \times X \to X$, such that all of the following conditions hold: (1) *associativity*: for all $a,\ b,\ c \in X$, $a \times (b \times c) = (a \times b) \times c$; and (2) *existence of a multiplicative identity element*: there exists a particular element called *one* or *unit* denoted by $1 \in X$ such that, for all $x \in X$, $x \times 1 = x$ and $1 \times x = x$. If a unit exists then it is unique hence one writes *the unit*. Denote a multiplicative monoid by $(X, \times, 1)$.

Definition 17.2.2. A *group* $(G, \times)$, interpreted as a multiplicative group, consists of a nonempty set G and a function $\times : G \times G \to G$ called *product* such that: (1) *associativity* holds: $\forall\, a, b, c \in G$, $(a \times b) \times c = (a \times b) \times c$; (2) *existence of an identity element*: there exists a $1 \in G$ such that for all $a \in G$, $1 \times a = a = a \times 1$; and (3) *existence of an inverse*: $\forall\ a \in G$, $\exists\ b \in G$ such that $a \times b = 1 = b \times a$. Denote the multiplicative inverse of a by $b = a^{-1}$. Denote a group by $(G, \times, 1)$.

A *commutative group* is defined to be a group $(G, \times, 1)$ which satisfies: (4) *commutativity*: for all $x,\ y \in G$, $x \times y = y \times x$. A *commutative group* is also called an *Abelian group* after the Norwegian-born mathematician N.H. Abel (1802–1829).

Example 17.2.3. Consider the set of the integers $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. The additive identity is $0 \in \mathbb{Z}$. The set of the integers with the addition operation, $(\mathbb{Z}, +, 0)$, is a group.

The tuple $(\mathbb{Z}, \times, 1)$ is a multiplicative monoid but not a group because there does not exist a multiplicative inverse. Note that $2 \in \mathbb{Z}$ while $1/2 \notin \mathbb{Z}$.

The set of the rational numbers is defined as a quotient of integers,

$$\mathbb{Q} = \{\frac{p}{q} \in \mathbb{R}|\ p, q \in \mathbb{Z},\ q \neq 0\}.$$

Then $(\mathbb{Q}, +, 0)$ is a group. Also, $(\mathbb{Q}, \times, 1)$ is a multiplicative monoid, but not a group due to the element $0/q \in Q$ which does not admit an inverse in $\mathbb{Q}$.

The set of the real numbers with the addition operation $+$ is a group, $(\mathbb{R}, +)$.

The set of the complex numbers with the addition operation is a group, $(\mathbb{C}, +)$.

Example 17.2.4. The tuple $(\mathbb{R}^{k \times m}, +, 0)$ with $k,\ m \in \mathbb{Z}_+$ consisting of $k \times m$ matrices of elements of the real numbers with the addition operation and the zero matrix, is an additive group. Here $0 \in \mathbb{R}^{k \times m}$ denotes the zero matrix.

The tuple of square matrices $(\mathbb{R}^{k \times k}, \times, I)$ is a multiplicative monoid but it is not commutative: there exist matrices $A,\ B \in \mathbb{R}^{k \times k}$ such that $A \times B \neq B \times A$. Here $I \in \mathbb{R}^{k \times k}$ denotes the identity matrix.

The set of nonsingular square matrices $(\mathbb{R}^{n \times n}_{nsng}, \times, I)$ for $n \in \mathbb{Z}_+$, is a group.

Definition 17.2.5. A *subgroup* $(H, \times_H)$ of a group $(G, \times_G)$ consists of a set H and the operation $\times : H \times H \to H$ such that: (1) $H \subseteq G$; (2) $\times_H \subseteq \times_G$; (3) $1 \in H \subset G$; (4) $a, b \in H$ implies that $a \times b \in H$; (5) $a \in H$ implies that $a^{-1} \in H$.

Definition 17.2.6. A *ring* $(R, +, \times, 0, 1)$ consists of a set R and operations $+, \times : R \times R \to R$ called respectively *addition* and *multiplication*, such that:

1. *Associativity with respect to addition*: $\forall\ a,\ b,\ c \in R$, $(a+b)+c = a+(b+c)$;
2. *Commutativity with respect to addition*: $\forall\ a,\ b \in R$, $a+b = b+a$;
3. *Existence of an additive identity*: $\exists\ 0 \in R$ such that, $\forall\ a \in R$, $a+0 = a$;
4. *Existence of an additive inverse*: $\forall\ a \in R$, $\exists\ b \in R$ such that $a+b = 0$.
5. *Associativity with respect to multiplication*:
   $\forall\ a,\ b,\ c \in R$, $(a \times b) \times c = a \times (b \times c)$;
6. *Multiplication* distributes *over addition*:
   $\forall\ a,\ b,\ c \in R$, $a \times (b+c) = (a \times b) + (a \times c)$;

A *commutative ring* is a ring which is commutative with respect to multiplication: $\forall\ a,\ b \in R$, $a \times b = b \times a$.

A *commutative ring with a multiplicative identity* is a commutative ring for which there exists an element $1 \in R$ such that $\forall\ a \in R$, $a \times 1 = a$.

A commutative ring with a muliplicative identity is not required to have a multiplicative inverse.

Example 17.2.7. The set of square matrices with elements in the real numbers, $\mathbb{R}^{n\times n}$, is a ring. It is not a commutative ring because in general for two matrices $A,\ B\in\mathbb{R}^{n\times n}$, $A\times B\neq B\times A$. It has a multiplicative identity, the identity matrix $I\in\mathbb{R}^{n\times n}$.

Definition 17.2.8. A *semi-ring* $(R,+,\times,0,1)$ consists of a set R and functions $+,\times : R\times R\to R$ called respectively *addition* and *multiplication*, such that:

1. *Associativity with respect to addition* holds: $\forall\ a,\ b,\ c\in R$, $(a+b)+c=a+(b+c)$;
2. *Commutativity with respect to addition* holds: $\forall\ a,\ b\in R$, $a+b=b+a$;
3. *Existence of an additive identity*: $\exists\ 0\in R$ such that for all $a\in R$, $a+0=a$;
4. *Associativity with respect to multiplication* holds: $\forall\ a,\ b,\ c\in R$, $(a\times b)\times c=a\times(b\times c)$;
5. *Existence of a multiplicative identity*: $\exists\ 1\in R$ such that for all $a\in R$, $a\times 1=a=1\times a$;
6. Multiplication *distributes* over addition: $\forall\ a,\ b,\ c\in F$, $a\times(b+c)=(a\times b)+(a\times c)$;

However, there is neither a condition on the existence of an additive inverse, nor the existence of a multiplicative inverse, nor multiplicative commutativity holds.

A *commutative semi-ring* is a semi-ring which is commutative with respect to multiplication: $\forall\ a,\ b\in R$, $a\times b=b\times a$.

Example 17.2.9. The set of the positive real numbers $(\mathbb{R}_+,+,\times,0,1)$ is a commutative semi-ring. It is not a ring because the additive inverse is $-x$, which exists in the real numbers, does not exist in the positive real numbers. It does not have a multiplicative inverse because the real number $0\in\mathbb{R}_+$ does not have an inverse in that set. It is a multiplicatively commutative set.

Example 17.2.10. The set of square positive matrices $\mathbb{R}_+^{n\times n}$ is a semi-ring. The set does neither admit an additive inverse nor does it satisfy the commutativity relation with respect to multiplication.

Definition 17.2.11. Define an *integral domain* $(R,+,\times,0,1)$ to be a commutative ring such that $\forall\ a,b\in\mathbb{R}$, $a\times b=0$ implies that either $a=0$ or $b=0$.

Example 17.2.12. Examples of integral domains are: the set of the integers $\mathbb{Z}$ and the set of the real numbers $\mathbb{R}$.

The set of matrices $\mathbb{R}^{n\times n}$ is not an integral domain as the following example shows,

$$0=\begin{pmatrix}0&0\\0&0\end{pmatrix}=\begin{pmatrix}1&0\\0&0\end{pmatrix}\begin{pmatrix}0&0\\0&2\end{pmatrix}=A\times B.$$

Definition 17.2.13. A *field* $(F,+,\times,0,1)$ consists of a nonempty set F and the operations $+,\times : F\times F \to F$ called respectively *addition* and *multiplication* such that the following properties all hold:

1. *Associativity with respect to addition*: $\forall\ a,b,c \in F$, $(a+b)+c = a+(b+c)$;
2. *Commutativity with respect to addition*: $\forall\ a,b,c \in F$, $a+b = b+a$;
3. *Existence of an additive identity*: $\exists\ 0 \in F$ such that, $\forall\ a \in F$, $a+0 = a$;
4. *Existence of an additive inverse*: $\forall\ a \in F$, $\exists\ b \in F$, such that $a+b = 0$; denote $b = -a$;
5. *Associativity with respect to multiplication*: $\forall\ a,b,c \in F$, $(a\times b)\times c = a\times(b\times c)$;
6. *Commutativity with respect to multiplication*: $\forall\ a,b,c \in F$, $a\times b = b\times a$;
7. *Existence of a multiplicative identity*: $\exists\ 1 \in F$ such that, $\forall\ a \in F$, $a\times 1 = a$;
8. *Existence of a multiplicative inverse*: $\forall\ a \in F\backslash\{0\}$, $\exists\ c \in F$ such that $a\times c = 1$; if c is unique then denote $c = a^{-1}$; and
9. *distributivity of multiplication over addition*:
   $\forall\ a,b,c \in F$, $a\times(b+c) = (a\times b)+(a\times c)$.

An alternative definition is that a field is a nontrivial commutative ring in which, for every nonzero element, there exists an inverse element with respect to multiplication.

Example 17.2.14. Examples of fields are the set of the real numbers $\mathbb{R}$ and the set of the complex numbers, $\mathbb{C}$. In general, the set of the rational numbers, is also a field,

$$\begin{aligned}&\mathbb{Q} = \{\frac{p}{q} \in \mathbb{R}|\ p,q \in \mathbb{Z},\ q \neq 0\};\\ &\qquad \forall\ p/q \in \mathbb{Q} \text{ with } p/q \neq 0,\ \Rightarrow\ (q/p)\times(p/q) = 1.\end{aligned}$$

17.3 Linear Algebra and Linear Dependence

The aim of this section is to provide the readers of these notes with formulations of concepts and of results of linear algebra as far as is useful for this book. In a subsequent section matrices are discussed.

Definition 17.3.1. A *vector space V over a field F* is an algebraic structure of the form,

$$((F,+_F,\times_F,0_F,1_F),(V,+_V,0_V),\times_{sc}),$$

also denoted by (F,V) consisting of (1) a *set of scalars* denoted by F, which has the algebraic structure of a field with two binary operations: addition $+_F : F\times F \to F$ and multiplication $\times_F : F\times F \to F$, and with the special elements zero and one respectively denoted by 0_F and 1_F;
(2) a *set of vectors*, denoted by V, which has the algebraic structure defined below with the binary operation called *vector addition* $+_V : V\times V \to V$ and the additive identity element denoted by 0_V; an element of the set of vectors is called a *vector*;

(3) an operation relating F and V, $\times_{sc} : F \times V \to V$ called the *scalar product* or *scalar multiplication*; such that the following conditions all hold:

- *Additive commutativity.* $\forall\, x,y \in V,\, x +_V y = y +_V x.$
- *Additive associativity.* $\forall\, x,y,z \in V,\, (x +_V y) +_V z = x +_V (y +_V z).$
- *Existence of an additive identity in* V; there exists an element in V, denoted by 0_V, such that $\forall\, x \in V$, $x +_V 0 = x$.
- *Existence of an additive inverse of* V. For all $x \in V$ there exists a unique element in V, denoted by $-x$, such that $x +_V (-x) = 0_V$.
- *Associativity of scalar multiplication over the field* F. $\forall\, a,b \in F,\ x \in V,$ $(a \times_F b) \times_{sc} x = a \times_{sc} (b \times_{sc} x).$
- *Distributivity* of scalar multiplication over the field and the vector space,
$$\forall\, x,y \in V,\ a \in F,\ a \times_{sc} (x +_V y) = a \times_{sc} x +_V a \times_{sc} y;$$
$$\forall\, x \in V,\ a,b \in F,\ (a +_F b) \times_{sc} x = a \times_{sc} x +_V b \times_{sc} x.$$
- *Operation of the multiplicative identity of the field in scalar multiplication.* If $1 \in F$ is the multiplicative identity of F then, $\forall\, x \in V$, $1 \times_{sc} x = x$.

Below the elements $0_F,\ 1_F,\ 0_V$ and 0_V are denoted without the subscript if no confusion is possible. Similarly, the operations $+_F,\ \times_F,\ +_V,\ \times_{sc}$ will be denoted without subscript.

Example 17.3.2. Example. The vector space of n-tuples of a field. Let $n \in Z_+$. Let $(F,+,\times)$ be a field. Consider the structure $(F^n,F,+,\times)$ in which F^n consists of n-tuples of the field F with the operations of addition and of scalar multiplication defined as,

$$+ : F^n \times F^n \to F^n,\ \times : F \times F^n \to F^n,$$
$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix},\ y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},\ x+y = \begin{pmatrix} x_1+y_1 \\ x_2+y_2 \\ \vdots \\ x_n+y_n \end{pmatrix},\ a \times x = \begin{pmatrix} a \times x_1 \\ a \times x_2 \\ \vdots \\ a \times x_n \end{pmatrix}.$$

Then $((F,+_F,\times_F,0_F,1_F),(F^n,+_{F^n},0_{F^n}),\times_{sc})$ is a vector space. The shorter notation of $(F,\ F^n)$ will be used also.

Examples of vector spaces of n-tuples over a field are $(\mathbb{R},\ \mathbb{R}^n)$ and $(\mathbb{C},\ \mathbb{C}^n)$.

Definition 17.3.3. Let $((F,+_F,\times_F,0_F,1_F),(V,+_V,0_V),\times_{sc})$ be a vector space and let $W \subseteq V$ be a subset. Then $((F,+_F,\times_F,0_F,1_F),W,+_W,0_w),\times_{sc})$ is said to be a *vector subspace* of (F,V) if it is a vector space itself and if the operations of addition and of scalar multiplication of W are inherited from those of V by restriction.

Example 17.3.4. *Vector subspace*. Let $(F,+,\times,0,1)$ be a field. Consider the vector space $(F,V) = ((F,+_F,\times_F,0_F,1_F),(F^n,+_{F^n},0_{F^n}),\times_{sc})$ for some $n \in Z_+$. Define $W = \{x \in F^n | x_1 = 0\}$. Then $((F,+_F,\times_F,0_F,1_F),(W,+_W,0_w),\times_{sc})$ is a vector subspace of (F,V).

Definition 17.3.5. Let (F,V) and (F,W) be two vector spaces over the same field F. Consider a map $A : V \to W$. One says that A is a *linear transformation*, or a *linear map*, or a *linear function*, if

$$(\forall x,y \in V,\ \forall\ a,\ b \in F),\ \ A(a \times x + b \times y) = a \times A(x) + b \times A(y). \tag{17.1}$$

Equivalently, if $\forall\ x,\ y \in V,\ \forall\ a \in F,\ A(x+y) = Ax + Ay$ and $A(ax) = aAx$.

Definition 17.3.6. Let (F,V) and (F,W) be two vector spaces and $A : V \to W$ be a linear map. Define the *null space* of A, also called the *kernel* of A, and the *range space* of A, also called the *image* 0f A, respectively as,

$$N(A) = \ker(A) = \{v \in V |\ A(v) = 0_W\},$$
$$R(A) = \mathrm{Im}(A) = \{w \in W |\ \exists\ v \in V, \text{ such that } w = A(v)\}.$$

Proposition 17.3.7. *Consider two vector spaces V and W and a linear map $A : V \to W$. The map A is injective if and only if $N(A) = \{0_V\}$ if and only if A_b has full column rank. The map A is surjective if and only if $R(A) = W$ if and only if A_b has full row rank. The map A is invertible if and only if $N(A) = \{0_V\}$ and $R(A) = W$.*

Linear Dependence and Basis

Definition 17.3.8. Consider a vector space (F,V) and a finite and nonempty set of vectors $V_1 \subset V$.

One says that V_1 is a *linearly dependent set of vectors* of (F,V) if there exist $k \in \mathbb{Z}_m$ distinct elements $x_1,\ldots,x_k \in V_1$ and there exist $a_1,\ldots,a_k \in F$ at least one of which is nonzero, such that $\sum_{i=1}^{k} a_i x_i = 0$.

One says that V_1 is a *linearly independent set of vectors* if,

$$\forall\ x_1,\ldots,x_k \in V_1,\ \ a_1,\ldots,a_k \in F,\ \ \sum_{i=1}^{k} a_i x_i = 0\ \Rightarrow\ a_1 = a_2 = \ldots = a_n = 0.$$

Definition 17.3.9. Let $V_1 \subseteq V$ be a subset of V. The *span* of V_1 is defined as

$$\mathrm{span}(V_1) = \{\sum_{i=1}^{k} a_i x_i \in V |\ \exists\ k \in \mathbb{Z}_+,\ \exists\ x_1,\ldots,x_k \in V_1,\ \exists\ a_1,\ldots,a_k \in F\}.$$

Thus it is defined as the collection of all finite linear combinations of vectors in V_1.

Definition 17.3.10. A *basis* of the vector space (F,V) is a set of vectors $V_B \subset V$ such that: (1) V_B is a linear independent set of vectors; (2) $\mathrm{span}(V_B) = V$. The elements of V_B are in this case called the *basis vectors* of V. Denote the basis as
$V_B = \{x_i \in V,\ \forall i \in I \subseteq \mathbb{Z}_+\}$.

A consequence of the above definition is that if $V_B \subset V$ is a finite basis of V with the presentation above and if $x \in V = \mathrm{span}(V_B)$, then there exist scalars $a_1,\ldots,a_n \in F$

such that, $x = \sum_{i=1}^{n} a_i x_i$. The n-tuple $\{a_1, \ldots, a_n\}$ is called the representation of x with respect to the basis V_B. This representation is not unique in general.

Definition 17.3.11. A vector space (F,V) is called *n-dimensional*, for $n \in \mathbb{Z}_+ \cup \{+\infty\}$, if there exists a basis V_B of V which has n vectors and there does not exist a basis with strictly less than n vectors. If $n < \infty$ then V is called *finite-dimensional*; otherwise, when $n = +\infty$, it is called *infinite-dimensional*.

A linear map $A : V \to W$ between two vector spaces can then be presented in terms of the bases of V and W by a matrix. This is well known and not detailed in this book.

17.4 Matrices

Linear Transformation and Their Matrix Representations

. A linear map $A : V \to W$ between two vector spaces can be presented in terms of the bases of V and W by a matrix A_b according to $A(v) = A_b v$. This is well known and not detailed in this book.

Definition 17.4.1. (a)A matrix $B \in \mathbb{R}^{n \times m}$ is said to have *full column rank* if the columns of the matrix B are linearly independent. Equivalently, if for all vectors $v \in \mathbb{R}^m$, $Bv = 0 \;\Rightarrow\; v = 0$.
(b)A matrix $C \in \mathbb{R}^{m \times n}$ is said to have *full row rank* if the rows of the matrix C are linearly independent. Equivalently, if for all vectors $y \in \mathbb{R}^m$, $y^T C = 0 \;\Rightarrow\; y = 0$.

Proposition 17.4.2. *Consider two vector spaces over the real numbers* $(F_V, V) = (\mathbb{R},\ \mathbb{R}^k)$ *and* $(F_W, W) = (\mathbb{R},\ \mathbb{R}^m)$ *for* $k,\ m \in \mathbb{Z}_+$*, and a linear map* $A : V \to W$*. Denote the matrix representation of the linear map A with respect to bases of V and W by the matrix* $A_b \in \mathbb{R}^{k \times m}$*.*

The map A is injective if and only if $m \le k$ *and* $\operatorname{rank}(A_b) = m$*. The map A is surjective if and only if* $k \le m$ *and* $\operatorname{rank}(A_b) = k$*. The map A is bijective if and only if* $k = m$ *and* $\operatorname{rank}(A_b) = m = k$*. In the latter case there exists an inverse map* $A^{-1} : W \to V$ *which is represented with respect to the basis by the inverse matrix* A_b^{-1}*.*

Lemma 17.4.3. *(a)Consider a matrix* $B \in \mathbb{R}^{n \times m}$*. The matrix B has full column rank if and only if* $B^T B \succ 0$*; or, equivalently, if for all* $u \in \mathbb{R}^m$*,* $u \neq 0 \;\Rightarrow u^T B^T B u > 0$*.*
(b)Consider a matrix $C \in \mathbb{R}^{m \times n}$*. The matrix C has full row rank if and only if* $CC^T \succ 0$*; or, equivalently, if for all* $y \in \mathbb{R}^m$*,* $y \neq 0 \;\Rightarrow y^T CC^T y > 0$*.*

Proof. (a) $(\Rightarrow)$ Suppose that $B^T B \not\succ 0$. Then there exists a vector $u \in \mathbb{R}^m$ which is not identically zero, $u \neq 0$, such that $u^T B^T B u = 0$. Hence $Bu = 0$. Because the matrix B has full column rank it then follows that $u = 0$. This is a contradiction of the supposition.

($\Leftarrow$) Let $u \in \mathbb{R}^m$ be such that $Bu = 0$. Then $u^T B^T Bu = 0$. It follows from the assumption that $B^T B \succ 0$ that $u = 0$. Thus B has full column rank.
(b) The proof can be constructed from the proof of (a) by transposition. □

Multiplicative Factorization of a Matrix

Theorem 17.4.4. Singular value decomposition.
Consider a real-valued nonsquare matrix $A \in \mathbb{R}^{k\times m}$ *for* $k,\ m \in \mathbb{Z}_+$. *Then the matrix* A *admits a* singular value decomposition*,*

$$\exists\, U \in \mathbb{R}^{k\times k}_{ortg},\ \exists\, V \in \mathbb{R}^{m\times m}_{ortg},\ \exists\, S \in \mathbb{R}^{k\times m}_{+},\ \exists\, D \in \mathbb{R}^{n\times n}_{s+diag,+},\ \textit{such that,}$$
$$A = USV^T,\ \ S = \begin{pmatrix} D\,0 \\ 0\ \ 0 \end{pmatrix},$$
$$D = \text{Diag}(d) = \text{Diag}(d_1,\ d_2,\ \ldots,\ d_n),\ \ 1 \geq d_1 \geq d_2 \geq \ldots \geq d_n > 0.$$

In general $0 \leq n \leq \min\{k,m\}$. *If* $n = k < m$ *then the block row with zeroes is missing while if* $n = m < k$ *then the column with zeroes is missing. In general the decomposition is not unique.*

Classically the concept of a rank of a matrix is defined for the row-echelon form of that matrix. Below it is related to the singular values of a matrix. The reason to do so is that by now the singular value decomposition of a matrix is well known while few researchers still know the construction of a row-echelon form.

Example 17.4.5. Consider a matrix $A \in \mathbb{R}^{k\times m}$ for integers $k,\ m \in \mathbb{Z}_+$, $k \leq m$. Consider its singular value decomposition of the form,

$$A = USV^T,\ U \in \mathbb{R}^{k\times k}_{ortg},\ V \in \mathbb{R}^{m\times m}_{ortg},$$
$$S = \begin{pmatrix} D\ 0 \end{pmatrix},\ D = \text{Diag}(d_1, d_2, \ldots, d_k) \in \mathbb{R}^{k\times k},$$
$$r = \begin{cases} 0, \text{ if } A = 0, \\ k, \text{ if } d_1 \geq d_2 \geq \ldots \geq d_k > 0, \\ n, \text{ if } d_1 \geq \ldots d_n > 0 = d_{n+1} = \ldots = d_k. \end{cases}$$

Then $n = \text{rank}(A)$. The example illustrates that for a linear map described by a matrix, the set of singular values is more expressive than the rank.

The following special result is needed in the body of the book.

Proposition 17.4.6. *Consider a matrix* $A \in \mathbb{R}^{k\times m}$ *for* $k,m \in \mathbb{Z}_+$ *and* $k \leq m$. *If* $\text{rank}(A) = n < k$ *then,*

$$\exists\, B \in \mathbb{R}^{k\times m},\ \exists\, V \in \mathbb{R}^{m\times m}_{ortg}\ \textit{such that}\ \text{rank}(B) = k,$$
$$A\,V(\mathbb{R}^n \oplus 0_{k-n} \oplus 0_{m-k}) = B\,V(\mathbb{R}^n \oplus 0_{k-n} \oplus 0_{m-k}).$$

Proof. The singular value decomposition of A has the form,

$$\begin{aligned}
A &= U \begin{pmatrix} D_n & 0 \\ 0 & 0 \end{pmatrix} V^T = \begin{pmatrix} U_{11}D_nV_{11} & U_{11}D_nV_{12} & U_{11}D_nV_{13} \\ U_{21}D_nV_{11} & U_{21}D_nV_{12} & U_{21}D_nV_{13} \end{pmatrix}, \\
U &= \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix},\ V^T = \begin{pmatrix} V_{11} & V_{12} & V_{13} \\ V_{21} & V_{22} & V_{23} \end{pmatrix}, \\
D_n &= \text{Diag}(d_{n,1}, d_{n,2}, \ldots, d_{n,n}),\ 1 > d_{n,1} \geq d_{n,2} \geq \ldots \geq d_{n,n} > 0; \text{ define,} \\
D_{k-n} &= \text{Diag}(d_{k-n,1}, \ldots, d_{k-n,k-n}),\ 1 > d_{k-n,1} \geq \ldots \geq d_{k-n,k-n} > 0; \\
B &= U \begin{pmatrix} D_n & 0 & 0 \\ 0 & D_{k-n} & 0 \end{pmatrix} V^T,\ \text{rank}(B) = k, \\
& A\, V(\mathbb{R}^n \oplus 0_{k-n} \oplus 0_{m-k}) = U \begin{pmatrix} D_n & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} (\mathbb{R}^n \oplus 0_{k-n} \oplus 0_{m-k}) \\
&= U \begin{pmatrix} D_n & 0 & 0 \\ 0 & D_{k-n} & 0 \end{pmatrix} (\mathbb{R}^n \oplus 0_{k-n} \oplus 0_{m-k}) = B\, V(\mathbb{R}^n \oplus 0_{k-n} + \oplus 0_{m-k}).
\end{aligned}$$

A choice for D_{k-n} is to take for all $i \in \mathbb{Z}_{k-n}$, $d_{k-n,i} = \sum_{j=1}^{n} d_{n,j}/n$. Because of the distribution of the values of the elements of $(d_{n,1}, \ldots, d_{n,n})$, other choices can be considered. □

Square Matrices

Definition 17.4.7. The set of matrices with entries in the real numbers with k rows and m columns for $k,\ m \in \mathbb{Z}_+$ is denoted by $\mathbb{R}^{k \times m}$. Denote the (i, j)-th element of a matrix $A \in \mathbb{R}^{k \times m}$ for $i \in \mathbb{Z}_k$ and for $j \in \mathbb{Z}_m$ by $A_{i,j}$ or A_{ij}. A matrix in $\mathbb{R}^{k \times k}$ for $k \in \mathbb{Z}_+$ is called a *square matrix*. The corresponding set of matrices with entries in the complex numbers is denoted by $\mathbb{C}^{k \times m}$.

A matrix $A \in \mathbb{R}^{n \times n}$ is called *diagonal* if $\forall i, j \in Z_n,\ i \neq j \mapsto A_{ij} = 0$; or, equivalently, if all its off-diagonal elements are zero. Denote several sets of diagonal matrices by, $\mathbb{R}^{n \times n}_{diag}$, $\mathbb{R}^{n \times n}_{s+,diag}$, and $\mathbb{R}^{n \times n}_{st,diag}$ depending on whether the diagonal elements are respectively real numbers, or strictly positive real numbers, or the elements of a probability vector.

Define the *signature matrix* of a square matrix $A \in \mathbb{R}^{n \times n}$, regarded now as related to a system matrix of a continuous-time system, as the diagonal matrix of the form,

$$A = \begin{pmatrix} I_{n_+} & 0 & 0 \\ 0 & 0_{n_0} & 0 \\ 0 & 0 & -I_{n_-} \end{pmatrix} \in \mathbb{R}^{n \times n}, \text{ denoted by } n_{sgn}(A) = (n_+, n_0, n_-).$$

Determinant, Trace, and Norm

Definition 17.4.8. Define the *determinant* of a square matrix $A \in \mathbb{R}^{n \times n}$ recursively in $n \in \mathbb{Z}_+$ as,

$$\begin{aligned}
\det(A) &= A_{1,1}, \text{ if } n = 1, \\
&\quad \forall\, n \geq 2,\ \forall\, i,\ j \in \mathbb{Z}_n,\ \text{define the } \textit{minor} \text{ of } A_{i,j} \text{ as,} \\
M_{i,j} &= (-1)^{i+j} \det(A|_{\mathbb{Z}\setminus\{i\}\times\mathbb{Z}_n\setminus\{j\},}) \text{ where,} \\
&\quad A|_{\mathbb{Z}\setminus\{i\}\times\mathbb{Z}_n\setminus\{j\},} \in \mathbb{R}^{(n-1)\times(n-1)}, \\
&\quad \text{is obtained from } A \text{ by deleting row } i \text{ and column } j, \\
\det(A) &= \sum_{j=1}^{n} A_{i,j} M_{i,j}, \ \text{ which is independent of } i \in \mathbb{Z}_n.
\end{aligned}$$

Define the *trace* of a square matrix $A \in \mathbb{R}^{n\times n}$ by the formula $\text{tr}(A) = \sum_{i=1}^{n} A_{ii}$.

Example 17.4.9. If $A \in \mathbb{R}$ then $\det(A) = A_{1,1}$. If $B \in \mathbb{R}^{2\times 2}$ then,

$$\begin{aligned}
\det(B) &= \det\begin{pmatrix} B_{1,1} & B_{1,2} \\ B_{2,1} & B_{2,2} \end{pmatrix} = B_{1,1}\det(B_{2,2}) - B_{1,2}\det(B_{2,1}) \\
&= B_{1,1}B_{2,2} - B_{1,2}B_{2,1}; \\
\text{tr}(B) &= B_{1,1} + B_{2,2}.
\end{aligned}$$

Theorem 17.4.10. The determinant and the trace of special matrices.
Consider square matrices $A,\ B \in \mathbb{R}^{n\times n}$.

(a)$\det(I_n) = 1$.
(b)$\det(AB) = \det(A)\det(B)$.
(c)If the matrix $A \in \mathbb{R}^{n\times n}_{nsng}$*, hence* $\det(A) \neq 0$*, then* $\det(A^{-1}) = 1/\det(A)$.
(d)For square matrices A_{11} *and* A_{22} *there holds,*

$$\det\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} = \det(A_{11}) \times \det(A_{22}).$$

(e)If $D \in \mathbb{R}^{n\times n}_{diag} = \text{Diag}(d_1,\ldots,d_n)$ *then* $\det(D) = \prod_{i=1}^{n} d_i$.
(f) If $H \in \mathbb{R}^{n_1\times n_2}$ *and* $J \in \mathbb{R}^{n_2\times n_1}$ *for* $n_1,\ n_2 \in \mathbb{Z}_+$ *then* $\text{tr}(HJ) = \text{tr}(JH)$.

Proposition 17.4.11. Determinants of special matrices arising in information theory.

(a)For all $v,\ w \in \mathbb{R}^n$, $\det(I + vw^T) = 1 + w^T v$.
(b)For any $D \in \mathbb{R}^{n\times n}_{s+,diag}$ *with* $D = \text{Diag}(d_1,\ldots,d_n)$ *and for any* $v,\ w \in \mathbb{R}^n$,

$$\det(D + vw^T) = (\prod_{i=1}^{n} d_i)[1 + \sum_{j=1}^{n} \frac{v_j w_j}{d_j}].$$

Proof. Note that,

$$D + vw^T = D^{1/2}[I + D^{-1/2}vw^T D^{-1/2}]D^{1/2},$$

and the result then follows from Theorem 17.4.10.(b) and from part (a). □

Definition 17.4.12. *Norms of a real vector space*. Consider the vector space of tuples of the real numbers of the form $(F,V) = (\mathbb{R},\mathbb{R}^n)$ for an integer $n \in \mathbb{Z}_+$. A *norm* of this vector space is a map $\|.\| : V \to \mathbb{R}_+$ such that the following three conditions all hold:

1. $\|x\| = 0$ if $x = 0$ and $\|x\| > 0$ if $x \neq 0$;
2. for all $x \in V$ and for all $a \in \mathbb{R}$, $\|ax\| = |a|\ \|x\|$; and
3. for all $x,\ y \in \mathbb{R}^n$ the triangle inequality holds: $\|x+y\| \leq \|x\| + \|y\|$.

The norms $\|.\|_a,\ \|.\|_b$ are called *equivalent norms* if

$$\exists\ c_{min},\ c_{max} \in \mathbb{R}_{s+},\ \text{such that}\ \forall\ x \in V,\ c_{min}\|x\|_a \leq \|x\|_b \leq c_{max}\|x\|_a;$$
$$\text{if so then}\ (1/c_{max})\|x\|_b \leq \|x\|_a \leq (1/c_{min})\|x\|_b.$$

The latter inequalities show that the concept of equivalent norms is a symmetric relation.

On the real vector space $(\mathbb{R}, \mathbb{R}^n)$ define the functions,

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|,\ \ \|x\|_2 = (\sum_{i=1}^{n} |x_i|^2)^{1/2},\ \ \|x\|_q = (\sum_{i=1}^{n} |x_i|^q)^{1/q},\ \forall\ q \in (0,\infty),$$
$$\|x\|_\infty = \max_{i \in \mathbb{Z}_n} |x_i|.$$

Define the *inner product* on $(\mathbb{R}, \mathbb{R}^n)$ as the function, $(x,\ y) = \sum_{i=1}^{n} x_i\, y_i$.

Theorem 17.4.13. *The functions* $\|.\|_1,\ \|.\|_2,\ \|.\|_q$ *for* $q \in \mathbb{R}_{s+}$ *, and* $\|.\|_\infty$ *are norms. The norms* $\|.\|_1,\ \|.\|_2,\ \|.\|_\infty$ *are equivalent with the inequalities,*

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\ \|x\|_2,\quad \|x\|_\infty \leq \|x\|_1 \leq n\ \|x\|_\infty,$$
$$\frac{1}{\sqrt{n}}\|x\|_1 \leq \|x\|_2 \leq \|x\|_1,\quad \|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\ \|x\|_2,$$
$$\frac{1}{n}\|x\|_1 \leq \|x\|_\infty \leq \|x\|_1,\quad \frac{1}{\sqrt{n}}\|x\|_2 \leq \|x\|_\infty \leq \|x\|_2.$$

Definition 17.4.14. *Matrix norms*. Define a *matrix norm* as a function $\|.\|$ of a square matrix in $\mathbb{R}^{n\times n}$ for an integer $n \in \mathbb{Z}_+$ such that the following conditions all hold:

1. $\|A\| = 0$ if $A = 0$ and $\|A\| > 0$ if $A \neq 0$;
2. $\forall\ A \in \mathbb{R}^{n\times n}$ and $\forall\ a \in \mathbb{R}$ there holds $\|aA\| = |a|\ \|A\|$;
3. $\forall\ A,\ B \in \mathbb{R}^{n\times n}$, $\|A+B\| \leq \|A\| + \|B\|$;
4. $\forall\ A,\ B \in \mathbb{R}^{n\times n}$, $\|A \times B\| \leq \|A\| \times \|B\|$.

Call a matrix norm a *natural matrix norm* associated with a vector norm on $\mathbb{R}^n$ if the following supremization is attained,

$$\|A\|_{nm} = \sup_{z \in \mathbb{R}^n,\ \|z\|=1} \|Az\| = \max_{z \in \mathbb{R}^n,\ \|z\|=1} \|Az\|. \tag{17.2}$$

For a vector norm and its associated natural matrix norm the following inequality holds, $\|Ax\| \leq \|A\|_{nm}\|x\|$.

Theorem 17.4.15. *If the maximum is attained in equation (17.2) then* $\|.\|_{nm}$ *is a matrix norm.*

The natural matrix norm associated with the corresponding vector norms are,

$$\|A\|_1 = \max_{j\in\mathbb{Z}_n} \sum_{i=1}^{n} |A_{i,j}|, \ \ \|A\|_2 = \max_{\lambda\in\text{spec}(A^TA)} |\lambda|^{1/2}, \ \ \|A\|_\infty = \max_{i\in\mathbb{Z}_n} \sum_{j=1}^{n} |A_{i,j}|.$$

$\|A\|_2$ *is called the* spectral norm *of* A.

For any natural matrix norm, $\text{specrad}(a) \leq \|A\|_{nm}$*. There exists matrices for which* $\text{specrad}(a) < \|A\|_2$ *thus the spectral radius can be strictly smaller than the spectral norm of a matrix.*

Spectral Theory

Definition 17.4.16. Consider a square matrix $A \in \mathbb{R}^{n\times n}$ for an integer $n \in \mathbb{Z}_+$.

An *eigenvalue* of A is an element $\lambda \in \mathbb{C}$ such that $\det(A-\lambda I) = 0$. The *spectrum* of a matrix is defined to be the set of all eigenvalues of the matrix,

$$\text{spec}(A) = \{\lambda \in \mathbb{C} | \ \det(\lambda I - A) = 0\}.$$

The *spectral radius* of a matrix A is the maximum of the absolute modulus of an eigenvalue of a matrix A, $r(A) = \max_{\lambda\in\text{spec}(A)} |\lambda| \in \mathbb{R}_+$.

Define the *spectral index*, also called the *signature*, of a square matrix, regarded as the system matrix of a discrete-time system, as the triple of integers,

$$\begin{aligned} &\forall\, A \in \mathbb{R}^{n\times n}, \ \ n_{si}(A) = (n_+, n_1, n_-) \in \mathbb{N}^3, \ n = n_+ + n_1 + n_-, \\ n_+ &= \text{number of eigenvalues } \lambda \in \text{spec}(A) \text{ such that } |\lambda| > 1, \\ n_1 &= \text{number of eigenvalues } \lambda \in \text{spec}(A) \text{ such that } |\lambda| = 1, \\ n_- &= \text{number of eigenvalues } \lambda \in \text{spec}(A) \text{ such that } |\lambda| < 1. \end{aligned}$$

A matrix $A \in \mathbb{R}^{n\times n}$ is called (discrete-time) *exponentially stable* if, $\text{spec}(A) \subset \text{D}_o = \{c \in \mathbb{C} ||c| < 1\}$. Equivalently, its spectral index has the value $n_{si}(A) = (0,0,n)$ or if its spectral radius satisfies $r(A) < 1$.

Definition 17.4.17. Let $K, L \in \mathbb{R}^{n\times n}$.

(a) A *generalized eigenvalue* of K, L is an element $\lambda \in \mathbb{C}$ such that $\det(K - \lambda L) = 0$.

(b) A *generalized eigenvector* associated with a generalized eigenvalue $\lambda \in \mathbb{C}$ of K, L, is a vector $x \in \mathbb{C}^n$ such that $(K - \lambda L)x = 0$.

(c) A *chain of generalized eigenvectors of order k,* with $k \in \mathbb{Z}_+\backslash\{1\}$, associated with a generalized eigenvalue $\lambda \in \mathbb{C}$ of K, L is a set of vectors $\{x_1, x_2, \ldots, x_k\} \subset \mathbb{C}^n$, such that $(K - \lambda L)x_1 = 0$; for all $i = 2, 3, \ldots, k$, $(K - \lambda L)x_i = Lx_{i-1}$.

The *matrix pencil* of the pair $(K, L) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$ is the function $f : \mathbb{C} \to \mathbb{C}^{n\times n}$, $f(\lambda) = K - \lambda L$.

Inverses of Nonsingular Square Matrices

Definition 17.4.18. A square matrix $A \in \mathbb{R}^{n\times n}$ is called *singular* if $\det(A) = 0$ and called *nonsingular* if $\det(A) \neq 0$. Denote the set of nonsingular square real matrices by $\mathbb{R}^{n\times n}_{nsng}$.

Definition 17.4.19. Consider a square matrix $A \in \mathbb{R}^{n\times n}$. A *left-inverse* of the matrix A is a square matrix $B_L \in \mathbb{R}^{n\times n}$ such that $B_L A = I_n$. A *right-inverse* of the matrix A is a matrix $B_R \in \mathbb{R}^{n\times n}$ such that $AB_R = I_n$.

Example 17.4.20. Consider the rather special matrix,

$$A = \begin{pmatrix} D | 0 \end{pmatrix} \in \mathbb{R}^{k\times n},\ n < k,\ k,\ n \in \mathbb{Z}_+,$$

$$D = \begin{pmatrix} d_1 & 0 & \ldots & 0 & 0 \\ 0 & d_2 & & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \ldots & d_{k-1} & 0 \\ 0 & 0 & \ldots & 0 & d_k \end{pmatrix} \in \mathbb{R}^{k\times k},\ d_1 \geq d_2 \geq \ldots \geq d_{k-1} \geq d_k > 0.$$

Note that by the conditions imposed above, there exists a matrix,

$$D^{-1} = \begin{pmatrix} 1/d_1 & 0 & \ldots & 0 & 0 \\ 0 & 1/d_2 & & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \ldots & 1/d_{k-1} & 0 \\ 0 & 0 & \ldots & 0 & 1/d_k \end{pmatrix} \in \mathbb{R}^{k\times k},\ I_k = DD^{-1} = D^{-1}D.$$

The matrix A has a right-inverse,

$$A_{i,r} = \begin{pmatrix} D^{-1} \\ 0 \end{pmatrix} \in \mathbb{R}^{n\times k},\ \ AA_{i,r} = I_k.$$

It does not admit a left-inverse because,

$$B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} \in \mathbb{R}^{n\times k}\ \Rightarrow\ BA = \begin{pmatrix} B_1 D & 0 \\ B_2 D & 0 \end{pmatrix} \neq I_n.$$

Correspondingly, the following matrix admits a left-inverse but no right-inverse,

$$C = \begin{pmatrix} D \\ 0 \end{pmatrix} \in \mathbb{R}^{n\times k},\ \ C_{i,l} = \begin{pmatrix} D^{-1} & 0 \end{pmatrix} \in \mathbb{R}^{k\times n},\ C_{i,l}C = I_k.$$

Proposition 17.4.21. *Consider a square matrix $A \in \mathbb{R}^{n\times n}$. If there exists both a left-inverse B_L and a right-inverse B_R of the matrix A then these are equal. Thus there exists a unique matrix which is both a left-inverse and a right-inverse.*

Definition 17.4.22. Consider a square matrix $A \in \mathbb{R}^{n\times n}$. If there exists both a left-inverse and a right-inverse of the matrix A then denote the unique inverse matrix by $A^{-1} \in \mathbb{R}^{n\times n}$. Then $A^{-1}A = I_n = AA^{-1}$.

Theorem 17.4.23. *Consider two square matrices $A,\ B \in \mathbb{R}^{n \times n}$ and assume that their inverses $A^{-1},\ B^{-1} \in \mathbb{R}^{n \times n}$ exist.*

(a)There exists an inverse of the matrix A^{-1} and that matrix satisfies $(A^{-1})^{-1} = A$.
(b)There exists an inverse of the matrix $A \times B$ and that inverse satisfies
$(A \times B)^{-1} = B^{-1} \times A^{-1}$.

Theorem 17.4.24. *Consider a square matrix $A \in \mathbb{R}^{n \times n}$. Then the inverse A^{-1} of A exists if and only if* $\det(A) \neq 0$ *if and only if A is nonsingular. Consequently, by logical negation, the inverse A^{-1} of A does not exists if and only if* $\det(A) = 0$ *if and only if A is singular.*

Definition 17.4.25. A matrix $U \in \mathbb{R}^{n \times n}$ is called *orthogonal* if $UU^T = I$. Note that then $U,\ U^T \in \mathbb{R}^{n \times n}_{nsng}$, and $X = U^T U$ implies that $XU^T = U^T UU^T = U^T$ hence $X = XU^T (U^T)^{-1} = U^T (U^T)^{-1} = I$ and $UU^T = I = X = U^T U$. An orthogonal matrix is thus non-singular. Denote the set of orthogonal matrices by $\mathbb{R}^{n \times n}_{ortg}$.

Example 17.4.26. Note that,

$$U = \begin{pmatrix} +1 & 0 \\ 0 & -1 \end{pmatrix} \in \mathbb{R}^{2 \times 2},\ U^2 = I,\ U^{-1} = U^T = U.$$

Consider two sets of orthogonal matrices in $\mathbb{R}^{2 \times 2}_{ortg}$,

$$\begin{aligned} U_+(a) &= \begin{pmatrix} \cos(a) & -\sin(a) \\ \sin(a) & \cos(a) \end{pmatrix} \in \mathbb{R}^{2 \times 2},\ \forall\, a \in [0, 2\pi); \\ & U_+(a)U_+(a)^T = I_2,\ \det(U_+(a)) = +1; \\ U_-(b) &= \begin{pmatrix} \cos(b) & \sin(b) \\ \sin(b) & -\cos(b) \end{pmatrix} \in \mathbb{R}^{2 \times 2},\ \forall\, b \in [0, 2\pi), \\ & U_-(b)U_-(b)^T = I_2,\ \det(U_-(b)) = -1. \end{aligned}$$

Proposition 17.4.27. *If $U \in \mathbb{R}^{n \times n}_{ortg}$ then either* $\det(U) = +1$ *or* $= -1$.

Proof. By definition of an orthogonal matrix, $UU^T = I$, hence $1 = \det(I)$ $= \det(UU^T) = \det(U)\det(U^T) = (\det(U))^2$ from which the conclusion follows. □

Lemma 17.4.28. Matrix inversion lemma. *Let $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times p}$, $C \in \mathbb{R}^{p \times p}$, $D \in \mathbb{R}^{p \times n}$. If A, D, and $A + BDC$ are nonsingular then,*

$$[A + BDC]^{-1} = A^{-1} - A^{-1}B[D^{-1} + CA^{-1}B]^{-1}CA^{-1}.$$

Symmetric Square Matrices

Definition 17.4.29. A matrix $Q \in \mathbb{R}^{n \times n}$ is called *symmetric* if $Q = Q^T$; or, equivalently, if $\forall\, i, j \in \mathbb{Z}_n,\ Q_{ij} = Q_{ji}$. Denote the set of symmetric matrices by $\mathbb{R}^{n \times n}_s$.

A symmetric matrix $Q \in \mathbb{R}^{n \times n}$ is called *positive-definite symmetric* if $\forall\ x \in \mathbb{R}^n,\ 0 \leq x^T Q x$, and this is denoted by $0 \preceq Q$; and called *strictly-positive-definite symmetric* if $\forall\ x \in \mathbb{R}^n,\ x \neq 0\ \Rightarrow\ 0 < x^T Q x$, and this is denoted by $0 \prec Q$. The terms of a matrix being *negative definite* and *strictly-negative definite* are defined correspondingly. Denote the set of symmetric positive-definite matrices by $\mathbb{R}^{n \times n}_{pds}$ and the strictly-positive-definite matrices by $\mathbb{R}^{n \times n}_{spds}$.

Define and denote the *partial-order relation of positive-definiteness* on the set of symmetric square-matrices by $Q_1 \preceq Q_2$ if $Q_1 - Q_2 \preceq 0$ and the *partial-order relation of strict-positive-definiteness* by $Q_1 \prec Q_2$ if $Q_1 - Q_2 \prec 0$.

Any positive-definite symmetric matrix admits a decomposition of the form, $Q = UDU^T$ with an orthogonal matrix $U \in \mathbb{R}^{n \times n}_{ortg}$ and a positive diagonal matrix $D = \text{Diag}(d) \in \mathbb{R}^{n \times n}_{diag,+}$ with $d \in \mathbb{R}^n_+$. This follows from the singular value decomposition for example. If in addition the matrix Q is strictly-positive definite then $D \in \mathbb{R}^{n \times n}_{s+,diag}$.

For a positive-definite matrix $Q \in \mathbb{R}^{n \times n}$ define its *square root* as the matrix: if $Q = UDU^T$ then $Q^{1/2} = UD^{1/2}U^T$ with $D = \text{Diag}(d_1, \ldots, d_n)$ and $d \in \mathbb{R}^n_+$, $D^{1/2} = \text{Diag}(d_1^{1/2}, \ldots, d_n^{1/2})$. Note that then, $Q^{1/2}Q^{1/2} = UD^{1/2}U^T UD^{1/2}U^T = UDU^T = Q$.

A particular form is needed for a positive-definite symmetric matrix which is used on the body of the book.

Proposition 17.4.30. *Consider the set $\mathbb{R}^{n \times n}_{pds}$ of positive-definite symmetric square matrices for an integer $n \in \mathbb{Z}_n$.*

For any positive-definite symmetric square matrix there exists a decomposition of the form,

$$\begin{aligned}
&\forall\ Q \in \mathbb{R}^{n \times n}_{pds},\ \exists\ U \in \mathbb{R}^{n \times n}_{ortg},\ D \in \mathbb{R}^{n \times n}_{+,diag},\ \textit{such that,} \\
Q &= UDU^T = (UD^{1/2})(UD^{1/2})^T,\ D = \begin{pmatrix} D_1 & 0 \\ 0 & 0 \end{pmatrix}, \\
&D_1 \in \mathbb{R}^{n_1 \times n_1}_{s+,diag},\ n_1 = \text{rank}(D_1) \in \mathbb{N}_n.
\end{aligned}$$

In general such a decomposition is not unique. For example, if $n_1 < n$ then the following transformation produces another decomposition,

$$\begin{aligned}
L &= \begin{pmatrix} I & 0 \\ 0 & L_2 \end{pmatrix} \in \mathbb{R}^{n \times n}_{ortg},\ \ L_2 \in \mathbb{R}^{(n-n_1)\times(n-n_1)}_{ortg}, \\
&L^T D L = D\ \Rightarrow \\
Q &= UDU^T = ULL^T DLL^T U^T = ULDL^T U^T = (ULD^{1/2})(ULD^{1/2})^T.
\end{aligned}$$

In addition, if the matrix D contains a block of the form,

$$D = \begin{pmatrix} D_1 & 0 & 0 \\ 0 & D_2 & 0 \\ 0 & 0 & D_3 \end{pmatrix}, \; D_2 = d_2 I_{n_2}, \; then,$$

$$U_2 \in \mathbb{R}_{ortg}^{n_2 \times n_2}, \; \overline{U} = \begin{pmatrix} I & 0 & 0 \\ 0 & U_2 & 0 \\ 0 & 0 & I \end{pmatrix} \; \Rightarrow \; \overline{U} D \overline{U}^T = D,$$

$$Q = (U\overline{U}) D (U\overline{U})^T.$$

The transformation of an arbitrary matrix $Q \in \mathbb{R}_{pds}^{n \times n}$ to the above decomposition is provided for example by the singular value decomposition,

$$Q = UDU^T.$$

Proof. This follows from the singular value decomposition. □

Any symmetric square matrix $Q \in \mathbb{R}^{n \times n}$, not necessarily positive-definite, may be decomposed as $Q = UD_sU^{-1}$, with $U \in \mathbb{R}_{nsng}^{n \times n}$ a nonsingular matrix and in which D_s is a diagonal matrix with, for all $i \in \mathbb{Z}_n$, $D_{s,ii} \in \{-1, 0, +1\}$, [8, Ch. X].

Definition 17.4.31. Define the *upper-diagonal canonical form of a strictly positive-definite symmetric matrix* by the Cholesky factorization according to the formulas,

$$Q = LD(LD)^T \in \mathbb{R}^{n \times n},$$
$$L \in \mathbb{R}_{updiag}^{n \times n}, \text{ with } +1 \text{ on the diagonal},$$
$$D = \text{Diag}(d) \in \mathbb{R}_{s+,diag}^{n \times n} \text{ a diagonal matrix such that } d \in \mathbb{R}_{s+}^n.$$

The elements of the matrices L and D are not further restricted than described above. Thus there are $n + (n-1)(n-2)/2$ independent parameters which describe the free elements of the matrices L and D. These parameters are the strictly-positive diagonal elements of D and the off-diagonal elements of the upper diagonal matrix L.

Proposition 17.4.32. *The* upper-diagonal canonical form of a strictly positive-definite symmetric matrix *is a well defined canonical form.*

Proof. (1) Let the matrix tuple (L, D) satisfy the definition of the candidate canonical form of Def. 17.4.31. Then $Q = LD(LD)^T$ is a strictly positive-definite symmetrix matrix.

(2) For any $Q \in \mathbb{R}_{spds}^{n \times n}$ there exists a Cholesky decomposition of the form $Q = LD(DL)^T$ where $L \in \mathbb{R}_{nsng}^{n \times n}$ is a nonsingular upper-diagonal matrix with every diagonal element equal to $+1$ and $D_Q \in \mathbb{R}_{s+,diag}^{n \times n}$ thus with strictly positive diagonal elements. This follows from [12, Cor. 7.2.9] if one extracts the diagonal elements from the upper diagonal matrix into the D matrix. Moreover, the quoted corollary states that the matrix LD is unique. Then $Q = LD(LD)^T = LD^2L^T$ belongs to the set of the canonical forms.

(3) As stated above, in the decomposition $Q = (LD)(DL)^T$ the product LD is unique. Because L has $+1$ on the diagonal, the matrices L and D are unique.

The points (1)-(3) establish that for any strictly positive-definite symmetric matrix Q there exists an unique decomposition in terms of the canonical form defined. Hence the defined form is a true canonical form. □

Proposition 17.4.33. Schur complement of a symmetric matrix *Consider the decomposed symmetric square matrix,*

$$Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{12}^T & Q_{22} \end{pmatrix} \in \mathbb{R}^{n\times n},\ Q_{22} \succ 0,$$
$$n_1,\ n_2 \in \mathbb{Z}_+,\ n = n_1 + n_2,\ Q_{11} \in \mathbb{R}^{n_1\times n_1},\ Q_{12} \in \mathbb{R}^{n_1\times n_2},\ Q_{22} \in \mathbb{R}^{n_2\times n_2}.$$

Define the matrix L displayed below. An algebraic calculation shows that,

$$L = \begin{pmatrix} I & -Q_{12}Q_{22}^{-1} \\ 0 & I \end{pmatrix},\ LQL^T = \begin{pmatrix} Q_{11} - Q_{12}Q_{22}^{-1}Q_{12}^T & 0 \\ 0 & Q_{22} \end{pmatrix} = \begin{pmatrix} \widetilde{Q} & 0 \\ 0 & Q_{22} \end{pmatrix}.$$

(a)The matrix L is well defined and nonsingular. Then the matrix LQL^T has the form displayed above.
(b)The matrix $\widetilde{Q} = Q_{11} - Q_{12}Q_{22}^{-1}Q_{12}^T \in \mathbb{R}_s^{n_1\times n_1}$ is well defined.
(c)$Q \succeq 0$ if and only if $\widetilde{Q} \succeq 0$.
(d)$Q \succ 0$ if and only if $\widetilde{Q} \succ 0$.
(e)Assume that $Q \succeq 0$ and $Q_{22} \succ 0$. Then $\text{rank}(Q) = n_2 + \text{rank}(\widetilde{Q})$.

The symmetric matrix $\widetilde{Q} \in \mathbb{R}^{n_1\times n_1}$ is called the Schur complement *of the matrix Q after the mathematician I. Schur (1875 – 1941).*

Proof. (a) - (d) By assumption the matrix $Q_{22} \succ 0$ hence the inverse matrix Q_{22}^{-1} exists. A calculation then determines the displayed form of LQL^T. Hence, because $Q_{22} \succ 0$, $Q \succeq 0$ if and only if $\widetilde{Q} \succeq 0$; and $Q \succ 0$ if and only if $\widetilde{Q} \succ 0$.
(e) Note that $\text{rank}(Q) = \text{rank}(LQL^T) = \text{rank}(Q_{22}) + \text{rank}(\widetilde{Q})$. □

Lemma 17.4.34. *Consider the matrices $Q_1,\ Q_2 \in \mathbb{R}^{n\times n}_{spds}$.*

(a)$Q_1 \succeq I$ if and only if $I \succeq Q_1^{-1}$.
(b)$Q_1 \succeq Q_2$ if and only if $Q_2^{-1} \succeq Q_1^{-1}$.

Proof. (a) Because $Q_1 \in \mathbb{R}^{n\times n}_{sspd}$ there exists a square root matrix $Q_1^{1/2} \in \mathbb{R}^{n\times n}_{spds}$ such that $Q_1 = Q_1^{1/2}Q_1^{1/2}$. Then, $Q_1 \succeq I$ if and only if, by pre- and post-multiplication of $Q_1^{-1/2}$, $I = Q_1^{-1/2}Q_1Q_1^{-1/2} \succeq Q_1^{-1/2}Q_1^{-1/2} = Q_1^{-1}$.
(b) As in the proof of (a), there exists a square-root matrix $Q_2^{1/2} \in \mathbb{R}^{n\times n}_{spds}$. Then,

$$Q_1 \succeq Q_2 \Leftrightarrow Q_2^{-1/2}Q_1Q_2^{-1/2} \succeq I \Leftrightarrow I \succeq (Q_2^{-1/2}Q_1Q_2^{-1/2})^{-1}, \text{ by (a)},$$
$$\Leftrightarrow I \succeq Q_2^{1/2}Q_1^{-1}Q_2^{1/2} \Leftrightarrow Q_2^{-1} \succeq Q_1^{-1}.$$

□

Contra Gradient Transform

The following result on the contra gradient transform is needed elsewhere in this book.

Theorem 17.4.35. *Let $Q_1, Q_2 \in \mathbb{R}^{n\times n}_{s+,pds}$, thus $Q_1 = Q_1^T \succ 0$ and $Q_2 = Q_2^T \succ 0$.*

(a)

$$\begin{aligned} &\exists\, L \in \mathbb{R}^{n\times n}_{nsng},\ \exists\, D = \text{Diag}(d_1, d_2, \ldots, d_n) \in \mathbb{R}^{n\times n}_{s+,diag}, \\ &\textit{such that } d_1 \geq d_2 \geq \ldots \geq d_n > 0, \qquad (17.3) \\ LQ_1L^T &= L^{-T}Q_2L^{-1} = D. \qquad (17.4) \end{aligned}$$

Call then the matrix L the contra gradient transform *and D the* diagonal characteristic.

(b)If $L_1, L_2 \in \mathbb{R}^{n\times n}$ are two contragradient transforms satisfying the conditions of (a), then there exists an orthogonal matrix $U \in \mathbb{R}^{n\times n}_{ortg}$ such that,

$$\begin{aligned} L_1 &= D^{1/2}UD^{-1/2}L_2, \qquad (17.5) \\ UD^2 &= D^2U, \quad U^TU = I. \qquad (17.6) \end{aligned}$$

Conversely, if L_2 is a contragradient transform and U satisfies equation (17.6), then L_1, defined by equation (17.5), is also a contragradient transform.

(c)If for all $i, j \in \mathbb{Z}_n$, $i \neq j$ implies $d_i \neq d_j$, then U is a nonsingular signature matrix.

Procedure 17.4.36 Computation of a contragradient transform.
Data: $n \in Z_+$, $Q_1, Q_2 \in \mathbb{R}^{n\times n}$ satisfying $Q_1 = Q_1^T \succ 0$, and $Q_2 = Q_2^T \succ 0$.

1. *Determine $U_1 \in \mathbb{R}^{n\times n}$ orthogonal and $D_1 \in \mathbb{R}^{n\times n}_{s+}$ a diagonal matrix with the diagonal elements ordered in a decreasing way as in (17.3) such that* $Q_1 = U_1D_1^2U_1^T$.
2. *Determine $U_2 \in \mathbb{R}^{n\times n}_{ortg}$ and $D_2 \in \mathbb{R}^{n\times n}_{+}$ a diagonal matrix with the diagonal elements ordered in a decreasing way as in (17.3) such that* $D_1U_1^TQ_2U_1D_1 = U_2D_2^4U_2^T$.
3. *Compute $L = D_2U_2^TD_1^{-1}U_1^T$ and $D = D_2^2$.*
4. *Return (D, L).*

For a numerical procedure for the above defined computations see [17].

Proof. Of Theorem 17.4.35.
(a) Consider Procedure 17.4.36 and the notation defined there. Then,

$$\begin{aligned} LQ_1L^T &= (D_2U_2^TD_1^{-1}U_1^T)(U_1D_1^2U_1^T)(U_1D_1^{-1}U_2D_2), \\ &\quad \text{by definition of } L \text{ in Step 3 and by Step 1}, \\ &= D_2U_2^TU_2D_2 = D_2^2 = D, \\ L^{-T}Q_2L^{-1} &= (D_2^{-1}U_2^TD_1U_1^T)Q_2(U_1D_1U_2D_2^{-1}), \text{by definition of } L, \\ &= D_2^{-1}U_2^TU_2D_2^4U_2^TU_2D_2^{-1}, \text{by Step 2}, \\ &= D_2^2 = D. \end{aligned}$$

(b) Define,

$$\begin{aligned}
U &= D^{-1/2}L_1L_2^{-1}D^{1/2} \in \mathbb{R}^{n\times n}, \text{ which is nonsingular;}\\
D &= L_1Q_1L_1^T = L_1L_2^{-1}L_2Q_1L_2^TL_2^{-T}L_1^T\\
&= D^{1/2}D^{-1/2}L_1L_2^{-1}D^{1/2}D^{1/2}L_2^{-T}L_1^TD^{-1/2}D^{1/2} = D^{1/2}UU^TD^{1/2},\\
&\Rightarrow UU^T = D^{-1/2}DD^{-1/2} = I; \;\Rightarrow\; U^{-1} = U^T;\\
D &= L_1^{-T}Q_2L_1^{-1} = L_1^{-T}L_2^TL_2^{-T}Q_2L_2^{-1}L_2L_1^{-1}\\
&= D^{-1/2}D^{1/2}L_1^{-T}L_2^TD^{-1/2}D^2D^{-1/2}L_2L_1^{-1}D^{1/2}D^{-1/2}\\
&= D^{-1/2}U^{-T}D^2U^{-1}D^{-1/2},\\
&\Rightarrow U^TD^2U = D^2 \;\Rightarrow\; D^2U = U^{-T}D^2 = UD^2.
\end{aligned}$$

(c) This is a direct verification. □

17.5 Analysis

Definition 17.5.1. Consider a nonempty set X. A *distance* on X or, equivalently, a *metric*, is a map $d : X \times X \to \mathbb{R}_+$ which satisfies the conditions:

(a)*positivity*: $d(x,y) > 0$ if $x \neq y$ and $d(x,y) = 0$ if $x = y$.
(b)*symmetry* $\forall\, x,y \in X$, $d(x,y) = d(y,x)$;
(c)*triangle inequality* $\forall\, x,y,z \in X$, $d(x,y) + d(y,z) \geq d(x,z)$.

Define a *metric space* as a tuple (X,d) of a nonempty set X and a metric $d : X \times X \to \mathbb{R}_+$. In this book the symbol d is used for a distance and a metric rather than the symbol m, because the latter symbol is used for the mean of a random variable.

A *separable metric space* is a metric space for which there exists an at most denumerable dense subset.

Example 17.5.2. Define the *discrete metric space* for a set X and the metric $d(x,y) = +1$ if $x \neq y$ and $d(x,y) = 0$ if $x = y$. Then (X,d) is a metric space.

Example 17.5.3. Consider the set of the real numbers $\mathbb{R}$. Define the absolute-value function $d : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $d(x,y) = |x-y|$. Then $(\mathbb{R},d)$ is a metric space.

Example 17.5.4. Consider the vector space $(\mathbb{R},\mathbb{R}^n)$ of n-tuples of the real numbers for an integer $n \in \mathbb{Z}_+$. As described in Section 17.3, it is possible to define vector addition and scalar multiplication for these vectors.

Define the *Euclidian distance* $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ as,

$$d(x,y) = \left(\sum_{i=1}^{n} (x_i - y_i)^2\right)^{1/2} = \|x-y\|_2 = (x-y,x-y)^{1/2}.$$

Then $((\mathbb{R},\mathbb{R}^n),d)$ is a metric space.

In the above $(x,y) = \sum_{i=1}^n x_iy_i$ denotes the *inner product* of x and y and $\|x\|_2$ denotes the two-norm of $x \in \mathbb{R}^n$.

Convergence is a major concept of analysis which will not be summarized in this book. The definitions of a few concept follow. The concept of a Cauchy sequence is useful, it avoids the use of the limit of the sequence which is often unknown.

Definition 17.5.5. Consider a metric space (X,d). A *Cauchy sequence* denoted by $\{x_n \in X,\ n \in \mathbb{Z}_+\}$, is a sequence such that for all $\varepsilon \in (0,\infty)$ there exists an integer $k \in \mathbb{Z}_+$ such that for all $n,m \in \mathbb{Z}_+$, $n \geq k$ and $m \geq k$ imply that $d(x_n,x_m) < \varepsilon$.

Definition 17.5.6. A *complete metric space* is defined to be a metric space in which any Cauchy sequence is a convergent sequence.

A *complete separable metric space* is a separable metric space which is also a complete metric space.

Example 17.5.7. The metric space consisting of the real numbers $((\mathbb{R},\mathbb{R}),\|.\|_2)$, is a complete metric space.

The metric space consisting of the set of n-tuples of the real numbers, denoted by $((\mathbb{R},\mathbb{R}^n),\|.\|_2)$ for $n \in \mathbb{Z}_+$, is a complete metric space.

The following results are used in the body of the book.

Proposition 17.5.8. Sums of exponentials. *Consider a real number* $r \in (0,1)$. *Then,*

$$\sum_{s=0}^{t} r^s = \frac{1-r^{t+1}}{1-r},\ \sum_{s=0}^{\infty} r^s = \frac{1}{1-r},\ \sum_{s=t+1}^{\infty} r^s = \frac{r^{t+1}}{1-r}.$$

Proof. The first sum is directly proven by induction starting with $t=0$. Then,

$$\begin{aligned}
\sum_{s=0}^{\infty} r^s &= \lim_{t\to\infty} \sum_{s=0}^{t} r^s = \lim_{t\to\infty} \frac{1-r^{t+1}}{1-r} = \frac{1}{1-r};\\
\sum_{s=t+1}^{\infty} r^s &= [\sum_{s=0}^{\infty} r^s] - [\sum_{s=0}^{t} r^s] = \frac{1}{1-r} - \frac{1-r^{t+1}}{1-r} = \frac{r^{t+1}}{1-r}.
\end{aligned}$$

□

Proposition 17.5.9. *Consider a function* $f : \mathbb{Z}_+ \to \mathbb{R}_+$. *The following implication holds,*

$$\begin{aligned}
&\exists\, r \in (0,1),\ \exists\, c \in (0,\infty),\ \exists\, t_1 \in \mathbb{Z}_+,\ \exists\, \overline{f} \in \mathbb{R}_+,\\
&\textit{such that}\ \forall\, t \geq t_1,\ |f(t)-\overline{f}| \leq c\, r^{t-t_1}\\
\Rightarrow\ &\lim_{t\to\infty} f(t) = \overline{f},\ \lim_{t\to\infty} \frac{1}{t}\sum_{s=0}^{t-1} f(s) = \overline{f}.
\end{aligned}$$

Proof. (a) $\lim_{t\to\infty} |f(t)-\overline{f}| \leq \lim_{t\to\infty} c\, r^{t-t_1} = 0$ because $r \in (0,1)$. Note that,

$$\forall\, n,\ m \in \mathbb{N},\ m \leq n,\ \sum_{k=0}^{n} r^k = \frac{1-r^{n+1}}{1-r},\ \ \sum_{k=m}^{n} r^k = \frac{r^m - r^{n+1}}{1-r},$$

$$\forall\, t \geq t_1,\ \frac{1}{t}\sum_{s=0}^{t-1} |f(s)-\overline{f}| = \frac{1}{t}\sum_{s=0}^{t_1-1} |f(s)-\overline{f}| + \frac{1}{t}\sum_{s=t_1}^{t-1} |f(s)-\overline{f}|$$

$$\lim_{t\to\infty} \frac{1}{t}\sum_{s=0}^{t-1} |f(s)-\overline{f}| = \lim \frac{1}{t}\sum_{s=0}^{t_1-1} |f(s)-\overline{f}| + \lim \frac{1}{t}\sum_{s=t_1}^{t-1} |f(s)-\overline{f}|$$

$$\leq 0 + \lim_{t\to\infty} \frac{1}{t}\sum_{s=t_1}^{t-1} c\, r^{s-t_1} = \lim_{t\to\infty} \frac{cr^{t_1}}{t}\frac{r^{t+1}}{1-r^{t_1}} = 0.$$

17.6 Geometry

Geometry is useful besides algebra and analysis for the development of concepts for control theory. The focus in this section is on convex sets and on affine subsets.

Definition 17.6.1. *A convex set.* A set $X \subset \mathbb{R}^n$ for an integer $n \in \mathbb{Z}_+$ is said to be *convex set* if for any $x, y \in X$ and any $\lambda \in [0,1]$, $\lambda x + (1-\lambda)y \in X$.

There follow concepts and results for convex sets.

Proposition 17.6.2. *If $\{K_i \subseteq \mathbb{R}^n,\ \forall\, i \in I\}$ is an arbitrary family of convex sets, not necessarily a countable family, then the intersection $K = \cap_{i=1}^n K_i \subseteq \mathbb{R}^n$ is a convex set.*

Definition 17.6.3. Consider a subset $V \subseteq \mathbb{R}^n$ for an integer $n \in \mathbb{Z}_+$. The *convex hull* of V, denoted by $\text{convh}(V) \subseteq \mathbb{R}^n$, is defined as the intersection of all convex subsets of $\mathbb{R}^n$ which contain V. Because $\mathbb{R}^n$ is a convex set which contains V, the intersection is not empty. Hence the convex hull is well defined.

Definition 17.6.4. Define a *closed half space* of the vector space $\mathbb{R}^n$ for an integer $n \in \mathbb{Z}_+$, and parametrized by $(h,c) \in \mathbb{R}^n \times \mathbb{R}$, as the set,

$$H(h,c) = \{x \in \mathbb{R}^n |\ h^T x \leq c\},\ \ (h,c) \in \mathbb{R}^n \times \mathbb{R}.$$

Note that a plane in $\mathbb{R}^n$ is also a closed half space because,

$$\{x \in \mathbb{R}^n |\ h^T x = c\} = \{x \in \mathbb{R}^n |\ \ h^T x \leq c,\ h^T x \geq c\ \}$$
$$= \{x \in \mathbb{R}^n |\ h^T x \leq c,\ (-h)^T x \leq (-c)\ \}.$$

The representation of a closed half space in terms of linear inequalities is preferred over the representation with a combination of inequalities and equalities.

It is elementary to prove that a closed-half space is a convex set,

$$\forall\, x,\ \overline{x} \in H(h,c),\ \forall\, \lambda \in [0,1],$$
$$\Rightarrow h^T[\lambda x + (1-\lambda)\overline{x}] = \lambda(h^T x) + (1-\lambda)h^T\overline{x} \leq\ \lambda c + (1-\lambda)c = c,$$
$$\Rightarrow [\lambda x + (1-\lambda)\overline{x}] \in H(h,c).$$

Definition 17.6.5. A *polyhedron* or a *polyhedral set* $X_{ph} \subseteq \mathbb{R}^n$ for an integer $n \in \mathbb{Z}_+$ is defined to be the nonempty intersection of a finite number of closed half spaces. Equivalently, there exists an integer $m \in \mathbb{Z}_+$ and a finite set,

$$\{(h_i, c_i) \in \mathbb{R}^n \times \mathbb{R}, \ \forall \, i \in \mathbb{Z}_m\} \text{ such that,}$$
$$X_{ph} = \cap_{i \in \mathbb{Z}_m} H(h_i, c_i) \neq \emptyset; \text{ define the representation,}$$
$$X_{ph} = \{x \in \mathbb{R}^n | \, H^T x \leq c\}, \ H = \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_m \end{pmatrix}, \ c = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{pmatrix},$$

and where the vector inequality $H^T x \leq c$ is interpreted componentwise.

Define a *polytope* as a bounded polyhedron.

Example 17.6.6. Consider the polytope,

$$X = \mathbb{R}^2, \ h = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$
$$X_{ph} = \{x \in X | \, x_1 = h^T x = 0\} = \{x \in X | \, h^T x \leq 0, \ (-h)^T x \leq 0\}.$$

Thus X_{ph} corresponds to the vertical axis in the plane which is an unbounded subset of the plane. Hence there exists an unbounded polytope!

Additional theory on polyhedral sets and cones is provided in Section 18.4.

Affine Geometry

Definition 17.6.7. *Affine geometry*. Consider the vector space $(\mathbb{R}, \mathbb{R}^n)$ for an $n \in \mathbb{Z}_+$.

An *affine set* $\mathrm{X}_{\text{affst}}$, more accurately *an affine subset* of the considered vector space, is defined by the condition that, for all $x_1, \ x_2 \in \mathrm{X}_{\text{affst}}$, the line piece connecting these vectors belongs to the affine set; equivalently, if for all $\lambda \in [0,1] \subseteq \mathbb{R}$, $\lambda x_1 + (1-\lambda)x_2 \in \mathrm{X}_{\text{affst}}$.

An *affine space* of this vector space is defined by the condition that, for all $x_1, \ x_2 \in \mathrm{X}_{\text{affsp}}$, the entire line through these vectors belongs to the affine space; equivalently, if for all $\lambda \in \mathbb{R}$, $\lambda x_1 + (1-\lambda)x_2 \in \mathrm{X}_{\text{affsp}}$.

Define the *affine hull* of a subset $X_s \subseteq \mathbb{R}^n$ of the smallest affine subspace containing X_s. The existence of such affine hull can be proven.

A finite set of vectors $\{x_0, \ x_1, \ x_2, \ldots, x_k \in \mathbb{R}^n\}$ is said to be *affinely independent* if the following set of vectors is linearly independent,
$\{x_1 - x_0, \ x_2 - x_0, \ldots, x_{k-1} - x_0 \in \mathbb{R}^n\}$.

It follows from the definitions that an affine space is also an affine subset but there exists an example of an affine subset which is not an affine subspace.

Example 17.6.8. Consider the vector space $(\mathbb{R},\mathbb{R}^3)$ and the subset

$$X_1 = \begin{pmatrix}1\\2\\3\end{pmatrix} + \left\{ \begin{pmatrix}0&0\\1&0\\0&1\end{pmatrix} \begin{pmatrix}y_2\\y_3\end{pmatrix} \in \mathbb{R}^3 | \; \forall\, y_2,y_3 \in \mathbb{R} \right\}.$$

Then $X_1 \subseteq \mathbb{R}^3$ is an affine subspace. The elementary proof is omitted.

Proposition 17.6.9. *Consider a subset* $\mathrm{X}_{\mathrm{aff}} \subseteq \mathbb{R}^n$. *The following statements are equivalent:*

(a)$\mathrm{X}_{\mathrm{aff}}$ *is an affine subspace of* $\mathbb{R}^n$.
(b)There exists a vector $x_0 \in \mathrm{X}_{\mathrm{aff}}$ *and a linear subspace* $X_L \subseteq \mathbb{R}^n$ *such that,*

$$\mathrm{X}_{\mathrm{aff}} = x_0 + X_L = \{x_0 + x \in \mathbb{R}^n | \; \forall\, x \in X_L\}$$

(c)

$$\begin{aligned} &\exists\, m \in \mathbb{Z}_+,\; m \leq n, \exists\, A \in \mathbb{R}^{m\times n},\; \exists\, a \in \mathbb{R}^m, \textit{such that,} \\ \mathrm{X}_{\mathrm{aff}} &= \{x \in \mathbb{R}^n | \; Ax = a\}. \end{aligned}$$

Example 17.6.10. Consider the affine subspace,

$$\begin{aligned} x_0 &= \begin{pmatrix}1\\2\end{pmatrix},\; L = \left\{ \begin{pmatrix}x\\3x\end{pmatrix} \in \mathbb{R}^2 | \; \forall\, x \in \mathbb{R} \right\}, \\ \mathrm{X}_{\mathrm{aff}} &= x_0 + L = \left\{ \begin{pmatrix}1+x\\2+3x\end{pmatrix} \in \mathbb{R}^2 | \; \forall\, x \in \mathbb{R} \right\} \\ &= \left\{ y \in \mathbb{R}^2 | \; \begin{pmatrix}-3 & 1\end{pmatrix} y = -1 \right\} = \left\{ y \in \mathbb{R}^2 | \; Ay = a \right\},\; A = \begin{pmatrix}-3 & 1\end{pmatrix},\; a = -1. \end{aligned}$$

17.7 Optimization

The reader finds in this section elementary concepts and results on the infimization of real-valued functions.

Problem 17.7.1. *Unconstrained optimization problem.* Consider a real-valued function $f : X \subseteq \mathbb{R}^n \to \mathbb{R}_+$. Call X the *domain* of the function f. Define the *unconstrained infimization problem* by the expression,

$$J^* = \inf_{x \in D_f} f(x) \in \mathbb{R}_+.$$

The problem amounts to determine a *minimum* or *optimal value* $x^* \in D_f$ such that $J^* = f(x^*) \leq f(x)$ for all $x \in X$ and the *value* $J^* = f(x^*) \in \mathbb{R}_+$. A minimum need not be unique in general. Denote the *set of optimal values* by $X_{opt} = \{x \in X | \; f(x) = J^*\}$.

A *local minimum* is defined to be an element $x^*_{loc} \in X$ such that there exists an open set,

$$X_{loc} \subseteq X, \text{ such that } x^*_{loc} \in X_{loc} \text{ and } \forall\, x \in X_{loc},\; f(x^*_{loc}) \leq f(x).$$

A local minimum need not be unique either.

It may the case that there does not exist a vector x^* such that $J^* = f(x^*)$. In this case determine, for any $\varepsilon \in (0,\infty)$, an element $x_\varepsilon \in X$ such that $J^* < f(x_\varepsilon) < J^* + \varepsilon$.

In general, a minimum need not exist. An example of an optimization problem not admitting a minimal value is,

$$\inf_{x \in D_f = (0,1]} x = 0.$$

Note that the infimal value 0 does not belong to the set of admissible input values, $0 \notin (0,1]$. Therefore the concept of an ε-optimal value is necessary. Define,

$$x^*_\varepsilon = \varepsilon/2; \text{ then, } J^* = 0 = \inf_{x \in X = (0,1]} x < \varepsilon/2 = x^*_\varepsilon < J^* + \varepsilon.$$

Below the concept of a convex function is defined for functions on a continuous subset of tuples of the real numbers.

Definition 17.7.2. The function $f : X \subset \mathbb{R}^n \to \mathbb{R}$ is said to be a *convex function* if
(1) X is a convex set and
(2) $(\forall\, x,y \in X,\ \forall\, \lambda \in [0,1]),\ f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$.
It is said to be *strictly convex function* if:
(1) X is a convex set and
(2) if $(\forall\, x,\ y \in X,\ x \neq y,\ \lambda \in (0,1)),\ f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y)$.

The function $f : X \subset \mathbb{R}^n \to \mathbb{R}$ for a convex set X is said to be *concave function* if $(-f)$ is convex; and *strictly concave function* if $(-f)$ is strictly convex.

Proposition 17.7.3. *Consider an open convex set $X \subseteq \mathbb{R}^n$ and a function $f : X \to \mathbb{R}$ which is assumed to be twice continuously differentiable on its domain of definition. Define the* Hessian matrix *at the variable $y \in X$ as the map,*

$$H(y) = \frac{\partial^2 f(x)}{\partial x^2}|_{x=y} \in \mathbb{R}^{n \times n},\ \ H : X \to \mathbb{R}^{n \times n}.$$

From the theory of differentiation follows that, for fixed $y \in X$, the matrix $H(y)$ is symmetric. If, for all $y \in X$, the matrix $H(y)$ is positive definite then the function f is convex while if for all $y \in X$ the matrix $H(y)$ is strictly-positive-definite then the function f is strictly convex. Correspondingly, if for all $y \in X$ the matrix $H(y)$ is negative definite then the function f is concave etc.

Example 17.7.4. *A quadratic function is convex.* Let $f : \mathbb{R}^n \to \mathbb{R}$ be such that there exists a matrix $Q \in \mathbb{R}^{n \times n}_{pds}$ such that $f(x) = x^T Q x$. If $0 \preceq Q$ then f is a convex function; and if $0 \prec Q$ then f is strictly convex. These results follow directly from Proposition 17.7.3.

Example 17.7.5. *Natural logarithm strictly concave.*
Let $f : (0,\infty) \to \mathbb{R}$, $f(x) = \ln(x)$. The function f is strictly concave. This follows from Proposition 17.7.3 and the calculation,

$$J(x) = \frac{df(x)}{dx} = \frac{1}{x},\ H(x) = \frac{d^2 f(x)}{dx^2} = \frac{-1}{x^2} < 0,\ \forall\, x \in (0,\infty).$$

Example 17.7.6. *Convexity of a special function*. Define the function,

$$f(x) = x - \ln(x) - 1, \; f : (0,\infty) \to \mathbb{R}.$$

Then f is a strictly convex function.

This is directly proven because the set $(0,\infty)$ is a convex set and $df(x)/dx = 1 - 1/x$ and $d^2 f(x)/dx^2 = 1/x^2 > 0$ on the domain. The result then follows from Proposition 17.7.3.

Proposition 17.7.7. *Consider a function $f : \mathbb{R}^n \to \mathbb{R}_+$ which is continuous and strictly convex. Assume that a global minimum exists denoted by $x^* \in \mathbb{R}^n$.*

For any $\varepsilon \in (0,\infty)$, $\exists \, x_a \in \mathbb{R}^n$ such that $f(x^*) < f(x_a) < f(x^*) + \varepsilon$.

Proof. Consider an $\varepsilon \in (0,\infty)$. Define the sublevel set of the domain, $X(f(x^*)+\varepsilon) = \{x \in \mathbb{R}^n | \; f(x) < f(x^*) + \varepsilon\}$. Clearly $x^* \in X_\varepsilon$. The set X_ε is by definition an open set which contains the global minimum in its interior, due to the continuity of f. Also due to the continuity of f there exists a real number $r \in (0,\infty)$ such that $B(x^*,r) = \{x \in \mathbb{R}^n | \; \|x - x^*\| \leq r\} \subseteq X_\varepsilon$. Then $B(x^*,r)$ is a closed and bounded set, hence a compact set. If $X = \mathbb{R}$ it follows from the intermediate value theorem, [5, Thm. 3.15, Thm. 3.16], that the infimum and the supremum over $B(x^*,r)$ are attained. If $X = \mathbb{R}^n$ then the closed and bounded subset $B(x^*,r)$ is compact by [5, Cor. 6.64], $f(B(x^*,r))$ is a compact set, and, similar to [5, Cor. 6.57], the supremum and the infimum over $B(x^*,r)$ are attained. Denote by $x_{min}, \; x_{max} \in B(x^*,r)$ the values such that $f(x_{max}) = \max_{x \in B(x^*,r)} \; f(x)$ and $f(x_{min}) = \min_{x \in B(x^*,r)} \; f(x) = f(x^*)$. Because f is strictly convex, $f(x_{min}) < f(x_{max})$. Because of continuity of f, there exists a real number $y \in B(x^*,r)$ such that $f(x^*) < y < f(x_{max})$. Because f is a continuous function, there exists an element $x_a \in B(x^*,r)$ such that $f(x^*) < y = f(x_a) < f(x_{max}) \leq f(x^*) + \varepsilon$. □

If the set over which one optimizes is either finite or countable then convexity is not the proper concept. For those sets the concep of modularity and submodularity play a corresponding role.

17.8 Further Reading

Elementary algebra. Recommended is the book on algebra by G. Birkhoff and A. MacLane, [1], which is quite rich in concepts. Another book on elementary or universal algebra is [28]. The concept of a denumerable set is provided in [20, Def. 2.6].

Algebraic Structures A basic introduction to algebraic structures is the book [1]. At an advanced level see [2, 3, 13]. For the algebraic structures of a module and a ring see [14, 15]. For semi-rings see [9]. A history of algebra and mathematical structures is provided by Leo Corry, [6].

Linear algebra. Linear algebra at an introductory level is presented well in [25]. At a more advanced level is the book [11].

Matrices. Several standard textbooks are, [8, 16, 22]. Books on numerical linear algebra include [10, 18]. A standard reference on matrix analysis is [12]. A useful source for this section is [22].

Analysis. The author recommends the book [5]. At an advanced level see [7]. Measure theory in analysis is treated in [24].

Geometry – Convex Sets and Affine Sets. Convexity is well introduced in the book [4]. A classical book on convexity is [23].

Optimization. The author recommends the book by S. Boyd and L. Vandenberghe, [4]. The book of Y. Nesterov provides a different emphasis, [21]. Older books include [19, 23]. For modularity and submodularity, see [26, 27].

References

1. G. Birkhoff and A. MacLane. *A survey of modern algebra, fourth edition*. MacMillan Publ. Co. Inc., New York, 1977. 635
2. N. Bourbaki. *Elements of mathematics - Algebra I - Chapters 1-3*. Springer, Berlin, 1989. 635
3. N. Bourbaki. *Elements of mathematics - Algebra II - Chapters 4-7*. Springer, Berlin, 1990. 635
4. Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, U.K., (with corrections) edition, 2007. 426, 636
5. A. Browder. *Mathematical analysis - An introduction*. Undergraduate texts in mathematics. Springer-Verlag, New York, 1996. 30, 49, 424, 426, 475, 526, 635, 636, 677, 815
6. Leo Corry. *Modern algebra and the rise of mathematical structures*. Birkhäuser, Basel, 2004. 635
7. J. Dieudonné. *Foundations of modern analysis*. Academic Press, New York, 1969. 49, 636
8. F.R. Gantmacher. *The theory of matrices, volume 1, 2*. Chelsea Publ. Co., New York, 1959. 626, 636, 697
9. J.S. Golan. *The theory of semirings with applications in mathematics and theoretical computer science*. Longman, Harlow, 1992. 635
10. G.H. Golub and C.F. Van Loan. *Matrix computations*. The Johns Hopkins University Press, Baltimore, 1983. 636
11. P.R. Halmos. *Finite-dimensional vector spaces*. Springer, New York, 1993. 635, 801
12. R.A. Horn and C.R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, UK, 2nd. edition, 2013. 626, 636
13. N. Jacobson. *Basic algebra, Volumes 1 (2nd edition)*. W.H. Freeman and Company, New York, 1985. 635
14. T.Y. Lam. *A first course in noncommutative rings*. Number 131 in Graduate texts in mathematics. Springer, Berlin, 1991. 635
15. T.Y. Lam. *Lectures on modules and rings*. Number 189 in Graduate Texts in Mathematics. Springer, Berlin, 1999. 635
16. P. Lancaster and M. Tismenetsky. *The theory of matrices - Second edition with applications*. Academic Press, San Diego, 1985. 636, 849
17. A.J. Laub. Computation of balancing transformations. In *Proceedings 1980 Joint Automatic Control Conference*, pages Paper FA8–E, San Francisco, 1980. 628
18. Alan J. Laub. *Computational matrix analysis*. SIAM, Philadelphia, 2012. 636
19. D.G. Luenberger. *Optimization by vector space methods*. John Wiley & Sons, New York, 1969. 636
20. Y.N. Moschovakis. *Notes on set theory*. Undergraduate Text in Mathematics. Springer Verlag, Berlin, 1994. 635

21. Y. Nesterov. *Lectures on convex optimization*. Springer, Berlin, 2018. 636

22. B. Noble. *Applied linear algebra*. Prentice-Hall, Englewood Cliffs, NJ, 1969. 636, 663

23. R.T. Rockafellar. *Convex analysis*. Princeton University Press, Princeton, 1970. 636, 697

24. H.L. Royden. *Real analysis, 2nd edition*. MacMillan Co., New York, 1968. 49, 526, 636

25. G. Strang. *Linear algebra and its applications, 2nd. ed.* Academic Press, New York, 1980. 635

26. D.M. Topkis. Minimizing a submodular function on a lattice. *Operations Research*, 26:305–321, 1978. 636

27. D.M. Topkis. *Supermodularity and complementarity*. Princeton University Press, Princeton, 1998. 636

28. W. Wechler. *Universal algebra for computer scientists*. Springer-Verlag, Berlin, 1992. 635

Chapter 18
Appendix B Positive Matrices

Abstract This chapter concerns positive matrices which are matrices with elements of the positive real numbers. The motivation for the inclusion of the algebraic structure of positive matrices are the problems (1) of stability of the system of probability measures of the Markov process of a finite stochastic systems; and (2) of minimal positive matrix factorization motivated by the stochastic realization problem of an output-finite stochastic system. Using the concepts of similarity, of equivalence, of a prime in the positive matrices, and of an extremal cone, one can investigate a minimal positive matrix factorization. This appendix is to be considered as a reference chapter.

Key words: Positive matrices. Stochastic matrices. Similarity. Equivalence.

The state-transition matrix of a finite stochastic system is a stochastic matrix and hence a positive matrix. Positive matrices are used to determine the performance of finite stochastic systems, to characterize stochastic observability and stochastic controllability, and to derive results on stochastic realization of finite stochastic systems. Therefore concepts and results on positive matrices are used in this book.

The literature on positive matrices is vast but also spread out over many books and papers. The author has therefore decided to include in this book a chapter on positive matrices focused on the results needed in the body of the book.

18.1 Problems

The chapter is motivated by problems of finite stochastic system. In this section use is made of notation and of concepts only introduced in the subsequent sections. The reader should take this section as a motivation and come back to this section after reading other sections of this chapter. This chapter is motivated by the following subproblems.

Problem 18.1.1. Solve the following two subproblems for a positive matrix:

1. Determine necessary and sufficient conditions on a stochastic matrix for a sequence of probability measures on the positive vector space $\mathbb{R}_{st}^n$ generated by the stochastic matrix, to converge and determine its limit.
2. Classify or describe all positive factorizations of a positive matrix and from this deduce the positive rank of a positive matrix.

Below the above formulated problems are successively discussed in more detail.

Problem 18.1.2. *Existence of an invariant probability distribution and convergence of a sequence of probability distributions to an invariant distribution, for a finite state Markov process associated with a stochastic matrix.*

Consider a stochastic matrix $A \in \mathbb{R}_{st}^{n \times n}$ for an integer $n \in \mathbb{Z}_+$.

(a) Does there exist an eigenvector $p_s \in \mathbb{R}_{st}^n$ such that

$$Ap_s = p_s, \ 1_n^T p_s = 1?$$

Does there exists a unique such vector or, if there exist two or more, what are all of them? A vector p_s is characterized by the above equations will be called an *invariant distribtution* associated with the stochastic matrix A.

(b) Define the sequence of probability measures,

$$p(t+1) = Ap(t), \ p(0) = p_0 \in \mathbb{R}_{st}^n.$$

For which initial measures p_0 does the following limit exist,

$$\lim_{t \to \infty} p(t) = p_\infty \in \mathbb{R}_{st}^n.$$

If the limit exists, does the limit p_∞ then equal the invariant distribution $p_s \in \mathbb{R}_{st}^n$ defined in (a)?

The above problem is solved in Section 18.8

The second subproblem requires the introduction of the concept of a positive rank of a positive matrix.

Definition 18.1.3. Consider a positive matrix $A \in \mathbb{R}_+^{k \times m}$ for $k, \ m \in \mathbb{Z}_+$.

Define the *positive rank* of the matrix A as the *smallest integer* $n \in \mathbb{N} = \{0, 1, 2, \ldots\}$ such that there exists positive matrices $B \in \mathbb{R}_+^{k \times n}$ and $C \in \mathbb{R}_+^{n \times m}$ such that $A = B \times C$. Denote by $n = \text{pos} - \text{rank}(A)$ this integer.

Call any factorization of A of the form $A = B \times C$ with $B \in \mathbb{R}_+^{k \times n}$, $C \in \mathbb{R}_+^{n \times m}$, and with $n = \text{pos} - \text{rank}(A)$, a *minimal positive-matrix factorization* of A.

Most important is a positive-matrix factorization, the positive-rank is a by product of the factorization.

Problem 18.1.4. Consider a nonsquare positive matrix $A \in \mathbb{R}_+^{k \times m}$ for integers $k, \ m \in \mathbb{Z}_+$. Determine: (1) the positive rank $\text{pos} - \text{rank}(A)$; and (2) classify or describe all minimal positive-matrix factorizations of the matrix A.

A solution will be proposed for the minimal factorization of positive matrices using the analogon of the singular value decomposition of a real-valued matrix. Recall the singular value decomposition of a real-valued matrix, see Theorem 17.4.4,

$$\begin{aligned}
&\forall\, A \in \mathbb{R}^{k \times m},\ \exists\, U \in \mathbb{R}^{k \times k}_{ortg},\ V \in \mathbb{R}^{m \times m}_{ortg},\ \exists\, S \in \mathbb{R}^{k \times m},\\
&\exists\, n \in \mathbb{Z}_{\max\{k,m\}},\ D \in \mathbb{R}^{n \times n}_{s+,diag},\ \text{such that},\\
A &= USV^T = USV^{-1} = U \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} V^T.
\end{aligned}$$

It will be proven in Section 18.9.6 that a square positive matrix admits a decomposition of the form,

$$\begin{aligned}
A &= M_1 \begin{pmatrix} B & 0 \\ 0 & I_{n_2} \end{pmatrix} M_2 \in \mathbb{R}^{n \times n}_{+},\\
&n_1,\ n_2 \in \mathbb{Z}_n,\ n_1 + n_2 = n,\ M_1,\ M_2 \in \mathbb{R}^{n \times n}_{mon},\ \text{are monomial matrices},\\
&B \in \mathbb{R}^{n_1 \times n_1}_{dst},\ \text{is a doubly stochastic matrix, fully indecomposible,}\\
&\text{and a prime in the positive matrices,}\\
&I_{n_2} \in \mathbb{R}^{n_2 \times n_2}_{+}\ \text{is an identity matrix.}
\end{aligned}$$

The above decomposition is analogous to the singular valued decomposition of a real-valued matrix. The novel element is the first diagonal block of the decomposed matrix. A further decomposition of the doubly stochastic matrix will be provided in Section 18.9.6.

Example 18.1.5. An example of a decomposed positive matrix is the doubly stochastic matrix,

$$A = \frac{1}{2} \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \in \mathbb{R}^{4 \times 4}_{dst};\ \ \text{pos} - \text{rank}(A) = 4 > 3 = \text{rank}(A).$$

18.2 The Positive Real Numbers and a Positive Vector Space

The set of the positive real numbers requires a discussion. The definition of this set is rather delicate and the literature is not always clear on this. A discussion follows.

The reader is reminded of the adopted convention that a real number in the subset $\mathbb{R}_+ = [0, \infty) \subset \mathbb{R}$ is called a *positive real number* and an element of $\mathbb{R}_{s+} = (0, \infty) \subset \mathbb{R}$ is called a *strictly positive real number*. The term *nonnegative real number* is not used in this book.

Most readers think of the set of the positive real numbers $\mathbb{R}_+ \subset \mathbb{R}$ as a restriction of the set of the real numbers. Then $\mathbb{R}_+$ inherits from the real numbers the addition and the product operations, $+,\ \times : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$. Note that for any $x \in (0, \infty)$ there exists an element $x^{-1} \in (0, \infty) \subset \mathbb{R}_+$.

However, for any $x \in \mathbb{R}_+$, $-x \notin \mathbb{R}_+$. Therefore, subtraction is not defined in the positive real numbers as it is defined in the real numbers. Subtraction by multiplication is the procedure to be used. A subtraction operation can be defined according to,

$$x - y = \begin{cases} r \times x, & \text{if } x - y \geq 0, \text{ if } x > 0, \text{ and if } r = (x-y)/x, \\ \text{not defined, else.} \end{cases}$$

The set of the positive real numbers is algebraically different from the set of the real numbers. Below use is made of the concepts of associativity and of commutativity, see Def. 17.2.1.

Definition 18.2.1. Define the algebraic structure of the set of the *positive real numbers* as the subset of the real numbers $\mathbb{R}_+ \subset \mathbb{R}$ with the operations of: (1) *addition*: $+ : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ satisfying associativity, commutativity, and the existence of an additive neutral element $0 \in \mathbb{R}_+$ such that for all $x \in \mathbb{R}_+$, $x + 0 = x$; and (2) *multiplication* or *product*, $\times : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ satisfying associativity and commutativity; (3) the existence of a multiplicative neutral element $1 \in \mathbb{R}_+$ such that for all $x \in \mathbb{R}_+$, $1 \times x = x$; and (4) multiplication distributes over addition.

The algebraic structure of the positive real numbers is a *commutative semi-ring*, see Def. 17.2.8.

Geometrically the set of the positive real numbers $\mathbb{R}_+$ is a polyhedral cone with only one spanning vector, see the definition of a cone in Section 18.4.

The Positive Vector Space

The algebraic structure of the set of tuples of positive real numbers can be defined abstractly as a positive vector space.

Definition 18.2.2. *Positive vector space*. Define a *positive vector space* as an algebraic structure denoted by $((R, +_R, \times_R, 0_R, 1_R), (V, +_V, 0_V), \times_{sc})$ consisting of a set of *scalars* $(R, +_R, \times_R, 0_R, 1_R)$, a set of *vectors* $(V, +_V, 0_V)$, and the *scalar multiplication* $\times_{sc} : R \times V \to V$ satisfying:
(1) $(R, +_R, \times_R, 0_R, 1_R)$ is a semi-ring, see Def. 17.2.8; (2) $(V, +_V, 0_V)$ in which $+_V : V \times V \to V$ is a function which is commutative, associative, and there exists an additive identity 0_V; (3) the rules for scalar multiplication of vectors hold: (3.1) for all $a \in R$ and $x,\ y \in V$, $a \times_{sc} (x + y) = a \times_{sc} x + a \times_{sc} y$ and (3.2) for all $a,\ b \in R$ and $x \in V$, $(a + b) \times_{sc} x = a \times_{sc} x + b \times y$. Neither in R nor in V does there exist an additive inverse.

Below the multiplication operations $\times_R$ and $\times_{sc}$ are both denoted by $\times$ hence the reader has to determine from the context which operation is used. Similarly, the neutral elements 0_R, 0_V are both denoted by 0 and the neutral element 1_R is denoted by 1.

Definition 18.2.3. *Positive vector space of n-tuples of the positive real numbers*. Consider the positive integer $n \in \mathbb{Z}_+$. Denote the *positive vector space of n-tuples of the positive real numbers* by

$$\left((\mathbb{R}_+, +_{\mathbb{R}_+}, \times_{\mathbb{R}_+}, 0_{\mathbb{R}_+}, 1_{\mathbb{R}_+}), (\mathbb{R}_+^n, +_{\mathbb{R}_+}, 0_{\mathbb{R}_+^n}), \times_{sc} \right).$$

It is denoted also by $(\mathbb{R}_+, \mathbb{R}_+^n)$ to keep the list short. The algebraic structure of this positive vector space is defined as: (1) the set of scalars inherits its algebraic structure from the set of the positive real numbers as defined in Def. 18.2.1; (2) the set of vectors consists of n tuples of the positive realnumbers where $(\mathbb{R}_+^n, +_{\mathbb{R}_+}, 0_{\mathbb{R}_+^n})$ is an additive monoid, see Def. 17.2.1; (3) scalar-vector multiplication is defined components wise as in the set of the positive real numbers. Then this is a positive vector space as defined in Def. 18.2.2.

Note that a positive vector space is not a vector space as defined in Def. 17.3.1 because the set of the positive real numbers is not a field and the set of vectors does not admit an additive inverse.

Definition 18.2.4. One says that a positive vector $a \in \mathbb{R}_+^n$ for $n \in \mathbb{Z}_+$ is of *order* $k \in \mathbb{N}_n$ if exactly k elements are strictly positive, with the understanding that then $n-k$ elements are zero. Denote this property by $n_{s+}(a) = k$. Denote the *index set of strictly positive elements* of a vector $a \in \mathbb{R}_+$ by

$$i_{s+}(a) = \{j \in \mathbb{Z}_n | \; a_j \in \mathbb{R}_{s+}, \text{ or, equivalently, } a_j > 0\}.$$

Define the *one vector* $1_n \in \mathbb{R}_+^n$ as the vector of which all elements equal one and, for the size $n \in \mathbb{Z}_+$, denote this vector by $1_n \in \mathbb{R}_+^n$. When the size is clear from the context, the subindex n may be omitted.

Example 18.2.5. The concept of the order and of the index set of strictly positive elements of a positive vector are illustrated by two cases,

$$a = \begin{pmatrix} a_1 \\ a_2 \\ 0 \\ 0 \end{pmatrix} \in \mathbb{R}_+^4, \; a_1, \; a_2 \in \mathbb{R}_{s+}, \; n_{s+}(a) = 2, \; i_{s+}(a) = \{1,2\};$$

$$b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ 0 \end{pmatrix} \in \mathbb{R}_+^4, \; b_1, \; b_2, \; b_3 \in \mathbb{R}_{s+}, \; n_{s+}(b) = 3, \; i_{s+}(b) = \{1,2,3\}.$$

The concept of a simplex is of importance for finite-state Markov processes.

Definition 18.2.6. The *unit simplex and the probability simplex*. Define and denote the *unit simplex* in $\mathbb{R}_+^n$ for a $n \in \mathbb{Z}_+$, as the set,

$$\mathbb{R}_{st}^n = \{x \in \mathbb{R}_+^n | \; 1_n^T x \leq 1\}.$$

Define and denote the *probability simplex* in $\mathbb{R}_+^n$ for a $n \in \mathbb{Z}_+$, as the set,

$$\mathbb{R}_{st}^n = \{x \in \mathbb{R}_+^n | \; 1_n^T x = 1\}.$$

A probability distribution on a finite set with $n \in \mathbb{Z}_+$ elements is always a vector in the probability simplex $\mathbb{R}_{st}^n$. The simplex was historically defined in geometry as a unit simplex. In probability and stochastic processes one uses the probability simplex.

Example 18.2.7. The probability simplex in $\mathbb{R}_+^2$ is a line,

$$\mathbb{R}_{st}^2 = \left\{ x \in \mathbb{R}_+^2 | \; 1_2^T x = x_1 + x_2 = 1 \right\}.$$

Example 18.2.8. The probability simplex in $\mathbb{R}_+^3$ is a plane,

$$\mathbb{R}_{st}^3 = \left\{ x \in \mathbb{R}_+^3 | \; 1_3^T x = x_1 + x_2 + x_3 = 1 \right\}.$$

The two examples of simplices can easily be drawn on paper and it can be done also for $n = 4$, but a drawing is not so clearly done for $n \geq 5$.

Definition 18.2.9. Define the *decreasing arrangement* of the elements of a positive vector $x \in \mathbb{R}_+^n$ for $n \in \mathbb{Z}_+$ as the vector,

$$x_\downarrow = \begin{pmatrix} x_{[1]} \\ x_{[2]} \\ \vdots \\ x_{[n-1]} \\ x_{[n]} \end{pmatrix} \in \mathbb{R}_+^n, \quad x_{[1]} \geq x_{[2]} \geq x_{[3]} \geq \ldots \geq x_{[n-1]} \geq x_{[n]} \geq 0.$$

If the latter order relations are strict, equivalently, if $x_{[1]} > x_{[2]} > x_{[3]} > \ldots > x_{[n]}$, then there exists a unique permutation matrix $Q \in \mathbb{R}_{perm}^{n \times n}$ such that $x_\downarrow = Qx$. Otherwise the permutation matrix Q is not unique; in this case the set of permutation matrices achieving the transformation is simple to describe, it allows an extra permutation for each set of two or more order relations as stated above in which subsequent terms are related by equalities. The details are omitted.

Define the *majorization order* on the simplex $\mathbb{R}_{st}^n$ by the conditions:

$$x \succeq y \Leftrightarrow \begin{cases} \sum_{i=1}^k x_{[i]} \geq \sum_{i=1}^k y_{[i]}, & \text{if } k = 1, 2, \ldots, n-1, \\ \sum_{i=1}^n x_{[i]} = \sum_{i=1}^n y_{[i]} = 1, & \text{if } k = n. \end{cases}$$

One then says that *x majorizes y*.

The majorization order satisfies that (1) for all $x \in \mathbb{R}_{st}^n$, $x \succeq x$ and (2) it is transitive, hence it is an order relation, see Def. 17.1.6. However, the majorization order is not a total-order relation of the simplex because there exist vectors x and y such that neither $x \succeq y$ nor $y \succeq x$ hold.

Example 18.2.10. Consider the following positive vector and its decreasing arrangement,

$$x = \begin{pmatrix} 0.2 \\ 0.5 \\ 0.3 \end{pmatrix} \in \mathbb{R}_{st}^3, \; x_\downarrow = \begin{pmatrix} 0.5 \\ 0.3 \\ 0.2 \end{pmatrix} \in \mathbb{R}_{st}^3,$$
$$x_{[1]} = 0.5 > 0.3 = x_{[2]} > 0.2 = x_{[3]}.$$

Example 18.2.11. Consider the following positive vectors and their decreasing arrangements,

$$
\begin{aligned}
x &= \begin{pmatrix} 0.5 \\ 0.3 \\ 0.2 \end{pmatrix} \in \mathbb{R}^3_{st},\ y = \begin{pmatrix} 0.4 \\ 0.3 \\ 0.3 \end{pmatrix} \in \mathbb{R}^3_{st},\ \sum_{i=1}^{1} x_{[i]} = 0.5 > 0.4 = \sum_{i=1}^{1} y_{[i]}, \\
&\sum_{i=1}^{2} x_{[i]} = 0.8 > 0.7 = \sum_{i=1}^{2} y_{[i]},\ \sum_{i=1}^{3} x_{[i]} = 1.0 = 1.0 = \sum_{i=1}^{3} y_{[i]}, \\
&x \succeq y.
\end{aligned}
$$

That neither $x \succeq y$ nor $y \succeq x$ need to hold is seen in the following special case,

$$
\begin{aligned}
x &= \begin{pmatrix} 0.3 \\ 0.3 \\ 0.3 \\ 0.1 \end{pmatrix} \in \mathbb{R}^4_{st},\ y = \begin{pmatrix} 0.4 \\ 0.2 \\ 0.2 \\ 0.2 \end{pmatrix} \in \mathbb{R}^4_{st}, \\
&\sum_{i=1}^{1} x_{[i]} = 0.3 < 0.4 = \sum_{i=1}^{1} y_{[i]},\ \sum_{i=1}^{2} x_{[i]} = 0.6 = 0.6 = \sum_{i=1}^{2} y_{[i]}, \\
&\sum_{i=1}^{3} x_{[i]} = 0.9 > 0.8 = \sum_{i=1}^{3} y_{[i]},\ \sum_{i=1}^{4} x_{[i]} = 1.0 = 1.0 = \sum_{i=1}^{4} y_{[i]}, \\
&x \not\succeq y,\ y \not\succeq x.
\end{aligned}
$$

Theorem 18.2.12. *Consider two stochastic vectors $x,\ y \in \mathbb{R}^n_{st}$ for an $n \in \mathbb{Z}_+$. The following statements are equivalent:*

(a)$x \succeq y$;
(b)there exists a doubly stochastic matrix $A \in \mathbb{R}^{n \times n}_{dst}$ such that $y = Ax$;
(c)for all continuous convex functions f, and for all integers $k \in \mathbb{Z}_n$,
$\sum_{i=1}^{k} f(x_i) \geq \sum_{i=1}^{k} f(y_i)$.

18.3 Definitions of Positive Matrices

The reader finds in this section and the following sections many concepts and results for positive matrices used elsewhere in the book. The set of positive matrices is rich in algebra, graph theory, and geometry. The algebraic theory has advanced based on the research of O. Perron and of G. Frobenius. Both were mathematicians living in Germany early in the 20th century.

There follow definitions of several subsets of positive matrices.

Definition 18.3.1. Denote the set of *positive matrices* of size $n \times m$ with elements in the positive real numbers and with sizes $n,\ m \in \mathbb{Z}_+$ by $\mathbb{R}^{n \times m}_+$. Hence each element belongs to the set of the positive real numbers $\mathbb{R}_+ = [0, \infty) \subset \mathbb{R}$. The set of the positive matrices consists of $\mathbb{R}^{n \times m}_+$ for all integers $n,\ m \in \mathbb{Z}_+$.

Define the following subsets of positive matrices. Afterwards the notation of these subsets is introduced.

A square positive matrix $Q \in \mathbb{R}^{n \times n}_+$ of size n by n is called a *permutation matrix* if every row and every column contains exactly one element equal to one while all other elements in the row or column are equal to zero. A permutation matrix in this book is denoted by the symbol Q rather than by P because the symbol P is used for a probability measure. The set of permutation matrices in $\mathbb{R}_+^{n\times n}$ has $n!$ elements.

A square matrix is called a *diagonal positive matrix* of size n by n for $n \in \mathbb{Z}_+$ if it is a positive matrix and a diagonal matrix. It is called a *diagonal positive matrix with a strictly-positive diagonal* of size n by n if it is a diagonal matrix, and the diagonal elements are strictly positive.

A square positive matrix is called a *monomial matrix* of size n by $n \in \mathbb{Z}_+$ if every column contains exactly one element which is strictly positive while all other elements in the column are equal to zero, and if every row contains exactly one element which is strictly positive while all other elements in the row are equal to zero.

A positive matrix $A \in \mathbb{R}_+^{n\times m}$, not necessarily square, of sizes $n,\ m \in \mathbb{Z}_+$ with $m \leq n$, is called *part of a monomial* in $\mathbb{R}_+^{n\times n}$ if there exists a positive matrix $A_e \in \mathbb{R}_+^{n\times (n-m)}$ such that $\begin{pmatrix} A \; A_e \end{pmatrix} \in \mathbb{R}_+^{n\times n}$ is a monomial matrix as defined above. Alternatively, A is part of a monomial if every column of the matrix A contains exactly one strictly positive element and zeroes otherwise, and if every row of the matrix A contains at most one strictly positive element and zeroes otherwise. A positive matrix $G \in \mathbb{R}_+^{m\times n}$ with $m \leq n$ is also called *part of monomial* if G^T is part of a monomial as defined above. In the book [3, p. 67] a matrix which above is called a part of a monomial, is called an *m-monomial* for the A matrix used above.

A square matrix is called a *stochastic matrix* of size $n \times n$ for $n \in \mathbb{Z}_+$ if it is a positive matrix and if, of every column, the sum of all column elements equals one. That of every column, the sum of all column elements equals one, is equivalent to the condition that $1_n^T A = 1_n^T$. The time-transition matrix of a finite-state Markov chain is a stochastic matrix.

A square matrix is called a *doubly-stochastic matrix* of size n by n with $n \in \mathbb{Z}_+$ if it is a positive matrix, if, of every column, the sum of all column elments equals one, and if, of every row, the sum of all row elements equals one. The latter conditions are summarized by the conditions that $1_n^T A = 1_n^T$ and $A 1_n = 1_n$. The state transition matrix of a time-reversible Markov process is a doubly stochastic matrix.

A square matrix is called a *circulant positive matrix* of size $n \times n$ for $n \in \mathbb{Z}_+$, in the downward direction, if it is a positive matrix and if it has the pattern described by the next formula,

$$A = \begin{pmatrix} A_1 & A_n & A_{n-1} & \ldots & A_2 \\ A_2 & A_1 & A_n & \ldots & A_3 \\ \vdots & & & \ddots & \vdots \\ A_{n-1} & A_{n-2} & A_{n-3} & \ldots & A_n \\ A_n & A_{n-1} & A_{n-2} & \ldots & A_1 \end{pmatrix} \in \mathbb{R}_+^{n\times n},\ \forall\, i \in \mathbb{Z}_n,\ A_i \in \mathbb{R}_+.$$

Define a *unit-shift (downward-) circulant positive matrix*, or also the *shift matrix* if it is a matrix with one downward shift, specified by the formula,

$$W_n = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 1 \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix} \in \mathbb{R}_+^{n \times n}.$$

Then W_n^2 is a downward-circulant shift matrix by two elements, etc. Consequently, $W_n^{n-1} = I_n$. The shift matrix W_n and its powers W_n^k for all $k \in \mathbb{Z}_+$, are all permutation matrices. Denote the sets of all shift matrices by,

$$\mathbb{R}_{crcl,+}^{n \times n} = \left\{ W_n^i \in \mathbb{R}_+^{n \times n} | \; i \in \mathbb{N} = \{0, 1, \dots, n-1\} \right\}.$$

Define a *circulant doubly-stochastic matrix* as a doubly-stochastic matrix which is also a circulant matrix.

Define the *one matrix* as the square positive matrix of which all elements are equal to one,

$$E_n \in \mathbb{R}_+^{n \times n}, \;\; \forall \, i, j \in \mathbb{Z}_+, \; E_{n,i,j} = 1; \;\; \text{for example, } E_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \in \mathbb{R}_+^{4 \times 4}.$$

Finally the notation of the positive matrices $\mathbb{R}_+^{k \times m}$ of size $k \times m$ for $k, \; m \in \mathbb{Z}_+$ and their subsets are collected,

$\mathbb{R}_+^{k \times m}$	set of positive matrices,
$\mathbb{R}_{perm}^{n \times n}$	set of permutation matrices,
$\mathbb{R}_{+diag}^{n \times n}$	set of diagonal positive matrices with positive diagonal,
$\mathbb{R}_{s+diag}^{n \times n}$	set of diagonal matrices with a strictly-positive diagonal,
$\mathbb{R}_{mon}^{n \times n}$	set of the monomial positive matrices,
$\mathbb{R}_{crcl,+}^{n \times n}$	set of the circulant positive matrices,
$\mathbb{R}_{sh-crcl,+}^{n \times n}$	set of the shift-circulant positive matrices,
$\mathbb{R}_{st}^{n \times m}$	set of stochastic matrices,
$\mathbb{R}_{dst}^{n \times n}$	set of doubly-stochastic matrices,
$\mathbb{R}_{crcl,dst}^{n \times n}$	set of circulant doubly-stochastic matrices,
$\mathbb{R}_{sh-crcl,dst}^{n \times n}$	set of shift-circulant doubly-stochastic matrices.

Example 18.3.2. Consider the following positive matrices.

$$Q_1 = \begin{pmatrix} 0\,1\,0\,0 \\ 1\,0\,0\,0 \\ 0\,0\,0\,1 \\ 0\,0\,1\,0 \end{pmatrix},\ D_1 = \begin{pmatrix} 1\,0\,0\,0 \\ 0\,4\,0\,0 \\ 0\,0\,0\,0 \\ 0\,0\,0\,3 \end{pmatrix},\ D_2 = \begin{pmatrix} 5\,0\,0\,0 \\ 0\,4\,0\,0 \\ 0\,0\,2\,0 \\ 0\,0\,0\,3 \end{pmatrix},$$

$$M_1 = \begin{pmatrix} 0\,9\,0\,0 \\ 8\,0\,0\,0 \\ 0\,0\,0\,7 \\ 0\,0\,6\,0 \end{pmatrix},\ S_1 = \begin{pmatrix} 0.4\ 0.0\ 0.5\ 0.1 \\ 0.3\ 0.3\ 0.0\ 0.3 \\ 0.2\ 0.3\ 0.2\ 0.6 \\ 0.1\ 0.4\ 0.3\ 0.0 \end{pmatrix},\ DS_1 = \begin{pmatrix} 0.4\ 0.1\ 0.2\ 0.3 \\ 0.3\ 0.4\ 0.1\ 0.2 \\ 0.2\ 0.3\ 0.4\ 0.1 \\ 0.1\ 0.2\ 0.3\ 0.4 \end{pmatrix},$$

$$DSC_1 = \begin{pmatrix} 0.4\ 0.1\ 0.2\ 0.3 \\ 0.3\ 0.4\ 0.1\ 0.2 \\ 0.2\ 0.3\ 0.4\ 0.1 \\ 0.1\ 0.2\ 0.3\ 0.4 \end{pmatrix},\ DSC_2 = \begin{pmatrix} 0\,0\,0\,1 \\ 1\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,1\,0 \end{pmatrix}.$$

Then Q_1 is a permutation matrix, D_1 is a diagonal positive matrix, D_2 is a diagonal positive matrix with a strictly-positive diagonal, M_1 is a monomial matrix, S_1 is a stochastic matrix, DS_1 is a doubly-stochastic matrix, DSC_1 is a circulant doubly-stochastic matrix, and DSC_2 is a shift-circulant doubly-stochastic matrix.

It is elementary to prove that any monomial matrix admits two decompositions, each as a product of a permutation matrix and a diagonal matrix with a strictly-positive diagonal. In terms of notation, $M_1 = Q_1 D_1 = D_2 Q_2$, with Q_1, Q_2 permutation matrices and D_1 and D_2 diagonal positive matrices with strictly-positive diagonal.

Proposition 18.3.3. *Elementary relations between the subsets of positive matrices are described below.*

(a)$\{I\} \subset \mathbb{R}^{n\times n}_{sh-circ,dst} \subset \mathbb{R}^{n\times n}_{circ,dst} \subset \mathbb{R}^{n\times n}_{dst}$.
(b)$\{I\} \subset \mathbb{R}^{n\times n}_{sh-circ,dst} \subset \mathbb{R}^{n\times n}_{perm}$.
(c)$\mathbb{R}^{n\times n}_{dst} \subset \mathbb{R}^{n\times n}_{st}$.
(d)$\mathbb{R}^{n\times n}_{s+diag} \subset \mathbb{R}^{n\times n}_{+diag}$.
(e)$\mathbb{R}^{n\times n}_{s+diag} \cup \mathbb{R}^{n\times n}_{perm} \subset \mathbb{R}^{n\times n}_{mon} \subset \mathbb{R}^{n\times n}_{+}$.

Proposition 18.3.4. *A matrix $A \in \mathbb{R}^{n\times n}_{+}$ is a circulant positive matrix if and only if there exists a positive vector $a \in \mathbb{R}^{n}_{+}$ such that $A = \sum_{i=1}^{n} a_i W_n^i$.*

The elementary proof is omitted.

Proposition 18.3.5. *(a)The product of two square stochastic matrices is again a stochastic matrix. Thus $\times : \mathbb{R}^{n\times n}_{st} \times \mathbb{R}^{n\times n}_{st} \to \mathbb{R}^{n\times n}_{st}$.*
(b)The matrix product of two square doubly stochastic matrices is again a doubly stochastic matrix.

Proof. Recall that the positive matrix $A \in \mathbb{R}^{n\times n}_{+}$ is stochastic if $1_n^T A = 1_n^T$; and that it is doubly stochastic if in addition it satisfies $A 1_n = 1_n$. Note that,

$$A,\ B \in \mathbb{R}^{n\times n}_{st} \ \Rightarrow\ 1_n^T (A \times B) = (1_n^T A) \times B = 1_n^T B = 1_n^T;$$
$$A,\ B \in \mathbb{R}^{n\times n}_{dst} \ \Rightarrow\ (A \times B) 1_n = A \times (B 1_n) = A 1_n = 1_n.$$

□

Positive Matrices – Graph Theory

There exists a relation between a positive matrix and a graph. This relation is very valuable because concepts and results of the graph domain can be transported to the matrix domain and conversely.

Definition 18.3.6. A *directed graph* consists of a tuple (V,E) in which $V \subseteq \mathbb{Z}_+$ is a finite set of vertices or nodes and $E \subset V \times V$ a finite set of edges. Then $(i,j) \in E$ denotes that there exists a directed edge from node i to node j.

A *subgraph* of a graph (V,E) is a graph (V_1,E_1) such that $V_1 \subseteq V$ and $E_1 \subseteq E$.

A *path* in a directed graph is a sequences of edges, for example $\{(i_0,i_1),\ (i_1,i_2),\ldots,\ (i_{m-1},i_m)\}$ where for all $k = 1,2,\ldots,m$, $(i_{k-1},i_k) \in E$.

A subgraph (V_1,E_1) of a graph (V,E) is called a *strongly connected component* if for any tuple $(i,j) \in E_1$ there exists a path from node i to node j passing only nodes in the subgraph V_1, *and* a path from node j to node i passing only nodes in the subgraph.

In graph theory a graph (V,E) is called a *reducible graph* if the set of vertices can be *partitioned* into two nonempty subsets, $V_1,\ V_2 \subset V$ such that there may exist edges from V_1 to V_2 but there do not exist edges from V_2 to V_1. A graph is called an *irreducible graph* if it is not reducible.

Example 18.3.7. It is shown how to relate a positive matrix to a graph and then a graph back to a positive matrix.

$$A_1 = \begin{pmatrix} 0 & 6 & 0 & 4 \\ 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \end{pmatrix} \in \mathbb{R}^{4\times 4};\ \ G_1 = (V_1,E_1),\ V_1 = \{1,2,3,4\},$$

$$E_1 = \{(1,2),\ (2,1),\ (2,3),\ (3,4),\ (4,1)\},\ \ (i,j) \in E_1 \text{ if } A_{1,i,j} > 0;$$

$$B_1 = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \in \mathbb{R}^{4\times 4},\ \ B_{1,i,j} = +1 \text{ if } (i,j) \in E_1.$$

Another example follows.

$$A_2 = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix} \in \mathbb{R}^{4\times 4},\ \ G_2 = (V_2,E_2),\ V_1 = \{1,2,3,4\},$$

$$E_1 = \{(1,2),\ (1,3),\ (2,3),\ (2,4),\ (3,1),\ (3,4),\ (4,1),\ (4,2)\}.$$

The concept of a primitive matrix is related to the graph of a matrix. It is of interest for the investigation of the low-integer powers of the associated matrix.

Definition 18.3.8. Call a square positive matrix $A \in \mathbb{R}_+^{n\times n}$ *primitive* if there exists an integer $m \in \mathbb{Z}_+$ such that A^m is a matrix with only strictly positive elements; in terms of notation, $A^m \in \mathbb{R}_{s+}^{n\times n}$.

Define the *index of primitivity* of the matrix $A \in \mathbb{R}_+^{n\times n}$ and denote it by $n_{prim}(A) \in \mathbb{Z}_+$, as the smallest integer $m \in \mathbb{Z}_+$ such that $A^m \in \mathbb{R}_{s+}^{n\times n}$.

Example 18.3.9. Consider a mammillary matrix of size 4 × 4 and its square,

$$A = (1/4) \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 3 & 0 & 0 \\ 1 & 0 & 3 & 0 \\ 1 & 0 & 0 & 3 \end{pmatrix} \in \mathbb{R}_{dst}^{4\times 4}, \; A^2 = (1/16) \begin{pmatrix} 4 & 4 & 4 & 4 \\ 4 & 10 & 1 & 1 \\ 4 & 1 & 10 & 1 \\ 4 & 1 & 1 & 10 \end{pmatrix} \in \mathbb{R}_{s+,dst}^{4\times 4}.$$

Thus the matrix A is primitive irreducible stochastic matrix with $n_{prim}(A) = 2$.

That the matrix A^2 has only strictly positive elements can also be concluded from the fact that from any of the four states of the associated finite system, one can reach by the graph associated with the A matrix any other state in at most two transitions.

Example 18.3.10. Consider the matrix, Consider the shift irreducible stochastic matrix and several of its powers,

$$A = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \in \mathbb{R}_{st}^{3times3}, \; A^2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \; A^3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = I_3, \; A^4 = A, \text{etc.}$$

It is then clear that powers of the stochastic matrix A are periodic. Note that, for all $k \in \mathbb{Z}_+$ of the form $k = m \times 3 + k_1$ where $m \in \mathbb{Z}_+$ and $1 \leq k_1 \leq 2 < 3$, $A^k = A^{k_1} \times A^{3m} = A^{k_1} \times (A^3)^m = A^{k_1}$. Thus there does not exist an integer $k \in \mathbb{Z}_+$ such that A^k is a stochastic matrix with strictly positive elements. Hence the shift irreducible stochastic matrix A is periodic but not primitive.

18.4 Geometry and Cones

A positive matrix is geometrically described by a polyhedral cone. The geometric viewpoint is useful for the theory to be developed.

Definition 18.4.1. A *ray* is a half line $Y \subseteq \mathbb{R}_+^n$ if there exists a vector $x \in Y$, $x \neq 0$, such that for all $c \in \mathbb{R}_+$ one has $c \times x \in Y$ and if $y \in Y$ then there exists a $c \in \mathbb{R}_+$ such that $y = c \times x$. Below $c \times x$ is denoted by $c\,x$.

A *cone* is a nonempty subset $C \subseteq \mathbb{R}_+^n$ satisfying moreover that $C \neq \{0\}$, such that (1) if $x \in C$ and $c \in \mathbb{R}_+$ then $c\,x \in \mathbb{R}_+^n$; and (2) if $x,\; y \in \mathbb{R}_+^n$ then $x + y \in \mathbb{R}_+^n$. By definition, the *zero vector* of $\mathbb{R}_+^n$ is always an element of a cone due to $c\,x \in Y$ for $c = 0$. The zero vector is called the *apex* of the cone.

Define on the set of cones an order relation by the inclusion relation. Thus if $C_1,\, C_2 \subseteq \mathbb{R}_+^n$ are cones than define the order relation on this set of cones by $C_2 \subseteq C_1$ if the inclusion relation holds set wise.

Call the set $C_2 \subseteq \mathbb{R}^n$ a *subcone* of the cone $C_1 \subseteq \mathbb{R}^n$, if C_2 is a cone and if the order relation $C_2 \subseteq C_1$ holds.

A cone $C \subseteq \mathbb{R}_+^n$ is called a *polyhedral cone* if (1) it is a cone and (2) it is a polyhedral set, see Def. 17.6.5.

A cone can be represented in an *implicit representation* or in an *explicit representation*. The *implicit representation* of the polyhedral cone $C \subseteq \mathbb{R}_+^n$ is defined as,

$$\begin{aligned}
&\exists\, m \in \mathbb{Z}_+,\ \exists\, v_1, \ldots, v_m \in C, \text{ such that,} \\
C &= \{x \in \mathbb{R}_+^n \,|\, \exists\, a \in \mathbb{R}_+^m \text{ such that } x = Va\}, \\
V &= \begin{pmatrix} v_1 \ v_2 \ \ldots \ v_m \end{pmatrix} \in \mathbb{R}_+^{n \times m},\ a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} \in \mathbb{R}_+^m.
\end{aligned}$$

The representation is called *implicit* because each element x of the cone is represented as the solution of a linear equation $x = Va$.

Denote in this case the cone by $C = \text{cone}(V)$ or, in stead of V, using the finite set V_f of vectors generating the cone, $C = \text{cone}(V_f)$. One says that the cone $C = \text{cone}(V)$ is *generated* by the columns of the matrix V or by the corresponding set of vectors of V_f. Equivalently, if there exists an integer $m \in \mathbb{Z}_+$ and a finite set of vectors $V_f = \{v_1,\ v_2,\ \ldots,\ v_m \in \mathbb{R}_+^n\}$ such that $C = \text{cone}(V_f)$. Then, for any $x \in C$, there exist positive real numbers $a_i \in \mathbb{R}_+$ for $i = 1, \ldots, m$ such that $x = \sum_{i=1}^m\ a_i\ v_i$. Note that one can have either $m \leq n$ or $m > n$. For example, in $\mathbb{R}_+^3$ with $m = 4$ such a cone has four extremal rays in the space $\mathbb{R}^3$.

Define the *explicit representation* of the polyhedral cone C by the formulas,

$$\begin{aligned}
C &= \cup_{i=1}^k\ \{x \in \mathbb{R}_+^n \,|\, h_i^T x \leq c_i\} = \{x \in \mathbb{R}^n \,|\, Hx \leq c\}, \\
H &= \begin{pmatrix} h_1^T \\ h_2^T \\ \vdots \\ h_m^T \end{pmatrix},\ \ c = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{pmatrix}.
\end{aligned}$$

This formula follows from the fact that the cone is polyhedral hence the intersection of a finite number of half spaces with each half space described by $\{x \in \mathbb{R}_+^n \,|\, h_i^T x \leq c_i\}$.

A *boundary ray* of a cone is a ray that lies on the boundary of the cone. A ray lies on the boundary of a cone if for any $\varepsilon \in (0,1) \subseteq \mathbb{R}_+$ sufficiently small and for every x on the ray, the ball

$$B(x,\varepsilon) = \{y \in \mathbb{R}_+^n \,|\, \|x - y\| < \varepsilon\},$$

includes an element outside the cone.

A ray is called an *extremal boundary ray* of a cone C if it cannot be written as a strict convex combination of two boundary rays. Thus, x is an extremal boundary ray and if $x = cy_1 + (1-c)y_2$ with $y_1,\ y_2$ boundary rays of the cone and $c \in (0,1) \subseteq \mathbb{R}_+$, implies that $y_1 = y_2$. In Example 18.4.2 each of the columns of the matrix B is an extremal boundary ray.

Conversely, any polyhedral cone can be related to a positive matrix.

Consider a polyhedral cone C and denote its extremal vectors by $\{y_1,\, y_2,\, \ldots,\, y_m \in \mathbb{R}_+^n\}$. Define then the matrix $A \in \mathbb{R}_+^{n\times m}$ such that the columns of A equal the extremal rays of C. Then A is a positive matrix and $C = \text{cone}(A)$.

A cone is called a *nonpolyhedral cone* if it is not a polyhedral cone. Thus there do not exist an integer $m \in \mathbb{Z}_+$ and a matrix $A \in \mathbb{R}_+^{n\times m}$ such that $C = \text{cone}(A)$. An example of a nonpolyhedral cone is the well know ice cream cone, [3, Ex. 1.2.2]. See Example 18.4.3.

Example 18.4.2. An example of a polyhedral cone is provided by the formulas,

$$C = \{x \in \mathbb{R}_+^4 |\ \exists\, a \in \mathbb{R}_+^4,\ x = Ba\}, \quad B = \begin{pmatrix} 1\,0\,0\,1 \\ 1\,1\,0\,0 \\ 0\,1\,1\,0 \\ 0\,0\,1\,1 \end{pmatrix}.$$

That the spanning vectors of cone(B) are extremal rays of the corresponding cone is directly obvious from the elements of the columns of the matrix B.

Example 18.4.3. The reader has to recall in his/her mind the image of an ice cream cone which, before the consumption of the ice, was filled with coups of ice cream. The object below is a mathematical representation of this image though it stretches to infinity, would contain too much ice cream, hence is not quite realistic.

The *ice cream cone* in the positive real numbers of size $3 \in \mathbb{Z}_+$. The cone is generated by defining rays from the origin to all the points on a circle in the probability simplex.

Recall the probability simplex in $\mathbb{R}_+^3$ defined as,

$$S^3 = \{x \in \mathbb{R}_+^3 |\ 1_3^T x = 1\}.$$

Choose the center of the probability simplex as the element $c = \begin{pmatrix} 1/3\ 1/3\ 1/3 \end{pmatrix}^T$ and define the circle with center c and radius $r = 0.4$ in the probability simplex,

$$B(c, 0.4) = \{x \in S^3 |\ \|x - c\|_2 = 0.4\}.$$

Construct all rays which start in the origin 0 and pass through one of the points of the circle $B(c, 0.4)$. The set containing all those rays is called the *ice cream cone*. Note that any ray on the boundary of that cone is an extremal ray because it cannot be written as a convex combination of two other boundary rays.

Definition 18.4.4. A finite set of vectors of $\mathbb{R}_+^n$ for $n \in \mathbb{Z}_+$, say $\{v_1, \ldots, v_m \in \mathbb{R}^n\}$ for $m \in \mathbb{Z}_+$, is said to be *positively dependent* if,

$$\exists\, i \in \mathbb{Z}_m,\ \exists\, a_1, \ldots, a_{i-1},\ a_{i+1}, \ldots, a_m \in \mathbb{R}_+, \text{ not all zero, such that,}$$
$$v_i = \sum_{j \in \mathbb{Z}_m \setminus \{i\}} a_j v_j.$$

The equivalent geometric definition is that,

$$v_i \in \text{cone}(V_{\mathbb{Z}_m \setminus \{i\}})$$
$$V_{\mathbb{Z}_M \setminus \{i\}} = \begin{pmatrix} v_1\ v_2\ \ldots\ v_{i-1}\ v_{i+1}\ \ldots\ v_m \end{pmatrix},\ \forall\, i \in \mathbb{Z}_m.$$

It is said to be *positively independent* if it is not positively dependent and, equivalently geometrically, no v_i is a member of the cone spanned by the other vectors.

Example 18.4.5. The columns of the matrix B of Example 18.4.2 are positively independent. For example, if one wants to compute whether the first column is positively dependent on the set consisting of the second, third, and fourth column then one computes,

$$\begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0\;0\;1 \\ 1\;0\;0 \\ 1\;1\;0 \\ 0\;1\;1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix},$$

$$\Rightarrow 0 = a_1 + a_2,\ 0 = a_2 + a_3,\ \Rightarrow\ a_1 = a_2 = a_3 = 0,$$

hence no solution exists. The other combinations of columns is similar.

Example 18.4.6. The following calculation shows that the vector x is positively dependent on the indicated three vectors,

$$x = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix} = \begin{pmatrix} 1\;0\;1 \\ 1\;1\;0 \\ 0\;1\;1 \end{pmatrix} \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix} \frac{1}{2}.$$

Definition 18.4.7. *Frame of a polyhedral cone*. A finite set of vectors $V_{fr} = \{v_1, v_2, \ldots, v_m \in \mathbb{R}_+^n\}$ for $n,\ m \in \mathbb{Z}_+$ and with $|V_{fr}| = m$, is said to be a *frame* of the polyhedral cone $C \subseteq \mathbb{R}_+^n$ if (1) $C = \text{cone}(V_{fr})$; and (2) V_{fr} is a positively independent set of vectors. Call the smallest integer $m \in \mathbb{Z}_+$ such that $C = \text{cone}(V_{fr})$ the *frame size* and denote it by $\text{Framesz}(V_{fr}) = m$.

For integers $n,\ m \in \mathbb{Z}_+$ denote the set of all polyhedral cones in $\mathbb{R}_+^n$ with a frame of size m by,

$$C_{n,m} = \{C \subseteq \mathbb{R}_+^n |\ \exists\ V_{fr}\ \subseteq \mathbb{R}_+^n \text{ such that } \text{Framesz}(V_{fr}) = m,\ C = \text{cone}(V_{fr})\}.$$

Example 18.4.8. Consider Example 18.4.2 where the cone is spanned by the column vectors of the B matrix. It follows from Example 18.4.5 that the columns of B are positively independent. The cone cone(B) generated by the B matrix has thus a frame consisting of the columns of the B matrix. The positive independence of the columns the B matrix is thus necessary for the vectors to be a frame. The framesize is $\text{Framesz}(\text{cone}(V(B))) = 4$.

A polyhedral cone has faces and facets defined below. The terminology is not always consistent in the literature.

Definition 18.4.9. *Faces of a polyhedral cone*. Consider a polyhedral cone $C \in C_{n,m}$ for integers $n,\ m \in \mathbb{Z}_+$ and with a frame $V_{fr} \subseteq \mathbb{R}_+^n$ of frame size $\text{Framesz}(V_{fr}) = m$, hence $C = \text{cone}(V_{fr})$. Denote the rank of the matrix of which the columns are the elements of V_{fr} by $n_d = \text{rank}(V_{fr})$.

Define the n_d-face of the cone C as the cone itself and $\mathbb{F}_{n_d}(C)$ as the set containing all n_d-faces which thus contains only C. There do not exist r-faces of C for $r \in \mathbb{Z}_+$ with $r > n_d$.

For $r = n_d - 1,\ n_d - 2,\ \ldots,\ 1,\ 0 \in \mathbb{N}$ define an *r-face* $F_r \subseteq \mathbb{R}_+^n$ with $\dim(F_r) = r$ by the conditions: (1) there exists a $r+1$-face of the set $F_{r+1} \in \mathbb{F}_{r+1}(C)$ such that

F_r is a subcone of F_{r+1}; (2) $F_r \subseteq \partial F_{r+1}$; (3) no subcone of F_{r+1} contained in ∂F_{r+1} strictly contains F_r; and (4) $F_r \neq \emptyset$.

Denote by $\mathbb{F}_r(C)$ the set of r-dimensional faces of the cone C. Call the faces of the set $\mathbb{F}_{n_d-1}$, the *facets* of the cone(C). Only faces of dimension equal to the dimension of the cone minus one, are called facets.

Possibly the definition has to be revised later by replacing the rank of a face by the positive rank of that face.

Example 18.4.10. Consider a cone in the plane,

$$V_{fr} = \left\{ v_1 = \begin{pmatrix} 2/3 \\ 1/3 \end{pmatrix},\ v_2 = \begin{pmatrix} 1/3 \\ 2/3 \end{pmatrix} \right\} \in \mathbb{R}^2_+,\ C = \text{cone}(V_{fr}) \subseteq \mathbb{R}^2_+.$$

Then $\mathbb{F}_2 = \{C\}$ is the set of 2-faces which contains only the cone C, cone$(\{v_1\})$ and cone$(\{v_2\}) \in \mathbb{R}^1_+$ are the only facets of $\mathbb{F}_1(C)$ of dimension one, and $0 \in \mathbb{R}^2_+$ is the only face of the cone of dimension zero.

Example 18.4.11. *Faces of a unit cube*. The reader is expected to know that a cube in $\mathbb{R}^3_+$ can be presented as the body generated by positive combinations of the eight vectors, represented as the columns of a matrix. Note that a cube is not a cone. However, the concepts of faces and facets can be well explained for a cube.

$$\begin{aligned} V &= \begin{pmatrix} 0\,1\,0\,1\,0\,1\,0\,1 \\ 0\,0\,1\,1\,0\,0\,1\,1 \\ 0\,0\,0\,0\,1\,1\,1\,1 \end{pmatrix} \\ &= \begin{pmatrix} 0 & e_1 & e_2 & e_1+e_2 & e_3 & e_1+e_3 & e_2+e_3 & e_1+e_2+e_3 \end{pmatrix} \in \mathbb{R}^{3\times 8}, \\ C_3 &= \{x \in \mathbb{R}^3_+ \,|\, \exists\, a \in \mathbb{R}^8_{st},\ x = Va\}. \end{aligned}$$

The set of three-dimensional faces $\mathbb{F}_3(C_3)$ contains only the cube C_3

The set of two-dimensional faces, called the *set of facets*, contains the boundary planes of the cube each of which is generated by a quadruple of vectors,

$$\begin{aligned} \mathbb{F}_2(C_3) &= \{F_{2,1},\ \ldots,\ F_{2,6}\}, \\ F_{2,1} &= \{x \in \mathbb{R}^2_+ \,|\, \exists\, a \in \mathbb{R}^4_{st}\ x = V_{2,1}a\},\ V_{2,1} = \{0,\ e_1,\ e_2,\ e_1+e_2\}, \\ F_{2,2} &= \{x \in \mathbb{R}^2_+ \,|\, \exists\, a \in \mathbb{R}^4_{st}\ x = V_{2,2}a\},\ V_{2,2} = \{0,\ e_1,\ e_3,\ e_1+e_3\}, \text{etc.} \end{aligned}$$

The set of one-dimensional faces contains the line pieces on the boundary of the cube.

$$\begin{aligned} \mathbb{F}_1(C_3) &= \{F_{1,1},\ \ldots,\ F_{1,12}\}, \\ F_{1,1} &= \{x \in \mathbb{R}^2_+ \,|\, \exists\, a \in \mathbb{R}^2_{st}\ x = V_{1,1}a\},\ V_{1,1} = \{0,\ e_1\}, \\ F_{1,2} &= \{x \in \mathbb{R}^2_+ \,|\, \exists\, a \in \mathbb{R}^2_{st}\ x = V_{1,2}a\},\ V_{1,2} = \{e_1,\ e_1+e_2\},\ \text{etc.} \end{aligned}$$

The one-dimensional faces contains the extremal points of the cube,

$$\begin{aligned} \mathbb{F}_0(C_3) &= \{F_{0,1}, F_{0,2},\ \ldots,\ F_{0,8}\}, \\ &\quad F_{0,1} = \{0\},\ F_{0,2} = \{e_1\},\ F_{0,3} = \{e_2\},\ F_{0,4} = \{e_1+e_2\}, \text{etc.} \end{aligned}$$

18.5 Units

Positive matrices can be multiplied and, possibly, be inverted.

Definition 18.5.1. Consider a monoid $(X, \times, 1)$. Call an element $x \in X$ *invertible* or a *unit* if there exists an element $y \in X$ such that $x \times y = 1 = y \times x$. If there exists, for any such $x \in X$ a unique element $y \in X$ such that $x \times y = 1$, then denote $y = x^{-1}$. Denote the subset of X of elements for which an inverse exists the *group of units* and denote it by $X_U \subseteq X$. That X_U is a group, see Def. 17.2.2, is easily proven.

Example 18.5.2. The group of units of the positive real numbers $\mathbb{R}_+$ is the set of strictly positive real numbers $\mathbb{R}_{s+} = (0, \infty)$.

It is of interest to determine for each subset of the positive matrices, what is the associated group of units.

Theorem 18.5.3. *(a)For any permutation matrix $Q \in \mathbb{R}^{n \times n}_{perm}$ there exists an inverse which is also a permutation matrix, $Q^{-1} \in \mathbb{R}^{n \times n}_{perm}$. In fact, the inverse of a permutation matrix is its transposed matrix, $Q^{-1} = Q^T$. Thus the set of permutation matrices is a group of units.*

(b)For any diagonal positive matrix with strictly-positive diagonal $D \in \mathbb{R}^{n \times n}_{s+diag}$ there exists an inverse matrix which is also a diagonal positive matrix with strictly-positive diagonal. In fact, D^{-1} satisfies that for all $i \in \mathbb{Z}_n$, $(D^{-1})_{ii} = (D_{ii})^{-1}$. The group of units in the diagonal positive matrices, $\mathbb{R}^{n \times n}_{diag}$, is the set of diagonal positive matrices with strictly-positive diagonal, $\mathbb{R}^{n \times n}_{s+diag}$.

(c)For any monomial positive matrix $M \in \mathbb{R}^{n \times n}_{mon}$ there exists an inverse matrix which is also a monomial matrix, $M^{-1} \in \mathbb{R}^{n \times n}_{mon}$. If $M = DQ$ with $Q \in \mathbb{R}^{n \times n}_{perm}$ and $D \in \mathbb{R}^{n \times n}_{s+}$ then $M^{-1} = Q^T D^{-1}$. The group of units of the monomial matrices is thus the set of monomial matrices.

(d)For a positive matrix $A \in \mathbb{R}^{n \times n}_{+}$ there exists a multiplicative inverse matrix which is also a positive matrix, $A^{-1} \in \mathbb{R}^{n \times n}_{+}$ if and only if A is a monomial matrix. The group of units of the positive matrices is therefore the set of monomial matrices.

(e)For a doubly stochastic matrix $A \in \mathbb{R}^{n \times n}_{dst}$ there exists a multiplicative inverse matrix which is also doubly stochastic if and only if A is a permutation matrix. Thus the group of units of the doubly stochastic matrices is the set of permutation matrices.

(f) For a circulant doubly stochastic matrix $A \in \mathbb{R}^{n \times n}_{circ,dsc}$ there exists an inverse matrix which is also a circulant doubly stochastic if and only if A is a unit-shift-circulant doubly stochastic matrix. Because a unit-shift-circulant doubly stochastic matrix is a permutation matrix, $A^{-1} = A^T$. Thus the group of units of the doubly stochastic circulant matrices is the set of shift-circulant doubly stochastic matrices.

(g)For a shift-circulant doubly stochastic matrix $A \in \mathbb{R}^{n \times n}_{circ,dst}$ there always exists an inverse matrix which is also a shift-circulant doubly stochastic matrix. If $A = W_n^i$ then $A^{-1} = W_n^j = W_n^{n-i| \mod n} = A^T$ where $i,\ j \in \mathbb{Z}_n$ are such that $i + j = n$.

Proof. (a) It is to be shown that $QQ^T = I$. Note that for all $i,\ j \in \mathbb{Z}_n$,

$$(QQ^T)_{i,j} = \sum_{k=1}^{n} Q_{i,k}Q^T_{k,j} = \sum_{k=1}^{n} Q_{i,k}Q_{j,k} = \left\{ \begin{array}{l} 1,\ \text{if } i = j, \\ 0,\ \text{if } i \neq j, \end{array} \right\} = I_{i,j}.$$

If there exists an inverse $X \in \mathbb{R}^{n \times n}$ such that $QX = I$ then, $\forall\ i,\ j \in \mathbb{Z}_n$,

$$\begin{aligned} I_{i,j} &= (QX)_{i,j} = \sum_{k=1}^{n} Q_{i,k}X_{k,j} = \left\{ \begin{array}{l} 1,\ \text{if } i = j, \\ 0,\ \text{if } i \neq j, \end{array} \right. \Rightarrow \\ &\forall\ i,\ j \in \mathbb{Z}_n,\ i = j,\ \exists\ k \in \mathbb{Z}_N,\ Q_{i,k}X_{k,i} = 1 \ \Rightarrow\ X_{k,i} = Q_{i,k}, \\ &\forall\ i,\ j \in \mathbb{Z}_n,\ i \neq j,\ \forall\ k \in \mathbb{Z}_n,\ Q_{i,k}X_{k,j} = 0 \ \Rightarrow\ X_{k,j} = 0, \\ &\Rightarrow X = Q^T. \end{aligned}$$

(b) The elementary proof is omitted.

(c) Any monomial matrix admits a decomposition as $A = DQ$ according to the paragraph above Proposition 18.3.4 with Q and D as specified in the (c) statement. Then $Q^{-1} = Q^T$ by (a) and D^{-1} is well defined by (b) and $D^{-1} \in \mathbb{R}^{n \times n}_{s+diag}$. Then $X = Q^T D^{-1} \in \mathbb{R}^{n \times n}_{mon}$. It is to be shown that $AX = I$. Note that $AX = DQQ^T D^{-1} = DD^{-1} = I$ by (a) and by (b), and $XA = Q^T D^{-1} DQ = Q^T Q = I$. Thus $A^{-1} = X = Q^T D^{-1}$.

(d) This is proven in [3, Lemma 3.4.3] similar to (a) and (c).

(e) ($\Rightarrow$) Note that for a doubly stochastic matrix, for all $i,\ j \in \mathbb{Z}_n$, $A_{i,j} \in [0,1]$. For all $i \in \mathbb{Z}_n$ and because the column sums of A equal one, there exists a $j \in \mathbb{Z}_n$ such that $A_{i,j} > 0$. Let $j(i) \in \mathbb{Z}_N$ be the smallest integer such that $A_{i,j(i)} > 0$. Then for $k \in \mathbb{Z}_n \backslash \{j\}$, $0 = I_{k,j(i)} = \sum_{m=1}^n A^{-1}_{k,m}A_{m,j} \geq A^{-1}_{k,i}A_{i,j(i)}$, $A_{i,j(i)} > 0$, and $A^{-1}_{i,j(i)} \in \mathbb{R}_+$ imply that $A^{-1}_{i,k} = 0$; for $k = j(i)$, $1 = I_{j(i),j(i)} = \sum_{m=1}^n A^{-1}_{j(i),m}A_{m,j(i)} = A^{-1}_{j(i),i}A_{i,j(i)}$. This implies that $A^{-1}_{j(i),i} = 1/A_{i,j(i)}$ and, because $A_{i,j(i)} \in [0,1]$ and $A^{-1}_{j(i),i} \in [0,1]$, it follows that $A^{-1}_{j(i),i} = 1$ and $A_{i,j(i)} = 1$. Because the sum of row i of a doubly stochastic matrix equals 1, it follows that $A_{i,k} = 0$ for all $k \neq j(i)$. This then holds for all rows. Because the matrix A is doubly stochastic, both A and A^{-1} are permutation matrices, in fact $A^{-1} = A^T$.

($\Leftarrow$) If A is a permutation matrix then $A^{-1} = A^T$ is also a permutation matrix and hence doubly stochastic.

(f) The proof is similar to that of (e).

(g) This is directly obvious from the definition of the unit-shift circulant matrices.

□

Example 18.5.4. Consider the following monomial matrix and its inverse,

$$\begin{aligned} A_1 &= \begin{pmatrix} 0 & 0 & 1/4 \\ 1/2 & 0 & 0 \\ 0 & 1/3 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1/2 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/4 \end{pmatrix}, \\ A_1^{-1} &= \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 3 \\ 4 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}. \end{aligned}$$

The following doubly stochastic matrix is invertible because it is a permutation matrix,

$$A_2 = \begin{pmatrix} 0\,1\,0\,0 \\ 1\,0\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,1 \end{pmatrix}, \; A_2^{-1} = A_2^T = \begin{pmatrix} 0\,1\,0\,0 \\ 1\,0\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,1 \end{pmatrix} = A_2.$$

The next doubly stochastic matrix is singular and the second doubly stochastic matrix is nonsingular but its inverse is neither a stochastic matrix nor a positive matrix.

$$A_3 = \begin{pmatrix} 1/2 \; 1/2 \\ 1/2 \; 1/2 \end{pmatrix} \in \mathbb{R}_{dst}^{2\times 2},$$

$$A_4 = \begin{pmatrix} 0.6 \; 0.4 \\ 0.4 \; 0.6 \end{pmatrix} \in \mathbb{R}_{dst}^{2\times 2}, \; A_4^{-1} = \begin{pmatrix} 0.6 & -0.4 \\ -0.4 & 0.6 \end{pmatrix} \in \mathbb{R}^{2\times 2}.$$

The following circulant doubly stochastic matrix is invertible and its inverse is displayed.

$$A_5 = \begin{pmatrix} 0\,0\,0\,1 \\ 1\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,1\,0 \end{pmatrix} \in \mathbb{R}_{dst}^{4\times 4}, \; A_5^{-1} = \begin{pmatrix} 0\,1\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,1 \\ 1\,0\,0\,0 \end{pmatrix} \in \mathbb{R}_{dst}^{4\times 4}.$$

18.6 Similarity

In the body of the book representations are used of finite stochastic system. The state process of a time-invariant finite system is a finite-state Markov process with as typical state space the finite set $\mathbb{Z}_n = \{1, 2, \ldots, n\}$ for an integer $n \in \mathbb{Z}_+$. In the indicator representation the state is then represented by a vector $x(t) \in X_{uv} = \{e_i \in \mathbb{R}_+^n, \; \forall \; i \in \mathbb{Z}_n\}$. The indicator representation of the dynamics of this system is then provided by the formula,

$$E[x(t+1)|F_t^x] = Ax(t), \; x(0) = x_0.$$

The state set $\mathbb{Z}_n$ can be changed only by a permutation matrix and the state set X_{uv} similarly can be changed only by a permutation matrix. Thus if $Q \in \mathbb{R}_{perm}^{n\times n}$ is a permutation matrix then the formula $\bar{x}(t) = Qx(t)$ determines another stochastic system with the state set $\mathbb{Z}_n$ in which the states are permuted with respect to the old state set. For the indicator representation a corresponding transformation takes place.

If there is this choice of the ordering of the states of the state set, what is then the best order to select? This motivates the concept of similarity by permutations as defined below. The concept was formulated by G. Frobenius including the decompositions defined below.

Definition 18.6.1. Consider two square positive matrices $A, \; B \in \mathbb{R}_+^{n\times n}$. Call these matrices:

- *permutation similar* if there exists a permutation matrix $Q \in \mathbb{R}^{n \times n}_{perm}$ such that $A = QBQ^{-1} = QBQ^T$;
- *strictly-positive diagonal similar* if there exists a diagonal positive matrix with a strictly-positive diagonal, $D \in \mathbb{R}^{n \times n}_{s+diag}$, such that $A = DBD^{-1}$;
- *monomially similar* if there exists a monomial matrix $M \in \mathbb{R}^{n \times n}_{mon}$ such that $A = MBM^{-1}$;

That similarity is defined for permutation matrices, diagonal matrices with strictly positive diagonal, and for monomial matrices, is that each of these is a groups of units for a particular subset of positive matrices, see Theorem 18.5.3

Characterizations of positive matrices which are similar by permutation matrices, by diagonal positive matrices with strictly-positive diagonal, or by monomial matrices, follow in this subsection.

The relation of similarity of positive matrices defined above, is an equivalence relation as defined in Def. 17.1.2 which is easily proven using the properties of the corresponding group of units.

Permutation similarity is investigated first.

Definition 18.6.2. Consider a positive matrix $A \in \mathbb{R}^{n \times n}_+$ for $n \in \mathbb{Z}_+$. Call this matrix *reducible* if,

$$\exists\, Q \in \mathbb{R}^{n \times n}_{perm},\ \exists\, n_1, n_2 \in \mathbb{Z}_+,\ \exists\, A_{11} \in \mathbb{R}^{n_1 \times n_1},\ A_{12} \in \mathbb{R}^{n_1 \times n_2},\ A_{22} \in \mathbb{R}^{n_2 \times n_2},$$

$$\text{such that } n = n_1 + n_2 \text{ and } QAQ^T = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}; \tag{18.1}$$

$$\text{hence } A = Q^T \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} Q. \tag{18.2}$$

Call the matrix A *irreducible* if (1) $A \neq 0$ and (2) A is not reducible.

Call the matrix A *fully reduced* if either $n = 1$ or there exists a transformation by a permutation matrix Q such that QAQ^T has a decomposition in upper-block diagonal form with on the diagonal only blocks of irreducible submatrices. Thus the lower-block diagonal matrices are all zero. The notation for a fully reduced matrix follows,

$$\exists\, m \in \mathbb{Z}_+,\ \exists\, n_1, \ldots, n_m \in \mathbb{Z}_+ \text{ such that } n_1 + \ldots + n_m = n;$$
$$\forall\, i,\ j \in \mathbb{Z}_n \text{ with } j \geq i,\ \exists\, A_{i,j} \in \mathbb{R}^{n_i \times n_j}_+;$$
$$\forall\, i \in \mathbb{Z}_n\ \exists\, A_{i,i} \in \mathbb{R}^{n_i \times n_i}_+, \text{ irreducible, such that,}$$

$$A = \begin{pmatrix} A_{1,1} & A_{1,2} & A_{1,3} & \ldots & A_{1,m-1} & A_{1,m} \\ 0 & A_{2,2} & A_{2,3} & \ldots & A_{2,m-1} & A_{2,m} \\ 0 & 0 & A_{2,3} & \ldots & A_{2,m-1} & A_{2,m} \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \ldots & A_{m-1,m-1} & A_{m-1,m} \\ 0 & 0 & 0 & \ldots & 0 & A_{m,m} \end{pmatrix} \in \mathbb{R}^{n \times n}_+.$$

A reducible matrix corresponds to a reducible graph as defined above. An irreducible positive matrix corresponds to a strongly connected component of a graph.

A fully-reduced positive matrix corresponds to graph with $m-1$ subgraphs in which every such subgraph there may exist a transition to another subgraph with a lower index number.

Example 18.6.3. The following two matrices are irreducible.

$$A = \begin{pmatrix} 1\,0\,0\,1 \\ 1\,1\,0\,0 \\ 0\,1\,1\,0 \\ 0\,0\,1\,1 \end{pmatrix} \in \mathbb{R}_+^{4\times 4}, \;\; B = \begin{pmatrix} 2\,1\,1\,1 \\ 1\,2\,0\,0 \\ 1\,0\,2\,0 \\ 1\,0\,0\,2 \end{pmatrix} \in \mathbb{R}_+^{4\times 4}.$$

Of interest for classification of finite-state stochastic systems is a classification of irreducible positive matrices. This is partly developed in Section 18.8. There follow two subsets of irreducible positive matrices which arise in applications.

Definition 18.6.4. Consider a positive matrix $A \in \mathbb{R}_+^{n\times n}$ for an integer $n \in \mathbb{Z}_+$.

An *order-two circulant matrix* of an order-two positive vector as defined below, is an irreducible matrix of the form,

$$A = \text{crcl}(a) = \begin{pmatrix} a_1 & 0 & 0 & \ldots & 0 & a_2 \\ a_2 & a_1 & 0 & \ldots & 0 & 0 \\ 0 & a_2 & a_1 & \ldots & 0 & 0 \\ 0 & 0 & a_2 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & a_1 & 0 \\ 0 & 0 & 0 & \ldots & a_2 & a_1 \end{pmatrix} = a_1 I_n + a_2 W_n, \;\; a = \begin{pmatrix} a_1 \\ a_2 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \in \mathbb{R}_+^n, \tag{18.3}$$

$$a_1,\; a_2 \in \mathbb{R}_{s+},\; n_{s+}(a) = 2. \tag{18.4}$$

Another irreducible matrix is the *mammillary matrix*, or an *arrow matrix*, or a *star matrix*, defined to have the form,

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} & A_{14} & \ldots & A_{1,n-1} & A_{1,n} \\ A_{21} & A_{22} & 0 & 0 & \ldots & 0 & 0 \\ A_{31} & 0 & A_{33} & 0 & \ldots & 0 & 0 \\ A_{41} & 0 & 0 & A_{44} & \ldots & 0 & 0 \\ \vdots & \vdots & \ldots & \ldots & \ddots & \ddots & \vdots \\ A_{n-1,1} & 0 & 0 & \ldots & \ldots & A_{n-1,n-1} & 0 \\ A_{n1} & 0 & 0 & \ldots & \ldots & 0 & A_{n,n-1} \end{pmatrix} \in \mathbb{R}_+^{n\times n}, \tag{18.5}$$

$$(\forall\, i \in \mathbb{Z}_n,\;\; A_{ii} \in \mathbb{R}_{s+}),\;\; (\forall\, i \in \mathbb{Z}_n,\; i \neq 1,\;\; A_{1,i},\, A_{i,1} \in \mathbb{R}_{s+}).$$

The expression mammillary matrix is derived from the life-science expression *mammals* which are living creatures who directly after birth are fed by sucking milk from their mother. A mammilary matrix is named after the concept of a mammilary system, see [38]. A mammillary matrix is for example the state transition matrix of a linear mammillary system where the first state-component represents the concentration of a substance in the mother compartment, often the blood, and the second and higher state-components represent the concentration of a substance in the child compartments.

The transitions of a system with a mammillary matrix are thus such that from Subsystem 1 one can transit to any other subsystem; and from any substem not equal to Subsystem 1, one can either stay at the same subsystem or go to Subsystem 1 but not to any other subsystem. The form of a mammillary matrix is also called an *arrow matrix* based on the shape of the nonzero elements of the matrix with a NorthWest arrow.

Theorem 18.6.5. *A doubly-stochastic matrix is either irreducible or reducible by a permutation similarity to a block-diagonal matrix of which each diagonal block-matrix is an irreducible doubly-stochastic matrix.*

18.7 Eigenvalues and Eigenvectors of Positive Matrices

There follows the Perron-Frobenius theorem. O. Perron proved it for strictly positive matrices, of which every element is strictly positive, and G. Frobenius defined the concepts of reducible and irreducible positive matrices, and proved the theorem for irreducible positive matrices.

Theorem 18.7.1. Perron - Frobenius theorem. *Let $A \in \mathbb{R}_+^{n\times n}$ be an irreducible positive matrix. Recall that the spectral radius is denoted by $r(A) \in \mathbb{R}_+$, see Def 17.4.16.*

(a)The spectral radius $r(A) \in \mathbb{R}_+$ is an algebraically-simple root of the characteristic equation of A.

(b)Consider the case where there are precisely $k \in \mathbb{Z}_+$ eigenvalues $\lambda \in \text{spec}(A)$ such that $|\lambda| = r(A)$, hence $1 \leq k \leq n$. Each of those eigenvalues is also an algebraically-simple root of the characteristic equation of the matrix A and these eigenvalues are the solution of the equation $\lambda^k = r(A)^k$ in $\lambda \in \mathbb{C}$.

In this case, the subset of the spectrum of the matrix A consisting of those eigenvalues whose modulus equal $r(A)$,

$$\Lambda(r(A)) = \{\lambda \in \text{spec}(A) \subset \mathbb{C} | \; |\lambda| = r(A)\}.$$

is mapped to itself by a rotation around zero with any integer multiple of $2\pi/k$. In this case the matrix A is permutation similar to the following matrix,

$$A = Q \begin{pmatrix} 0 & 0 & 0 \ldots 0 & 0 & A_{1,k} \\ A_{2,1} & 0 & 0 \ldots 0 & 0 & 0 \\ 0 & A_{3,2} & 0 \ldots 0 & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 \ldots A_{k-1,k-2} & 0 & 0 \\ 0 & 0 & 0 \ldots 0 & A_{k,k-1} & 0 \end{pmatrix} Q^T,$$

$$\exists\, n_1, \; \ldots, \; n_k \in \mathbb{Z}_n, \text{ such that } n_1 + n_2 + \ldots + n_k = n,$$
$$Q \in \mathbb{R}_{perm}^{n\times n}, \; \forall \; i \; \in \mathbb{Z}_k, \; A_{i,i-1| \text{ mod } k} \in \mathbb{R}_{s+}^{n_i \times n_{i-1}}.$$

where the dimensions are such that all the blocks on the diagonal of the decomposed matrix are square.

(c) There exists an eigenvector associated with the eigenvalue of the spectral radius, which is a real vector in $\mathbb{R}^n$ (hence not complex valued), of which all components are nonzero and have the same sign; in terms of mathematical notation,

$$\exists\ v \in \mathbb{R}^n \text{ such that } Av = r(A)v,$$
$$\text{either } \{\forall\ i \in \mathbb{Z}_n,\ v_i \in (0,\infty)\},\ \text{or } \{\forall\ j \in \mathbb{Z}_n,\ v_j \in (-\infty,0)\}.$$

The vector v is unique upto a scalar multiple, meaning that if $v \in \mathbb{R}^n$ is a solution and $c \in \mathbb{R}\backslash\{0\}$ then $c\ v$ is also a solution.

The case of a positive matrix which is reducible is not treated in this book. But see Section 18.8 for the case of reducible stochastic matrices.

The reader is reminded of the concept of the spectral index of a matrix, see Def. 17.4.16. The formal definition is repeated below and the notation is changed because of the special form of the spectral index of a stochastic matrix.

Definition 18.7.2. Consider a stochastic matrix $A \in \mathbb{R}_{st}^{n\times n}$. Because by Theorem 18.8.2 the spectral radius of the matrix A equals $r(A) = 1$, the matrix A has no eigenvalues $\lambda \in \text{spec}(A)$ such that $|\lambda| > 1$. Therefore the spectral index of a stochastic matrix simplifies.

Denote the *spectral index of a stochastic matrix* $A \in \mathbb{R}_{st}^{n\times n}$ as a special case of the spectral index of the positive matrix A by the tuple,

$$\begin{aligned} n_{st,si}(A) &= (n_1,n_-) \in \mathbb{N}_n,\ n = n_1 + n_-, \\ n_1 &= \text{number of eigenvalues } \lambda \in \text{spec}(A) \text{ such that } |\lambda| = 1 = r(A), \\ n_- &= \text{number of eigenvalues } \lambda \in \text{spec}(A) \text{ such that } |\lambda| < 1 = r(A). \end{aligned}$$

Note that each of the eigenvalues λ such that $|\lambda| = 1$, is of a single algebraic complexity. The index of the eigenvalues $\lambda \in \text{spec}(A)$ such that $|\lambda| = 1 = r(A)$ is also called the *index of cyclicity* of the matrix A and denoted by $n_{cyc}(A) = n_1 \in \mathbb{Z}_n$.

Because $r(A) = 1$ one concludes that $n_1 \geq 1$. Consequently, $0 \leq n_- \leq n-1$.

Call a square stochastic matrix $A \in \mathbb{R}_+^{n\times n}$ *substochastic* if there exists a $j \in \mathbb{Z}_n$ such that the sum of the elements of the j-th column satisfies $\sum_{i=1}^n A_{i,j} < 1$. Call it *column-wise substochastic* if, for all $j \in \mathbb{Z}_n$, the column sum of the j-th column satisfies $\sum_{i=1}^n A_{i,j} < 1$.

Define the *spectral gap* $r_{gap}(A)$ of the stochastic matrix A by the expression,

$$\begin{aligned} \text{spec}_{subr}(A) &= \{\lambda \in \text{spec}(A) \subset \mathbb{C}|\ |\lambda| < r(A) = 1\}, \\ & n_- > 0\ \Rightarrow\ \text{spec}_{subr}(A) \neq \emptyset, \\ r_{gap}(A) &= \begin{cases} r(A) - \max_{\lambda\in\text{spec}_{subr}(A)}\ |\lambda|, & \text{if spec}_{subr}(A) \neq \emptyset, \\ 0, & \text{if spec}_{subr}(A) = \emptyset. \end{cases} \end{aligned}$$

There is a relation for an irreducible positive matrix between its column sums, its spectral radius, and the unique eigenvector corresponding to its spectral radius.

Proposition 18.7.3. *Consider an irreducible positive matrix $A \in \mathbb{R}_+^{n\times n}$. From Theorem 18.7.1 follows that there exists an eigenvector $v \in \mathbb{R}^n$ corresponding to the spectral radius $r(A)$, $Av = r(A)v$. All elements of the vector v have the same sign. If* *the sign is not positive then multiply the vector with* -1 *to get a vector of which all elements of the vector* v *are strictly positive which is assumed below.*

Define the vector of column sums, and the maximum and the minimum of these column sums, and the maximal quotient q_{max} *of the elements of the eigenvector* v *as the variables,*

$$\text{colsum} \in \mathbb{R}_+^n, \ \forall\, i \in \mathbb{Z}_n, \ \text{colsum}_i = \sum_{j=1}^{n} A_{i,j};$$

$$\text{colsum}_{max} = \max_{i \in \mathbb{Z}_n} \text{colsum}_i \in (0,\infty), \ \text{colsum}_{min} = \min_{i \in \mathbb{Z}_n} \text{colsum}_i \in (0,\infty);$$

$$q_{max}(v) = \max_{i,\ j \in \mathbb{Z}_n} \left(\frac{v_i}{v_j} \right) \in (0,\infty),$$

$$v_{min} = \min_{i \in \mathbb{Z}_n} v_i \in (0,\infty), \ v_{max} = \max_{j \in \mathbb{Z}_n} v_j \in (0,\infty).$$

Then the following inequalities hold,

$$\text{colsum}_{min} \leq r(A) \leq \text{colsum}_{max}; \quad \frac{\text{colsum}_{max}}{\text{colsum}_{min}} \leq q_{max}(v)^2.$$

Equality holds in the above inequalities if and only if $\text{colsum}_{min} = \text{colsum}_{max}$. *Consequently, if equality holds,* $r(A) = \text{colsum}_{min} = \text{colsum}_{max}$.

Proof. The set of eigenvalues is defined by the determinant equation $0 = \det(\lambda I - A)$. From Theorem 18.7.1.(a) follows that the spectral radius $r(A)$ is an eigenvalue hence,

$$0 = \det(rI - A) = \det \begin{pmatrix} r - A_{1,1} & -A_{1,2} & \ldots & -A_{1,n-1} & -A_{1,n} \\ -A_{2,1} & r - A_{2,2} & \ldots & -A_{2,n-1} & A_{2,n} \\ \vdots & & \ddots & & \vdots \\ -A_{n-1,1} & -A_{n-1,2} & \ldots & r - A_{n-1,n-1} & -A_{n-1,n} \\ -A_{n,1} & -A_{n,2} & \ldots & -A_{n,n-1} & r - A_{n,n} \end{pmatrix}.$$

Define a linear transformation L which has as effect that the above structured matrix is transformed to one in which the columns 2 to n are all added to column 1. One then obtains the relation,

$$
\begin{aligned}
0 &= \det(L)\det(rI-A) = \det(L(rI-A)),\ L = \left(\begin{array}{c|ccc} 1 & 1 & \dots & 1 \\ \hline 0 & & & \\ \vdots & & I_{n-1} & \\ 0 & & & \end{array}\right) \in \mathbb{R}^{n\times n}_{nsng}, \\
&= \det \begin{pmatrix} r-\text{colsum}_1 & r-\text{colsum}_2 & \dots & r-\text{colsum}_{n-1} & r-\text{colsum}_n \\ -A_{2,1} & r-A_{2,2} & \dots & -A_{2,n-1} & A_{2,n} \\ \vdots & & \ddots & & \vdots \\ -A_{n-1,1} & -A_{n-1,2} & \dots & r-A_{n-1,n-1} & -A_{n-1,n} \\ -A_{n,1} & -A_{n,2} & \dots & -A_{n,n-1} & r-A_{n,n} \end{pmatrix} \\
&= \sum_{i=1}^{n} (r-\text{colsum}_i)\ \text{Adj}\ B_{i,1}, \qquad (18.6) \\
B_{1,j} &= (-1)^{1+j}\, M_{1,j},\ \text{called the } (1,j)\text{-th co-factor of } L(rI-A), \\
M_{1,j} &= \text{called the } (1,j)\text{-the } \textit{minor} \text{ of } L(rI-A),
\end{aligned}
$$

where the minor $M_{1,j}$ is defined as the determinant of the matrix obtained from $L(rI-A)$ be deleting the 1-th row and the j-th column. See for the minor and the co-factor, [53, Def. 7.1, 7.2].

The definition of an adjoint matrix is such that $(\lambda I - A)\ B(\lambda) = I\ \det(\lambda I - A)$ for all $\lambda \in \mathbb{C}$. Because the spectral radius is an eigenvalue of A it follows that $(rI-A)B(r) = (rI-A)\det(rI-A) = 0$ hence $A\,B(r) = r\,B(r)$. The columns of $B(r)$ are thus eigenvectors of A corresponding to the eigenvalue r. It follows from Theorem 18.7.1.(c) that any eigenvector of A corresponding to the spectral radius has no zero components and all components are of the same sign. The same argument applied to the transposed matrix A^T establishes that the elements of all rows of $B(r)$ have the same sign. Below it is supposed that the sign is positive hence $B_{1,j} > 0$ for all $j \in \mathbb{Z}_n$ though the proof also works in case it is negative.

Then it follows from equation (18.6) that,

$$
\begin{aligned}
&\text{colsum}_{min} = \text{colsum}_{i_{min}} \ \Rightarrow\ r - \text{colsum}_{i_{min}} \geq 0, \\
&\text{colsum}_{max} = \text{colsum}_{i_{max}} \ \Rightarrow\ r - \text{colsum}_{i_{max}} \leq 0, \\
\Rightarrow\ &\text{colsum}_{min} \leq r \leq \text{colsum}_{max}.
\end{aligned}
$$

Equality in the latter equation holds if and only if $\text{colsum}_{min} = \text{colsum}_{max}$.

Note that, because A is a positive matrix,

$$
\begin{aligned}
Av = r\,v &\Rightarrow \forall\, i \in \mathbb{Z}_n,\ r\,v_i = \sum_{j=1}^{n} A_{j,i} v_i \ \geq \ \sum_{j=1}^{n} A_{j,i} v_{min} = \text{colsum}_i\ v_{min} \\
&\Rightarrow \forall\, i \in \mathbb{Z}_n,\ \text{colsum}_i \leq \frac{r\,v_i}{v_{min}} \leq r\,q_{max} \ \Rightarrow\ \text{colsum}_{max} \leq r\,q_{max}; \\
&\quad \forall\, i \in \mathbb{Z}_n,\ r\,v_i = \sum_{j=1}^{n} A_{j,i} v_j \ \leq \ \sum_{j=1}^{n} A_{j,i} v_{max} = \text{colsum}_i\ v_{max} \\
&\Rightarrow \forall\, i \in \mathbb{Z}_n,\ \frac{1}{\text{colsum}_i} \ \leq \ \frac{v_{max}}{r\,v_i} \leq \frac{q_{max}}{r} \ \Rightarrow\ \frac{1}{\text{colsum}_{min}} \ \leq \ \frac{q_{max}}{r}; \\
&\Rightarrow \frac{\text{colsum}_{max}}{\text{colsum}_{min}} \ \leq \ q_{max}^2.
\end{aligned}
$$

□

18.8 Eigenvalues and Eigenvectors of Stochastic Matrices

This section is motivated by the following problem of state-finite stochastic systems.

Problem 18.8.1. *Convergence of a sequence of stochastic vectors and the equation of the invariant stochastic vector.* Consider the state process of a time-invariant finite stochastic system with a state-transition stochastic matrix $A \in \mathbb{R}_{st}^{n \times n}$. Define the sequence of stochastic vectors,

$$
\begin{aligned}
& n \in \mathbb{Z}_+,\ A \in \mathbb{R}_{st}^{n \times n},\ p_0 \in \mathbb{R}_{st}^n,\ p : T = \mathbb{N} \to \mathbb{R}_{st}^n, \\
& p(t+1) = Ap(t),\ p(0) = p_0.
\end{aligned}
$$

Determine whether or not the following limit exists, and whether, for the indicated equation for the invariant stochastic vector p_s, there exists a solution, etc.,

$$
\begin{aligned}
& \forall\, p_0 \in \mathbb{R}_{st}^n,\ \lim_{t \to \infty} p(t;0,p_0) = p(\infty) \in \mathbb{R}_{st}^n,\ p(\infty) = Ap(\infty)? \\
& \exists\, p_s \in \mathbb{R}_{st}^n \text{ such that } p_s = Ap_s?
\end{aligned}
$$

The reader best distinguishes the following three problem issues because they have unexpected answers as proven in the remainder of this section:

1. Does the exist a solution p_s of the steady-state equation $Ap_s = p_s$?
2. Is a solution of the steady-state equation unique?
   In general there is no uniqueness.
3. Does the sequence of stochastic vectors defined above converge, and, if so, is the limit a solution of the steady state equation?
   No convergence needs to take place in general and if convergence takes place, the limit value may depend on the initial stochastic vector.

The problem is analogous to that treated in Theorem 22.1.2 for a time-invariant Gaussian system where the focus is on the convergence of both the mean function and the covariance function of a stationary Gaussian process.

Theorem 18.8.2. *Consider a stochastic matrix $A \in \mathbb{R}_{st}^{n \times n}$.*

(a)Because $A \in \mathbb{R}_{st}^{n \times n}$ for an integer $n \in \mathbb{Z}_+$ is a stochastic matrix, its spectral radius equals one, $r(A) = 1$. The spectral radius is an algebraically-simple root of the determinantal equation $\det(\lambda I - A) = 0$ in $\lambda \in \mathbb{C}$, and hence an eigenvalue of the stochastic matrix A.

(b)Then there exists a steady state,

$$\exists\ p_s \in \mathbb{R}_{st}^n \text{ such that, } A p_s = p_s,\ 1_n^T p_s = 1.$$

Proof. (a) Because the matrix has all column sums equal to one it follows from Proposition 18.7.3 that $1 = \text{colsum}_{min} \le r(A) \le \text{colsum}_{max} = 1$ and hence that $r(A) = 1$. The remaining conclusions follow from the Perron-Frobenius Theorem 18.7.1.

(b) This result is a generalization of Theorem 18.7.1 from the case where A is irreducible to the case where the matrix A is not irreducible. □

A Partition of the Set of Stochastic Matrices

The set of irreducible stochastic matrices is best partitioned further into several subsets because the convergence analysis differs per subset.

Definition 18.8.3. Consider an irreducible stochastic matrix $A \in \mathbb{R}_{st}^{n \times n}$. Distinguish the following subsets of irreducible stochastic matrices of size $n \times n$ based on the spectral index of the corresponding stochastic matrix:

- an *irreducible and nonperiodic stochastic matrix*, also called a *primitive stochastic matrix*: $(n_1, n_-) = (1, n-1)$, hence there is only one eigenvalue $\lambda \in \text{spec}(A)$ satisfying $|\lambda| = r(A) = 1$ while $n_- = n-1$ eigenvalues satisfy $|\lambda| < 1$; this seems to be the most interesting special case of the three;
- an *irreducible and partly-periodic stochastic matrix*: $(n_1, n_-) = (k, n-k)$ for $k \in \mathbb{Z}_n$ satisfying $1 < k < n$; and
- an *irreducible and pure-periodic stochastic matrix*: $(n_1, n_-) = (n, 0)$, hence all eigenvalues $\lambda \in \text{spec}(A)$ satisfy $|\lambda| = r(A) = 1$; this is the least interesting of the three cases.

Proposition 18.8.4. *The classification of irreducible stochastic matrices of Def. 18.8.3 by the spectral indices, is a complete classification and a partition.*

Proof. From Theorem 18.8.2 follows that a stochastic matrix $A \in \mathbb{R}_{st}^{n \times n}$ is such that its spectral radius is an eigenvalue of the matrix, equivalently, $r(A) \in \text{spec}(A)$. The classification of the definition then exhausts the values for n_r by the cases: (1) $n_{st,si}(A) = (1, n-1)$, (2) $n_{st,si}(A) = (k, n-k)$ for $2 \le k \le n-1$; and (3) $n_{st,si}(A) = (n, 0)$. Thus the definition is a complete classification. □

Theorem 18.8.5. Equivalent conditions for an irreducible and nonperiodic stochastic matrix. *Consider an irreducible and nonperiodic stochastic matrix A hence $n_{si}(A) = (1, n-1)$. The following statements are equivalent:*

(a)the matrix A is an irreducible and nonperiodic stochastic matrix;
(b)there exists a matrix $A(\infty) \in \mathbb{R}_{st}^{n\times n}$ such that $\lim_{k\to\infty} A^k = A(\infty)$*;*
(c)there exists an integer $m \in \mathbb{Z}_+$ such that $A^m \in \mathbb{R}_{s+,st}^{n\times n}$, which denotes that all its elements are strictly positive; and
(d)for all $k \in \mathbb{Z}_+$, the stochastic matrix A^k is irreducible.

Overview of Convergence Results

To assist the reader with his or her comprehension, there follows an overview of the results for Problem 18.8.1 based on the above-defined classification of the transition matrix A of a time-invariant state-finite stochastic system. The definition below is an overview of the results stated below in this section.

Definition 18.8.6. *An overview of the results for convergence of a sequence of stochastic vectors based on algebraic and spectral concepts.*
Consider Problem 18.8.1 with a stochastic matrix $A \in \mathbb{R}_{st}^{n\times n}$ which is the state-transition matrix of a state-finite stochastic system.

- Stochastic matrices – Irreducible stochastic matrices.
  - Case of an irreducible and nonperiodic stochastic matrix with spectral index $n_{si}(A) = (1, n-1)$.
    Theorem 18.8.7 states that $\lim_{t\to\infty} p(t) = p(\infty)$, the equation $p_s = Ap_s$ has a unique solution, that $p_s \in \mathbb{R}_{s+,st}^{n}$, and that $p(\infty) = p_s$.
  - Case of an irreducible and partly-periodic stochastic matrix with spectral index $n_{si}(A) = (k, n-k)$ where $1 < k < n$.
    Theorem 18.8.8 establishes that in this case the sequence of stochastic vectors is periodic in general and hence does not converge. There exists an invariant stochastic vector $p_s \in \mathbb{R}_{st}^{n_x}$. The equation $Ap_s = p_s$ does not have a unique solution.
  - Case of an irreducible and pure-periodic stochastic matrix with spectral index $n_{si}(A) = (n, 0)$.
    Theorem 18.8.9 establishes that in this case the transition matrix A is permutation similar to a unit-shift circulant stochastic matrix. The sequence $\{p(t),\, t \in T\}$ is in general periodic and hence does not converge. There exists an invariant stochastic vector $p_s \in \mathbb{R}_{st}^{n}$ such that $p_s = Ap_s$.
- Stochastic matrices – Fully-reduced. Proposition 18.8.12 shows that in general the stochastic matrix A, which is reducible, is permutation similar to a fully-reduced stochastic matrix consisting of the following diagonal blocks: (1) $m_1 \in \mathbb{Z}_n$ terminal irreducible and nonperiodic stochastic matrices and (2) $m_2 \in \mathbb{Z}_n$ irreducible substochastic matrices.
  The sequence of stochastic vectors converges. The limit value depends in general on the initial stochastic vector p_0. The equation $Ap_s = p_s$ does in general not have a unique solution.

Irreducible Matrices

The consequences of the Perron-Frobenius theorem are summarized for the three subsets of irreducible stochastic matrices. Each theorem describes the three problem items of the case considered even if that means that there is an overlap with other theorems.

Theorem 18.8.7. *Consider an irreducible and nonperiodic stochastic matrix $A \in \mathbb{R}_{st}^{n \times n}$ hence with spectral index $n_{si}(A) = (1, n-1)$.*

(a)Then,

$$\exists \text{ a unique strictly-positive eigenvector } p_s \in \mathbb{R}^n_{s+,st} \text{ such that,}$$

$$A p_s = p_s, \;\; 1_n^T p_s = 1; \;\; (v \in \mathbb{R}^n_{s+,st} \;\Rightarrow\; (\forall\, i \in \mathbb{Z}_n, \; v_i > 0)).$$

(b)Recall the definition of the sequence of stochastic vectors,

$$p(t+1; 0, p_0) = A p(t; 0, p_0), \; p(0; 0, p_0) = p_0 \in \mathbb{R}^n_{st}.$$

Then, for any initial stochastic vector $p_0 \in \mathbb{R}^n_{st}$, the sequence of stochastic vectors converges to the unique steady state p_s,

$$\lim_{t \to \infty} p(t; 0, p_0) = p_s.$$

Proof. (a) It follows from Theorem 18.8.2 that the spectral radius of A is one, $r(A) = 1$. It follows from Theorem 18.7.1 and from the matrix A being irreducible that there exists a vector $v \in \mathbb{R}^n_+ \backslash \{0\}$ such that $Av = v$, that $\sum_{i=1}^n \; v_i > 0$, hence $(\sum_{i=1}^n v_i)^{-1}$ is well defined. Define $p_s = v(\sum_{i=1}^n \; v_i)^{-1}$ then $p_s \in \mathbb{R}^n_{s+,st}$. From the definition of p_s follows that,

$$A p_s = A v (\sum_{i=1}^n \; v_i)^{-1} = v (\sum_{i=1}^n \; v_i)^{-1} = p_s, \;\; 1_n^T p_s = 1_n^T v (\sum_{i=1}^n v_i)^{-1} = 1.$$

The uniqueness follows from Theorem 18.7.1.(c).
(b) Note that,

$$p(t+1) - p_s = A(p(t) - p_s), \; p(0) - p_s = p_0 - p_s.$$

Define the linear transformation L such that,

$$L = \begin{pmatrix} I_{n-1} & 0 \\ 1_{n-1}^T & 1 \end{pmatrix} \in \mathbb{R}_{nsng}^{n \times n},\ L^{-1} = \begin{pmatrix} I_{n-1} & 0 \\ -1_{n-1}^T & 1 \end{pmatrix},\ Lp = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_{n-1} \\ \sum_{i=1}^n p_i \end{pmatrix},$$

$$LAL^{-1} = \begin{pmatrix} A_{11} & A_{12} \\ 0 & 1 \end{pmatrix},$$

$$A_{11} \in \mathbb{R}_+^{(n-1)\times(n-1)},\ A_{12} \in \mathbb{R}_+^{n-1},$$

$$\text{spec}(A_{11}) = \text{spec}(A)\backslash\{1\} \subset \text{D}_o,\ \ n_{si}(A_{11}) = (0, n-1),$$

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = L(p(t) - p_s),\ x(t) \in \mathbb{R}^{n-1},\ y(t) = y(0) = 1^T(p_0 - p_s) = 0 \in \mathbb{R},$$

$$\begin{pmatrix} x(t+1) \\ y(t+1) \end{pmatrix} = LAL^{-1} \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x(t) \\ 0 \end{pmatrix} = \begin{pmatrix} A_{11}x(t) \\ 0 \end{pmatrix},$$

$$x(t+1) = A_{11}x(t),\ x(0) = x_0 \in \mathbb{R}^{n-1}.$$

Choose any matrix $W \in \mathbb{R}_{spds}^{(n-1)\times(n-1)}$, hence $0 \prec W$, and let $B \in \mathbb{R}^{(n-1)\times(n-1)}$ be such that $W = B^T B$ and such that (A_{11}, B) is a controllable pair. For example $W = I_{n-1}$ and $B = I_{n-1}$ satisfy these conditions. Because $\text{spec}(A_{11}) \subseteq \text{D}_o$ there exists by Theorem 22.1.2 a matrix,

$$Q \in \mathbb{R}_{spds}^{(n-1)\times(n-1)} \text{ such that } Q = A_{11}^T Q A_{11} + W,$$

and Q is the unique solution of the above equation and (A_{11}, B) a controllable tuple implies that $0 \prec Q$.

It follows from Proposition 21.9.5 that $\lim_{t\to\infty} x(t) = 0$. Hence,

$$\lim_{t\to\infty} x(t) = 0;\ \forall\, t \in T,\ y(t) = 0;\ p(t) - p_s = L^{-1} \begin{pmatrix} x(t) \\ y(t) \end{pmatrix},$$

$$\Rightarrow \lim_{t\to\infty} (p(t) - p_s) = 0 \ \Leftrightarrow\ \lim_{t\to\infty} p(t) = p_s.$$

□

Consider next the partial-periodic case of an irreducible stochastic matrix.

Theorem 18.8.8. The case of an irreducible and partial-periodic stochastic matrix. *Consider an irreducible and partly-periodic stochastic matrix $A \in \mathbb{R}_{st}^{n\times n}$ and recall that its spectral index is $n_{si}(A) = (k, n-k)$ for an integer $k \in \mathbb{N}_n$ satisfying $1 < k < n$.*

(a)The matrix A is permutation similar to a block-shift matrix with k shifts,

$$\exists\, Q \in \mathbb{R}_{perm}^{n \times n}\ \exists\, n_1, n_2, \ldots, n_k \in \mathbb{Z}_+,\ n = n_1 + n_2 + \ldots + n_k,$$
$$\forall\, i \in \mathbb{Z}_{k-1},\ \exists\, A_{i,i+1} \in \mathbb{R}_{st}^{n_i \times n_{i+1}},\ \exists\, A_{1,k} \in \mathbb{R}_{st}^{n_1 \times n_k},\ \textit{such that,}$$
$$A = Q \begin{pmatrix} 0 & 0 & 0 \ldots 0 & A_{1,k} \\ A_{12} & 0 & 0 \ldots 0 & 0 \\ 0 & A_{32} & 0 \ldots 0 & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & 0 \ldots 0 & 0 \\ 0 & 0 & 0 \ldots A_{k,k-1} & 0 \end{pmatrix} Q^T.$$

(b) *The matrix A^k is a block-diagonal matrix with as the k diagonal blocks the submatrices,*

$$B_{1,1} = A_{12}A_{23}\ldots A_{k,k-1}A_{k,1} \in \mathbb{R}_{st}^{n_1 \times n_1},$$
$$B_{2,2} = A_{23}\ldots A_{k,k-1}A_{k,1}A_{12} \in \mathbb{R}_{st}^{n_2 \times n_2},\ \ldots,$$
$$B_{k,k} = A_{k,1}\ A_{12}A_{23}\ldots A_{k-1,k} \in \mathbb{R}_{st}^{n_k \times n_k}.$$

Each of those submatrices of the diagonal-block has the spectral index, $n_{si}(B_{i,i}) = (1, n_i - 1)$ for all $i \in \mathbb{Z}_k$.

(c) *There exists an invariant stochastic vector $p_s \in \mathbb{R}_{s+,st}^n$ which is a solution of the steady-state equations $Ap_s = p_s$ and $1_n^T p_s = 1$, see Theorem 18.8.2(b).*

(d) *There exists an invariant stochastic vector $p_s \in \mathbb{R}_{st}^n$ with components $p_{s,i} \in \mathbb{R}_+^{n_i}$ such that $A_{i,i+1}\ldots A_{i-1,i}p_{s,i} = p_{s,i}$ for all $i \in \mathbb{Z}_k$, and for all $m \in \mathbb{Z}_+$, $(A^k)^m p_s = p_s$.*

(e) *In general, there exist initial stochastic vectors $p_0 \in \mathbb{R}_{st}^n$ such that the sequence has a periodically shifting support. Consequently, for such an initial stochastic vector,* no convergence *takes place.*

Proof. (a) This follows from Theorem 18.7.1.(b).
(b) It follows from the decomposition of (a), the fact that permutation matrices are stochastic matrices, and from the closure of stochastic matrices with respect to multiplication, that $Q^T A^k Q$ is a stochastic matrix. Thus each of the block-diagonal submatrices $B_{i,i}$ of the diagonal-block matrix $Q^T A^k Q$ is a stochastic matrix. The spectral radius of A^k is equal to one, $r(A^k) = r(A)^k = 1^k = 1$ and consequently $n_{si}(A^k) = (k, n-k)$. Because for all $i \in \mathbb{Z}_k$, $B_{i,i} \in \mathbb{R}_{st}^{n_i \times n_i}$ is a stochastic matrix it follows that its spectral radius $1 = r(B_{i,i}) \in \text{spec}(B_{i,i})$. Because $\text{spec}(A^k) = \cup_{i=1}^k \text{spec}(B_{i,i})$ it follows that for any $i \in \mathbb{Z}_k$, $n_{si}(B_{i,i}) = (1, n_i - 1)$.
(c) This follows from Theorem 18.8.2.(b).
(d) Because by (b) each of the block diagonal submatrices $B_{i,i}$ of A^k is an irreducible and nonperiodic stochastic matrix with spectral index $(1, n_i - 1)$, it follows from Theorem 18.8.2.(b) that there exists a unique vector $p_{s,i}$ such that the relation of part (d) holds. If each such vector is scaled to a length of $1/k$ then the combined vector satisfies statement (d),

$$p_s = \left(p_{s,1}^T(1/k)\ p_{s,2}^T(1/k)\ \ldots\ p_{s,k}^T(1/k) \right)^T.$$

(e) This follows from Example 18.8.10.

□

Consider finally the case of a pure-cyclic irreducible stochastic matrix.

Theorem 18.8.9. The case of a pure-cyclic irreducible stochastic matrix. *Consider an irreducible and pure-periodic stochastic matrix $A \in \mathbb{R}_{st}^{n \times n}$ and recall that its spectral index equals $(n_1, n_-) = (n, 0)$ hence $n_{cyc}(A) = n$.*

(a)There exists a permutation matrix $Q \in \mathbb{R}_{perm}^{n \times n}$ such that the matrix A is permutation similar to a unit-shift circulant stochastic matrix,

$$A = QW_nQ^T = Q \begin{pmatrix} 0 \, 0 & 0 \dots 0 \, 1 \\ 1 \, 0 & 0 \dots 0 \, 0 \\ 0 \, 1 & 0 \dots 0 \, 0 \\ \vdots \ \ddots & \ddots \quad \vdots \\ 0 \, 0 & 0 \dots 0 \, 0 \\ 0 \, 0 & 0 \dots 1 \, 0 \end{pmatrix} Q^T.$$

(b)There exists a unique invariant stochastic vector $p_s \in \mathbb{R}_{s+,st}^{n_x}$ satisfying $Ap_s = p_s$, defined by $p_s = (1/n)1_n$.

(c)There exist initial stochastic vectors $p_0 \in \mathbb{R}_{st}^n$ such that the sequence of stochastic vectors $p(.;0,p_0) : \mathbb{N} \to \mathbb{R}_{st}^n$ is a periodic sequence with period n, hence no convergence takes place in general.

Proof. (a) The matrix decomposition follows from Theorem 18.7.1.(b). Because by assumption the index of cyclicity equals the size of the matrix, $n_{cyc}(A) = n$, it follows that each of the blocks in the decomposition is a real number and not a matrix. Because the matrix A is a stochastic matrix and the relation between A and the decomposed matrix is by permutation similarity, the transformed matrix Q^TAQ is again a stochastic matric, hence every column contains only one strictly positive real number which thus equals one. It then follows from the definition of the unit-shift circulant stochastic matrix that the decomposed matrix equals W_n, see Def. 18.3.1.

(b) Clearly $W_np_s = W_n(1/n)1_n = (1/n)1_n = p_s$.

(c) The statement is illustrated in Example 18.8.10 for the case of a 6×6 matrix. □

Example 18.8.10. *A stochastic vector sequence with periodically shifting support.* Consider the irreducible stochastic matrix A, an initial stochastic vector p_0, and compute the stochastic vector sequence,

$$A = \left(\begin{array}{cc|c|ccc} 0 & 0 & 0 & 0.7 & 0.5 & 0.2 \\ 0 & 0 & 0 & 0.3 & 0.5 & 0.8 \\ \hline 1 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.3 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0 & 0 & 0 \end{array} \right) \in \mathbb{R}_{st}^{6\times 6}, \;\; n_{cyc}(A) = 3,$$

$$p_0 = p_x(0) = \left(\begin{array}{c} 0.7 \\ 0.3 \\ \hline 0 \\ \hline 0 \\ 0 \\ 0 \end{array} \right); \;\; p_x = \left[p_x(0) \; p_x(1) \; p_x(2) \; p_x(3) \right] = \left[\begin{array}{cccc} 0.7 & 0 & 0 & 0.54 \\ 0.3 & 0 & 0 & 0.46 \\ \hline 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0.3 & 0 \\ 0 & 0 & 0.2 & 0 \end{array} \right].$$

The state set of the system can be partitioned into the state subsets as,

$$X = \{1,2\} \cup \{3\} \cup \{4,5,6\} = X_1 \cup X_2 \cup X_3.$$

Hence the support of the stochastic measure sequence circulates periodically along the sequence of state subsets $X_1 \mapsto X_2 \mapsto X_3 \mapsto X_1$ etc.

Note that the third power of the matrix A equals,

$$A^3 = \begin{pmatrix} A_{13}A_{32}A_{21} & 0 & 0 \\ 0 & A_{21}A_{13}A_{32} & 0 \\ 0 & 0 & A_{32}A_{21}A_{13} \end{pmatrix}.$$

Then convergence takes place along the sequence,

$$\{p_{x_1}(3t) = A^{3t} p_0 \in \mathbb{R}_{st}^{6}, \; t \in \mathbb{N}\}, \;\; \lim_{t \to \infty} A^{3t} p_0 = p_s,$$

where p_s is as defined in Theorem 18.8.8.(d).

The conclusion of this example is: if there exists an index of cyclicity $n_{cyc}(A) \in \{2,3,\ldots,n-1\}$ then there may exists initial stochastic vectors p_0 such that the sequence of stochastic vectors has periodically varying support, hence no convergence can take place to the steady state vector p_s.

Reducible Matrices

The case of a reducible stochastic matrix is by a decomposition reduced to that of a set of irreducible stochastic matrices.

Definition 18.8.11. *Frobenius canonical form.*

Define the *Frobenius canonical form* of a reducible stochastic matrix by the representation,

$$A = \left(\begin{array}{cccc|cccc}
A_{11} & 0 & 0 & 0 & A_{1,n_1+1} & A_{1,n_1+2} & \ldots A_{1,n_2-1} & A_{1,n_2} \\
0 & A_{22} & \ldots & 0 & A_{2,n_1+1} & A_{2,n_1+2} & \ldots A_{2,n_2-1} & A_{2,n_2} \\
0 & 0 & \ddots & 0 & \vdots & \vdots & \vdots \;\; \vdots & \vdots \\
0 & 0 & \ldots & A_{n_1,n_1} & A_{n_1,n_1+1} & A_{n_1,n_1+1} & \ldots A_{n_1,n_2-1} & A_{n_1,n_2} \\
\hline
0 & 0 & \ldots & 0 & A_{n_1+1,n_1+1} & A_{n_1+1,n_1+2} & \ldots A_{n_1+1,n_2-1} & A_{n_1+1,n_2} \\
0 & 0 & \ldots & 0 & 0 & A_{n_1+2,n_1+2} & \ldots A_{n_1+2,n_2-1} & A_{n_1+2,n_2} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \vdots & \vdots \\
0 & 0 & \ldots & 0 & 0 & 0 & \ldots A_{n_2-1,n_2-1} & A_{n_2-1,n_2} \\
0 & 0 & \ldots & 0 & 0 & 0 & \ldots 0 & A_{n_2,n_2}
\end{array}\right)$$

$$\begin{aligned}
&\in \mathbb{R}_{st}^{n\times n}, \;\; \exists\, m_1,\, m_2 \in \mathbb{Z}_+,\; m_1 + m_2 = n, \\
&\exists\, n_1,\, n_2,\; \ldots,\; n_{m_1} \in \mathbb{Z}_+,\; n_{m_1+1},\; n_{m_1+2}, \ldots,\; n_{m_1+m_2} \in \mathbb{Z}_+, \\
&n_1 + \ldots + n_{m_1} + n_{m_1+1} + \ldots + n_{m_2} = n, \\
&\forall\, i \in \mathbb{Z}_{m_1},\; A_{i,i} \in \mathbb{R}_{st}^{n_i\times n_i} \text{ irreducible stochastic matrix}, \\
&\forall\, i \in \mathbb{Z}_{m_1},\; \forall\, j = n_1+1, \ldots, n_2,\; A_{i,j} \in \mathbb{R}_+^{n_i\times n_j}, \\
&\quad \forall\, k_2 = n_1+1, \ldots, n_2,\; \exists\, k_1 \in \{1, \ldots, k_2-1\},\; A_{(i,j),(k_1,k_2)} \neq 0; \\
&\text{in words, for every column } k_2 \text{ there exists a row element } k_1 \\
&\text{such that } A_{k_1,k_2} \neq 0, \\
&\forall\, i = n_1+1, \ldots, n_2,\; A_{i,i} \in \mathbb{R}_+^{n_i\times n_i} \text{ is substochastic.}
\end{aligned}$$

Call the submatrices $A_{i,i}$ for all $i \in \mathbb{Z}_{n_1}$ the *terminal submatrices* and the submatrices $A_{i,i}$ for $i = n_1+1, \ldots, n_2$ the *transient submatrices*.

Theorem 18.8.12. Permutation similarity of a reducible stochastic matrix.

(a)Any reducible stochastic matrix is permutation similar to a matrix in the Frobenius canonical form of Def. 18.8.11.
(b)The square submatrix bordered by the submatrices A_{n_1+1,n_1+1}, A_{n_1+1,n_2}, A_{n_2,n_2} and a zero matrix, has a spectral radius strictly less than one, $r(A_{n_1+1:n_2,n_1+1:n_2}) < 1$.

Proof. (a) If the A matrix is reducible then there exists a permutation similarity transformation such that the transformed A matrix has the form displayed in Def. 18.6.2. If in the obtained new stochastic matrix there exist a block-diagonal matrix which is not irreducible then apply another permutation similarity transform to reduce that block further. Because of the matrix has a finite size $n \in \mathbb{Z}_+$ the process of reducing ends after a finite number of steps with a block matrix with all diagonal blocks being irreducible matrices. It could be that one or more of such blocks are of size 1×1. By construction, there exists at least one irreducible block-diagonal matrix which is the $(1,1)$ block and has no other blocks above it. Then apply another permutation similarity transform to put, after the first block-column, all block-columns with only a diagonal block, meaning there are no nonzero off-diagonal blocks in the same column. The remaining columns are moved to the right in the blocked matrix. For all $i = n_1+1, \ldots, n_2$ and for all k_2 there exists a

k_1 such that $A_{(i,j),(k_1,k_2)} \neq 0$ hence $A_{i,i}$ is substochastic. Then the form displayed in the proposition is obtained. □

To save space, below the case of a fully-reduced stochastic matrix is treated with only two diagonal blocks of irreducible stochastic matrices.

Example 18.8.13. Consider the stochastic matrix,

$$A = \left(\begin{array}{cc|cc} 0.2 & 0.3 & 0.25 & 0 \\ 0.8 & 0.7 & 0 & 0.3 \\ \hline 0 & 0 & 0.25 & 0.3 \\ 0 & 0 & 0.50 & 0.4 \end{array} \right) = \left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline 0 & A_{22} \end{array} \right);$$

$$\sum_{i=1}^{2} A_{22,(i,1)} = 0.75 < 1, \; \sum_{i=1}^{2} A_{22,(i,2)} = 0.7 < 1.$$

Then A_{22} is a column-wise substochastic matrix. The matrix,

$$B = \left(\begin{array}{cc|cc} 0.2 & 0.3 & 0 & 0 \\ 0.8 & 0.7 & 0 & 0.3 \\ \hline 0 & 0 & 0.5 & 0.3 \\ 0 & 0 & 0.5 & 0.4 \end{array} \right) = \left(\begin{array}{c|c} B_{11} & B_{12} \\ \hline 0 & B_{22} \end{array} \right);$$

$$\sum_{i=1}^{2} B_{22,(i,1)} = 1, \; \sum_{i=1}^{2} B_{22,(i,2)} = 0.7 < 1.$$

is substochastic but not column-wise substochastic.

Note further that,

$$\text{spec}(B_{11}) = \{1, \, 0.5\}, \; p_{1,2} = \begin{pmatrix} 3 \\ 8 \end{pmatrix} (1/11) \in \mathbb{R}^2_{st}, \; B_{11} p_{1,2} = p_{1,2}.$$

The convergence is first treated for an elementary example.

Example 18.8.14. Consider the elementary fully-reduced stochastic matrix,

$$A = \begin{pmatrix} 1 & 1-s \\ 0 & s \end{pmatrix} \in \mathbb{R}^{2\times 2}_{st}, \;\; \text{with } s \in (0,1) \subset \mathbb{R}.$$

Then the sequence of measures converges according to,

$$\begin{aligned} & \text{spec}(A) = \{s, 1\}, \; \lambda_1(A) = r(A) = 1, \; \lambda_2(A) = s < 1 = r(A), \\ & \forall \, t \in \mathbb{Z}_+, \, t \geq 2, \, A^t = \begin{pmatrix} 1 & 1-s^t \\ 0 & s^t \end{pmatrix}; \;\; \forall \, p_0 \in \mathbb{R}^2_{st}, \\ & \lim_{t\to\infty} p_x(t) = \lim A^t p_x(0) = \lim \begin{pmatrix} p_{0,1} + p_{0,2}(1-s^t) \\ p_{0,2} s^t \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \in \mathbb{R}^2_{st}. \end{aligned}$$

Theorem 18.8.15. *Consider a fully-reduced stochastic matrix of the particular form,*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \; n \in \mathbb{Z}_+, \; n_1, \; n_2 \in \mathbb{Z}_n, \; n = n_1 + n_2,$$

$A_{11} \in \mathbb{R}_{st}^{n_1 \times n_1}$, *stochastic, irreducible, and nonperiodic,*
hence with $n_{st,si}(A_{11}) = (1, n-1)$,
$A_{22} \in \mathbb{R}_+^{n_2 \times n_2}$ *is irreducible and column-wise substochastic, and*
$A_{12} \in \mathbb{R}_+^{n_1 \times n_2}$, $A_{12} \neq 0$.

Define the sequence of stochastic vectors on $\mathbb{R}_{st}^n$*,*

$$p_x : T = \mathbb{N} \rightarrow \mathbb{R}_{st}^n, \; p_{x,1} : T \rightarrow \mathbb{R}_+^{n_1}, \; p_{x,2} : T \rightarrow \mathbb{R}_+^{n_2}, \; p_0 \in \mathbb{R}_{st}^n,$$

$$p_x(t+1; p_0) = A p_x(t; p_0), \; p_x(0; p_0) = p_0, \;\; p_x(t; p_0) = \begin{pmatrix} p_{x,1}(t; p_0) \\ p_{x,2}(t; p_0) \end{pmatrix}.$$

(a)For the set of steady-state equations of the first state component $p_{x,1}$*,*

$p_1 \in \mathbb{R}_{st}^{n_1}, \; A_{11} p_1 = p_1, \; 1_{n_1}^T p_1 = 1,$
there exists a unique solution denoted by $p_{s,1} \in \mathbb{R}_{s+,st}^{n_1}$.

(b)For the set of steady state equations for the matrix A,

$p = Ap, \; 1_n^T p = 1,$ *there exists a unique solution which has the form,*

$$p_s = \begin{pmatrix} p_{s,1} \\ 0 \end{pmatrix},$$

where $p_{s,1} \in \mathbb{R}_{s+,st}^{n_1}$ *is determined as in (a).*

(c)For any $p_0 \in \mathbb{R}_{st}^n$*, the sequence of stochastic vectors of the second component of* p_x*,* $p_{x,2}$*, converges to the unique steady state* $\lim_{t \rightarrow \infty} \; p_{x_2}(t; p_0) = 0$.

(d)$\lim_{t \rightarrow \infty} \; p_{x,1}(t) = p_{s,1}$.

(e)Assume that there exists an eigenvalue $\lambda(A) \in \mathbb{C}$ *of the matrix A such that* $|\lambda| < r(A) = 1$*. Then, the sequence of stochastic vectors converges to the unique steady state,*

$$\forall \; p_0 \in \mathbb{R}_{st}^n, \; \lim_{t \rightarrow \infty} \; p_x(t; p_0) = \lim_{t \rightarrow \infty} A^t p_0 = p_s,$$

where p_s *is as determined in (b).*

The same conclusions (a)-(d) hold if A_{22} *is irreducible and substochastic.*

Proof. (a) Because by assumption the matrix $A_{11} \in \mathbb{R}_{st}^{n_1 \times n_1}$ is stochastic, irreducible, and nonperiodic, the result follows from Theorem 18.8.7.(a).
(b) The steady state equations of item (b) are equivalent with,

$$p_1 = A_{11} p_1 + A_{12} p_2, \;\; p_2 = A_{22} p_2, \; 1_{n_1}^T p_1 + 1_{n_2}^T p_2 = 1.$$

Clearly the vector $p_{s,2} = 0 \in \mathbb{R}_+^{n_2}$ is a solution of the equation for p_2 of the form $p_2 = A_{22} p_2$.

Suppose that there exists a vector $p_2 \in \mathbb{R}_+^{n_2 \times n_2}$ satisfying $p_2 = A_{22} p_2$ which is nonzero. It will be proven that this supposition leads to a contradiction.

If $p_2 \neq 0$ then there exists an integer $k \in \mathbb{Z}_{n_2}$ such that $p_{2,k} > 0$. Because by assumption the matrix A_{22} is column-wise substochastic, for all $j \in \mathbb{Z}_{n_2}$, $\text{colsum}_j = \sum_{i=1}^{n_2} A_{22,i,j} < 1$. Then,

$$1_{n_2}^T p_2 = 1_{n_2}^T A_{22} p_2 = \begin{pmatrix} \text{colsum}_1 \dots \text{colsum}_{n_2} \end{pmatrix} p_2 < 1_{n_2}^T p_2.$$

This is a contradiction of the supposition that a vector $p_2 \neq 0$ exists. Thus $p_2 = 0$.

The first steady-state equation now becomes, $p_1 = A_{11}p_1 + A_{22}p_2 = A_{11}p_1$. It then follows from (a) that the steady state equation $p_1 = A_{11}p_1$ and $1_{n_1}^T p_1 = 1$ has a unique solution $p_{s,1} \in \mathbb{R}_{st}^{n_1}$. Then the combined vector p_s is a solution of the steady state equation,

$$p_s = \begin{pmatrix} p_{s,1} \\ 0 \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} \begin{pmatrix} p_{s,1} \\ 0 \end{pmatrix} = Ap_s.$$

(c) The second component of the stochastic vector sequence satisfies the recursion,

$$p_{x_2}(t+1) = A_{22}p_{x_2}(t),\ p_{x_2}(0) = p_{0,2},\ \ 0 = p_{s,2} = A_{22}p_{s,2}.$$

Define the candidate Lyapunov function,

$$\begin{aligned}
V : \mathbb{R}_+^{n_2} \to \mathbb{R}_+,\ V(p) = 1_{n_2}^T p;\ \forall\ p \in \mathbb{R}_+^{n_2},\ V(p) \geq 0;& \\
0 = V(p) = 1_{n_2}^T p \ \Leftrightarrow\ p = 0;\ \text{and } V \text{ is a continuous function};& \\
V(p_{x,2}(t+1)) &= 1_{n_2}^T p_{x,2}(t+1) = 1_{n_2}^T A_{22} p_{x,2}(t) \\
&= \begin{pmatrix} \text{colsum}_1 \dots \text{colsum}_{n_2} \end{pmatrix} p_{x,2}(t) \\
&\begin{cases} < 1_{n_2}^T p_{x,2}(t) = V(p_{x,2}(t)), & \text{if } p_{x,2}(t) \neq 0, \\ = 0, & \text{if } p_{x_2}(t) = 0. \end{cases}
\end{aligned}$$

In the above strict inequality, the assumption is used that the matrix A_{22} is column-wise substochastic.

It then follows from the Lyapunov theorem that $\lim_{t\to\infty} V(p_{x,2}(t)) = 0$. Then $V(p) = 0 \ \Leftrightarrow\ p = 0$ and continuity of V imply that $\lim_{t\to\infty} p_{x,2}(t) = 0$.

(d) This follows from Theorem 18.8.7.

(e) The limit for the full sequence $\{p_x(t) \in \mathbb{R}_{st}^n,\ \forall\ t \in \mathbb{N}\}$ decomposes into the following two limits, $\lim_{t\to\infty} p_{x,2}(t)$ and $\lim_{t\to\infty} p_{x,1}(t)$. From (c) follows that for any $p_0 \in \mathbb{R}_{st}^n$, $\lim_{t\to\infty} p_{x,2}(t;p_0) = 0$. From the system equations then follows that,

$$\lim_{t\to\infty} p_{x,1}(t+1) = \lim A_{11}p_{x,1}(t) + \lim A_{12}p_{x,2}(t) = \lim A_{11}p_{x,1}(t),$$

where the conclusion of (c) for the limit of the sequence $p_{x,2}$ is used. From (d) then follows that, $\lim_{t\to\infty} p_{x,1}(t) = p_{s,1}$, hence that,

$$\lim_{t\to\infty} p_x(t) = \begin{pmatrix} \lim p_{x,1}(t) \\ \lim p_{x,2}(t) \end{pmatrix} = \begin{pmatrix} p_{s,1} \\ 0 \end{pmatrix} = p_s.$$

□

Theorem 18.8.16. *Consider a fully-reduced stochastic matrix A of the particular form with four blocks,*

$$A = \begin{pmatrix} A_{11} & 0 & A_{13} & A_{14} \\ 0 & A_{22} & A_{23} & A_{24} \\ 0 & 0 & A_{33} & A_{34} \\ 0 & 0 & 0 & A_{44} \end{pmatrix}, \; n \in \mathbb{Z}_+, \; n_1, \; n_2, n_3, \; n_4 \in \mathbb{Z}_n, \; n = n_1 + n_2 + n_3 + n_4,$$

$A_{11} \in \mathbb{R}_{st}^{n_1 \times n_1}$, $A_{22} \in \mathbb{R}_{st}^{n_2 \times n_2}$,
irreducible and nonperiodic stochastic matrices,
with $n_{si}(A_{11}) = (1, n_1 - 1)$, $n_{si}(A_{22}) = (1, n_2 - 1)$,
$A_{33} \in \mathbb{R}_+^{n_3 \times n_3}$ $A_{44} \in \mathbb{R}_+^{n_4 \times n_4}$ *are both irreducible and*
both column-wise substochastic, and
$\forall \; (i,j) = (1,3), \; (1,4), \; (2,3), \; (2,4), \; (3,4), \; A_{ij} \in \mathbb{R}_+^{n_i \times n_j}$.

Define the sequence of stochastic vectors on $\mathbb{R}_{st}^n$ *as stated before. Denote a decomposition of the stochastic vector* $p(t)$ *by,*

$$p(t) = \begin{pmatrix} p_1(t) \\ p_2(t) \\ p_3(t) \\ p_4(t) \end{pmatrix} \in \mathbb{R}_{st}^n, \; \forall \; i \in \mathbb{Z}_4, \; p_i(t) \in \mathbb{R}_+^{n_i}.$$

(a)For the steady state stochastic vector associated with the matrix A, the last two components have the unique solution,

$$p = Ap, \; 1_n^T p = 1, \; p_s = \begin{pmatrix} p_{s,1} \\ p_{s,2} \\ 0 \\ 0 \end{pmatrix}.$$

(b)There exist unique vectors $p_{s,1} \in \mathbb{R}_{st}^{n_1}$ *and* $p_{s,2} \in \mathbb{R}_{st}^{n_2}$ *such that* $A_{11} p_{s,1} = p_{s,1}$ *and* $A_{22} p_{s,2} = p_{s,2}$. *Then,*

$$\forall \; c \in [0,1], \; p_s = c \begin{pmatrix} p_{s,1} \\ 0 \\ 0 \\ 0 \end{pmatrix} + (1-c) \begin{pmatrix} 0 \\ p_{s,2} \\ 0 \\ 0 \end{pmatrix} \; \Rightarrow \; Ap_s = p_s, \; 1_n^T p_s = 1.$$

Therefore the equation $Ap_s = p_s$ *does not have a unique solution.*

(c)For any $p_0 \in \mathbb{R}_{st}^n$, *the sequence of stochastic vectors for the third and the fourth component of p, converge to the unique steady state* $\lim_{t \to \infty} \; p_3(t; p_0) = 0$ *and* $\lim_{t \to \infty} \; p_4(t; p_0) = 0$.

The same conclusions (a)-(b) hold if the matrices A_{33} *and* A_{44} *are irreducible and substochastic.*

The proof is analogous to that of the previous theorem.

Weighted Average of Powers

The next result is used in stochastic control of a state-finite stochastic control system with complete observations on an infinite horizon with the average cost function.

Theorem 18.8.17. Convergence of the weighted average of the power of a stochastic transition matrix. *Consider a stochastic transition matrix $A \in \mathbb{R}_{st}^{n \times n}$ for an integer $n \in \mathbb{Z}_+$. Define the sequence of matrices,*

$$\{S(t) \in \mathbb{R}_{+}^{n \times n},\ \forall\, t \in \mathbb{N}\},\ \ S(0) = A,\ S(t) = \frac{1}{t} \sum_{s=0}^{t-1} A^s.$$

(a)The following limit exists and satisfies the stated equation,

$$\lim_{t \to \infty} S(t) = Q^* \in \mathbb{R}_{st}^{n \times n}, \tag{18.7}$$

$$Q^* = AQ^* = Q^*A. \tag{18.8}$$

Call the positive matrix Q^* the weighted power-average *of A in fact the weighted average of the sum of the powers of the stochastic matrix A.*

(b)The equation (18.8) for a matrix $Q^* \in \mathbb{R}_{st}^{n \times n}$, has a unique solution.

(c)If in addition the state transition matrix A is irreducible and nonperiodic, hence its spectral index equals $n_{si}(A) = (1, n-1)$, then there exists a strictly-positive vector $p \in \mathbb{R}_{s+,st}^{n}$ such that $Ap = p$ and,

$$Q^* = \begin{pmatrix} p & p & \dots & p \end{pmatrix} \in \mathbb{R}_{st}^{n \times n}.$$

Proof. (1) Because A is a stochastic matrix, $A1_n = 1_n$ and, for all $t \in T$, $S(t)1_n = \sum_{s=0}^{t-1} A^s\ 1_n/t = 1_n$ hence $S(t)$ is a stochastic matrix.
(2) For all $i,\ j \in \mathbb{Z}_n$ and for all $t \in T,\ \ S(t)_{i,j} \in [0,1]$ which is a compact subset of $\mathbb{R}$. It follows from a theorem of analysis, see [6, Thm. 2.17], that there exists a subsequence,

$$\{t_k \in \mathbb{Z}_+,\ k \in \mathbb{Z}_+\} \text{ such that } \lim_{k \to \infty} S(t_k)_{i,j} = Q_{1,i,j} \in [0,1].$$

Thus one obtains a matrix $Q_1 \in \mathbb{R}_{+}^{n \times n}$ such that $\lim_{k \to \infty} S(t_k) = Q_1$. Because for all $t \in T$, $S(t) \in \mathbb{R}_{st}^{n \times n}$ it follows that $Q_1 \in \mathbb{R}_{st}^{n \times n}$. For another subsequence one obtains,

$$\{s_k \in \mathbb{Z}_+,\ k \in \mathbb{Z}_+\}\ \lim_{k \to \infty} S(s_k) = Q_2 \in \mathbb{R}_{st}^{n \times n}.$$

(3) Note that,

$$A \times S(t) - S(t) = \frac{1}{t} \sum_{s=1}^{t} A^s - \frac{1}{t} \sum_{s=0}^{t-1} A^s = [A^t - I]/t,\ \ \forall\, t \in T,$$
$$AQ_1 - Q_1 = \lim_{k \to \infty} [AS(t_k) - S(t_k)] = \lim\ [A^{t_k} - I]/t_k = 0,$$

where the latter limit is zero because $A^{t_k} \in \mathbb{R}_{st}^{n \times n}$ is a stochastic matrix and $\lim_{k \to \infty} t_k = +\infty$. Thus one obtains $AQ_1 = Q_1$. Similarly one proves that $Q_1 A = Q_1$. By induction one proves that $A^2 Q_1 = A(AQ_1) = AQ_1 = Q_1 = Q_1 A = (Q_1 A)A =$

Q_1A^2 and, for all $t \in \mathbb{Z}_+$, $A^tQ_1 = Q_1 = Q_1A^t$. Similarly one proves that, for all $k \in \mathbb{Z}_+$, $Q_2^kA = Q_2 = AQ_2^k$.
(4) Consider two matrices $Q_1,\ Q_2 \in \mathbb{R}_{st}^{n\times n}$ such that $AQ_1 = Q_1 = Q_1A$ and $AQ_2 = Q_2 = Q_2A$, Then it follows from the above that for all $t \in \mathbb{Z}_+$, $A^tQ_1 = Q_1 = Q_1A^t$ and $A^tQ_2 = Q_2 = Q_2A^t$. These equalities and the assumed limits along the subsequences, imply that,

$$\begin{aligned} Q_2Q_1 &= \lim_{k\to\infty} S(s_k)Q_1 = \lim \frac{1}{s_k}\sum_{r=0}^{s_k-1} A^rQ_1 = Q_1 \lim \frac{1}{s_k}\sum_{r=0}^{s_k-1} A^r = Q_1Q_2 \\ &= \lim \frac{1}{s_k}\sum_{r=0}^{s_k-1} A^r = Q_2,\ \Rightarrow\ Q_2Q_1 = Q_2 = Q_1Q_2; \\ & Q_1Q_2 = Q_1 = Q_2Q_1,\ \text{by symmetry; hence } Q_1 = Q_2Q_1 = Q_2. \end{aligned}$$

Thus the two limits along subsequences are equal and the limit exists. Denote the limit from now by $Q^* \in \mathbb{R}_{st}^{n\times n}$. This proves (a) and (b)
(c) If the matrix A has the properties of the theorem statement then it follows from Theorem 18.8.7 that there exists a unique vector $p \in \mathbb{R}_{s+,st}^n$ such that $Ap = p$. Define,

$$Q = \begin{pmatrix} p\ p\ \ldots\ p \end{pmatrix} \in \mathbb{R}_{st}^{n\times n}.$$

From the definition of Q and the definition of the vector p, follows that $AQ = Q$ while QA is such that for all $i,\ k \in \mathbb{Z}_n$, $(QA)_{i,k} = \sum_{j=1}^n p_iA_{j,k} = p_i$ hence $QA = Q$. Thus $AQ = Q = QA$ and the uniqueness of Step (4) above shows that $Q^* = Q$. □

18.9 Multiplicative Factorization

18.9.1 Problem

For any real matrix $A \in \mathbb{R}^{k\times m}$ there exists a singular value decomposition $A = USV^T$ where U and V are orthogonal matrices and S is the matrix with the singular values of the matrix A which has in general a partly diagonal form. The matrix of singular values determines the rank of the matrix A and the relative scaling of the singular values determines the magnitude of the operations on various subspace of the domain.

Is there an analog of the singular value decomposition for positive matrices? This question leads to the concept of equivalence of a tuple of positive matrices. The analogue in the positive matrices of the singular value matrix for real matrices is more complex than a diagonal matrix.

As an introduction to the following sections, the multiplicative factorization of a positive matrix is briefly summarized by three concepts and theorems.

The first concept to be used is equivalence of positive matrices. Two square positive matrices $A,\ B \in \mathbb{R}_+^{n\times n}$ are called *monomially equivalent* if there exist two monomial matrices $M_1,\ M_2 \in \mathbb{R}^{n\times n}$, they may be different, such that $A = M_1BM_2$. It is proven below that any fully-indecomposible positive matrix is diagonally-equivalent to a doubly stochastic matrix $B \in \mathbb{R}_{dst}^{n \times n}$.

A second concept and theorem to be used is that any doubly-stochastic matrix is a convex sum of permutation matrices. $B = \sum_{i=1}^{n!} a_i Q_i$. This use of this result requires a study on the factorization of convex sums of permutation matrices.

A third concept and theorem of positive matrices refers to the definition of a prime in the positive matrices, ect. This concept puts the algebraic framework of factorizations of square positive matrices on a sound algebraic basis.

With these results it is possible to derive a description of the multiplicative decomposition of positive matrices from which the internal structure can be directly read from the factorization. The author is indebted to a large set of researchers who have formulated these concepts and derived the theorems. The role of the author is limited to the integration of this theory.

The theory of positive matrices proceeds with the concept of primes in the positive matrices. This concept was formulated and investigated by D.J. Richman and H. Schneider. This algebraic concept has been quite useful in unraveling the algebraic structure of positive matrices.

Definition 18.9.1. *Primes in the positive matrices.* Consider a subset of the positive matrices $X \subseteq \mathbb{R}_{+}^{n \times n}$ and denote its set of units by $X_{unit} \subseteq X$. Call a positive matrix $A \in X$ a *prime in the subset* X if (1) it is not a unit; or, equivalently, $A \notin X_{unit}$; and (2) if there exists a factorization $A = B \times C$ with $B,\ C \in X$, then either B or C is a unit hence belongs to X_{unit}.

Recall that the group of units of the positive matrices is the subset of monomial matrices, $\mathbb{R}_{mon}^{n \times n} \subset \mathbb{R}_{+}^{n \times n}$.

A positive matrix $A \in \mathbb{R}_{+}^{n \times n}$ is called a *prime in the positive matrices* if the following two conditions both hold: (1) A is not a monomial matrix; and (2) if $A = B \times C$ where $B,\ C \in \mathbb{R}_{+}^{n \times n}$ then either B or C is a monomial matrix.

Correspondingly one defines *prime in the doubly-stochastic matrices* with as set of units the permutation matrices and a *prime in the circulant doubly-stochastic matrices* with as group of units the unit-shift circulant doubly-stochastic matrices.

Example 18.9.2. In the positive matrices of size 3×3 there is only one prime matrix of the form,

$$\exists\, M_1,\ M_2 \in \mathbb{R}_{mon}^{n \times n},\ \text{such that,}\ A = M_1 \frac{1}{2} \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} M_2 \in \mathbb{R}_{+}^{3 \times 3}.$$

Example 18.9.3. The following matrices are primes in the positive matrices of size 4×4.

$$A = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \in \mathbb{R}_{prime,+}^{4 \times 4};\ B = \begin{pmatrix} 0.4 & 0 & 0.3 & 0.3 \\ 0.3 & 0.4 & 0 & 0.3 \\ 0.3 & 0.3 & 0.4 & 0 \\ 0 & 0.3 & 0.3 & 0.4 \end{pmatrix} \in \mathbb{R}_{prime,crcl,dst}^{4 \times 4}.$$

These examples terminate the introduction to positive-matrix factorization.

18.9.2 Equivalence

Note that equivalence is a relation between positive matrices which differs from similarity as is directly clear from the definition.

Definition 18.9.4. Consider two square positive matrices $A,\ B \in \mathbb{R}_{+}^{n\times n}$. Call these matrices:

(a)*permutation equivalent* if there exists two permutation matrices $Q_1,\ Q_2 \in \mathbb{R}_{perm}^{n\times n}$ such that $A = Q_1 B Q_2$;

(b)*diagonally equivalent* if there exists two diagonal-positive matrices with strictly-positive diagonals $D_1,\ D_2 \in \mathbb{R}_{s+,diag}^{n\times n}$ such that $A = D_1 B D_2$; and

(c)*monomially equivalent* if there exists two monomial matrices $M_1,\ M_2 \in \mathbb{R}_{mon}^{n\times n}$ such that $A = M_1 B M_2$.

One also says in case (a) that $Q_1 B Q_2$ is a *permutation scaling* of the positive matrix B, and, correspondingly, a *diagonal scaling* for case (b) above, and a *monomial scaling* in case of (c).

The difference between similarity and equivalence of positive matrices is in the formula relating the matrices A and B, for equivalence one needs Q_1 and Q_2 which need not be related while for similarity one must have $Q_2 = Q_1^T$. Both concepts have their particular use in the decomposition of positive matrices, similarity for the state transition matrix and its decomposition, and equivalence for the multiplicative factorization of a positive matrix.

The relation of equivalence for the various subsets of positive matrices is in each case an equivalence relation as defined in Def. 17.1.2.

Analogous to reducibility of positive matrices one defines decomposibility of positive matrices.

Definition 18.9.5. Call a square positive matrix $A \in \mathbb{R}_{+}^{n\times n}$ for $n \in \mathbb{Z}_{+}$ *partly decomposible* if,

$$\begin{aligned}
&\exists\, n_1,\ n_2 \in \mathbb{Z}_{+},\ n_1 + n_2 = n,\ \ \exists\, Q_1,\ Q_2 \in \mathbb{R}_{perm}^{n\times n},\\
&\exists\, A_{11} \in \mathbb{R}_{+}^{n_1\times n_1},\ A_{22} \in \mathbb{R}_{+}^{n_2\times n_2},\ A_{12} \in \mathbb{R}_{+}^{n_1\times n_2},\ \text{such that,}\\
A &= Q_1 \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} Q_2.
\end{aligned}$$

Call the matrix A *fully indecomposible* or *indecomposible* if it is not partly decomposible. Call the matrix A *decomposed* or *fully decomposed* if

$$\begin{aligned}
&\exists\, k \in \mathbb{Z}_n\ (k = 1 \text{ is allowed}),\ \exists\, n_1,\ n_2,\ \ldots,\ n_k \in \mathbb{Z}_n,\\
&\forall\, i \in \mathbb{Z}_k,\ \exists\, A_{ii} \in \mathbb{R}_{+}^{n_i\times n_i},\ \text{which is indecomposible, such that,}\\
A &= \begin{pmatrix}
A_{11} & A_{12} & A_{13} & \ldots & A_{1,k-1} & A_{1,k} \\
0 & A_{22} & A_{23} & \ldots & A_{2,k-1} & A_{2,k} \\
0 & 0 & A_{33} & \ldots & A_{3,k-1} & A_{3,k} \\
\vdots & & & \ddots & & \vdots \\
0 & 0 & 0 & \ldots & A_{k-1,k-1} & A_{k-1,k} \\
0 & 0 & 0 & \ldots & 0 & A_{kk}
\end{pmatrix},\ \ n = \sum_{i=1}^{k} n_i. \qquad (18.9)
\end{aligned}$$

The above definition differs from that of the book [3, p. 75].

The reader should carefully distinguish the terms of a positive matrix being partly decomposible, fully indecomposible, and decomposed. The use of these terms differs slightly between references.

Theorem 18.9.6. *Consider a positive matrix $A \in \mathbb{R}^{n \times n}$ which is partly decomposible. There exist permutation matrices $Q_1,\ Q_2 \in \mathbb{R}^{n \times n}_{perm}$ such that matrix $Q_1 A Q_2$ has the representation of equation (18.9).*

Proof. Because the matrix A is partly decomposible, there exists permuation matrices $Q_1,\ Q_2$ such that $Q_1 A Q_2$ has the form displayed in Def. 18.9.5. If any of the two diagonal block matrices is not indecomposible then apply to each of the two block-matrices a further permutation equivalence. Because the matrix has finite size, this procedure stops after a finite number of steps with the indicated fully decomposed matrix. □

Theorem 18.9.7. A characterization of diagonal equivalence of a positive matrix. *Consider a positive matrix $A \in \mathbb{R}^{n \times n}_{+}$ for an integer $n \in \mathbb{Z}_+$.*

(a)If A is a fully indecomposible positive matrix then there exist two diagonal matrices each with a strictly-positive diagonal, $D_1,\ D_2 \in \mathbb{R}^{n \times n}_{s+,diag}$, such that,

$$B = D_1 A D_2 \in \mathbb{R}^{n \times n}_{dst}, \tag{18.10}$$

is a doubly stochastic matrix. Then the formula $A = D_1^{-1} B D_2^{-1}$ shows that A is diagonally equivalent to a doubly stochastic matrix.

(b)The diagonal matrices $D_1,\ D_2 \in \mathbb{R}^{n \times n}_{diag,s+}$ of part (a) are unique up to a scaling factor. Thus, if the tuple (D_1, D_2) are diagonal matrices which achieve the transformation and if $c \in (0, \infty) \subset \mathbb{R}_+$ then $(cD_1, c^{-1} D_2)$ is another such tuple.

(c)There exists a procedure which, starting from the fully indecomposible matrix A, constructs two sequences of positive matrices $\{D_1(k),\ D_2(k) \in \mathbb{R}^{n \times n}_{s+,diag},\ k \in \mathbb{Z}_+\}$ such that the following transformation sequence converges to a doubly stochastic matrix,

$$\lim_{k \to \infty} D_1(k) A D_2(k) = D_1(\infty) A D_2(\infty) = B \in \mathbb{R}^{n \times n}_{dst}.$$

18.9.3 Permutation Matrices

Needed in the subsequent sections is an additive decomposition over a set of permutation matrices and a Latin square positive matrix. These concepts are defined in this section.

Definition 18.9.8. Consider for an integer $n \in \mathbb{Z}_+$ and the set of permutation matrices of size $n \times n$ denoted by $\mathbb{R}^{n \times n}_{perm}$. Note that there are $n!$ distinct permutation matrices. For any subset denote,

$$J \subset \mathbb{Z}_{n!},\ \ \mathbb{R}^{n\times n}_{perm}(J) = \{Q_j \in \mathbb{R}^{n\times n}_{perm},\ \forall\ j \in J\}.$$

A *permutation cover* of the set of $n \times n$ positive matrices is defined as the subset of permutations,

$$\exists\ J_c \subset \mathbb{Z}_{n!},\ \text{ such that } |J_c| = n,\ E_n = \sum_{j \in J_c} Q_j,\ \text{and denote these properties by}$$

$$\{Q_j \in \mathbb{R}^{n\times n}_{perm},\ \forall\ j \in J\} \in \text{CoverPerm}(\mathbb{R}^{n\times n}_{+}),$$

where the matrix E_n is the one matrix, see Def. 18.3.1.

Call a subset of permutations $\mathbb{R}^{n\times n}_{perm}(J)$ *closed with respect to multiplication* if for any two matrices $Q_i,\ Q_j \in \mathbb{R}^{n\times n}_{perm}(J)$ it is true that $Q_iQ_j \in \mathbb{R}^{n\times n}_{perm}(J)$

Call a subset of permutations $\mathbb{R}^{n\times n}_{perm}(J)$ *closed with respect to inversion* if for any matrix $Q_i \in \mathbb{R}^{n\times n}_{perm}(J)$ it is true that $Q_i^{-1} = Q_i^T \in \mathbb{R}^{n\times n}_{perm}(J)$.

If a subset of permutations is a multiplicative group then it is closed with respect to multiplication and to inversion.

Example 18.9.9. The set of the downward unit-shifts for $n \in \mathbb{Z}_+$, $\{W_n^i \in \mathbb{R}^{n\times n}_{+},\ \forall\ i \in \mathbb{N}_{n-1}\}$, is a permutation covering of the positive matrices of size $n \times n$. In case of $n = 4$ one has,

$$W_4^1 = \begin{pmatrix} 0\,0\,0\,1 \\ 1\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,1\,0 \end{pmatrix},\ W_4^2 = \begin{pmatrix} 0\,0\,1\,0 \\ 0\,0\,0\,1 \\ 1\,0\,0\,0 \\ 0\,1\,0\,0 \end{pmatrix},\ W_4^3 = \begin{pmatrix} 0\,1\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,1 \\ 1\,0\,0\,0 \end{pmatrix},\ W_4^4 = W_4^0 = I_4,$$

$$\{W_4^1,\ W_4^2,\ W_4^3,\ W_4^4 = I_4\} \in \text{CoverPerm}(\mathbb{R}^{4\times 4}_{+}).$$

Example 18.9.10. Consider the set of permutations, based on powers of the unit shifts,

$$J = \{1,\ 2,\ 3\},\ Q_{6,2} = \{W_6^2,\ W_6^4,\ I_6\}.$$

Then this set of permutations is closed with respect to multiplications and with respect to inversion. But it is not a permutation cover of the 6×6 positive matrices.

18.9.4 Circulant Doubly-Stochastic Matrices

The focus of this section is on circulant doubly-stochastic matrices. Note that the set of such matrices of size $n \times n$ for $n \in \mathbb{Z}_+$ is closed with respect to multiplication. The group of units is the set of unit-shift circulant doubly-stochastic matrices.

The main research issue is: What is a characterization of all primes in the circulant doubly-stochastic matrices? The results below are derived using an algebraic approach by associating any matrix in this set by a polynomial.

Definition 18.9.11. Recall the definition of a *circulant doubly-stochastic matrix* as a matrix A for which there exists a decomposition of the form,

$$A = \sum_{i=1}^{n} a_i W_n^i \in \mathbb{R}_{dst}^{n\times n},\ \exists\, a \in \mathbb{R}_{st}^n,\ \{W_n^i \in \mathbb{R}_{shift}^{n\times n},\ \forall\, i \in \mathbb{Z}_n\};$$

$$\mathbb{R}_{crcl,dst}^{n\times n} = \{\sum_{i=1}^{n} a_i W_n^i \in \mathbb{R}_{dst}^{n\times n} |\ a \in \mathbb{R}_{st}^n\}.$$

The set of matrices $\{W_n^i,\ i \in \mathbb{N}_n\}$ are powers of the unit shift matrix W_n, see Def. 18.3.1. Note that $W_n^n = I_n$.

It is then obvious from the definition of the unit shift matrix that the relation $a \in \mathbb{R}_{st}^n \mapsto A = \sum_{i=1}^n\ a_i W_n^i \in \mathbb{R}_{cdst}^{n\times n}$ is a bijection.

Example 18.9.12. A *circulant doubly-stochastic matrix*. Consider the matrix,

$$A = \begin{pmatrix} a_3 & a_2 & a_1 \\ a_1 & a_3 & a_2 \\ a_2 & a_1 & a_3 \end{pmatrix} = a_1 W_3 + a_2 W_3^2 + a_3 W_3^3,\ a = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \in \mathbb{R}_{st}^n.$$

Definition 18.9.13. Define for any circulant doubly stochastic matrix its *polynomial representer* by the relation,

$$A = \sum_{i=1}^{n} a_i W_n^i \in \mathbb{R}_{cdst}^{n\times n} \mapsto\ q_a(z) = \sum_{i=0}^{n-1} a_i z^i;\ W_n^n = I\ \Rightarrow\ z^n = 1 = z^0,$$

$$a\ \mapsto\ \sum_{i=0}^{n-1} a_i z^i/(z^n-1) = q_a(z)/(z^n-1) \in \mathbb{R}_{st}^n[z].$$

Define the sets,

$$\mathbb{R}_{st}^n[z]/(z^n-1) = \{\sum_{i=1}^{n} a_i z^i/(z^n-1)|\ \exists\, a \in \mathbb{R}_{st}^n\},$$

$$\mathbb{R}_{+}^n[z]/(z^n-1) = \{\sum_{i=1}^{n} a_i z^i/(z(n-1)|\ \exists\, a \in \mathbb{R}_{+}^n\}.$$

Then $\mathbb{R}_+^n[z]/(z^n-1)$ is a ring because it is closed with respect to addition and to multiplication. However, $\mathbb{R}_{st}^n[z]/(z^n-1)$ is not a ring, it is not closed with respect to addition.

The map $A = \sum_{i=1}^n\ a_i W_n^i\ \mapsto\ \sum_{i=0}^{n-1}\ a_i z^i/(z^n-1) = q_a(z)/(z^n-1)$ for $a \in \mathbb{R}_+$ is a ring homomorphism.

The matrix $A \in \mathbb{R}_{cdst}^{n\times n}$ is a unit in $\mathbb{R}_{cdst}^{n\times n}$ if and only if there exists a natural number $k \in \mathbb{N}_{n-1}$ such that $q_a(z) = z^k$.

There follow several results on primes in the circulant doubly stochastic matrices.

The theory for primes in the positive matrices was primarily developed by J.M. van den Hof. The theory makes use of polynomials in an indeterminate variable and factorization of such polynomials, which framework is not described in this book.

Definition 18.9.14. Define a *prime in the circulant doubly-stochastic matrices* if Def. 18.9.1 holds for circulant doubly-stochastic matrices and its group of units. The group of units equals the set of downward-shift doubly-stochastic matrices.

Definition 18.9.15. A polynomial $q \in \mathbb{R}_+[z]/(z^n - 1)$ is a prime in this set if (1) it is not a monomial ($\nexists\ k \in \mathbb{N}_{n-1}$ such that $q(z) = z^k$); (2) if $q(z) = b(z) \times c(z)$ mod $(z(n-1))$ with $b,\ c \in \mathbb{R}_+[z]$ then neither b nor c is a monomial.

Proposition 18.9.16. Necessary condition for a prime in the circulant doubly-stochastic matrices. *If $A = \text{crcl}(a) \in \mathbb{R}^{n\times n}_{circdst}$, see Def. 18.3.1, for a $n \in \mathbb{Z}_+$ and a vector $a \in \mathbb{R}^n_{st}$, is a prime in the circulant doubly-stochastic matrices then $1 < n_{s+}(a) < n$.*

Theorem 18.9.17. Sufficient condition of a prime in the circulant doubly-stochastic matrices. *If $A = \text{crcl}(a) \in \mathbb{R}^{n\times n}_{circdst}$, $n \geq 3$, $a \in \mathbb{R}^n_{st}$, and $n_{s+}(a) = 2$, then A is a prime in the circulant doubly-stochastic matrices.*

Theorem 18.9.18. Characterization of primes in the order-two circulant doubly-stochastic matrices. *Consider a circulant doubly-stochastic matrix $A \in \mathbb{R}^{4\times 4}_{dsc}$ with $A = \text{crcl}(a)$, $a \in \mathbb{R}^4_{st}$, and $n_{s+}(a) = 2$.*

Then A is a prime in the circulant doubly-stochastic matrices if and only if either case (1) or case (2) below holds,

$$(1)\ \exists\ a = \begin{pmatrix} a_1 \\ a_2 \\ 0 \\ 0 \end{pmatrix} \in \mathbb{R}^4_{st},\ a_1 + a_2 = 1,\ n_{s+}(a) = 2,\ i(a) = \{1,\ 2\},$$

$$\exists\ j,\ k \in \mathbb{N}_3 \textit{ such that,}$$

$$A = W_4^k \text{crcl}(a) W_4^j = W_4^k \begin{pmatrix} a_1 & 0 & 0 & a_2 \\ a_2 & a_1 & 0 & 0 \\ 0 & a_2 & a_1 & 0 \\ 0 & 0 & a_2 & a_1 \end{pmatrix} W_4^j;$$

$$(2)\ \exists a = \begin{pmatrix} a_1 \\ 0 \\ a_3 \\ 0 \end{pmatrix} \in \mathbb{R}^4_{st},\ a_1 + a_3 = 1,\ n_{s+}(a) = 2,\ i(a) = \{1,\ 3\},$$

$$\exists\ j,\ k \in \mathbb{N}_3,\ \textit{such that,}$$

$$A = W_4^k \text{crcl}(a) W_4^j = W_4^k \begin{pmatrix} a_1 & 0 & a_3 & 0 \\ 0 & a_1 & 0 & a_3 \\ a_3 & 0 & a_1 & 0 \\ 0 & a_3 & 0 & a_1 \end{pmatrix} W_4^j.$$

Theorem 18.9.19. Characterization of primes in the order-three circulant doubly-stochastic matrices. *Consider a circulant doubly-stochastic matrix,*

$$A = \text{crcl}(a) = \sum_{i=1}^{n} a_1 W_4^i,\ \ a = \begin{pmatrix} a_1 & a_2 & a_3 & 0 \end{pmatrix}^T \in \mathbb{R}^4_{st},\ n_{s+}(a) = 3.$$

Then A is a prime in the circulant doubly-stochastic matrices if and only if $a_2^2 < 4a_1a_3$.

Example 18.9.20. If

$$a = \begin{pmatrix} 1/3 \; 1/3 \; 1/3 \; 0 \end{pmatrix}^T \in \mathbb{R}^4_{st},\ a_2^2 = (1/3)^2 < 4(1/3)^2 = 4a_1a_3, \text{ then,}$$

$$A = \text{crcl}(a) = \begin{pmatrix} 1/3 & 0 & 1/3 & 1/3 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1/3 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 1/3 \end{pmatrix} \in \mathbb{R}^{4\times 4}_{dst}.$$

is a prime in the circulant doubly-stochastic matrices.

For circulant doubly-stochastic matrices of size 5 × 5 and larger one can similarly classify the existence of primes in the circulant doubly-stochastic matrices.

It is unfortunate that the existence of a prime in the circulant doubly-stochastic matrices requires an analytic condition as $a_2^2 < 4a_1a_3$. This makes the structure of the theory less clear. But if the condition is not met then a different factorization exists. This issue requires further investigation.

18.9.5 Doubly-Stochastic Matrices

Additive Factorizations

Concepts and results are collected on additive decompositions of positive matrices. This extends the definition of a circulant matrix of Def. 18.3.1.

The motivation to study additive decompositions is a theorem of G. Birkhoff published in 1946. The reader is expected to know the concept of a polyhedron, see Def. 17.6.5.

Theorem 18.9.21. *Consider the set of doubly-stochastic matrices $\mathbb{R}^{n\times n}_{dst}$ with an integer $n \in \mathbb{Z}_+$.*

(a)The set of doubly-stochastic matrices equals a convex polyhedron whose vertices are the permutation matrices.
(b)

$$\begin{aligned} &\forall\, n \in \mathbb{Z}_+,\ \ \forall\, A \in \mathbb{R}^{n\times n}_{dst},\ \ \exists\, n_a \in \mathbb{Z}_+,\ n_a \leq n^2 - 2n + 2, \\ &\exists\, a \in \mathbb{R}^{n_a}_{st},\ \exists\, \{Q_i \in \mathbb{R}^{n\times n}_{perm},\ i \in J\} \in \text{CoverPerm}(\mathbb{R}^{n\times n}_+) \textit{ such that,} \\ A = &\sum_{i \in J} a_j Q_j,\ \ J \subset \mathbb{Z}_{n!},\ |J| = n_a. \end{aligned}$$

Definition 18.9.22. A doubly-stochastic matrix $A \in \mathbb{R}^{n\times n}_{dst}$ for $n \in \mathbb{Z}_+$ is called a *convex sum of permutations* if there exists a stochastic vector $a \in \mathbb{R}^n_{st}$ such that $A = \sum_{i=1}^{n!} a_i Q_i$ where $\{Q_i \in \mathbb{R}^{n\times n}_{perm},\ i \in \mathbb{Z}_{n!}\}$ is an enumeration of all permutations of the indicated size.

Define a *Latin Square in the positive matrices* by the expressions,

$$L : \mathbb{R}^n_+ \times \text{Pwrset}(\mathbb{Z}_{n!}) \to \mathbb{R}^{n\times n}_+,\ L(a, J_c) = \sum_{j \in J_c} a_j Q_j,\ |J_c| = n.$$

The doubly-stochastic matrix $A \in \mathbb{R}_{dst}^{n \times n}$ is called a *doubly-stochastic Latin Square* or a *Latin Square in the doubly-stochastic matrices* if there exists a permutation cover of the positive matrices $\mathbb{R}_{+}^{n \times n}$ such that,

$$\begin{aligned}
&\exists\ \mathbb{R}_{perm}^{n \times n}(J_c) = \{Q_j \in \mathbb{R}_{perm}^{n \times n},\ \forall\ j \in J\} \in \text{CoverPerm}(\mathbb{R}_{+}^{n \times n}),\\
&\exists\ a \in \mathbb{R}_{st}^{n},\ \text{such that}\ A = \sum_{i \in I} a_i Q_i \in \mathbb{R}_{dst}^{n \times n};\\
&\text{define}\ L : \mathbb{R}_{st}^{n} \times \text{CoverPerm}(\mathbb{R}_{+}^{n \times n}) \to \mathbb{R}_{dst}^{n \times n},\ L(a, J_c) = \sum_{j \in J_c} a_j Q_j.
\end{aligned}$$

Proposition 18.9.23. Characterizations of a sum of permutation matrices. *Consider the set of all permutation matrices of size $n \times n$, for an integer $n \in \mathbb{Z}_{+}$, $\{Q_i \in \mathbb{R}_{perm}^{n \times n},\ i \in \mathbb{Z}_{n!}\}$.*

(a)For any stochastic vector $a \in \mathbb{R}_{st}^{n!}$ the following matrix is doubly stochastic, $A = \sum_{k=1}^{n!}\ a_k Q_k \in \mathbb{R}_{dst}^{n \times n}$.
(b)For any doubly-stochastic matrix $A \in \mathbb{R}_{dst}^{n \times n}$ there exists a stochastic vector $a \in \mathbb{R}_{dst}^{n!}$ such that $A = \sum_{i=1}^{n!}\ a_i Q_i$. In general, the vector a is not uniquely determined by A.
(c)Consider a permutation cover $\text{Cover}(\mathbb{R}_{+}^{n \times n}, \mathbb{R}_{perm}^{n \times n}) = \{Q_i,\ j \in J_c\}$ *with $J_c \subset \mathbb{Z}_{n!}$ and $|J_c| = n$. If $A \in \mathbb{R}_{dst}^{n \times n}$ and $a \in \mathbb{R}_{st}^{n}$ satisfy $A = \sum_{k \in J_c}\ a_k Q_k$ then the relation from a to A is injective; equivalently, $\sum_{k \in J_c}\ a_k Q_k = A = \sum_{j \in J_c}\ b_j Q_j$ implies that for all $j \in J_c$, $a_j = b_j$.*
(d)Assume that $A = \sum_{i=1}^{n!}\ a_i Q_i$ for a vector $a \in \mathbb{R}_{st}^{n \times n}$
Then A is not a permutation matrix if and only if the a vector is of order at least two; or, equivalently, $2 \leq n_{s+}(a)$.

Proof. (a) Note that,

$$A 1_n = \sum_{k=1}^{n!} a_k Q_k 1_n = (\sum a_k) 1_n = 1_n,\ 1_n^T A = 1_n^T \sum_{k=1}^{n!} a_k Q_k = (\sum a_k) 1_n^T = 1_n^T.$$

(b) This follows from Theorem 18.9.21.
(c) This result follows from the definition of a permutation cover.
(d) This follows because the permutation cover is an exhaustive description of all permutations of the indicated size. If there exist two different permutations in a sum then there exist both a row and a column each of which has two nonzero elements.
□

Example 18.9.24. Consider the following doubly-stochastic matrix and its decomposition as a sum of permutation matrices,

$$
\begin{aligned}
A &= \begin{pmatrix} a_1 & a_2+a_4 & a_3 \\ a_2 & a_3 & a_1+a_4 \\ a_3+a_4 & a_1 & a_2 \end{pmatrix} \in \mathbb{R}^{3\times 3}_{dst}, \;\; a = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} \in \mathbb{R}^4_{st}, \\
&= a_1 \begin{pmatrix} 1&0&0 \\ 0&0&1 \\ 0&1&0 \end{pmatrix} + a_2 \begin{pmatrix} 0&1&0 \\ 1&0&0 \\ 0&0&1 \end{pmatrix} + a_3 \begin{pmatrix} 0&0&1 \\ 0&1&0 \\ 1&0&0 \end{pmatrix} + a_4 \begin{pmatrix} 0&1&0 \\ 0&0&1 \\ 1&0&0 \end{pmatrix} \\
&= a_1Q_1 + a_2Q_2 + a_3Q_3 + a_4Q_4.
\end{aligned}
$$

Note that when considering a general A matrix, the elements of the vector a of the above representation are uniquely determined.

For the factorization of Latin Square in the positive matrices the following concept is useful.

Definition 18.9.25. The *Latin Square induced by multiplication of permutations* in the positive matrices of size $n \times n$ where $n \in \mathbb{Z}_+$ is defined as,

$$
\begin{aligned}
&\exists\, J_c,\;\; \mathrm{CoverPerm}(\mathbb{R}^{n\times n}) = \{Q_k \in \mathbb{R}^{n!\times n!}_{perm},\; \forall\, k \in J_c \subseteq \mathbb{Z}_{n!}\}, \\
&L_{n!} : \mathbb{R}^{n!}_{+} \to \mathbb{R}^{n!\times n!}_{+},\;\; [L_{n!}(x)_{k,j}] = x_i,\;\; \text{if } Q_iQ_j = Q_k \;(\Leftrightarrow\; Q_i = Q_kQ_j^T), \\
&\forall\, Q_i,\, Q_j,\, Q_k \in \mathbb{R}^{n!\times n!}_{perm}.
\end{aligned}
$$

Example 18.9.26. Consider the following enumeration of the elements of the permutation matrices of size 3×3,

$$
\begin{aligned}
&\mathbb{R}^{3\times 3}_{perm} = \{Q_1,\, Q_2,\, \ldots,\, Q_6\}, \\
&Q_1 = I_3,\; Q_2 = W_3 = \begin{bmatrix} 0&0&1 \\ 1&0&0 \\ 0&1&0 \end{bmatrix},\; Q_3 = W_3^2 = \begin{bmatrix} 0&1&0 \\ 0&0&1 \\ 1&0&0 \end{bmatrix}, \\
&Q_4 = \begin{bmatrix} 1&0&0 \\ 0&0&1 \\ 0&1&0 \end{bmatrix},\; Q_5 = \begin{bmatrix} 0&0&1 \\ 0&1&0 \\ 1&0&0 \end{bmatrix},\; Q_6 = \begin{bmatrix} 0&1&0 \\ 1&0&0 \\ 0&0&1 \end{bmatrix}.
\end{aligned}
$$

Then the Latin-Square operator induced by the permutation matrices of size 3×3 has the representation with respect to the specified permutation covering of the form,

$$
L_6(x) = \begin{bmatrix} x_1 & x_3 & x_2 & x_4 & x_5 & x_6 \\ x_2 & x_1 & x_3 & x_6 & x_4 & x_5 \\ x_3 & x_2 & x_1 & x_5 & x_6 & x_4 \\ x_4 & x_6 & x_5 & x_1 & x_2 & x_3 \\ x_5 & x_4 & x_6 & x_3 & x_1 & x_2 \\ x_6 & x_5 & x_4 & x_2 & x_3 & x_1 \end{bmatrix} = \sum_{i=1}^{6} x_iQ_{6,i},\; L_6 : \mathbb{R}^6_+ \to \mathbb{R}^{6\times 6}_+,
$$

$$
\text{where } Q_{6,1} = I_6,\;\; Q_{6,2} = \begin{pmatrix} 0&0&1&0&0&0 \\ 1&0&0&0&0&0 \\ 0&1&0&0&0&0 \\ 0&0&0&0&1&0 \\ 0&0&0&0&0&1 \\ 0&0&0&1&0&0 \end{pmatrix}, \text{ etc.}
$$

Lemma 18.9.27. *[58, Prop. A.5]. Consider the map of a doubly-stochastic Latin Square, $L : \mathbb{R}^n_{st} \to \mathbb{R}^{n\times n}_{dst}$ for an integer $n \in \mathbb{Z}_+$. Consider three stochastic vectors $a, b, c \in \mathbb{R}^n_{st}$.*

If $a = L(b)c$ and a is a vector of order $n_{s+}(a) \in \mathbb{Z}_+$ then both b and c are vectors of orders at most $n_{s+}(a)$; equivalently,

$$a = L(b)c \;\Rightarrow\; n_{s+}(b) \le n_{s+}(a),\; n_{s+}(c) \le n_{s+}(a).$$

Proof. The assumptions that $b \in \mathbb{R}^n_{st}$ and that L is the map of a Latin Square in the doubly-stochastic matrices imply that $L(b) \in \mathbb{R}^{n\times n}_{dst}$. It follows from Theorem 18.2.12 and from $a = L(b)c$ that $a \prec c$ or, equivalently, that,

$$\sum_{i=1}^{k} a_{[i]} \le \sum_{j=1}^{k} c_{[j]},\; k = 1,\, 2, \ldots,\, n-1,\; 1 = \sum_{i=1}^{n} a_{[i]} = \sum_{j=1}^{n} c_{[j]};$$

$$n_{s+}(a) = m \;\Rightarrow\; 1 = \sum_{i=1}^{m} a_{[i]} \le \sum_{j=1}^{m} c_{[j]} \;\Rightarrow\; 1 = \sum_{j=1}^{m} c_{[j]},\; n_{s+}(c) \le m = n_{s+}(a).$$

Because $a = L(b)c$ there exists another Latin square map L_1 such that $a = L_1(c)b$. From the above then follows that $n_{s+}(b) \le n_{s+}(a)$. □

Multiplicative Factorizations

Recall from Theorem 18.5.3.(e) that the group of units of the doubly-stochastic matrices is the set of the permutation matrices. Recall further from Def. 18.9.1 that a doubly-stochastic matrix $A \in \mathbb{R}^{n\times n}_{dst}$ is called a *prime in the doubly-stochastic matrices* if the following two conditions both hold: (1) A is not a permutation matrix; and (2) if $A = B \times C$ where $B,\, C \in \mathbb{R}^{n\times n}_{dst}$ then either B or C is a permutation matrix.

Theorem 18.9.28. Characterization of primes in the doubly-stochastic matrices.

(a)The matrix $A \in \mathbb{R}^{n\times n}_{dst}$ for $n \ge 3$ is a prime in the doubly-stochastic matrices if and only if it is permutation equivalent to the following form,

$$\exists\, n_1,\, n_2 \in \mathbb{Z}_+,\; n_1 + n_2 = n,\;\; \exists\, Q_1,\, Q_2 \in \mathbb{R}^{n\times n}_{perm},$$
$$\exists\, A_1 \in \mathbb{R}^{n_1\times n_1}_{dst},\ \textit{which is indecomposible and}$$
$$\textit{a prime in the doubly-stochastic matrices, such that}$$
$$A = Q_1 \begin{pmatrix} A_1 & 0 \\ 0 & I_{n_2} \end{pmatrix} Q_2 = Q_1(A_1 \oplus I_{n_2})Q_2.$$

(b)If there exist two decompositions as in (a) of the form,

$$A = Q_1 \begin{pmatrix} A_1 & 0 \\ 0 & I_{n_2} \end{pmatrix} Q_2 = Q_4 \begin{pmatrix} A_2 & 0 \\ 0 & I_{n_2} \end{pmatrix} Q_4,$$

$n_1,\ n_2,\ n_3,\ n_4 \in \mathbb{N},\ n_1 + n_2 = n = n_3 + n_4,$

$\exists\, A_1 \in \mathbb{R}_{dst}^{n_1 \times n_1},\ A_2 \in \mathbb{R}_{dst}^{n_3 \times n_3}$, *both indecomposible and primes in the doubly-stochastic matrices,*

then $n_1 = n_3,\ n_2 = n_4$, *and* $\exists\, Q_5,\ Q_6 \in \mathbb{R}_{perm}^{n_1 \times n_1}$, *such that,*

$$A_1 = Q_5 A_2 Q_6.$$

According to Theorem 18.9.28 the multiplicative decomposition of a prime in the doubly-stochastic matrices consists of the direct sum of an identity matrix and a doubly-stochastic matrix which is both indecomposible and a prime in the doubly stochastic matrices. The appearance of the identity matrix is to be expected because in singular value decomposition there is also a diagonal matrix with on the diagonal strictly positive elements. Due to the definition of a doubly-stochastic matrix, the diagonal elements are units, thus one. The search for the multiplicative factorization and for a prime of the doubly-stochastic matrices should now be focused on indecomposible such matrices. This is described below.

The proof of the above theorem is based on the following lemma.

Lemma 18.9.29. *Consider an integer* $n \in \mathbb{Z}_+$ *and a doubly stochastic matrix* $A = \sum_{k=1}^{n!} a_k Q_k \in \mathbb{R}_{dst}^{n \times n}$ *with* $a \in \mathbb{R}_+^{n!}$. *Denote the set of all permutations of size* $n \times n$ *by* $\mathbb{R}_{perm}^{n \times n} = \{Q_k \in \mathbb{R}_{perm}^{n \times n},\ \forall\ k \in \mathbb{Z}_{n!}\}$. *Recall the definition and notation of the Latin Square in the doubly-stochastic matrices* L_m *induced by the multiplication of permutation matrices, Def. 18.9.25.*

(a)Assume that the relation between the matrix $A \in \mathbb{R}_{dst}^{n \times n}$ *and the vector* a *by* $A = \sum_{k=1}^{n!} a_k Q_k$ *is a bijection and that* a *is at least of order 2.*
If there do not exist vectors $b,\ c \in \mathbb{R}_{st}^{n!}$ *such that* $a = L_m(b)\ c$, *and* b *and* c *are both of order at least two, then* A *is a prime in the doubly-stochastic matrices.*

(b)If $A \in \mathbb{R}_{dst}^{n \times n}$ *is a prime in the doubly-stochastic matrices then there do not exist vectors* $b,\ c \in \mathbb{R}_{st}^{n!}$ *both of which are at least of order two such that* $a = L_m(b)\ c$.

Proof. (a) Suppose that A is not a prime in the doubly-stochastic matrices. Because a is at least of order 2, A is not a permutation. That A is not a prime in the doubly-stochastic matrices implies by definition of such a prime that there exists a factorization $A = B \times C$ where $B,\ C \in \mathbb{R}_{dst}^{n \times n}$ and neither B nor C is a permutation. Because B and C are doubly-stochastic matrices, there exist vectors $b,\ c \in \mathbb{R}_{st}^{n!}$ such that $B = \sum_{i=1}^{n!} b_i Q_i$ and $C = \sum_{i=1}^{n!} c_i Q_i$. Because neither B nor C is a permutation, both b and c are at least of order 2. Note that,

$$\sum_{k=1}^{n} a_k Q_k = A = B \times C = (\sum_{i=1}^{n} b_i Q_i)(\sum_{j=1}^{n} c_j Q_j) = \sum_{k=1}^{n}\ [\sum_{j=1}^{n} L_m(b)_{kj} c_j]\ Q_k.$$

Because the relation of A and a is assumed to be a bijection, it follows that,

$$a_k = \sum_{j=1}^{n!} L_m(b)_{k,j}\ c_j,\ \forall\ k \in \mathbb{Z}_{n!}\ \Leftrightarrow\ a = L_m(b) c.$$

Thus there exists b and c of order at least 2 such that $a = L_m(b)\ c$. This is a contradiction of the assumption that such vectors do not exist.
(b) Suppose that there do exist vectors $b,\ c \in \mathbb{R}_{st}^{n!}$ each of which is at least of order 2 such that $a = L_m(b)\ c$. Define $B = \sum_{i=1}^{n!}\ b_iQ_i$ and $C = \sum_{j=1}^{n!}\ c_jQ_j$. Then,

$$B \times C = (\sum_{i=1}^{n}\ b_iQ_i)(\sum_{j=1}^{n}\ c_jQ_j) = \sum_{i=1}^{n}\ \sum_{j=1}^{n}\ b_ic_jQ_iQ_j = \sum_{k=1}^{n}\ L_m(b)_{kj}\ c_jQ_k$$
$$= \sum_{k=1}^{n}\ a_kQ_k = A.$$

Because $b,\ c$ are of order at least two, neither B nor C is a permutation. This and $A = B \times C$ imply by definition that A is not a prime in the doubly-stochastic matrices. This is a contradiction of the assumption.

□

Because of Lemma 18.9.29, the search for a prime in the doubly-stochastic matrices has to be focused on the factorization of a convex sum of permutations. The solution procedure to be used is: (1) Determine a subset of permutations for the factorization. (2) Solve the equation $a = L(b)\ c$.

Theorem 18.9.30. Characterization of primes in the order-two doubly-stochastic matrices. *Consider the doubly-stochastic matrix* $A = \sum_{i=1}^{n!}\ a_i\ Q_i \in \mathbb{R}_{dst}^{n \times n}$ *for a vector* $a \in \mathbb{R}_{st}^{n!}$ *with* $n \geq 3$ *and* $n_{s+}(a) = 2$.

Then A is a prime in the doubly-stochastic matrices and indecomposible if and only if,

$$\exists\ s \in (0,1),\ \exists\ Q_1,\ Q_2 \in \mathbb{R}_{perm}^{n \times n},\ \textit{such that},$$
$$A\ = Q_1[sI + (1-s)W_n]Q_2 = Q_1 \begin{pmatrix} s & 0 & 0 \ldots 0 & 1-s \\ 1-s & s & 0 \ldots 0 & 0 \\ 0 & 1-s & s\ \ldots 0 & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & 0 \ldots s & 0 \\ 0 & 0 & 0 \ldots 1-s & s \end{pmatrix} Q_2.$$

The matrix $sI + (1-s)W_n$ *belongs to the circulant doubly-stochastic matrices.*

As an aside, there exist primes in the doubly-stochastic matrices which are indecomposible, and yet do not belong to the circulant doubly-stochastic matrices.

Proof. (⇒) Let $A = \sum\ a_iQ_i \in \mathbb{R}_{dst}^{n \times n}$ be a prime in the doubly-stochastic matrices with $a \in \mathbb{R}_{st}^{n!}$ of order 2. By pre- and postmultiplication of permutation matrices there exists an integer $j \in \mathbb{Z}_{n!}$ such that $Q_1AQ_2 = a_1I + a_jQ_j$. From the fact that A is a prime and from Lemma 18.9.29.(b) follows that there do not vectors $b,\ c \in \mathbb{R}_{st}^{n!}$ both of order two such that $a = L_{n!}(b)c$.

From Theorem 18.9.28 follows that the matrix A is permutation equivalent to a matrix of the form $A_1 \oplus I$. Because by assumption the matrix A is indecomposible, the identity part of $A_1 \oplus I$ cannot be present. From [10, Section 4.2] follows that

$a_1 I + a_j Q_j$ is indecomposible if and only if Q_j is irreducible. But Q_j is irreducible if and only if it is permutation equivalent to the unit-shift matrix, $\exists\ Q \in \mathbb{R}_{perm}^{n\times n}$ such that $QQ_jQ^T = W_n$. Hence A is permutation equivalent to $a_1 I + a_j Q_j$.
($\Leftarrow$) Consider $A = sI + (1-s)W_n = \sum_{i=1}^{n!} a_i Q_j$. From [58, Lemma C.13] and the property that $W_n \neq W_n^T$ follows that there do not exist $b,\ c \in \mathbb{R}_{st}^{n!}$ both of order at least two such that $a = L_m(b)c$. The map $(a_1 I + a_2 W_n) \mapsto a \in \mathbb{R}_{st}^{n!}$ is a bijection. From Lemma 18.9.29.(a) follows that A is a prime in the doubly-stochastic matrices. □

There follows a result which illustrates how to construct primes in the doubly-stochastic matrices.

Theorem 18.9.31. *Consider the set of doubly-stochastic matrices in $\mathbb{R}_{dst}^{3\times 3}$. It follows from Theorem 18.9.21.(a) that any such matrix admits a decomposition as a convex sum of permutation matrices of the form,*

$$A = \sum_{i=1}^{6} b_i Q_i \in \mathbb{R}_{dst}^{3\times 3},\ \ \text{Cover}(\mathbb{R}_{+}^{3\times 3}, \mathbb{R}_{perm}^{3\times 3}),\ b \in \mathbb{R}_{st}^{n}.$$

In example 18.9.26 there is an enumeration of all permutations of $\mathbb{R}_{perm}^{3\times 3}$ which notation is used below.

The doubly-stochastic matrix $A \in \mathbb{R}_{dst}^{3\times 3}$ is indecomposible and a prime in the doubly-stochastic matrices if and only if the matrix A is permutation equivalent to the form,

$$\exists\ a \in \mathbb{R}_{st}^{n},\ n_{s+} = 3,\ i(a) = \{1,2,5\},\ \exists\ Q_i,\ Q_j \in \mathbb{R}_{perm}^{3\times 3},\ \textit{such that,}$$

$$\begin{aligned} A &= Q_i(\sum_{k=1}^{6} a_k Q_k)Q_j = Q_i[a_1 I + a_2 Q_2 + a_5 Q_5]Q_j = \\ &= Q_i \begin{pmatrix} a_1 & 0 & a_2 + a_5 \\ a_2 & a_1 + a_5 & 0 \\ a_5 & a_2 & a_1 \end{pmatrix} Q_j \in \mathbb{R}_{dst}^{3\times 3}. \end{aligned}$$

The classification of primes in the doubly-stochastic matrices which are in addition indecomposible is now an investigation into the solvability of a linear equation in terms of a Latin-Square in the doubly-stochastic matrices. Details on how to proceed are stated in the reference paper.

18.9.6 Positive Matrices

In this section and the following one, the focus is on the formulation of theory for the positive matrix factorization Problem 18.1.4.

Multiplicative Factorizations

Theorem 18.9.32. Characterization of primes in the positive matrices. *Consider a positive matrix $A \in \mathbb{R}_{+}^{n\times n}$.*

(a)The positive matrix A is a prime in the positive matrices if and only if,

$$\exists\, n_1,\, n_2 \in \mathbb{N},\, n_1 \geq 2,\, n_1+n_2 = n;\ \ \exists\, M_1,\, M_2 \in \mathbb{R}^{n\times n}_{mon},$$

$\exists\, B \in \mathbb{R}^{n\times n}_{dst}$, *which is a doubly-stochastic matrix, indecomposible, and a prime in the positive matrices such that,*

$$A = M_1 \begin{pmatrix} B\, 0 \\ 0\ I_{n_2} \end{pmatrix} M_2. \tag{18.11}$$

(b)If A admits two factorizations as described in (a) with the notation,

$$A = M_1 \begin{pmatrix} B_1\ 0 \\ 0\ \ I_{n_2} \end{pmatrix} M_2 = M_3 \begin{pmatrix} B_2\ 0 \\ 0\ \ I_{n_4} \end{pmatrix} M_4,$$

$$n_1,\, n_2,\, n_3,\, n_4 \in \mathbb{N},\, n_1 \geq 2, n_3 \geq 2,\, n = n_1+n_2 = n_3+n_4,$$

$$B_1 \in \mathbb{R}^{n_1\times n_1}_{dst},\, B_2 \in \mathbb{R}^{n_3\times n_3}_{dst},$$

both fully indecomposible and primes in the positive matrices;

then $n_1 = n_3$, $\exists\, Q_1,\, Q_2 \in \mathbb{R}^{n\times n_1}_{perm}$, *such that,* $B_1 = Q_1 B_2 Q_2$.

An interpretation of the above theorem follows. Theorem 18.9.32 is for positive matrices the analogon of the singular value decomposition of square real-valued matrices in $\mathbb{R}^{n\times n}$, see Theorem 17.4.4. The singular value decomposition has the form,

$$H = USV^T \in \mathbb{R}^{k\times m},\ \ U \in \mathbb{R}^{k\times k}_{ortg},\, V \in \mathbb{R}^{m\times m}_{ortg},\, D \in \mathbb{R}^{k\times m}_{diag},\, S = \begin{pmatrix} D\, 0 \\ 0\ \ 0 \end{pmatrix}.$$

For real-valued matrices, the role of the units is played by the orthogonal matrices and the nucleus of the decomposition is the matrix of singular values.

Correspondingly, a square positive matrix admits a decomposition with the units being the monomial matrices and as nucleus a block-diagonal matrix with as second block an identity matrix and as first block an indecomposible doubly-stochastic matrix which is a prime in the positive matrices. That the second block is a diagonal matrix is due to the scaling with a monomial matrix hence the second block corresponds to what one expects based on the singular value decomposition. Of interest is now the first diagonal block and why this doubly stochastic matrix cannot be multiplicatively factorized except by the units of the positive matrices.

To carry further the investigation of the primes in the positive matrices it follows from Theorem 18.9.32 that one has to investigate the set of indecomposible doubly-stochastic matrices which are primes in the positive matrices. But by Theorem 18.9.21, any doubly stochastic matrix is a convex sum of permutation matrices. This result allows a further analysis.

There follows a subset of the primes in the positive matrices.

Proposition 18.9.33. Characterization of primes in the order-two positive matrices. *Consider a doubly-stochastic matrix with representation,*

$$A = \sum_{i=1}^{n!} a_i\, Q_i \in \mathbb{R}^{n\times n}_{dst},\ \ n \in \mathbb{Z}_+,\, n \geq 3,\, a \in \mathbb{R}^n_{st},\, n_{s+}(a) = 2.$$

Then A is an indecomposible matrix and a prime in the positive matrices if and only if,

$$
\begin{aligned}
&\exists\, s \in (0,1),\ \exists\, M_1,\ M_2 \in \mathbb{R}^{n\times n}_{mon},\ \textit{such that,}\\
A &= M_1[sI_n + (1-s)W_n]M_2 = M_1
\begin{pmatrix}
s & 0 & 0 & \ldots 0 & 1-s\\
1-s & s & 0 & \ldots 0 & 0\\
0 & 1-s & s & \ldots 0 & 0\\
0 & 0 & 1-s & 0 & 0\\
\vdots & & \ddots & \ddots & \vdots\\
0 & 0 & 0 & \ldots s & 0\\
0 & 0 & 0 & \ldots 1-s & s
\end{pmatrix}
M_2,\\
&= M_1BM_2.
\end{aligned}
\tag{18.12}
$$

Proof. ($\Leftarrow$) From [3, Th. 2.6] follows that the matrix of equation (18.12) is a prime in the positive matrices. From [10, Section 4.2] follows that that matrix is indecomposible.
($\Rightarrow$) The matrix A is doubly stochastic. From Theorem 18.9.30 follows that A is a prime in the doubly-stochastic matrices with decomposition $A = M_1BM_2$. Thus the matrix A is permutation equivalent to the matrix B displayed in equation (18.12) hence A is monomially equivalent to that matrix. □

Positive Matrices – Extremal Cones

The problem of this section was motivated and described in Section 18.1. An example is shown.

Example 18.9.34. Consider the following doubly-stochastic matrix.

$$
A = \frac{1}{2}\begin{pmatrix} 1\ 0\ 0\ 1\\ 1\ 1\ 0\ 0\\ 0\ 1\ 1\ 0\\ 0\ 0\ 1\ 1 \end{pmatrix} \in \mathbb{R}^{4\times 4}_{dst};\ \text{ then } \operatorname{rank}(A) = 3 < 4 = \operatorname{pos}-\operatorname{rank}(A).
$$

From Theorem 18.9.30 follows that the doubly-stochastic matrix A is a prime in the doubly-stochastic matrices and from Theorem 18.9.33 that it is a prime in the positive matrices. From Proposition 18.9.40 and Proposition 18.9.41 then follows that $\operatorname{pos}-\operatorname{rank}(A) = 4$ equals the size of the matrix.

From this example follows that a different approach to the concept of positive rank is required than the approach of the rank of a real-valued matrix.

Below the concept of a prime in the positive matrices, defined only for square matrices, is generalized to nonsquare matrices. The new concept is strict factorizability of a nonsquare positive matrix. The equivalent geometric concept is that of an extremal cone. A relation is then formulated between an extremal cone with a non-strictly factorizable matrix of positive rank.

Definition 18.9.35. The nonsquare positive matrix $A \in \mathbb{R}_+^{k \times m}$ is called *strictly factorizable* if there exists a positive-matrix factorization of the form,

$$A = B \times C,$$

with $n \in \mathbb{Z}_+$, $n \leq \min\{k,m\}$, $B \in \mathbb{R}_+^{k \times n}$, $C \in \mathbb{R}_+^{n \times m}$, and neither B nor C is part of a monomial, see Def. 18.3.1.

It is called *not strictly factorizable* otherwise; equivalently, if there exists a factorization $A = B\, C$ with $n \leq \min\{k,m\}$ then either B or C is part of a monomial.

Definition 18.9.36. Consider integers $k,\ m \in \mathbb{Z}_+$ with $k \geq m$. A cone in $C_{k,m}$, see Def. 18.4.7, is said to be an *extremal cone* if it is a maximal element in $C_{k,m} \backslash \mathbb{F}_m(\mathbb{R}_+^k)$ with respect to the inclusion relation on the set of cones. Denote the set of extremal cones in this space by,

$$CE_{k,m} = \{C \in C_{k,m} \backslash \mathbb{F}_m(\mathbb{R}_+^k) |\ C \text{ is an extremal cone}\}.$$

The only maximal element in $C_{k,k}$ is $\mathbb{R}_+^k$.

Proposition 18.9.37. *Consider the integers $k,\ m \in \mathbb{Z}_+$ with $k \geq m$, and $A \in \mathbb{R}_+^{k \times m}$.*

Then $\text{cone}(A)$ *is a m-face of $\mathbb{R}_+^k$, see Def. 18.4.9, if and only if A is part of a monomial.*

Example 18.9.38. Please note the 3-face $F_{3,1}$ and the 2-face $F_{2,1}$.

$$A_{3,1} = \begin{pmatrix} 0\,0\,0 \\ 1\,0\,0 \\ 0\,0\,1 \\ 0\,1\,0 \end{pmatrix} \in \mathbb{R}_+^{4 \times 3},\ \ F_{3,1} = \text{cone}(A_{3,1}),$$

$$A_{2,1} = \begin{pmatrix} 0\,0 \\ 1\,0 \\ 0\,0 \\ 0\,1 \end{pmatrix} \in \mathbb{R}_+^{4 \times 2},\ \ F_{2,1} = \text{cone}(A_{2,1}).$$

Proposition 18.9.39. Characterization of an extremal cone in terms of its matrix properties. *Consider integers $k,\ m \in \mathbb{Z}_+$ and a positive matrix $A \in \mathbb{R}_+^{k \times m}$.*

Then $\text{cone}(A) \in CE_{k,m}$ *if and only if (1) A is not strictly factorizable and (2) A is not part of a monomial.*

Proposition 18.9.40. Characterization of a particular extremal cone as generated by a prime matrix. *Consider an integer $k \in \mathbb{Z}_+$ and a square positive matrix $A \in \mathbb{R}_+^{k \times k}$.*

Then $\text{cone}(A) \in CE_{k,k}$ *if and only if A is a prime in the positive matrices.*

At this time there is no complete characterization of extremal cones in $CE_{k,m}$ with $k > m$. But see the subsection *Embedding in an extremal cone* below for a construction procedure of extremal cones.

The following results describe sufficient and necessary conditions for a positive matrix to have a particular positive rank.

Proposition 18.9.41. Sufficient condition for the positive rank. *Consider the integers $k,\ m \in \mathbb{R}_+$ with $k \geq m$ and a positive matrix $A \in \mathbb{R}_+^{k \times m}$.*

If $\text{cone}(A) \in CE_{k,m} \cup \mathbb{F}_m(\mathbb{R}_+^k)$*, thus* $\text{cone}(A)$ *is either an extremal cone or a face of the positive orthant, then* $\text{pos} - \text{rank}(A) = m$*.*

Theorem 18.9.42. Characterization of minimal positive matrix factorizations.
Consider the integers $k,\ m \in \mathbb{R}_+$ and a positive matrix $A \in \mathbb{R}_+^{k \times m}$.

If A is not strictly factorizable then $\text{pos} - \text{rank}(A) = \min\{k, m\}$*.*

Assume that $k \geq m$ without loss of generality. Any minimal positive-matrix factorization is given by one of the following two decompositions:

1. *$A = M \times C$ with $M \in \mathbb{R}_+^{k \times m}$ and $C \in \mathbb{R}_+^{m \times m}$, in which M is part of a monomial. Moreover, C is a prime in the positive matrices if and only if A is not part of a monomial, otherwise C is a monomial. A necessary condition for this decomposition is that A contains exactly $k - m$ zero rows.*
2. *$A = B \times N$ with $B \in \mathbb{R}_+^{k \times m}$ and $N \in \mathbb{R}_+^{m \times m}$, in which N is a monomial. Moreover,* $\text{cone}(B) \in CE_{k,m}$ *if and only if A is not part of a monomial, otherwise* $\text{cone}(B)$ *is a m-facet of $\mathbb{R}_+^k$.*

In both cases, $\text{cone}(A) \in CE_{k,m}$ *if and only if A is not part of a monomial. Otherwise* $\text{cone}(A)$ *is an m-facet of $\mathbb{R}_+^k$.*

The conclusion of the above theorem is that for the determination of positive matrix factorizations one has to determine the extremal cones in the positive orthant. This issue is investigated in more generality in the next subsection.

Embedding in an Extremal Cone

What to do if a positive matrix $A \in \mathbb{R}_+^{k \times m}$ is strictly factorizable? Below it is shown that any polyhedral cone can either be embedded in or is already equal to, either an extremal cone or a facet of the positive orthant.

Recall the notation from Def. 18.2.4 that for a vector $a \in \mathbb{R}_+^n$, $n_{s+}(a) \in \mathbb{N}_n$ denotes the number of elements of a which are strictly positive and that the set
$i_{s+}(a) = \{j \in \mathbb{Z}_n | a_j > 0\}$ denotes the index set of the strictly positive components of the vector a.

Denote by $n_{s+}(D_{.,j}) \in \mathbb{N}$ the number of strictly positive elements of the j-th colum vector of the matrix D for $j \in \mathbb{Z}_n$. Denote, for a positive matrix $A \in \mathbb{R}_+^{k \times m}$, $n_{s+}(A) = \sum_{j=1}^{m} n_{s+}(A_{.j})$ the number of nonzero elements of the matrix A. There is a slight abuse of notation here because now $n_{s+}(a)$ and $n_{s+}(A)$ are both defined. Note that in this chapter use is made of the symbol $n_{s+}(A)$ denoting the number of strictly positive elements while in the paper [71] notation is used for the number of zero elements of a positive matrix.

Proposition 18.9.43. *Consider a positive matrix $B \in \mathbb{R}_+^{k \times n}$ which satisfies $n =$* $\text{pos} - \text{rank}(B)$*. Consider the minimal positive rank factorization $B = D \times E$ with*

$D \in \mathbb{R}_+^{k \times n}$ and $E \in \mathbb{R}_+^{n \times n}$. Then there exists a permutation map $\sigma : \mathbb{Z}_n \to \mathbb{Z}_n$ such that,

$$\forall\, i \in \mathbb{Z}_n,\ \ n_{s+}(D_{.,i}) \leq n_{s+}(B_{.,\sigma(i)}).$$

Proposition 18.9.44. Embedding in a polyhedral cone*, Consider integers $k,\ n \in \mathbb{Z}_+$ such that $k \geq n$ and a positive matrix $B \in \mathbb{R}_+^{k \times n}$. Assume that B is strictly factorizable and that* $n = \text{pos} - \text{rank}(B)$.

Then there exist positive matrices $B_2 \in \mathbb{R}_+^{k \times n}$ and $C_2 \in \mathbb{R}_+^{n \times n}$ such that $B = B_2 \times C_2$ and $n_{s+}(B_2) < n_{s+}(B)$. Consequently, $\text{cone}(B) \subset \text{cone}(B_2)$ *hence* $\text{cone}(B)$ *is embedded in* $\text{cone}(B_2)$.

The proof of the above proposition in reference [71] contains a procedure to compute the matrices B_2 and C_2.

Proposition 18.9.45. Relation of positive rank of a matrix with the positive rank of its minimal factor matrices. *Consider a positive matrix $A \in \mathbb{R}_+^{k \times m}$ and assume that* $\text{pos} - \text{rank}(A) = n \in \mathbb{Z}_+$. *If $A = S \times R$ with $S \in \mathbb{R}_+^{k \times n}$ and $R \in \mathbb{R}_+^{n \times m}$ then* $\text{pos} - \text{rank}(S) = n = \text{pos} - \text{rank}(R)$.

Theorem 18.9.46. Embedding of a cone in an extremal cone or a facet. *Consider a positive matrix $A \in \mathbb{R}_+^{k \times m}$.*

If $\text{pos} - \text{rank}(A) = n \in \mathbb{Z}_+$ *then there exist positive matrices $B \in \mathbb{R}_+^{k \times n}$ and $F \in \mathbb{R}_+^{n \times m}$ such that $A = B \times F$ and* $\text{cone}(B) \in CE_{k,n} \cup F_n(\mathbb{R}_+^k)$.

18.10 Computations

The number of computations for positive matrices and stochastic matrices is enormous. No list will be formulated, almost every section of this chapter has procedures which can be programmed.

Of interest to theory of computations for stochastic matrices is the issue that a stochastic vector has the number one for all elements of the positive vector. After any computation, the sum will no longer be one but differ from one. Useful is possibly a simple convergence procedure to bring the length of the stochastic vector which is the result of the computation as close one as preferred.

18.11 Further Reading

History. O. Perron and G.F. Frobenius (1849–1917), [57] and [25, 26, 27], have considerably advanced the early research on positive matrices.

Major contributors to the theory of positive matrices in the latter part of the 20th century include H. Schneider and R.A. Brualdi. The current theory of positive matrices is primarily due to: G. Birkhoff, R.A. Brualdi, M. Marcus, A.W. Marshall and I. Olkin, L. Mirsky, H. Schneider, and R. Sinkhorn.

Positive real numbers and positive vector spaces. The majorization order is defined in [35]. See also the book [49].

Positive matrices. The literature on positive matrices is very large. In this chapter concepts and results of positive are summarized which are useful in the body of the book, in particular for the decomposition of state-finite stochastic systems and for stochastic realization of such systems.

Books on positive matrices. A major reference is the book of A. Berman and R.J. Plemmons, [3], which has been republished as [4]. Another basic resource is the book of F.R. Gantmacher, [29, Ch. XIII]. Other books on positive matrices are [1, 46, 50, 64, 65] For circulant matrices see [16]. For latin squares see [10, Chapter 8] and [17]. On the algebraic theory of positive matrices see below but also [14, 24].

Books not devoted exclusively to positive matrices but relevant for such matrices include [10] for the relation of matrices and graphs, and [49] for stochastic matrices and doubly stochastic matrices.

Books on control and system theory of positive linear and nonlinear systems which include text on positive matrices are, [2, 21, 33, 44].

Positive matrix factorization problem is formulated in [48] though it could have been formulated earlier. Other papers are [11, 12, 13, 15, 34, 39, 42, 69, 71]. An approximation procedure is proposed in [23].

Stochastic vectors. The concept of the decreasing arrangement and the majorization order are adapted from [49, p. xx; Def. A.1], while they are due to [35]. Theorem 18.2.12 is adapted from [49, Th. 2.B.2 and 4.B.1]. An upperbound on the index of imprimitivity is provided in [3, Th. 2.4.14] while in [3, Ex. 2.4.15] it is proven that the bound of the above theorem is attained for a subset of the square positive matrices.

Geometry and graphs. The concept of faces of a polyhedral cone or a polyhedral set is adapted from M. Gerstenhaber, [30], and R.T. Rockafellar, [60]. Graphs and matrices are described in the book [31]. For polyhedral cones, the most expensive computations are the transformation from an implicit representation to an explicit representation and conversely, [28, 72].

Similarity. See the book [3] for permutation similarity to irreducible and completely reduced matrices. Similarity is described for finite-state Markov processes in [22, 40]. Similarity of a square irreducible positive matrix to an irreducible doubly-stochastic matrix by diagonal matrices, is treated in [19]. Similarity of doubly-stochastic matrix to an irreducible such matrix or to a reduced such matrix is established in [63, Lemma 1], see Theorem 18.6.5. The book of R.A. Brualdi and H.J. Ryser, [10, pp. 96-102], provides a procedure to transform a reducible stochastic matrix to the Frobenius canonical form. Generalizations of irreducible matrices to nonlinear operators include [54].

Eigenvalues and eigenvectors. Theorem 18.7.1 is a generalization of the famous Perron-Frobenius theorem which is stated and proven in [3, Thm. 2.1.3, Thm. 2.1.4, Thm, 2.2.20, Thm. 2.2.33], [1, Thm. 1.4.4], and [29, Thm. XIII.2.1]. Proposition 18.7.3 is related to [3, Th. 2.2.35] though the proof is adjusted at a particular point. Theorem 18.8.2.(b) is stated in [3, Th. 2.5.3 and Th. 8.3.11]. Theorem 18.8.5 is a standard equivalence result for irreducible and nonperiodic matrices, a proof is provided in [1, Thm. 1.8.2]. Theorem 18.8.12.(b) is stated in [3, Lemma 8.3.20]. Theorem 18.8.17 is adjusted from [1, Thm. 1.9.7]. Theorem 18.8.17.(c) is proven in [32, Thm. 11.11].

Multiplicative factorization of a positive matrix. The concept of a prime in the positive matrices was proposed and investigated by D.J. Richman and H. Schneider, [59]. The examples which follow the definition are stated in [59, 3].

Equivalence. See for a characterization of irreducible and of fully indecomposible matrices the paper of H. Schneider with a history of the subject, [61]. Diagonal equivalence of a positive matrix by a diagonal matrix with strictly-positive diagonal, Theorem 18.9.7, may be found in (a) [9, 66]; (b) [9, 66]; and (c) [55, 68]. The result quoted above is based on the study of doubly stochastic matrices for which sources are [51, 56]. The papers on transformations start with the transformation to a stochastic matrix by [9, 66]. Papers on the transformation to a doubly stochastic matrix are also, [43, 55, 67, 68]. The computational issues of scaling to a doubly stochastic form are discussed in [55].

Diagonal equivalence of a fully indecomposible positive matrix with a doubly-stochastic matrix. The result is due to R. Sinkhorn, [66]. A different approach to the transformation to a stochastic matrix is [9]. Related papers on the diagonal equivalence of a positive matrix to a doubly stochastic matrix are, [36, 47, 62, 43, 62, 67, 68]. References on the computational issues of this equivalence include [11, 55].

Circulant doubly-stochastic matrices. The research on primes in the circulant doubly-stochastic matrices was primarily developed by J.M. van den Hof, see [58, 70].

Doubly-stochastic matrices. See the books [49], [3], and [1, Ch. 2]. See also the papers [7, 8, 20, 45, 51, 52, 56, 63]. See for bounds on the eigenvalues of doubly stochastic matrices [18].

Infinite doubly-stochastic matrices are discussed in [37, 41, 56].

The result of G. Birkhoff, [5], that any doubly stochastic matrix admits an additive decompositino as a convex sum of permutation matrices has been the basis for much further research. See the paper of R. Brualdi, [7] for the issues of nonuniqueness, the set of permutations involved, the faces of the polyhedron, etc. Theorem 18.9.21 is a generalization of [3, Th. 2.5.6], [47], and [3, Th. 2.5.7].

The results on primes in the doubly-stochastic matrices are adjusted from [58].

Primes in the positive matrices. All results are adapted from the paper [58]. The results on extremal cones generated by positive matrices are related to [71, Prop. 3.7] and [70, Prop. 3.15].

References

1. R.B. Bapat and T.E.S. Raghavan. *Nonnegative matrices and applications*. Encyclopedia of mathematics and its applications. Cambridge University Press, Cambridge, 1997. 697, 698
2. A. Berman, M. Neumann, and R.J. Stern. *Nonnegative matrices in dynamic systems*. John Wiley & Sons, New York, 1989. 169, 697

3. A. Berman and R.J. Plemmons. *Nonnegative matrices in the mathematical sciences*. Academic Press, New York, 1979. 646, 652, 656, 681, 693, 697, 698
4. A. Berman and R.J. Plemmons. *Nonnegative matrices in the mathematical sciences*. Number 9 in Classics in Applied Mathematics. SIAM, Philadelphia, 1993. 697
5. G. Birkhoff. Tres observaciones sobre el algebra lineal. *Univ Nac. Tucuman Rev. Ser. A*, 5:147–150, 1946. 698
6. A. Browder. *Mathematical analysis - An introduction*. Undergraduate texts in mathematics. Springer-Verlag, New York, 1996. 30, 49, 424, 426, 475, 526, 635, 636, 677, 815
7. R.A. Brualdi. Notes on the Birkhoff theorem for doubly stochastic matrices. *Can. Math. Bull.*, 25:191–199, 1982. 698
8. R.A. Brualdi. Some applications of doubly stochastic matrices. *Linear Algebra & its Applications*, 107:77–89, 1988. 698
9. R.A. Brualdi, S.V. Parter, and H. Schneider. The diagonal equivalence of a nonnegative matrix to a stochastic matrix. *J. Math. Anal. Appl.*, 16:31–50, 1966. 698
10. R.A. Brualdi and H.J. Ryser. *Combinatorial matrix theory*. Cambridge University Press, Cambridge, 1991. 690, 693, 697
11. S.L. Campbell and G.D. Poole. Computing nonnegative rank factorizations. *Linear Algebra Appl.*, 35:175–182, 1981. 697, 698
12. M. Catral, L. Han, M. Neumann, and R.J. Plemmons. On reduced rank nonnegative matrix factorizations of symmetric nonnegative matrices. *Linear Algebra & its Applications*, 393:107–126, 2004. 697
13. C.J. Cheng. The non-negative rank factorizations of nonnegative matrices. *Linear Algebra & its Applications*, 62:207–217, 1984. 697
14. A.H. Clifford and G.B. Preston. *The algebraic theory of semigroups - Vol. 1*. Number 7 in Math. Surveys. Amer. Math. Soc., Providence, 1961. 697
15. J.E. Cohen and U.G. Rothblum. Nonnegative ranks, decompositions, and factorizations of nonnegative matrices. *Linear Algebra & its Applications*, 190:149–168, 1993. 697
16. P.J. Davis. *Circulant matrices*. J. Wiley, New York, 1979. 697
17. J. Dénes and A.D. Keedwell. *Latin squares: New developments in the theory and applications*. North-Holland, Amsterdam, 1991. 697
18. P. Diaconis and D. Stroock. Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Probab.*, 1:36–61, 1991. 169, 698
19. B. Curtis Eaves, Alan J. Hoffman, Uri G. Rothblum, and Hans Schneider. Line-sum-symmetric scalings of square nonnegative matrices. *Math. Programming Stud.*, 25:124–141, 1985. 697
20. H.K. Farahat. The semigroup of doubly-stochastic matrices. *Proc. Glasgow Math. Assoc.*, 7:178–183, 1966. 698
21. L. Farina and S. Rinaldi. *Positive linear systems: Theory and applications*. Pure and Applied Mathematics. John Wiley & Sons, New York, 2000. 697
22. W. Feller. *An introduction to probability theory and its applications – Volume 1*. Wiley, New York, 1968. 697
23. Lorenzo Finesso and Peter Spreij. Nonnegative matrix factorization and I-divergence alternating minimization. *Linear Algebra & its Applications*, 416:270–287, 2006. 697
24. P. Flor. On groups of nonnegative matrices. *Compositio Mathematica*, 21:376–382, 1969. 697
25. G. Frobenius. Über matrizen aus positiven elementen. *S.-B. Preuss. Akad. Wiss. (Berlin)*, pages 471–476, 1908. 696
26. G. Frobenius. Über matrizen aus positiven elementen ii. *S.-B. Preuss. Akad. Wiss. (Berlin)*, pages 514–518, 1909. 696
27. G. Frobenius. Über matrizen aus nicht negativen elementen. *S.-B. Preuss. Akad. Wiss. (Berlin)*, pages 456–477, 1912. 696
28. K. Fukuda. Frequently asked questions in polyhedral computation. Report http://www.ifor.math.ethz.ch/ fukuda/polyfaq/polyfaq.html, ETH, Zürich, 2000. 697
29. F.R. Gantmacher. *The theory of matrices, volume 1, 2*. Chelsea Publ. Co., New York, 1959. 626, 636, 697

30. M. Gerstenhaber. Theory of convex polyhedral cones. In Tj.C. Koopmans, editor, *Activity analysis of production and allocation*, pages 298–316. Wiley & Sons, New York, 1951. 697
31. M. Gondran and M. Minoux. *Graphs and algorithms*. John Wiley & Sons, Chichester, 1984. 697
32. C.M. Grinstead and J.L. Snell. *Introduction to probability, 2nd revised edition*. American Mathematical Society, Boston, 1997. 49, 698
33. Wassim M. Haddad, VijaySekhar Chellaboina, and Qing Hui. *Nonnegative and Compartmental Dynamical Systems*. Princeton University Press, Princeton, 2010. 169, 697
34. J. Hannah and T.J. Laffey. Nonnegative factorization of completely positive matrices. *Linear Algebra Appl.*, 55:1–9, 1983. 697
35. G.H. Hardy, J.E. Littlewood, and G. Polya. *Inqualities (2nd Ed.)*. Cambridge University Press, London and New York, 1952. 697
36. D. Hershkowitz, U.G. Rothblum, and H. Schneider. Classifications of nonnegative matrices using diagonal equivalence. *SIAM J. Matrix Anal. Appl.*, 9:455–460, 1988. 698
37. J.R. Isbell. Infinite doubly stochastic matrices. *Canadian Math. Bull.*, 5:1–4, 1962. 698
38. J.A. Jacquez. *Compartmental analysis in biology and medicine, 2nd Ed.* The University of Michigan Press, Ann Arbor, 1985. 659
39. M.W. Jeter and W.C. Pye. A note on non-negative rank factorizations. *Linear Algebra & its Applications*, 38:171–173, 1981. 697
40. J.G. Kemeny, J.L. Snell, and A.W. Knapp. *Denumerable Markov chains*. Springer-Verlag, Berlin, 1976. 73, 697
41. D.G. Kendall. On infinite doubly-stochastic matrices and Birkhoff's Problem 111. *J. London Math. Soc.*, 35:81–84, 1960. 698
42. D. Lee and H. Seung. Algorithms for non-negative matrix factorization. *NIPS*, 13:556–562, 2001. 697
43. D. London. On matrices with a doubly stochastic pattern. *J. Math. Anal. Appl.*, 34:648–652, 1971. 698
44. D.G. Luenberger. *Introduction to dynamic systems - Theory, models and applications*. J. Wiley & Sons, New York, 1979. 697
45. M. Marcus, K. Kidman, and M. Sandy. Products of elementary doubly stochastic matrices. *Linear and Multilinear Algebra*, 15:331–340, 1984. 698
46. M. Marcus and H. Minc. *A survey of matrix theory and matrix inequalities*. Allyn and Bacon, Boston, 1964. 697
47. M. Marcus and R. Ree. Diagonals of doubly stochastic matrices. *Quart. J. Math. Oxford, Ser. 2*, 10:296–302, 1959. 698
48. T.L. Markham. Factorizations of nonnegative matrices. *Proc. Amer. Math. Soc.*, 32:45–47, 1972. 697
49. A.W. Marshall and I. Olkin. *Inequalities: Theory of majorization and its applications*. Academic Press, New York, 1979. 697, 698
50. H. Minc. *Nonnegative matrices*. Wiley, New York, 1988. 697
51. L. Mirsky. Results and problems in the theory of doubly-stochastic. *Z. Wahrscheinlichkeitstheorie*, 1:319–334, 1962/1963. 698
52. J.S. Montague and R.J. Plemmons. Doubly stochastic matrix equations. *Israel J. Math.*, 15:216–229, 1973. 698
53. B. Noble. *Applied linear algebra*. Prentice-Hall, Englewood Cliffs, NJ, 1969. 636, 663
54. R.P. Nussbaum and S.M. Verduyn Lunel. *Generalizations of the Perron-Frobenius theorem for nonlinear maps*. Number 659 in Memoirs of the A.M.S. A.M.S., Providence, 1999. 697
55. B.N. Parlett and T.L. Landis. Methods for scaling to doubly stochastic form. *Linear Algebra Appl.*, 48:53–79, 1982. 698
56. H. Perfect and L. Mirsky. The distribution of positive elements in doubly stochastic matrices. *J. London Math. Soc.*, 40:689–698, 1965. 698
57. O. Perron. Zur theorie der über matrizen. *Math. Ann.*, 64:248–263, 1907. 696
58. G. Picci, J.M. van den Hof, and J.H. van Schuppen. Primes in several classes of the positive matrices. *Linear Algebra & its Applications*, 277:149–185, 1998. 688, 691, 698

59. D.J. Richman and H. Schneider. Primes in the semigroup of non-negative matrices. *Linear and Multilinear Algebra*, 2:135–140, 1974. 698

60. R.T. Rockafellar. *Convex analysis*. Princeton University Press, Princeton, 1970. 636, 697

61. H. Schneider. The concepts of irreducibility and full indecomposability of a matrix in the works of Frobenius, König, and Markov. *Linear Algebra & its Applications*, 18:139–162, 1977. 698

62. M.H. Schneider. Matrix scaling, entropy minimization, and conjugate duality 1. Existence conditions. *Linear Algebra & its Applications*, 114/115:785–813, 1989. 698

63. S. Schwarz. On the structure of the semigroup of stochastic matrices. *Publ. Math. Inst. Hung. Acad. Sci., Ser.A*, 9:297–311, 1964. 697, 698

64. E. Seneta. *Non-negative matrices*. Wiley, New York, 1973. 697

65. E. Seneta. *Nonnegative matrices and Markov chains*. Springer-Verlag, New York, 1981. 697

66. R. Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.*, 35:876–879, 1964. 698

67. R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. ii. *Proc. Amer. Math. Soc.*, 45:195–198, 1974. 698

68. R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.*, 21:343–348, 1967. 698

69. L.B. Thomas. Solution to problem 73-14: Rank factorization of nonsingular matrices. *SIAM Rev.*, 16:393–394, 1974. 697

70. J.M. van den Hof. *System theory and system identification of compartmental systems*. PhD thesis, University of Groningen, Groningen, 1996. 698

71. J.M. van den Hof and J.H. van Schuppen. Positive matrix factorization via extremal polyhedral cones. *Linear Algebra & its Applications*, 293:171–186, 1999. 695, 696, 697, 698

72. D.K. Wilde. A library for doing polyhedral operations. Publication Interne 785, IRISA, Rennes, 1993. 697

Chapter 19
Appendix C Probability

Abstract Concepts and results of probability theory are presented in this chapter which complement those of Chapter 2. Concepts covered in detail include: the canonical variable decomposition of a tuple of Gaussian random variables, a set of stable probability distribution functions, conditional probability, conditional Gaussian random variables, the conditional independence relation, a measure transformation, the P-essential infimum, and metrics on the set of probability measures.

Key words: Probability theory. Conditional independence. Measure Transformations.

The main topics of this chapter which are relevant for the body of the book are: the theory of the conditional independence relation, the concept of a measure transformation, and the concept of a P-essential infimum. Proofs are provided only if they are not readily available elsewhere or are relatively important for the theory of the book.

19.1 Sets and the Monotone Class Theorems

Let X,Y be sets. If $A \subseteq X$, $B \subseteq Y$ then the *rectangle* of A and B is defined by

$$A \times B = \{(a,b) \in X \times Y | a \in A, \ b \in B\}.$$

If (X,G) and (Y,H) are measurable spaces and $A \in G, \ B \in H$ then $A \times B$ is called a *measurable rectangle.* Let

$$F_0 = \left\{ \begin{array}{l} \cup_{k \in \mathbb{Z}_n} (A_k \times B_k) | \exists n \in \mathbb{Z}_+ \text{ such that} \{A_k \in G, k \in \mathbb{Z}_n\}, \\ \{B_k \in H, k \in \mathbb{Z}_n\}, \text{ and each of these families is disjoint} \end{array} \right\}.$$

It can be verified that F_0 is an algebra. Let $G \otimes H = F(F_0)$ be the σ-algebra generated by F_0. It is to be called the *product σ-algebra of* G and H. Then $(X \times Y, G \otimes H)$ is called the *measurable product space* of (X,G) and (Y,H).

The monotone class theorem is an important technical tool in probability theory. It may be used to prove a property of a σ-algebra on a set Ω by proving these properties for monotone families of subsets. In this subsection the statement of the theorem is presented.

Definition 19.1.1. A family of subsets M of a set Ω is called a *monotone class* if it satisfies:

(1) if $\{A_n \in M,\ n \in \mathbb{Z}_+\}$ is a countable increasing family then $(\cup_{n \in \mathbb{Z}_+} A_n) \in M$;
(2) if $\{A_n \in M,\ n \in \mathbb{Z}_+\}$ is a countable decreasing family then $(\cap_{n \in \mathbb{Z}_+} A_n) \in M$.

Note that a σ-algebra is by definition a monotone class.

Proposition 19.1.2. *Let G be a family of subsets of Ω not necessarily countable. Then there exists a smallest monotone class, denoted by $M(G)$, such that $G \subseteq M(G)$. It will be called* the monotone class generated by G.

Proof. Define the set,

$$M(G) = \{M_1 \subseteq \text{Pwrset}(\Omega) |\ G \subseteq M_1 \text{ and } M_1 \text{ is a monotone class}\},$$
$$M_s = \cap_{M \in M(G)} M.$$

From the definition of $M(G)$ and M_s then follows directly that $G \subseteq M_s$ and that M_s is a monotone class. From the definition of M_s follows that it is the smallest monotone class containing G. □

Theorem 19.1.3. *Consider a set Ω and a family of subsets of Ω, $F \subseteq Pwrset(\Omega)$. Then F is a σ-algebra if and only if (1) F is an algebra and (2) F is a monotone class.*

Proof. ($\Rightarrow$) Condition (1). A σ-algebra is also an algebra. Condition (2). Because F is a σ-algebra, it is closed with respect to a countable increasing sequence of sets. Consider a countable decreasing sequence of sets $\{A_n \in F,\ n \in \mathbb{Z}_+\}$. Then $\cap_{n \in \mathbb{Z}_+} A_n = (\cup_{n \in \mathbb{Z}_+} A_n^c)^c \in F$ because of the De Morgan laws, $A_n \in F \Rightarrow A_n^c \in F$, because of the closure of F with respect to complementation, because $\{A_n^c \in F,\ n \in \mathbb{R}_+\}$ is a countable increasing sequence, and because eF is a σ-algebra.
($\Leftarrow$) Note that $\Omega \in F$ because F is an algebra. Because F is an algebra, $A \in F$ implies that $A^c \in F$. Consider a countable sequence $\{A_n \in F, n \in \mathbb{Z}_+\}$. Construct the increasing sequence $\{B_n \in F, n \in \mathbb{Z}_+\}$ by the formulas $B_1 = A_1$ and for all $n \in \mathbb{Z}_+$, $B_n = \cup_{k=1}^n A_k \in F$, which is in F because F is an algebra. Then this sequence is increasing and $\cup_{n=1}^\infty A_n = \cup_{n=1}^\infty B_n \in F$ because the sequence $\{B_n,\ n \in \mathbb{Z}_+\}$ is increasing and because F is a monotone class. □

Theorem 19.1.4. *[The monotone class theorem] Let G be an algebra. Then $M(G) = F(G)$, or, equivalently, the monotone class generated by G equals the σ-algebra generated by G.*

Proof. (1) It will be proven that $M(G) \subseteq F(G)$. From Theorem 19.1.3 follows that $F(G)$ is a monotone class. By definition of $F(G)$ being the smallest σ-algebra containing G, $G \subseteq F(G)$. Thus $F(G)$ is a monotone class containing G. Then $M(G) \subseteq F(G)$ because $M(G)$ is the smallest monotone class containing G.
(2) It will be proven that $F(G) \subseteq M(G)$. The first step is to prove that $M(G)$ is an algebra. Because G is an algebra, $\Omega \in G \subseteq M(G)$.
(3) Define

$$M_1 = \{B \in M(G) |\ \forall A \in G,\ A \cap B \in M(G)\},$$
$$M_2 = \{C \in M(G) |\ \forall A \in M(G),\ A \cap C \in M(G)\}.$$

Because $G \subseteq M(G)$ and G is an algebra, $G \subseteq M_1 \subseteq M(G)$. Consider an increasing sequence $\{B_n \in M_1,\ \forall\ n \in \mathbb{Z}_+\}$, a decreasing sequence $\{C_n \in M_1,\ \forall\ n \in \mathbb{Z}_+\}$, and $A \in G$. Then

$$A \cap (\cup_{n=1}^{\infty} B_n) = \cup_{n=1}^{\infty} (A \cap B_n) \in M(G),$$
$$A \cap (\cap_{n=1}^{\infty} C_n) = \cap_{n=1}^{\infty} (A \cap C_n) \in M(G),$$

because $M(G)$ is a monotone class. Hence $\cup_{n=1}^{\infty} B_n$ and $\cap_{n=1}^{\infty} C_n \in M_1$ and M_1 is a monotone class containing G. From the definition of $M(G)$ then follows that $M(G) \subseteq M_1$ hence $M_1 = M(G)$. If $C \in G$ and $A \in M(G)$ then it follows from $M_1 = M(G)$ that $A \cap C \in M(G)$, hence $C \in M_2$ and $G \subseteq M_2 \subseteq M(G)$. As above for M_1, one proves that M_2 is a monotone class. Then the fact that $M(G)$ is the smallest monotone class containing G implies that $M(G) \subseteq M_2$ hence $M_2 = M(G)$ and $M(G)$ is closed with respect to binary intersection.
(4) Define $M_3 = \{A \in M(G) | A^c \in M(G)\}$. Because G is an algebra, $G \subseteq M_3 \subseteq M(G)$. Consider the sequence of sets, either increasing or decreasing, $\{A_n \in M_3, n \in \mathbb{Z}_+\}$. Then $(\cup_{n=1}^{\infty} A_n)^c = \cap A_n^c \in M(G)$ and $(\cap_{n=1}^{\infty} A_n)^c = \cup A_n^c \in M(G)$ imply that $\cup_{n=1}^{\infty} A_n$ and $\cap_{n=1}^{\infty} A_n$ belong to M_3 hence M_3 is a monotone class containing G. Then the definition of $M(G)$ implies that $M(G) \subseteq M_3$ hence that $M(G) = M_3$, and $M(G)$ is closed with respect to complementation.
(5) From (2), (3), and (4) follows that $M(G)$ is an algebra. From Theorem 19.1.3 follows that $M(G)$ is a σ-algebra. Then the definition of $F(G)$ implies that $F(G) \subseteq M(G)$. Thus $F(G) \subseteq M(G)$ and $F(G) = M(G)$. □

In certain applications the following version of a monotone class is useful.

Definition 19.1.5. A family of sets D of a set Ω is said to be a *D-monotone class* if it satisfies:

(1)if $\{A_n \in D, n \in \mathbb{Z}_+\}$ is a countable increasing collection of sets in D then $(\cup_{n \in \mathbb{Z}_+} A_n) \in D$;
(2)if $A_1, A_2 \in D$ with $A_2 \subseteq A_1$ then $(A_1 \cap A_2^c) \in D$.

Note that a σ-algebra is a D-monotone class. The symbol D stands for the family name of the mathematician E.B. Dynkin who has introduced the concept.

Proposition 19.1.6. *Let G be a family of subsets of Ω. Then there exists a smallest D-monotone class generated by G, denoted by $D(G)$, such that $G \subseteq D(G)$. It will be called the* D-monotone class generated by G.

Theorem 19.1.7 (The D-monotone class theorem). *Let G be a family of subsets of Ω such that $\Omega \in G$ and G is closed with respect to finite intersections. Then $F(G) = D(G)$, or, equivalently, the σ-algebra generated by G is equal to the D-monotone class generated by G.*

The proofs of the above two results are similar to those of Proposition 19.1.2 and Theorem 19.1.4 respectively.

19.2 Probability Measures

Consider a probability space (Ω,F,P). A *null set* is a set $N \in F$ such that $P(N) = 0$. A *negligible set* $A \subseteq \Omega$ of (Ω,F,P) is a subset of a null set: there exists a set $N \in F$ such that $A \subseteq N$ and $P(N) = 0$. In general, a negligible set is not in the σ-algebra F. This may cause trouble, for example in case of an intersection over an uncountable subset of measureable sets. To prevent difficulties, negligible sets must therefore be in the σ-algebra. A probability space (Ω,F,P) is said to be *complete* with respect to P if all negligible sets of the probability space (Ω,F,P) belong to the σ-algebra F.

Theorem 19.2.1. *Any probability space (Ω,F,P) can be transformed into a complete probability space $(\Omega,\overline{F},\overline{P})$ where $F \subseteq \overline{F}$, $\overline{F}$ contains all the negligible sets of (Ω,F,P), and $\overline{P}$ extends P meaning that $\overline{P}(A) = P(A)$ for all $A \in F$.*

If (Ω,F,P) is a complete probability space then one says that $A = B$ *almost surely* with respect to P, denoted by *a.s. P* or *a.s.*, if $A,B \in F$ and $(A \cap B^c) \cup (A^c \cap B)$ is a null set. So, for any null set one can write $N = \emptyset$ *a.s. P*.

Definition 19.2.2. Let (Ω_1,F_1) and (Ω_2,F_2) be measurable spaces. A *probability kernel* is a function $Q : \Omega_1 \times F_2 \to \mathbb{R}_+$ such that:

(1)for all $\omega_1 \in \Omega_1$, $Q(\omega_1,.) : F_2 \to \mathbb{R}_+$ is a probability measure on (Ω_2,F_2);
(2)for all $A_2 \in F_2$, $Q(.,A_2) : \Omega_1 \to \mathbb{R}_+$ is a measurable function on (Ω_1,F_1).

Theorem 19.2.3. *Let (Ω_1,F_1,P_1) be a σ-finite measurable space, (Ω_2,F_2) a measurable space, and $Q : \Omega_1 \times F_2 \to \mathbb{R}_+$ be a probability kernel. Then there exists a unique measure P_{12} on $(\Omega_1 \times \Omega_2, F_1 \otimes F_2)$ such that for any $A_1 \in F_1$, $A_2 \in F_2$ the following equality holds*

$$P_{12}(A_1 \times A_2) = \int_{A_1} Q(\omega_1,A_2)P_1(d\omega_1).$$

If P_1 is a probability measure then so is P_{12}.

19.3 Stable Subsets of Probability Distribution Functions

In the theory of limits of sums of independent random variables the concept of a stable distribution has been defined. In stochastic system theory the following concept is more useful.

Definition 19.3.1. (a)The set of probability distribution functions D is said to be *stable with respect to addition* if it is closed with respect to addition of a finite number of independent random variables, each having a probability distribution function in the set. Thus, if for any two independent random variables x_1, x_2 with distributions in D, the sum $(x_1 + x_2)$ has also a probability distribution in the set D of which the parameter values may be different from those of x_1 and x_2.
(b)The class of distributions D is said to be *stable with respect to multiplication* by elements of a set F if it is closed with respect to multiplication by elements of F. Thus, if x has a distribution in the set D and if $s \in F$ then $s\,x$ is a well defined random variable and has a distribution in the set D.

Whether or not a set of probability distributions is stable with respect to addition can be proven by checking their characteristic functions. Note that if $x_1,\ x_2 : \Omega \to \mathbb{R}$ are independent random variables then,

$$E\left[\exp\left(iw\ (x_1 + x_2)\right)\right] = E\left[\exp\left(iw\, x_1\right)\right] \times E\left[\exp\left(iw\, x_2\right)\right],\ \ \forall\, w \in \mathbb{R}.$$

Proposition 19.3.2. *Consider the set of Poisson distributions.*

This set is stable with respect to addition of two independent random variables each having a Poisson probability distribution. If x_1 has a Poisson distribution with parameter λ_1, x_2 has a Poisson distribution with parameter λ_2, and x_1 and x_2 are independent then $x_1 + x_2$ has a Poisson distribution with parameter $\lambda_1 + \lambda_2$. The set of Poisson distributions is not stable with respect to multiplication by real numbers for any nontrivial subset of $\mathbb{R}$.

Proof.

$$\begin{aligned} & E\left[\exp\left(iw\ (x_1 + x_2)\right)\right] = \exp\left((\lambda_1 + \lambda_2)\ \left(\exp(iw) - 1\right)\right) \\ &= E\left[\exp\left(iw\, x_1\right)\right] \times E\left[\exp\left(iw\, x_2\right)\right],\ \forall\, w \in \mathbb{R}. \end{aligned}$$

□

Proposition 19.3.3. *The class of Gaussian distributions on $\mathbb{R}^n$ is stable with respect to addition and stable with respect to multiplication by elements of $\mathbb{R}$.*

Let $x_1,\ x_2 : \Omega \to \mathbb{R}^n$, $x_1 \in G(m_1, Q_1)$, $x_2 \in G(m_2, Q_2)$ with $F^{x_1},\ F^{x_2}$ independent. Then $x_1 + x_2 \in G(m_1 + m_2,\ Q_1 + Q_2)$. Thus the set of Gaussian distributions is stable with respect to addition of independent random variables.

If follows from Proposition 2.7.2 that the set of Gaussian distributions is closed with respect to matrix multiplication.

Proof.

$$E[\exp(iw^T\ x_1)]E[\exp(iw^T\ x_2)] = \exp(iw^T\ (m_1+m_2) - \frac{1}{2}w^T\ (Q_1+Q_2)\ w)$$
$$= E[\exp(iw^T\ (x_1+x_2))],\ \forall\ w \in \mathbb{R}^n.$$

□

The elementary proof of these statements follows immediately from the expression of the characteristic function of a Gaussian random variable. The details are omitted.

Proposition 19.3.4. *Consider the class of Gamma distributions with density function*

$$p(v) = v^{\gamma_1 - 1}\exp(-v/\gamma_2)\lambda_2^{-\gamma_1}/\Gamma(\gamma_1),\ (\gamma_1,\gamma_2) \in (0,\infty)^2.$$

(a)If the random variables x and y have Gamma probability distributions with respectively the parameters, $(\gamma_{x,1},\gamma_{x,2})$ *and* $(\gamma_{y,1},\gamma_{y,2})$ *and are independent, then,*

$$E[\exp(iw\ (x+y))] = (1-iw\ \gamma_{x,2})^{\gamma_{x,1}}\ (1-iw\ \gamma_{y,2})^{\gamma_{y,1}},\ \forall\ w \in \mathbb{R}.$$

If in addition $\gamma_{x,2} = \gamma_{y,2}$ *then* $x+y$ *has a Gamma probability distribution with parameters* $(\gamma_{x,1}+\gamma_{y,1},\gamma_{x,2})$,

$$E[\exp(iw\ (x+y))] = (1-iw\ \gamma_{x,2})^{\gamma_{x,1}+\gamma_{y,1}},\ \ \forall\ w \in \mathbb{R}.$$

(b)If the random variable x has a Gamma probability distribution with parameters $(\gamma_{x,1},\gamma_{x,2})$ *and if is* $a \in \mathbb{R}_{s+} = (0,\infty)$ *is a strictly positive real number then the random variable* $a \times x$ *has a Gamma probability distribution with parameters* $(\gamma_{x,1}, a \times \gamma_{x,2})$*. Thus the set of Gamma distribution is closed with respect to scalar multiplication by strictly positive real numbers.*

The elementary proof is omitted.

19.4 Gaussian Random Variables

In this section results for Gaussian random variables are collected which complement those of Section 2.7.

The geometric approach to Gaussian random variables will be introduced below. The focus of the geometric approach is on the spaces which generate the random variables rather on the representations of the variables themselves. In this section, the spaces concerned are the σ-algebras which the Gaussian random variables generate. The spaces are *not* the corresponding Hilbert spaces.

Proposition 19.4.1. *Let* $x : \Omega \to \mathbb{R}^n$ $x \in G(m,Q)$, $L \in \mathbb{R}^{n_1 \times n}$.

(a)Then $F^{Lx} \subseteq F^x$.
(b)$F^{Lx} = F^x$ *if and only if* $ker(Q) = ker(LQ)$.

Proof. (a) This follows from the fact that the map $x \mapsto Lx$ is Borel measurable, and from the property that a measurable function of a random variable is a random variable.
(b) By Proposition 2.5.14 $F^{f(x)} = F^x$ for any measurable function $f : \mathbb{R}^n \to \mathbb{R}^m$ if and only if f is injective on the support of x. The support of x is the range of Q. Now f is injective on the support of x if and only if u in the range of Q and $f(u) = Lu = 0$ imply $u = 0$ if and only if for any $y \in \mathbb{R}^n$, $LQy = 0$ implies $Qy = 0$, if and only if $\ker(LQ) \subseteq \ker(Q)$. The inclusion $\ker(Q) \subseteq \ker(LQ)$ always holds. □

The above result motivates the following definition.

Definition 19.4.2. Let $x : \Omega \to \mathbb{R}^n$, $x \in G$ and consider the σ-algebra F^x.

(a) A *basis* for F^x is a triple $(n_1, m_1, Q_1) \in \mathbb{N} \times \mathbb{R}^{n_1} \times \mathbb{R}^{n_1 \times n_1}_{pds}$ such that there exists a $x_1 : \Omega \to \mathbb{R}^{n_1}$ satisfying $x_1 \in G(m_1, Q_1)$, $0 \preceq Q_1$, and $F^x = F^{x_1}$.
(b) A *minimal basis* for F^x is a basis (n_1, m_1, Q_1) of F^x such that $\text{rank}(Q_1) = n_1$.
(c) A *basis transformation* of F^x is a map $x \mapsto Lx$, with $L \in \mathbb{R}^{n_1 \times n_1}$ a nonsingular matrix such that $F^{Lx} = F^x$.

By use of linear algebra one can, for a basis (n_1, m_1, Q_1), always construct a minimal basis for F^x.

Proposition 19.4.3. *Let $x_1 : \Omega \to \mathbb{R}^{n_1}$, $x_1 \in G(0, Q_1)$, and $0 \preceq Q_1$. Then there exists a $n_2 \in \mathbb{N}$ and a basis transformation $L \in \mathbb{R}^{n_2 \times n_1}$ such that, if $x_2 : \Omega \to \mathbb{R}^{n_2}$, $x_2 = Lx_1$, then $x_2 \in G(0, Q_2)$, $0 \prec Q_2$, and $F^{x_2} = F^{x_1}$.*

Proof. Let $n_2 \in \mathbb{N}$, $U \in \mathbb{R}^{n_1 \times n_1}$ be orthogonal, and $D_1 \in \mathbb{R}^{n_2 \times n_2}$ be such that,

$$Q_1 = U \begin{pmatrix} D_1 & 0 \\ 0 & 0 \end{pmatrix} U^T, \;\; U^T = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}, \; U_1 \in \mathbb{R}^{n_2 \times n_1}, \; U_2 \in \mathbb{R}^{(n_1 - n_2) \times n_1},$$
$$D_1 = \text{Diag}(d_1, d_2, \ldots, d_{n_2}), \;\; d_1 \geq d_2 \geq \ldots \geq d_{n_2} > 0.$$

Let $L_1 = \begin{pmatrix} I_{n_2} & 0 \end{pmatrix} U^T$, $x_2 : \Omega \to \mathbb{R}^{n_2}$, and $x_2 = L_1 x_1$. Then $x_2 \in G(0, Q_2)$ and,

$$Q_2 = L_1 Q_1 L_1^T = \begin{pmatrix} I & 0 \end{pmatrix} U^T U \begin{pmatrix} D_1 & 0 \\ 0 & 0 \end{pmatrix} U^T U \begin{pmatrix} I \\ 0 \end{pmatrix} = D_1 = Q_2^T \succ 0.$$

Let $r : \Omega \to \mathbb{R}^{n_1}$,

$$\begin{aligned} r \;&=\; x_1 - U \begin{pmatrix} x_2 \\ 0 \end{pmatrix} = x_1 - U \begin{pmatrix} L_1 \\ 0 \end{pmatrix} x_1 = [I - U \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} U^T] x_1 \\ &=\; U \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} U^T x_1; \\ &\Rightarrow E[rr^T] = U \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} U^T U \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} U^T U \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} U = 0, \\ &\Rightarrow r = 0 \; a.s. \;\Rightarrow\; x_1 = U \begin{pmatrix} x_2 \\ 0 \end{pmatrix} \text{ a.s.; with } x_2 = L_1 x_1, \;\Rightarrow\; F^{x_1} = F^{x_2}. \end{aligned}$$

□

Below the canonical variable decomposition of a tuple of Gaussian random variables is defined.

Problem 19.4.4. Let $y_1 : \Omega \to \mathbb{R}^{k_1}$ and $y_2 : \Omega \to \mathbb{R}^{k_2}$ be jointly Gaussian random variables with $(y_1, y_2) \in G(0, Q)$. Determine a canonical form for the spaces F^{y_1}, F^{y_2}.

Note that a basis transformation of the form $L = \text{Block} - \text{diag}(L_1,\ L_2)$ with $L_1,\ L_2$ nonsingular, leaves the spaces $F^{y_1},\ F^{y_2}$ invariant. Therefore this introduces an equivalence relation on the spaces F^{y_1}, F^{y_2}, hence one can speak about a canonical form which corresponds to a quotient set of a set with an equivalence relation. The above problem has been formulated and solved by H. Hotelling [19].

Definition 19.4.5. Let $y_1 : \Omega \to \mathbb{R}^{n_{y_1}}$ and $y_2 : \Omega \to \mathbb{R}^{n_{y_2}}$ be jointly Gaussian random variables with $(y_1, y_2) \in G(0, Q)$. Then (y_1, y_2) are said to be in *canonical variable form* if

$$Q_{cvf} = \begin{pmatrix} I & Q_{12} \\ Q_{12}^T & I \end{pmatrix} \in \mathbb{R}^{(n_{y_1}+n_{y_2})\times(n_{y_1}+n_{y_2})},\ Q_{12} = \begin{pmatrix} I & 0 & 0 \\ 0 & D & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$\begin{aligned} & n_{y_{11}},\ n_{y_{12}},\ n_{y_{13}},\ n_{y_{21}},\ n_{y_{22}},\ n_{y_{23}} \in \mathbb{N}, \\ & n_{y_1} = n_{y_{11}} + n_{y_{12}} + n_{y_{13}},\ n_{y_2} = n_{y_{21}} + n_{y_{22}} + n_{y_{23}}, \\ & n_{y_{11}} = n_{y_{21}},\ n_{y_{12}} = n_{y_{22}},\ \ I \in \mathbb{R}^{n_{y_{11}} \times n_{y_{11}}}, \\ & D \in \mathbb{R}^{n_{y_{12}} \times n_{y_{12}}},\ D = \text{Diag}(d_1, \ldots, d_{n_{y_{12}}}),\ 1 > d_1 \geq \ldots \geq d_{n_{y_{12}}} > 0. \end{aligned}$$

One then says that $(y_{11}, \ldots, y_{1,n_{y_1}})$, $(y_{21}, \ldots, y_{2,n_{y_2}})$ are the *canonical variables* and $(d_1, \ldots, d_{k_{12}})$ the *canonical correlation coefficients* of the considered tuple of random variables.

Theorem 19.4.6. *Let $y_1 : \Omega \to \mathbb{R}^{k_1}$ and $y_2 : \Omega \to \mathbb{R}^{k_2}$ be jointly Gaussian random variables with $(y_1, y_2) \in G(0, Q)$.*

(a)Then there exists a basis transformation $L = \text{Block} - \text{diag}(L_1,\ L_2)$ such that with respect to the new basis $(L_1 y_1, L_2 y_2) \in G(0, Q_1)$ has the canonical variable form presented in 19.4.5.

(b)Assume that the pair $(y_1, y_2) \in G(0, Q_1)$ is in canonical variable form. Then the basis transformation $L = \text{Block} - \text{diag}(L_1,\ L_2)$ leaves the canonical variable form invariant if and only if, when

$$\begin{aligned} D &= \text{Block} - \text{diag}(D_1, \ldots, D_m),\ \textit{with,} \\ D_i &= \text{Diag}(d_i, \ldots, d_i) = d_i I,\ i \neq j\ \Rightarrow\ d_i \neq d_j,\ \textit{then,} \\ L_1 &= \text{Block} - \text{diag}(L_{1,1}, \ldots, L_{1,m}, L_{1,m+1}), \\ L_2 &= \text{Block} - \text{diag}(L_{2,1}, \ldots, L_{2,m}, L_{2,m+1}), \\ & \quad \textit{compatible with the block decomposition of } D \textit{ such that,} \\ & \quad \forall\, i \in \mathbb{Z}_m,\ L_{1,i}^T L_{1,i} = I,\ L_{2,i}^T L_{2,i} = I,\ D_i L_{2,i} = L_{1,i} D_i, \\ & \quad L_{1,m+1}^T L_{1,m+1} = I,\ L_{2,m+1}^T L_{2,m+1} = I; \end{aligned}$$

$(in\ case\ the\ canonical\ variables\ are\ all\ distinct\ \Leftrightarrow\ i \neq j\ \Rightarrow d_i \neq d_j);$
then $L_1 = \text{Block} - \text{diag}(L_{1,1}, L_{1,m+1}),\ L_2 = \text{Block} - \text{diag}(L_{2,1}, L_{2,m+1}),$
$L_{1,1}, L_{2,1},$ *sign matrices.*

Procedure 19.4.7 Transformation of a variance matrix to canonical variable representation.
Data: $n_{y_1}, n_{y_2} \in \mathbb{Z}_+$, $Q \in \mathbb{R}_{pds}^{(n_{y_1}+n_{y_2})\times(n_{y_1}+n_{y_2})}$ *with decomposition,*

$$Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{12}^T & Q_{22} \end{pmatrix},\ Q_{11} \in \mathbb{R}^{n_{y_1}\times n_{y_2}},\ Q_{22} \in \mathbb{R}^{n_{y_2}\times n_{y_2}},\ Q_{12} \in \mathbb{R}^{n_{y_1}\times n_{y_2}}.$$

(1)Perform singular value decompositions:

$$\begin{aligned} Q_{11} &= U_1 D_1 U_1^T,\ U_1 \in \mathbb{R}_{ortg}^{n_{y_1}\times n_{y_2}},\ U_1 U_1^T = I, \\ D_1 &= \text{Diag}(d_{1,1}, ..., d_{1,n_{y_1}}) \in \mathbb{R}^{n_{y_1}\times n_{y_1}},\ \ d_{1,1} \geq d_{1,2} \geq \ldots \geq d_{1,n_{y_1}} > 0, \\ Q_{22} &= U_2 D_2 U_2^T,\ corresponding\ conditions. \end{aligned}$$

(2)Perform another singular value decomposition of,

$$\begin{aligned} & D_1^{-1/2} U_1^T Q_{12} U_2 D_2^{-1/2} = U_3 D_3 U_4^T, \\ & U_3 \in \mathbb{R}_{ortg}^{n_{y_1}\times n_{y_1}},\ U_4 \in \mathbb{R}_{ortg}^{n_{y_2}\times n_{y_2}},\ U_3 U_3^T = I,\ U_4 U_4^T = I, \\ Q_{12} &= \begin{pmatrix} I & 0 & 0 \\ 0 & D_4 & 0 \\ 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{n_{y_1}\times n_{y_2}}, \\ D_4 &= \text{Diag}(d_{4,1}, ..., d_{4,n_{y_{12}}}) \in \mathbb{R}_{spds}^{n_{y_{12}}\times n_{y_{12}}},\ 1 > d_{4,1} \geq d_{4,2} \geq \ldots \geq d_{4,n_{y_{12}}} > 0. \end{aligned}$$

(3)The transformation to canoncial variable representation is then,

$$(y_1 \mapsto L_1 y_1,\ y_2 \mapsto L_2 y_2),\ L_1 = U_3^T D_1^{-\frac{1}{2}} U_1^T,\ L_2 = U_4^T D_2^{-\frac{1}{2}} U_2^T.$$

The canonical variable form is specified by the integers and the matrices,

$$\begin{aligned} & (n_{y_{11}}, n_{y_{12}}, n_{y_{13}}),\ (n_{y_{21}}, n_{y_{22}}, n_{y_{23}}), \\ Q_{y_1,y_2} &= \begin{pmatrix} I & Q_{12} \\ Q_{12} & I \end{pmatrix},\ Q_{12} = \begin{pmatrix} I & 0 & 0 \\ 0 & D_4 & 0 \\ 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{n_{y_1}\times n_{y_2}}. \end{aligned}$$

The following technical result is used in one of the chapters on stochastic control.

Proposition 19.4.8. *Let* $y : \Omega \to \mathbb{R}^{n_1}$, $y \in G(m, Q_y)$ *with* $m \in \mathbb{R}^{n_1}$, $Q_y \in \mathbb{R}_{spds}^{n_1\times n_1}$. *Let* $w \in \mathbb{R}^{n_2}$, $c \in \mathbb{R}$, *and*

$$Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{12}^T & Q_{22} \end{pmatrix} \in \mathbb{R}_s^{(n_1+n_2)\times(n_1+n_2)}. \tag{19.1}$$

Assume that $(Q_y^{-1} - cQ_{11}) \succ 0$. *Then*

$$
\begin{aligned}
&E\left[\exp\left(\frac{1}{2}c\begin{pmatrix} y \\ w \end{pmatrix}^T L\begin{pmatrix} y \\ w \end{pmatrix}\right)\right] \\
&= \left(\frac{\det\left(\left(Q_y^{-1}-cQ\right)^{-1}\right)}{\det(Q_y)}\right)^{\frac{1}{2}} \exp\left(\frac{1}{2}\begin{pmatrix} m \\ w \end{pmatrix}^T M\begin{pmatrix} m \\ w \end{pmatrix}\right)
\end{aligned}
$$

$$
\begin{aligned}
M = &\begin{pmatrix} Q_y^{-1}(Q_y^{-1}-cQ_{11})^{-1}Q_y^{-1}-Q_y^{-1} & Q_y^{-1}(Q_y^{-1}-cQ_{11})^{-1}cQ_{12} \\ cQ_{12}^T(Q_y^{-1}-cQ_{11})^{-1}Q_y^{-1} & cQ_{22}+c^2Q_{12}^T(Q_y^{-1}-cQ_{11})^{-1}Q_{12} \end{pmatrix} \\
&\in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}.
\end{aligned}
$$

Proof. The proof is a lengthy calculation. Let

$$
r = -\left[Q_y^{-1}-cQ_{11}\right]^{-1}\left(-Q_y^{-1} - cQ_{12}\right)\begin{pmatrix} m \\ w \end{pmatrix}. \text{ Then,}
$$

$$
\begin{aligned}
&E\left[\exp\left(\frac{1}{2}c\ \begin{pmatrix} y \\ w \end{pmatrix}^T L\begin{pmatrix} y \\ w \end{pmatrix}\right)\right] \\
&= \int (2\pi)^{-\frac{1}{2}p}\det(Q_y)^{-\frac{1}{2}} \\
&\quad \exp(\frac{1}{2}c\begin{pmatrix} v \\ w \end{pmatrix}^T L\begin{pmatrix} v \\ w \end{pmatrix}^T - \frac{1}{2}(v-m)^T Q_y^{-1}(v-m))\,dv;
\end{aligned}
$$

$$
\begin{aligned}
a &= -c\begin{pmatrix} v \\ w \end{pmatrix}^T L\begin{pmatrix} v \\ w \end{pmatrix} + (v-m)^T Q_y^{-1}(v-m) \\
&= \begin{pmatrix} v \\ w \end{pmatrix}^T \begin{pmatrix} -cQ & -cQ_{12} \\ -cQ_{12}^T & -cQ_{22} \end{pmatrix}\begin{pmatrix} v \\ w \end{pmatrix} + \begin{pmatrix} v \\ m \end{pmatrix}^T \begin{pmatrix} Q_y^{-1} & -Q_y^{-1} \\ -Q_y^{-1} & Q_y^{-1} \end{pmatrix}\begin{pmatrix} v \\ m \end{pmatrix} \\
&= \begin{pmatrix} v \\ m \\ w \end{pmatrix}^T \begin{pmatrix} Q_y^{-1}-cQ_{11} & -Q_y^{-1} & -cQ_{12} \\ -Q_y^{-1} & Q_y^{-1} & 0 \\ -cQ_{12}^T & 0 & -cQ_{22} \end{pmatrix}\begin{pmatrix} v \\ m \\ w \end{pmatrix} \\
&= \begin{pmatrix} v \\ r \end{pmatrix}^T \begin{pmatrix} (Q_y^{-1}-cQ_{11}) & -(Q_y^{-1}-cQ_{11}) \\ -(Q_y^{-1}-cQ_{11}) & (Q_y^{-1}-cQ_{11}) \end{pmatrix}\begin{pmatrix} v \\ r \end{pmatrix} \\
&\quad - \begin{pmatrix} m \\ w \end{pmatrix}^T M\begin{pmatrix} m \\ w \end{pmatrix}, \text{ by definition of } r, \\
&= (v-r)^T(Q_y^{-1}-cQ_{11})(v-r)\ -\ \begin{pmatrix} m \\ w \end{pmatrix}^T M\begin{pmatrix} m \\ w \end{pmatrix}.
\end{aligned}
$$

$$
\begin{aligned}
& E\left[\exp\left(\frac{1}{2}c\begin{pmatrix} y \\ w \end{pmatrix}^T L \begin{pmatrix} y \\ w \end{pmatrix}^T\right)\right] \\
& = \int (2\pi)^{-\frac{1}{2}p}(\det(Q_y))^{-\frac{1}{2}} \; \exp(-\frac{1}{2}(v-r)^T(Q_y^{-1}-cQ_{11})(v-r)) \\
& \quad \times \exp\left(\frac{1}{2}\begin{pmatrix} m \\ w \end{pmatrix}^T M \begin{pmatrix} m \\ w \end{pmatrix}\right) dv \\
& = \left(\frac{\det((Q_y^{-1}-cQ_{11})^{-1})}{\det(Q_y)}\right)^{1/2} \times \exp\left(\frac{1}{2}\begin{pmatrix} m \\ w \end{pmatrix}^T M \begin{pmatrix} m \\ w \end{pmatrix}\right).
\end{aligned}
$$

□

Nonlinear Functions of Gaussian Random Variables

Problem 19.4.9. *Nonlinear functions of Gaussian random variables which are also Gaussian random variables*. Describe or classify all nonlinear functions of a multivariable Gaussian random variable which have a Gaussian probability distribution.

The above formulated problem was first published in a paper of L.A. Shepp, [31], who stated an elementary example. The problem is motivated by the extension of Kalman filtering for Gaussian systems to filtering of nonlinear stochastic systems driven by Gaussian disturbances. The author learned of this problem from A.N. Shiryaev (also spelled as Shiryayev) during a personal meeting but Shiryaev did not show the author any example. The above problem is still open. The solution of the problem will provide information how Gaussian random variables can be transformed nonlinearly to other Gaussian random variables. The solution will be useful to information theory and to communication theory.

There follows an example of such a function, different from that of [31].

Example 19.4.10. *A nonlinear function of a multivariable Gaussian random variable which function is itself also a Gaussian random variable.*

Consider a multivariable Gaussian random variable with representation,

$$
x \in G(0,I_4), \; x : \Omega \to \mathbb{R}^4, n_x = 4, \; x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}.
$$

Thus x_1, x_2, x_3, x_4 are independent standard Gaussian random variables. Define a function of the random variable x by the formula,

$$
\begin{aligned}
& y = f(x) = \left(\frac{x_1^2}{x_1^2+x_2^2}\right)^{1/2} x_3 + \left(q_y + \left(\frac{x_2^2}{x_1^2+x_2^2}\right)\right)^{1/2} x_4, \\
& \quad f : \mathbb{R}^4 \to \mathbb{R}, \; y : \Omega \to \mathbb{R}, \; q_y \in (0,\infty) \subset \mathbb{R}.
\end{aligned}
$$

Then the random variable y is a Gaussian random variable with probability distribution function $G(0,(q_y+1))$. The example can be generalized in several ways.

The proof of the above claim follows. From the assumption $x \in G(0,I_4)$ follows that,

$$0 < \frac{x_1^2}{x_1^2+x_2^2},\; 0 < \frac{x_2^2}{x_1^2+x_2^2},\; 0 < q_y + \frac{x_2^2}{x_1^2+x_2^2},\; a.s.$$

Define the functions $g_1,\ g_2 : \mathbb{R}^2 \to \mathbb{R}_+$ and $f : \mathbb{R}^4 \to \mathbb{R}$,

$$\begin{aligned} g_1(x_1,x_2) &= \left(\frac{x_1^2}{x_1^2+x_2^2}\right)^{1/2},\; g_2(x_1,x_2) = \left(q_y + \frac{x_2^2}{x_1^2+x_2^2}\right)^{1/2}, \\ y &= f(x) = g_1(x_1,x_2)\, x_3 + g_2(x_1,x_2)\, x_4. \end{aligned}$$

These functions are well defined by the above calculations, the square roots exist because of strict positivity of the expressions. The Gaussian random variables $x_3,\ x_4$ are thus scaled by the functions $g_1(x_1,x_2)$ and $g_2(x_1,x_2)$ respectively.

Below the notation of conditional independence is used. Because $x \in G(0,I_4)$, the component random variables $x_1,\ x_2,\ x_3,\ x_4$ are independent random variables. It follows from Proposition 2.9.3.(b) that $(F^{x_3}, F^{x_4} | \, F^{x_1,x_2}) \in$ CI and then from Proposition 19.8.2.(f) that $(F^{x_1,x_2,x_3}, F^{x_1,x_2,x_4} | \, F^{x_1,x_2}) \in$ CI.

Note the calculations,

$$\begin{aligned} & E[\exp(iw\, y)] = E[\exp(iw[g_1(x_1,x_2)\, x_3 + g_2(x_1,x_2)\, x_4])] \\ = \; & E\,[E[\exp(iwg_1(x_1,x_2)x_3)\; \exp(iwg_2(x_1,x_2)x_3])|\, F^{x_1,x_2}]] \\ = \; & E\,[E[\exp(iwg_1(x_1,x_2)x_3)|\, F^{x_1,x_2}]\; E[\exp(iwg_2(x_1,x_2)x_4)|\, F^{x_1,x_2}]] \\ & \text{by the above conditional independence relation,} \\ = \; & E[\exp(-w^2 g_1(x_1,x_2)^2/2)\; \exp(-w^2 g_2(x_1,x_2)^2/2)], \\ & \text{because } g_1(x_1,x_2)\, x_3 \text{ is conditionally Gaussian conditioned on } F^{x_1,x_2} \text{ etc.,} \\ = \; & E[\exp(-w^2[g_1(x_1,x_2)^2 + g_2(x_1,x_2)^2]/2)] \\ = \; & \exp(-w^2[q_y+1]/2),\; \text{by definition of } g_1 \text{ and } g_2,\; \forall\, w \in \mathbb{R}, \\ \Rightarrow \; & y \in G(0,[q_y+1]). \end{aligned}$$

19.5 Spaces and Sequences of Random Variables

Consider the vector space $(\mathbb{R},\ \mathbb{R}^n)$ with norm denoted by $\|.\|_n$. Let $p \in (0,\infty)$. Define the function,

$$\|.\|_p : \{x : \Omega \to \mathbb{R}^n\} \mapsto (\mathbb{R}_+ \cup +\infty),\text{ by } \|x\|_p = (E[\|x\|_n^p])^{1/p},$$

which may take the value $+\infty$. Let,

$$\begin{aligned} L_p(\Omega,\mathbb{R}^n) &= L_p((\Omega,F),(\mathbb{R}^n,B(\mathbb{R}^n))) = \{x : \Omega \to \mathbb{R}^n |\ \|x\|_p < \infty\}, \\ L_\infty(\Omega,\mathbb{R}^n) &= L_\infty((\Omega,F),(\mathbb{R}^n,B(\mathbb{R}^n)) = \{x : \Omega \to \mathbb{R}^n |\ \mathrm{P}-\mathrm{esssup}\, x < \infty\}. \end{aligned}$$

If the range space of the random variables is known from the context then one writes only L_p. Let,

$$L_+ = L(\Omega,\mathbb{R}_+) = \{x : \Omega \to \mathbb{R}_+ | \ \forall \omega \in \Omega, \ x(\omega) \geq 0\}.$$

Theorem 19.5.1. *For any $n \in \mathbb{Z}_+$, $p \in [1,\infty)$, the space $L_p(\Omega,\mathbb{R}^n)$ is a vector space over the field $\mathbb{R}$ with semi-norm $\|.\|_p$. Furthermore, $(L_p(\Omega,\mathbb{R}^n), \|.\|_p)$ is a complete vector space, hence a Banach space. The space $L_2(\Omega,\mathbb{R})$ with the inner product $(x,y) = \|xy\|_1$ is a Hilbert space.*

Theorem 19.5.2. *The following inequalities hold:*

(a)The Hölder inequality: *if $p,q \in (1,\infty)$ are such that $p^{-1}+q^{-1} = 1$, if $x \in L_p$, and if $y \in L_q$ then $\|xy\|_1 \ \leq \ \|x\|_p\|y\|_q$ and $xy \in L_1$.*
*(b)*Cauchy-Schwartz inequality *(a special case of (a) with $p = q = 1/2$): if $x,y \in L_2$ then $xy \in L_1$ and $\|xy\|_1 \leq \ \|x\|_2\|y\|_2$.*
*(c)*Minkowski inequality *: if $p \in (1,\infty)$, $x,y \in L_p$ then $(x+y) \in L_p$ and $\|x+y\|_p \leq \|x\|_p + \|y\|_p$.*

Proposition 19.5.3 (Doob's inequalities). *Let $x,y \in L(\Omega,\mathbb{R}_+)$ and assume that for all $u \in (0,\infty)$, $P(x > u) \leq E[I_{(x>u)}y]/u$.*

(a)Then,

$$E[x] \leq \frac{e}{e-1} + \frac{e}{e-1}E[yln^+(y)].$$

(b)For $p, \ q \in (1,\infty)$ with $p^{-1}+q^{-1} = 1$, $\|x\|_p \leq q\|y\|_p$.

Proposition 19.5.4. *Let $x : \Omega \to \mathbb{R}$, and let $f,g : \mathbb{R} \to \mathbb{R}$ be increasing functions (f is said to be* increasing *if $u < v$ implies that $f(u) \leq f(v)$). Then,*

$$E[f(x)]E[g(x)] \ \leq \ E[f(x)g(x)].$$

Sequences of Random Variables

A sequence of real-valued random variables $\{x_n : \Omega \to \mathbb{R}, n \in \mathbb{Z}_+\}$ is called:

- *increasing* if for all $\omega \in \Omega$, $n \in \mathbb{Z}_+$, $x_n(\omega) \leq x_{n+1}(\omega)$;
- *decreasing* if for all $\omega \in \Omega$, $n \in \mathbb{Z}_+$, $x_n(\omega) \geq x_{n+1}(\omega)$;
- *monotone* if it is either increasing or decreasing.

The limit of an increasing sequence of real-valued random variables

$$\begin{aligned}
& \{x_n(\omega),\ \forall\, n \in \mathbb{Z}_+\},\ \text{ is defined as:} \\
\lim_{n\to\infty} x_n(\omega) &= \sup_{n\in\mathbb{Z}_+} x_n(\omega),\ \forall\ \omega \in \Omega; \\
& \text{similarly, for a decreasing sequence one defines,} \\
\lim_{n\to\infty} x_n(\omega) &= \inf_{n\in\mathbb{Z}_+} x_n(\omega),\ \forall\ \omega \in \Omega; \\
& \text{for non-monotone sequences define,} \\
\limsup\, x_n(\omega) &= \lim_{n\to\infty} \sup_{m\ge n} x_m(\omega), \\
\liminf x_n(\omega) &= \lim_{n\to\infty} \inf_{m\ge n} x_m(\omega).
\end{aligned}$$

Note that if for any $n \in \mathbb{Z}_+$, $y_n = \sup_{m\le n} x_n(\omega)$, then $\{y_n, n \in \mathbb{Z}_+\}$ is a decreasing sequence hence $\lim_{n\to\infty} y_n$ is well defined by the above definition. One says that the limit exists if $\limsup x_n(\omega) = \liminf x_n(\omega)$ for all $\omega \in \Omega$, and it is then defined as $\lim\ x_n = \limsup x_n = \liminf x_n$. The limit will be allowed to take values in the extended real line $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty, -\infty\}$.

The following result allows extensions of concepts from simple random variables to positively real-valued random variables.

Proposition 19.5.5. *Consider a positive real-valued random variable $x : \Omega \to \mathbb{R}_+$. There exists an increasing sequence of simple positive random variables,*

$$\{x_n : \Omega \to \mathbb{R}_+, n \in \mathbb{Z}_+\},\ \textit{ such that } \forall \omega \in \Omega,\ \ x(\omega) = \lim_{n\to\infty} x_n(\omega).$$

Proof. Define for $n \in \mathbb{Z}_+$ and for $k \in \{0, 1, 2, \ldots, n2^n - 1\}$ the random variables

$$x_n(\omega) = \begin{cases} k2^{-n}, & \text{if } k2^{-n} \le x(\omega) < (k+1)2^{-n},\ 0 \le k < n2^{-n} - 1, \\ n, & \text{if } n \le x(\omega). \end{cases}$$

Then $\{x_n, n \in \mathbb{Z}_+\}$ are positive simple random variables and by construction increasing. Then for $n \in \mathbb{Z}_+$,

$$\begin{aligned}
& |x(\omega) - x_n(\omega)| \\
&= |\sum_{k=0}^{n2^n-1} (x(\omega) - k2^{-n}) I_{k2^{-n} \le x(\omega) < (k+1)2^{-n})} + (x(\omega) - n) I_{\{n < x(\omega)\}}| \\
&= \sum_{k=0}^{n2^n-1} |x(\omega) - k2^{-n}| I_{k2^{-n} \le x(\omega) < (k+1)2^{-n})} + |x(\omega) - n| I_{\{n < x(\omega)\}}| \\
&\le 2^{-n} + |x(\omega) - n| I_{\{n < x(\omega)\}},\ \text{ hence} \\
0 &= \lim_{n\to\infty} |x(\omega) - x_n(\omega)|,\ \forall \omega \in \Omega,\ \text{ because } \lim_{n\to\infty} I_{\{n < x(\omega)\}} = 0.
\end{aligned}$$

□

The concept of a monotone class of a collection of sets can be extended to one for random variables. This concept can then be used to prove properties of random variables.

Theorem 19.5.6. *Consider set Ω and a family G of subsets of Ω such that $\Omega \in G$ and G is closed with respect to finite intersections. Let*

$$L(\Omega, F(G)) = \left\{ \begin{array}{l} x : \Omega \to \mathbb{R} | x \text{ a random variable} \\ \text{measurable with respect to } F(G) \end{array} \right\}.$$

Assume that there is a set $H \subseteq L(\Omega, F(G))$ such that:

(1)H is a vector space over $\mathbb{R}$;
(2)$I_A \in H$ for all $A \in G$;
(3)H is closed with respect to increasing limits: if $\{x_n \in H, n \in \mathbb{Z}_+\}$ is an increasing sequence of nonnegative random variables and if $\lim x_n = x < \infty$, then $x \in H$.

Under these conditions $H = L(\Omega, F(G))$.

Corollary 19.5.7. *Given the sub-σ-algebras $G_1,\ G_2$. Let*

$$G = \{A_1 \cap A_2 | A_1 \in G_1, A_2 \in G_2\}.$$

Denote by $L_b(\Omega, G_1 \vee G_2)$ the set of all bounded real valued random variables that are measurable with respect to $G_1 \vee G_2$. Let $H \subset L_b(\Omega, G_1 \vee G_2)$ be such that:

(1)H is a vector space over the field $\mathbb{R}$;
(2)$I_A \in H$ for all $A \in G$;
(3)H is closed with respect to increasing limits where the limit is required to be bounded.

Then $H = L_b(\Omega, G_1 \vee G_2)$.

For an application of the above result see the proof of Proposition 2.9.2.

Convergence of Sequences of Random Variables

Let $\{x_k : \Omega \to \mathbb{R},\ k \in \mathbb{Z}_+\}$ be a sequence of real-valued random variables. The definition of convergence of random variables used up to this point in this appendix is $\lim_{k \to \infty} x_k(\omega) = x(\omega)$ for all $\omega \in \Omega$. This definition does not depend on the probability measure. Below other convergence concepts are introduced that do depend on the probability measure.

Definition 19.5.8. The sequence $\{x_k : \Omega \to \mathbb{R}^n,\ k,\ n \in \mathbb{Z}_+\}$ of real-valued random variables converges to a random variable x:

(a)*in probability* if, for all $\varepsilon \in (0, \infty)$, $\lim_{k \to \infty} P(\{\|x - x_k\| > \varepsilon\}) = 0$;
Notation $P - \lim_{k \to \infty}\ x_k = x$;
(b)*in p-th mean* for $p \in (0, \infty]$ if $\lim_{k \to \infty}\ E\|x - x_k\|^p = 0$;
notation $L_p - \lim_{k \to \infty}\ x_k = x$;
(c)*almost surely* if $P(\{\omega \in \Omega | \lim_{k \to \infty} x_k(\omega) = x(\omega)\}) = 1$ where the convergence has to take place componentwise; notation $\text{a.s.} - \lim x_k = x$;

(d) *in distribution* or *weakly* if $\lim_{k\to\infty} f_{x_k}(u) = f_x(u)$, for all $u \in \mathbb{R}$ for which f_x is continuous. Here f_{x_k}, f are respectively the pdf's of $x_k,\ x$.
Notation $D - \lim_{k\to\infty} x_k = x$.

As in real analysis, one can avoid the use of the possibly unknown limit random variable x by defining mutual convergence.

Definition 19.5.9. The sequence of real-valued random variables
$\{x_k : \Omega \to \mathbb{R}^n,\ k \in \mathbb{Z}_+\}$ converges mutually:

(a) *in probability* if for all $\varepsilon \in (0,\infty)$, $\lim_{k,m\to\infty} P(\{\|x_k - x_m\| > \varepsilon\}) = 0$;
(b) *in p-th mean* for $p \in (0,\infty]$ if $\lim_{k,\ m\to\infty} E\|x_k - x_m\|^p = 0$; and
(c) *almost surely* if $P(\{\omega \in \Omega \,|\, \lim_{k,\ m\to\infty} \|x_k(\omega) - x_m(\omega)\| = 0\}) = 1$.

It can be proven that for the above defined convergence concepts mutual convergence is equivalent to convergence.

Proposition 19.5.10. *Consider a sequence of real-valued random variables* $\{x_k : \Omega \to \mathbb{R},\ k \in \mathbb{Z}_+\}$.

(a) Convergence almost surely implies convergence in probability.
(b) Convergence in p-th mean implies convergence in probability.
(c) If the sequence converges in probability, then there exists a subsequence $\{x_{k_m},\ m \in \mathbb{Z}_+\}$ *converging almost surely to the same limit.*

There exist examples that show that:

- convergence in probability does not imply convergence in p-th mean
- convergence in p-th mean does not imply convergence almost surely
- convergence almost surely does not imply convergence in p-th mean

The most important convergence concept is that of convergence almost surely. This definition is difficult to verify directly. Below follows a sufficient condition.

Proposition 19.5.11. *If* $\{x_k : \Omega \to \mathbb{R},\ k \in \mathbb{Z}_+\}$ *is a sequence of real-valued random variables and if there exists a sequence* $\{\varepsilon_k \in (0,\infty),\ k \in \mathbb{Z}_+\}$ *of strictly positive real numbers such that,*

$$\sum_{k\in\mathbb{Z}_+} \varepsilon_k < \infty \text{ and } \sum_{k\in\mathbb{Z}_+} P(\{|x_{k+1} - x_k| > \varepsilon_k\}) < \infty,$$

then $a.s. - \lim_{k\to\infty} x_k = x$ *and* $x < \infty$ *a.s..*

Let $\{x_k : \Omega \to \mathbb{R},\ k \in \mathbb{Z}_+\}$ be a sequence of random variables such that a.s. $-\lim x_k = x$. The question to be discussed in this subsection is whether $\lim_{k\to\infty} E[x_k] = E[\lim_{k\to\infty} x_k] = E[x]$. The following results are known for this question.

Lemma 19.5.12. *[Fatou's lemma] Let* $\{x_k : \Omega \to \mathbb{R}_+,\ k \in \mathbb{Z}_+\}$ *be a sequence of real-valued positive integrable random variables. Then,*

$$E[\liminf_{k\to\infty} x_k] \le \liminf_{k\to\infty} E[x_k].$$

Theorem 19.5.13. Monotone convergence theorem. *Let* $\{x_k : \Omega \rightarrow \mathbb{R}_+,\ k \in \mathbb{Z}_+\}$ *be an increasing sequence of positive and integrable random variables such that* $\lim_{k\rightarrow\infty} x_k = x$*. Then,*

$$E[x] = E[\lim_{k\rightarrow\infty} x_k] = \lim_{k\rightarrow\infty} E[x_k].$$

Theorem 19.5.14. Dominated convergence theorem. *Let* $\{x_k : \Omega \rightarrow \mathbb{R},\ k \in \mathbb{Z}_+\}$ *be a sequence of random variables such that* $\lim_{k\rightarrow\infty} x_k = x$*, and assume that there exists an integrable random variable y such that*
$|x_k(\omega)| \leq y(\omega)$ *for all* $\omega \in \Omega$ *and all* $k \in \mathbb{Z}_+$*. Then,*

$$E[x] = E[\lim_{k\rightarrow\infty} x_k] = \lim_{n\rightarrow\infty} E[x_k].$$

Definition 19.5.15. The family of real-valued random variables $\{x_i : \Omega \rightarrow \mathbb{R}, i \in I\}$, with the index set I not necessarily countable, is said to be *uniformly integrable* and *uniform integrability* is said to hold, if

$$\lim_{c\rightarrow\infty} \sup_{i\in I} E[|x_i|I_{\{|x_i|>c\}}] = 0.$$

Proposition 19.5.16. *The family of real-valued random variables* $\{x_i : \Omega \rightarrow \mathbb{R}, i \in I\}$ *is uniformly integrable if:*

(a)and only if:

(1)for all $\varepsilon \in (0,\infty)$ *there exists a* $\delta(\varepsilon) \in (0,\infty)$ *such that if* $P(A) > \delta(\varepsilon)$ *then*

$$\sup_{i\in I} E[|x_i|I_A] \leq \varepsilon;$$

(2)$\sup_{i\in I} E|x_i| < \infty$*;*

(b)there exists a random variable $x \in L_1(\Omega,\mathbb{R}_+)$ *such that for all* $i \in I$*,* $|x_i| \leq x$ *a.s.*
In particular, every finite family of random variables, for all $n \in \mathbb{Z}_+$*,*
$\{x_m : \Omega \rightarrow \mathbb{R},\ m \in \mathbb{Z}_n\}$ *is uniformly integrable;*

(c)there exists a Borel measurable function $g : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ *that is increasing and such that* $\lim_{u\rightarrow\infty} g(u)/u = +\infty$*,* $\{x_i : \Omega \rightarrow \mathbb{R}, i \in I\}$ *is in* $L_1(\Omega,\mathbb{R}^n)$*, and*
$\sup_{i\in I} E[g(|x_i|)] < \infty$*.*
In particular, if there exists a $p \in (1,\infty)$ *such that* $\sup_{i\in I} E\|x_i\|^p < \infty$*.*

Theorem 19.5.17. *Let* $\{x_k : \Omega \rightarrow \mathbb{R},\ k \in \mathbb{Z}_+\} \subseteq L_1(\Omega,\mathbb{R})$ *be uniformly integrable and such that* $\lim_{k\rightarrow\infty} x_k = x$*. Then* $x \in L_1(\Omega,\mathbb{R})$ *and* $\lim_{k\rightarrow\infty} E|x - x_k| = 0$*. Furthermore, if the random variables of the sequence are positive then* $E[x] = \lim_{k\rightarrow\infty} E[x_k]$*.*

19.6 Conditional Expectation and Conditional Probability

In this section the projection operator of one σ-algebra on another is defined and its properties are derived. In addition, several inequalites for conditional expectation are stated. This section is an extension of Section 2.8.

Definition 19.6.1. Consider a complete probability space (Ω,F,P) and two sub-σ-algebras G, H of F. Define the *projection* of H on G as the σ-algebra,

$$\sigma(H|G) = \sigma(\{E[h|G] \mid \forall\, h \in L((\Omega,H),\mathbb{R}_+,B(\mathbb{R}_+)\}) \subseteq G \subseteq F.$$

It is assumed that all the null sets of F are included in the projection. The projection is therefore dependent on the probability measure.

See Proposition 19.8.6 for properties of the projection operator.

Proposition 19.6.2. Jensen's inequality for conditional expectation.
Let $x : \Omega \to \mathbb{R}$ be an integrable random variable, $G \subseteq F$ be a sub-σ-algebra, and $f : \mathbb{R} \to \mathbb{R}$ be a convex and Borel measurable function such that $f(x) \in L_1$. Then

$$f(E[x|G]) \leq E[f(x)|G].$$

In particular, $|E[x|G]| \leq E[|x||G]$.

Proposition 19.6.3. *(Hölder's inequality for conditional expectation))*

(a)Let $p,q \in (1,\infty)$ be such that $p^{-1}+q^{-1} = 1$, and $x,y : \Omega \to \mathbb{R}$, $x \in L_p$, $y \in L_q$. Then

$$E[|xy||G] \;\leq\; (E[|x|^p|G])^{1/p}(E[|y|^q|G])^{1/q} \;\; a.s.$$

(b)Let $x \in L_2$. Then $E[x^2|G] = (E[x|G])^2$ if and only if x is G measurable.

In section 19.5 the question was examined when the limit operator and the expectation operator commute. Similar results as for this question also hold for the interchange of the limit operation and the conditional expectation operation. Below only one such result is stated.

Theorem 19.6.4 (Monotone convergence for conditional expectation). *If $\{x_k : \Omega \to \mathbb{R}_+,\ k \in \mathbb{Z}_+\} \subseteq L_1$ is an increasing sequence of positive random variables such that $x = \lim_{k\to\infty} x_k$, and $G \subseteq F$ is a sub-σ-algebra, then,*

$$\lim_{k\to\infty} E[x_k|\ G] = E[x|G] \ a.s.$$

Definition 19.6.5. Given a probability space (Ω,F,P) and a sub-σ-algebra $G \subseteq F$. A *regular conditional probability* given G is defined to be a function $P(.|G) : \Omega \times F \to [0,1]$ such that:

(1)for $A \in F$ fixed, $P(A|G) : \Omega \to [0,1]$ is a G measurable random variable such that for all $B \in G$, $E[P(A|G)I_B] = P(A \cap B)$;

(2)for all $\omega \in \Omega$ fixed, $P(.|G) : F \to [0,1]$ is a probability measure on F.

A regular conditional probability is a probability kernel as defined in 19.2.2. A regular conditional probability given a sub-σ-algebra G does not always exists, see for an example [14, p. 624].. By restricting attention to real-valued random varaibles one can prove existence.

Definition 19.6.6. Consider a probability space (Ω, F, P), a sub-σ-algebra $G \subseteq F$, and a real-valued random variable $x : \Omega \to \mathbb{R}$.

A *regular conditional probability for x given G* is defined to be a function $P_{\mathbb{R}}(.|G) : \Omega \times G \to [0,1]$ such that:

(1)for $A \in B(\mathbb{R})$ fixed, $P_{\mathbb{R}}(A|\ G) : \Omega \to [0,1]$ is a version of $P(\{x(\omega) \in A\}|\ G)$; and
(2)for any $\omega \in \Omega$ fixed, $P_{\mathbb{R}}(.|G) : B(\mathbb{R}) \to [0,1]$ is a probability measure.

Theorem 19.6.7. *There always exists a regular conditional probability for a real-valued random variable x given G.*

Proposition 19.6.8. *Let $x : \Omega \to \mathbb{R}$ be an integrable random variable, $f : \mathbb{R} \to \mathbb{R}$ be Borel measurable such that $f(x) \in L_1$, $G \subseteq F$ be a sub-σ-algebra, and $P_{\mathbb{R}}(.|G) : \Omega \times F \to [0,1]$ be a regular conditional probability for x given G. Then,*

$$E[f(x)|\ G] = \int_{\mathbb{R}} f(w)\ P_{\mathbb{R}}(dw|\ G)\ \ a.s.$$

19.7 Conditionally Gaussian Random Variables

The concepts and results below are used to derive the conditional Kalman filter and related results.

Definition 19.7.1. The random variable $x : \Omega \to \mathbb{R}^n$ will be called *conditionally Gaussian* with respect to the σ-algebra $G \subseteq F$ with parameters $m : \Omega \to \mathbb{R}^n$ and $Q : \Omega \to \mathbb{R}^{n \times n}_{pds}$ if (1) the random variables m and Q are G measurable and (2) the conditional characteristic function of x given G equals

$$E[\exp(iw^T\ x)|G] = \exp(iw^T\ m - \frac{1}{2} w^T\ Q\, w),$$
$$\forall\ w : \Omega \to \mathbb{R}^n \text{ which is } G\text{-measureable.}$$

The notation $x \in CG(m, Q|\ G)$ denotes that the random variable x is conditionally Gaussian with respect to G with parameters (m, Q).

Proposition 19.7.2. *Let the random variable $x : \Omega \to \mathbb{R}^n$ be conditionally Gaussian given the σ-algebra $G \subseteq F$ with $x \in CG(m, Q|G)$.*

(a)If $E\|x\|_2 < \infty$ then $m = E[x|G]$ a.s.
(b)If $E\|x - m\|_2 < \infty$ then $Q = E[(x-m)(x-m)^T|G]$ a.s.

Proof. This follows along the lines of Proposition 2.6.3. □

Proposition 19.7.3. *Let the random variable $x : \Omega \to \mathbb{R}^n$ be conditionally Gaussian with respect to the σ-algebra $G \subseteq F$ with $x \in CG(m, Q|G)$. Consider the G-measurable random variables $A : \Omega \to \mathbb{R}^{p \times n}$ and $b : \Omega \to \mathbb{R}^p$.*

(a)Then $(Ax+b) \in CG(Am+b, AQA^T|G)$.
(b)If in addition $E\|Ax\|_2 < \infty$ and $E\|b\|_2 < \infty$ then $E[Ax+b|G] = Am+b$.
(c)If in addition to (b) $E\|A(x-m)\|_2 < \infty$, then
$E[(A(x-m))(A(x-m))^T|G] = AQA^T$.

Proof. (a) Let $w \in \mathbb{R}^n$. Then

$$\begin{aligned}
&E[\exp(iw^T(Ax+b))|G] \\
&= E[\exp(i(A^Tw)^Tx)|G]\exp(iw^Tb), \text{ because } b \text{ is } G \text{ measurable,} \\
&= \exp(i(A^Tw)^Tm - \frac{1}{2}w^TAQA^Tw + iw^Tb), \\
&\quad \text{because } A \text{ is } G \text{ measurable hence } Aw \text{ is } G \text{ measurable and, } x \in CG(m,Q|G), \\
&= \exp(iw^T(Am+b) - \frac{1}{2}w^TAQA^Tw).
\end{aligned}$$

(b) From (a) and Proposition 19.7.2 follow that $E[Ax+b|G] = Am+b$.
(c) Similarly, $E[(A(x-m))(A(x-m))^T|G] = AQA^T$. □

Proposition 19.7.4. *Consider the random variables $x : \Omega \to \mathbb{R}^{n_x}$, $y : \Omega \to \mathbb{R}^{n_y}$, and the sub-$\sigma$-algebra $G \subseteq F$. Assume that:*

(1)$(x,y) \in CG(m,Q|G)$ is conditionally Gaussian with respect to G and with the parameters,

$$m = \begin{pmatrix} m_x \\ m_y \end{pmatrix}, \; Q = \begin{pmatrix} Q_{xx} & Q_{xy} \\ Q_{xy}^T & Q_{yy} \end{pmatrix}.$$

(2)$0 \prec Q_{yy}$ a.s.
(3)$E\|x-m_x\|_2 < \infty$ and $E\|x - m_x - Q_{xy}Q_{yy}^{-1}(y-m_y)\|_2^2 < \infty$.

Then x is conditionally Gaussian with respect to $F^y \vee G$ with the parameters,

$$\begin{aligned}
E[\exp(iw^Tx)|F^y \vee G] &= \exp(iw^TE[x|F^y \vee G] - \frac{1}{2}w^TQ_{x|y}w), \\
E[x|F^y \vee G] &= m_x + Q_{xy}Q_{yy}^{-1}(y-m_y), \\
Q_{x|y} &= Q_{xx} - Q_{xy}Q_{yy}^{-1}Q_{xy}^T, \; Q_{x|y} : \Omega \to \mathbb{R}_{pds}^{n_x \times n_x}.
\end{aligned}$$

Proof. Let $z : \Omega \to \mathbb{R}^{n_x}$, $z = x - Q_{xy}Q_{yy}^{-1}y$. Note that

$$
\begin{aligned}
& E\left[\exp(iw_x^T z + iw_y^T y)|G\right] \\
& = E\left[\exp(iw_x^T x + i(w_y - Q_{yy}^{-1} Q_{xy}^T w_x)^T y)|G\right] \\
& = \exp\left(i \begin{pmatrix} w_x \\ w_y - Q_{yy}^{-1} Q_{xy}^T w_x \end{pmatrix}^T \begin{pmatrix} m_x \\ m_y \end{pmatrix} - \frac{1}{2} w_x^T Q_{xx} w_x \right) \times \\
& \quad \times \exp\left(-\frac{1}{2} (w_y - Q_{yy}^{-1} Q_{xy}^T w_x)^T Q_{yy} (w_y - Q_{yy}^{-1} Q_{xy}^T w_x) - w_x^T Q_{xy} (w_y - Q_{yy}^{-1} Q_{xy}^T w_x) \right) \\
& = \exp\left(iw_x^T (m_x - Q_{xy} Q_{yy}^{-1} m_y) - \frac{1}{2} w_x^T Q_1 w_x \right) \times \exp\left(iw_y^T m_y - \frac{1}{2} w_y^T Q_{yy} w_y \right) \\
& = E\left[\exp(iw_x^T z)|G\right] E\left[\exp(iw_y^T y)|G\right],
\end{aligned}
$$

where the last equality follows by using the derived expression, setting either $(w_x, w_y) = (w_x, 0)$ or $(w_x, w_y) = (0, w_y)$, and then concluding that the displayed equality holds. From the above factorization follows that $(F^z, F^y|G) \in CI$. From Proposition 2.9.2 then follows that,

$$
\begin{aligned}
E[\exp(iw_x^T z)|F^y \vee G] &= E[\exp(iw_x^T z)|G], \text{ by conditional independence}, \\
&= \exp(iw_x^T (m_x - Q_{xy} Q_{yy}^{-1} m_y) - \frac{1}{2} w_x^T Q_{x|y} w_x), \\
E[\exp(iw_x^T x)|F^y \vee G] &= \exp(iw_x^T (m_x + Q_{xy} Q_{yy}^{-1} (y - m_y)) - \frac{1}{2} w_x^T Q_{x|y} w_x), \\
&= \exp(iw_x^T E[z|G] - \frac{1}{2} w_x^T Q_{x|y} w_x),
\end{aligned}
$$

because $Q_{xy} Q_{yy}^{-1} y$ is $F^y \vee G$ measurable. □

19.8 Conditional Independence Continued

This section of the appendix is an extension of Section 2.9 where an introduction to conditional independence is provided. The concepts and results described in this section are motivated by the stochastic realization problem.

The results are due to [26, 28, 27], see also the book [16], and to C. van Putten etal. [36].

Recall the definition of the projection of one σ-algebra on another, see Def. 19.6.1. Consider two sub-σ-algebras $G,\ H \subseteq F$ of a probability space (Ω, F, P). Define the *projection* of H on G as,

$$
\sigma(H|G) = \sigma(\{E[h|G] \mid \forall\, h \in L(H, \mathbb{R}_+)\} \subseteq G \subseteq F,
$$

with the understanding that all null sets of F are included in $\sigma(H|G)$.

Theorem 19.8.1. Maximality of the second σ-algebra.
Consider $F_1, F_2, F_3, G \subseteq F$ with $F_2 \subseteq F_3$.

$$
(F_1, F_3|G) \in \text{CI} \;\Leftrightarrow\; \{(F_1, F_2|G) \in \text{CI} \text{ and } \sigma(F_1|F_3 \vee G) \subseteq (F_2 \vee G)\}.
$$

Proof. $(\Rightarrow)$ $(F_1, F_3 | \, G) \in \mathrm{CI}$ implies by $F_2 \subseteq F_3$ and by restriction that $(F_1, F_2 | \, G) \in$ CI. Using Proposition 2.9.2.(a-c) it follows that

$$\sigma(F_1 | F_3 \vee G) \subseteq \sigma(F_1 | G) \subseteq G \subseteq (F_2 \vee G).$$

$(\Leftarrow)$ Consider $x_1 \in L_+(F_1)$. Then,

$$\begin{aligned} E[x_1 | F_3 \vee G] &= E[E[x_1 | F_3 \vee G] | F_2 \vee G], \text{ by } \sigma(F_1 | F_3 \vee G) \subseteq F_2 \vee G, \\ &= E[x_1 | F_2 \vee G], \text{ by } F_2 \subseteq F_3 \ \Rightarrow \ F_2 \vee G \subseteq F_3 \vee G, \\ &= E[x_1 | G], \text{ by } (F_1, F_2 | \, G) \in \mathrm{CI}, \end{aligned}$$

and the result follows with Proposition 2.9.2. □

Proposition 19.8.2. Applications of the preceeding theorem.
Consider the sub-σ-algebras $F_1, F_2, F_3, G \subseteq F$.

(a)

$$(F_1, F_2 \vee F_3 | G) \in \mathrm{CI} \ \Leftrightarrow \ \{(F_1, F_2 | G) \in \mathrm{CI} \ \textit{and} (F_1, F_3 | F_2 \vee G) \in \mathrm{CI}\}.$$

(b)

$$\begin{aligned} & (F_1, F_2 | G) \in \mathrm{CI} \ \textit{and} \ (F_1, F_3 | F_2 \vee G) \in \mathrm{CI} \\ \Leftrightarrow \ & (F_1, F_3 | G) \in \mathrm{CI} \ \textit{and} \ (F_1, F_2 | F_3 \vee G) \in \mathrm{CI}. \end{aligned}$$

(c)Assume that $G \subseteq F_2$.

$$(F_1, F_2 | G) \in \mathrm{CI} \ \Leftrightarrow \ \sigma(F_1 | F_2) \subseteq G.$$

(d)Assume that $F_1 \vee F_2 \vee G$ is independent of F_3.
Then $(F_1, F_2 \vee F_3 | G) \in \mathrm{CI}$ if and only if $(F_1, F_2 | G) \in \mathrm{CI}$.
(e)

$$\begin{aligned} & (F_1 \vee F_3, F_2 | G) \in \mathrm{CI} \ \textit{and} \ (F_1, F_3 | G) \in \mathrm{CI} \\ \Leftrightarrow \ & (F_1, F_3 \vee F_2 | G) \in \mathrm{CI} \ \textit{and} \ (F_3, F_2 | G) \in \mathrm{CI}. \end{aligned}$$

(f) Assume that $F_4 \subseteq (F_2 \vee G)$.
If $(F_1, F_2 | G) \in \mathrm{CI}$ then $(F_1, F_4 | G) \in \mathrm{CI}$. Similarly, $F_3 \subseteq F_1 \vee G \Rightarrow (F_3, F_2 | G) \in \mathrm{CI}$.
In particular, $(F_1 \vee G, \ F_2 \vee G | \, G) \in \mathrm{CI}$ if and only if $(F_1, \ F_2 | \, G) \in \mathrm{CI}$.

Proof. (a) It follows from Proposition 2.9.2 that $\sigma(F_1 | \ F_2 \vee F_3 \vee G) \subseteq F_2 \vee G$ is equivalent to $(F_1, F_3 | \ F_2 \vee G) \in \mathrm{CI}$. The result then follows from Proposition 19.8.1.
(b) By (a) both sides are equivalent with $(F_1, F_2 \vee F_3 | \ G) \in \mathrm{CI}$.
(c) By Proposition 2.9.2, $(F_1, F_2 | \ G) \in \mathrm{CI}$ if and only if $\sigma(F_1 | F_2 \vee G) \subseteq G$. From $G \subseteq F_2$ follows that $\sigma(F_1 | F_2) = \sigma(F_1 | F_2 \vee G)$.
(d) $(\Rightarrow)$ This follows by restriction. $(\Leftarrow)$ $F_1 \vee F_2 \vee G$, F_3 independent σ-algebras and Proposition 2.9.3.(b) imply that $(F_1, F_3 | \ F_2 \vee G) \in \mathrm{CI}$. The conclusion follows from (a).
(e) Note that by applying (a) it follows that,

$$\begin{aligned}
& (F_1 \vee F_3, F_2 |\, G) \in \mathrm{CI},\ (F_1, F_3 | G) \in \mathrm{CI}, \\
\overset{(a)}{\Leftrightarrow}\ & (F_3, F_2 |\, G) \in \mathrm{CI},\ (F_1, F_2 |\, F_3 \vee G) \in \mathrm{CI},\ (F_1, F_3 | G) \in \mathrm{CI}, \\
\overset{(a)}{\Leftrightarrow}\ & (F_3, F_2 |\, G) \in \mathrm{CI},\ (F_1, F_2 \vee F_3 |\, G) \in \mathrm{CI}.
\end{aligned}$$

(f) This follows directly from Theorem 19.8.1 by taking $F_3 = F_2 \vee G$ and by a restriction to $F_4 \subseteq F_2 \vee G$. □

Theorem 19.8.3. Interchange of two state σ-algebras. *Consider a probability space (Ω, F, P) and sub-σ-algebras F_1, F_2, G_1, G_2 such that $G_2 \subseteq G_1$. Then,*

$$\begin{aligned}
& (F_1, F_2 | G_1) \in \mathrm{CI} \text{ and } \sigma(F_1 | G_1) \subseteq G_2 \\
\Leftrightarrow\ & (F_1, F_2 \vee G_1 |\, G_2) \in \mathrm{CI} \\
\Leftrightarrow\ & (F_1, F_2 | G_2) \in \mathrm{CI} \text{ and } \sigma(F_1 | F_2 \vee G_1) \subseteq F_2 \vee G_2.
\end{aligned}$$

Proof. (1) If $(F_1, G_1 | G_2) \in \mathrm{CI}$ then $\sigma(F_1 | G_1) = \sigma(F_1 | G_1 \vee G_2)$ by $G_2 \subseteq G_1$ hence $\sigma(F_1 | G_1) = \sigma(F_1 | G_1 \vee G_2) = \sigma(F_1 | G_2) \subseteq G_2$ by the assumed conditional independence. Conversely, if $\sigma(F_1 | G_1) \subseteq G_2$ then $\sigma(F_1 | G_1 \vee G_2) = \sigma(F_1 | G_1)$ by $G_2 \subseteq G_1$, hence $\sigma(F_1 | G_1 \vee G_2) = \sigma(F_1 |\, G_1) \subseteq G_2$ which implies by Proposition 2.9.3 and $G_2 \subseteq G_1$ that $(F_1, G_1 | G_2) \in \mathrm{CI}$.
(2) Note that $(F_1, F_2 \vee G_1 | F_2 \vee G_2) \in \mathrm{CI}$ if and only if $\sigma(F_1 | F_2 \vee G_1) \subseteq F_2 \vee G_2$ because of Proposition 2.9.3 and $G_2 \subseteq G_1$.
(3) Note the equivalences,

$$\begin{aligned}
& (F_1, F_2 | G_1) \in \mathrm{CI} \text{ and } \sigma(F_1 | G_1) \subseteq G_2 \\
\Leftrightarrow\ & (F_1, F_2 | G_1) \in \mathrm{CI} \text{ and } (F_1, G_1 | G_2) \in \mathrm{CI}, \text{ by Step (1)}, \\
\Leftrightarrow\ & (F_1, F_2 |\, G_1 \vee G_2) \in \mathrm{CI},\ (F_1, G_1 |\, G_2) \in \mathrm{CI}, \text{ by } G_2 \subseteq G_1, \\
\Leftrightarrow\ & (F_1, F_2 \vee G_1 | G_2) \in \mathrm{CI}, \text{ by Proposition 19.8.2.(a)}, \\
\Leftrightarrow\ & (F_1, F_2 | G_2) \in \mathrm{CI} \text{ and } (F_1, G_1 | F_2 \vee G_2) \in \mathrm{CI}, \text{ by Proposition 19.8.2.(a)}, \\
\Leftrightarrow\ & (F_1, F_2 | G_2) \in \mathrm{CI} \text{ and } (F_1, F_2 \vee G_1 | F_2 \vee G_2) \in \mathrm{CI}, \text{ by Proposition 19.8.2.(a)}, \\
\Leftrightarrow\ & (F_1, F_2 | G_2) \in \mathrm{CI} \text{ and } \sigma(F_1 | F_2 \vee G_1) \subseteq F_2 \vee G_2, \\
& \text{because } G_2 \subseteq G_1 \ \Rightarrow\ F_2 \vee G_2 \subseteq F_2 \vee G_1 \text{ and because of Step (2).}
\end{aligned}$$

□

Proposition 19.8.4. Consequences of the preceeding theorem. *Consider a probability space (Ω, F, P) and sub-σ-algebras F_1, F_2, F_3, G_1, G_2.*

(a)The relation $\{\ (F_1, F_2 | G_1) \in \mathrm{CI}$ *and* $(F_1, G_2 | F_2 \vee G_1) \in \mathrm{CI}\ \}$ *hold if and only if* $\{\ (F_1, F_2 | G_1 \vee G_2) \in \mathrm{CI}$ *and* $(F_1, G_2 | G_1) \in \mathrm{CI}$ *hold* $\}$.
(b)The relations $\{\ (F_1, F_2 | G_1) \in \mathrm{CI}$*,* $(F_1, G_2 | F_2 \vee G_1) \in \mathrm{CI}$*, and* $(F_1, G_1 | G_2) \in \mathrm{CI}\ \}$ *hold if and only if the relations* $\{\ (F_1, G_2 | G_1) \in \mathrm{CI}$*,* $(F_1, G_1 | F_2 \vee G_2) \in \mathrm{CI}$*, and* $(F_1, F_2 | G_2) \in \mathrm{CI}\ \}$ *hold.*
(c)If $(F_1, F_2 | G_1) \in \mathrm{CI}$ *and* $F_1 \subseteq F_3$ *then* $(F_1, F_2 | \sigma(F_3 | G_1)) \in \mathrm{CI}$.
(d)For any sub-σ-algebras F_1, F_2, $(F_1, F_2 | \sigma(F_1 | F_2)) \in \mathrm{CI}$ *and* $(F_1, F_2 | \sigma(F_2 | F_1)) \in \mathrm{CI}$.

(e)If $(F_1,F_2|G_1) \in \mathrm{CI}$ *and* $\sigma(F_1|G_1) \subseteq G_2 \subseteq (F_2 \vee G_1)$ *then* $(F_1,F_2|G_2) \in \mathrm{CI}$.
(f) If $(F_1,F_2|G_1) \in \mathrm{CI}$ *then* $(F_1,F_2|\sigma(G_1|F_1)) \in \mathrm{CI}$. *Hence* $\sigma(F_2|F_1) \subseteq \sigma(G_1|F_1)$.
(g)Assume that $F_1 \vee F_2 \vee G_1$ *and* G_2 *are independent* σ*-algebras.*
Then $(F_1,F_2|G_1 \vee G_2) \in \mathrm{CI}$ *if and only if* $(F_1,F_2|G_1) \in \mathrm{CI}$.
(h)For any sub-σ*-algebras* F_1, F_2, $(F_1,F_2|\sigma(F_2|F_1) \vee \sigma(F_1|F_2)) \in \mathrm{CI}$.

Proof. (a) It follows from Proposition 19.8.2 that both sides are equivalent with $(F_1,F_2 \vee G_2|\ G_1) \in \mathrm{CI}$.
(b) By applying (a) twice one obtains,

$$\begin{aligned}
&\{(F_1,F_2|G_1) \in \mathrm{CI},\ (F_1,G_1|\ G_2) \in \mathrm{CI},\ (F_1,G_2|\ F_2 \vee G_1) \in \mathrm{CI}\}\\
\Leftrightarrow\ &\{(F_1,F_2|G_1 \vee G_2) \in \mathrm{CI},\ (F_1,G_2|\ G_1) \in \mathrm{CI},\ (F_1,G_1|\ G_2) \in \mathrm{CI}\}\\
\Leftrightarrow\ &\{(F_1,G_2|\ G_1) \in \mathrm{CI},\ (F_1,G_1|\ F_2 \vee G_2) \in \mathrm{CI},\ (F_1,F_2|\ G_2) \in \mathrm{CI}\}.
\end{aligned}$$

(c) $F_1 \subseteq F_3$ implies that $\sigma(F_1|G) \subseteq \sigma(F_3|G) \subseteq G$. The result then follows from Theorem 19.8.3.
(d) For any random variable $x_1 \in L_+(F_1)$, $\sigma(F_1|F_2) \subseteq F_2$ implies that $E[x_1|F_2 \vee \sigma(F_1|F_2)] = E[x_1|F_2]$ which is $\sigma(F_1|F_2)$ measurable. The result follows from Proposition 2.9.2.(d). By symmetry, $(F_1,F_2|\ \sigma(F_1|F_2)) \in \mathrm{CI}$.
(e) $F_2 \subseteq (F_2 \vee G_1)$ and Proposition 2.9.3 then imply that $(F_1,F_2|\ F_2 \vee G_1) \in \mathrm{CI}$. From Proposition 2.9.2 follows that $\sigma(F_1|F_2 \vee G_1) = \sigma(F_1|G_1) \subseteq G_2 \subseteq (F_2 \vee G_2)$. Now apply Theorem 19.8.3 with G_1 replaced by $F_2 \vee G_1$ to obtain $(F_1,F_2|\ G_2) \in \mathrm{CI}$.
(f) Take in (e) $G_2 = \sigma(G_1|F_1) \vee \sigma(F_2|G_1)$. Then it follows from (e) with F_1 and F_2 interchanged that $(F_1,F_2|\ \sigma(G_1|F_1) \vee \sigma(F_2|G_1)) \in \mathrm{CI}$. From (d) follows that $(F_1,G_1|\ \sigma(G_1|F_1)) \in \mathrm{CI}$ hence $(F_1,\sigma(F_2|G_1)|\sigma(G_1|F_1)) \in \mathrm{CI}$. Combining the above results with Proposition 19.8.2.(a) yields $(F_1,F_2 \vee \sigma(F_2|G_1)|\ \sigma(G_1|F_1)) \in \mathrm{CI}$ hence $(F_1,F_2|\ \sigma(G_1|F_1)) \in \mathrm{CI}$. This and Proposition 19.8.2.(c) yield that $\sigma(F_2|F_1) \subseteq \sigma(G_1|F_1)$.
(g) Independence of $F_1 \vee F_2 \vee G_1$ and G_2 and Proposition 2.9.3.(b) imply that $(F_1,G_2|\ G_1 \vee F_2) \in \mathrm{CI}$ and $(F_1,G_2|\ G_1) \in \mathrm{CI}$. The conclusion then follows using (a). The converse is a direct consequence of independence and the conditional expectation operator, Thm. 2.8.2.
(h) From (d) follows that $(F_1,F_2|\sigma(F_2|F_1)) \in \mathrm{CI}$. Then,

$$\begin{aligned}
&\sigma(F_1|\ F_2 \vee \sigma(F_2|F_1) \vee \sigma(F_1|F_2)) = \sigma(F_1|\ F_2 \vee \sigma(F_2|F_1))\\
\subseteq\ &\sigma(F_2|F_1),\ \text{by the conditional independence,}\ \subseteq (\sigma(F_2|F_1) \vee \sigma(F_1|F_2)),
\end{aligned}$$

and the result follows from Proposition 2.9.2. □

Proposition 19.8.5. Conditional independence invariant under a finite independent addition. *Consider sub-*σ*-algebras* F_1,F_2,F_3,F_4,G_1,G_2. *Assume that the* σ*-algebras* $F_1 \vee F_2 \vee G_1$ *and* $F_3 \vee F_4 \vee G_2$ *are independent.*
Then

$$\begin{aligned}
&(F_1 \vee F_3, F_2 \vee F_4|G_1 \vee G_2) \in \mathrm{CI}\\
\Leftrightarrow\ &(F_1,F_2|G_1) \in \mathrm{CI}\ \textit{and}\ (F_3,F_4|G_2) \in \mathrm{CI}.
\end{aligned}$$

Proof. ($\Rightarrow$) By restriction $(F_1, F_2|\ G_1 \vee G_2) \in \mathrm{CI}$. From Proposition 19.8.4.(g) follows that $(F_1, F_2|\ G_1) \in \mathrm{CI}$. By symmetry one obtains $(F_3, F_4|\ G_2) \in \mathrm{CI}$.
($\Leftarrow$) By Proposition 19.8.2.(d) $(F_1, G_2 \vee F_3 \vee F_2 \vee F_4|\ G_1) \in \mathrm{CI}$ and by Proposition 19.8.2.(a) $(F_1, F_3 \vee F_2 \vee F_4|\ G_1 \vee G_2) \in \mathrm{CI}$. Analogously one proves that, $(F_3, F_3 \vee F_2 \vee F_4|\ G_1 \vee G_2) \in \mathrm{CI}$ hence $(F_3, F_2 \vee F_4|\ G_1 \vee G_2) \in \mathrm{CI}$. The result then follows from Proposition 19.8.2.(e). □

Several properties of the projection operator of σ-algebras are stated.

Proposition 19.8.6. Conditional independence and the projection operator. *Consider sub-σ-algebras $F_1,\ F_2,\ F_3, G$ of (Ω, F). The following properties all hold.*

(a)If $F_1 \subseteq F_2$ then $\sigma(F_1|F_2) = F_1$.
(b)If $F_1 \supseteq F_2$ then $\sigma(F_1|F_2) = F_2$.
(c)If $(F_1, F_2|\ G) \in \mathrm{CI}$ then $\sigma(F_1|F_2 \vee G) = \sigma(F_1|G)$.
(d)$\sigma(F_1|\sigma(F_1|F_2)) = \sigma(F_1|F_2)$.
(e)$\sigma(F_1|\sigma(F_1|F_2) \vee \sigma(F_2|F_1)) = \sigma(F_1|F_2)$.
(f) If $(F_1, F_2|G) \in \mathrm{CI}$ then $\sigma(\sigma(G|F_1)|F_2) = \sigma(F_1|F_2)$.
(g)
(h)$\sigma(\sigma(F_1|F_2)|\sigma(F_2|F_1)) = \sigma(F_2|F_1)$.
(i) If $F_1 \subseteq F_3$ then $F_1 \vee \sigma(F_2|F_3) = \sigma(F_1 \vee F_2|F_3)$.
(j) $\sigma(\sigma(F_2|F_1) \vee \sigma(F_1|F_2)|F_1) = \sigma(F_2|F_1)$.

Proof. Denote the σ-algebras $F_{12} = \sigma(F_1|F_2)$ and $F_{21} = \sigma(F_2|F_1)$.
(a-b) This is directly obvious from the assumed inclusion relations and the definition of the projection operator.
(c) For $x_1 \in L_+(F_1)$, $(F_1, F_2|G) \in \mathrm{CI}$ implies that $E[x_1|F_2 \vee G] = E[x_1|G]$. The result then follows from consideration of the generators of the two projections.
(d) From Proposition 19.8.2(c) follows that $(F_1, F_2|F_{12}) \in \mathrm{CI}$ and then the result follows from (c).
(e) Again $(F_1, F_2|F_{21}) \in \mathrm{CI}$ and by restriction $(F_1, F_{12}|F_{21}) \in \mathrm{CI}$. Then

$$\sigma(F_1|F_{21} \vee F_{12}) = \sigma(F_1|F_{21}),\ \text{ by } (F_1, F_{12}|F_{21}) \in \mathrm{CI} \text{ and (c)},$$
$$= F_{21},\ \text{ by } F_{21} \subseteq F_1 \text{ and (b)}.$$

(f) $(\sigma(G|F_1) \subseteq F_1$ implies by (a) that $\sigma(\sigma(G|F_1)|F_2) \subseteq F_{12}$. $(F_1, F_2|G) \in \mathrm{CI}$ and Proposition 19.8.4.(f) imply that $(F_1, F_2|\sigma(G|F_1)) \in \mathrm{CI}$. Again by
Proposition 19.8.2(f) $(F_1, F_2|\sigma(\sigma(G|F_1)|F_2) \in \mathrm{CI}$. From this and from Proposition 19.8.2.(c) follows that $F_{12} \subseteq \sigma(\sigma(G|F_1)|F_2)$.
(g) By Proposition 2.9.3.(a), $(F_1, F_2|F_2) \in \mathrm{CI}$ and then the result follows from (f).
(h) $F_{21} = \sigma(F_{12}|F_1)$ by (g) hence $F_{21} = \sigma(F_{12}|F_1 \vee F_{21}) = \sigma(F_{12}|F_{21})$ by
$(F_1, F_{12}|F_{21}) \in \mathrm{CI}$.
(i) By assumption $F_1 \subseteq \sigma(F_1 \vee F_2|F_3)$ and also $\sigma(F_2|F_3) \subseteq \sigma(F_1 \vee F_2|F_3)$, hence $F_1 \vee \sigma(F_2|F_3) \subseteq \sigma(F_1 \vee F_2|F_3)$. Let $x_1 \in L_+(F_1)$ and $x_2 \in L_+(F_2)$. Then $E[x_1 x_2|F_3] = x_1 E[x_2|F_3]$ is $F_1 \vee \sigma(F_2|F_3)$ measureable. A monotone class argument shows that for all $y \in L_+(F_1 \vee F_2)$, $E[y|F_3]$ is $F_1 \vee \sigma(F_2|F_3)$ measureable, hence $\sigma(F_1 \vee F_2|F_3) \subseteq F_1 \vee \sigma(F_2|F_3)$.
(j) By (i) , $\sigma(F_{21} \vee F_{12}|F_1) = F_{21} \vee \sigma(F_{12}|F_1) = F_{21}$ where in the latter equality (g) is used. □

Conditional Independence and Finite Probability Spaces

In this subsection a framework is formulated to establish conditional independence for finite probability spaces. An introduction follows.

A *finite probability space* consists of the probability space (Ω, F, P) where, for an integer $n \in \mathbb{Z}_+$, $\Omega = \mathbb{Z}_n = \{1,2,\ldots,n\}$, $F \subseteq \text{Pwrset}(\Omega)$ is the σ-algebra generated by all subsets of Ω, and $P : F \to [0,1]$ is a probability measure. On a finite space no σ-algebra is needed the set of all subsets suffices. To avoid misunderstandings the term σ-algebra is used for the set of all subsets of the finite set Ω. Call P a *uniform probability measure* of the finite probability space if $P(\{i\}) = 1/n$ for all $i \in \Omega$.

For any sub-σ-algebra $G \subseteq F$, call A_G the *set of atoms of G* and call every element of A_G an *atom* of G if A_G is the smallest collection $A_G = \{A_1, \ldots, A_{k_G} \in F\}$ such that $G = \sigma(A_G)$, the elements of A_G are disjoint, and, for all $A_i \in A_G$, $P(A_i) > 0$. The empty set is omitted from inclusion in the set of atoms A_G. This definition implies that A_G is a partition of Ω, hence $\Omega = \cup_{i=1}^{k_G} A_i$ and, for all $i,\ j \in \Omega$ with $i \neq j$, $A_i \cap A_j = \emptyset$. The symbol k_x is used rather than k_G if the σ-algebra $G = F^x$ is generated by a finite-valued random variable x.

Proposition 19.8.7. *Consider a probability space and finite-valued random variables* $y_1 : \Omega \to Y_1 \subset \mathbb{R}^{n_{y_1}}$, $y_2 : \Omega \to Y_2 \subset \mathbb{R}^{n_{y_2}}$, *and* $x : \Omega \to X \subset \mathbb{R}^{n_x}$.

The random variables $y_1,\ y_2$ *are conditionally independent given x, or* $(F^{y_1}, F^{y_2} |\ F^x) \in \text{CI}$, *if and only if the conditional probability measure* $p_{y_1,y_2|x}$ *factorizes as,*

$$\begin{aligned}&p_{y_1,y_2|x} = p_{y_1|x} \times p_{y_2|x};\ \textit{or, equivalently,}\\ &\forall\ k_1 \in Y_1,\ k_2 \in Y_2,\ k_x \in X,\ p_{y_1,y_2|x}(k_1,k_2|k_x) = p_{y_1|x}(k_1|k_x) \times p_{y_2|x}(k_2|x).\end{aligned}$$

It is possible to reformulate the result and require equality of the factorization only for those values in the support set of the random variables. Outside the support set the values are either zero or the conditional probabilities are not defined.

Proof. $(\Leftarrow)$ That conditional independence holds follows by writing out the definition of conditional independence with, in stead of arbitrary random variables, the indicator functions of the atoms of F^{y_1} and F^{y_2}, and writing out the conditional expectations component wise. The conditional independence then follows from the factorization of the conditional probability.
$(\Rightarrow)$ The proof starts with the conditional independence and by writing out the relation as in the converse direction, one obtains the factorization of the conditional probabilities. □

Below are described several operators for finite probability spaces which are used in this chapter.

(1) Recall from Def. 2.5.9 the vector notation of a finite-valued random variable. Let $A_G = \{A_1,\ A_2, \ldots, A_{k_y}\}$ be a set of atoms of F^y. Thus,

$$x = C_x y = \begin{pmatrix} C_1 \; C_2 \; \ldots \; C_{k_y-1} \; C_{k_y} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{k_y-1} \\ y_{k_y} \end{pmatrix},$$
$$\forall\, i \in \mathbb{Z}_{k_y},\; y_i(\omega) = I_{A_i}(\omega),\; F^y = \sigma(\{A_i \in F,\; \forall\, i \in \mathbb{Z}_{k_y}\}),\; P(A_i) > 0.$$

Call y the indicator representation of the finite-valued random variable x.

(2) The conditional expectation for finite-valued random variables on a finite probability space is treated next. The formula of the calculation of the conditional expectation of one finite-valued random variable on another is provided by Proposition 2.8.4. For the benefit of the reader, the formula follows. Let $x,\; y : \Omega \to \mathbb{R}$ be two finite-valued random variables and suppose that both are written in the indicator representation. By definition of the atoms of F^x, for all $j \in \mathbb{Z}_{k_y}$, $E[y_j] > 0$. Then,

$$E[x|F^y](\omega) = C_{x|y} y,$$
$$C_{x|y,i,j} = \frac{E[x_i(\omega) y_j(\omega)]}{E[y_j]},\; \forall\, i \in \mathbb{Z}_{k_x},\; j \in \mathbb{Z}_{k_y},\;\; C_{x|y} \in \mathbb{R}_+^{k_x \times k_y}.$$

(3) The σ-algebra $\sigma(F^x|F^y)$ can be calculated from the conditional expectation. For any atom of F^x, calculate $E[I_A|F^y]$. The set of random variables $\{E[I_{A_x}|F^y],\; \forall\, A_x \in A_{G_x}\}$ then generates $\sigma(F^x|F^y)$ by definition.
See Example 19.8.8 how to calculate the σ-algebra.

(4) The inclusion of two finite σ-algebras $G_2 \subseteq G_1$ holds if any atom $A_2 \in G_2$ is the union of a finite number of atoms of G_1, for example $A_2 = \cup_{i=1}^{n_2} A_{1,i}$ for $A_{1,i} \in G_1$ for all $i = 1, \ldots, n_2$. Such a check is then a simple calculation as explained in examples in this section.

(5) Does equality $\sigma(F^x|F^y) = F^y$ hold? Because $\sigma(F^x|F^y) \subseteq F^y$ always holds, the equality holds if $F^y \subseteq \sigma(F^x|F^y)$. This is proven by showing that every atom of F^y is a finite union of the atoms of $\sigma(F^x|F^y)$.

(6) To check whether or not conditional independence holds of a triple of finite-valued random variables, one has to calculate and check equality of,

$$y_1 : \Omega \to \mathbb{R}^{k_{y_1}},\; y_2 : \Omega \to \mathbb{R}^{k_{y_2}},\; x : \Omega \to \mathbb{R}^{k_x},$$
$$(F^{y_1}, F^{y_2}|F^x) \in \mathrm{CI}$$
$$\Leftrightarrow \forall\, (i,j) \in \mathbb{Z}_{k_{y_1}} \times \mathbb{Z}_{k_{y_2}},\; E[y_{1,i} y_{2,j}|F^x] = E[y_{1,i}|F^x] E[y_{2,j}|F^x].$$

The three conditional expectations can be calculated as described above. For random variables with many values one can use a computer program to check conditional independence.

The calculations for finite-valued random variables are laborious and take much space. Only for Example 19.8.8, the calculations are carried out in detail.

In examples conditional independence can be checked in another way. It is conjectured that for a finite probability space with a uniform probability measure, in which the atoms of F^{y_1} are displayed by horizontal rectangles and the atoms of F^{y_2} are displayed by vertical rectangles, then $(F^{y_1}, F^{y_2}|F^x) \in \mathrm{CI}$ holds if the σ-algebra F^x has atoms which satisfy: (1) they are rectangles and (2) the union of all such atoms covers the set Ω, meaning that $\Omega = \cup_{i=1}^{n_x} A_i$. In addition, $(F^{y_1}, F^{y_2}|F^x) \in \text{CI}_{\min}$ holds if none of the atoms of F^x which are rectangles can be combined into a new rectangle which new set of atoms then meets the conditions (1) and (2) above.

Example 19.8.8. *Conditional independence for a finite probability space.*

Consider the finite probability space with a uniform measure, specified by $\Omega = \mathbb{Z}_9$. See Fig. 19.1 for a diagram of the σ-algebras. Define the following sub-σ-algebras of F by their atoms according to,

$$F^+ = \sigma(\{1,2,3\},\{4,5,6\},\{7,8,9\}),\ F^- = \sigma(\{1,4,7\},\{2,5,8\},\{3,6,9\}),$$
$$G_1 = \{\Omega,\emptyset\},\quad G_2 = \sigma(\{2,3\},\{6,9\},\{7,8\},\{1,4\},\{5\}).$$

The example of G_2 is due to J.C. Willems in a personal communication with the author.

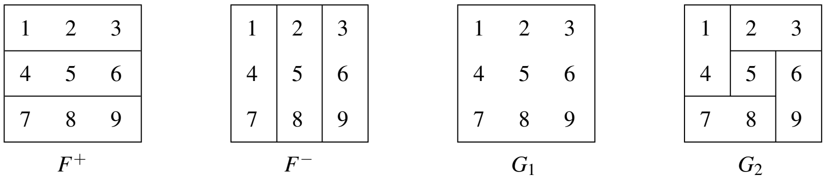


Fig. 19.1 Diagram of the finite probability space and sub-σ-algebras of Example 7.3.16. Each smallest rectangle of a σ-algebra is an atom of the corresponding σ-algebra. From the orientation of the atoms of F^+ and F^- and their overlap, combined with the uniform probability measure, it is directly clear that F^+ and F^- are independent σ-algebras.

Then,

(a) $F^+ \vee F^-$ is a σ-algebra generated of which the atoms are the numbers 1 through 9. Further, $F^{-+} = \sigma(F^-|F^+) = G_1$, $F^{+-} = \sigma(F^+|F^-) = G_1$, $F_f = F^{-+} \vee F^{+-} = G_1$;

(b) $(F^+, F^-|G_1) \in \text{CI}_{\min}$; and

(c) $(F^+, F^-|G_2) \in \text{CI}$ and $\sigma(F^+|G_2) = G_2 = \sigma(F^-|G_2)$. From the previous equality follows that in this case stochastic observability and stochastic co-observability hold for
$(F^+, F^-|G_2) \in \text{CI}$.
But $G_1 \subsetneq G_2 \subsetneq F^+ \vee F^-$ hence G_2 is not minimal for F^+ and F^-.

The proofs of the above statements follow.

Proof. (a) Introduce the notation,

$$\begin{aligned}
& F^+ = F^{y^+},\ F^- = F^{y^-},\ G_2 = F^x,\ y^+ : \Omega \to \mathbb{R}^3, y^- : \Omega \to \mathbb{R}^3, x : \Omega \to \mathbb{R}^5, \\
& y_1^+ = I_{\{1,2,3\}},\ y_2^+ = I_{\{4,5,6\}},\ y_3^+ = I_{\{7,8,9\}}, \\
& y_1^- = I_{\{1,4,7\}},\ y_2^- = I_{\{2,5,8\}},\ y_3^- = I_{\{3,6,9\}}, \\
& x_1 = I_{\{2,3\}},\ x_2 = I_{\{6,9\}},\ x_3 = I_{\{7,8\}},\ x_4 = I_{\{1,4\}},\ x_5 = I_{\{5\}}.
\end{aligned}$$

Because G_1 is the trivial σ-algebra, conditional independence holds as follows from,

$$\begin{aligned}
& E[y^+|G_1] = E[y^+] = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix} = E[y_i^-] = E[y_j^-|F^x] \\
& E[y_i^+(y_j^-)^T] = 1/9,\ \forall\ (i,j) \in \mathbb{Z}_3 \times \mathbb{Z}_3, \\
& \Rightarrow E[y^+(y^-)^T|G_1] = E[y^+(y^-)^T] = E[y^+]E[y^-]^T\ = E[y^+|G_1](E[y^-|G_1])^T \\
& \Rightarrow (F^+, F^-|G_1) = (F^{y^+}, F^{y^-}|G_1) \in \text{CI}_{\min}.
\end{aligned}$$

That minimality holds for G_1 is directly obvious because G_1 has only one atom hence is the sub-σ-algebra of F with the least number of atoms.
(b) Conditional independence is checked first. Recall that, for sets $A,\ B \in F$, $I_A\ I_B = I_{A \cap B}$. Note also that for all $i \in \mathbb{Z}_5$, $E[x_i] > 0$.

$$\begin{aligned}
& E[y_1^+|F^x] = \frac{E[y_1^+ x_1]}{E[x_1]} x_1 + \frac{E[y_1^+ x_2]}{E[x_2]} x_2 + \frac{E[y_1^+ x_3]}{E[x_3]} x_3 + \frac{E[y_1^+ x_4]}{E[x_4]} x_4 + \frac{E[y_1^+ x_5]}{E[x_5]} x_5 \\
= & \frac{E[I_{\{1,2,3\}} I_{\{2,3\}}]}{E[x_1]} x_1 + \frac{E[I_{\{1,2,3\}} I_{\{6,9\}}]}{E[x_2]} x_2 + \frac{E[I_{\{1,2,3\}} I_{\{7,8\}}]}{E[x_3]} x_3 + \\
& + \frac{E[I_{\{1,2,3\}} I_{\{1,4\}}]}{E[x_4]} x_4 + \frac{E[I_{\{1,2,3\}} I_{\{5\}}]}{E[x_5]} x_5 \\
= & x_1 + 0.5\, x_4 = \begin{pmatrix} 1 & 0 & 0 & 0.5 & 0 \end{pmatrix} x, \\
& \text{because } E[I_{\{1,2,3\}} I_{\{1,4\}}]/E[x_4] = E[I_{\{1\}}]/E[x_4] = (1/7)/(2/7) = 1/2 = 0.5, \\
& E[I_{\{1,2,3\}} I_{\{5\}}]/E[x_5] = 0, \text{ etc.}, \\
& E[y_2^+|F^x] = \begin{pmatrix} 0 & 0.5 & 0 & 0.5 & 1 \end{pmatrix} x,\ E[y_3^+|F^x] = \begin{pmatrix} 0 & 0.5 & 1 & 0 & 0 \end{pmatrix} x, \\
& E[y_1^-|F^x] = \begin{pmatrix} 0 & 0 & 0.5 & 1 & 0 \end{pmatrix} x,\ E[y_2^-|F^x] = \begin{pmatrix} 0.5 & 0 & 0.5 & 0 & 1 \end{pmatrix} x, \\
& E[y_3^-|F^x] = \begin{pmatrix} 0.5 & 1 & 0 & 0 & 0 \end{pmatrix} x.
\end{aligned}$$

The check of conditional independence is illustrated by Table 19.1. This is done only for one atom of F^+ and for one atom of F^- because of limitations of space. Note that one has to check the equality of $E[y_i^+ y_j^-|F^x] = E[y_i^+|F^x] \times E[y_j^-|F^x]$ which is equivalent to checking whether the equality of the product of the entries of the last two columns equals to the value of the fourth column. The conclusion is that $(F^+, F^-|G_2) \in \text{CI}$.

Next it is shown that $\sigma(F^+|G_2) = G_2$, equivalently, $\sigma(F^{y^+}|F^x) = F^x$. From the calculation of $E[y^+|F^x]$ see above, follows that,

Atoms of F^x	y_i^+	y_j^-	$E[y_i^+ y_j^- \vert F^x]$	$E[y_i^+ \vert F^x]$	$E[y_j^- \vert F^x]$
$\{1,2,3\}$	$\{1,4,7\}$	$\{2,3\}$	0	1	0
$\{1,2,3\}$	$\{1,4,7\}$	$\{6,9\}$	0	0	0
$\{1,2,3\}$	$\{1,4,7\}$	$\{7,8\}$	0	0	0.5
$\{1,2,3\}$	$\{1,4,7\}$	$\{1,4\}$	0.5	0.5	1
$\{1,2,3\}$	$\{1,4,7\}$	$\{5\}$	0	0	0

Table 19.1 Table for checking conditional independence in a special case of Example 7.3.16.

$$\begin{aligned}
E[y_1^+|F^x] &= E[I_{\{1,2,3\}}|F^x] = I_{\{2,3\}} + 0.5I_{\{1,4\}}, \\
E[y_2^+|F^x] &= E[I_{\{4,5,6\}}|F^x] = 0.5I_{\{6,9\}} + 0.5I_{\{1,4\}} + I_{\{5\}}, \\
E[y_3^+|F^x] &= E[I_{\{7,8,9\}}|F^x] = 0.5I_{\{6,9\}} + I_{\{7,8\}}, \\
& x_1 = I_{\{2,3\}} = I_{\{E[y_1^+|F^x]=1\}},\ x_4 = I_{\{1,4\}} = I_{\{E[y_1^+|F^x]=0.5\}}, \\
& x_5 = I_{\{5\}} = I_{\{E[y_2^+|F^x]=1\}}, \\
& x_3 = I_{\{7,8\}} = I_{\{E[y_3^+|F^x]=1\}},\ x_2 = I_{\{6,9\}} = I_{\{E[y_3^+|F^x]=0.5\}}, \\
& \Rightarrow G_2 \subseteq \sigma(F^+|G_2) \ \Rightarrow \ \sigma(F^+|G_2) = G_2.
\end{aligned}$$

Similarly one proves that $\sigma(F^-|G_2) = G_2$. Thus (b) is proven. □

19.9 Measure Transformations

The purpose of this section is to introduce concepts and results on absolute continuity of probability measures which will be used elsewhere in this book.

Definition 19.9.1. Given a measurable space (Ω, F) and two probability measures P_0, P_1 defined on it.

(a) The probability measure P_1 is said to be *absolutely continuous* with respect to P_0 on F if for all $A \in F$, $P_0(A) = 0$ implies that $P_1(A) = 0$. This is denoted by $P_1 \ll P_0$.

(b) The probability measures are said to be *equivalent* on F if they are mutually absolutely continuous; or, equivalently, $P_1 \ll P_0$ and $P_0 \ll P_1$. This is denoted by $P_1 \sim P_0$.

(c) The probability measure P_1 is said to be *singular* with respect to P_0 if there exists a set $A \in F$ such that $P_0(A) = 0$ and $P_1(A^c) = 0$. This is denoted by $P_1 \perp P_0$.

Expectation with respect to the measures P_0, P_1 is denoted by E_0 respectively E_1.

Example 19.9.2. Consider two probability measures defined on a finite set. They are shown to be absolute continuous,

$$\begin{aligned}
\Omega &= \mathbb{Z}_6 = \{1,2,3,4,5,6\},\ P_a : \Omega \to \mathbb{R}_+,\ P_b : \Omega \to \mathbb{R}_+, \\
P_a(3) &= P_a(4) = P_a(5) = 1/3,\ P_b(4) = P_b(5) = 1/2, \\
A_1 &= \{1,2\},\ P_a(A_1) = 0 = P_b(A_1),\ A_2 = \{6\},\ P_a(A_2) = 0 = P_b(A_2).
\end{aligned}$$

Example 19.9.3. Consider two probability measures induced by probability density functions. They are shown to be absolut continuous.

$$\begin{aligned}
&\Omega = \mathbb{R},\ F = B(\mathbb{R}),\ p_a,\ p_b : \mathbb{R} \to \mathbb{R}_+,\\
&p_a(w) = 1/4,\ v \in [0,4],\ = 0,\ \text{else},\ p_b(w) = 1/2,\ v \in [1,3],\ = 0,\ \text{else},\\
&P_a(A) = \int_A p_a(w)dw,\ P_b(A) = \int_A p_b(w)dw,\\
&\quad P_b \ll P_a,\ A_1 \subseteq (4,\infty)\ \Rightarrow\ P_a(A_1) = 0 = P_b(A_1),\\
&\quad A_2 \subseteq (-\infty,0)\ \Rightarrow\ P_a(A_2) = 0 = P_b(A_2).
\end{aligned}$$

Note that the values of the functions p_a and p_b when these are strictly positive, do not play a role in the proof of absolute continuity.

Theorem 19.9.4. Radon-Nikodym theorem. *Let (Ω,F,P_0) be a probability space.*

(a)P_1 is a probability measure on (Ω,F) and $P_1 \ll P_0$ on F if and only if there exists a real valued random variable $r : \Omega \to \mathbb{R}_+$ satisfying $E_0[r] = 1$, such that

$$P_1(A) = E_0[rI_A],\ \textit{for all } A \in F,\ \textit{denoted by } \frac{dP_1}{dP_0} = dP_1/dP_0 = r.$$

Given P_1, r is unique up to an almost sure modification with respect to P_0.

(b)P_1 is a probability measure on (Ω,F) and $P_1 \sim P_0$ on F if and only if there exists a real valued random variable $r : \Omega \to \mathbb{R}_+,\ r > 0$ a.s. P_0 satisfying $E_0[r] = 1$, such that $dP_1/dP_0 = r$ and $dP_0/dP_1 = 1/r$.

The random variable r will be called the Radon-Nikodym derivative *or* density *of P_1 with respect to P_0.*

The above result establishes a bijective relation between probability measures P_1 that are absolutely continuous with respect to a given measure P_0, and non-negative random variables r satisfying $E_0[r] = 1$. This correspondence allows one to work with random variables rather than with measures. A *measure transformation* is the procedure that given a probability space (Ω,F,P_0) and a random variable $r \geq 0,\ E_0[r] = 1$, constructs a probability measure P_1 by the formula $dP_1/dP_0 = r$.

Example 19.9.5. Consider the probability measures P_a and P_b of Example 19.9.3 and recall that $P_b \ll P_a$. Then their Radon-Nikodym derivative is,

$$r_{b|a} = \begin{cases} 2 = \frac{1/2}{1/4},\ v \in [1,3], \\ 0, \qquad\quad v \in \mathbb{R}\backslash[1,3]. \end{cases}$$

Example 19.9.6. Consider two Gaussian probability measures on $(\mathbb{R},B(\mathbb{R}))$, $P_a = G(m_a,1)$ and $P_b = G(m_b,1)$. Then $P_b \sim P_a$ and the Radon-Nikodym derivative is the quotient of the two probability densities,

$$\begin{aligned}
r_{b|a}(w) &= \frac{p_b(w;(m_b,1))}{p_a(w;(m_a,1))} = \frac{\exp(-(w-m_b)^2/2)}{\exp(-(w-m_a)^2/2)}\\
&= \exp((m_b-m_a)w-(m_b^2-m_a^2)/2).
\end{aligned}$$

That it is the correct formula can be proven also by the construction of the density function of p_b from p_a and $r_{b|a}$ according to,

$$\begin{aligned}p_b(w) &= p_a(w)r_{b|a}(w)\\ &= \exp((m_b - m_a)w - (m_b^2 - m_a^2)/2)\exp(-(w-m_a)^2/2)(2\pi 1)^{-1/2}\\ &= \exp(-(w-m_b)^2/2)(2\pi 1)^{-1/2}.\end{aligned}$$

Proposition 19.9.7. Absolute continuity with respect to a sub-σ-algebra.
Let (Ω,F) be a measurable space with two probability measures P_0,P_1 defined on it. Assume that $P_1 \ll P_0$ on F with $r = dP_1/dP_0$. Let G be a sub-σ-algebra of F. Then $P_1 \ll P_0$ on G and

$$\begin{aligned}&P_1^G(A) = P_1(A),\ P_0^G(A) = P_0(A),\ \forall\, A \in G,\\ r_G &= dP_1^G/dP_0^G = E_0[r|G] \text{ a.s. } P_0^G.\end{aligned}$$

Proposition 19.9.8. Transformation of expectation.
Assume the notation and conditions of Theorem 19.9.4.

(a)If $x : \Omega \to \mathbb{R}_+$ then $E_1[x] = E_0[xr]$;
(b)If $x : \Omega \to \mathbb{R}$ satisfies $E_0|xr| < \infty$, then

$$E_1|x| = E_0|xr| < \infty \text{ and } E_1[x] = E_0[xr].$$

A reason for performing a measure transformation is that the expectation $E_0[xr]$ may be easier to calculate than $E_1[x]$. The proofs of the above two propositions are omitted because they are easily deduced from the definitions.

Proposition 19.9.9. *Consider a measurable space (Ω,F) with two probability measures P_0,P_1 defined on it. Assume that $P_1 \sim P_0$ on F with $r = dP_1/dP_0$.*
Then $dP_0/dP_1 = r^{-1}$ a.s. P_1 and a.s. P_0.

Theorem 19.9.10. Transformation of measures and conditional expectation. *Consider a measurable space (Ω,F) with two probability measures P_0,P_1 defined on it. Assume that $P_1 \ll P_0$ on F with $r = dP_1/dP_0$. Let $x : \Omega \to \mathbb{R}$ be such that $E_0|xr| < \infty$, and let $G \subset F$ be a sub-σ-algebra. Then*

$$E_1[x|G] = \frac{E_0[xr|G]}{E_0[r|G]} \text{ a.s. } P_1.$$

Proof. If $P_0(\{\omega \in \Omega|\ E_0[r|\ G] = 0\}) = 0$ then it follows from absolute continuity that $P_1(\{\omega \in \Omega|\ E_0[r|\ G] = 0\}) = 0$. Suppose therefore that $E_0[r|\ G] > 0$ *a.s.*.

Consider first a positive random variable $x : \Omega \to \mathbb{R}_+$ and a set $A \in G$. Note that then,

$$\begin{aligned}&E_1[I_A\ \frac{E_0[xr|\ G]}{E_0[r|\ G]}] = E_0[I_A\ \frac{E_0[xr|\ G]}{E_0[r|\ G]}\ r]\\ &= E_0[I_A\ \frac{E_0[xr|\ G]}{E_0[r|\ G]}\ E_0[r|\ g]\ ] \text{ by reconditioning on } G,\\ &= E_0[E_0[I_A\ x\ r|\ G]], \text{ because } A \in G,\end{aligned}$$

$$\begin{aligned}
&= E_0[I_A\, x\, r],\ \text{by conditional expectation,}\\
&= E_1[I_A\, x] \ \Rightarrow\ E_1[x|\ G] = \frac{E_0[x\, r|\ G]}{E_0[r|\ G]}.
\end{aligned}$$

The case of a real-valued random variable then follows from decomposing the random variable into the sum of its positive and its negative parts and using the above proven result. □

Theorem 19.9.11. Measure transformation and conditional independence.
Let F_1, F_2, G be sub-σ-algebra's of F, and $P_0,\ P_1 : F \to [0,1]$ be equivalent probability measures on (Ω, F). Let $r : \Omega \to \mathbb{R}_{s+}$ be the projection of the Radon-Nikodym derivative dP_1/dP_0 on $F_1 \vee F_2 \vee G$,

$$r = E[\frac{dP_1}{dP_0}|F_1 \vee F_2 \vee G].$$

Assume that $(F_1, F_2|G) \in CI(P_0)$.

Then $(F_1, F_2|G) \in CI(P_1)$ if and only if there exist random variables $r_1 \in L_+(F_1 \vee G)$ and $r_2 \in L_+(F_2 \vee G)$ such that $r = r_1 r_2 > 0$ a.s. P_0 and P_1. This decomposition is non unique in general.

Proof. (⇐) From Proposition 19.8.2.(f) follows that $(F_1 \vee G, F_2 \vee G|\ G) \in \mathrm{CI}(P_0)$. Consider $x_1 \in L_+(F_1)$. Then,

$$\begin{aligned}
E_1[x_1|F_2 \vee G] &= E_0[x_1 r_1 r_2|F_2 \vee G]/E_0[r_1 r_2|F_2 \vee G]\\
&= r_2 E_0[x_1 r_1|F_2 \vee G]/(r_2 E_0[r_1|F_2 \vee G]) = E_0[x_1 r_1|G]/E_0[r_1|G],\\
&\quad \text{because, } P_1(\{r_1 = 0\}) \le P_1(\{r_1 r_2 = 0\}) = 0,\\
&\qquad (F_1 \vee G, F_2 \vee G|\ G) \in \mathrm{CI}(P_0).
\end{aligned}$$

Thus $E_1[x_1|F_2 \vee G]$ is G measurable and the result follows from
Proposition 2.9.2.(d).
(⇒) Define the random variables,

$$r_1 = E_0[r|F_1 \vee G],\ r_2 = E_0[r|F_2 \vee G]/E_0[r|G].$$

Consider $A_1 \in F_2 \vee G$ and $A_2 \in F_2 \vee G$. Then,

$$\begin{aligned}
&E_0[I_{A_1} I_{A_2} r_1 r_2|G]\\
&= E_0[I_{A_1} r_1|G]\ E_0[I_{A_2} r_2|G],\ \text{by } (F_1 \vee G, F_2 \vee G|\ G) \in \mathrm{CI}(P_0),\\
&= E_0[I_{A_1} E_0[r|F_1 \vee G]|G]\ E_0[I_{A_2} E_0[r|F_2 \vee G]\ /\ E_0[r|G]|G],\ \text{by def. of } r_1,\ r_2,\\
&= E_0[I_{A_1} r|G]\ E_0[I_{A_2} r|G]\ /\ E_0[r|G]\\
&= E_1[I_{A_1}|G]\ E_1[I_{A_2}|G]\ E_0[r|G]\ = E_1[I_{A_1} I_{A_2}|G]\ E_0[r|G],\\
&\quad \text{by } (F_1, F_2|\ G) \in \mathrm{CI}(P_1),\\
&\quad E_0[I_{A_1} I_{A_2}\ r_1 r_2] = E_0[E_0[I_{A_1} I_{A_2}\ r_1 r_2|G]] = E_0[E_1[I_{A_1} I_{A_2}|G]\ E_0[r|G]]\\
&= E_0[E_1[I_{A_1} I_{A_2}|G]\ r] = E_1[I_{A_1} I_{A_2}].
\end{aligned}$$

An application of the monotone class theorem then yields that for all $A \in (F_1 \vee G) \vee (F_2 \vee G) = F_1 \vee F_2 \vee G$, $E_0[I_A r_1 r_2] = E_1[I_A]$, hence $r = r_1 r_2$ *a.s.*(P_0, P_1). □

19.10 The Family of Exponential Probability Distributions

The concept of an exponential family of probability distribution appears at several places in this book. In this section the concept is defined and illustrated. The formal definition is necessarily abstract and it makes use of absolute continuity of measures discussed in the previous section. Below the definitions, several examples are stated.

Definition 19.10.1. A family of probability measures P_{exp} on a measureable space (Ω, F) is said to be an *exponential family* if there exist: (1) a σ-finite measure P_σ on (Ω, F); (2) an integer $k \in \mathbb{Z}_+$; (3) a finite set of functions $a,\ \alpha_1,\ \alpha_2,\ \ldots,\ \alpha_k : P_{exp} \rightarrow \mathbb{R}$; (4) a finite set of measurable functions $t_1, \ldots, t_k : \Omega \rightarrow \mathbb{R}$ and $b : \Omega \rightarrow \mathbb{R}_+$; such that, for all $P \in P_{exp}$, absolute continuity holds $P \ll P_\sigma$, and the Radon-Nikodym derivative has the form of an *exponential representation*,

$$\frac{dP}{dP_\sigma}(\omega) = a(P)b(\omega)\exp(\alpha(P)^T\ t(\omega)), \quad \alpha(P)^T t(\omega) = \sum_{i=1}^{k} \alpha_i(P)t_i(\omega).$$

Define the *order* of the exponential family as the smallest integer $k \in \mathbb{Z}_+$ for which the above-defined representation holds. An exponential represention is called *minimal* if k equals the order.

Call the exponential representation *canonical* if,

$$P_{exp} = \{P(\theta) |\ \forall\ \theta \in \Theta\},\ \Theta \subseteq \mathbb{R}^k,\ \ \frac{dP(\theta)}{dP_\sigma} = a(P)b(\omega)\exp(\theta^T t(\omega)).$$

Proposition 19.10.2. *An exponential representation of an exponential family of probability measures is minimal if and only if (1) the function $\{\alpha_1, \ldots, \alpha_k\}$ are affinely independent; and (2) the functions $\{t_1, \ldots, t_k\}$ are affinely independent.*

Example 19.10.3. Consider the Poisson probability measures with parameter $\lambda \in \mathbb{R}_{s+} = (0, \infty)$ defined by their the frequency functions in the form of a Radon-Nikodym derivative with respect to the counting measure on,

$$\begin{aligned} \Omega &= \mathbb{N} = \{0, 1, \ldots\},\ A = \mathbb{R} \times \mathbb{R}_{s+}, \\ p(k) &= \lambda^k \exp(-\lambda)/k! = \exp(k\ \ln(\lambda) - \lambda)/k!,\ \forall\ k \in \mathbb{N}, \\ &\quad \alpha(\lambda) = (\ln(\lambda), \lambda),\ t(k) = (k, -1),\ b(k) = 1/k! \end{aligned}$$

Then the set of Poisson measures is an exponential family. Because $(\lambda, \ln(\lambda))$ are affinely independent and $(k,\ -1)$ are affinely independent, the order of the exponential representation is 2.

Example 19.10.4. Consider the set of Gamma probability density functions with parameters $(\gamma_1, \gamma_2) \in (0, \infty)^2$. The σ-finite measure is the Lebesgue measure and the Radon-Nikodym derivative has the form,

$$\begin{aligned} \Omega &= \mathbb{R}_+,\ A = \left\{ \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \in \mathbb{R}_{s+} \times \mathbb{R} |\ t_2 \leq \ln(t_1) \right\}, \\ p(v) &= v^{\gamma_1 - 1} \exp(-v/\gamma_2)/\Gamma(\gamma_1) = \exp((\gamma_1 - 1)\ln(v) - v/\gamma_2)/\Gamma(\gamma_1), \\ &\quad \alpha(\gamma_1, \gamma_2) = (\gamma_1 - 1, -1/\gamma_2),\ \ t(v) = (\ln(v), v),\ a(\gamma_1, \gamma_2) = 1/\Gamma(\gamma_1). \end{aligned}$$

Then the set of Gamma measures is an exponential family with the above indicated exponential representation. Because the functions $(v, \ln(v))$ are affinely independent and (γ_1, γ_2) are affinely independent, the order of this exponential representation is 2.

Example 19.10.5. Consider the set of Gaussian probability measures on $\mathbb{R}^n$ for $n \in \mathbb{Z}_+$ with parameters $m \in \mathbb{R}^n$ and $Q \in \mathbb{R}^{n \times n}_{spds}$, with the set of density functions specified by,

$$\begin{aligned}
\Omega &= \mathbb{R}^n, \\
&\quad \alpha_1(m,Q) = -Q^{-1}, \alpha_2(m,Q) = 2m^T Q^{-1}, \\
&\quad \alpha_3(m,Q) = -m^T Q^{-1} m - \ln(\det(Q))/2, \\
&\quad t_1(v) = v^T Q^{-1} v,\ t_2(v) = v,\ t_3(v) = 1,\ a(.) = (2\pi)^{-1/2}, \\
p(v) &= (2\pi \det(Q))^{-1/2} \exp(-(x-m)^T Q^{-1}(x-m)) \\
&= a(.) \exp(v^T Q^{-1} v + 2m^T Q^{-1} v - [m^T Q^{-1} m + \ln(\det(Q))/2] * 1).
\end{aligned}$$

Then the set of Gaussian measures on $\mathbb{R}^n$ is an exponential family and its order is 3.

19.11 Pseudo-Distances on the Set of Probability Measures

The purpose of this section is to define metrics on the set of probability measures. For the definition of a metric and of a distance, see Section 17.5.

Definition 19.11.1. Let X be a set. A *pseudo-distance* or, equivalently, a *pseudo-metric*, on a set X is a function $d : X \times X \to \mathbb{R}$ such that: (1) $d(x,y) \geq 0$ for all $x, y \in X$; and (2) $d(x,y) = 0$ if and only if $x = y$.

Because the function defined above does not satisfy the triangle inequality of a distance, each such function is termed a *pseudo distance*. If a pseudo-distance d is in general not symmetric.

Definition 19.11.2. Define the set of functions,

$$F_{2s} = \left\{ \begin{array}{l} f : (0,\infty) \to \mathbb{R} |\ f \in C^2((0,\infty), \mathbb{R}),\ f(1) = 0, \\ (\forall x \in (0,\infty),\ \partial^2 f(x)/\partial x^2) > 0 \end{array} \right\}.$$

It follows from Proposition 17.7.3 that any function in the set F_{2s} is strictly convex on its domain of definition.

Definition 19.11.3. Given a measurable space (Ω, F), let

$$\mathbf{P}(\Omega, F) = \{P : F \to \mathbb{R}_+ |\ P \text{ is a probability measure }\}.$$

Define the pseudo-distance d_f by the formulas,

$$\begin{aligned}
&f \in F_{2s},\ d_f : \mathbf{P}(\Omega,F) \times \mathbf{P}(\Omega,F) \to \mathbb{R},\\
&Q \text{ a } \sigma\text{-finite measure, } P_a,\ P_b \in \mathbf{P}(\Omega,F),\\
&P_1 \ll Q \ \text{ with } \frac{dP_a}{dQ} \ = \ r_a,\ P_b \ll Q \ \text{ with } \frac{dP_b}{dQ} = r_b,\\
d_f(P_a,P_b) = E_Q[f(\frac{r_a}{r_b})r_b] = E_{P_b}[f(\frac{r_a}{r_b})],
\end{aligned}$$

where Q is a σ-finite measure on (Ω,F). The pseudo-distance d_f is also called the *f-information measure*, the *f-entropy*, or the *f-divergence.*

A σ-finite measure Q as mentioned above always exists, for example $Q = P_1 + P_2$ will do. An example of a σ-finite measure is the counting measure on the set of the natural numbers which assigns mass one to every natural number, $P_c : \mathbb{N} \to \mathbb{R}_+$, for all $k \in \mathbb{N}, P_c(k) = 1$. Another example of a σ-finite measure is the Lebesgue measure on $(\Omega,F) = (\mathbb{R},B(\mathbb{R}))$. Because $r_2 > 0$ *a.s.* P_2, the above expression is well defined.

Proposition 19.11.4. *(a)The function d_f defined in 19.11.3 is a pseudo-distance. (b)The pseudo-distance d_f does not depend on the choice of the σ-finite measure Q.*

Definition 19.11.5. The *Kullback-Leibler* pseudo-distance is defined by,

$$\begin{aligned}
f(x) = \begin{cases} x\ln(x), & x > 0,\\ 0, & x = 0; \end{cases} \ f : (0,\infty) \to \mathbb{R},\ f \in F_{2s},\\
&d_{KL} : \mathbf{P}(\Omega,F) \times \mathbf{P}(\Omega,F) \to \mathbb{R},\\
&Q \text{ a } \sigma\text{-finite measure, } P_a,\ P_b \in \mathbf{P}(\Omega,F),\\
&P_1 \ll Q \text{ with } \frac{dP_1}{dQ} = r_a,\ P_2 \ll Q \text{ with } \frac{dP_2}{dQ} = r_b,\\
d_{KL}(P_a,P_b) = E_{P_b}[\frac{r_a}{r_b}\ \ln(\frac{r_a}{r_b})] = E_Q[r_a \ln(\frac{r_a}{r_b})].
\end{aligned}$$

In information theory, the above defined pseudo-distance is called *divergence* and *relative entropy*, and it is denoted there by $D(P_a \| P_b)$.

Definition 19.11.6. The *Hellinger* pseudo-distance is defined as,

$$\begin{aligned}
f_H(x) = (\sqrt{x}-1)^2,\ f_H : \mathbb{R}_+ \to \mathbb{R}_+,\ d_H : \mathbf{P}(\Omega,F) \times \mathbf{P}(\Omega,F) \to \mathbb{R},\\
&Q \text{ a } \sigma\text{-finite measure, } P_a,\ P_b \in \mathbf{P}(\Omega,F),\\
&P_1 \ll Q \text{ with } \frac{dP_1}{dQ} = r_a,\ P_2 \ll Q \text{ with } \frac{dP_2}{dQ} = r_b,\\
d_H(P_a,P_b) = E_{P_b}[(\sqrt{\frac{r_a}{r_b}} - 1)^2] = E_Q[(\sqrt{r_a} - \sqrt{r_b})^2].
\end{aligned}$$

The Hellinger pseudo-distance is symmetric.

Definition 19.11.7. Consider a probability space (Ω,F). Define the *total variation pseudo-distance* by the formula,

$$d_{TV}(P_a,P_b) = \sup_{A \in F} |P_a(A) - P_b(A)|,\ d_{TV} : \mathbf{P} \times \mathbf{P} \to \mathbb{R}_+.$$

Note that no absolute continuity is needed, neither $P_a \ll P_b$ nor $P_b \ll P_a$, while absolute continuity of the two measures was required in the definition of the pseudo-distances introduced earlier in this section.

Proposition 19.11.8. Properties of the total-variation pseudo-distance.

(a)The function d_{TV} is a pseudo-distance.
(b)For all $P_a,\ P_b \in P(\Omega,F)$, $d_{TV}(P_a,P_b) \leq 1$.
(c)If there exists a σ-finite measure Q such that absolute continuity holds as in Def. 19.11.3 then the total variation pseudo-distance equals the L_1 distance, $d_{TV}(P_a,P_b) = \|P_a - P_b\|_1$.

Proof. (a) Positivity of the function d_{TV} follows from the definition.
If $d_{TV}(P_a,P_b) = 0$ then by definition, for all $A \in F$, $P_a(A) - P_b(A) = 0$ hence $P_a = P_b$.
(b) $|P_a(A) - P_b(A)| \leq \max\{P_a(A),\ P_b(A)\} \leq 1$.
(c) This follows from Proposition 19.11.9. □

Proposition 19.11.9. *Consider two probability measures induced by two probability density functions $p_a,\ p_b : \mathbb{R} \to \mathbb{R}_+$ on the real line. The corresponding measures are then defined by,*

$$P_a(A) = \int_A p_a(w)dw,\ P_b(A) = \int_A p_b(w)dw,\ A \in B(\mathbb{R}).$$

Then,

$$d_{TV}(P_a,P_b) = \int_{\mathbb{R}} |p_a(w) - p_b(w)|dw = 2\int_{\mathbb{R}} (p_a(w) - p_b(w))^+ dw.$$

Proof. Note that,

$$\begin{aligned}
1 &= \int_{\mathbb{R}} p_a(w)dw = \int_{\mathbb{R}} p_b(w)dw \ \Rightarrow \\
0 &= \int_{\mathbb{R}} (p_a(w) - p_b(w))dw \\
&= \int_{\{w\in\mathbb{R}|\ p_a(w)>p_b(w)\}} (p_a(w) - p_b(w))dw + \\
&\quad + \int_{\{w\in\mathbb{R}|\ p_a(w)\leq p_b(w)\}} (p_a(w) - p_b(w))dw \\
&= \int_{\{p_a>p_b\}} (p_a(w) - p_b(w))dw + \int_{\{p_a\leq p_b\}} (p_a(w) - p_b(w))dw, \\
&\int_{\{p_a\leq p_b\}} (p_a(w) - p_b(w))dw = -\int_{\{p_a>p_b\}} (p_a(w) - p_b(w))dw;
\end{aligned}$$

$$\begin{aligned}
&\int_{\mathbb{R}} |p_a(w) - p_b(w)| dw \\
&= \int_{\{p_a > p_b\}} (p_a(w) - p_b(w)) dw - \int_{\{p_a \le p_b\}} (p_a(w) - p_b(w)) dw \\
&= 2 \int_{\{p_a > p_b\}} (p_a(w) - p_b(w)) dw = 2 \int (p_a(w) - p_b(w))^+ dw; \\
&d_{TV}(P_a, P_b) = \sup_{A \in F} | \int_A p_a(w) dw - \int_A p_b(w) dw| \\
&= \sup | \int_{A \cap \{p_a > p_b\}} (p_a(w) - p_b(w)) dw + \int_{A \cap \{p_a \le p_b\}} (p_a(w) - p_b(w)) dw| \\
&\le \max\{ \int_{\{p_a > p_b\}} (p_a(w) - p_b(w)) dw, \ \int_{\{p_a \le p_b\}} (p_b(w) - p_a(w)) dw\} \\
&= \int_{\{p_a > p_b\}} (p_a(w) - p_b(w)) dw = \int_{\{p_a > p_b\}} |p_a(w) - p_b(w)| dw \\
&= 2 \int_{\mathbb{R}} (p_a(w) - p_b(w))^+ dw \\
&\le \sup_{A \in F} \int_{A \cap \{p_a > p_b\}} (p_a(w) - p_b(w)) dw = d_{TV}(P_a, P_b).
\end{aligned}$$

□

19.12 P-essential Infima

Needed in control theory is the concept of an infimum over a noncountable set of positive-valued random variables. The concept is called the P-essential infimum. It is defined below. The frame work was developed by C. Striebel, see the section *Further Reading* for the references.

Using the concept of $\mathrm{P}-essinf$ a more satisfactory formulation can be provided of the optimality conditions of optimal stochastic control.

Consider two real-valued random variables $x, y : \Omega \to \mathbb{R}$ on the measurable space (Ω, F). Then their minimum is defined by the expression,

$$\begin{aligned}
x \wedge y(\omega) &= \min\{x, y\}(\omega) \\
&= x(\omega) I_{\{\omega_1 \in \Omega | x(\omega_1) < y(\omega_1)\}} + y(\omega) I_{\{\omega_2 \in \Omega | x(\omega_2) \ge y(\omega_2)\}}.
\end{aligned}$$

Because x and y are random variables, the sets at the indicator functions of the above expression belong to the σ-algebra F. Hence $x \wedge y : \Omega \to \mathbb{R}$ is a random variable. Similarly, the infimum of a countable collection of random variables is a random variable.

However, when $\{x_i : \Omega \to \mathbb{R}, \ i \in I\}$, is a collection of random variables which is not countable then the above procedure cannot be applied to prove that the infimum is a random variable. A new concept is needed.

Definition 19.12.1. Consider a collection of positive real-valued random variables, $\{x_i : \Omega \to \mathbb{R}_+,\ i \in I\}$. The *P-essential infimum* of this collection is defined as the function

$$x = \mathrm{P} - essinf_{i \in I} x_i : \Omega \to \mathbb{R}_+, \text{ such that,}$$

1. $x : (\Omega, F) \to (\mathbb{R}_+, B(\mathbb{R}_+))$ is a random variable;
2. $x \leq x_i$ a.s. P for all $i \in I$; and
3. if $\bar{x} : \Omega \to \mathbb{R}_+$ is another function which satisfies (1) and (2) then $\bar{x} \leq x$ a.s. P.

The *P-essential supremum* of an upper bounded set of random variables is defined correspondingly.

A P-essential infimum of a collection of random variables is thus like the infimum for a subset of the real numbers except that the measurability and the probability measure enter into the definition.

Theorem 19.12.2. *For any family of positive real-valued random variables,* $\{x_i : \Omega \to \mathbb{R}_+,\ i \in I\}$*,*

(a)the P-essential infimum exists, is a positive random variable $x : \Omega \to \mathbb{R}_+$, and denoted by $x = \mathrm{P} - essinf_{i \in I} x_i$; and
(b)the P-essential infimum is unique upto an almost sure modification. Thus, if $\bar{x} : \Omega \to \mathbb{R}_+$ also satisfies the definition of the P-essential infimum then $x = \bar{x}$ a.s. P.

With respect to an additional condition, one can prove that the interchange of conditional expectation and the $\mathrm{P} - essinf$ is possible with equality. The reader is referred to the basic sources mentioned in next section.

19.13 Further Reading

See also the section *Further Reading* of Chapter 2. In this section references are stated to the theory presented only in this chapter.

Probability measures. A source on probability measures is the book [3].

Spaces and sequences of random variables. Proposition 19.5.4 is proven in [32]. Proposition 19.5.11 is proven in [29, II.4.2]. Proposition 19.5.16 on uniform integrability is related to the references [3, 7.5.3], [29, II.5.2], [29, II.5.1], and [25, II.T22]. Theorem 19.5.17 is related to [3, 7.5.2], [25, II.T21], and [29, II.5.4]. Theorem 19.5.6 is an exercise in the book [9, Exercise 2.1.11].

Conditional expectation and conditional probability. The concept of projection of one σ-algebra on another is used in [24, p. 343], [16, Section 4.3], and [36, Def. 2.1]. For a definition of the projection of a σ-algebra on another, see also the book [16, Section 4.3] which definition differs from the one used in this book.

A useful source for conditional probability is [6]. The problem has been investigated whether on a particular conditional probability space there exists an independent complement in the form of a σ-algebra, see [30]. A characterization is proven.

Conditional Gaussian random variables. An early reference is [22, Ch. 11]. The concepts were developed for the conditional Kalman filter for which see Section 8.9 and the paper [8] which paper deals with a slightly different system. The main definitions and results of this section were developed by the author, and the statements and proofs are different from those of the literature.

Conditional independence. Many results are adjusted from the papers and the report [26, 28, 27], and from the book [16, Section 2.2], The related work of C. van Putten etal. is [36]. Papers of interest include [10, 11, 13].

Conditional independence and finite probability spaces. Example 19.8.8 was adjusted from [37, Ex. 4.4].

Measure transformations. That one can relate two probability measures via a measure transformation is due to J. Radon and O.M. Nikodym. For continuous-time Wiener integrals this was developed [7]. A. Girsanov [18] formulated the version in terms of continuous-time Ito integrals. The extensions to continuous-time martingales is described in [38]. See also the book by I. Karatzas and S.E. Shreve, [20, Section 3.5]. For continuous-time processes see the books [22, 20]. Theorem 19.9.10 is adjusted from [23, 24.4]. Theorem19.9.11 is adjusted from [36, Th. 3.6].

For the family of *exponential probability distributions* see the book of Barndorff-Nielsen, [4]. For conjugate priors of members of exponential families, [2].

Metrics on probability spaces. Several metrics are described in the book by L. Devroye, [12]. See also the paper [17]. Def. 19.11.3 is adjusted from [1]. The total variation metric is described in [12, Sec. 1.1] and in papers of I. Tzortzis, [35, 34]. A metric on the set of finite probability spaces of different cardinalities is proposed by M. Vidyasagar, [39].

P-essential infima. A basic source is the book of C. Striebel, [33]. See also the book [5]. Theorem 19.12.2 is based on the Radon-Nikodym theorem. See for the proofs [33, Th. A.2.1] and [21, Th. A.3]. The result may be deduced from a theorem on linear operators, see [15, Ch. III 7.5, 7.6; Ch. IV 8.22].

References

1. N.L. Aggrawal. Sur l'information de fisher. In J. Kampe De Feriet and C.F. Picard, editors, *Théories de l'Information*, volume 398 of *Lecture Notes in Mathematics*, pages 111–117. Springer-Verlag, Berlin, 1974. 742
2. P. Diaconis anad D. Ylvisaker. Conjugate priors for exponential families. *Ann. Statist.*, 7:269–281, 1979. 352, 742
3. R.B. Ash. *Real analysis and probability*. Academic Press, New York, 1972. 12, 49, 741
4. O.E. Barndorff-Nielsen. *Information and exponential families in statistical theory*. Wiley, London, 1978. 353, 742
5. D.P. Bertsekas and S.E. Shreve. *Stochastic optimal control: The discrete time case*. Athena Scientific, Belmont, MA, 1996. 742
6. L. Breiman. *Probability*. Addison-Wesley Publ. Co., Reading, MA, 1968. 49, 73, 741, 758
7. R.H. Cameron and W.T. Martin. Transformation of Wiener integrals under translations. *Ann. Math.*, 45:386–396, 1944. 742
8. Han-Fu Chen, P.R. Kumar, and J.H. van Schuppen. On Kalman filtering for conditionally Gaussian systems with random matrices. *System & Control Lett.*, 13:397–404, 1989. 742

9. K.L. Chung. *A course in probability theory*. Academic Press, New York, 1974. 12, 49, 741, 758
10. A.P. Dawid. Some misleading arguments involving conditional independence. *J. Royal Statist. Soc., Series B*, 41:249–252, 1979. 742
11. A.P. Dawid. Conditional independence for statistical operations. *Ann. Math. Statist.*, 8:598–617, 1980. 49, 276, 742
12. L. Devroye. *A course in density estimation*. Birkhäuser Verlag, Basel, 1987. 9, 742
13. R. Döhler. On the conditional independence of random events. *Theory Probab. Appl.*, 25:628–634, 1980. 276, 742
14. J.L. Doob. *Stochastic processes*. Wiley, New York, 1953. 72, 73, 721, 747, 754, 758
15. N. Dunford and J.T. Schwartz. *Linear operators (3 volumes)*. Interscience, New York, 1958–1971. 742
16. J.P. Florens, M. Mouchart, and J.M. Rolin. *Elements of Bayesian statistics*. Routeldge, Baton Rouge, FL, 1990. 253, 276, 723, 741, 742
17. A.L. Gibbs and F.E. Su. On choosing and bounding probability metrics. *Int. Statist. Rev.*, 70:419–435, 2002. 742
18. I.V. Girsanov. On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. *Th. Probab. Appl.*, 5:285–301, 1960. 352, 604, 742
19. H. Hotelling. Relation between two sets of variates. *Biometrika*, 28:321–377, 1936. 710
20. I. Karatzas and S. Shreve. *Brownian motion and stochastic calculus*. Springer-Verlag, Berlin, 1988. 742
21. I. Karatzas and S.E. Shreve. *Methods of mathematical finance*. Number 39 in Applications of Mathematics. Springer, Berlin, 1998. 411, 605, 742
22. R.S. Liptser and A.N. Shiryayev. *Statistics of random processes: I. General theory; II. Applications*. Springer-Verlag, Berlin, 1977,1978. 49, 742
23. M. Loève. *Probability theory, 3rd edition*. Van Nostrand Reinhold Co. Inc., New York, 1963. 49, 72, 742
24. H.P. McKean Jr. Brownian motion with a several dimensional time. *Theory Probab. Appl.*, 8:335–354, 1963. 276, 741
25. P.A. Meyer. *Probability and Potentials*. Blaisdell Publishing Company, Waltham, MA, 1966. 49, 336, 741, 758
26. M. Mouchart and J.-M. Rolin. A note on conditional independence (with statistical applications). Report 129, Institut de Mathématique Pure et Appliquée, Université Catholique de Louvain, Louvain-la-Neuve, 1979. 49, 276, 723, 742
27. M. Mouchart and J.-M. Rolin. A note on conditional independence with statistical applications. *Statistica*, 44:557–584, 1984. 49, 276, 723, 742
28. M. Mouchart and J.-M. Rolin. On the σ-algebraic realization problem. Report 8604, Center for Operations Research and Econometrics, Université de Louvain, Louvain-la-Neuve, 1986. 253, 276, 723, 742
29. J. Neveu. *Mathematical foundations of the calculus of probability*. Holden-Day Inc., San Francisco, 1965. 49, 741
30. D. Ramachandran. Existence of independent complements in regular conditional probability spaces. *Ann. Probab.*, 7:433–443, 1979. 741
31. L.A. Shepp. Normal functions of normal random variables. *SIAM Rev.*, 6:459–460, 1964. 713
32. L.A. Shepp and A.M. Odlyzko. A probabilistic inequality. *SIAM Review*, 21:564–565, 1979. 741
33. C. Striebel. *Optimal control of discrete time stochastic systems*, volume 110 of *Lecture Notes in Economic and Mathematical Systems*. Springer-Verlag, Berlin, 1975. 431, 468, 575, 595, 603, 605, 742
34. Ioannis Tzortzis, Charalambos D. Charalambous, and Themistokles Charalambous. Infinite horizon average cost dynamic programming subject to total variation distance ambiguity. *SIAM J. Control & Opt.*, 57:2843–2872, 2019. 526, 742

35. Ioannis Tzortzis, Charalambos D. Charalambous, and Themistoklis Charalambous. Dynamic programming subject to total variation distance ambiguity. *SIAM J. Control & Opt.*, 53:2040–2075, 2015. 742
36. C. van Putten and J.H. van Schuppen. Invariance properties of the conditional independence relation. *Ann. Probab.*, 13:934–945, 1985. 49, 276, 723, 741, 742
37. J.H. van Schuppen. The strong finite stochastic realization problem - Preliminary results. In A. Bensoussan and J.L. Lions, editors, *Analysis and optimization of systems*, volume 44 of *Lecture Notes in Control and Information Sciences*, pages 179–190, Berlin, 1982. Springer-Verlag. 120, 276, 742
38. J.H. van Schuppen and E. Wong. Transformations of local martingales under a change of law. *Ann. Probab.*, 2:879–888, 1974. 352, 742
39. M. Vidyasagar. A metric between probability distributions on finite sets of different cardinalities and applications to order reduction. *IEEE Trans. Automatic Control*, 57:2464–2477, 2012. 742

Chapter 20
Appendix D Stochastic Processes

Abstract Specialized topics of the theory of stochastic processes are described which are used in the body of this book. Defined are a filtration and stochastic processes relative to a filtration. Elementary martingale theory is discussed. Stopping times and a stochastic process indexed by a stopping time are defined. The supermartingale convergence theorem is established. A brief introduction to ergodicity is provided.

Key words: Stochastic process. Filtration. Martingale. Stopping time.

The reader finds in this chapter several concepts and theorems of the theory of stochastic processes which are less elementary than those described in Chapter 3. The character of this appendix is that of an introduction to various topics with concepts and elementary results.

20.1 Stochastic Processes and Filtrations

In modeling of phenomena of signals by stochastic processes it is important to be able to denote at any time the information available about the past and the future of the stochastic process. The appropriate concept for this is a filtration, a collection of σ-algebras.

Definition 20.1.1. A *filtration* is a family of σ-algebras $\{F_t, t \in T\}$ on $T = \mathbb{N} = \{0, 1, \ldots\}$ such that:

(a) for all $t \in T$, $F_t \subseteq F$; or, equivalently, F_t is a sub-σ-algebra of F;
(b) the family is increasing: for all $s, t \in T$, $s < t$ implies that $F_s \subseteq F_t$;
(c) F_0 contains all negligible subset of Ω.

Define $F_\infty = \sigma(\cup_{t \in T} F_t)$ as the smallest σ-algebra containing for all times $t \in T$, F_t.

Definition 20.1.2. *Filtration generated by a stochastic process.* Consider a stochastic process $x : \Omega \times T \to \mathbb{R}^n$. Define the filtration,

$$\{F_t^x, t \in T\}, \;\; F_t^x = \sigma(\{x(s), \forall s \in [0,t]\} \vee N), \; \forall t \in T,$$

where $N \subseteq \text{Pwrset}(\Omega)$ contains all negligigle subsets of Ω. Thus, F_t^x is the smallest σ-algebra with respect to which $x(s)$ for all $s \in [0,t]$ are measurable. Clearly, F_t^x is a sub-σ-algebra of F and the family is increasing. Hence it is a filtration. It will be called the *filtration generated by the stochastic process x*.

A filtration can also be defined without being generated by a stochastic process in which case the relation between a process and a filtration needs to be defined.

Definition 20.1.3. Consider a probability space (Ω, F, P), a measurable space (X, G), and a filtration $\{F_t, t \in T\}$.

(a) The stochastic process $x : \Omega \times T \to \mathbb{R}^n$ is said to be an *adapted process* to the filtration $\{F_t, t \in T\}$ if for all $t \in T$, $x(t)$ is an F_t measurable random variable. Thus, if for all $t \in T$ and for all $A \in B(\mathbb{R}^n)$, $\{\omega \in \Omega | \; x(\omega, t) \in A\} \in F_t$. Equivalently also, for all $t \in T$, $F^{x(t)} \subseteq F_t$. Denote by $\{x(t), F_t, t \in T\}$ that the process x is adapted to the displayed filtration.

(b) The stochastic process $x : \Omega \times T \to \mathbb{R}^n$ is said to be a *predictable process* with respect to the filtration if $x(0) \in \mathbb{R}^n$ is deterministic and for all $t \in T$, $x(t+1)$ is F_t measurable.

Example 20.1.4. Any stochastic process $x : \Omega \to \mathbb{R}^n$ is adapted to the filtration generated by this process, $\{x(t), F_t^x, t \in T\}$.

Definition 20.1.5. Consider a filtration and two stochastic process,

$$\begin{aligned}
&\{x(t), F_t, t \in T\}, \; x : \Omega \times T \to \mathbb{R}, \text{ adapted,} \\
&\{h(t), F_t, t \in T\}, \; h : \Omega \times T \to \mathbb{R}, \text{ predictable.}
\end{aligned}$$

Define the *transform process* of x by h as the stochastic process,

$$\begin{aligned}
&\{(h.x)(t), F_t, t \in T\}, \; (h.x) : \Omega \times T \to \mathbb{R}, \\
&(h.x)(t) = h_0 x_0 + \sum_{s=1}^{t} h(s)(x(s) - x(s-1)),
\end{aligned}$$

which by definition is an adapted process.

Proposition 20.1.6. *Consider an adapted process $\{x(t), F_t^x, t \in T\}$ with $T = \mathbb{N}$ and the stochastic process y which is a modification of x. Then $\{y(t), F_t, t \in T\}$ is an adapted process.*

Proof. For $t \in T$ and $A \in B(\mathbb{R}^n)$,

$$\begin{aligned}
& \{\omega \in \Omega | y(\omega,t) \in A\} \\
= & \; (\{\omega \in \Omega | x(\omega,t) \in A\} \cap \{\omega \in \Omega | x(\omega,t) = y(\omega,t)\}) \cup \\
& \cup (\{\omega \in \Omega | y(\omega,t) \in A\} \cap \{\omega \in \Omega | x(\omega,t) \neq y(\omega,t)\}), \\
& P(\{\omega \in \Omega | x(\omega,t) \neq y(\omega,t)\}) = 0, \\
\Rightarrow & \; \{x(\omega,t) \neq y(\omega,t)\} \in F_0 \subseteq F_t, \text{ because negligible sets belong to } F_0 \subseteq F_t; \\
\Rightarrow & \; \{y(\omega,t) \in A\} \cap \{x(\omega,t) \neq y(\omega,t)\} \in F_t; \\
& \{\omega \in \Omega | x(\omega,t) \in A\} \cap \{\omega \in \Omega | x(\omega,t) = y(\omega,t)\} \in F_t, \\
& \{\omega \in \Omega | y(\omega,t) \in A\} \\
= & \; (\{\omega \in \Omega | x(\omega,t) \in A\} \cap \{\omega \in \Omega | x(\omega,t) = y(\omega,t)\}) \cup \\
& \cup (\{y(\omega,t) \in A\} \cap \{x(\omega,t) \neq y(\omega,t)\}) \in F_t.
\end{aligned}$$

□

20.2 Martingale Theory

The reader finds in this section the elementary concepts of martigale theory. This is complemented in Section 20.4 by convergence theory of supermartingales.

The term martingale was introduced into mathematics by J.L. Doob, see [3]. The term was used in France for a gambling procedure in horse races. It is also used for part of the briddle of a horse. A martingale is a generalization of the concept of a stochastic process with independent increments. The theory of martingales admits a relative simple proof for almost sure convergence seen as an extension of processes with independent increments.

Definition 20.2.1. Consider a stochastic process $m : \Omega \times T \to \mathbb{R}^n$ and a filtration $\{F_t, t \in T\}$. The process is called a *martingale* with respect to the filtration if,

(a)$\{m(t), F_t, t \in T\}$ is adapted;
(b)m is integrable: for all $t \in T$, $E|m(t)| < \infty$;
(c)$\forall \, s,t \in T, \, s \leq t \;\Rightarrow E[m(t)|F_s] = m(s), \; a.s. \; P.$

The process $\{x(t), F_t, t \in T\}$ is called a *supermartingale* respectively a *submartingale* if (a) and (b) above hold and if

$$\begin{aligned}
(c1) \;\; & \forall \, s,t \in T, \; s \leq t \;\Rightarrow\; E[x(t)|F_s] \leq x(s), \; a.s. \; P, \\
(c2) \;\; & \forall \, s,t \in T, \; s \leq t \;\Rightarrow\; E[x(t)|F_s] \geq x(s), \; a.s. \; P.
\end{aligned}$$

Notation. $\{x(t), F_t, t \in T\} \in M_1, \; M_{1u}, \; SubM_1, SupM_1$, if x is respectively a martingale, a uniformly-integrable martingale, a submartingale, or a supermartingale. Such a process is called L_1-*bounded* if in addition $\sup_{t \in T} E|x(t)| < \infty$.

A discrete-time adapted integrable process $\{m(t), F_t, t \in T\}$ is a martingale if and only if $E[m(t+1)|F_t] = m(t), \;\; \forall t \in T$. The proof is a simple induction argument. An analogous condition holds for sub- and supermartingales. If $\{x(t), F_t, t \in T\} \in$

$SupM_1$ then $\{-x(t), F_t, t \in T\} \in SubM_1$ and conversely. Therefore in the theory, most results are stated for supermartingales.

Proposition 20.2.2. *Consider a sequence of independent random variables $v : \Omega \times T \to \mathbb{R}$ on $T = \mathbb{N}$. and a random variable $x_0 : \Omega \to \mathbb{R}$. Assume that F^{x_0} and F^v_∞ are independent. Define the process and the filtration,*

$$x(t) = x_0 + \sum_{s=1}^{t} v(s), \;\; F_t = F^{x_0} \vee F^v_t \vee N, \;\; \{F_t, t \in T\}.$$

Assume that $E|x_0| < \infty$, and for all $s \in T$, $E|v(s)| < \infty$ and $E[v(s)] = 0$. Then $\{x(t), F_t, t \in T\} \in M_1$

Proof. Because of the assumptions, $\{x(t), F_t, t \in T\}$ is adapted, for all $t \in T = \mathbb{N}$, $E|x(t)| < \infty$, and

$$\begin{aligned} & E[x(t+1) - x(t)|F_t] = E[v(t+1)|F_t] = E[v(t+1)], \\ & \quad \text{because } v \text{ is a sequence of independent random variables,} \\ & \quad \text{and } F^{v(t+1)} \text{ is independent of } F_t = F^{x_0} \vee F^v_t \text{ are independent,} \\ & = 0, \;\text{ by assumption, } \Rightarrow \; E[x(t+1)|F_t] = x(t). \end{aligned}$$

□

A corresponding result holds if $x : \Omega \times T \to \mathbb{R}^n$ and if x has independent increments.

Proposition 20.2.3. *Consider a filtration $\{F_t, t \in T\}$ and an integrable random variable $x_\infty : \Omega \to \mathbb{R}$. Define the process $x : \Omega \times T \to \mathbb{R}$, $x(t) = E[x_\infty|F_t]$, $\forall t \in T$. Then $\{x(t), F_t, t \in T\} \in M_1$.*

Proof. $\{x(t), F_t, t \in T\}$ is adapted, for all $t \in T$,

$$E|x(t)| = E|E[x_\infty|F_t]| \le E[E[|x_\infty||F_t]] = E|x_\infty| < \infty.$$

$$E[x(t+1)|F_t] = E[E[x_\infty|F_{t+1}]|F_t] = E[x_\infty|F_t] = x(t) \text{ by Theorem 2.8.2.(d).}$$

□

Proposition 20.2.4. *Consider a submartingale $\{x(t), F_t, t \in T\} \in SubM_1$. Then x is a L_1-bounded submartingale if and only if $\sup_{t \in T} E[x^+(t)] < \infty$.*

Proof.

$$\begin{aligned} \Leftarrow \quad & E|x(t)| = E[x^+(t) + x^-(t)] = E[2x^+(t) - (x^+(t) - x^-(t))] \\ & = 2E[x^+(t)] - E[x(t)] \le 2E[x^+(t)] - E[x(1)], \\ & \text{because } x \in SubM_1 \text{ implies that } E[x(t)] \ge E[x(1)], \\ & \sup_{t \in T} E|x(t)| \le \sup_{t \in T} E[x^+(t)] - E[x(1)] < \infty; \\ \Rightarrow \quad & E[x^+(t)] = E[x^+(t) + x^-(t) - x^-(t)] = E|x(t)| - E[x^-(t)] \le E|x(t)|, \\ & \sup_{t \in T} E[x^+(t)] \le \sup_{t \in T} E|x(t)| < \infty. \end{aligned}$$

□

Proposition 20.2.5. *Consider a filtration* $\{F_t, t \in T\}$.

(a)The tuple $(M_1(\Omega, F, P, \{F_t, t \in T\}, \mathbb{R}^n), \mathbb{R}, +, \times, 1, 0)$ *is a vector space.*
(b)The tuple $(SupM_1(\Omega, F, P, \{F_t, t \in T\}, \mathbb{R}^n), \mathbb{R}_+, +, \times, 1, 0)$ *is a vector space over the semi-ring* $\mathbb{R}_+$. *The same conclusion holds for positive submartingales.*

Proposition 20.2.6. Martingales and convexity. *Consider a filtration* $\{F_t, t \in T\}$*, an adapted integrable stochastic process* $x : \Omega \times T \to \mathbb{R}$*,* $\{x(t), F_t, t \in T\}$*, and a Borel measurable function* $f : \mathbb{R} \to \mathbb{R}$ *such that for all* $t \in T$*,* $E|f(x(t))| < \infty$.

(a)If $\{x(t), F_t, t \in T\} \in M_1$ *and if* f *is convex then* $\{f(x(t)), F_t, t \in T\} \in SubM_1$.
(b)If $\{x(t), F_t, t \in T\} \in SubM_1$*,* f *is convex and increasing then* $\{f(x(t)), F_t, t \in T\} \in SubM_1$.
(c)If $\{x(t), F_t, t \in T\} \in SupM_1$ *and if* f *is concave and increasing then* $\{f(x(t)), F_t, t \in T\} \in SupM_1$.

Proof. The proof uses Jensen's inequality for conditional expectation.
(a) For all $t \in T$,

$$\begin{aligned} E[f(x(t+1))|F_t] &\geq f(E[x(t+1)|F_t]), \text{ by Proposition 19.6.2,} \\ &= f(x(t)), \text{ because } x \in M_1. \end{aligned}$$

(b) For all $t \in T$, in addition to (a),

$$E[f(x(t+1))|F_t] \geq f(E[x(t+1)|F_t]) \geq f(x(t)),$$

because $x \in SubM_1$ implies that $E[x(t+1)|\ F_t] \geq x(t)$ and because f is increasing.
(c)

$$\begin{aligned} &E[f(x(t))|F_t] \leq f(E[x(t+1)|F_t]), \text{ by Proposition 19.6.2.} \\ &\leq f(x(t)), \end{aligned}$$

because $x \in SupM_1$ implies that $E[x(t+1)|\ F_t] \leq x(t)$, and because f is increasing. □

Corollary 20.2.7. *(a)If* $\{x(t), F_t, t \in T\} \in SubM_1$ *then* $\{x^+(t), F_t, t \in T\} \in SubM_1$.
(b)If $\{m(t), F_t, t \in T\} \in M_1$ *then* $\{|m(t)|, F_t, t \in T\} \in SubM_1$*; and if in addition, for all* $t \in T$*,* $|m(t)| > 0$*, then* $\{|m(t)| \ln^+(|m(t)|), F_t, t \in T\} \in SubM_1$.
(c)If $\{x(t), F_t, t \in T\} \in SubM_1$*, for a* $p \in [1, \infty)$ *and for all* $t \in T$*,* $E|x(t)|^p < \infty$ *then* $\{|x(t)|^p, F_t, t \in T\} \in SubM_1$.

Proof. This follows directly from Proposition 20.2.6 with the following functions and their properties. One may use Proposition 17.7.3 to prove that a function is convex.
(a) The function $f(x) = x^+$, $f : \mathbb{R} \to \mathbb{R}_+$ is convex and increasing (though not strictly convex).
(b) The function $f_1(x) = |x|$ is convex hence one obtains that $|m| \in SubM_1$ and with $f_2(x) = x\ln(x)$, $f_2 : (0, \infty) \to \mathbb{R}$ convex and increasing one obtains the result.
(c) The function $f_3(x) = |x|^p$ is convex. □

Proposition 20.2.8. *If $x, y \in SupM_1$ then $x \wedge y \in SupM_1$.*

Proof. For all $t \in T$,

$$\begin{aligned} E[x(t+1) \wedge y(t+1)|F_t] \ &\leq \ E[x(t+1)|F_t] \leq x(t), \text{ similarly, } \leq y(t), \\ &\Rightarrow E[x(t+1) \wedge y(t+1)|F_t] \leq x(t) \wedge y(t). \end{aligned}$$

□

Theorem 20.2.9. *Consider a filtration $\{F_t,\ t \in T\}$.*

(a)If $\{m(t), F_t, t \in T\} \in M_1$, $m : \Omega \times T \to \mathbb{R}$, $\{h(t), F_t, t \in T\}$, $h : \Omega \times T \to \mathbb{R}$, is predictable, and if $E|(h.m)(t)| < \infty$, then the martingale transform *is a martingale, $\{(h.m)(t), F_t, t \in T\} \in M_1$.*

(b)If $\{x(t), F_t, t \in T\} \in SupM_1$, $x : \Omega \times T \to \mathbb{R}$, $\{h(t), F_t, t \in T\}$, $h : \Omega \times T \to \mathbb{R}_+$, is positive and predictable, and if $E|(h.x)(t)| < \infty$, then the supermartingale transform *is a supermartingale, $\{(h.x)(t), F_t, t \in T\} \in SupM_1$.*

Proof.

$$\begin{aligned} &(a) \ \ E[(h.m)(t+1) - (h.m)(t)|F_t] = h(t+1)E[m(t+1) - m(t)|F_t] = 0; \\ &(b) \ \ E[(h.x)(t+1) - (h.x)(t)|F_t] = h(t+1)E[x(t+1) - x(t)|F_t] \leq 0. \end{aligned}$$

□

20.3 Stochastic Processes and Stopping Times

The time a stochastic process enters a particular subset in the range space depends on the value of $\omega \in \Omega$. Therefore the time is a random variable. The special case where the random variable is also related to a filtration deserves a special name, that of a stopping time. The concept of a stopping time is best contrasted with that of a random time.

Definition 20.3.1. (a)A *random time* is a random variable $\tau : \Omega \to \mathbb{N} \cup \{\infty\}$; hence,

$$A \in \mathbb{N} \cup \{\infty\} \ \Rightarrow \ \{\omega \in \Omega | \tau(\omega) \in A\} \in F.$$

(b)A random time $\tau : \Omega \to T \cup \{\infty\}$ is said to be a *stopping time* of the filtration if,

$$\forall\, t \in T \ \{\omega \in \Omega | \tau(\omega) \leq t\} \in F_t.$$

Equivalently, if for all $t \in T$, $\{\tau = t\} \in F_t$.

Denote by $T_{st}(\{F_t,\ t \in T\})$ the *set of all stopping times* with respect to the listed filtration. If the filtration is clear then it may be deleted from the notation.

(c)Let $x : \Omega \times T \to \mathbb{R}^n$ be a stochastic process and $\tau : \Omega \to \mathbb{N} \cup \{\infty\}$ be a random time. Define the process at this random time as the function,

$$\begin{aligned} &x(\tau) : \{\tau < \infty\} \to \mathbb{R}^n, \\ &x(\omega, \tau) = x(\omega, \tau(\omega)), \text{ if } \omega \in \{\omega_1 \in \Omega | \tau(\omega_1) < \infty\}. \end{aligned}$$

The function $x(\tau)$ is not defined on the set $\{\omega \in \Omega | \tau(\omega) = \infty\}$. If there exists a random variable $x_\infty : \Omega \to \mathbb{R}^n$ then define the function

$$x(\omega, \tau) = \begin{cases} x(\omega, \tau(\omega)), & \text{if } \omega \in \{\omega_1 \in \Omega | \tau(\omega_1) < \infty\}, \\ x_\infty(\omega), & \text{if } \omega \in \{\omega_1 \in \Omega | \tau(\omega_1) = \infty\}. \end{cases}$$

A random time which is equal to a constant, say $\tau(\omega) = s \in \mathbb{N}$, is a stopping time of any filtration which follows directly from the definitions.

Proposition 20.3.2. *If x is a stochastic process and τ is a random time then $x(\tau)$ is a random variable.*

Proof. This follows from the composition of measurable maps,
$\omega \mapsto (\omega, \tau(\omega)) \mapsto x(\omega, \tau(\omega))$. □

Example 20.3.3. Consider the price of a stock. A mathematical model for this price is a stochastic process, $x : \Omega \times T \to \mathbb{R}_+$. Suppose that the process is adapted to a filtration, $\{x(t), F_t, t \in T\}$. It may be of interest to formulate the first time the price of the stock exceeds a particular value, $c \in \mathbb{R}_+$,

$$\tau_c(\omega) = \begin{cases} \inf\{t \in T | x(\omega, t) > c\}, & \text{if set not empty,} \\ +\infty & \text{if set empty.} \end{cases}$$

It will then be useful if for any $t \in T$, the event $\{\omega \in \Omega | \tau(\omega) \leq t\}$ belongs to F_t. Note that the set denotes those $\omega \in \Omega$ for which the price has exceeded the value $c \in \mathbb{R}_+$ in the interval $[0, t]$.

Definition 20.3.4. Consider a filtration $\{F_t, t \in T\}$ and a stochastic process x, adapted $\{x(t), F_t, t \in T\}$. For the subset $A \in B(\mathbb{R}^n)$, define the *hitting time* as the function

$$h_A(\omega) = \begin{cases} \inf\{t \in T | x(\omega, t) \in A\}, & \text{if set not empty,} \\ +\infty, & \text{otherwise.} \end{cases}$$

In the following, a hitting time will be referred to by the indicated set and it will be assumed that, in case the set is empty, the value $+\infty$ is assigned to the hitting time.

Proposition 20.3.5. *Consider a filtration, an adapted process $\{x(t), F_t, t \in T\}$, and a set $A \in B(\mathbb{R}^n)$. Then the hitting time h_A is a stopping time of the filtration.*

Proof. For $t \in T$, $\{\omega \in \Omega | h_A(\omega_) \leq t\} = \cup_{s=0}^t \{x(s) \in A\} \in F_t$, because for all $s \in T$, $s \leq t$, $\{x(s) \in A\} \in F_s \subseteq F_t$, the process being adapted. □

Definition 20.3.6. Consider the stochastic process $x : \Omega \to \mathbb{R}$ with $\{x(t), F_t, t \in T\}$ adapted. Define the *last exit time* of x with respect to the set $A \in B(\mathbb{R})$ as the function, $\tau_e : \Omega \to T \cup \{+\infty\}$

$$\tau_e = \begin{cases} \sup\{t \in T | x(\omega,t) \in A\}, & \\ +\infty & \text{if } x(\omega,t) \in A \text{ infinitely often.} \end{cases}$$

Infinitely often here means that for all $t \in T$ there exists an $s \in (t,\infty)$ such that $x(\omega,s) \in A$.

Proposition 20.3.7. *Consider a filtration and an adapted process,* $\{x(t), F_t, t \in T\}$.

(a)The last exit time, τ_e*, is a random time.*
(b)In general, the last exit time τ_e *is not a stopping time of the filtration.*

Proof. (b) Note that

$$\{\tau_e \leq t\} = \cap_{s \in (t,+\infty)} \{x(s) \notin A\} \in F_t^{x+} = \sigma(\{x(r), \forall r \in (t,+\infty)\}),$$

which set in general does not belong to F_t. □

Proposition 20.3.8. *If* τ *is a stopping time of the filtration* $\{F_t, t \in T\}$ *and if* $k \in \mathbb{Z}_+$ *then* $\tau + k$ *is also a stopping time.*

Proof. $\forall\, t \in T$ and for all $k \in \mathbb{Z}_+$,

$$\{\tau + k \leq t\} = \begin{cases} \emptyset \in F_t, & t < k, \\ \{\tau \leq t - k\} \in F_{t-k} \subseteq F_t, & t \geq k. \end{cases}$$

□

Proposition 20.3.9. *If* τ, s *are stopping times then so are* $\tau \wedge s$, $\tau \vee s$*, and* $\tau + s$.

Proof. Let $t \in T$. Then

$$\begin{aligned} \{\tau \wedge s \leq t\} &= \{\tau \leq t\} \cup \{s \leq t\} \in F_t, \\ \{\tau \vee s\} &= \{\tau \leq s\} \cap \{s \leq t\} \in F_t. \\ \{\tau + s \leq t\}^c &= \{\tau + s > t\} \\ &= \{\tau = 0,\ s > t\} \cup \{0 < \tau < t,\ \tau + s > t\} \\ &\quad \cup \{\tau > t,\ s = 0\} \cup \{\tau \geq t,\ s > 0\}. \end{aligned}$$

The first, third, and the fourth event are in F_t. Then,

$$\begin{aligned} &\{0 < \tau < t,\ \tau + s > t\} \\ &= \cup_{r \in \mathbb{Z}_+, 0<r<t} \{r < \tau < t\} \cap \{s > t - r\}, \\ &\{s > t - r\} = \{s \leq t - r\}^c \in F_{t-r} \subseteq F_t, \\ &\{r < \tau < t\} = \{r < \tau\} \cap \{\tau < t\} = \{\tau \leq r\}^c \cap \{\tau < t\} \in F_r \cap F_t \subseteq F_t. \end{aligned}$$

□

Proposition 20.3.10. *Let* $\{\tau_n, n \in \mathbb{Z}_+\}$ *be stopping times. Then the following expressions are also stopping times,*

$$\sup_{n \in \mathbb{Z}_+} \tau_n, \quad \inf_{n \in \mathbb{Z}_+} \tau_n, \quad \limsup_{n \in \mathbb{Z}_+} \tau_n, \quad \liminf_{n \in \mathbb{Z}_+} \tau_n.$$

Proof.

$$\{\sup_{n \in \mathbb{Z}_+} \tau_n \leq t\} = \cap_n \{\tau_n \leq t\} \in F_t; \; \{\inf_{n \in \mathbb{Z}_+} \tau_n < t\} = \cup \{\tau_n < t\} \in F_t.$$
$$\{\sup \tau_n < t\}^c = \{\sup \tau_n \geq t\} = \cup \{\tau_n \geq t\} \in F_t.$$

□

Definition 20.3.11. Consider the filtration $\{F_t, t \in T\}$ and a stopping time τ of this filtration. Define the σ-algebra F_τ of *events prior to* τ as the collection,

$$F_\tau = \{A \in F | \forall t \in T, \, A \cap \{\tau \leq t\} \in F_t\}. \tag{20.1}$$

Denote the *filtration indexed by stopping times* by $\{F_\tau, \, \forall \, \tau \in T_{st}\}$.

The definition of F_τ is motivated by the need for a σ-algebra family of information up to a stopping time τ. If $A \in F$ and at time $t \in T$, τ is observed, hence $\{\tau \leq t\} = \Omega$, then one must be able to determine whether the event A has occured on the basis of F_t only. Thus one wants that, $A \cap \{\tau \leq t\} \in F_t$ and $A^c \cap \{\tau \leq t\} \in F_t$. Note that,

$$A \cap \{\tau \leq t\} \in F_t \; \Rightarrow \; A^c \cap \{\tau \leq t\} = \{\tau \leq t\} \cap (A \cap \{\tau \leq t\})^c \in F_t.$$

Hence it suffices to require only the first of these two conditions.

Proposition 20.3.12. *Consider a filtration $\{F_t, t \in T\}$ and a stopping time τ of this filtration.*

(a)F_τ is a σ-algebra.
(b)τ is F_τ measurable.
(c)If $\tau(\omega) = t \in T$ for all $\omega \in \Omega$ then $F_\tau = F_t$.

Proposition 20.3.13. *Let τ, s be stopping times of the filtration $\{F_t, t \in T\}$.*

(a)If $A \in F_s$ then $A \cap \{s \leq \tau\} \in F_\tau$.
(b)If $s(\omega) \leq \tau(\omega)$ for all $\omega \in \Omega$ then $F_s \subseteq F_\tau$.
(c)$F_{\tau \wedge s} = F_\tau \cap F_s$.
(d)

$$\{\tau < s\}, \; \{\tau > s\}, \; \{\tau \leq s\}, \; \{\tau \geq s\}, \; \{\tau = s\} \in F_\tau \cap F_s.$$

Definition 20.3.14. Consider an adapted process $\{x(t), F_t, t \in T\}$ and a stopping time $\tau \in T_{st}$ of the filtration. Define the *process x stopped at the stopping time τ* as the stochastic process,

$$x^\tau(\omega, t) = x(\omega, \tau(\omega) \wedge t), \; x^\tau : \Omega \times T \to \mathbb{R}.$$

Example 20.3.15. If $\{x(t), F_t, t \in T\}$ is an adapted process and $\tau \in T_{st}$ is a stopping time then the stopped process x^τ is a process transform,

$$\begin{aligned} x^{\tau} &= (h.x), \ \ h(t) = I_{\{t \leq \tau\}}, \text{ because,} \\ (h.x)(t) &= h_0 x_0 + \sum_{s=1}^{t} I_{\{s \leq \tau\}}(x(s) - x(s-1)) = x(\tau \wedge t) = x^{\tau}(t). \end{aligned}$$

Corollary 20.3.16. *If $\{x(t), F_t, t \in T\} \in M_1$, respectively $SupM_1$, $\tau \in T_{st}$ then $x^{\tau} = \{x^{\tau}(t), F_t, t \in T\} \in M_1$ respectively $SupM_1$.*

Proof.

$$\begin{aligned} x^{\tau} &= (h.x)(t), \ h(t) = I_{\{t \leq \tau\}}, \ \ \{t \leq \tau\} = \{t > \tau\}^c = \{\tau \leq t-1\}^c \in F_{t-1}, \\ & E|x^{\tau}(t)| \leq \sum_{s=0}^{t} E|x(s)| < \infty, \end{aligned}$$

hence the results follows from Theorem 20.2.9. □

20.4 Supermartingale Convergence

Theorem 20.4.1. Doob's optional sampling theorem - Bounded stopping time case. *Consider a supermartingale respectively a martingale, $\{x(t), F_t, t \in T\} \in SupM_1$ respectively $\in M_1$, $x : \Omega \times T \to \mathbb{R}$, $s, \tau \in T_{st}$, there exists $t_1 \in T$, such that $\tau \leq t_1$. Then $x(s), x(\tau) \in L_1$, and $\{x(t), F_t, t \in \{s, \tau\}\} \in SupM_1, \ \in M_1$ respectively.*

The proof of the convergence of a supermartingale makes use of the concept of an uncrossing random variable as explained next. The concept was used by J.L. Doob and is described in his book, [3].

Consider a stochastic process $x : \Omega \times T \to \mathbb{R}$. Suppose there exists a time $t \in T$ such that a.s. $- \lim_{s \to \infty} x(s)$ does not exist. Then, by the definition of a limit, $\liminf_{s \to \infty} x(s) < \limsup_{s \to \infty} x(s)$. Therefore, there exist rational numbers a, $b \in \mathbb{Q}$ such that $\liminf x(s) < a < b < \limsup x(s)$ and the number of upcrossings from below a to above b has to be infinite because of the concept of lim inf and lim sup.

It will be proven below that for a submartingale x the expected number of upcrossings has a finite expectation and therefore is finite. Therefore, the supposition that the limit does not exist is not correct hence the almost-sure limit exists.

Theorem 20.4.2. Submartingale inequalities. *Consider the submartingale $x : \Omega \times T \to \mathbb{R}$, $\{x(t), F_t, t \in T\} \in SubM_1$, $T = \{0, 1, \ldots, t_1\}$.*

(a)For all $\lambda \in \mathbb{R}$,

$$\lambda P(\{\max_{s \in T} x(s) \geq \lambda\}) \leq E[I_{(\max_{s \in T} x(s) \geq \lambda)} x(t_1)] \leq E[x(t_1)^+].$$

(b)For all $\lambda \in \mathbb{R}$,

$$\begin{aligned} \lambda P(\{\min_{s \in T} x(s) \leq -\lambda\}) &\leq E[x(t_1) - x(0)] - E[I_{(\min_{s \in T} x(s) \leq -\lambda)} x(t_1)] \\ &\leq E[x^+(t_1)] - E[x(0)]. \end{aligned}$$

Theorem 20.4.3. *Consider the L_1-bounded submartingale $\{x(t), F_t, t \in T\} \in SubM_1$. Then*

$$\text{a.s.} - \lim_{t\to\infty} x(t) = x(\infty), \;\; x(\infty) : \Omega \to \mathbb{R}. \tag{20.2}$$

Proof. Define the upcrossing variable,

$$\begin{aligned}
u([0,\infty),a,b,x) &= \lim_{t\to\infty} u([0,t],a,b,x). \text{ Then,}\\
E[u([0,t],a,b,x)] &\leq \frac{E[x^+(t)] + |a|}{b-a}, \text{ because of Theorem 20.4.2;}\\
E[u([0,\infty),a,b,x)] &= \lim_{t\to\infty} E[(u[0,t],a,b,x)],\\
& \text{because of the monotone convergence theorem,}\\
&\leq \lim_{t\to\infty} \frac{E[x^+(t)] + |a|}{b-a} \leq \frac{\sup_{t\in T} E[x^+(t)] + |a|}{b-a} < \infty.\\
& \text{Thus, for all } a,b \in \mathbb{Q},\; a<b,\\
\Lambda([a,b]) &= \{\omega \in \Omega \,|\, \liminf x(t) < a < b < \limsup x(t)\},\\
P(\Lambda([a,b])) &= 0, \;\Rightarrow\; P(\cup_{a,b\in\mathbb{Q},\; a<b} \Lambda([a,b])) = 0, \;\Rightarrow\\
0 &= P(\{\omega \in \Omega \,|\, \liminf x(t) < \limsup x(t)\}),\\
\Rightarrow &\quad \text{a.s.} - \lim_{t\to\infty} x(t) = x(\infty), \;\text{ exists.}\\
E|x(\infty)| &= E[\text{a.s.} - \lim |x(t)|] \leq \lim E|x(t)|, \text{ by Lemma 19.5.12,}\\
&\leq \sup E|x(t)| < \infty,
\end{aligned}$$

hence $x(\infty) \in L_1$ and this random variable is real valued. □

Theorem 20.4.4. *Consider the submartingale $\{x(t), F_t, t \in T\} \in SubM_1$. The following statements are equivalent:*

(a)$\{x(t), F_t, t \in T\}$ is uniformly integrable.
(b)$L_1 - \lim_{t\to\infty} x(t) = x(\infty)$.
(c)$\text{a.s.} - \lim_{t\to\infty} x(t) = x(\infty)$, $\{x(t), F_t, t \in T \cup \{\infty\}\} \in SubM_1$, and
$lim_{t\to\infty} E[x(t)] = E[x(\infty)]$.

Proof. (a) ⇒ (b). Uniform integrability implies with Proposition 20.2.4 that $\sup_{t\in T} E|x(t)| < \infty$. From Theorem 20.4.3 follows that $\text{a.s.} - \lim x(t) = x(\infty)$. This, [7, Lemma IV-2-5], and uniform integrability imply that $L_1 - \lim_{t\to\infty} x(t) = x(\infty)$.
(b) ⇒ (c). From (b) and the fact that x is a submartingale follows that $\lim_{t\to\infty} E|x(t)| = E|x(\infty)| < \infty$. From Theorem 20.4.3 follows that $\text{a.s.} - \lim_{t\to\infty} x(t) = x(\infty)$. Let $s,t \in T$, $t < s$, and $A \in F_t$. Then

$$\begin{aligned}
E[x(s)|F_t] &\geq x(t) \;\;\Rightarrow\\
E[I_A x(t)] &\leq E[I_A E[x(s)|F_t]] = E[I_A x(s)]\\
\lim_{s\to\infty} E[I_A x(s)] &= E[I_A x(\infty)], \text{ because of the } L_1\text{-convergence. Hence,}\\
E[I_A x(t)] &\leq E[I_A x(\infty)], \;\; \forall A \in F_T, \;\Rightarrow\\
E[x(\infty)|F_t] &\geq x(t), \;\; \{x(t), F_t, t \in T \cup \{\infty\}\} \in SubM_1.
\end{aligned}$$

The L_1-convergence of (b) implies convergence in expectation.
(c) $\Rightarrow$ (a). From $\{x(t), F_t, t \in T \cup \{\infty\}\} \in SubM_1$ and Corollary 20.2.7.(a) follows that $\{x^+(t), F_t, t \in T \cup \{\infty\}\} \in SubM_1$. Let $\lambda \in (0,\infty)$. Then

$$E[I_{(x^+(t)>\lambda)} x^+(t)] \leq E[I_{(x^+(t)>\lambda)} x^+(\infty)],$$

and from $x(\infty) \in L_1$ and from Proposition 19.5.16.(b) follows that $\{x^+(t), t \in T\}$ is uniformly integrable. Then,

$$\text{a.s.} - \lim_{t\to\infty} x^+(t) = x^+(\infty), \quad \lim_{t\to\infty} E[x^+(t)] = E[x^+(\infty)].$$

From (c) follows that $\lim_{t\to\infty} E[x(t)] = E[x(\infty)]$, hence $\lim_{t\to\infty} E[x^-(t)] = E[x^-(\infty)]$. This, a.s. $-\lim x^-(t) = x^-(\infty)$, and Theorem ... imply that $\{x^-(t), t \in T\}$ is uniformly integrable. Thus $x = x^+ - x^-$ is uniformly integrable.
□

Theorem 20.4.5. *Consider a martingale $\{m(t), F_t, t \in T\} \in M_1$. Then (a) and (b) of Theorem 20.4.4 are equivalent with*

(c') a.s. $-\lim_{t\to\infty} m(t) = m(\infty)$, *$m(\infty) \in L_1$, and $\{m(t), F_t, t \in T \cup \{\infty\}\} \in M_1$.*
(d) There exists a random variable $y : \Omega \to \mathbb{R}$, $y \in L_1$, such that $m(t) = E[y|F_t]$ for all $t \in T$.

Definition 20.4.6. The stochastic process $\{a(t), F_t, t \in T\}$, $a : \Omega \times T \to \mathbb{R}$ is called *increasing* if,

$$0 = a(0) \leq a(1) \leq \ldots \leq a(t) \leq \ldots, \ \forall\, t \in T.$$

It is called *integrable increasing* if for all $t \in T$, $E|a(t)| < \infty$. Define the random variable $a(\infty) : \Omega \to \mathbb{R} \cup \{\infty\}$ as $\lim_{t\to\infty} a(t) = a(\infty) \in \mathbb{R} \cup \{\infty\}$. The integrable increasing process is said to be *integrable at infinity* if $E|a(\infty)| < \infty$.

Theorem 20.4.7. Doob's submartingale decomposition. *Every submartingale $\{x(t), F_t, t \in T\} \in SubM_1$ admits a unique decomposition as a sum of a martingale and a predictable increasing process,*

$$x(t) = a(t) + m(t), \ \forall\, t \in T,$$

where $\{a(t), F_t, t \in T\}$ is predictable, integrable, and increasing; and $\{m(t), F_t, t \in T\} \in M_1$ is an integrable martingale.

Proof. Define the stochastic processes,

$$a : \Omega \times T \to \mathbb{R}, \ a(0) = E[x(0)],$$
$$a(t) = a(0) + \sum_{s=1}^{t} \left(E[x(s)|\, F_{s-1}] - x(s-1)\right), \ \forall\, t \in T\backslash\{0\},$$
$$m : \Omega \times T \to \mathbb{R}, \ m(0) = x(0) - E[x(0)],$$
$$m(t) = m(0) + \sum_{s=1}^{t} \left(x(s) - E[x(s)|\, F_{s-1}]\right), \ \forall\, t \in T\backslash\{0\}.$$

These processes have the following properties,

$$\begin{aligned}
& \{a(t),F_t,t \in T\} \text{ is a predictable process and } x \in SubM_1 \ \Rightarrow \\
& \forall\, t \in T\backslash\{0\},\ a(t+1)-a(t) = E[x(t+1)|F_t]-x(t) \geq 0, \\
& \{m(t),F_t,t \in T\} \text{ is an adapted process, } \forall\, t \in T\backslash\{0\}, \\
& E[m(t+1)-m(t)|\, F_t] = E[x(t+1)-E[x(t+1)|F_t]|\, F_t] \\
=\ & E[x(t+1)|\, F_t]-E[x(t+1)|\, F_t] = 0, \text{ hence } m \in M_1; \\
x(0) =\ & E[x(0)]+(x(0)-E[x(0)]) = a(0)+m(0), \\
& \text{suppose that for } s = 0,\ldots,t,\ x(t) = a(t)+m(t), \text{ then,} \\
x(t+1) =\ & x(t)+(E[x(t+1)|\, F_t]-x(t))+(x(t+1)-E[x(t+1)|\, F_t) \\
=\ & a(t+1)+m(t+1), \text{ and, by induction, } \forall\, r \in T,\ x(r) = a(r)+m(r); \\
& \text{suppose there exist two decompositions with} \\
x(t) =\ & a_1(t)+m_1(t) = a_2(t)+m(t),\ \forall\, t \in T,\ a_1(0) = E[x(0)] = a_2(0); \\
& \text{then } m_1(0) = x(0)-a_1(0) = x(0)-a_2(0) = m_2(0); \\
& \text{suppose that } a_1(s) = a_2(s),\ m_1(s) = m_2(s)\ \forall\, s = 0,\ldots,t; \\
0 =\ & E[(m_1(t+1)-m_1(t))-(m_2(t+1)-m_2(t))|F_t] \\
=\ & E[(a_2(t+1)-a_2(t))-(a_1(t+1)-a_1(t))|F_t] \\
=\ & (a_2(t+1)-a_2(t))-(a_1(t+1)-a_1(t)), \\
& \text{because } a_1 \text{ and } a_2 \text{ are predictable,} \\
=\ & a_2(t+1)-a_1(t+1) \ \Rightarrow\ a_1(t+1) = a_2(t+1), \\
\Rightarrow\ & m_1(t+1) = m_2(t+1).
\end{aligned}$$

□

20.5 Ergodicity

To estimate the finite-dimensional distributions of a stochastic process from observations of a phenomenon for which the process is a model, the process must be ergodic. The concept of ergodicity has been deeply investigated.

Definition 20.5.1. Let (Ω,F,P) be a probability space and $S:\Omega \to \Omega$ be a measurable map. Define $S^{-1}A = \{\omega \in \Omega|S(\omega) \in A\}$. A set $A \in F$ is called a *S-invariant set* if $S^{-1}A = A$. Let $F_S \subseteq F$ be the set of all S-invariant sets of F. The map S is called *measure-preserving* if for all $A \in F$, $P(S^{-1}A) = P(A)$.

Proposition 20.5.2. *The class F_S is a σ-algebra.*

The proof of the above proposition is an elementary verification.

Given a probability space (Ω,F,P), a measurable map $S:\Omega \to \Omega$, and a random variable $x:\Omega \to R$. Define $x_1(\omega) = x(\omega)$, $x_2(\omega) = x(S\omega),\ldots,x_{n+1} = x(S^n\omega),\ldots$. In time, $\{S^n(\omega),\ n \in \mathbb{Z}_+\}$ wanders all over Ω. Suppose there exists a probability measure P_1 on Ω that describes the way $\{S^n(\omega),\ n \in \mathbb{Z}_+\}$ is distributed. Then one expects that,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} x_k \ = \ \int_{\Omega} x(\omega)P_1(d\omega) \ = \ E_1[x].$$

If there is a set $A \in F$ such that A is S-invariant and $0 < P_1(A) < 1$ then $\{S^n(\omega),\ n \in \mathbb{Z}_+\}$ will not be distributed over all Ω because the trajectory $\{S^n(\omega),\ n \in \mathbb{Z}_+\}$ will remain in A. Hence one must demand that S-invariant sets have either probability zero or probability one.

Definition 20.5.3. A measure-preserving map $S : \Omega \to \Omega$ is called an *ergodic map* if for all $A \in F_S$ either $P(A) = 0$ or $P(A) = 1$.

Definition 20.5.4. Let $S : \Omega \to \Omega$ be a measure-preserving map and $x : \Omega \to \mathbb{R}$ be a random variable. Then x is called a *S-invariant random variable* if for all $\omega \in \Omega$, $x(\omega) = x(S(\omega))$.

Proposition 20.5.5. *Let $S : \Omega \to \Omega$ be a measure-preserving map and $x : \Omega \to \mathbb{R}$ be a random variable.*

(a)x is a S-invariant random variable if and only if x is measurable on F_S.
(b)S is ergodic if and only if every S-invariant random variable is almost surely equal to a constant.

Theorem 20.5.6. *Let $S : \Omega \to \Omega$ be a measure-preserving map and $x : \Omega \to \mathbb{R}$ be a random variable. Assume that $E|x| < \infty$. Then*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} x(S^{k-1}(\omega)) = E[x|F_S](\omega) \ \textit{a.s.}$$

If S is ergodic then the above limit equals $E[x]$ a.s.

20.6 Further Reading

See also the section *Further Reading* of Chapter 3.

Filtrations. For filtrations see [4, 5, 6] on which the presentation of this chapter is based.

Martingale theory including stopping times and filtrations. The classical books include [2, 3, 6]. Discrete-time martingales are treated in the French-language book of J. Neveu, [7]. The presentation of this chapter is based on these references.

Ergodicity is treated in [1, Sec. 6.3] on which the presentation of this chapter is based.

References

1. L. Breiman. *Probability*. Addison-Wesley Publ. Co., Reading, MA, 1968. 49, 73, 741, 758
2. K.L. Chung. *A course in probability theory*. Academic Press, New York, 1974. 12, 49, 741, 758
3. J.L. Doob. *Stochastic processes*. Wiley, New York, 1953. 72, 73, 721, 747, 754, 758
4. P.A. Meyer. *Probability and Potentials*. Blaisdell Publishing Company, Waltham, MA, 1966. 49, 336, 741, 758
5. P.A. Meyer. *Martingales and stochastic integrals I*, volume 284 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1972. 758
6. P.A. Meyer. Un cours sur les intégrales stochastiques. In P.A. Meyer, editor, *Séminaire de Probabilités*, number 511 in Lecture Notes in Mathematics, pages 245–400. Springer, Berlin, 1972. 758
7. J. Neveu. *Martingales à temps discrets*. Masson et Cie, Paris, 1972. 39, 49, 755, 758

Chapter 21
Appendix E Control and System Theory of Deterministic Systems

Abstract Concepts and theorems of the system theory of deterministic linear systems are summarized. Controllability, observability, and a realization are formulated. Realization theory includes necessary and sufficient conditions for the existence of a realization, a characterization of the minimality of a realization, and a classification and description of all minimal realizations. Attention is restricted to linear control systems after a general introduction.

Key words: Linear system. Controllability. Observability. Realization.

The reader finds in this appendix an introduction to system theory of linear control systems, primarily of time-invariant linear control systems. Although the theory can be found in various books, the results needed for this book are not easily available in one reference.

Realization theory of finite-dimensional linear deterministic systems is described in this chapter. The reader is expected to have read the introduction to realization theory provided in Section 6.1.

21.1 Deterministic Control Systems

In this section the classes of linear and nonlinear control systems are defined. Properties of such systems are established. Major sources for the formulation of a control system include those of R.E. Kalman [34], the book by L.A. Zadeh and C.A. Desoer [77], and the paper by M. Arbib [2] who includes automata with partial observations. In general, a control system has as objects: trajectories of an input, a state, and an output; the state-transition map from a state and an input to the next state; and the output map from a state and an input to the output. A condition is that the system satisfies the semi-group property for the state.

In this book first linear control systems are defined and subsequently more general classes of systems. In the light of the historic development of system theory it is more natural to go from specific classes of systems to generalizations.

The *discrete-time time-index set* is in this chapter defined to be either the set $T = T(0:t_1) = \{0,1,2,\ldots,t_1\}$ for a strictly positive integer $t_1 \in \mathbb{Z}_+$ or the set of the natural numbers $T = \mathbb{N} = \{0,1,\ldots\}$. For T as defined above, denote the set of functions from T to X by $F(T,X) = \{x : T \rightarrow X\}$.

Definition 21.1.1. A *discrete-time time-invariant finite-dimensional real linear control system* (LS), (if the context is understood, this will be called a *linear control system* or a *linear system*) is a tuple,

$$\begin{aligned}
& (T,X,U,Y,A,B,C,D), \\
T &= \mathbb{N}, \text{ is the } \textit{time index set}; \; n_x,n_u,n_y \in \mathbb{N}; \\
X &= \mathbb{R}^{n_x}, \text{ is a vector space over } \mathbb{R}, \text{ called the } \textit{state space}, \\
U &= \mathbb{R}^{n_u}, \text{ is a vector space over } \mathbb{R}, \text{ called the } \textit{input space}, \\
Y &= \mathbb{R}^{n_y}, \text{ is a vector space over } \mathbb{R} \text{ called the } \textit{output space}, \\
& A \in \mathbb{R}^{n_x \times n_x}, \; B \in \mathbb{R}^{n_x \times n_u}, \; C \in \mathbb{R}^{n_y \times n_x}, \; D \in \mathbb{R}^{n_y \times n_u}.
\end{aligned}$$

Define the *transition map* of this system by the sets and functions,

$$\begin{aligned}
& \phi : D_\phi \rightarrow R_\phi, \\
& \text{over the interval } T_1 = \{t_0, t_0+1, t_0+2, \ldots, t_1\} \subseteq T, \\
D_\phi &= \left\{ \, (T_1, x_0, u_{T_1}) \in \text{Pwrset}(T) \times X \times F(T_1,U) \, \right\}, \\
& \text{is called the } \textit{domain} \text{ of } \phi, \\
R_\phi &= \left\{ \, (x_{T_1}, y_{T_1}, x_1) \in F(T_1,X) \times F(T_1,Y) \times X \; \right\}, \text{ is called the } \textit{range} \text{ of } \phi, \\
& \text{if } (T_1, x_0, u_{T_1}) \in D_\phi, \text{ then} \\
& T_1 = \{t_0, \; t_0+1, \; \ldots, \; t_1-1, \; t_1\} \in \textit{Init}(\mathbb{N}) \\
& \text{is called the } \textit{time interval} \text{ of the transition}, \\
& x_0 \in X, \text{ is called the } \textit{initial state}, \\
& u_{T_1} \in F(T_1,U), \text{ is called the } \textit{input function}, \\
u_{T_1} &= \{u(t_0), u(t_0+1), \ldots, u(t_1-1)\}, \\
& \text{and the value of the transition map is defined as} \\
& (x_{T_1}, y_{T_1}, x_1) = \phi(T_1, x_0, u_{T_1}), \text{ by the recursion}, \\
x(t+1) &= Ax(t) + Bu(t), \; x(t_0) = x_0, \\
y(t) &= Cx(t) + Du(t), \text{ for } t = t_0, t_0+1, \ldots, t_1-1, \\
x_{T_1} &= \{x(t_0), x(t_0+1), \ldots, x(t_1-1)\}, \\
& \text{called the } \textit{state function } x : T_1 \rightarrow X, \\
y_{T_1} &= \{y(t_0), y(t_0+1), \ldots, y(t_1-1)\}, \\
& y : T_1 \rightarrow Y, \text{ called the } \textit{output function}, \\
x_1 &= x(t_1) \in X, \text{ called the } \textit{terminal state}.
\end{aligned}$$

A transition will also be denoted as $(t_0,x_0) \mapsto^{u_{T_1}} (t_1,x_1)$.

The system is assumed to satisfy the conditions of *finite concatenation* and *sequential concatenation* defined next. Consider two consecutive time intervals $T_1 = \{0,1,2,\ldots,t_1\}$ and $T_2 = \{t_1,t_1+1,t_1+2,\ldots,t_2\}$. Consider the initial state $x_0 \in X$ and an input trajectory $u : T_1\backslash\{t_1\} \to U$ which define by the above definitions the transition $(0,x_0) \mapsto^{u_{T_1}} (t_1,x_1)$. Next consider the terminal state x_1 which becomes the initial state of the time interval T_2, and consider the input trajectory $u : T_2\backslash\{t_2\} \to U$. According to the above definitions, there then exists a transition $(t_1,x_1) \mapsto^{u_{T_2}} (t_2,x_2)$. Next define the combined time interval $T_{12} = T_1 \cup T_2$ and the *concatenated input trajectory* $u : T_{12} \to U$ $u_{T_{12}}(t) = u_{T_1}(t)$ for all $t \in T_1\backslash\{t_1\}$ and $u_{T_{12}}(s) = u_{T_2}(s)$ for all $s \in T_2$.

The system is said to be closed with respect to *finite concatenation* if T_{12} is an admitted interval of the system and the following transition is defined in the system, $(0,x_0) \mapsto^{u_{T12}} (t_2,x_2)$.

The system is said to be closed with respect to *sequential concatenation* if the system is well defined as described above and it is closed with respect to finite concatenation for a sequence of consecutive time intervals $\{T_{i,i+1},\ \forall\, i \in \mathbb{Z}_+\}$.

In the literature the finite concatenation is also described as the semi-group property of a system.

Denote the set of time-invariant linear systems in the vector space of the real numbers by LS when its attributes are understood, and denote its parameters by the set,

$$(T,\mathbb{R}^{n_x},\mathbb{R}^{n_u},\mathbb{R}^{n_y},A,B,C,D) \in \mathrm{LS},$$
$$\mathrm{LSP}(n_x,n_y,n_u) = \left\{(A,B,C,D) \in \mathbb{R}^{n_x\times n_x} \times \mathbb{R}^{n_x\times n_u} \times \mathbb{R}^{n_y\times n_x} \times \mathbb{R}^{n_y\times n_u}\right\}.$$

A linear system is a linear control system in which there is no input hence the matrices $B,\ D$ do not appear in the representation.

A linear system, once the definition and the notation is understood, is often specified only by the recursion for the state function and the output function, thus by the equations,

$$x(t+1) = Ax(t) + Bu(t),\ x(t_0) = x_0,$$
$$y(t) = Cx(t) + Du(t).$$

The definition of a linear system may be extended to a system over an arbitrary field $\mathbb{F}$, thus with $X = \mathbb{F}^{n_x}$, $U = \mathbb{F}^{n_u}$, and $Y = \mathbb{F}^{n_y}$. A particular case of this is a linear system over the complex numbers, $X = \mathbb{C}^{n_x}$.

Definition 21.1.2. A *finite-dimensional real control system*, defined in finite-dimensional vector spaces of tuples of the real numbers, is a collection,

$(T,X,U,Y,\mathbb{R},f,h)$ where $T = \mathbb{N}$, is the time index set,

$X \subseteq \mathbb{R}^{n_x}$, $U \subseteq \mathbb{R}^{n_u}$, $Y \subseteq \mathbb{R}^{n_y}$, *are subsets of the indicated vector spaces,*

$f : T \times X \times U \to X$, $h : T \times X \times U \to Y$, are nonlinear maps.

Define the *state transition map* and the *output equation* of this system as,

$$\begin{aligned}
& \phi : D_\phi \rightarrow R_\phi, \text{ over the interval } T_1 = \{t_0, t_0+1, \ldots, t_1\}, \\
D_\phi &= \left\{ \, (T_1, x_0, u_{T_1}) \in \mathbb{N}_{t_1} \times X \times F(T_1, U) \, \right\} \\
R_\phi &= \left\{ \, (x_{T_1}, y_{T_1}, x_1) \in F(T_1, X) \times F(T_1, Y) \times X \, \right\}, \\
(x_{T_1}, y_{T_1}, x_1) &= \phi(T_1, x_0, u_{T_1}), \\
x(t+1) &= f(t, x(t), u(t)), \; x(t_0) = x_0, \\
y(t) &= h(t, x(t), u(t)), \\
x_{T_1} &= \{x(0), x(1), \ldots, x(t_1-1)\}, \; y_{T_1} = \{y(0), y(1), \ldots, y(t_1-1)\}, \\
x_1 &= x(t_1).
\end{aligned}$$

This definition may be generalized to a system in arbitrary sets X, U, Y and maps. Such a system is then called a *control system*.

Many engineering systems have the time-invariance property. A discrete-time time-invariant finite-dimensional linear system has the time-invariance property.

Definition 21.1.3. Consider a finite-dimensional linear system. It is called *time-invariant* if,

$$\begin{aligned}
& \forall \; t_0, \; t_1 \in T, \; t_0 < t_1, \; \forall \; s \in \mathbb{Z}, \text{ such that, } t_0+s, \; t_1+s \in T, \\
& (x_{T_1}, y_{T_1}, x(t_1)) = \phi(T_1, x_0, u_{T_1}), \text{ and,} \\
& (x_{T_1+s}, y_{T_1+s}, x(t_1+s)) = \phi(T_1+s, x_0, u_{T_1+s}), \text{ where,} \\
T_1 &= \{t_0, \ldots, t_1-1\}, \; T_1+s = \{t_0+s, \ldots, t_1-1+s\}, \\
u_{T_1} &= \{u(t_0), \ldots, u(t_1-1)\}, \; u_{T_1+s} = \{u(t_0+s), \ldots, u(t_1-1+s)\}, \\
y_{T_1} &= \{y(t_0), \ldots, y(t_1-1)\}, \; y_{T_1+s} = \{y(t_0+s), \ldots, y(t_1-1+s)\}.
\end{aligned}$$

Thus, if the systems starts at time t_0+s with an input signal which is delayed or advanced by a duration $s \in \mathbb{Z}_+$ then the state function and the output fuction are also delayed by duration s.

A time-invariant linear system has time-invariant input-output trajectories as proven below.

Proposition 21.1.4. Time-invariance. *Consider a time-invariant linear control system specified by,*

$$\begin{aligned}
& (T, \mathbb{R}^{n_x}, \mathbb{R}^{n_u}, \mathbb{R}^{n_y}, A, B, C, D), \; T = \mathbb{N}, \\
x(t+1) &= Ax(t) + Bu(t), \; x(t_0) = x_0, \\
y(t) &= Cx(t) + Du(t).
\end{aligned}$$

Then,

$$\begin{aligned}
& \forall \; t_0, \; t_1 \in T, \; t_0 < t_1, \; \forall \; s \in \mathbb{Z}_+, \textit{ such that } t_0+2, \; t_1+s \in T, \\
(x_{T_2}, y_{T_2}, x_2) &= \phi(T_2, x_0, u_2) = \phi(T_1, x_0, u_1) = (x_{T_1}, y_{T_1}, x_1), \textit{ where,} \\
T_1 &= \{t_0, t_0+1, \ldots, t_1-1\}, \; s \in \mathbb{Z}, \\
T_1+s &= \{t_0+s, t_0+1+s, \ldots, t_1-1+s\} \subseteq T, \\
& u_1 \in F(T_1, U), \; u_2 \in F(T_2, U), \; u_2(t+s) = u_1(t), \; \forall t \in T_1, \\
x_{T_1+2} &= x_{T_1} = \{x_1(t_0), \ldots, x_1(t_1-1)\}, \; y_{T_1+s} = y_{T_1}, \; x_2 = x_1 = x_1(t_1).
\end{aligned}$$

Proof. Define for (T_1, x_o, u_1)

$$\begin{aligned}
x_1(t+1) &= Ax_1(t) + Bu_1(t),\ x_1(t_0) = x_0,\\
y_1(t) &= Cx_1(t) + Du_1(t),\\
x_{1,T_1} &= \{x_1(t_0), x_1(t_0+1), \ldots, x_1(t_1-1)\},\ x_{1,1} = x_1(t_1),\\
y_{1,T_1} &= \{y_1(t_0), y_1(t_0+1), \ldots, y_1(t_1-1)\}.
\end{aligned}$$

Then consider and note that,

$$\begin{aligned}
x_2(t+1) &= Ax_2(t) + Bu_2(t),\ x_2(t_0+s) = x_0,\\
y_2(t) &= Cx_2(t) + Du_2(t);\\
x_2(t_0+s) &= x_0,\ u_2(t_0+s) = u_1(t_0),\\
x_2(t_0+s+1) &= Ax_2(t_0+s) + Bu_2(t_0+s) = Ax_0 + Bu_1(t_0) = x_1(t_0+1),\\
y_2(t_0) &= Cx_2(t_0+s) + Du_2(t_0+s) = Cx_1(t_0) + Du_1(t_0) = y_1(t_0).
\end{aligned}$$

By induction it may be proven that,

$$\begin{aligned}
&\forall\ r = 0, 1, \ldots, t_1 - 1,\ x_2(t_0+s+r) = x_1(t_0+r),\ y_2(t_0+s+r) = y_1(t_0+r);\\
&\Rightarrow\ \phi(T_1+s, x_0, u_2) = \phi(T_1, x_0, u_1).
\end{aligned}$$

□

Definition 21.1.5. Consider a linear control system,

$$\begin{aligned}
&(\mathbb{N},\ \mathbb{R}^{n_x},\ \mathbb{R}^{n_u},\ \mathbb{R}^{n_y},\ A,\ B,\ C,\ D),\\
x(t+1) &= Ax(t) + Bu(t),\ x(t_0) = x_0,\\
y(t) &= Cx(t) + Du(t).
\end{aligned}$$

Define the *impulse response function* of this system as the map

$$\begin{aligned}
&H : T \to \mathbb{R}^{n_y \times n_u},\ T_0 = \{0, 1, 2, \ldots\} = N,\\
H(t) &= \begin{cases} D, & \text{if } t = 0,\\ CA^{t-1}B, & \text{if } t = 1, 2, \ldots \end{cases}
\end{aligned}$$

In general, an *impulse response function* is a map $H : T \to \mathbb{R}^{n_y \times n_u}$ without reference to a linear control system.

Proposition 21.1.6. *Consider a linear control system,*

$$\begin{aligned}
&(\mathbb{N},\ \mathbb{R}^{n_x},\ \mathbb{R}^{n_u},\ \mathbb{R}^{n_y},\ A,\ B,\ C,\ D),\\
x(t+1) &= Ax(t) + Bu(t),\ x(0) = x_0,\\
y(t) &= Cx(t) + Du(t).
\end{aligned}$$

Formulas for the state and the output function at any time in terms of the initial state, the input function, and the impulse response function are:

$$\begin{aligned}
x(t) &= A^t x_0 + \Sigma_{s=0}^{t-1} A^{t-1-s} Bu(s),\\
y(t) &= CA^t x_0 + \Sigma_{s=0}^{t-1} CA^{t-1-s} Bu(s) + Du(t)\\
&= CA^t x_0 + \sum_{s=0}^{t} H(t-s)\ u(s),\ t \in T(0:t_1);\\
&x_0 = 0\ \Rightarrow\ y(t) = \Sigma_{s=0}^{t} H(t-s)\ u(s),\ t \in T.
\end{aligned}$$

Proof. By induction,

$$\begin{aligned}
x(1) &= Ax(0) + Bu(0) = Ax_0 + Bu(0), \\
x(t-1) &= A^{t-1}x_0 + \Sigma_{s=0}^{t-2}\, A^{t-2-s}Bu(s), \;\Rightarrow \\
x(t) &= Ax(t-1) + Bu(t-1) = A^t x_0 + \Sigma_{s=0}^{t-2}\, A^{t-1-s}Bu(s) + Bu(t-1) \\
&= A^t x_0 + \Sigma_{s=0}^{t-1}\, A^{t-1-s}Bu(s), \\
y(t) &= Cx(t) + Du(t) = CA^t x_0 + \Sigma_{s=0}^{t-1}\, CA^{t-1-s}Bu(s) + Du(t) \\
&= CA^t x_0 + \sum_{s=0}^{t-1} H(t-s)\; u(s).
\end{aligned}$$

□

21.2 Controllability

The existence of a control law for a particular control objective depends on a condition that is often called controllability. This concept arises in the realization problem also and this occurence is understandable from realization theory. In this subsection the concept of controllability is defined and characterized.

A subset $S \subset \mathbb{C}$ is called *symmetric* with respect to the real axis if for all $\lambda \in S$ its complex conjugate also belongs to S; equivalently, $\lambda = x + iy \in S \;\Rightarrow \overline{\lambda} = x - iy \in S$.

Problem 21.2.1. *Eigenvalue assignment problem.* Consider a time-invariant linear control system with representation,

$$x(t+1) = Ax(t) + Bu(t), \; x(t_0) = x_0.$$

Consider any subset of the complex numbers $S \subset \mathbb{C}$ which is symmetric with respect to the real axis. Does there exist a linear control law of the form $g(x) = Fx$ for a matrix $F \in \mathbb{R}^{n_u \times n_x}$ such that $\text{spec}(A + BF) \subset S$?

A necessary and sufficient condition for the above problem is the concept of controllability.

Controllability is below first formulated in the form of a set theoretic description which is useful for extension to other sets of systems than linear control systems. The reader should recall the terminology and the notation of a state transition, see Def. 21.1.1, with $(t_0, x_0) \mapsto^{u_{t_0:t_1-1}} (t_1, x_1)$.

The reader has to distinguish: the concept of (1) controllability from that of (2) co-controllability, see Fig. 21.1 and Fig. 21.2.

There is a major problem with the terminology of controllability. Originally R.E. Kalman introduced the term of *reachability* for what will be called controllability below. Then there was a need for backward reachability which has also been called co-reachability. Reachability and controllability involve the use of the input trajectory. In computer science the term of reachability is used in case of an autonomous

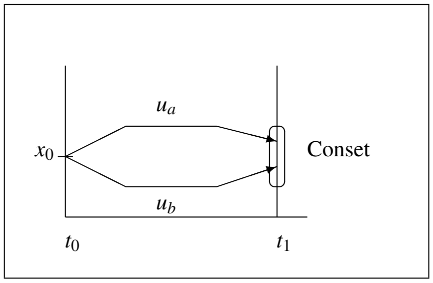


Fig. 21.1 Diagram of the controllable set Conset$(t_1;(t_0,x_0),u)$.

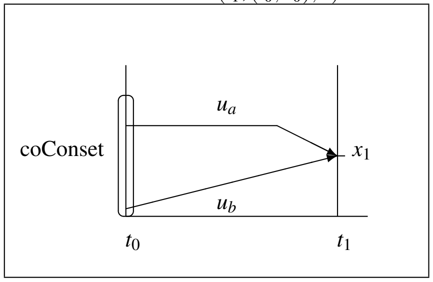


Fig. 21.2 Diagram of the co-controllable set coConset$(t_0;(t_1,x_1),u)$.

system without input, and of co-reachability for the backward version. There have also been formulated concepts like strong controllability and weak controllability.

To avoid confusion with the readers on the issue of the terms, the following definitions of terms will be used in this book:

- The term *reachability* refers to the autonomous system $x(t+1) = f(t,x(t))$, $x(0) = x_0 \in X$ where x_0 is fixed.
  A state $x_1 \in X$ is called a *reachable state* by the system considered on the interval $T_r = \{0,1,\ldots,t_1\}$ if $x(t_1) = x_1$. The system is called a *reachable system* if for every state $x_1 \in X$ there exists an interval T_r such $x_1 \in X$ is a reachable state on that interval.
- A state $x_1 \in X$ is called a *co-reachable state* on the interval $T_r = \{0,-1,\ldots,-t_1\}$ if $x(-t_1;(0,x_0)) = x_1$. The system is called a *co-reachable system* if for every state $x_1 \in X$ there exists an interval T_r such that x_1 is a co-reachable state on that interval.
- Consider next a linear control system with representation,
  $x(t+1) = A(t)x(t) + B(t)u(t),\ x(0) = x_0$. Regard x_0 as fixed.
  A terminal state $x_1 \in X$ is called a *controllable state* on the interval
  $T_c = \{0,1,\ldots,t_1\} \subset T$ if there exists an input trajectory $u : T_c \to \mathbb{R}^{n_u}$ such that $x_1 = x(t_1;(0,x_0),u)$. The system is called a *controllable system* if for every state $x_1 \in X$ there exists an interval T_c such that the state $x_1 \in X$ is a controllable state on that interval.
- A terminal state is called a *co-controllable state* on the interval
  $T_c = \{0,-1,\ldots,-t_1\}$ if there exists an input $u : T_c \to \mathbb{R}^{n_u}$ such that
  $x(-t_1;(0,x_0),u) = x_1$. The system is called a *co-controllable system* if for ev-

ery state $x_1 \in X$ there exists an interval T_c such that the state $x_1 \in X$ is a co-controllable state on that interval.

Example 21.2.2. The following time-invariant linear system with an idempotent system transition matrix shows that the set of reachable states and the set of co-reachable states are different.

$$x(t+1) = Ax(t),\ x(0) = x_0,\ A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},\ n_x = 2.$$

Definition 21.2.3. *Controllable subset and co-controllable subset.* Consider a control system of Def. 21.1.2, not necessarily a linear control system, with a time-index set $T = T(0:t_1) \subset \mathbb{N}$, a state set X, and set of input functions $\mathbf{U}$.

Define the *controllable subset* at time $t_1 \in T$ from the initial time $t_0 \in T$ and from a subset of initial states $X_0 \subseteq X$ as the set,

$$\text{Conset}(\{t_1\};\ t_0,X_0) = \left\{ \begin{array}{l} x_1 \in X |\ \exists\, x_0 \in X_0,\ \exists\, u(t_0:t_1-1) \in \mathbf{U} \\ \text{such that } (t_0,x_0) \mapsto^{u(t_0:t_1-1)} (t_1,x_1) \end{array} \right\},$$

$$\text{Conset}((t_0:t_1);\ t_0,X_0) = \cup_{s\in(t_0,t_1)}\text{Conset}(\{s\};t_0,X_0);\ \text{in general}$$

$$\text{Conset}(\{t_1+1\};t_0,X_0) \not\subseteq \text{Conset}((t_0:t_1);t_0,X_0),$$

because it may not be possible to stay at a particular state.

Define the *co-controllable subset* at time $t_0 \in T$ to the terminal time t_1 and to the subset of terminal states X_1 as the set,

$$\text{coConset}(\{t_0\};\ t_1,X_1) = \left\{ \begin{array}{l} x_0 \in X |\ \exists\, x_1 \in X_1,\ \exists\, u(t_0:t_1-1) \\ \text{such that } (t_0,x_0) \mapsto^{u(t_0:t_1-1)} (t_1,x_1) \end{array} \right\},$$

$$\text{coConset}(\{t_0,t_1\};\ t_1,X_1) = \cup_{s\in\{t_0,t_1\}}\text{coConset}(\{s\};t_1,X_1).$$

Note that the above definition is formulated for an arbitrary dynamic system. The following definition is however particular.

Definition 21.2.4. *Controllable system and co-controllable system.* Consider a time-invariant finite-dimensional linear control system without output,

$$x(t+1) = Ax(t) + Bu(t),\ x(t_0) = x_0,\ x(t) \in X,\ t_0,\ t_1 \in \mathbb{N}.$$

- The system is said to be *controllable from the subset of initial states $X_0 \subseteq X$ on the time interval* $T = T(t_0:t_1) \subset \mathbb{Z}$ if $\text{Conset}(\{t_1\};t_0,X_0) = X$.
- The system is said to be *controllable on the time interval $T = T(t_0,t_1)$ to the terminal time* if for all $x_0 \in X$ it is controllable from $\{x_0\} \subset X$ to X on $T = T(t_0:t_1)$. Equivalently, if $\forall\, x_0 \in X, X = \text{Conset}(\{t_1\},t_0,\{x_0\})$.
- The system is said to be *controllable* on $T = \mathbb{N}$ if there exists a $t_1 \in T$ such that it is controllable on the interval $T = T(t_0:t_1)$.
- The system is said to be *co-controllable to the set of terminal states $X_1 \subseteq X$* on the time interval $T = T(t_0;t_1)$ if $\text{coConset}(t_0;t_1,X_1) = X$.
- The system is said to be *co-controllable on the time interval $T(t_0:t_1)$* if for all $x_1 \in X$ it is co-controllable to the subset $\{x_1\} \subset X$ on $T(t_0:t_1)$.

- The system is said to be *co-controllable* if there exists a $t_0 \in T(0:\infty) = \mathbb{N}$ such that for all $x_1 \in X$, it is controllable to $\{x_1\}$ on the interval $T(t_0:t_1)$.

Theorem 21.2.5. *Consider a time-invariant linear control system,*

$$x(t+1) = Ax(t) + Bu(t),\ x(t_0) = x_0.$$

(a)The system is controllable on $T = T(t_0,t_1)$ if and only if the system is controllable from the initial state $0 \in X$ on $T = T(t_0,t_1)$.

(b)If for all $x_1 \in X$ the system is controllable to $\{x_1\}$ on the interval $T = T(t_0,t_1)$ then it is controllable to the zero state, $\{0\} \subset X$. However, the converse is not true.

(c)The system is controllable if and only if the system is co-controllable.

Proof. (a) ($\Rightarrow$) By definition. ($\Leftarrow$) Consider $x_0,\ x_1 \in X$. Define $\bar{x}_1 = x_1 - A^{t_1-t_0}x_0 \in X$ The assumption that the system is controllable from $\{0\}$ implies that there exists an input function $u \in U_{Tbackslash\{t_1\}}$ such that,

$$\bar{x}_1 = \sum_{s=t_0}^{t_1-1} A^{t_1-1-t_0}Bu(s) \ \Rightarrow\ x_1 = A^{t_1-t_0}x_0 + \sum_{s=t_0}^{t_1-1} A^{t_1-1-t_0}Bu(s),$$

implies that $x_1 \in \text{Conset}(t_1;t_0;\{x_0\})$ hence $X = \text{Conset}(t_1;t_0,\{x_0\})$.
(b) The implication holds because $x_1 = 0 \in X$ is a particular case. The following linear system,

$$x(t+1) = \begin{pmatrix} 0\ 0 \\ 1\ 0 \end{pmatrix} x(t) + \begin{pmatrix} 0 \\ 0 \end{pmatrix} u(t),\ x(t_0) = x_0 \in X = \mathbb{R}^2,$$

shows that any initial state is controllable to the zero state by a zero input function. Yet, for any state $x_1 \in X$, $x_{1,1} \neq 0$, there exists a $x_0 \in X$, $x_0 \neq x_1$, such that $(t_0,x_0) \not\mapsto (t_1,x_1)$ because,

$$x(t_0+1) = \begin{pmatrix} 0 \\ x_{0,1} \end{pmatrix},\ \forall t \geq t_0+2,\ x(t) = \begin{pmatrix} 0 \\ 0 \end{pmatrix};\ \forall\ t \in T,\ x(t) \neq x_1.$$

(c) ($\Rightarrow$) Consider $x_0 \in X$. Then there exists a $t_1 \in \{t_0,\ldots,\infty\}$ such that the system is controllable from $\{0\}$ on $T(t_0:t_1)$ and by (a) it is controllable from $\{x_0\}$ to X on $T = T(t_0:t_1)$. $X = \text{Conset}(t_1;t_0,\{x_0\})$. Take $x_1 \in X$. Then,

$$\Rightarrow\ \exists u \in U \text{ such that } x_1 = A^{t_1-t_0}x_0 + \sum_{s=t_0}^{t_1-1} A^{t_1-1-s}Bu(s)$$

$$\Rightarrow\ x_0 \in \text{coConset}(t_0;t_1,\{x_1\}) \ \Rightarrow\ X \subseteq \text{coConset}(t_0;t_1,\{x_1\}),$$

hence equality holds and the system is co-controllable.
($\Leftarrow$) Consider $x_1 \in X$. By assumption there exists a $t_1 \in [t_0,\infty)$ such that the system is co-controllable from $0 \in X$ to $\{x_1\}$ on $T = [t_0,t_1]$. From (b) follows that the system is controllable to $\{0\} \subset X$ on $T = [t_0,t_1]$. Then there exists a $u \in U$ such that,

$$\begin{aligned} x_1 &= \sum_{s=t_0}^{t_1-1} A^{t_1-1-s}Bu(s) \Rightarrow x_1 \in \mathrm{Conset}(t_1;t_0,\{0\}), \\ &\Rightarrow X \subseteq \mathrm{Conset}(t_1;t_0,\{0\}), \end{aligned}$$

hence the system is controllable from $\{0\}$ on $T=[t_0,t_1]$ and consequently the system is controllable. □

Definition 21.2.6. *Controllable tuple* $(A,\ B)$ *of system matrices*. Consider the linear control system without output,

$$x(t+1) = Ax(t)+Bu(t),\ x(t_0)=x_0.$$

The matrix tuple (A,B) is said to satisfy the the condition of *controllability* if one of the following equivalent conditions holds:

(a) Define the *controllability matrix* of this system as the matrix,

$$\mathrm{conmat}(A,B) = \left(B\ AB\ A^2B\ \dots\ A^{n_x-1}B \right) \in \mathbb{R}^{n_x\times(n_xn_u)}.$$

The pair of matrices $(A,B)\in\mathbb{R}^{n_x\times n_x}\times\mathbb{R}^{n_x\times n_u}$ will be called a *controllable pair* if $\mathrm{rank}(\mathrm{conmat}(A,B)) = n_x$.

(b) The complex number $\lambda\in\mathrm{spec}(A)$ will be called an (A,B)*-controllable eigenvalue* if,

$$n_x = \mathrm{rank}\left(A-\lambda I\ B \right).$$

The matrix tuple (A,B) will be called *eigenvalue controllable* if every eigenvalue $\lambda\in\mathrm{spec}(A)$ is (A,B)-controllable.

(c) The complex number $\lambda\in\mathrm{spec}(A)$ will be called an *spectrally* (A,B)*-controllable eigenvalue* if $x\in\mathbb{C}^{n_x}$ with $x^TA = x^T\lambda$ and $x^TB=0$ implies that $x=0$. The pair (A,B) is called a *spectrally controllable pair* if every eigenvalue $\lambda\in\mathrm{spec}(A)$ is spectrally (A,B)-controllable.

Theorem 21.2.7. Existence control law equivalent with system controllability. *Consider the time-invariant linear control system,*

$$x(t+1) = Ax(t)+Bu(t),\ x(t_0)=x_0.$$

(a) The controllability conditions of Def. 21.2.6 are equivalent.

(b) The following statements are equivalent:

(b.a) For any subset $S\subset\mathbb{C}$ with n_x elements which is symmetric with respect to the real axis, there exists a linear control law $g(x)=Fx$ with $g: X\to U$, $F\in\mathbb{R}^{n_u\times n_x}$ such that $\mathrm{spec}(A+BF)\subset S$*;*

(b.b) The system is controllable.

(b.c) The matrix tuple (A,B) satisfies any of the equivalent conditions of controllability of Def. 21.2.6.

Example 21.2.8. Consider the linear control system,

$$\begin{aligned} &(T,\ \mathbb{R}^2,\ \mathbb{R},\ \mathbb{R},\ \mathbb{R},\ A,\ B,\ C,\ D),\ T=\mathbb{N},\ n_x=2,\\ x(t+1) &= \begin{pmatrix} a_{11} & a_{12} \\ 0 & a_{22} \end{pmatrix} x(t) + \begin{pmatrix} b_1 \\ 0 \end{pmatrix} u(t),\ x(t_0)=x_0,\\ y(t) &= \begin{pmatrix} c_1 & c_2 \end{pmatrix} x(t) + du(t). \end{aligned}$$

Then

$$\text{conmat}(A,B) = (B \;\; AB) = \begin{pmatrix} b_1 & a_{11}b_1 \\ 0 & 0 \end{pmatrix},\ \text{rank}(\text{conmat}(A,B)) = 1 < 2 = n_x.$$

Thus the system is not controllable. Note that,

$$x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} \Rightarrow x_2(t+1) = a_{22}x_2(t),\ x_2(t_0) = x_0,$$

thus x_2 is not influenced at all by the input.

Proposition 21.2.9. Kalman control canonical form. *For any time-invariant linear control system*

$$x(t+1) = Ax(t) + Bu(t),\ x(t_0) = x_0,\ x : T \to \mathbb{R}^{n_x},$$

there exists a state-space transformation of the form, see Proposition 17.4.6,

$$\bar{x}(t) = Lx(t),\ L \in \mathbb{R}^{n_x \times n_x}_{nsng}$$

such that with respect to the new basis the system representation is in the Kalman control-canonical form*,*

$$\begin{aligned} \bar{x}(t+1) &= \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} \bar{x}(t) + \begin{pmatrix} B_1 \\ 0 \end{pmatrix} u(t),\ \bar{x}(t_0) = \bar{x}_0 = Lx_0,\\ &\exists\, n_{x_1}, n_{x_2} \in \mathbb{N},\ \textit{such that } n_{x_1} + n_{x_2} = n_x,\\ &A_{11} \in \mathbb{R}^{n_{x_1} \times n_{x_1}},\ A_{22} \in \mathbb{R}^{n_{x_2} \times n_{x_2}},\ A_{12} \in \mathbb{R}^{n_{x_1} \times n_{x_2}},\ B_1 \in \mathbb{R}^{n_{x_1} \times n_u},\\ &(A_{11}, B_1) \textit{ is a controllable pair.} \end{aligned}$$

If $n_{x_1} = n_x$ hence $n_{x_2} = 0$ then the control system is controllable. If $n_{x_1} = 0$ hence $n_{x_2} = n_x$ then the system is fully uncontrollable. For $0 < n_{x_1} < n_x$ the system is decomposed in two subsystems one of which is controllable and the other is fully uncontrollable because $B_2 = 0$.

Proof. If $\text{rank}(\text{conmat}(A,B)) = n_{x_1} < n_x$ then choose a basis for,

$$\text{conmat}(A,B) = \begin{pmatrix} B & AB & \ldots & A^{n_x-1}B \end{pmatrix} \subseteq \mathbb{R}^{n_x},$$

and complete that basis to a basis of $\mathbb{R}^{n_x}$. See Proposition 17.4.6 for a construction.

Note that then by construction,

$$\begin{aligned} \bar{x}(t+1) &= LAL^{-1}\bar{x}(t) + LBu(t),\\ &\text{conmat}(LAL^{-1}, LB) = \begin{pmatrix} \text{conmat}(A_{11}, B_1) \\ 0 \end{pmatrix}. \end{aligned}$$

□

Stabilizability

Recall from Section 17.4 and in particular from the subsection of Spectral Theory, that the *spectrum* of a matrix $A \in \mathbb{R}^{n_x \times n_x}$, denoted by $\text{spec}(A)$, is the set of eigenvalues of A. The matrix A is called *exponentially stable* (in discrete-time) if $\text{spec}(A) \subset \text{D}_o = \{c \in \mathbb{C} | \, |c| < 1\}$.

A matrix with elements in the real numbers is such that if an eigenvalue is complex valued as $\lambda = x + iy$ then also the complex number $\overline{\lambda} = x - iy$ is an eigenvalue. Hence complex eigenvalues occur in conjugate pairs. Keep this in mind in the following definition.

Definition 21.2.10. *Stabilizability conditions.*
Consider a time-invariant linear control system with $(A,B,C,D) \in \text{LSP}(n_x,n_u,n_y)$. The matrix tuple (A,B) will be called a *stabilizable tuple* if one of the following equivalent conditions hold.

(a)The matrix tuple (A,B) satisfies the condition that, if the system is transformed to the Kalman control canonical form of Proposition 21.2.9, then the condition $\text{spec}(A_{22}) \subset \text{D}_o$ holds.
(b)The complex number $\lambda \in \text{spec}(A)$ will be called (A,B)-*stabilizable* if either $\lambda \in \text{D}_o$ or if it is (A,B)-controllable.
(c)The complex number $\lambda \in \text{spec}(A)$ will be called (A,B)-*spectrally stabilizable* if $|\lambda| \geq 1$ and $x \in \mathbb{C}^{n_x}$ with $x^T A = x^T \lambda, \; x^T B = 0$, imply that $x = 0$. The tuple $(A, \; B)$ is called a *spectrally stabilizable pair* if every eigenvalue $\lambda \in \text{spec}(A)$ is spectrally (A,B) stabilizable.

Theorem 21.2.11. Existence of a stabilizing control law equivalent with stabilizability of the system. *Consider a time-invariant linear control system with* $(A,B,C,D) \in \text{LSP}(n_x,n_u,n_y)$.

(a)The stabilizability conditions of Def. 21.2.10 are equivalent.
(b)The following statements are equivalent:

(b.a)There exists a matrix $F \in \mathbb{R}^{n_u \times n_x}$ such that $\text{spec}(A+BF) \subset \text{D}_o$*;*
(b.b)One of the equivalent conditions of stabilizability of Def. 21.2.10 holds.

Proof. (a) These properties are well known in the theory of linear systems.
(b) One uses (a) with $S = \text{D}_o$.
$(\Leftarrow)$ Let $\lambda \in \text{spec}(A) \cap (\text{D}_o)^c$. Then $|\lambda| \geq 1$ and if $x \in \mathbb{C}^{n_x}$, with $x^T A = x^T \lambda, x^T B = 0$, then by assumption $x = 0$. Hence $\lambda \in \text{spec}(A) \cap \text{D}_o$ is (A,B) controllable. The result follows from (a).
$(\Rightarrow)$ By (a) every $\lambda \in \text{spec}(A) \subset (\text{D}_o)^c$ is (A,B) controllable. Thus if $\lambda \in \text{spec}(A)$, $|\lambda| \geq 1$, and $x \in \mathbb{C}^{n_x}$, with $x^T A = x^T \lambda, x^T B = 0$, then $x = 0$. Hence $\lambda \in \text{spec}(A)$ is (A,B) stabilizable.

□

Controllability after Feedback

The following technical result will be used elsewhere in this book.

Proposition 21.2.12. *Consider a time-invariant linear control system and system matrices of the sizes,*
$A \in \mathbb{R}^{n_x \times n_x}$, $G \in \mathbb{R}^{n_x \times n_{u_1}}$, $H \in \mathbb{R}^{n_x \times n_{u_2}}$, $J \in \mathbb{R}^{n_x \times n_{u_2}}$, *and* $V \in \mathbb{R}^{n_{u_2} \times n_x}$. *Recall the concept of an* (A,B) *stabilizable pair, see Def.21.2.10.*

(a)If (A,G) *is a spectrally stabilizable pair and if* $GG^T + HH^T = JJ^T$,
then $((A+HV),J)$ *is a stabilizable pair.*
(b)If (A,G) *is a controllable pair and if* $GG^T + HH^T = JJ^T$,
then $((A+HV),J)$ *is a controllable pair.*

Proof. (a) Let $\lambda \in \text{spec}(A+HV), x \in \mathbb{C}^{n_x}$ with $|\lambda| \geq 1, x^T(A+HV) = x^T\lambda, x^T J = 0$. Then, $x^T GG^T x + x^T HH^T x = x^T JJ^T x = 0$, hence $x^T G = 0$ and $x^T H = 0$. This and $x^T(A+HV) = x^T\lambda$ imply that $x^T A = x^T\lambda$. Because (A,G) is a stabilizable pair, $\lambda \in \text{spec}(A)$ is (A,G) stabilizable. This, $|\lambda| \geq 1$, $x^T A = x^T\lambda$, and $x^T G = 0$ imply that $x = 0$. Then $\lambda \in \text{spec}(A+HV)$ is $((A+HV),J)$ stabilizable, and by Def. 21.2.10 $(A+HV,J)$ is a spectrally stabilizable pair and by Theorem 21.2.11 a stabilizable pair.
(b) The proof is analogous to that of (a). □

Proposition 21.2.12 can be used to establish stabilizability or controllability of a system after the system has been changed due to a linear control law based on state feedback.

Corollary 21.2.13. *Consider a linear control system and a linearcontrol law based on state feedback,*

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t),\ x(0) = x_0, \\ g(x) &= Fx + Lv, \\ x(t+1) &= (A+BF)x(t) + BLv(t), \end{aligned}$$

where v *is a new input signal possible of a different dimension that the input* u. *Assume that,* $BLL^T B^T = BB^T + BB^T$. *The latter condition can be satisfied if* $L = I_{n_u}\sqrt{2}$.

It then follows from Proposition 21.2.12 that, if (A,B) *is a stabilizable pair then* $(A+BF,BL)$ *is a stabilizable pair.*

The simple proof is omitted.

Controllability of a Time-Varying Linear Control System

A characterization is needed elsewhere in the book of controllability of a time-varying linear control system. The relevant concepts are defined.

Definition 21.2.14. *Reachability matrix and reachability map*. Consider a time-varying linear control system,

$$\begin{aligned} x(t+1) = A(t)x(t) + B(t)u(t),\ x(0) = x_0, \\ T = \{0,1,\ldots,t_1\},\ \text{or } T = \mathbb{N}, \\ X = \mathbb{R}^{n_x},\ U = \mathbb{R}^{n_u},\ Y = \mathbb{R}^{n_y},\ n_x,\ n_u,\ n_y \in \mathbb{Z}_+. \end{aligned}$$

Define the time-varying *state transition function* as the map,

$$\begin{aligned} &\Phi : T \times T \to \mathbb{R}^{n_x \times n_x}, \\ \Phi(t,s) = &\begin{cases} A(t-1)A(t-2)\ldots A(s+1)A(s), & \text{if } s < t, \\ A(s), & \text{if } s = t-1, \\ I, & \text{if } s = t, \\ \text{undefined}, & \text{if } s > t. \end{cases} \\ &\Phi(t,t) = I,\ \Phi(t+1,t) = A(t), \\ &\Phi(t+1,s) = A(t)\Phi(t,s),\ \forall\ s,\ t \in T,\ s < t. \end{aligned}$$

Recall that then,

$$x(t) = \Phi(t,0)x_0 + \sum_{s=0}^{t-1} \phi(t-1,s)B(s)u(s).$$

Define respectively the *reachability matrix* and the *reachability map* by the formulas,

$$\begin{aligned} &\{t_0, t_0+1, \ldots, t_1\} \subseteq T, \\ L_{rmat}(t_0 : t_1 - 1) = &\begin{pmatrix} B(t_1-1) & \Phi(t_1-1 : t_1-2)B(t_1-2) & \ldots \\ \ldots & \Phi(t_1-1, t_0+2)B(t_0+1) & \Phi(t_1-1 : t_0+1)B(t_0) \end{pmatrix} \\ &\in \mathbb{R}^{n_x \times (n_u(t_1 - t_0 - 1))}, \\ L_r(t_0 : t_1) = &\ L_{rmat}(t_0 : t_1 - 1)u(t_0 : t_1 - 1), \\ u(t_0 : t_1 - 1)^T = &\left(u(t_1-1)\ u(t_1-2)\ \ldots\ u(t_0+1)\ u(t_0) \right). \end{aligned}$$

Theorem 21.2.15. Characterization of controllability of a time-varying linear control system.

(a)The following statements are equivalent:

1. *A time-varying linear control system is controllable on the interval $T(t_0 : t_1)$.*
2. *the corresponding reachability matrix is full rank,*

   $$n_x = \text{rank}(L_{rmat}(t_0 : t_1 - 1)).$$

   A necessary condition is that $t_1 - t_0 - 1 \geq n_x$.
3. *The* controllability Grammian matrix *is nonsingular,*

   $$\begin{aligned} &\det(W_r(t_0 : t_1)) \neq 0, \\ W_r(t_0 : t_1) = &\sum_{s=0}^{t_1 - 1} \Phi(t_0 : s)B(s)B(s)^T \Phi(t_0 : s)^T. \end{aligned}$$

(b)If the control system is controllable then for any initial state $x_0 \in X$ there exists an input trajectory $u(t_0 : t_1 - 1) \in U(t_0 : t_1 - 1)$ transferring the system to the zero state; equivalently, such that,

$$\begin{aligned} &\forall\, x_0 \in X,\ \exists\, u(t_0 : t_1 - 1) \in U(t_0 : t_1 - 1), \\ 0 &= \Phi(t_1 : t_0)x_0 + L_{rmat}(t_0 : t_1)u(t_0 : t_1 - 1). \end{aligned}$$

21.3 Observability

Consider a linear control system,

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t),\ x(t_0) = x_0, \\ y(t) &= Cx(t) + Du(t). \end{aligned}$$

According to Proposition 21.1.6

$$y(t) = CA^{t-t_0}x_0 + \Sigma_{s=t_0}^{t-1} CA^{t-1-s}Bu(s) + Du(t).$$

Observability of a control system concerns the question: Is it possible to uniquely determine the initial state $x_0 \in X$ from the output trajectory on the time interval T_1, $y_{T_1} = \{y(t_0), y(t_0+1), \ldots, y(t_1-1)\}$, and from the input trajectory $u_{T_1} = \{u(t_0), u(t_0+1), \ldots, u(t_1-1)\}$?

According to the above formula, the answer to this question depends on the input trajectory $u_{T_1} \in F(T_1, U)$.

Another motivation to investigate observability is the problem of constructing a stable observer defined next.

Problem 21.3.1. *Observer of a linear system which has a stable error system.* Consider a linear system with representation,

$$\begin{aligned} x(t+1) &= Ax(t),\ x(0) = x_0, \\ y(t) &= Cx(t),\ \ T = \mathbb{N},\ X = \mathbb{R}^{n_x},\ Y = \mathbb{R}^{n_y}. \end{aligned}$$

Consider a linear observer of this system with the system representation,

$$\hat{x}(t+1) = A\hat{x}(t) + K(y(t) - C\hat{x}(t)),\ \hat{x}(0) = \hat{x}_0,\ K \in \mathbb{R}^{n_x \times n_y}.$$

Define the error of the observer as the system,

$$\begin{aligned} e(t) &= x(t) - \hat{x}(t),\ e : T \to \mathbb{R}^{n_x}, \\ e(t+1) &= (A - KC)e(t),\ e(0) = x_0 - \hat{x}_0. \end{aligned}$$

Does there exist a matrix K such that the system matrix $(A-KC)$ of the error system satisfies, $\text{spec}(A-KC) \subset \text{D}_o$? If $S \subset \text{D}_o \subset \mathbb{C}$ is symmetric with respect to the real axis does there then exist a matrix K such that $\text{spec}(A-KC) \subset S \subset \text{D}_o$?

Observability will be related to the above problem later in this section.

Definition 21.3.2. *Observability and co-observability of a system.* Consider a control system.

(a) The initial states $x_0,\ \bar{x}_0 \in X$ are said to be *indistinguishable on the interval* T_1 *and for the particular input trajectory* u_{T_1},

$$\begin{aligned} T_1 &= \{t_0, t_0+1, \ldots, t_1-1\} \subseteq T = \mathbb{N},\ u_{T_1} \in F(T_1, U) \text{ if} \\ y_{T_1} &= \bar{y}_{T_1},\ \text{ where} \\ &\quad (x_{T_1}, y_{T_1}, x_1) = \phi(T_1, x_0, u_{T_1}),\ (\bar{x}_{T_1}, \bar{y}_{T_1}, \bar{x}_1) = \phi(T_1, \bar{x}_0, u_{T_1}). \end{aligned}$$

(b) The states $x_0, \bar{x}_0 \in X$ are said to be *indistinguishable on the interval* T_1 if for all input functions $u_{T_1} \in F(T_1, U)$ they are indistinguishable on the interval T_1. They are said to be *indistinguishable* if there exists a time interval T_1 such that they are indistinguishable on the interval T_1.

(c) The *system* is said to be *observable on the interval* T_1 if there does not exist a pair of distinct states $x_0,\ \bar{x}_0 \in X$ which is indistinguishable on T_1. The *system* is said to be *observable* if there exists a time interval $T_1 \subseteq T$ such that every tuple of states $x_0,\ \bar{x}_0 \in X$ is observable on that interval.

Definition 21.3.3. *Observability matrix and an* (A,C)*-observable tuple of system matrices.* Consider a time-invariant linear control system,

$$\begin{aligned} &(T, \mathbb{R}^{n_x}, \mathbb{R}^{n_u}, \mathbb{R}^{n_y}, A, B, C, D), \\ x(t+1) &= Ax(t) + Bu(t),\ x(t_0) = x_0, \\ y(t) &= Cx(t) + Du(t). \end{aligned}$$

(a) Define the *observability matrix* of this system as the matrix,

$$\text{obsm}(A,C) = \begin{pmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n_x-1} \end{pmatrix} \in \mathbb{R}^{(n_y n_x) \times n_x}.$$

Call the pair of matrices $(A,C) \in \mathbb{R}^{n_x \times n_x} \times \mathbb{R}^{n_y \times n_x}$ an *observable pair* if $\text{rank}(\text{obsm}(A,C)) = n_x$.

(b) The eigenvalue $\lambda \in \text{spec}(A)$ will be called (A,C)*-observable eigenvalue* if,

$$n_x = \text{rank} \begin{pmatrix} A - \lambda I \\ C \end{pmatrix}.$$

The matrix tuple (A,C) is called *eigenvalue-observable* if all eigenvalues of the matrix A are (A,C)-observable.

(c) The complex number $\lambda \in \text{spec}(A)$ will be called (A,C) *spectrally observable eigenvalue* if $x \in \mathbb{C}^{n_x}$ with $Ax = \lambda x$ and $Cx = 0$, implies that $x = 0$. The pair (A,C) is called a *spectrally observable pair* if every eigenvalue $\lambda \in \text{spec}(A)$ is a (A,C)-spectrally observable eigenvalue.

Theorem 21.3.4. *Consider a time-invariant linear control system,*

$$\begin{aligned}
&(T,\mathbb{R}^{n_x},\mathbb{R}^{n_u},\mathbb{R}^{n_y},A,B,C,D),\\
x(t+1) &= Ax(t)+Bu(t),\ x(t_0)=x_0,\\
y(t) &= Cx(t)+Du(t).
\end{aligned}$$

(a)The observability conditions of Def. 21.3.3 are equivalent.
(b)The following statements are equivalent:

- *(b.i)For every symmetric subset $S \subset \mathbb{C}$ there exists a matrix $K \in \mathbb{R}^{n_x \times n_y}$ such that* $\text{spec}(A-KC) \subset S$.
- *(b.ii)The matrix tuple (A,C) satisfies one of the equivalent definitions of Def. 21.3.3 of being an observable pair.*

Example 21.3.5. *A non-observable linear system.* Consider the particular time-invariant linear control system,

$$\begin{aligned}
&(\mathbb{Z},\mathbb{R}^2,\mathbb{R},\mathbb{R},\mathbb{R},A,B,C,D),\ a_{11},\ a_{21},\ a_{22},\ c_1 \in \mathbb{R},\ c_1 \neq 0,\\
x(t+1) &= \begin{pmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{pmatrix} x(t)+Bu(t),\ x(t_0)=x_0,\\
y(t) &= \begin{pmatrix} c_1 & 0 \end{pmatrix} x(t)+Du(t);\\
\text{obsm}(A,C) &= \begin{pmatrix} C \\ CA \end{pmatrix} = \begin{pmatrix} c_1 & 0 \\ c_1 a_{11} & 0 \end{pmatrix},\ \text{rank}(\text{obsm}(A,C)) = 1 < 2 = n_x,
\end{aligned}$$

thus the system is not observable. Take $u \equiv 0$. Then,

$$\begin{aligned}
y(t) &= CA^{t-t_0}x_0 = \begin{pmatrix} c_1 a_{11}^{t-t_0} & 0 \end{pmatrix} x_0;\ \ x_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix},\ \bar{x}_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \ \Rightarrow\\
y(t;t_0,x_0) &= CA^{t-t_0}x_0 = c_1 a_{11}^{t-t_0} x_{0,1} = c_1 a_{11}^{t-t_0} \bar{x}_{0,1} = CA^{t-t_0}\bar{x}_0 = \bar{y}(t;t_0,\bar{x}_0).
\end{aligned}$$

hence x_0 and $\bar{x}_0$ are indistinguishable.

Proposition 21.3.6. Kalman observable-canonical form. *For any time-invariant linear control system,*

$$\begin{aligned}
x(t+1) &= Ax(t)+Bu(t),\ x(t_0)=x_0,\ x: T \to \mathbb{R}^n,\\
y(t) &= Cx(t)+Du(t),
\end{aligned}$$

there exists a state-space transformation $\bar{x}(t) = Lx(t)$, $L \in \mathbb{R}^{n_x \times n_x}_{nsng}$, see Proposition 17.4.6, such that with respect to the new basis the system representation is in the Kalman observable-canonical form*,*

$$\begin{aligned}
\bar{x}(t+1) &= \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix} \bar{x}(t)+LBu(t),\ \ \bar{x}(t_0)=\bar{x}_0=Lx_0,\\
y(t) &= \begin{pmatrix} C_1 & 0 \end{pmatrix} \bar{x}(t)+Du(t),\\
&\quad n_{x_1},n_{x_2} \in \mathbb{N},\ n_{x_1}+n_{x_2}=n_x,\\
&\quad A_{11} \in \mathbb{R}^{n_{x_1} \times n_{x_1}},\ A_{22} \in \mathbb{R}^{n_{x_2} \times n_{x_2}},\ A_{21} \in \mathbb{R}^{n_{x_2} \times n_{x_1}},\ C_1 \in \mathbb{R}^{n_y \times n_{x_1}},\\
&\quad (A_{11},C_1) \text{ is an observable pair.}
\end{aligned}$$

If $n_{x_1} = n_x$ and hence $n_{x_2} = 0$ then the system is observable. If $n_{x_1} = 0$ and hence $n_{x_2} = n_x$ then the system is called fully unobservable. *For $0 < n_{x_1} < n_x$ the system is decomposed into two subsystems one of which is observable and the other is fully unobservable because $C_2 = 0$.*

Proof. If $\text{rank}(\text{obsmat}(A,C)) = n_{x_1} < n_x$ then choose a basis for the range space of the observability matrix,

$$\text{obsmat}(A,C) = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n_x} \end{pmatrix},$$

and complete that basis to a basis of $\mathbb{R}^{n_x}$. The result then follows by simple algebraic operations. Note that then by construction,

$$\begin{aligned} \bar{x}(t+1) &= LAL^{-1}\bar{x}(t) + LBu(t),\ \bar{x}(0) = \bar{x}_0 = Lx_0, \\ y(t) &= CL^{-1}\bar{x}(t) + Du(t), \\ &\quad \text{obsmat}(LAL^{-1}, CL^{-1}) = \begin{pmatrix} \text{obsmat}(A_{11}, C_1) & 0 \end{pmatrix}. \end{aligned}$$

□

Definition 21.3.7. A time-invariant linear control system is said to be in the *Kalman decomposition* if it has the representation,

$$\begin{aligned} x(t+1) &= \begin{pmatrix} A_{11} & 0 & A_{13} & 0 \\ A_{21} & A_{22} & A_{23} & A_{24} \\ 0 & 0 & A_{33} & 0 \\ 0 & 0 & 0 & A_{44} \end{pmatrix} x(t) + \begin{pmatrix} B_1 \\ B_2 \\ 0 \\ 0 \end{pmatrix} u(t),\ x(t_0 = x_0, \\ y(t) &= \begin{pmatrix} C_1 & 0 & C_3 & 0 \end{pmatrix} x(t) + Du(t),\ \text{with} \\ &\quad \left(\begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}, \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} \right),\ \text{a controllable pair,} \\ &\quad \left(\begin{pmatrix} A_{11} & A_{13} \\ 0 & A_{33} \end{pmatrix}, \begin{pmatrix} C_1 & C_3 \end{pmatrix} \right),\ \text{an observable pair,} \\ &\quad n_{x_1},\ n_{x_2},\ n_{x_3},\ n_{x_4} \in \mathbb{N},\ n = n_{x_1} + n_{x_2} + n_{x_3} + n_{x_4}. \end{aligned}$$

In particular cases, one or more of the components of x may not be present.

Any linear system can be transformed to the Kalman decomposition though in particular cases one or more components of the state may be missing.

If one is only interested in the relation between inputs and outputs, then the above system can also be described by the controllable and observable system representation,

$$\begin{aligned} x_1(t+1) &= A_{11}x_1(t) + B_1u(t),\ x_1(0) = x_{0,1}, \\ y(t) &= C_1x_1(t) + Du(t), \\ &\quad (A_{11}, B_1)\ \text{controllable pair and}\ (A_{11}, C_1)\ \text{observable pair.} \end{aligned}$$

This form of system reduction is further explored in the realization theory of linear systems below in this chapter.

Next a weaker form of observability of a linear system is investigated, detectability.

Definition 21.3.8. *Detectability conditions*. Consider a time-invariant linear control system with $\{A,B,C,D\} \in \text{LSP}(n_x,n_u,n_y)$. The matrix tuple (A,C) is called *detectable* if one of the following equivalent conditions holds.

(a) If the system representation is in the form of a the Kalman observable-canonical form of Proposition 21.3.6 then $\text{rank}(A_{22}) \subset \text{D}_o$.

(b) Any eigenvalue of the system matrix A, $\lambda \in \text{spec}(A)$, either satisfies $\lambda \in \text{D}_o$ or is an (A,C)-observable eigenvalue hence satisfies,

$$n_x = \text{rank}\begin{pmatrix} A - \lambda I \\ C \end{pmatrix}.$$

(c) The complex number $\lambda \in \text{spec}(A)$ is called an (A,C)-*detectable eigenvalue* if $|\lambda| \geq 1$ and $x \in \mathbb{C}^{n_x}$ with $Ax = \lambda x,\ Cx = 0$ imply that $x = 0$. The pair (A,C) is called a *spectrally detectable pair* if every eigenvalue $\lambda \in \text{spec}(A)$ either satisfies $\lambda \in \text{D}_o$ or is an (A,C)-detectable eigenvalue.

Theorem 21.3.9. Characterization of the existence of a stable observer error system. *Consider a time-invariant linear control system with* $\{A,B,C,D\} \in \text{LSP}(n_x,n_u,n_y)$.

(a) The conditions of Def. 21.3.8 are equivalent.

(b) There exists a matrix $K \in \mathbb{R}^{n_x \times n_y}$ *such that* $\text{spec}(A - KC) \subset \text{D}_o$ *if and only if the pair of matrices* (A,C) *is a detectable pair.*

The proof is analogous to that of Theorem 21.2.11 and therefore omitted.

21.4 Geometric Approach to Linear Systems

A linear system is defined to have vector spaces as state space and as input space and to operate with linear maps. Therefore one can work with the geometric viewpoint of linear systems and regard a linear system as defined in vector spaces with linear maps. This geometric viewpoint has the advantage that the system does not directly depend on its representation with respect to a particular basis. Invariance properties are simple to formulate and to investigate. The approach also allows the use of abstract algebra for the operations. A minor negative aspect is that robustness properties cannot be handled directly. Robustness requires a relation with analysis which can of course be formulated.

W.M. Wonham, [75] and later editions, with co-workers, has formulated the geometric framework of linear systems and derived the main theorems. The approach has inspired other researchers to develop a geometric approach to infinitely-differentiable nonlinear control systems, [27], and to other sets of systems.

Geometric framework of linear system is defined by vector spaces, which represents the state space, the input space, and the output space, and by linear maps of the system. This applies to both continuous-time and a discrete-time linear systems.

Concepts of the geometric viewpoint of linear systems are: an invariant subspace of an autonomous linear system, a controllable subspace, and the unobservable subspace. For particular problems, such as zero-output dynamics, or input-to-state maps, or existence of a left-inverse of a linear system, one can associate more specific subspaces.

Control theory of linear systems then relates the existence of a solution of a control problem to the existence of a particular subspace. A procedure to construct that particular subspace is then provided. In case one wants to carry out the construction for a particular linear control system then one has to choose a basis and carry out the calculations with respect to that basis.

Control of Gaussian stochastic control systems does not make much use of the geometric approach because the effects of the noise are usually assume to affect the entire state space so there is little need for subspaces. It is possible to formulate concepts of subspaces of stochastic systems in terms of probability distribution on those spaces.

Control theory has become enriched by the geometric approach to linear control systems because of the emphasis on geometry, the use of algebra, and the generalizations to other sets of control systems.

21.5 Zero-Output Dynamics

The reader may learn in this section about the dynamics of a linear system when the output trajectory is constrained to be zero. The concepts for the description of this phenomenon are useful for several system theoretic problems. This section is a precursor to the section on inverses of linear systems which follows.

The description of this section is incomplete. Several subjects of system theory for discrete-time linear systems have been insufficiently investigated. The author does not have the time to develop the required results. Needed are more explicit concepts and characterizations of those concepts. The concepts are used to formulate necessary and sufficient conditions for the existence of a left-inverse and of a right-inverse of a linear system. For continuous-time linear systems the literature is in a better state.

The formulations of the concepts of this section are in terms of trajectories of the input, the state, and the output of a linear system. This formulation allows generalization to other subsets of control systems. Therefore the definitions and the conditions are different from those of observability of a linear system. The zero-output dynamics of a linear system involves states, inputs, and outputs. It relates to the map from an initial state and an input to the resulting output trajectory.

The early motivation to study the dynamics of a linear system when the output trajectory was restrained to be zero, came from H. Rosenbrock, [55], and from circuit theory. See also the book by F.M. Callier and C.A. Desoer, [8, Sec. 11.7]. A broader view is offered in the book by T. Kailath, [30, Subsection 6.5.3, Sections 6.5 and 6.7].

For discrete-time linear systems several concepts were formulated by B.P. Molinari, [50, 51]. There does not seem a more recent reference where the discrete-time case is developed. Therefore the definitions and theorems stated below are referred to the book [68, Ch. 7] which deals only with the case of a continuous-time linear system. Several of the definitions and theorems stated below are reformulated for the discrete-time case.

Definition 21.5.1. Consider a time-invariant linear system with representation,

$$\begin{aligned} x(t+1) &= Ax(t)+Bu(t),\ x(0)=x_0, \\ y(t) &= Cx(t)+Du(t),\ \ T=\mathbb{N},\ X=\mathbb{R}^{n_x},\ U=\mathbb{R}^{n_u}. \end{aligned}$$

Call the state $x_0 \in X$ a *weakly-unobservable state* if,

$$\exists\, u : T \to U, \text{ such that } \forall\, t \in T,\ y(t;(0,x_0),u) = 0.$$

In words, there exists an input function such that, when the system is started at time $t=0$ at state x_0, the output is identically zero for all times. Define then the *weakly-unobservable subspace* of the state space X as the set,

$$X_{wuo} = \{x_0 \in X |\ x_0 \text{ is a weakly-unobservable state}\}.$$

Because the state trajectory and the output trajectory are linear functions of the initial state and of the sequence of inputs, the set X_{wuo} is a subspace of the state space.

Below is illustrated how the interest into zero-output dynamics was first formulated as an extension of the corresponding transfer function.

Definition 21.5.2. Consider the time-invariant linear control system of Def. 21.5.1. Call the real number $z \in \mathbb{R}$ a *zero* of the linear system if,

$$\text{rank}\begin{pmatrix} A-zI_{n_x} & B \\ C & D \end{pmatrix} < \min\{(n_x+n_u),\ (n_x+n_y)\}.$$

If $n_u \leq n_y$ and if $z \in \mathbb{R}$ is a zero then there exists a nonzero *state-input vector* such that,

$$0 = \begin{pmatrix} A-zI_{n_x} & B \\ C & D \end{pmatrix} \begin{pmatrix} x_0 \\ u_0 \end{pmatrix}.$$

If $n_y \leq n_u$ and if $z \in \mathbb{R}$ a zero then there exists a nonzero *state-output vector*, called an *annihilator*, such that,

$$0 = \begin{pmatrix} x_1 \\ y_0 \end{pmatrix}^T \begin{pmatrix} A-zI_{n_x} & B \\ C & D \end{pmatrix}.$$

Call the linear system as having *only strictly stable zeros* if any zero of this linear system is located strictly inside the unit disc D_o. In the context of transfer functions, this condition is described as a *minimum phase condition* of the associated transfer function in the complex domain.

The above definition has to be extended to the case where there exists a tuple of complex-conjugate zeroes, even more general, a set of complex-conjugate zeroes.

Example 21.5.3. *A linear system and its zero*. Consider a time-invariant linear system with the representation,

$$\begin{aligned} n_y &= 1,\ n_x = 2,\ n_u = 1, \\ x(t+1) &= Ax(t) + Bu(t),\ x(0) = x_0, \\ y(t) &= Cx(t) + Du(t), \\ A &= \begin{pmatrix} 0 & 1 \\ 1/4 & 0 \end{pmatrix},\ B = \begin{pmatrix} 1 \\ 1 \end{pmatrix},\ C = \begin{pmatrix} 1 & 1 \end{pmatrix},\ D = 1, \\ & \text{spec}(A) = \{1/2,\ -1/2\},\ z = 1/3,\ x_0 = \begin{pmatrix} -6 \\ -1/3 \end{pmatrix},\ u_0 = 1. \end{aligned}$$

This linear system has a zero $z = 1/3$ while the initial state x_0 belongs to the weakly unobservable subspace hence there exists an input starting with $u_0 = 1$ such that the trajectory of the system output is identically zero.

Example 21.5.4. *Single-input-single-output linear system with a zero.* Consider the linear system,

$$\begin{aligned} x(t+1) &= ax(t) + bu(t),\ x(0) = x_0, \\ y(t) &= cx(t) + du(t), \\ & n_y = 1,\ n_x = 1,\ n_u = 1,\ a,\ b,\ c,\ d \in \mathbb{R},\ d \neq 0. \end{aligned}$$

This system has a zero,

$$\begin{aligned} z_0 &= \frac{ad - bc}{d}, \\ A(z_0) &= \begin{pmatrix} a - z_0 & b \\ c & d \end{pmatrix},\ \text{rank}(A(z_0)) = 1 < 2,\ \det(A(z_0)) = 0. \end{aligned}$$

Proposition 21.5.5. *Consider the time-invariant linear control system of Def. 21.5.1. If $n_u \leq n_y$ and if there exists a zero $z \in \mathbb{R}$ of the linear control system then,*

$$\exists\, x_0 \in X,\ \exists\, u \in U^T,\ \textit{such that}\ \forall\, t \in T,\ y(t; (0, x_0), u) = 0.$$

hence x_0 is a weakly-observable state.

Proof. If $z \in \mathbb{R}$ is a zero of the linear control system then by Def. 21.5.2 the system matrix of that system is column-rank deficient hence,

$$\exists\, x_0 \in X,\ \exists\, u_0 \in U,\ (x_0,\ u_0) \neq 0 \ \Rightarrow\ \begin{pmatrix} A - zI_{n_x} & B \\ C & D \end{pmatrix} \begin{pmatrix} x_0 \\ u_0 \end{pmatrix} = 0.$$

Define the input function, $u(t) = z^t\ u_0$ for all $t \in T$. Then,

$$\begin{aligned}
\begin{pmatrix} x(1) \\ y(0) \end{pmatrix} &= \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x_0 \\ u_0 \end{pmatrix} = \begin{pmatrix} A - zI & B \\ C & D \end{pmatrix} \begin{pmatrix} x_0 \\ u_0 \end{pmatrix} + \begin{pmatrix} zx_0 \\ 0 \end{pmatrix} = \begin{pmatrix} zx_0 \\ 0 \end{pmatrix}; \\
&\text{suppose that } s = 0,1,\ldots t,\ x(t) = z^t\, x_0;\ \Rightarrow \\
\begin{pmatrix} x(t+1) \\ y(t) \end{pmatrix} &= \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x(t) \\ u(t) \end{pmatrix} = \begin{pmatrix} A - zI & B \\ C & D \end{pmatrix} \begin{pmatrix} x_0 \\ u_0 \end{pmatrix} z^t + \begin{pmatrix} z^{t+1}x_0 \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} z^{t+1}x_0 \\ 0 \end{pmatrix};\ \forall\, t \in T,\ y(t;(0,x_0),u) = 0.
\end{aligned}$$

□

Proposition 21.5.6. *Consider the time-invariant linear control system of Def. 21.5.1. If the initial state $x_0 \in X_{wuo}$ belongs to the weakly-unobservable subspace and the input trajectory $u \in U^T$ is such that the output trajectory is zero for all times then the state trajectory corresponding to x_0 and u satisfies that for all $t \in T$, $x(t;(0,x_0),u) \in X_{wuo}$, hence belongs also to the weakly-unobservable subspace.*

Proof. Let $u \in U^T$ be such as in the proposition statement. Then,

$$\begin{aligned}
\begin{pmatrix} x(1) \\ y(0) \end{pmatrix} &= \begin{pmatrix} x(1) \\ 0 \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x_0 \\ u_0 \end{pmatrix};\ u_s : T \to U,\ u_s(t) = u(t+1), \\
&\Rightarrow \forall\, t \in T,\ y(t;(0,x(1)),u_s) = y(t+1;(0,x_0),u) = 0; \\
&\quad \Rightarrow\ x(1) \in X_{wuo}.
\end{aligned}$$

The proof completes by induction. □

The weakly-unobservable subspace can be characterized in terms of a subspace condition as is the custom in the geometric approach to linear control systems.

Theorem 21.5.7. *Consider the time-invariant linear control system of Def. 21.5.1.*

(a)The weakly-unobservable subspace X_{wuo} is the largest subspace $V \subseteq X$ such that,

$$\begin{pmatrix} A \\ C \end{pmatrix} \subset (V \times \{0\}) + \mathrm{Im} \begin{pmatrix} B \\ D \end{pmatrix}.$$

(b)The weakly-unobservable subspace X_{wuo} is the largest subspace $V \subseteq X$ for which,

$$\exists\, F : X \to U, \text{ a linear map such that } (A+BF)V \subseteq V,\ (C+DF)V = 0.$$

Definition 21.5.8. Consider the time-invariant linear control system of Def. 21.5.1. Define the *controllable weakly-unobservable subspace* $X_{cwuo} \subseteq X$ as the set,

$$X_{cwuo} = \left\{ \begin{array}{l} x_0 \in X \mid \exists\, u \in U^T,\ \exists\, t_1 \in T, \text{ such that} \\ \forall\, t \in \{0,\ldots,t_1\},\ y(t;(0,x_0),u) = 0; \text{ and } x(t_1;(0,x_0),u) = 0 \end{array} \right\}.$$

The interpretation of this subspace is that from any initial state in this subset the system can be reduced to the zero state in a finite number of steps while keeping the output equal to zero during that interval.

Theorem 21.5.9. *Consider any feedback matrix F which makes the subspace X_{wuo} invariant, see Thm. 21.5.7.(b). Then the controllable weakly-unobservable subspace X_{cwuo} is characterized as the smallest subspace containing the subspace shown below and closed with respect to the indicated system matrix,*

$$X_{cwuo} = < (A+BF) |\, X_{wuo} \cap (B \ker(D)) >,$$
$$\ker(D) = \{u_s \in U |\, Du_s = 0\},\; B\ker(D) = \{x \in X |\, \exists\, u_s \in U,\; x = Bu_s,\; Du_s = 0\}.$$

For any such linear feedback map F the following relations hold,

$$(A+BF)X_{cwuo} \subseteq X_{cwuo},\;\; (C+DF)X_{cwuo} = 0.$$

Definition 21.5.10. Consider the time-invariant linear control system of Def. 21.5.1. This linear control system is called *strongly-observable system* if,

$$\forall\, x_0 \in X,\; \forall\, u \in U^T,\; (\forall\, t \in T,\; y(t;(0,x_0),u) = 0) \;\Rightarrow\; x_0 = 0.$$

Theorem 21.5.11. *The following statements are equivalent:*

(a)The linear control system is strongly observable.
(b)$X_{wuo} = \{0\}$.
(c)For any feedback matrix $F \in \mathbb{R}^{n_u \times n_x}$, the matrix tuple $(A+BF, C+DF)$ is an observable pair.

The next concept of a strongly co-controllable state was introduced by B.P. Molinari for a discrete-time linear system in [51]. See also for a continuous-time linear system the book [68, Section 8.3].

Definition 21.5.12. Consider a time-invariant linear system. A state $x_{scc} \in X$ is called *strongly co-controllable with zero output* if there exists a time $-t_1 \in \mathbb{Z}_-$ and a finite input trajectory, $u : T_1 = \{0,-1,-2,\ldots,-t_1\} \subset \mathbb{Z}_-$ such that $x(0) = 0$, $x_{-t_1} = x_{scc}$, and $y(s) = 0$ for $s = 0,-1,\ldots,-t_1$. Denote the *strongly co-controllable subspace with zero output* as,

$$X_{scc} = \{x_{scc} \in X |\, \text{is strongly co-controllable with zero output}\}.$$

Proposition 21.5.13. *Consider a time-invariant linear system. The state $x_{scc} \in X$ is strongly co-controllable with zero output if and only if there exists a time $-t_1 \in \mathbb{Z}_-$ and a finite input trajectory, $u : T_1 = \{0,-1,-2,\ldots,-t_1\} \subset \mathbb{Z}_-$ such that the following backward linear system generates the following system trajectories,*

$$x(t-1) = Ax(t) + Bu(t),\; x(0) = 0,$$
$$y(t-1) = Cx(t) + Du(t) = 0,\; \forall\, t = 0,-1,-2,\;\ldots,\; -t_1,$$
$$x(-t_1) = x_{scc}.$$

The following result has been proven to hold for continuous-time linear systems. It is conjectured to hold as well for discrete-time linear systems.

Proposition 21.5.14. *Consider a time-invariant linear system.*

(a)The strongly co-controllable subspace with zero output X_{scc} is the smallest subspace $V \subseteq X$ *for which there exists a linear map* $K : Y \to X$ *such that,*

$$(A+KC)V \subseteq V,\ \mathrm{Im}(C+KD) \subseteq V.$$

(b)The strongly co-controllable subspace with zero output is the smallest subspace $V \subseteq X$ *for which*

$$\begin{pmatrix} A & B \end{pmatrix} (V \times U) \cap \ker \begin{pmatrix} C & D \end{pmatrix} \subseteq V.$$

There has been formulated an algorithm to construct the strongly co-controllable subspace with zero output, see [68, Sec. 8.3].

The subject of an *input to state map* of a linear system is complementary to the zero-output dynamics. The concepts used are useful for control theory. The author has not found the required concepts for time-invariant linear systems in the literature.

21.6 Inverse of a Linear System

The theory of filtering and of stochastic control requires the concept of the inverse of a linear system. This concept is also used in information theory and in the research area of communications.

The reader may want to read, before proceeding, Def. 17.4.19 for the inverse of a matrix and subsequent results.

Definition 21.6.1. Consider a time-invariant linear control system representation,

$$\begin{aligned} x(t+1) &= Ax(t)+Bu(t),\ x(0)=x_0, \\ y(t) &= Cx(t)+Du(t). \end{aligned}$$

The following linear control system is said to be a *left-inverse system* of the above considered linear system with delay $t_d \in \mathbb{N}$ if,

$$\begin{aligned} &\forall\, u : T \to U = \mathbb{R}^{n_u}\ \exists\, \overline{x}_0 \in \overline{X}, \\ \overline{x}(t+1) &= F\overline{x}(t)+Gy(t),\ \overline{x}(0)=\overline{x}_0,\ n_{\overline{x}} \in \mathbb{N},\ \overline{x}(t) \in \overline{X} = \mathbb{R}^{n_{\overline{x}}}, \\ \overline{u}(t) &= H\overline{x}(t)+Jy(t); \\ \begin{pmatrix} x(t+1) \\ \overline{x}(t+1) \end{pmatrix} &= \begin{pmatrix} A & 0 \\ GC & F \end{pmatrix} \begin{pmatrix} x(t) \\ \overline{x}(t) \end{pmatrix} + \begin{pmatrix} B \\ GD \end{pmatrix} u(t); \\ \overline{u}(t) &= \begin{pmatrix} JC & H \end{pmatrix} \begin{pmatrix} x(t) \\ \overline{x}(t) \end{pmatrix} + JDu(t); \text{such that,} \\ \overline{u}(t+t_d) &= u(t),\ \forall\, t \in T. \end{aligned}$$

A *right-inverse system* is defined correspondingly and requires that $y(t+t_d) = \overline{y}(t)$ for all $t \in T$.

See [56, Ex. 5.7, p. 203] for an example where no left-inverse system exists for $t_d = 0,\ 1,\ 2,\ 3$, but one exists for $t_d = 4$.

A procedure to construct a left-inverse system of a linear system in a finite state set was formulated by J.L. Massey and M.K. Sain, [47]. The first publication of an inverse of a linear control system in the tuples of the real numbers of L.M. Silverman, [58], dates from approximately the same time. L.M. Silverman, [61], has also formulated a procedure which, if it completes successfully, establishes that a left-inverse system exists and produces a left-inverse system.

Needed is a simple algebraic condition which is equivalent to the existence of a left-inverse of a *discrete-time linear control system*. Such an equivalent condition was formulated for a continuous-time linear control system for the existence of a left-inverse system and of a right-inverse system, see the paper [24] and the book [68, Sec. 8.2]. The conditions are in terms of linear subspaces and injectiveness of the map,

$$u \mapsto \begin{pmatrix} B \\ D \end{pmatrix} u \in X \times Y.$$

No publication is know to the author where these conditions are formulated for a discrete-time linear control system.

Other references on invertibility of linear control systems are [5, 56, 47, 72].

If the system matrix D satisfies a rank condition then a left-inverse system and a right-inverse system exist both with zero delay, as is formulated below.

Proposition 21.6.2. *Consider a time-invariant linear control system with representation,*

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t),\ x(0) = x_0, \\ y(t) &= Cx(t) + Du(t),\ \ D \in \mathbb{R}^{n_y \times n_u},\ n_u \leq n_y,\ \text{rank}(D) = n_u. \end{aligned}$$

Then there exists an inverse system with the roles of the input u and the output y interchanged, having the representation,

$$\begin{aligned} x_{inv}(t+1) &= [A - B(D^TD)^{-1}D^TC]x_{inv}(t) + B(D^TD)^{-1}D^Ty(t),\ x(0) = x_0, \\ u_{inv}(t) &= -(D^TD)^{-1}D^TCx_{inv}(t) + (D^TD)^{-1}D^Ty(t). \end{aligned}$$

Proof. From the fact that $\text{rank}(D) = n_u$ it follows with Lemma 17.4.3 that $D^TD \succ 0$, hence $(D^TD)^{-1} \in \mathbb{R}^{n_u \times n_u}$ exists. One calculates,

$$\begin{aligned} (D^TD)^{-1}D^Ty(t) &= (D^TD)^{-1}D^TCx(t) + (D^TD)^{-1}(D^TD)u(t) \\ &= (D^TD)^{-1}D^TCx(t) + u(t) \\ x(t+1) &= Ax(t) + Bu(t) \\ &= [A - B(D^TD)^{-1}D^TC]x(t) + B(D^TD)^{-1}D^Ty(t), \\ u(t) &= -(D^TD)^{-1}D^TCx(t) + (D^TD)^{-1}D^Ty(t); \\ u_{inv}(t) &= u(t). \end{aligned}$$

□

Definition 21.6.3. Consider a time-invariant linear control system with representation,

$$\begin{aligned} x(t+1) &= Ax(t)+Bu(t),\ x(0)=x_0, \\ y(t) &= Cx(t)+Du(t),\ \ \text{rank}(D)=n_u. \end{aligned}$$

(a) Call the system *inverse exponentially stable* if the associated inverse system of Proposition 21.6.2, is exponentially stable; thus if, $\text{spec}(A-B(D^TD)^{-1}D^TC) \subset \text{D}_o$. This condition is related to the corresponding linear system having only strictly stable zeros, Def. 21.5.2.

(b) Define the system matrices of the inverse system,

$$\begin{aligned} A_{inv} &= A-B(D^TD)^{-1}D^TC \in \mathbb{R}^{n_x \times n_x}, \\ C_{inv}^TC_{inv} &= C^TC - C^TD(D^TD)^{-1}D^TC,\ \ C_{inv} \in \mathbb{R}^{n_x \times n_x}. \end{aligned}$$

Call the system *inverse compensated-variance observable* if (A_{inv}, C_{inv}) is an observable pair. It can be proven that the condition does not depend on the factorization of C_{inv} chosen above. Call the system *inverse compensated-variance detectable* if (A_{inv}, C_{inv}) is a detectable pair.

Proposition 21.6.4. *Consider a time-invariant linear control system with representation,*

$$\begin{aligned} x(t+1) &= Ax(t)+Bu(t),\ x(0)=x_0, \\ y(t) &= Cx(t)+Du(t),\ \text{rank}(D)=n_u. \end{aligned}$$

Define the associated dual linear control system *by the representation,*

$$\begin{aligned} \bar{x}(t+1) &= A^T\bar{x}(t)+C^Tu(t),\ \bar{x}(0)=x_0, \\ \bar{y}(t) &= B^T\bar{x}(t)+D^Tu(t);\ \textit{assume that,} \\ & n_u = \text{rank}(D^T) = \text{rank}(D),\ \ D^T \in \mathbb{R}^{n_y \times n_u}. \end{aligned}$$

Then there exists an inverse control system of the dual control system in which the roles of the input u and the output y interchanged, having the representation,

$$\begin{aligned} \bar{x}(t+1) &= [A^T - C^T(DD^T)^{-1}DB^T]\bar{x}(t) + C^T(DD^T)^{-1}D\bar{y}(t),\ \bar{x}(0)=x_0, \\ u(t) &= -(DD^T)^{-1}DB^T\bar{x}(t) + (DD^T)^{-1}D\bar{y}(t). \end{aligned}$$

Proof. From the fact that $\text{rank}(D^T)=n_u$ it follows with Lemma 17.4.3 that $DD^T \succ 0$, hence $(DD^T)^{-1} \in \mathbb{R}^{n_u \times n_u}$ exists. One calculates,

$$\begin{aligned} (DD^T)^{-1}D\bar{y}(t) &= (DD^T)^{-1}DB^T\bar{x}(t) + (DD^T)^{-1}(DD^T)u(t) \\ &= (DD^T)^{-1}DB^T\bar{x}(t) + u(t) \\ \bar{x}(t+1) &= A^T\bar{x}(t) + C^Tu(t) \\ &= [A^T - C^T(DD^T)^{-1}DB^T]\bar{x}(t) + C^T(DD^T)^{-1}D\bar{y}(t), \\ u(t) &= -(DD^T)^{-1}DB^T\bar{x}(t) + (DD^T)^{-1}D\bar{y}(t). \end{aligned}$$

□

Definition 21.6.5. Consider a time-invariant linear control system with representation,

$$\begin{aligned} x(t+1) &= Ax(t)+Bu(t),\ x(0)=x_0, \\ y(t) &= Cx(t)+Du(t),\ \ \text{rank}(D)=n_u. \end{aligned}$$

Define the associated dual system with the representation,

$$\begin{aligned} x(t+1) &= A^T x(t)+C^T u(t),\ x(0)=x_0, \\ y(t) &= B^T x(t)+D^T u(t);\ \text{assume that,} \\ & n_u = \text{rank}(D^T)=\text{rank}(D),\ \ D^T \in \mathbb{R}^{n_y \times n_u}. \end{aligned}$$

Consider also the inverse of the dual system of Proposition 21.6.4.

(a) Call the system *inverse-dual exponentially stable* if the inverse of the dual system of Proposition 21.6.4, is exponentially stable,

$$\text{spec}(A^T - C^T(DD^T)^{-1}DB^T) = \text{spec}(A - BD^T(DD^T)^{-1}C) \subset \mathrm{D}_o.$$

This condition is called the *dual minimum-phase condition* of the transfer function of the dual system in the frequency domain.

(b) Define the system matrices of the inverse system,

$$\begin{aligned} A_{inv,d}^T &= A^T - C^T(DD^T)^{-1}DB^T \in \mathbb{R}^{n_x \times n_x}, \\ A_{inv,d} &= A - BD^T(DD^T)^{-1}C \in \mathbb{R}^{n_x \times n_x}, \\ B_{inv,d}B_{inv,d}^T &= BB^T - BD(DD^T)^{-1}DB^T,\ \ B_{inv,d}^T \in \mathbb{R}^{n_x \times n_x}. \end{aligned}$$

Call the system *inverse-dual compensated-variance observable* if $(A_{inv,d}^T, B_{inv,d}^T)$ is an observable pair which is equivalent with $(A_{inv,d}, B_{inv,d})$ being a controllable pair. Call the system *inverse-dual compensated-variance detectable* if $(A_{inv,d}^T, B_{inv,d}^T)$ is a detectable pair which is equivalent with $(A_{inv,d}, B_{inv,d})$ being an observable pair.

21.7 Canonical Factorization of a Deterministic Map

The reader finds in this section an exposition of the canonical factorization of a map. This is to be regarded as an introduction to realization theory which is based on this canonical factorization.

Definition 21.7.1. Consider two sets U and Y, and a map $f: U \to Y$.

(a) A *realization* of the triple (U, Y, f) is a tuple (X, U, Y, g, h) where X is a set, and $g: U \to X$ and $h: X \to Y$ are maps such that

$$\forall u \in U,\ \ f(u) = h(g(u)). \tag{21.1}$$

The factorization (21.1) is also described by the phrase 'the diagram commutes' where the corresponding diagram is displayed in Figure 21.3.

(b) A *canonical realization* of the triple (U,Y,f) is a realization (X,U,Y,g,h) such that (1) the function g is surjective, and (2) the function h is injective. It is also called a *canonical form* for (U,Y,f). The use of that term is justified below in this section.

(c) Two canonical realizations (X_1,U,Y,g_1,h_1) (X_2,U,Y,g_2,h_2) are said to be *bijectively related* if there exists a bijection $s : X_1 \to X_2$ such that (1) $g_2 = s \circ g_1$ and (2) $h_1 = h_2 \circ s$. Equivalently, if the diagrams of Figure 21.4 commute. (The map s is a bijection if it is surjective and injective.)

(d) A realization (X_2,U,Y,g_2,h_2) is said to be a *reduction* of realization (X_1,U,Y,g_1,h_1) if there exists a map $s : X_1 \to X_2$ which is surjective and for which the diagram Figure 21.4 commutes.

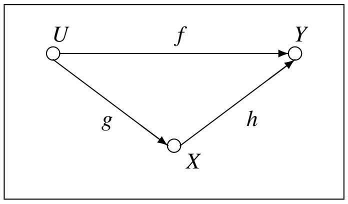


Fig. 21.3 Diagram of a canonical factorization of a map.

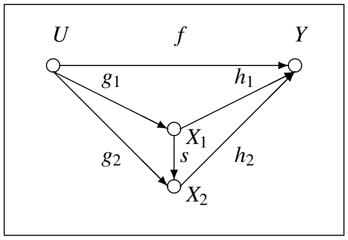


Fig. 21.4 Diagram of a bijective relation of a canonical factorization of a map.

Problem 21.7.2. *Realization problem of a map.* Consider a triple (U,Y,f) with U,Y sets and $f : U \to Y$.

(a) Does there exist a realization (X,U,Y,g,h) of (U,Y,f)?

(b) Does there exist a canonical realization of (U,Y,f)?

(c) Is a canonical realization unique or, if there are many canonical realizations, how are these related?

Consider a triple (U,Y,f). Suppose that the map f is not injective. Define the relation $\sim$ on U by $u_1 \sim u_2$ if $f(u_1) = f(u_2)$. It follows directly from the definition of this relation that it is an equivalence relation. Note that if $y \in f(U)$ then there may be two or many $u \in U$ such that $y = f(u)$. The approach is to construct a map, based on f, which is injective. The concept of a canonical factorization provides such a description. This concept equals the quotient set of set theory and of universal algebra.

Theorem 21.7.3. Existence of a canonical realization. *Consider the triple (U,Y,f) as defined above.*

(a)There exists a canonical realization (X,U,Y,g,h) of (U,Y,f). Procedure 21.7.4 produces a canonical realization (X_c,U,Y,g_c,h_c) from the considered realization.
(b)Consider a canonical realization (X,U,Y,g,h) of (U,Y,f).

- *If $u_1,\ u_2 \in U$ are such that $f(u_1) = f(u_2)$ then $g(u_1) = g(u_2)$.*
- *For any $u \in U$ there exists an $x \in X$ such that $f(u) = h(x)$. For any $x \in X$ there exists an $u \in U$ such that $h(x) = f(u)$. Hence $f(U) = h(X)$.*

Procedure 21.7.4 Procedure to reduce an arbitrary realization to a canonical one. *The procedure is described in the proof of Theorem 21.7.3.(a).*

Note that this is not an algorithm as regarded in computer science because there is no claim that the procedure yields a result after a finite number of steps.

Proof. Proof of Theorem 21.7.3.
(a) Define on (U,Y,f) the relation

$$\sim = \{(u_1,u_2) \in U \times U | f(u_1) = f(u_2)\}.$$

From this definition follows directly that it is an equivalence relation on U. Define the coset $U/\sim$ and denote for $u \in U$ the corresponding element of the coset by $[u] \in U/\sim$. Define,

$$\begin{aligned} X &= U/\sim, \\ g(u) &= [u],\ g : U \to X, \\ h(x) &= f(u),\ x \in X = U/\sim \ \Rightarrow\ \exists\, u \in U \text{ such that } x = [u],\ h : X \to Y. \end{aligned}$$

Because of the definition of $\sim$, $h : X \to Y$ is well defined; if $u_1,u_2 \in [u] = x$ then $f(u_1) = f(u_2)$ hence $h(x) = f(u_1) = f(u_2)$.

The properties (1) and (2) of a canonical factorization are proved next. First it is a realization because for $u \in U$ $h(g(u)) = h([u]) = h(x) = f(u)$ by definition of g and h.
(1) If $x \in X = U/\sim$ then there exists an $u \in U$ such that $x = [u]$. Hence $g(u) = [u] = x$ and g is surjective.
(2) Consider $x_1,x_2 \in X$ such that $h(x_1) = h(x_2)$. From the definition of h follows that $f(u_1) = h(x_1) = h(x_2) = f(u_2)$ for $x_1 = [u_1]$ and $x_2 = [u_2]$. But $f(u_1) = f(u_2)$ implies that $(u_1,u_2) \in \sim$ hence $x_1 = [u_1] = [u_2] = x_2$. Thus h is injective.
Item 1. Let $u_1,u_2 \in X$ satisfy $u_1 \sim u_2$ or $f(u_1) = f(u_2)$. Then by commutativity $h(g(u_1)) = f(u_1) = f(u_2) = h(g(u_2))$ and, because h injective, $g(u_1) = g(u_2)$.

Item 2. Consider $x \in X$. Because g is surjective by the above statements, there exists a $u \in U$ such that $x = g(u)$. Then, by commutativity, $f(u) = h(g(u)) = h(x) \in h(X)$. For the same reason $h(x) = h(g(u)) = f(u) \in f(U)$. Thus $h(X) = f(U)$. □

Theorem 21.7.5. Canonical realizations are bijectively related. *Consider the triple (U,Y,f). Suppose there exist two canonical realizations of this triple, (X_1,U,Y,g_1,h_1) and (X_2,U,Y,g_2,h_2). Then these realizations are bijectively related hence there exists a bijection $s : X_1 \to X_2$ such that the diagram of Figure 21.4 commutes.*

The proof of the above theorem follows from the following lemma whose statement is particular because of a partly corresponding result of the literature.

Lemma 21.7.6. *Consider a tuple (U,Y,f) and two realizations of it, (X_1,U,Y,g_1,h_1) and (X_2,U,Y,g_2,h_2).*

(a)If g_1 is surjective and h_2 is injective then there exists an unique $s : X_1 \to X_2$ such that $g_2 = s \circ g_1$ and $h_1 = h_2 \circ s$ (This statement is known as Zeiger's fill-in lemma.)
(b)Assume there exists a $s : X_1 \to X_2$ such that $g_2 = s \circ g_1$ and $h_1 = h_2 \circ s$. If g_2 is surjective then s is surjective.
(c)Assume there exists a $s : X_1 \to X_2$ such that $g_2 = s \circ g_1$ and $h_1 = h_2 \circ s$. If h_1 is injective then s is injective.

Proof. (a) (1) Because g_1 is surjective, for every $x_1 \in X_1$, there exists a $u \in U$ such that $x_1 = g_1(u)$. Thus, if

$$g_1^{-1}(x_1) = \{u \in U | g(u) = x_1\}$$

then $g_1^{-1}(x_1) \neq \emptyset$. Define $s : X_1 \to X_2$ by $s(x_1) = g_2(u)$ if $u \in g_1^{-1}(x_1)$.
(2) It will be proven that the function s is well defined. Consider $u_1, u_2 \in g_1^{-1}(x_1) \subseteq U$. Then

$$\begin{aligned}
x_1 &= g_1(u_1) = g_1(u_2), \text{ by definition of } g_1^{-1}(x_1),\\
h_2(g_2(u_1)) &= h_1(g_1(u_1)), \text{ because } h_2 \circ g_2 = h_1 \circ g_1,\\
&= h_1(g_1(u_2)), \text{ by the above,}\\
&= h_2(g_2(u_2)), \text{ by } h_2 \circ g_2 = h_1 \circ g_1,\\
&\Rightarrow s(x_1) = g_2(u_1) = g_2(u_2), \text{ because } h_2 \text{ is injective.}
\end{aligned}$$

Thus s is well defined.
(3) Consider $u \in U$ and $x_1 = g_1(u_1)$. Then $u \in g_1^{-1}(x_1)$, and by definition of s, $s(x_1) = g_2(u)$. Thus $g_2(u) = s(x_1) = s(g_1(u))$, or $g_2 = s \circ g_1$. Consider $x_1 \in X_1$. Because g_1 is surjective there exists a $u \in U$ such that $u \in g_1^{-1}(x_1)$. By definition of s, $s(x_1) = g_2(u)$. Then,

$$h_2(s(x_1)) = h_2(g_2(u)) = h_1(g_1(u)) = h_1(x_1),$$

or $h_2 \circ s = h_1$. Thus the diagram commutes.
(4) It will be proven that a function satisfying the properties of $s : X_1 \to X_2$ is unique.

Suppose there exist a $s_1 : X_1 \rightarrow X_2$ such that $g_2 = s_1 \circ g_1$, $h_1 = h_2 \circ s_1$. Consider $x_1 \in X_1$. Then $h_2(s_1(x_1) = h_1(x_1) = h_2(s(x_1))$, and injectivity of h_2 implies that $s(x_1) = s_1(x_1)$. Thus $s = s_1$.
(b) Consider $x_2 \in X_2$. Because g_2 is surjective there exists a $u \in U$ such that $x_2 = g_2(u)$. Define $x_1 = g_1(u)$. Then $x_2 = g_2(u) = s(g_1(u)) = s(x_1)$. Thus s is surjective.
(c) Consider $x_{1,1}, x_{1,2} \in X_1$ be such that $s(x_{1,1}) = s(x_{1,2})$. Then
$h_1(x_{1,1}) = h_2(s(x_{1,1})) = h_2(s(x_{1,2})) = h_1(x_{1,2})$, and injectivity of h_1 implies that $x_{1,1} = x_{1,2}$. Thus s is injective. □

21.8 Realization Theory for Linear Systems

Consider a linear control system,

$$\begin{aligned} &(T, X, U, Y, \mathbb{R}, A, B, C, D), \\ x(t+1) &= Ax(t) + Bu(t), \; x(t_0) = x_0, \\ y(t) &= Cx(t) + Du(t). \end{aligned}$$

Recall the definition of the impulse response function of this system but extend it to the time index set $T = \mathbb{Z}$ according to,

$$H(t) = \begin{cases} D, & \text{if } t = 0, \\ CA^{t-1}B, & \text{if } t = 1, 2, \ldots, \\ 0, & \text{if } t < 0, \end{cases} \quad H : \mathbb{Z} \rightarrow \mathbb{R}^{n_y \times n_u}.$$

The relation between an input and an output function of this system is then, in case the initial state at time $t = 0$ equals $x_0 = 0$,

$$y(t) = \sum_{s=0}^{\infty} H(t-s)u(s) = Du(t) + \sum_{s=t_0}^{t-1} CA^{t-s-1}Bu(s).$$

The realization problem is considered next. Consider an impulse response function, say obtained from observations of a phenomenon. The realization problem asks for the existence of a linear control system such that the impulse response function of the constructed control system equals the considered impulse response function.

First it is argued that from a phenomenon one can by choice of particular input functions, obtain in principle its impulse response function. Choose the input function,

$$u_i(t) = \begin{cases} 0, & \text{if } t < 0, \\ e_i, & \text{if } t = 0, \\ 0, & \text{if } \; t > 0, \end{cases}$$

where e_i is the i-th unit vector of $\mathbb{R}^{n_x}$. Then for $i = 1, 2, \ldots, n_y$,

$$y_i(t) = \Sigma_{s=0}^{\infty} H(t-s)u(s) = H(t)e_i,$$

hence one can in principle determine $H : \mathbb{N} \rightarrow \mathbb{R}^{n_y \times n_x}$.

The realization problem asks for: Is it possible to determine for an arbitrary impulse response function H an integer $n_x \in N$ and matrices $(A,B,C,D) \in \text{LSP}(n_x,n_u,n_y)$ of a linear system such that $H(0) = D$ and for all $t \in \mathbb{Z}_+$, $H(t) = CA^{t-1}B$? The realization problem is now motivated.

Problem 21.8.1. *The realization problem for a time-invariant finite-dimensional linear control system.* Consider a discrete-time impulse response function $H : \mathbb{N} \to \mathbb{R}^{n_y \times n_u}$, $n_y,\ n_u \in \mathbb{Z}_+$.

(a) Does there exist a linear control system with representation,

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t),\ x(t_0) = x_0, \\ y(t) &= Cx(t) + Du(t), \\ & (T,\mathbb{R}^{n_x},\mathbb{R}^{n_u},\mathbb{R}^{n_y},A,B,C,D) \in \text{LS},\ \text{with } n_x \in \mathbb{Z}_+,\ \text{such that}, \\ H(t) &= \begin{cases} D, & \text{if } t = 0, \\ CA^{t-1}B, & \text{if } t > 0? \end{cases} \end{aligned}$$

If so, then call this control system a *realization* of the impulse response function.

(b) Call a realization *minimal* if the dimension $n_x \in \mathbb{N}$ of the state space $X = \mathbb{R}^{n_x}$ is the smallest of all state-space dimensions of realizations of the impulse response function H. Characterize those linear control systems for which the system is a minimal realization of its impulse response function. Construct for an impulse response function H, a minimal realization.

(c) If there exist two minimal realizations of an impulse response function, what is the relation between these two realizations?

The above problem with the three major questions form the realization problem of system theory. Below the problem is discussed for a time-invariant finite-dimensional linear control system. The corresponding problem for other classes of control systems may be formulated also. The realization problem is the main problem of system theory. The solution to the realization problem is provided in below in this section after the introduction of new concepts.

Definition 21.8.2. An *input-output linear system* is a tuple,

$$\begin{aligned} & (T,U,Y,\mathbf{U},y_0,H,a) \in IOLS,\ \text{where}, \\ & T = \{0,1,\ldots,t_1\} \subset \mathbb{N},\ \text{or } T = \mathbb{N},\ \text{called the time-index set}, \\ U &= \mathbb{R}^{n_u},\ n_u \in \mathbb{Z}_+,\ \text{called the } \textit{input space}, \\ Y &= \mathbb{R}^{n_y},\ n_y \in \mathbb{Z}_+,\ \text{called the } \textit{output space},\ y_0 \in Y, \\ & H : T \to \mathbb{R}^{n_y \times n_u},\ \text{called the } \textit{impulse response function}, \\ \mathbf{U} & \subseteq \{u : T \to U\},\ \text{called the } \textit{set of input functions}, \\ a(T,u) &= y,\ a : T \times \mathbf{U} \to Y^T, \\ y(t) &= y_0 + \sum_{s=t_0}^{t_1-1} H(t,s)u(s),\ \forall t \in T. \end{aligned}$$

The system is called *time-invariant* if for all $t,s \in T$, $H(t,s) = H(t-s,0)$. In that case one denotes $H : T \to \mathbb{R}^{n_y \times n_u}$, $H(t) = H(t,0)$.

Proposition 21.8.3. *Consider a time-invariant finite-dimensional linear control system,*

$$\begin{aligned} &(T,\mathbb{R}^{n_x},\mathbb{R}^{n_u},\mathbb{R}^{n_y},x_0,\mathbf{U},a,b),\ \textit{with}\ X_0 \subseteq X, \\ x(t+1) &= Ax(t)+Bu(t),\ x(t_0)=x_0 \in X_0 \subseteq X = \mathbb{R}^{n_x}, \\ y(t) &= Cx(t)+Du(t). \end{aligned}$$

The zero-initial-state-response*, for $x_0 = 0 \in X_0$, is an input-output linear system with,*

$$\begin{aligned} &(T,\mathbb{R}^{n_u},\mathbb{R}^{n_y},\mathbf{U},y_0,H,a), \\ y_0 &= 0 \in \mathbb{R}^{n_y} = Y,\ H : T \to \mathbb{R}^{n_y \times n_u},\ H(t) = CA^{t-1}B, \\ a(T,u) &= y,\ \ y(t) = Du(t) + \sum_{s=t_0}^{t-1} H(t-s)u(s),\ \forall t \in T. \end{aligned}$$

Proof. This result follows directly from the solution of time-invariant linear system with $x_0 = 0$,

$$y(t) = Du(t) + \sum_{s=t_0}^{t-1} CA^{t-s-1}Bu(s) = Du(t) + \sum_{s=t_0}^{t-1} H(t-s)u(s).$$

□

In general terms, the realization problem for a finite-dimensional linear control system starts with a linear operator from inputs to outputs and asks whether there exists a finite-dimensional linear control system such that the input-output map of this system equals the considered map. The existence requires a condition. If a system exists then it will be called a *realization* of the considered input-output map.

A realization is not unique in general. Attention is therefore restricted to minimal realizations, for which the state space dimension is minimal, and this subclass needs to be characterized. The theory for this is described below in this chapter.

There follow two realization problems which differ in the specification of what is considered.

Problem 21.8.4. The *realization problem for a time-invariant finite-dimensional linear control system in terms of input-output functions*. Consider the sets,

$$\begin{aligned} &T \in T_{intv}(\mathbb{R}),\ T = [t_0,t_1],\ t_0 < t_1, \\ U &= \mathbb{R}^{n_u},\ Y = \mathbb{R}^{n_y},\ n_u,n_y \in \mathbb{Z}_+, \\ \mathbf{UY}(\mathbf{T}) &\subseteq \mathbf{U}(\mathbf{T}) \times \mathbf{Y}(\mathbf{T}),\ \text{a subset of tuples of input-output functions.} \end{aligned}$$

(a) Determine necessary and sufficient conditions for the existence of a time-invariant finite-dimensional linear control system,

$$\begin{aligned} &(T,X,U,Y,X_0,\mathbf{U},(A,B,C,D)) \in LS, \\ x(t+1) &= Ax(t)+Bu(t),\ x(t_0)=x_0 \in X_0,\ \ \text{with}\ X = \mathbb{R}^{n_x},\ n_x \in \mathbb{Z}_+, \\ y(t) &= Cx(t)+Du(t), \end{aligned}$$

such that, if $(\overline{u},\overline{y}) \in \mathbf{UY}(\mathbf{T})$, then there exists an initial condition $x_0 \in X_0$ such that for all $t_1 \in [t_0,\infty)$,

$$(t_1,x_1) = a(T,x_0,\overline{u}),\ y(t) = \overline{y}(t),\ \forall t \in T.$$

Call such a system, if it exists, a *realization of the set of input-output trajectories* $\mathbf{UY}(\mathbf{T})$.

(b) If a realization exists, characterize those systems which are minimal realizations of the set of input-output trajectories $\mathbf{UY}(\mathbf{T})$. A realization is said to be a *minimal realization* if there does not exist a realization with state space $X_1 = \mathbb{R}^{n_1}$ and $n_1 < n_x$.

(c) Classify all minimal realizations of the set $\mathbf{UY}(\mathbf{T})$.

Problem 21.8.5. The *realization problem for a time-invariant linear system in terms of the impulse response function*. Consider an input-output linear system,

$$\begin{aligned} &(T,\mathbb{R}^{n_u},\mathbb{R}^{n_y},\mathbf{U},y_0,H,a),\\ T &= [t_0,\infty),\ U = \mathbb{R}^{n_u},\ Y = \mathbb{R}^{n_y},\ n_u,n_y, \in \mathbb{Z}_+,\ L \in \mathbb{R}^{n_y \times n_u},\\ &y_0 \in Y,\ H : T \to \mathbb{R}^{n_y \times n_u},\\ a(T,u) &= y,\ a : T \times \mathbf{U} \to Y^T,\\ y(t) &= y_0 + Lu(t) + \sum_{s=t_0}^{t-1} H(t-s)u(s),\ \forall t \in T. \end{aligned}$$

An input-output linear system may also be considered as an operator,

$$\begin{aligned} &M : \mathbf{U} \to \mathbf{Y},\ \ \mathbf{U} \subseteq \{u : T \to U\},\ \mathbf{Y} \subseteq \{y : T \to Y\},\\ M(u)(t) &= y(t) = Lu(t) + \sum_{s=t_0}^{t-1} H(t-s)u(s),\ \ \forall t \in T. \end{aligned}$$

(a) Determine necessary and sufficient conditions for the existence of a time-invariant linear control system,

$$\begin{aligned} &(T,X,U,Y,X_0,\mathbf{U},(A,B,C,D)) \in TI.FDLS,\\ x(t+1) &= Ax(t) + Bu(t),\ x(t_0) = x_0 \in X_0,\ \ \text{with } X = \mathbb{R}^{n_x},\ n_x \in \mathbb{Z}_+,\\ y(t) &= Cx(t) + Du(t),\ \ \text{such that,}\\ H(t) &= CA^{t-1}B,\ \forall t \in \mathbb{Z}_+,\ L = D. \end{aligned}$$

Call such a system, if it exists, a *realization of the impulse response function.*

(b) Assume a realization exists. Characterize those systems which are minimal realizations. A realization is said to be a *minimal realization* if there does not exist a realization with state space $X_1 = \mathbb{R}^{n_1}$, $n_1 \in \mathbb{Z}_+$, and $n_1 < n_x$.

(c) Classify all minimal realizations of the impulse response function H.

Problem 21.8.5 is equivalent to a realization problem of a transfer function. The transfer function is the Laplace transform or the Fourier transform of the impulse response function.

Problem 21.8.4 is a realization problem formulated in terms of input-output trajectories. It was shown in the references [45] and [69] that Problem 21.8.4 can be reduced to Problem 21.8.5 if conditions are imposed on the set of input trajectories, such conditions have the character of requiring sufficient excitation of the phenomenon.

The solution of Problem 21.8.5 will be stated in Theorem 21.8.9.

Below the impulse response function is related to the Markov parameters.

Definition 21.8.6. Consider a time-invariant input-output linear system,

$$(T,U,Y,\mathbf{U},y_0,H,a),\ T=[t_0,\infty),\ H\in C^\infty(T),\ L\in\mathbb{R}^{n_y\times n_u},$$
$$y(t)=y_0+Lu(t)+\sum_{s=t_0}^{t-1}H(t-s)u(s).$$

Define the *Markov parameters* of the impulse response function H as the function,

$$M:\mathbb{N}\rightarrow\mathbb{R}^{n_y\times n_u},\ \{M(k)\in\mathbb{R}^{n_y\times n_u},\ \forall k\in\mathbb{N}\},$$
$$M(k)=H(k),\ \forall k\in\mathbb{Z}_+,\ M(0)=L.$$

Define the set of *block-Hankel matrices* of the impulse response function H as

$$\mathbf{H}=\{H(k,r)\in\mathbb{R}^{kn_y\times rn_u},\ \forall k,r\in\mathbb{Z}_+\},$$
$$H(k,r)=\begin{pmatrix} M(1) & M(2)\ \ldots & M(r)\\ M(2) & M(3)\ \ldots & M(r+1)\\ \vdots & & \vdots\\ M(k) & M(k+1)\ \ldots & M(k+r-1)\end{pmatrix}\in\mathbb{R}^{kn_y\times rn_u}.$$

The *rank* of the set of block-Hankel matrices is defined as,

$$\text{rank}(\mathbf{H})=\sup_{k,r\in\mathbb{Z}_+}\text{rank}(H(k,r))\in\mathbb{N}\cup\{\infty\}.$$

The statement of the solution of Problem 21.8.5 requires the following concepts for linear systems and for realizations.

Definition 21.8.7. Consider a linear control system,

$$(T,\mathbb{R}^{n_x},\mathbb{R}^{n_u},\mathbb{R}^{n_y},x_0,\mathbf{U},(A,B,C,D)).$$

(a) Define the *set of controllable and observable system matrices* of linear systems by,

$$\text{LSP}_{c,o}(n_x,n_u,n_y)=\left\{\begin{array}{l}(A,B,C,D)\in\text{LSP}(n_x,n_u,n_y)|\\ (A,B)\text{ controllable pair and }(A,C)\text{ observable pair}\end{array}\right\}.$$

(b) Call the following tuples *similar*

$$(A_1,B_1,C_1,D_1),\ (A_2,B_2,C_2,D_2)\in\text{LSP}(n_x,n_u,n_y),$$
$$\text{if }\exists\ L\in\mathbb{R}^{n_x\times n_x}_{nsng}\text{ such that,}$$
$$A_2=LA_1L^{-1},\ B_2=LB_1,\ C_2=C_1L^{-1},\ D_2=D_1;$$
$$\text{notation }(A_1,B_1,C_1,D_1)\overset{L}{\sim}(A_2,B_2,C_2,D_2).$$

The matrix L is called a *state-space transformation* because it corresponds to a change of basis of the state space of the first system to that of the second system.

Definition 21.8.8. (a)Consider a time-invariant linear control system,

$$\begin{aligned}
&(T,\mathbb{R}^{n_x},\mathbb{R}^{n_u},\mathbb{R}^{n_y},X_0,\mathbf{U},(A,B,C,D)) \in TI.FDLS, \\
x(t+1) &= Ax(t)+Bu(t),\ x(t_0)=x_0 \in X_0 \subseteq X, \\
y(t) &= Cx(t)+Du(t).
\end{aligned}$$

The system is said to be a *canonical realization* of its impulse response function if it is (1) controllable and (2) observable.

(b)Consider two time-invariant linear control systems which are realizatons of the same input-output map,

$$\begin{aligned}
&(T,\mathbb{R}^{n_x},\mathbb{R}^{n_u},\mathbb{R}^{n_y},X_{1,0},\mathbf{U},(A_1,B_1,C_1,D_1)) \in TI.FDLS, \\
&(T,\mathbb{R}^{n_x},\mathbb{R}^{n_u},\mathbb{R}^{n_y},X_{2,0},\mathbf{U},(A_2,B_2,C_2,D_2)) \in TI.FDLS, \\
x_1(t+1) &= A_1x_1(t)+B_1u(t),\ x_1(t_0)=x_{1,0} \in X_{1,0} \subseteq X_1, \\
y(t) &= C_1x_1(t)+D_1u(t). \\
x_2(t+1) &= A_2x_2(t)+B_2u(t),\ x_2(t_0)=x_{2,0} \in X_{2,0} \subseteq X_2, \\
y(t) &= C_2x_2(t)+D_2u(t).
\end{aligned}$$

These realizations are said to be *isomorphic* if there exists a state space isomorphism of the form $s: X_1 \to X_2$, $s(x_1)=Lx_1$ with a nonsingular matrix $L \in \mathbb{R}_{nsng}^{n_x \times n_x}$ such that the corresponding state trajectories are related as $x_2(t)=s(x_1(t))=Lx_1(t)$.

Realization Theorem and its Proof

Theorem 21.8.9. *Consider a time-invariant linear input-output system,*

$$(T=\mathbb{N},\mathbb{R}^{n_u},\mathbb{R}^{n_y},0,H,a).$$

(a)There exists a time-invariant finite-dimensional linear control system,

$$\begin{aligned}
&(T,\mathbb{R}^{n_x},\mathbb{R}^{n_u},\mathbb{R}^{n_y},X_0,\mathbf{U},(A,B,C,D)), \\
x(t+1) &= Ax(t)+Bu(t),\ x(t_0)=x_0 \in X_0 \subseteq X, \\
y(t) &= Cx(t)+Du(t),
\end{aligned}$$

such that the impulse response function of this system equals H, or, equivalently,

$$H(t) = CA^{t-1}B,\ \forall t \in \mathbb{Z}_+,\ H(0)=D,$$

if and only if $H(0)=D$ *and* $\text{rank}(\mathbf{H}) < \infty$.

(b)Assume there exists a realization of state-space dimension $n_x = \text{rank}(\mathbf{H}) \in \mathbb{Z}_+$ *as stated in (a). The following three statements are equivalent:*

(b.1)The realization is minimal.
*(b.2)*External characterization. $\text{rank}(\mathbf{H}) = n_x$
*(b.3)*Internal characterization. *The realization is canonical. Equivalently, (A,B) is a controllable pair and (A,C) is an observable pair.*

If one of the above three assumptions holds then there exists a minimal realization.

(c)If a realization is not minimal then the procedure consisting of (1) reduction to controllable form followed by (2) reduction to observable form, results in a minimal realization.

(d)Define the equivalence relation $LSREq_{min}$ on the set of minimal realizations by the condition that the system representations have the same impulse response function. In terms of notation,

$$((A_1,B_1,C_1,D_1),\ (A_2,B_2,C_2,D_2)) \in LSREq_{min},$$
$$\textit{if}\ \forall\ k \in \mathbb{N},\ C_1A_1^kB_1 = C_2A_2^kB_2\ \textit{and}\ D_1 = D_2.$$

The above equivalence relation is characterized by the condition that,

$$\begin{aligned} &((A_1,B_1,C_1,D_1),\ (A_2,B_2,C_2,D_2)) \in LSREq_{min}, \\ &\Leftrightarrow \exists\ L \in \mathbb{R}_{nsng}^{n_x \times n_x}\ \textit{nonsingular such that}, \\ A_2 &= LA_1L^{-1},\ B_2 = LB_1,\ C_2 = C_1L^{-1},\ D_2 = D_1. \end{aligned}$$

Proof. (a) ($\Rightarrow$) Assume that a realization of the impulse response function in the form a time-invariant finite-dimensional linear system exists. Then,

$$\begin{aligned} H(t) &= CA^{t-1}B,\ H(0) = D,\ M(k) = H(k) = CA^{k-1}B,\ \forall k \in \mathbb{Z}_+, \\ H(k,r) &= \begin{pmatrix} M(1) & M(2) & \ldots & M(r) \\ M(2) & M(3) & \ldots & M(r+1) \\ \vdots & & & \vdots \\ M(k) & M(k+1) & \ldots & M(k+r-1) \end{pmatrix} \\ &= \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{k-1} \end{pmatrix} \begin{pmatrix} B & AB & \ldots & A^{r-1}B \end{pmatrix} \in \mathbb{R}^{(kn_y) \times (rn_u)}, \\ &\Rightarrow \text{rank}(H(k,r)) \leq n_x\ \Rightarrow\ \text{rank}(\mathbf{H}) = \sup_{k,r \in \mathbb{Z}_+} \text{rank}(H(k,r)) \leq n_x. \end{aligned}$$

($\Leftarrow$) Define $n_x = \text{rank}(\mathbf{H}) \in \mathbb{Z}_+$. A realization of the impulse response function will be constructed. Define the vector spaces over $\mathbb{R}$,

$$U_\infty = \left\{ \begin{array}{l} u \in \mathbb{R}^\infty | u = (u_1, u_2, \ldots) \\ \forall i \in \mathbb{Z}_n,\ u_i \in \mathbb{R}^{n_u},\ \exists k \in \mathbb{Z}_+,\ \forall i \geq k,\ u_i = 0 \end{array} \right\},$$

$X_\infty = \mathbb{R}^\infty$, and $Y_\infty = \mathbb{R}^\infty$. Define the operator associated with the Hankel matrix $H : U_\infty \to Y_\infty$ as,

$$\begin{aligned}
u &= \begin{pmatrix} u_1 \\ u_2 \\ \vdots \end{pmatrix},\ y = H(u) = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \end{pmatrix}, \\
&\forall i \in \mathbb{Z}_+,\ u_i \in \mathbb{R}^{n_u},\ \forall i \in \mathbb{Z}_+,\ y_i \in \mathbb{R}^{n_y}, \\
y(r) &= \sum_{k=1}^{\infty} M(r+k-1)u(k),\ \text{considered as a formal power series}, \\
y(1) &= \sum_{k=1}^{\infty} M(k)u(k) = M(1)u(1) + M(2)u(2) + \ldots, \\
y(2) &= \sum_{k=1}^{\infty} M(k+1)u(k) = M(2)u(1) + M(3)u(2) + \ldots.
\end{aligned}$$

Then H is a linear operator. Define,

$$X_H = \text{Range}(H) = \{y \in Y_\infty | \exists u \in U_\infty \text{ such that } y = H(u)\} \subseteq\ Y_\infty = \mathbb{R}^\infty = X_\infty.$$

Then X_H is a subspace of X_∞. For $p \in \mathbb{Z}_+$ define the p-fold shift operator $S : X_\infty \to X_\infty$

$$\text{if } x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \end{pmatrix},\ \forall i \in \mathbb{Z}_+,\ x_i \in \mathbb{R},\ \text{ then } S(x) = \begin{pmatrix} x_{p+1} \\ x_{p+2} \\ \vdots \end{pmatrix}.$$

Then S is a linear operator on X_∞. Denote its restriction to X_H, $S : X_H \to X_\infty$, again by S.

Define the dual shift operator as,

$$S^*_{n_y} : X_\infty \to X_\infty,\ S^*_{n_y}(x) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ x_1 \\ x_2 \\ \vdots \end{pmatrix} \in X_\infty,\ \text{if } x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \end{pmatrix},$$

with n_y rows of zeroes in $S^*_{n_y}(x)$. It follows from the definitions of H, S, and S^* that for all $k \in \mathbb{Z}_+$,

$$(SH)_{.,k} = \begin{pmatrix} M(k) \\ M(k+1) \\ M(k+2) \\ \vdots \end{pmatrix} = (HS^*)_{.,k}.$$

Thus $SH = HS^*$ and for any $u \in U_\infty$ then $S(H(u)) = H(S^*(u))$. Consequently, $S : X_H \to X_H$ rather than $S : X_H \to X_\infty$.

Define the operators,

$$\begin{aligned}
&A_\infty : X_\infty \to X_\infty,\ A_\infty(x) = S(x),\\
&B_\infty : \mathbb{R}^{n_u} \to X_\infty,\ B_\infty(e_k) = k\text{-th column of } H;\\
&C_\infty : X_\infty \to \mathbb{R}^{n_y},\\
&C_\infty(x) = \begin{pmatrix} x_1 \\ \vdots \\ x_{n_y} \end{pmatrix} \in \mathbb{R}^{n_y} \text{ if } x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \end{pmatrix} \in X_\infty,\ \ \forall i \in \mathbb{Z}_+,\ x_i \in \mathbb{R},
\end{aligned}$$

the projection on the first n_y coordinates. Next define the operators,

$$\begin{aligned}
&A_H : X_H \to X_\infty,\ A_H(x) = A_\infty(x),\ \forall x \in X_H,\ \text{the restriction of } A_\infty \text{ to } X_H,\\
&B_H = B_\infty : \mathbb{R}^{n_u} \to X_\infty,\\
&C_H : X_H \to \mathbb{R}^{n_y},\ C_H(x) = C_\infty,\ \text{the restriction of } C_\infty \text{ to } X_H.
\end{aligned}$$

Then C_H is a linear operator.

It will be proven that $A_H : X_H \to X_H$. Let $y \in X_H$. Hence there exists a vector $u \in U_\infty = \mathbb{R}^\infty$ such that $y = H(u)$. Then $A_H(y) = A_\infty(y) = S(y)$. Define $v \in U_\infty$ as

$$v = \begin{pmatrix} 0 \\ u(1) \\ u(2) \\ u(3) \\ \vdots \end{pmatrix}.$$

Notice that A_H is well defined because $S : X_H \to X_\infty$ and $y \in X_H$. Then,

$$\begin{aligned}
A_H(y)(r) &= (Sy)(r) = y(r+1) = \sum_{k=1}^{\infty} M(r+1+k-1)u(k)\\
&= \sum_{k=1}^{\infty} M(r+k)u(k) = \sum_{s=2}^{\infty} M(r+s-1)u(s-1),\ s = k+1,\\
&= \sum_{s=1}^{\infty} M(r+s-1)v(s),\ \text{because } v(1) = 0,\\
&= H(v)(r),\ \forall r \in \mathbb{Z}_+ \ \Rightarrow\ A_H(y) \in X_H.
\end{aligned}$$

It will be proven by induction that $C_H A_H^k B_H(I_m) = M(k+1)$ for all $k \in \mathbb{N}$. The definition of $B_H = B_\infty$ implies that $B_H(I_m) \in X_H$ is the first column of H,

$$B_H(I_m) = \begin{pmatrix} M(1) \\ M(2) \\ M(3) \\ \vdots \end{pmatrix} \ \Rightarrow\ C_H B_H(I_m) = M(1);\ \forall\ k \in \mathbb{Z}_+,$$

$$A_H^k(B_H(I_m)) = A_H^k \begin{pmatrix} M(1) \\ M(2) \\ M(3) \\ \vdots \end{pmatrix} = \begin{pmatrix} M(k+1) \\ M(k+2) \\ M(k+3) \\ \vdots \end{pmatrix},\ C_H A_H^k B_H(I_m) = M(k+1).$$

The assumption that $n_x = \text{rank}(\mathbf{H}) < \infty$ implies that $\dim(X_H) = \text{rank}(\mathbf{H}) = n_x < \infty$. From [21, Theorem of paragraph 9] follows that X_H is isomorphic to the finite-dimensional vector space $(\mathbb{R}, \mathbb{R}^{n_x})$, hence there exists an isomorphism $f : X_H \to \mathbb{R}^{n_x}$. Define,

$$A : \mathbb{R}^{n_x} \to \mathbb{R}^{n_x},\ A = f A_H f^{-1},\ B : \mathbb{R}^{n_u} \to \mathbb{R}^{n_x},\ B = f B_H,$$
$$C : \mathbb{R}^{n_x} \to \mathbb{R}^{n_y},\ C = C_H f^{-1},\ D = L \in \mathbb{R}^{n_y \times n_u}.$$

Then for all $k \in \mathbb{N}$, $CA^kB = M(k+1) = H(k+1)$. Thus the impulse response function of a time-invariant finite-dimensional linear system,

$$(T, \mathbb{R}^{n_x}, \mathbb{R}^{n_u}, \mathbb{R}^{n_y}, X_0, \mathbf{U}, (A, B, C, D)),$$

is a realization of the input-output linear system.
(b) Because a realization of state-space dimension $n_x \in \mathbb{Z}_+$ is assumed to exist, for $s, t \in \mathbb{Z}_+$,

$$H(s,t) = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{s-1} \end{pmatrix} \begin{pmatrix} B\ AB\ \ldots\ A^{t-1}B \end{pmatrix},\ \Rightarrow\ \text{rank}(H(s,t)) \leq n_x,$$
$$\Rightarrow \quad \text{rank}(H) = \sup_{s,t \in \mathbb{Z}_+} \text{rank}(H(s,t)) \leq n_x.$$

(b.1) ⇒ (b.2) Suppose that $n_1 := \text{rank}(\mathbf{H}) < n_x$. From the proof of (a) then follows that there exists a realization of state-space dimension $n_1 < n_x$. But this contradicts that the realization of state-space dimension n_x is minimal. Hence $\text{rank}(\mathbf{H}) = n_x$.
(b.2) ⇒ (b.1) Suppose that the realization is not minimal. Then there exists a realization of state-space dimension $n_1 < n_x$. But, as argued above, then $\text{rank}(\mathbf{H}) \leq n_1 < n_x$ which contradicts $n_x = \text{rank}(\mathbf{H})$.
(b.2) ⇒ (b.3) If $n_1 := \text{rank}(\text{conmat}(A,B)) < n_x$ then $\text{rank}(\mathbf{H}) \leq \text{rank}(\text{conmat}(A,B)) = n_1 < n_x$ which contradicts $n_x = \text{rank}(\mathbf{H})$. Thus $\text{rank}(\text{conmat}(A,B)) = n_x$. Analogously, $\text{rank}(\text{obsm}(A,C)) = n_x$.
(b.3) ⇒ (b.2) For $s, t \in \mathbb{Z}_+$ sufficiently large,

$$\text{rank}(H(s,t)) = \text{rank}(\text{obsm}(A,C)\text{conmat}(A,C)) = n_x,$$
$$\Rightarrow \text{rank}(\mathbf{H}) = \sup_{s,t \in \mathbb{Z}_+} \text{rank}(H(s,t)) \geq n_x.$$

But, because a realization exists, it follows from the above that $\text{rank}(\mathbf{H}) \leq n_x$. Hence $\text{rank}(\mathbf{H}) = n_x$.
(c) (⇒) Denote

$$H(1:n_x, 1:n_x) = \begin{pmatrix} M(1) & \ldots\ M(n_x) \\ \vdots & \vdots \\ M(n_x) & \ldots\ M(2n_x - 1) \end{pmatrix},$$
$$H(2:n_x+1, 1:n_x) = \begin{pmatrix} M(2) & \ldots\ M(n_x+1) \\ \vdots & \vdots \\ M(n_x+1) & \ldots\ M(2n_x) \end{pmatrix}.$$

Because both quadrupels of system matrices are realizations of the same impulse response function, the following equations hold true,

$$\begin{aligned} H(1:n_x, 1:n_x) &= \text{obsm}(A_1, C_1)\text{conmat}(A_1, B_1) \qquad (21.2)\\ &= \text{obsm}(A_2, C_2)\text{conmat}(A_2, B_2),\\ H(2:n_x+1, 1:n_x) &= \text{obsm}(A_1, C_1)A_1\text{conmat}(A_1, B_1) \qquad (21.3)\\ &= \text{obsm}(A_2, C_2)A_2\text{conmat}(A_2, B_2). \end{aligned}$$

Because both quadrupels are minimal realizations, it follows from (b) that,

$$\begin{aligned} n_x &= \text{rank}(\text{conmat}(A_1, C_1)) = \text{rank}(\text{obsm}(A_1, B_1))\\ &= \text{rank}(\text{conmat}(A_2, C_2)) = \text{rank}(\text{obsm}(A_2, B_2)). \end{aligned}$$

Hence there exists a left inverse $O_{1,L} \in \mathbb{R}^{n_x \times n_u n_y}$ and a right inverse $R_{2,R} \in \mathbb{R}^{n_x n_u \times n_u}$ such that,

$$I_{n_x} = O_{1,L}\text{obsm}(A_1, C_1), \;\; I_{n_x} = \text{conmat}(A_2, B_2)R_{2,R}. \qquad (21.4)$$

Define $L \in \mathbb{R}^{n_x \times n_x}$, $L = (\text{conmat}(A_1, B_1)R_{2,R})^{-1}$. From (b) and (21.4) follows that $\text{rank}(L) = n_x$. From (21.2) and (21.3) follows that,

$$\begin{aligned} L^{-1} &= \text{conmat}(A_1, B_1)R_{2,R} = O_{1,L}\text{obsm}(A_1, C_1)\text{conmat}(A_1, B_1)R_{2,R}\\ &= O_{1,L}H(1:n, 1:n)R_{2,R} = O_{1,L}\text{obsm}(A_2, C_2)\text{conmat}(A_2, B_2)R_{2,R}\\ &= O_{1,L}\text{obsm}(A_2, C_2),\\ A_1L^{-1} &= A_1\text{conmat}(A_1, B_1)R_{2,R} = O_{1,L}\text{obsm}(A_2, C_2)A_2 = L^{-1}A_2,\\ &\quad \text{as for } L^{-1}. \end{aligned}$$

From $L^{-1} = \text{conmat}(A_1, B_1)R_{2,R}$ follows that
$\text{conmat}(A_2, B_2) = L\text{conmat}(A_1, B_1)$ and hence $B_2 = LB_1$. Similarly,
$L^{-1} = O_{1,L}\text{obsm}(A_2, C_2)$ implies that
$\text{obsm}(A_2, C_2) = \text{obsm}(A_1, C_1)L^{-1}$ hence that $C_2 = C_1L^{-1}$. Thus,

$$A_2 = LA_1L^{-1}, \; B_2 = LB_1, \; C_2 = C_1L^{-1}, \; D_2 = D_1.$$

($\Leftarrow$) It follows directly from the relation between the quadruples that,

$$M(0) = D_1 = D_2, \;\; M(k) = C_1A_1^kB_1 = C_1L^{-1}(LA_1L^{-1})^kLB_1 = C_2A_2^kB_2, \; \forall k \in \mathbb{N}.$$

Hence (A_2, B_2, C_2, D_2) are also system matrices of a realization. Because by assumption (A_1, B_1, C_1, D_1) are system matrices of a minimal realization,
$A_1, A_2 \in \mathbb{R}^{n_x \times n_x}$, the system matrices (A_2, B_2, C_2, D_2) also describe a minimal realization. □

Reduction of a Non-Minimal Realization

If a realization of an impulse response system is not a minimal realization then it can be reduced to a minimal realization by the following procedure.

Procedure 21.8.10 Reduction of a non-minimal realization of a linear impulse-response function to a minimal realization.
Date: System parameters of a linear control system $(n_y, n_x, n_u, A, B, C, D)$.

1. *If the rank of the controllability matrix satisfies,*

$$\begin{aligned} n_x &> \text{rank}(\text{conmat}(A,B)),\\ \text{conmat}(A,B) &= \left(B\ AB\ A^2B\ \ldots\ A^{n-1}B \right), \end{aligned}$$

then execute the following steps:

a. *construct a basis* $\{v_1, v_2, \ldots, v_{n_{x_1}} \in \mathbb{R}^n\}$ *for the matrix* conmat(A,B) *and denote by* $n_{x_1} \in \mathbb{N}$ *the dimension of this basis;*
b. *complete the basis of the previous step to a basis of* $\mathbb{R}^{n_x}$, $\{v_1, v_2, \ldots, v_n \in \mathbb{R}^{n_x}\}$ *and denote by L the matrix which has the basis vectors as its columns; see Proposition 17.4.6 for a procedure; and*
c. *compute the matrices,*

$$L_1 A L_1^{-1} = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix},\ L_1 B = \begin{pmatrix} B_1 \\ 0 \end{pmatrix},\ C L_1^{-1} = \left(C_1\ C_2 \right);$$

d. *set the system parameters of the reduced system to* $(n_y, n_{x_1}, n_u, A_{11}, B_1, C_1, D)$*;*

else proceed.

2. *If the observability matrix is rank deficient,*

$$\begin{aligned} n_{x_1} &> \text{rank}(\text{obsmat}(A_{11}, C_1)),\\ \text{obsmat}(A_{11}, C_1) &= \begin{pmatrix} C_1 \\ C_1 A_{11} \\ C_1 A_{11}^2 \\ \vdots \\ C_1 A_{11}^{n_{x_1}} \end{pmatrix}. \end{aligned}$$

then execute the following steps:

a. *construct a basis in* $\mathbb{R}^{n_{x_1}}$ *for*

$$\ker(\text{obsmat}(A_{11}, C_1)) = \text{span}\{w_1, w_2, \ldots, w_{n_{x_2}}\};$$

b. *extend the previous basis to a basis for* $\mathbb{R}^{n_{x_1}}$

$$L_2 = \left(w_1\ w_2\ w_3\ \ldots\ w_{n_{x_1}} \right) \in \mathbb{R}_{nsng}^{n_{x_1} \times n_{x_1}};$$

c. *compute the matrices,*

$$L_2 A_{11} L_2^{-1} = \begin{pmatrix} A_{2,11} & 0 \\ A_{2,12} & A_{2,22} \end{pmatrix}, \; B_2 = L_2 B_1 = \begin{pmatrix} B_{2,1} \\ B_{2,2} \end{pmatrix},$$
$$C_2 = C_1 L_2^{-1} = \begin{pmatrix} C_{2,1} & 0 \end{pmatrix};$$

d. *Set the system matrices of the reduced system to,*

$$(n_y, n_{x_3}, n_u, A_3, B_3, C_3, D) = (n_y, n_{x_2}, n_u, A_{2,11}, B_{2,1}, C_{2,1}, D).$$

3. *Output the system matrices* $(n_y, n_{x_3}, n_u, A_3, B_3, C_3, D)$. *These matrices then represent a minimal realization of the original impulse response function.*

The resulting reduced-dimension system is such that (A_3, B_3) is a controllable pair and (A_3, C_3) is an observable pair. Hence the produced linear system is a minimal realization of its impulse response function. This is straigthforward to prove. The steps 1 and 2 can be interchanged without affecting the minimality of the realization.

21.9 Stability

The reader finds in this section concepts of stability of a nonlinear deterministic system and sufficient conditions for these concepts using the approach of Lyapunov. This theory is used elsewhere in the book.

Definition 21.9.1. Consider a nonlinear time-varying deterministic system with the representation,

$$\begin{aligned} & x(t+1) = f(t, x(t)), \; x(0) = x_0, \qquad (21.5) \\ & T = \mathbb{N} = \{0, 1, \ldots\}, \; X = \mathbb{R}^{n_x}, \; x_0 \in X, \; f : T \times X \to X, \\ & x(t; (0, x_0)), \; x : T \to X, \end{aligned}$$

denotes the trajectory generated by the system when it starts at time $t = 0$ and at state x_0. Below $\|.\|$ denotes any norm on the state space $X = \mathbb{R}^{n_x}$. All norms on the finite-dimensional space X are known to be equivalent.

(a) A *steady state* is a state $x_s \in X$ of the system such that, for all times $t \in T$, $x_s = f(t, x_s)$. There may exist two or more steady states. Below a steady state is not converted to the zero state so as to avoid misunderstandings.

(b) A steady state $x_s \in X$ is called *stable* if,

$$\begin{aligned} & \forall\, \varepsilon \in (0, \infty), \; \forall\, t_0 \in T, \; \exists\, \delta \in (0, \infty) \\ & (\delta(\varepsilon, t_0) \text{ which indicates that } \delta \text{ may depend on } \varepsilon \text{ and on } t_0) \\ & \|x_0 - x_s\| < \delta \;\Rightarrow\; \forall\, t \geq t_0, \; \|x(t; (t_0, x_0)) - x_s\| < \varepsilon. \end{aligned}$$

(b) A steady state $x_s \in X$ is called *uniformly stable* if,

$$\begin{aligned} & \forall\, \varepsilon \in (0, \infty), \; \exists\, \delta \in (0, \infty), \\ & t_0 \in T, \; \|x_0 - x_s\| < \delta \;\Rightarrow\; \forall\, t \geq t_0, \; \|x(t; (t_0, x_0)) - x_s\| < \varepsilon. \end{aligned}$$

(c) A steady state $x_s \in X$ is called *attractive* if,

$$\begin{aligned}&\forall\, t_0 \in T,\ \exists\, b \in \mathbb{R}_{s+},\\ &\|x_0 - x_s\| < \delta \ \Rightarrow\ \lim_{t\to\infty} \|x(t;(t_0,x_0)) - x_s\| = 0.\end{aligned}$$

(d) A steady state $x_s \in X$ is called *asymptotically stable* if it is both stable and attractive.

(e) A steady state $x_s \in X$ is called *exponentially stable* if,

$$\begin{aligned}&\exists\, a,\ b \in (0,\infty),\ \exists\, r \in (0,1),\quad x_0 \in X,\ \|x_0 - x_s\| < b,\\ \Rightarrow\ &\forall\, t \geq 0,\ \|x(t;(0,x_0)) - x_s\|\ \leq\ a\, r^t\ \|x_0 - x_s\|.\end{aligned}$$

(f) A steady state $x_s \in X$ is called *globally exponentially stable* if,

$$\begin{aligned}&\exists\, a \in (0,\infty),\ \exists\, r \in (0,1),\ \forall\, t \in T,\ \forall\, x_0 \in X,\\ &\|x(t;(0,x_0)) - x_s\|\ \leq\ a\, r^t\ \|x_0 - x_s\|.\end{aligned}$$

Definition 21.9.2. A function $f_K : \mathbb{R}_+ \to \mathbb{R}_+$ is said to be of *class K* if $f_K(0) = 0$, it is strictly increasing ($a < b$ implies $f_K(a) < f_K(b)$), and it is a continuous function.

Definition 21.9.3. Consider the nonlinear deterministic system of Def. 21.9.1. Assume that there exists a steady state denoted by $x_s \in X = \mathbb{R}^{n_x}$. Consider a function $V : T \times X \to X$.

(a) Call the function V a *locally positive-definite function* (lpdf) if,

$$\begin{aligned}&(1) && \forall\, t \in T,\ V(t,0) = 0;\\ &(2) && \exists\, r \in (0,1),\ \exists\, \alpha \in K,\\ & && \forall\, t \in T,\ \forall\, x \in B(x_s,r) = \{x \in X \mid \|x - x_s\| < r\},\\ & && \Rightarrow\ \alpha(\|x - x_s\|) \leq V(t, x - x_s).\end{aligned}$$

(b) The function V is called a *positive-definite function* (pdf) if (1) $\forall\, t \in T$, $V(t,0) = 0$; (2) $\exists\, \alpha \in K$ such that for all $t \in T$ and $x \in X$, $\alpha(\|x - x_s\|) \leq V(t, x - x_s)$;

(c) The function V is called a *decrescent function* if there exists a function $\beta \in K$ and there exists a real number $r \in (0,1)$ such that
for all $t \in T$ and $x \in B(x_s,r)$, $V(t, x - x_s) \leq \beta(\|x - x_s\|)$;

(d) The function V is called *radially unbounded* if uniformly in $t \in T$,

$$\lim_{\|x - x_s\| \to \infty} V(t, x - x_s) = +\infty.$$

Theorem 21.9.4. *Consider a nonlinear deterministic system and assume that there exists a steady state,*

$$\begin{aligned}x(t+1) &= f(t,x(t)),\ x(0) = x_0 \in X,\\ x_s &= f(t,x_s),\ \forall\, t \in T,\ x_s \in X = \mathbb{R}_{n_x}.\end{aligned}$$

(a) The steady state is stable *if,*

$$\begin{aligned} &\exists\, V : T \times X \to \mathbb{R}, \exists\, r \in (0,1) \textit{ such that,} \\ (1) \quad & V \textit{ is a lpdf,} \\ (2) \quad & \forall t \in T,\ \forall\, x \in B(x_s, r), \\ \Delta V(t, x - x_s) &= V(t+1, x(t+1;(0,x)) - x_s) - V(t, x - x_s) \leq 0. \end{aligned}$$

(b)The steady state is uniformly asymptotically stable *if there exists function $V : T \times X \to \mathbb{R}$ such that V is a lpdf, descrescent, and $-\Delta V$ is a lpdf.*

(c)The steady state is globally asymptotically stable *if there exists function $V : T \times X \to \mathbb{R}$ such that (1) V is a positive-definite function, descrescent, and radially unbouded; and (2) $-\Delta V$ is a positive-definite function.*

(d)The steady state is exponentially stable *if there exists function $V : T \times X \to \mathbb{R}$ and if there exist real numbers $a,\ b,\ c,\ r \in (0,\infty)$ and an integer $p \in \mathbb{Z}_+$ such that*

$$\begin{aligned} &\forall\, t \in T,\ \forall\, x \in B(0,r), \\ &a\|x - x_s\|^p \leq V(t, x - x_s) \leq b\|x - x_s\|^p,\ \ \Delta V(t, x - x_s) \leq -c\|x - x_s\|^p. \end{aligned}$$

For the record the following result is stated and proven.

Proposition 21.9.5. *Consider a time-invariant linear system without input,*

$$\begin{aligned} &x(t+1) = Ax(t),\ \ x(0) = x_0, \\ &\qquad T = \mathbb{N} = \{0,1,\ldots\},\ X = \mathbb{R}^{n_x},\ n_x \in \mathbb{Z}_+,\ \text{spec}(A) \subseteq \mathrm{D}_o. \end{aligned}$$

Then $\lim_{t\to\infty} x(t;(0,x_0)) = 0$. *Note that $x_s = 0 \in X$ is the unique steady state of this system.*

Proof. The theorem above will be applied for global asymptotic stability. Consider a matrix $B \in \mathbb{R}^{n_x \times n_x}$ such that (A,B) is a controllable pair. For example, $B = I_{n_x}$ satisfies this condition. It is supposed that $A \neq 0$ because if $A = 0$ the conclusion is trivial. Consider the matrix Lyapunov equation for a matrix Q,

$$Q = A^T QA + B^T B,\ Q \in \mathbb{R}^{n_x \times n_x}.$$

It follows from Theorem 22.1.2 that this equation with the assumption $\text{spec}(A) \subset \mathrm{D}_o$ has a unique solution which satisfies $Q \in \mathbb{R}^{n_x \times n_x}_{spds}$ hence $0 \prec Q$.

Define the candidate Lyapunov function $V : X \to \mathbb{R}_+$, $V(x) = x^T Qx$. It will be proven that this function satisfies the conditions of Theorem 21.9.4.(c) Define the function $\alpha(\|x\|) = \lambda_{min}(Q)\|x\|$. Here $\lambda_{min}(Q)$ denotes the eigenvalue of the matrix Q with minimal real part. Then $0 \prec Q$ implies that $0 < \lambda_{min}(Q)$. Thus the function α satisfies $\alpha(\|x\|) = 0$, is strictly increasing in $\|x\|$, and continuous and is thus of class K, and $\alpha(\|x\|) \leq V(x)$. Because in addition $V(0) = 0$, V is a positive-definite function. The function $\beta(\|x\|) = \lambda_{max}(Q)\|x\|$, with $0 < \lambda_{max}(Q)$, satisfies $\beta(\|0\|) = 0$, is strictly increasing, continuous, and is thus of class K while $V(x) \leq \beta(\|x\|)$ for all $x \in \mathbb{R}^{n_x}$. Thus V is decrescent. Moreover, V is radially unbounded, $\lim_{\|x\|\to\infty} V(x) = +\infty$ which also follows from the inequality $\alpha(\|x\|) \leq V(x)$ and from the properties of α. Finally note that,

$$\begin{aligned}
-\Delta V(x) &= -V(x(t+1;(0,x_0)) + V(x(t;(0,x_0))) \\
&= -x(t+1)^T Qx(t+1) + x(t)^T Qx(t) = -x(t)^T [A^T QA - Q]x(t) \\
&= x(t)^T B^T Bx(t) > 0 \text{ if } x(t) \neq 0 \text{ because } 0 \prec B^T B.
\end{aligned}$$

Then $-\Delta V(0) = 0$, $\alpha_2(\|x\|) = \lambda_{min}(B^T B)\|x\|$, $0 < \lambda_{min}(B^T B)$, and $\alpha_2(\|x\|) \leq -\Delta V(x)$ show that $-\Delta V(x)$ is a positive-definite function.

From Theorem 21.9.4.(c) then follows that $\lim_{t\to\infty} x(t) = 0$. □

21.10 Further Reading

History. The mathematical formulation of a linear control system with inputs and outputs combined with the solution of the realization problem for this subset of systems, is due to R.E. Kalman. The relevant journal paper was published in 1963, [34]. An earlier conference paper of R.E. Kalman was presented in 1960 at the first IFAC World Congress held in Moscow, Soviet Union, in 1960, [32]. Kalman also formulated the canonical structure of linear systems, [33]. A paper on controllability is [37]. The book by R.E. Kalman, P. Falb, and Arbib contains a chapter on the formulation of algebraic system theory, [36].

Realization theory of linear systems has a root in computer science, in particular for recursive functions of the natural numbers, for which the papers of S.C. Kleene, [39, 40] are often quoted. A. Nerode, a researcher active in the USA, has formulated this in different terms and developed it further, [52]. Early papers are also by M.O. Rabin and D. Scott, [54] and by M.P. Schutzenberger, [57]. J.A. Brzozowski has formulated the concept of input derivative, also known as the Brzozowski derivative, which leads to an efficient procedure for minimal representation of regular expressions, [6, 7]. The relation of automata and control system has been described by M.A. Arbib, [1, 2]. The algebraic theory of machines was formulated by K. Krohn and J. Rhodes, [43, 41, 42, 44].

There is another root of system theory in circuit theory, see for example the two papers by B. McMillan, [48, 49]. Realization in those papers refers to the realization of a complex-valued matrix as the open-circuit impedance matrix of a finite passive electric circuit or electric network. This form of realization does not seem inspired by the realization theory of computer science though there is a direct analogy.

One of the contributions of Kalman to control and system theory is to have formulated realization theory for a time-invariant linear system defined on a real vector space and to solve the corresponding realization problem. Useful concepts and results of realization theory of a time-invariant linear control system are the concepts of controllability, of observability, of minimality of a realization, and of the characterization of algebraic equivalence of minimal realizations.

Subsequently system theory was developed for other classes of control systems. The approach for discrete-time polynomial systems is due to E.D. Sontag, [64].

The theory of this chapter is primarily due to R.E. Kalman, see [32, 33, 34]. The author has benefitted considerably for the preparation of this chapter from the publications of M.L.J. Hautus and of J.C. Willems, also from their unpublished lecture notes.

Books with system theory. A rather general book on control systems is that by E.D. Sontag, [65] where one finds realization theory in the Chapters 3 and 6, in particular in Section 6.5.

Books with system theory of linear systems at the level of this book include [8, 9, 29]. Most books on control theory also contain a part on system theory. Books on system theory of linear systems at the undergraduate or master level include [12, 68]. An early paper on realization of linear systems is [60].

Books on automata with a system theoretic flavor include [38, 17, 19, 67].

Books on system theory of algebraic systems include [53, 56]. System theory in the formalism of category theory has been formulated by M.A. Arbib and E. Manes, [3]. Category theory allows comparisons of system theoretic concepts and results of various domains of systems.

Books on control and system theory of nonlinear systems in a differential-geometric setting include [27, 28]. Papers on realization theory of subsets of non-linear systems are not listed here due to space limitations.

Realization theory for time-invariant linear systems was initially formulated by R.E. Kalman, [31, 33, 34, 35, 71]. Realization algorithms or produres are treated in [25, 26, 66]. An abstract algebraic viewpoint was developed by P. Zeiger, [78] and by A.J. Tether, [66]. Other papers on realization include [20, 76, 14, 73]

Conditions for structural identifiability of the parametrization of a linear control system are directly based on realization theory for the same set of systems. This research was initiated by C. Cobelli and co-authors, [4, 10, 15]. See also [69] and [45]. The literature on identifiability is very large but system identification is not the topic of this book.

Controllability Theorem 21.2.7 is based on [68, Section 3.3]. The definitions of stabilizability and detectability are due to M.L.J. Hautus, [22, 23]. Theorem 21.2.15 is related to [8, Th. 8.2.12].

Zero-output dynamics. The research topic of zero-output dynamics of linear systems has a long history. For continuous-time linear systems the reader is referred to [68, Ch. 7], to the book of C.A. Desoer and F.M. Callier [8], and to the book of T. Kailath, [30].

Papers on zeroes of linear systems include, [11, 13, 46]. The papers of B.P. Molinari deal with discrete-time linear systems, see for example [50, 51].

The eigenstructure of a multivariable linear system has been described by P.M. Van Dooren, [16] which is recommended by the author. The computation of system zeroes is described in [18].

Invertibility of a linear system. The inverse of a deterministic linear system defined on finite input, state, and output sets was treated in a paper by J.L. Massey and M.K. Sain, [47]. The inverse of a linear control system is treated by L.M. Silverman in [58] and a procedure to invert a discrete-time linear control system is stated by Silverman in [61]. Necessary and sufficient conditions for invertibility of a continuous-time linear system have been formulated by M.L.J. Hautus and L.M. Silverman in [24]. The conditions involve the concepts of strong controllability and its variants, and are mostly formulated in terms of subspaces of the state space according to the geometric viewpoint.

Realization theory. Realization of linear systems was formulated and proven by R.E. Kalman, [34]. The statement of this chapter is an extension and its proof is more detailed.

Realization theory of time-varying linear systems is treated in the tutorial paper [60] and in [20, 33, 34, 71, 59, 62, 63, 76].

The canonical factorization of a map is known in the algebra of sets. The reader may want to read [74, Sec.1.4] for a general framework.

Stability theory. The results are adjusted from the book of M. Vidyasagar, [70]. There are few references on stability of discrete-time deterministic systems.

References

1. M.A. Arbib. Automata theory and control theory – A rapprochement. In L.A. Zadeh and C.A. Desoer, editors, *Linear system theory: The state space approach*, pages 0–0. McGraw-Hill, New York, 1963. 807
2. M.A. Arbib. A common framework for automata theory and control theory. *SIAM J. Control*, 3:206–222, 1965. 761, 807
3. M.A. Arbib and E. Manes. Machines in a catagory: An expository introduction. *SIAM Review*, 57:163–192, 1974. 808
4. Stefania Audoly, Giuseppina Bellu, Leontina D'Agniò, Maria Pia Saccomani, and Claudio Cobelli. Global identifiability of nonlinear models of biological systems. *IEEE Trans. Biomedical Engineering*, 48:55–65, 2001. 808
5. R.W. Brockett and A.S. Willsky. Finite group homomorphic sequential systems. *IEEE Trans. Automatic Control*, 17:483–490, 1972. 353, 786
6. J.A. Brzozowski. A survey of regular expressions and their applications. *IEEE Trans. Electronic Computers*, 11:324–335, 1962. 174, 807
7. J.A. Brzozowski. Derivatives of regular expressions. *J. ACM*, 11:481–494, 1964. 174, 807
8. F.M. Callier and C.A. Desoer. *Linear system theory*. Springer-Verlag, New York, 1991. 217, 781, 808
9. Chi-Tsong Chen. *Linear system theory and design*. Holt, Rinehart and Winston, New York, 1970. 808
10. C. Cobelli and G. Romanin-Jacur. On the structural identifiability of biological compartmental systems in a general input-output configuration. *Math. Biosci.*, 30:139–151, 1976. 808
11. E.J. Davison and S.H. Wang. Properties and calculation of transmission zeros of linear multivariable systems. *Automatica*, 10:643–658, 1974. 808
12. C.A. Desoer. *Notes for a second course on linear systems*. Van Nostrand Reinhold Co., New York, 1970. 808
13. C.A. Desoer and J. Schulman. Zeros and poles of matrix functions and their dynamical interpretation. *IEEE Trans. Circuits & Systems*, 21:3 – 8, 1974. 808
14. C.A. Desoer and P.P. Varaiya. The minimal realization of a nonanticipative impulse response matrix. *SIAM J. Appl. Math.*, 15:754–764, 1967. 808
15. J.J. DiStefano III and C. Cobelli. On parameter and structural identifiability: nonunique observability/reconstructability for identifiable systems, other ambiguties, and new definitions. *IEEE Trans. Automatic Control*, 25:830–833, 1980. 808
16. Paul M. Van Dooren. The generalized eigenstructure problem in linear system theory. *IEEE Trans. Automatic Control*, 26:111–129, 1981. 193, 194, 808

17. S. Eilenberg. *Automata, languages, and machines (Volumes A and B)*. Academic Press, New York, 1974, 1976. 277, 808
18. A. Emami-Naeini and P. van Dooren. Computation of zeroes of linear multivariable systems. *Automatica*, 18:415–430, 1982. 808
19. F.Géseg and I. Peák. *Algebraic theory of automata*. Akadémiai Kiadó, Budapest, 1972. 808
20. E.G. Gilbert. Controllability and observability in multivariable control systems. *SIAM J. Control Ser. A*, 1:128–151, 1963. 808, 809
21. P.R. Halmos. *Finite-dimensional vector spaces*. Springer, New York, 1993. 635, 801
22. M.L.J. Hautus. Controllability and observability conditions of linear autonomous systems. *Indag. Math.*, 31:443–448, 1969. 808
23. M.L.J. Hautus. Stabilization, controllability and observability of linear autonomous systems. *Indag. Math.*, 32:448–455, 1970. 808
24. M.L.J. Hautus and L.M. Silverman. System structure and singular control. *Linear Algebra & its Applications*, 50:369–402, 1983. 786, 808
25. B.L. Ho and R.E. Kalman. *Effective construction of linear state-variable models from input/output data*, pages 449–459. University of Illinois, 1966. 808
26. B.L. Ho and R.E. Kalman. Effective construction of linear state-variable models from input/output functions. *Regelungstechnik*, 14:545–548, 1966. 808
27. A. Isidori. *Nonlinear control systems*. Springer, Berlin, 3rd ed. edition, 1995. 779, 808
28. A. Isidori. *Nonlinear control systems II*. Communications and Control Engineering Series. Springer-Verlag, London, 1999. 808
29. T. Kailath. *Linear systems*. Prentice-Hall Inc., Englewood Cliffs, 1980. 808
30. Thomas Kailath. Some alternatives in recursive estimation. *Int. J. Control*, 32:311–328, 1980. 781, 808
31. R.E. Kalman. A new approach to linear filtering and prediction problems. *J. Basic Eng.*, 82:35–45, 1960. 283, 310, 808
32. R.E. Kalman. On the general theory of control systems. In *Proc. 1st IFAC World Congress*, London, 1960. Butterworths. 807
33. R.E. Kalman. Canonical structure of linear dynamical systems. *Proc. Nat. Acad. Sci.*, 48:596–600, 1962. 807, 808, 809
34. R.E. Kalman. Mathematical description of linear dynamical systems. *SIAM J. Control*, 1:152–192, 1963. 174, 761, 807, 808, 809
35. R.E. Kalman. Irreducible representations and the degree of a rational matrix. *SIAM J. Control*, 3:520–544, 1965. 808
36. R.E. Kalman, P.L. Falb, and M.A. Arbib. *Topics in mathematical systems theory*. McGraw-Hill Book Co., New York, 1969. 78, 120, 807
37. R.E. Kalman, Y.C. Ho, and K.S. Narendra. Controllability of linear dynamic systems. *Contributions to differential equations*, 1:189–213, 1962. 807
38. Bakhadyr Khoussainov and Anil Nerode. *Automata theory and its applications*. Progress in Computer Science and Applied Logic. Birkhäuser, Boston, 2001. 808
39. S.C. Kleene. General recursive functions of natural numbers. *Mathematische Annalen*, 112:727–742, 1936. 174, 807
40. S.C. Kleene. Representation of events in nerve nets and finite automata. In *Automata Studies*, pages 3–42. Princetion University Press, Princeton, 1956. 807
41. K. Krohn and J. Rhodes. Algebraic theory of machines. I. Prime decomposition theorem for finite semigroups and machines. *Trans. Amer. Math. Soc.*, 116:450–464, 1965. 807
42. K. Krohn and J. Rhodes. Results on finite semigroups derived from the algebraic theory of machines. *Proc. Nat. Acad. Sc.*, 53:499–501, 1965. 807
43. K.B. Krohn and J.L. Rhodes. Algebraic theory of machines. In J. Fox, editor, *Proceedings of the Symposium on the Mathematical Theory of Automata*, pages 341–384. Polytechnic Institute of Brooklyn, New York, 1962. 807
44. K. Krohnn, J.L. Rhodes, and B.R. Tilson. The prime decomposition theorem of the algebraic theory of machines. In M.A. Arbib, editor, *Algebraic theory of machines, languages, and semigroups*, pages 81–125. Academic Press, New York, 1968. 807

45. R. Liu and L.-C. Suen. Minimal dimension realization and identifiability of input-output sequences. *IEEE Trans. Automatic Control*, 22:227–232, 1977. 796, 808

46. A. MacFarlane and N. Karcanias. Poles and zeros of linear multivariable systems: A survey of the algebraic, geometric, and complex-variable theory. *Int. J. Control*, 24:33 – 74, 1976. 808

47. J.L. Massey and M.K. Sain. Inverses of linear sequential circuits. *IEEE Tr. Computers*, 17:330–337, 1968. 786, 808

48. B. McMillan. Introduction to formal realizability theory I. *Bell System Techn. J.*, 31:217–279, 1952. 174, 807, 865

49. B. McMillan. Introduction to formal realizability theory II. *Bell System Techn. J.*, 31:541–600, 1952. 174, 807, 865

50. B.P. Molinari. Extended controllability and observability for linear systems. *IEEE Trans. Automatic Control*, 21:136–137, 1976. 781, 808

51. B.P. Molinari. On strong controllability and observability in linear multivariable control. *IEEE Trans. Automatic Control*, 21:761–764, 1976. 781, 784, 808

52. A. Nerode. Linear automaton transformations. *Proc. Amer. Math. Soc.*, 9:541–544, 1958. 174, 807

53. L. Padulo and M.A. Arbib. *System theory*. Hemisphere Publishing Corporation, Washington D.C., 1974. 808

54. M.O. Rabin and D. Scott. Finite automata and their decision problems. *IBM J.*, x:114–125, 1959. 174, 807

55. H.H. Rosenbrock. *State space and multivariable theory*. Wiley, New York, 1970. 174, 780

56. M.K. Sain. *Introduction to algebraic system theory*. Academic Press, New York, 1981. 786, 808

57. M.P. Schutzenberger. On the definition of a family of automata. *Information and Control*, 4:245–270, 1961. 807

58. L. Silverman. Inversion of multivariable linear systems. *IEEE Trans. Automatic Control*, 14:270–276, 1969. 786, 808

59. L.M. Silverman. Transformation of time-variable systems to canonical (phase-variable) form. *IEEE Trans. Automatic Control*, 11:300 – 303, 1966. 809

60. L.M. Silverman. Realization of linear dynamical systems. *IEEE Trans. Automatic Control*, 16:554–568, 1971. 808, 809

61. L.M. Silverman. Discrete Riccati equations: Alternative algorithms, asymptotic properties, and system theory interpretations. In C.T. Leondes, editor, *Control and Dynamic Systems*, pages 313–386. Academic Publishers, New York, 1976. 439, 786, 808

62. L.M. Silverman and H.E. Meadows. Controllability and observability in time-variable linear systems. *SIAM J. Control*, 5:64–73, 1967. 809

63. L.M. Silverman and H.E. Meadows. Equivalent realizations of linear systems. *SIAM J. Appl. Math.*, 17:393–408, 1969. 809

64. E.D. Sontag. *Polynomial response maps*, volume 13 of *Lecture Notes in Control and Information Sciences*. Springer-Verlag, Berlin, 1979. 807

65. E.D. Sontag. *Mathematical control theory: Deterministic finite dimensional systems (2nd. Ed.)*. Number 6 in Graduate Text in Applied Mathematics. Springer, New York, 1998. 217, 277, 808

66. A.J. Tether. Construction of minimal linear state-variable models from finite input-output data. *IEEE Trans. Automatic Control*, 15:427–436, 1970. 808

67. B.A. Trakhtnebrot and Ya.M. Barzdin. *Finite automata - Behavior and synthesis*. North-Holland Publishing Company, Amsterdam, 1973. 808

68. H.L. Trentelman, A.A. Stoorvogel, and M. Hautus. *Control theory for linear systems*. Springer, United Kingdom, 2001. 485, 781, 784, 785, 786, 808

69. J.M. van den Hof. Structural identifiability of linear compartmental systems. *IEEE Trans. Automatic Control*, 43:800–818, 1998. 796, 808

70. M. Vidyasagar. *Nonlinear systems analysis (2nd Ed.)*. Prentice Hall, Englewood Cliffs, 1993. 809

71. L. Weiss and R.E. Kalman. Contributions to linear system theory. *Int. J. Eng. Sci.*, 3:141–171, 1965. 808, 809
72. A.S. Willsky. Invertibility of finite-group homomorphic sequential systems. *Information and Control*, 27:126–147, 1975. 786
73. W.A. Wolovich and P.A. Falb. On the structure of multivariable systems. *SIAM J. Control*, 7:437–451, 1969. 808
74. W. Murray Wonham and Kai Cai. *Supervisory control of discrete-event systems*. Communications and Control Engineering. Springer, Cham, Switzerland, 2019. 809
75. W.M. Wonham. *Linear multivariable control: A geometric approach*. Springer-Verlag, Berlin, 1979. 310, 779
76. D.C. Youla. The synthesis of linear dynamical systems from prescribed weighting patterns. *SIAM J. Appl. Math.*, 14:527–549, 1966. 808, 809
77. L.A. Zadeh and C.A. Desoer. *Linear system theory: The state space approach*. McGrawHill, New York, 1963. 761
78. P. Zeiger. Ho's algorithm, commutative diagrams, and the uniqueness of minimal linear systems. *Inform. and Control*, 11:71–79, 1967. 808

Chapter 22
Appendix F Matrix Equations

Abstract The Lyapunov equation and the algebraic Riccati equation are treated in depth. The Lyapunov equation arises as the equation for the asymptotic covariance matrix of the state of a stationary Gaussian system. The algebraic Riccati equation arises in the Kalman filter, in stochastic control, and in stochastic realization of a Gaussian system. Results for both equations are provided on: the existence of a solution, uniqueness with respect to conditions, a description of the set of all solutions if applicable, particular properties of solutions, and on the computation of solutions.

Key words: Lyapunov equation. Algebraic Riccati equation.

22.1 Lyapunov Equation

The invariant distribution of a time-invariant Gaussian system is Gaussian with zero mean value and a state-variance matrix. The state variance matrix is a solution of a Lyapunov equation. In this section the existence, the uniqueness, and various properties of the solution of this Lyapunov equation are investigated.

Problem 22.1.1. *Convergence of the state variance of a Gaussian system.* Consider the time-invariant Gaussian system representation,

$$\begin{aligned} &\{n_y, n_x, n_v, A, C, M, N\} \in \text{GStocSP}, \\ x(t+1) &= Ax(t) + Mv(t),\ x_0, \\ y(t) &= Cx(t) + Nv(t), \end{aligned}$$

with $x_0 \in G(0, Q_0)$ and $v(t) \in G(0, I)$. Consider the variance function $Q : T \to \mathbb{R}^{n_x \times n_x}$ of the state process which by Theorem 4.3.5 satisfies the *Lyapunov recursion*,

$$Q_x(t+1) = AQ_x(t)A^T + MM^T,\ Q(0) = Q_0. \tag{22.1}$$

Consider the *Lyapunov equation* for the matrix $Q \in \mathbb{R}^{n_x \times n_x}$, where the matrices A, M are specified,

$$Q = AQA^T + MM^T. \tag{22.2}$$

The problem is to find conditions on the system matrices such that the following limit exists $\lim_{t\to\infty} Q(t) = Q(\infty)$, to show that the limit is a solution of the above formulated Lyapunov equation, to show that the Lyapunov equation has a unique solution, and to provide a necessary and sufficient condition for the limit matrix to be strictly positive definite.

Theorem 22.1.2. Lyapunov equation – Existence, uniqueness, and convergence. *Consider Problem 22.1.1.*

(a)If A is an exponentially stable matrix ($\Leftrightarrow$ $\text{spec}(A) \subseteq \text{D}_o$) then $\lim_{t\to\infty} Q_x(t) = Q$ exists and Q is a solution of the Lyapunov equation (22.2).
(b)If A is an exponentially stable matrix then the Lyapunov equation has a unique solution $Q \in \mathbb{R}^{n_x \times n_x}$ that, moreover, is symmetric and positive definite, hence $Q \in \mathbb{R}^{n_x \times n_x}_{pds}$.
(c)Assume that (A,M) is a stabilizable pair and that there exists a $Q \in \mathbb{R}^{n_x \times n_x}_{pds}$ such that $Q = AQA^T + MM^T$. Then A is an exponentially stable matrix.
(d)Let $Q \in \mathbb{R}^{n_x \times n_x}_{pds}$ be a solution of the Lyapunov equation. Any two of the following three statements implies the third statement:

1. *A is an exponentially stable matrix ($\Leftrightarrow$ $\text{spec}(A) \subset \text{D}_o$);*
2. *(A,M) is a controllable pair;*
3. *$Q \succ 0$.*

The formulation of this part is due to M.L.J. Hautus (private communication).
(e)If the system matrix A is exponentially stable then the convergence is exponential, meaning,

$$\exists\, c \in \mathbb{R}_+,\ \exists\, r \in (0,1),\ \textit{such that},\ \|Q(t) - Q(\infty)\|_2 \leq c|r|^t;$$
$$\Rightarrow \lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} Q(s) = Q(\infty) = Q,$$

where Q is the unique solution of the Lyapunov equation.
(f) If the system matrix A is exponentially stable then there exist solutions of the following two Lyapunov equations and the subsequent relation holds,

$$Q_o^* = AQ_o^*A^T + MM^T,\ \ Q_c^* = A^TQ_c^*A + C^TC,$$
$$\Rightarrow \text{tr}(CQ_o^*C^T) = \text{tr}(M^TQ_c^*M).$$

(g)Consider the more general Lyapunov equation,

$$Q = AQA^T + W \tag{22.3}$$

where $A,\ W,\ Q \in \mathbb{R}^{n_x \times n_x}$. Assume that $\text{spec}(A) \subset \text{D}_o$ and $W = W^T$. Then there exists a unique solution $Q \in \mathbb{R}^{n_x \times n_x}$ of the equation (22.3) satisfying $Q = Q^T$. Moreover, $0 \preceq Q$ if and only if $0 \preceq W$ and $Q \preceq 0$ if and only if $W \preceq 0$. Consequently,

$$\begin{aligned} &W_1,\ W_2 \in \mathbb{R}^{n_x \times n_x},\ W_1 = W_1^T,\ W_2 = W_2^T, \\ Q_1 &= AQ_1A^T + W_1,\ \ Q_2 = AQ_2A^T + W_2,\ W_1 \preceq W_2\ \Rightarrow\ Q_1 \preceq Q_2. \end{aligned}$$

(h)*Assume that A is exponentially stable. The following implication holds and this implication is useful if for the variance matrix $MM^T = W$ a lower and an upper-bound are known,*

$$\begin{aligned} &W_{min},\ W,\ W_{max} \in \mathbb{R}^{n_x \times n_x},\ W_{min} = W_{min}^T,\ W = W^T,\ W_{max} = W_{max}^T, \\ Q_{min} &= AQ_{min}A^T + W_{min},\ \ Q = AQA^T + W,\ Q_{max} = AQ_{max}A^T + W_{max}, \\ &W_{min} \preceq W \preceq W_{max}\ \Rightarrow\ Q_{min} \preceq Q \preceq Q_{max}. \end{aligned}$$

Proof. (a) It is a verification by induction that

$$Q_x(t) = A^tQ_0(A^T)^t + \sum_{s=0}^{t-1} A^sMM^T(A^T)^s.$$

Let A have the Jordan decomposition $A = L\Lambda L^{-1}$ with Λ consisting of Jordan blocks. Then,

$$Q_x(t) = L\Lambda^t(L^{-1}Q_0L)(\Lambda^T)^tL^{-1} + L\sum_{s=0}^{t-1} \Lambda^s[L^{-1}MM^T](\Lambda^T)^sL^{-1}.$$

Consider the matrix $\Lambda^t[L^{-1}Q_0L](\Lambda^T)^t$. Writing out the elements of this matrix one obtains expressions of the form,

$$\begin{aligned} &\alpha_{ij}\lambda_i^t\lambda_j^t, \text{ or,} \\ &\alpha_{ij}\lambda_i^{s-l}\lambda_j^{s-m}s(s-1)\ldots(s-l)s(s-1)\ldots(s-m),\ l,\ m \in \mathbb{Z}_{n_x} = \{1,2,\ldots,n_x\}. \end{aligned}$$

Because $\text{spec}(A) \subset \text{D}_o$ there exists an $r \in (0,1)$ such that for all $\lambda \in \text{spec}(A)$, $|\lambda| < r < 1$. Hence $\lim_{t\to\infty}\Lambda^t(L^{-1}Q_0L)(\Lambda^T)^t = 0$. Similarly one obtains expressions for,

$$\begin{aligned} &\sum_{s=0}^{t-1}\Lambda^s[S^{-1}MM^TL](\Lambda^T)^s, \text{ of the form, } \sum_{s=0}^{t-1}\beta_{ij}\lambda_i^s\lambda_j^s, \text{ or,} \\ &\sum_{s=0}^{t-1}\beta_{ij}\lambda_i^{s-l}\lambda_j^{s-m}s(s-1)\ldots(s-l_1)s(s-1)\ldots(s-m_1). \end{aligned}$$

The above expressions are illustrated for the case of dimension four,

$$A^2 = \begin{pmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{pmatrix}^2 = \begin{pmatrix} \lambda^2 & 2\lambda & 1 & 0 \\ 0 & \lambda^2 & 2\lambda & 1 \\ 0 & 0 & \lambda^2 & 2\lambda \\ 0 & 0 & 0 & \lambda^2 \end{pmatrix},\ A^3 = \begin{pmatrix} \lambda^3 & 3\lambda^2 & 3\lambda & 1 \\ 0 & \lambda^3 & 3\lambda^2 & 3\lambda \\ 0 & 0 & \lambda^3 & 3\lambda^2 \\ 0 & 0 & 0 & \lambda^3 \end{pmatrix}.$$

It is then clear that the limit exists by,

$$|\lambda_i| < r,\ |\lambda_j| < r \in (0,1),\ \ \lim_{s\to\infty}|\lambda_i\lambda_j| < \lim\ (r^2)^s = \frac{1}{1-r^2};$$

where the convergence follows from [6, Ex. 2.35]. Thus,

$$\lim_{t\to\infty} \Lambda^s[L^{-1}MM^TL^{-T}](\Lambda^T)^s, \text{ exists and hence,}$$

$$\lim_{t\to\infty} Q_x(t) = \sum_{s=0}^{\infty} A^sMQ_vM^T(A^T)^s = Q,$$

exists. Finally,

$$Q = \lim_{t\to\infty} Q_x(t+1) = \lim_{t\to\infty} [AQ_x(t)A^T + MM^T] = AQA^T + MM^T.$$

(b) Let $Q_1, Q_2 \in \mathbb{R}^{n_x\times n_x}_{pds}$ be two solutions of (22.2). Let,

$$\Delta Q = Q_1 - Q_2; \text{ then,}$$
$$\Delta Q = [AQ_1A^T + MM^T] - [AQ_2A^T + MM^T] = A\Delta QA^T, \text{ and by iteration,}$$
$$\Delta Q = A^t\Delta Q(A^T)^t, \ \forall\, t \in T.$$

Let again $A = L\Lambda L^{-1}$ be the Jordan decomposition of A. Then,

$$\Lambda^t[L^{-1}\Delta QL](\Lambda^T)^t = L^{-1}\Delta QL. \tag{22.4}$$

The same arguments used in the proof of (a) above then yield,

$$0 = \lim_{t\to\infty} \Lambda^t[L^{-1}\Delta QL](\Lambda^T)^t, \text{ because the limit exists,}$$
$$= L^{-1}\Delta QL, \text{ because of (22.4),}$$

from which follows that $\Delta Q = 0$. This establishes that the Lyapunov equation has a unique solution. By transposing equation (22.2) one obtains, $Q^T = AQ^TA^T + MM^T$. Using the uniqueness of the solution of this equation one obtains that $Q = Q^T$. Let $u \in \mathbb{R}^{n_x}$. From the expression obtained in the proof of (a),

$$Q_x(t) = A^tQ_0(A^T)^t + \sum_{s=0}^{t-1} A^sMM^T(A^T)^s,$$

it follows that $u^TQ_x(t)u \geq 0$. Then $u^TQ_xu = \lim_{t\to\infty} u^TQ_x(t)u \geq 0$, or $Q \succeq 0$.
(c) Suppose there exists a $\lambda \in \text{spec}(A)$, $|\lambda| \geq 1$, and a $u \in \mathbb{C}^{n_x}$, with $u^TA = u^T\lambda$ and $u \neq 0$. Then,

$$u^TQu = u^TAQA^Tu + u^TMM^Tu, \ 0 \geq [1-|\lambda|^2]u^TQu = u^TMM^Tu \geq 0.$$

Hence both sides are zero and thus $u^TM = 0$. Because (A,M) is a stabilizable pair, $\lambda \in \text{spec}(A)$, $|\lambda| \geq 1$, $u^TA = u^T\lambda$, $u^TM = 0$, Theorem 21.2.11 imply that $u = 0$. This is a contradiction of the assumption that $u \neq 0$. Thus for all $\lambda \in \text{spec}(A)$, $|\lambda| < 1$, or $\text{spec}(A) \subset \mathrm{D}_o$.
(d) (2) & (3) ⇒ (1). The condition (A,M) a controllable pair implies that (A,M) is a stabilizable pair. The result then follows from (c).
(1) & (2) ⇒ (3). The condition (A,G) a controllable pair is defined to be such that $\text{rank}(M, AM, \ldots, A^{n_x-1}M) = n_x$. This in turn is equivalent with, using a result on the rank of a product of matrices,

$$\sum_{s=0}^{n-1} A^sMM^T(A^T)^s \succ 0.$$

Using the expresssion obtained in the proof of (a) one obtains

$$Q = \sum_{s=0}^{\infty} A^s MM^T (A^T)^s \geq \sum_{s=0}^{n-1} A^s MM^T (A^T)^s \succ 0.$$

(1) & (3) ⇒ (2). Let $\lambda \in \text{spec}(A)$ and $u \in \mathbb{C}^n$ be such that $u^H A = u^H \lambda$ and $u^H M = 0$. Because Q is a solution

$$u^H Qu = u^H AQA^T u + u^H MM^T u, \;\; [1 - |\lambda|^2] u^H Qu = u^H MM^T u = 0.$$

Because $\lambda \in \text{spec}(A) \subset \text{D}_o$, $[1 - |\lambda|^2] > 0$, hence $u^H Qu = 0$. This and $Q \succ 0$ imply that $u = 0$. Then $\lambda \in \text{spec}(A)$ is (A,M) controllable, and by Theorem 21.2.11 (A,M) is a controllable pair.

(e) Because $\text{spec}(A) \subset \text{D}_o$ there exists a real number $r \in (0,1)$ such that, $\|A\|_2 < r$. It then follows from the proof of (a) and from Proposition 17.5.8 that,

$$\begin{aligned} \|Q(t) - Q(\infty)\|_2 \leq \; & \| \sum_{s=t+1}^{\infty} A^s MM^T (A^T)^s \| \leq \sum_{s=t+1}^{\infty} \|A\|_2^{2s} \|MM^T\|_2 \\ \leq \; c & \sum_{s=t+1}^{\infty} (r^2)^s \leq \; c \frac{(r^2)^{t+1}}{1 - r^2}. \end{aligned}$$

The second conclusion follows from the above inequality and from Proposition 17.5.9.

(f) Recall from Theorem 17.4.10.(f) that for a tuple of matrices, $\text{tr}(HJ) = \text{tr}(JH)$. Define the two sequences of matrices,

$$\begin{aligned} Q_o(t+1) &= AQ_o(t)A^T + MM^T, \; Q_o(0) = 0, \\ Q_c(t+1) &= A^T Q_c(t)A + C^T C, \; Q_c(0) = 0. \end{aligned}$$

Below one uses a relation of the proof of (a), that $Q_o = 0$ and $Q_c = 0$, and that trace is a linear operator,

$$\begin{aligned} Q_o(t) &= \sum_{s=0}^{t} A^{t-s-1} MM^T (A^T)^{t-s-1}, \; Q_c(t) = \sum_{s=0}^{t} (A^T)^{t-s-1} C^T CA^{t-s-1}, \\ \text{tr}(CQ_o^* C^T) &= \lim_{t \to \infty} \text{tr}(CQ_o(t)C^T), \text{ by (a)}, \\ &= \lim \sum_{s=0}^{t} \text{tr}(CA^{t-s-1} MM^T (A^T)^{t-s-1} C^T)) \\ &= \lim \sum_{s=0}^{t} \text{tr}(M^T (A^T)^{t-s-1} C^T CA^{t-s-1} M)) \\ &= \lim \text{tr}(M^T Q_c(t)M)) = \text{tr}(M^T Q_c^* M)). \end{aligned}$$

(g) This follows along the lines of the proof of (b) in which one uses the matrix W in stead of MM^T and one does not use that $0 \preceq W$. It follows also along the lines of the proof of (b) that if $0 \preceq W$ then $0 \preceq Q$, while if $W \preceq 0$ then $Q \preceq 0$. Suppose that

$0 \preceq Q$. Let $\lambda \in \mathbb{C}$ be an eigenvalue of A and $u \in \mathbb{C}^n$ a left eigenvector, $u^H A = \lambda u^H$. Then,

$$u^H W u = u^H [Q - AQA^T] u = u^H Q u [1 - |\lambda|^2] \succeq 0,$$

by the assumption that $\lambda \in \text{spec}(A) \subset \mathrm{D}_o$. Because this holds for any left eigenvector of A one concludes that $0 \preceq W$. The proof that $W \preceq 0$ implies that $Q \preceq 0$ proceeds analogously. Note that $(Q_2 - Q_1) = A(Q_2 - Q_1)A^T + (W_2 - W_1)$. The result then follows from the preceding statements.
(h) This follows from (g). □

Consider the Lyapunov equation,

$$Q = AQA^T + MM^T,\ Q \in \mathbb{R}^{n_x \times n_x},\ A \in \mathbb{R}^{n_x \times n_x},\ M \in \mathbb{R}^{n_x \times n_v}. \tag{22.5}$$

If $\text{spec}(A) \subset \mathrm{D}_o$ and (A, M) is a controllable pair then it follows from Theorem 22.1.2 that Equation (22.5) has a unique solution that moreover satisfies $Q \in \mathbb{R}^{n_x \times n_x}_{spds}$. Define the sets,

$$\begin{aligned} \text{LSP}_{cs} &= \left\{ \begin{array}{l} (A, M) \in \mathbb{R}^{n_x \times n_x} \times \mathbb{R}^{n_x \times n_v} | \\ (A, M) \text{ controllable pair, spec}(A) \subset \mathrm{D}_o \end{array} \right\}, \\ \mathbb{R}^{n_x \times n_x}_{spds} &= \{Q \in \mathbb{R}^{n_x \times n_x} | Q = Q^T \succ 0\}. \end{aligned}$$

Define the map,

$$L(A, M) = Q,\ L : \text{LSP}_{cs} \to \mathbb{R}^{n_x \times n_x}_{pds},\ \text{where } Q \text{ is the solution of (22.5).}$$

Theorem 22.1.3. *The map $L : \text{LSP}_{cs} \to \mathbb{R}^{n_x \times n_x}_{pds}$ is continuous, in fact an analytic function, and therefore infinitely differentiable.*

Proposition 22.1.4. *Consider the asymmetric Lyapunov equation for a matrix Q,*

$$Q = A_1 Q A_2^T + S,\ Q \in \mathbb{R}^{n_x \times n_x},\ A_1,\ A_2,\ S \in \mathbb{R}^{n_x \times n_x}.$$

(a)If for all $\lambda \in \text{spec}(A)$ and $\mu \in \text{spec}(B)$, the condition $\lambda\mu \neq 1$ holds, then there exists a unique solution $Q \in \mathbb{R}^{n \times n}$ of the above asymmetric Lyapunov equation.
(b)Consider,

$$\begin{aligned} f(X) &= X - AXB^T,\ f : \mathbb{R}^{n_x \times n_x} \to \mathbb{R}^{n_x \times n_x};\ \textit{then} \\ \text{spec}(f) &= \{1 - \lambda\mu \in \mathbb{C} |\ \forall\ \lambda \in \text{spec}(A),\ \forall\ \mu \in \text{spec}(B)\}. \end{aligned}$$

22.2 Algebraic Riccati Equations of Filtering and of Control

Existence and Characterization of Solution

Consider the time-invariant Gaussian system as specified by the equations:

$$
\begin{aligned}
x(t+1) &= Ax(t)+Mv(t),\ x(0)=x_0,\ \ x_0 \in G(m_{x_0},Q_{x_0}),\\
y(t) &= Cx(t)+Nv(t),\ \ v(t) \in G(0,I).
\end{aligned}
$$

The time-invariant Kalman filter has been defined in Section 8.5. The Kalman filter involved the filter Riccati recursion of a time-invariant Gaussian system defined by the recursion,

$$Q_f(t+1) = \mathrm{f}_{\mathrm{FARE}}(Q_f(t)),\ Q_f(0) = Q_{x_0}.$$

and the filter algebraic Riccati equation $Q_f^* = \mathrm{f}_{\mathrm{FARE}}(Q_f^*)$ for a matrix Q_f^*. In this section the following problem will be investigated.

Problem 22.2.1. 1. Does the limit $\lim_{t\to\infty} Q_f(t) = Q_f(\infty)$ exist?
2. Is $Q(\infty)$ a positive-definite solution of the algebraic Riccati equation?
3. Does there exist a solution $Q_f^* \in \mathbb{R}_{pds}^{n_x\times n_x}$ to the *algebraic filter Riccati equation*,

$$Q_f^* = \mathrm{f}_{\mathrm{FARE}}(Q_f^*).$$

4. Does the algebraic Riccati equation have a unique positive-definite solution?
5. Is then,

$$
\begin{aligned}
&\mathrm{spec}(A-KQ(\infty)C) \subset \mathrm{D}_o?;\ \text{where}\\
K(Q_\infty) &= [AQ(\infty)C^T+MN^T][CQ(\infty)C^T+NN^T]^{-1}.
\end{aligned}
$$

If so, then it follows that the error system of the filter error is asymptotically stable.

Theorem 22.2.2. Filter Algebraic Riccati Equation – Existence, uniqueness, and convergence. *Consider the time-invariant Gaussian system,*

$$
\begin{aligned}
x(t+1) &= Ax(t)+Mv(t),\ x(0)=x_0 \in G(m_{x_0},Q_{x_0}),\\
y(t) &= Cx(t)+Nv(t),\ v(t) \in G(0,I),\\
&\ \ n_y \le n_v,\ \mathrm{rank}(N) = n_y\ \Rightarrow\ NN^T \succ 0.
\end{aligned}
$$

Define the matrices and functions,

$$A_f = A - MN^T(NN^T)^{-1}C \in \mathbb{R}^{n_x\times n_x}, \tag{22.6}$$

$$M_f M_f^T = MM^T - MN^T(NN^T)^{-1}(MN^T)^T,\ M_f \in \mathbb{R}^{n_x\times n_x}. \tag{22.7}$$

$$K(Q) = [AQC^T+MN^T][CQC^T+NN^T]^{-1},\ K: \mathbb{R}_{pds}^{n_x\times n_x} \to \mathbb{R}^{n_x\times n_y} \tag{22.8}$$

$$A(Q) = A - K(Q)C,\ A: \mathbb{R}_{pds}^{n_x\times n_x} \to \mathbb{R}^{n_x\times n_x}, \tag{22.9}$$

$$\mathrm{f}_{\mathrm{FARE}}: \mathbb{R}_{pds}^{n_x\times n_x},$$

$$
\begin{aligned}
\mathrm{f}_{\mathrm{FARE}}(Q) = {} & AQA^T + MM^T + \\
& -[AQC^T+MN^T][CQC^T+NN^T]^{-1}[AQC^T+MN^T]^T.
\end{aligned} \tag{22.10}
$$

Because $NN^T \succ 0$ *and* $Q \in \mathbb{R}_{pds}^{n_x\times n_x}$ *the inverse in the function* $\mathrm{f}_{\mathrm{FARE}}(Q)$ *is well defined.*

Define respectively the filter Riccati recursion *or the* filter Riccati sequence *of matrices and the* filter algebraic Riccati equation *(FARE) for a matrix* Q_f^* *by the equations,*

$$Q_f(t+1) = \mathrm{f}_{\mathrm{FARE}}(Q_f(t)),\ \ Q_f(0) = Q_{x_0},\ Q_f : T \rightarrow \mathbb{R}^{n_x \times n_x}, \tag{22.11}$$
$$Q_f^* = \mathrm{f}_{\mathrm{FARE}}(Q_f^*),\ \ Q_f^* \in \mathbb{R}^{n_x \times n_x}. \tag{22.12}$$

(a)If $(A,\ C)$ is a detectable pair and $(A_f,\ M_f)$ is a stabilizable pair, then there exists a matrix $Q_f^* \in \mathbb{R}_{pds}^{n_x \times n_x}$ which is a solution of the filter algebraic Riccati equation (22.12).

(b)If $(A_f,\ M_f)$ is a stabilizable pair then the FARE (22.12) with the condition that $Q \in \mathbb{R}_{pds}^{n_x \times n_x}$, has at most one solution.

(c)Under the assumptions of (a) and with $Q_f(0) = 0$, the limit $\lim_{t\to\infty}\ Q_f(t) = Q_f^*$ exists and Q_f^* is the positive-definite solution of the algebraic Riccati equation.

(d)If $(A_f,\ M_f)$ is a stabilizable pair and if there exists a positive-definite solution Q_f^* of FARE then $\mathrm{spec}(A - K(Q_f^*)C) \subset \mathrm{D}_o$.

(e)Consider the FARE for the matrix $Q \in \mathbb{R}_s^{n_x \times n_x}$, $Q = \mathrm{f}_{\mathrm{FARE}}(Q)$, with the conditions that $CQC^T + NN^T \succ 0$ and $\mathrm{spec}(A(Q)) \subset \mathrm{D}_o$ (but without the condition that $Q \succeq 0$). The equation with these conditions has at most one solution which solution satisfies $Q \in \mathbb{R}_{pds}^{n_x \times n_x}$.

(f) Assume that $(A_f,\ M_f)$ is a controllable pair and that there exists a $Q \in \mathbb{R}_{pds}^{n_x \times n_x}$ such that $Q = \mathrm{f}_{\mathrm{FARE}}(Q)$. Then $Q \succ 0$.

(g)Assume that $(A_f,\ M_f)$ is a stabilizable pair. Then,

$$\begin{aligned} Q_1 &= \mathrm{f}_{\mathrm{FARE}}(Q_1),\ \ Q_1 = Q_1^T \succeq 0,\ Q_1 \in \mathbb{R}_{pds}^{n_x \times n_x}, \\ Q_2 &= \mathrm{f}_{\mathrm{FARE}}(Q_2),\ \ Q_2 = Q_2^T,\ \ CQ_2C^T + NN^T \succ 0,\ Q_2 \in \mathbb{R}^{n_x \times n_x}, \\ &\Rightarrow Q_1 \succeq Q_2;\ \ \textit{(note that } Q_2 \succeq 0 \textit{ is not imposed).} \end{aligned}$$

The theorem above answers the questions posed in Problem 22.2.1.

In the literature attention is usually restricted to the case where $MN^T = 0$. In that case $(A_f, M_f) = (A, M)$ and (A_f, M_f) is a stabilizable pair if (A, M) is a stabilizable pair. This implication is not true if $MN^T \neq 0$ as the following example shows.

Example 22.2.3. Consider the Gaussian system,

$$\begin{aligned} x(t+1) &= \frac{1}{2}x(t) + \begin{pmatrix} 1 & 1 \end{pmatrix} v(t),\ x_0, \\ y(t) &= 2x(t) + \begin{pmatrix} 1 & 1 \end{pmatrix} v(t),\ v(t) \in G(0,I),\ \ \ n_x = 1,\ n_v = 2,\ n_y = 1. \end{aligned}$$

Then

$$(A,M) = \left(\frac{1}{2}, \begin{pmatrix} 1 & 1 \end{pmatrix}\right), \text{ is a stabilizable pair, while,}$$
$$A_f = \frac{1}{2} - 22^{-1}2 = -1\frac{1}{2}, M_f M_f^T = 2 - 22^{-1}2 = 0,$$
$$(A_f, M_f) = (-1\frac{1}{2}, 0) \text{ is not a stabilizable pair.}$$

To avoid confusion with the readers, below is stated the theorem of the control algebraic Riccati equation which is dual to Theorem 22.2.2. The theorem below is used in the problem of optimal stochastic control with complete observations on an infinite horizon. The proof of the next theorem is obtained from that of Theorem 22.2.2 by transposition of the system matrices according to $(A,M,C,N) \mapsto (A^T,C^T,M^T,N^T) = (A^T,C_z^T,B^T,D_z^T)$.

Theorem 22.2.4. Control Algebraic Riccati Equation – Existence, uniqueness, and convergence.
Consider the time-invariant Gaussian control system,

$$
\begin{aligned}
x(t+1) &= Ax(t)+Bu(t)+Mv(t),\ x(0)=x_0 \in G(m_{x_0},Q_{x_0}),\\
z(t) &= C_z x(t)+D_z u(t),\ v(t) \in G(0,I),\\
n_u &\leq n_z,\ \mathrm{rank}(D_z)=n_u \ \Rightarrow\ D_z^T D_z \succ 0.
\end{aligned}
$$

Denote the right-hand side of the control algebraic Riccati equation by,

$$
\begin{aligned}
&\mathrm{f}_{\mathrm{CARE}} : \mathbb{R}_{pds}^{n_x \times n_x} \to \mathbb{R}_{pds}^{n_x \times n_x},\\
&\mathrm{f}_{\mathrm{CARE}}(Q)\\
&= A^T QA + C_z^T C_z - [A^T QB + C_z^T D_z][B^T QB + D_z^T D_z]^{-1}[A^T QB + C_z^T D_z]^T.
\end{aligned}
$$

Define the matrices and functions,

$$
\begin{aligned}
A_c &= A - B(D_z^T D_z)^{-1} D_z^T C_z,\ A_c \in \mathbb{R}^{n_x \times n_x}, && (22.13)\\
C_c^T C_c &= C_z^T C_z - C_z^T D_z (D_z^T D_z)^{-1} D_z^T C_z,\ C_c \in \mathbb{R}^{n_x \times n_x}, && (22.14)\\
F(Q) &= -[B^T QB + D_z^T D_z]^{-1}[A^T QB + C_z^T D_z]^T,\ F : \mathbb{R}^{n_x \times n_x} \to \mathbb{R}^{n_u \times n_x}, && (22.15)\\
A(Q) &= A - BF(Q),\ A : \mathbb{R}^{n_x \times n_x} \to \mathbb{R}^{n_x \times n_x}. && (22.16)
\end{aligned}
$$

Define respectively the control Riccati recursion *or the* control Riccati sequence*, and the* control algebraic Riccati equation *(CARE) for the matrix Q_c^* by,*

$$
\begin{aligned}
Q_c(t+1) &= \mathrm{f}_{\mathrm{CARE}}(Q_c(t)),\ \ Q_c(0) = C_z^T C_z,\ Q_c : T \to \mathbb{R}^{n_x \times n_x}, && (22.17)\\
Q_c^* &= \mathrm{f}_{\mathrm{CARE}}(Q_c^*),\ Q_c^* \in \mathbb{R}^{n_x \times n_x}. && (22.18)
\end{aligned}
$$

(a)If $(A,\ C)$ is a stabilizable pair and $(A_c,\ C_c)$ is a detectable pair, then there exists a matrix $Q_c^* \in \mathbb{R}_{pds}^{n_x \times n_x}$ which is a solution of the control algebraic Riccati equation (22.18).

(b)If $(A_c,\ C_c)$ is a detectable pair then the equation (22.18) with the condition that $Q_c \in \mathbb{R}_{pds}^{n_x \times n_x}$, has at most one solution.

(c)Under the assumptions of (a) and with $Q_c(0) = 0$, the limit $\lim_{t \to \infty} Q_c(t) = Q_c^*$ exists and the matrix Q_c^* is the positive definite-solution of the control algebraic Riccati equation.

(d)If $(A_c,\ C_c)$ is a detectable pair and if there exists a positive-definite solution Q_c of CARE then $\mathrm{spec}(A + BF(Q)) \subset \mathrm{D}_o$.

(e)Consider the control algebraic Riccati equation for $Q \in \mathbb{R}_s^{n_x \times n_x}$, $Q = f(Q)$, with the conditions that $B^T QB + D_z^T D_z \succ 0$ and $\mathrm{spec}(A(Q)) \subset \mathrm{D}_o$ (but without the condition that $Q \succeq 0$). The equation with these conditions has at most one solution and the solution satisfies $Q \in \mathbb{R}_{pds}^{n_x \times n_x}$.

(f) Assume that $(A_c,\ C_c)$ is an observable pair and that there exists a $Q_c \in \mathbb{R}_{pds}^{n_x \times n_x}$ such that $Q_c = \mathrm{f}_{\mathrm{CARE}}(Q_c)$. Then $Q_c \succ 0$.

(g)Assume that $(A_c,\, C_c)$ is a detectable pair. Then,

$$\begin{aligned} Q_1 &= \mathrm{f}_{\mathrm{CARE}}(Q_1),\;\; Q_1 = Q_1^T \succeq 0,\; Q_1 \in \mathbb{R}_{pds}^{n_x \times n_x}, \\ Q_2 &= \mathrm{f}_{\mathrm{CARE}}(Q_2),\;\; Q_2 = Q_2^T,\;\; B^T Q_2 B + D_z^T D_z \succ 0,\; Q_2 \in \mathbb{R}^{n_x \times n_x}, \\ &\Rightarrow Q_1 \succeq Q_2;\;\; \textit{(note that } Q_2 \succeq 0 \textit{ is not imposed).} \end{aligned}$$

There follows a discussion on the convergence rate of the solution of the filter Riccati recursion. The following concept is needed in optimal stochastic control with the average cost function.

Definition 22.2.5. Consider the filter Riccati recursion defined above. Assume that there exists a unique solution of the Filter Algebraic Riccati equation, which will be denoted by $Q_f(\infty)$. The filter Riccati recursion is said to satisfy the *condition of an exponential-asymptotic convergence rate* of the filter Riccati sequence if there exists a time $t_1 \in T = \mathbb{N}$, a positive constant $c \in \mathbb{R}_+$, and a convergence rate $r \in (0,1)$, such that,

$$t \geq t_1 \;\Rightarrow\; \|Q_f(t) - Q_f(\infty)\|_2 \leq c\, r^{t-t_1}.$$

In control theory one prefers a small convergence rate, r_1, over a large convergence rate, r_2, thus $r_1 < r_2$.

Does there exists an exponential-asymptotic convergence rate? A condition of controllability or stabilizability will establish the existence as will become clear later in this chapter. Can one bound the convergence rate? The answer in general is no!

J.C. Doyle has focused attention on the performance of LQ and of LQG control, [13]. These research issues are related to the above questions on the exponential-asymptotic convergence rate.

The framework for the asymptotic behavior of the eigenvalues of a closed-loop optimal control system has been investigated. But this part of control theory is not widely known. The theory may be found in the book of H. Kwakernaak and R. Sivan, [20, continuous-time case, Sec. 3.8, Thm. 3.11; discrete-time case, Subsec. 6.4.7, Thm. 6.37] for the single-input case and in the paper by H. Kwakernaak, [19], for the multivariable case. The basis for these results are books and papers on the algebraic geometry of algebraic curves, not stated here.

There follows a long description on the convergence rate.

Case of the filter problem. Consider a time-invariant Gaussian stochastic system with the representation,

$$\begin{aligned} x(t+1) &= Ax(t) + Mv(t),\; x(0) = x_0, \\ y(t) &= Cx(t) + rNv(t),\;\; r \in (0,\infty),\; \text{input-output system } v \mapsto y. \end{aligned}$$

Case of control problem. In case of a stochastic control problem, the linear system of interest is described by the equations,

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t),\; x(0) = x_0, \\ z(t) &= C_z x(t) + rD_z u(t),\;\; r \in (0,\infty) \text{ input-output system } u \mapsto z. \end{aligned}$$

Below mainly the case of a control problem is described. The reader can then relate this to the filter problem by duality of the equations.

Next the concept of a zero of a linear system is needed. This concept is formulated in Section 21.5. In fact, more useful for filtering and for control are the zeros of the associated Hamiltonian linear system, see Proposition 22.2.14 and Proposition 22.2.15 below in this chapter.

An example of a continuous-time linear system which has a system zero at the complex number zero, is the inverted pendulum, [20, Ex. 1.1].

If a zero of a discrete-time linear system exists then it is of interest to know whether the zero belongs to the subset D_o, or to the the unit circle $\{c \in \mathbb{C} | \ |c| = 1\}$, or to the outside of the unit disc, $\{c \in C | \ |c| > 1\}$.

Consider the optimal stochastic control problem Problem 13.2.8 for the infinite-horizon with average cost for a Gaussian stochastic control system with complete observations. The optimal control law leads to a closed-loop system of which the system matrix equals $A+BF$. Of interest is thus the question whether the eigenvalues of the closed-loop system are in the stable part of the complex place, thus in D_o. It is stated in Chapter 13 that if a condition of controllability or of stabilizability holds then the eigenvalues of the closed-loop system matrix $A+BF$ are in D_o.

For the performance analysis of the closed-loop system it is necessary to know where in the complex plane the eigenvalues of the closed-loop system matrix $A+BF$ are located. In particular, if the parameter r in the linear control system goes to zero. This question has been investigated.

If the parameter r of the cost on the input goes down to zero but is not zero yet, $\lim_{r \downarrow 0} \ rN$, then it is known that the eigenvalues of the filter error system in case of a control problem show the following behavior:

- there may exist one or more eigenvalues at the zero of the complex plane which do not change when r changes;
- there may exist one or more other eigenvalues which in the limit with r converge to the locations of the zeros of the linear system within D_o; here the deterministic linear system considered has as input u and as output the controlled output z;
- the remaining eigenvalues, if any, will converge to the origin of the complex plane according to the discrete-time Butterworth pattern with several radii.

The theory of this asymptotic behavior is stated in the book [20, continuous-time case, Sec. 3.8, Thm. 3.11; discrete-time case, Subsec. 6.4.7, Thm. 6.37] for the single-input case and in the paper by H. Kwakernaak, [19], for the multivariable case.

The *Butterworth pattern* of eigenvalues of a time-invariant continuous-time linear system, or of the poles of the corresponding transfer function, were described by S. Butterworth in a paper published in 1930, [10]. The corresponding discrete-time Butterworth pattern is described in [20, Subsec. 6.4.7, Thm. 6.37].

If the open-loop control system described above has a system zero located at $c = 1 \in \mathbb{C}$ on the stability boundary, or inside D_o but very near the instability boundary (for example in $1-\varepsilon$ for a small $\varepsilon \in (0,1)$), then for $r \downarrow 0$, at least one eigenvalue of the closed-loop system matrix $A+BF$ will move towards this zero of the linear system. Therefore the performance of the closed-loop system is directly affected by this eigenvalue and hence the exponential-asymptotic convergence rate will also be related to this eigenvalue and may therefore be very slow. For example, if the control system has a system zero located at $c = 1 \in \mathbb{C}$ then an eigenvalue of the closed-loop system matrix $A + BF$ will approach the value of the zero from inside the unit disc.

A direct consequence of this result is that one cannot establish a performance guarantee on the eigenvalues of the closed-loop system matrix $A + BF$ for general stochastic control systems. Meaning, that one cannot guarantee that the eigenvalues are a prespecified distance away from the unit circle. The same conclusion holds for the system matrix of the error filter system with the system matrix $A - KC$.

However, note that the controlled output z and hence the cost criterion of an optimal stochastic control problem will either not depend on the eigenvalue which converges to the linear system zero or will depend to an amount which disappears with diminshing r. This follows directly from the definition of a linear system zero which is based on a zero output for the particular mode described by the zero. Due to the use of numerical approximations, in practice there may be a slight dependence of z on the considered eigenvalue.

For any particular system as described above, one can compute whether there exists a system zero. If no zero exists or if the zero is significantly bounded away from the boundary of the instability in the complex plane, then performance bounds can be obtained but these are dependent on the value of the system zero.

J.C. Doyle is quite correct when he states that in general there are no performance guarantees for LQ control, [13]. More precisely, in general there are no performance guarantees in regard to the eigenvalues of the closed-loop system for the set of all controllable and observable linear deterministic systems if the zeros of the associated open-loop system cannot be bounded away from the unit circle. Only for particular systems after investigating the existence of system zeros, can one formulate a performance guarantee.

Proof for the Filter Algebraic Riccati Equation

A proof for the existence of a solution of the control algebraic Riccati equation in continuous-time was first published by W.M. Wonham, [36]. The proof of the discrete-time case of this section is based on that paper.

The proof of the convergence of the filter algebraic Riccati equation is explained for the control algebraic Riccati equation. The proof of convergence of the control algebraic Ricatti recursion is based on the convergence of the value function of the LQG optimal stochastic control with complete observations on an infinite-horizon with either the average-cost function or the discounted-cost function. The value function of such a problem has the form $V(t,x) = x^T Q_c(t) x$.

Thus convergence of the control Riccati sequence $\{Q(t) \in \mathbb{R}^{n_x \times n_x}_{pds},\ t \in T = \mathbb{N}\}$, $\lim_{t\to\infty} Q_c(t) = Q_c(\infty)$, is equivalent to the corresponding convergence of the value function V. The limit value of the sequence of value functions, $V(\infty,x) = \lim_{t\to\infty} V(t,x)$, satisfies a particular dynamic programming equation. But that dynamic programming equation, both for the average-cost and for the discounted-cost, is known not to have a unique solution. What one needs is the minimum of all solutions of that dynamic programming equation. One also needs a procedure to compute the minimal value function of the limit. Such a procedure was developed in control theory and in dynamic programming, and it is used to provide a proof of the convergence of the control algebraic Riccati equation.

It is known from control theory that the convergence of the value function is proven by the use of a lower bound and of an upper bound. This procedure is then also used for the convergence of the control Riccati recursion. Thus one constructs besides the control Riccati recursion two recursions, a lower bound and an upper bound, such that,

$$\begin{aligned}
&\{Q_{c,lb}(t) \in \mathbb{R}^{n \times n}_{pds},\ t \in T\},\ \{Q_{c,ub}(t) \in \mathbb{R}^{n \times n}_{pds},\ t \in T\},\\
&\forall\, t \in T,\ \ Q_{c,lb} \preceq Q_c(t) \preceq Q_{c,ub}.
\end{aligned}$$

The lower-bound sequence is constructing by starting at the zero matrix according to the formulas,

$$Q_{c,lb}(t+1) = f_{CARE}(Q_{c,lb}(t)),\ Q_{c,lb}(0) = 0.$$

It is then proven that the lower-bound sequence is monotonically increasing, that it converges, $Q_{c,lb}(\infty) = \lim_{t \to \infty} Q_{c,lb}(t)$, and that the limit is finite. The upper-bound sequence has then to be constructed so that it is an upper-bound and that it converges also to the same limit value, $\lim_{t \to \infty} Q_{c,ub}(t) = Q_{c,lb}(\infty)$. Such an upper-bound sequence can be constructed based on the assumption of controllability or of stabilizability, as shown later in this section for the filter Riccati recursion. Due to the lower-bound and the upper-bound sequences converging to the same limit, the convergence of the control Riccati recursion follows with the same limit value.

The procedure for the convergence of the value function and for the solution of the dynamic programming equation on an infinite-horizon with either discounted cost or average cost can also be used for other stochastic control systems than a time-invariant Gaussian stochastic control system. It is essential for the lower bound that the iteration at time zero starts at the zero value of the value function.

The proof of Theorem 22.2.2 is based on several preliminaries.

Definition 22.2.6. Consider the matrices A, C, M, N of Theorem 22.2.2. Define the *observation-Lyapunov map*,

$$\begin{aligned}
\mathrm{L_o} &: \mathbb{R}^{n_x \times n_x} \times \mathbb{R}^{n_x \times n_y} \to \mathbb{R}^{n_x \times n_x},\\
\mathrm{L_o}(Q, K) &= [A - KC]Q[A - KC]^T + [M - KN][M - KN]^T.
\end{aligned}$$

For a fixed matrix $K \in \mathbb{R}^{n_x \times n_y}$, the equation $Q = \mathrm{L_o}(Q, K)$ is the Lyapunov equation that has been analyzed in Theorem 22.1.2. Next useful relations are summarized.

Proposition 22.2.7. *Consider the matrix functions and matrices,*

$$\begin{aligned}
&\mathrm{f_{FARE}},\ \mathrm{L_o},\ Q_1,\ Q_2 \in \mathbb{R}^{n_x \times n_x}_{pds},\ L \in \mathbb{R}^{n_x \times n_y},\\
K(Q_1) &= [AQ_1C^T + MN^T][CQ_1C^T + NN^T]^{-1},\\
K(Q_2) &= [AQ_2C^T + MN^T][CQ_2C^T + NN^T]^{-1}.
\end{aligned}$$

(a)

$$\begin{aligned} \mathrm{f}_{\mathrm{FARE}}(Q_1) = {} & [A - K(Q_1)C]Q_1[A - K(Q_1)C]^T + \\ & +[M - K(Q_1)N][M - K(Q_1)N]^T \succeq 0. \end{aligned}$$

(b)$\mathrm{f}_{\mathrm{FARE}}(Q_1) = \mathrm{L_o}(Q_1, K(Q_1))$.
(c)$\mathrm{L_o}(Q_1, K) - \mathrm{f}_{\mathrm{FARE}}(Q_1) = (K - K(Q_1))[CQ_1C^T + NN^T](K - K(Q_1))^T$.
(d)

$$\begin{aligned} & \mathrm{L_o}(Q_2, K_2) - \mathrm{f}_{\mathrm{FARE}}(Q_1) = (A - K_2C)[Q_2 - Q_1](A - K_2C)^T + \\ & \quad +(K_2 - K(Q_1))[CQ_1C^T + NN^T](K_2 - K(Q_1))^T. \end{aligned}$$

(e)

$$\begin{aligned} & \mathrm{f}_{\mathrm{FARE}}(Q_2) - \mathrm{f}_{\mathrm{FARE}}(Q_1) = (A - K(Q_2)C)[Q_2 - Q_1](A - K(Q_2)C)^T + \\ & +(K(Q_2) - K(Q_1))[CQ_1C^T + NN^T](K(Q_2) - K(Q_1))^T. \end{aligned}$$

(f) If $Q_2 \geq Q_1$ *then* $\mathrm{f}_{\mathrm{FARE}}(Q_2) \geq \mathrm{f}_{\mathrm{FARE}}(Q_1)$.

Proof. Denote,

$$Q_{x^+,y} = AQ_1C^T + MN^T,\ Q_y = CQ_1C^T + NN^T,\ K_1 = K(Q_1) = Q_{x^+,y}Q_y^{-1}.$$

(a)

$$\begin{aligned} & [A - K(Q_1)C]Q_1[A - K(Q_1)C]^T + [M - K(Q_1)N][M - K(Q_1)N]^T \\ & = AQ_1A^T - AQ_1C^TK(Q_1)^T - K(Q_1)CQ_1A^T + K(Q_1)CQ_1C^TK(Q_1)^T \\ & \quad +MM^T - MN^TK(Q_1)^T - K(Q_1)NM^T + K(Q_1)NN^TK(Q_1)^T \\ & = AQ_1A^T + MM^T - K(Q_1)^TQ_y^{-1}K(Q_1)^T = \mathrm{f}_{\mathrm{FARE}}(Q_1). \end{aligned}$$

(b) This follows from (a) and the definition of the function $\mathrm{L_o}$.
(c) Let $K_1 = K(Q_1)$.

$$\begin{aligned} & L_o(Q_1, K) - f(Q_1) \\ & = [A - K_1C - (K - K_1)C]Q_1[A - K_1C - (K - K_1)C]^T \\ & \quad +[M - K_1N - (K - K_1)N][M - K_1N - (K - K_1)N]^T \\ & \quad -(A - K_1C)Q_1(A - K_1C)^T - (M - K_1N)(M - K_1N)^T \text{ by (a).} \\ & = -(A - K_1C)Q_1C^T(K - K_1)^T - (K - K_1)CQ_1(A - K_1C)^T \\ & \quad +(K - K_1)CQ_1C^T(K - K_1)^T + (K - K_1)NN^T(K - K_1)^T \\ & \quad -(M - K_1N)N^T(K - K_1)^T - (K - K_1)N(M - K_1N)^T \\ & = -(Q_{x^+,y} - K_1Q_y)(K - K_1)^T - (K - K_1)(Q_{x^+,y} - K_1Q_y)^T + \\ & \quad +(K - K_1)Q_y(K - K_1)^T \\ & = (K - K_1)Q_y(K - K_1)^T,\ \text{ by } K(Q_1) = Q_{x^+,y}Q_y^{-1}. \end{aligned}$$

(d)

$$\begin{aligned}
&\mathrm{L_o}(Q_2,K_2) - f(Q_1) \\
&= \mathrm{L_o}(Q_2,K_2) - \mathrm{L_o}(Q_1,K_2) + (K_2-K_1)[CQ_1C^T + NN^T](K_2-K_1)^T \text{ by (c)}, \\
&= (A-K_2C)[Q_2-Q_1](A-K_2C)^T + (K_2-K_1)[CQ_1C^T+NN^T](K_2-K_1)^T.
\end{aligned}$$

(e) This follows from (b) and (d).
(f) This follows from (e). □

Next a transformation is made from the case $MN^T \neq 0$ to the case with $M_fN^T = 0$.

Proposition 22.2.8. Transformation of system matrices to independent state noise and output noise. *Define (A_f, M_f) as in Theorem 22.2.2 and let $R = MN^T(NN^T)^{-1}$.*

(a)For any $K \in \mathbb{R}^{n_x \times n_y}$

$$\begin{aligned}
&\mathrm{L_o}(Q,K) \\
&= [A_f - (K-R)C]Q[A_f-(K-R)C]^T + M_fM_f^T + (K-R)NN^T(K-R)^T.
\end{aligned}$$

(b)Let,

$$\begin{aligned}
&K_1(Q) = A_fQC^T[CQC^T+NN^T]^{-1}. \textit{ Then,} \\
&K_1(Q) = K(Q) - R,\ A_f - K_1(Q)C = A - K(Q)C.
\end{aligned}$$

(c)

$$\mathrm{f_{FARE}}(Q) = A_fQA_f^T + M_fM_f^T - A_fQC^T[CQC^T+NN^T]^{-1}CQA_f^T,$$

which corresponds to an algebraic Riccati equation in which no term like MN^T is present. This has been the purpose of the transformation.

(d)

$$\mathrm{f_{FARE}}(Q) = [A_f - K_1(Q)C]Q[A_f-K_1(Q)C]^T + M_fM_f^T + K_1(Q)NN^TK_1(Q)^T.$$

Proof. (a)

$$\begin{aligned}
&A_f = A - RC, \\
&\quad A_f - (K-R)C = A - RC - (K-R)C = A - KC. \\
&\quad \mathrm{L_o}(Q,K) - [A_f-(K-R)C]Q[A_f-(K-R)C]^T - M_fM_f^T - (K-R)NN^T(K-R)^T \\
&\quad = (M-KN)(M-KN)^T - M_fM_f^T - (K-R)NN^T(K-R)^T, \\
&\qquad \text{by definition of } L_o(Q,K), \\
&\quad = -MN^TK^T - K(MN^T)^T + KNN^TK^T + RNN^TR^T \\
&\qquad -KNN^TK^T + KNN^TR^T + RNN^TK^T - RNN^TR^T, \\
&\qquad \text{by definition of } M_f, \\
&\quad = 0, \text{ by definition of } R = MN^T(NN^T)^{-1}.
\end{aligned}$$

(b)

$$\begin{aligned}
K(Q) &= [AQC^T + MN^T][CQC^T + NN^T]^{-1} = [AQC^T + MN^T]Q_y^{-1},\\
K(Q) - R &= [A_f QC^T + RCQC^T + MN^T]Q_y^{-1} - R, \text{ by definition of } A_f,\\
&= [A_f QC^T + RCQC^T + MN^T - R(CQC^T + NN^T)]Q_y^{-1}\\
&= A_f QC^T Q_y^{-1} = K_1(Q),\\
A_f - K_1(Q)C &= A_f - A_f QC^T[CQC^T + NN^T]^{-1}C, \text{ by def. of } K_1(Q),\\
&= A - MN^T(NN^T)^{-1}C - [K(Q) - R]C, \text{ by the above calculation},\\
&= A - K(Q)C.
\end{aligned}$$

(c)

$$\begin{aligned}
\mathrm{f}_{\mathrm{FARE}}(Q) &= \mathrm{L_o}(Q, K(Q))\\
&= A_f Q A_f^T - (K(Q) - R)CQA_f^T - A_f QC^T(K(Q) - R)^T\\
&\quad +(K(Q) - R)CQC^T(K(Q) - R)^T + M_f M_f^T\\
&\quad +(K(Q) - R)NN^T(K(Q) - R)^T \text{ by (a)},\\
&= A_f Q A_f^T + M_f M_f^T - A_f QC^T[CQC^T + NN^T]^{-1}CQA_f^T,\\
&\quad \text{by } A_f QC^T Q_y^{-1} = K(Q) - R.
\end{aligned}$$

(d)

$$\begin{aligned}
K &= R + K_1(Q) \Rightarrow A_f - (K - R)C = A_f - K_1(Q)C,\\
\mathrm{f}_{\mathrm{FARE}}(Q) &= \mathrm{L_o}(Q, K(Q)), \text{ by Theorem 22.2.7.(b)},\\
&= [A_f - K_1(Q)C]Q[A_f - K_1(Q)C]^T + M_f M_f^T + K_1(Q)NN^T K_1(Q)^T,\\
&\quad \text{by (a)}.
\end{aligned}$$

□

Proposition 22.2.9. *(a)If (A_f, M_f) is a stabilizable pair and if,*

$$\begin{aligned}
&\exists\, K \in \mathbb{R}^{n_x \times n_y},\ \exists\, Q_2 \in \mathbb{R}_{pds}^{n_x \times n_x}, \text{ such that,}\\
&Q_2 = \mathrm{L_o}(Q_2, K), \text{ then } \mathrm{spec}(A - KC) \subset \mathrm{D}_o.
\end{aligned}$$

(b)If in addition, there exists a $Q_1 \in \mathbb{R}_{pds}^{n_x \times n_x}$ such that $Q_1 = f(Q_1)$, then $Q_2 \succeq Q_1$.

Proof. (a)

$$\begin{aligned}
R &= MN^T(NN^T)^{-1},\ U = (K - R)(NN^T)^{\frac{1}{2}},\ Q_r = -(NN^T)^{-\frac{1}{2}}C,\\
JJ^T &= M_f M_f^T + (K - R)NN^T(K - R)^T = M_f M_f^T + UU^T;\ J \in \mathbb{R}^{n_x \times n_x};\\
&\Rightarrow A_f + UQ_r = A_f - (K - R)C = A - KC.
\end{aligned}$$

From (A_f, M_f) a stabilizable pair, the expresssion for JJ^T, and 21.2.12, it follows that $(A_f + UQ_v, J) = (A - KC, J)$ is a stabilizable pair. Then the definition of $L_o(Q_2, k)$, Proposition 22.2.8.(a), and the definition of J imply that,

$$Q_2 = \mathrm{L_o}(Q_2, K) = (A - KC)Q_2(A - KC)^T + JJ^T,\ Q_2 \in \mathbb{R}_{pds}^{n_x \times n_x}.$$

This equation, the fact that $Q_2 \in \mathbb{R}_{pds}^{n_x \times n_x}$, $(A_f + UQ_v, J) = (A - KC, J)$ is a stabilizable pair, and Thm. 22.1.2.(c), imply that $\text{spec}(A - KC) \subset \mathrm{D}_o$.
(b)

$$\begin{aligned} Q_2 - Q_1 &= \mathrm{L_o}(Q_2, K) - f(Q_1) \\ &= (A - KC)[Q_2 - Q_1](A - KC)^T + (K_1 - K)[CQ_1C^T + NN^T](K_1 - K)^T, \end{aligned}$$

by Proposition 22.2.7.(d). This, (a) above, and Thm. 22.1.2.(b) imply that $Q_2 - Q_1 \succeq 0$. □

Proposition 22.2.10. *Consider $Q : T \to \mathbb{R}_{pds}^{n_x \times n_x}$. Assume that the sequence Q is monotone: for all $t \in T$, either $Q(t+1) \succeq Q(t)$ or $Q(t+1) \preceq Q(t)$. Assume further that there exists an upper bound matrix $Q_{ub} \in \mathbb{R}_{pds}^{n_x \times n_x}$ such that for all $t \in T$ $0 \preceq Q(t) \preceq Q_{ub}$.*

Then the limit $\lim_{t \to \infty} Q(t) = Q(\infty)$ exists, $Q(\infty) \in \mathbb{R}_{pds}^{n_x \times n_x}$, and $0 \preceq Q(\infty) \preceq Q_{ub}$.

Proof. Suppose that the sequence Q is increasing. Take $u \in \mathbb{R}^n$. Then
$0 \leq u^T Q(t) u \leq u^T Q(t+1) u \leq u^T Q_{ub} u$. By a result in real analysis there exists a $w_u \in R$ such that
$\lim_{t \to \infty} u^T Q(t) u = w_u$. Define $Q(\infty) \in \mathbb{R}^{n_x \times n_x}$ as follows. If $i \in Z_n$ let e_i be the i-th unit vector. Then $Q_{ii}(\infty) = \lim_{t \to \infty} e_i^T Q(t) e_i \geq 0$. For $i, j \in Z_n$ let $u = e_i + e_j$,

$$Q_{ij}(\infty) = Q_{ji}(\infty) = \frac{1}{2}[\lim_{t \to \infty} u^T Q(t) u - Q_{ii}(\infty) - Q_{jj}(\infty)].$$

Then $Q_(\infty) = Q(\infty)^T$, $0 \preceq Q(\infty) \preceq Q_{ub}$, and $Q = \lim_{t \to \infty} Q(t)$. □

Proof. Proof of Theorem 22.2.2.
(*Step 1.*) (A, C) a detectable pair implies that there exists a $K_0 \in \mathbb{R}^{n \times p}$ such that $\text{spec}(A - K_0 C) \subset \mathrm{D}_o$. By Theorem 22.1.2 there exists a unique $P_0 \in \mathbb{R}^{n_x \times n_x}$ such that

$$P_0 = L_o(P_0, K_0) = (A - K_0C)P_0(A - K_0C)^T + (M - K_0N)(M - K_0N)^T \in \mathbb{R}_{pds}^{n_x \times n_x}.$$

Define $K_1 = [AP_0C^T + MN^T][CP_0C^T + NN^T]^{-1}$.
(*Step 2.*) It will be proven that $\text{spec}(A - K_1C) \subset \mathrm{D}_o$.

$$\begin{aligned} P_0 &= L_o(P_0, K_0) = f(P_0) + (K_0 - K_1)[CP_0C^T + NN^T](K_0 - K_1)^T \text{ by 22.2.7.c.} \\ &= [A - K_1C]P_0[A - K_1C]^T + M_f M_f^T + (K_1 - R)NN^T(K_1 - R)^T + \\ &\quad + (K_0 - K_1)[CP_0C^T + NN^T](K_0 - K_1)^T, \\ &\quad \text{by } f(P_0) = L_o(P_0, K_1) \text{ and Proposition 22.2.8.(b)+(d),} \\ &= (A - K_1C)P_0(A - K_1C) + JJ^T, \text{ where,} \\ J_1J_1^T &= M_f M_f^T + (K_1 - R)NN^T(K_1 - R)^T, \\ JJ^T &= J_1J_1^T + (K_0 - K_1)[CP_0C^T + NN^T](K_0 - K_1)^T. \end{aligned}$$

As in the proof of Proposition 22.2.9, (A_f, M_f) a stabilizable pair and $A_f - (K_1 - R)C = A - K_1C$ implies that $(A - K_1C, J_1)$ is a stabilizable pair. Using Proposition 21.2.12 one obtains that $(A - K_1C, J)$ is a stabilizable pair. This, the above

expression for P_0, and Theorem 22.1.2.(c) yield that $\text{spec}(A-K_1C) \subset \text{D}_o$.
(*Step 3.*) Again by Theorem 22.1.2 there exists a $P_1 \in \mathbb{R}_{pds}^{n_x \times n_x}$ such that $P_1 = g(P_1, K_1)$. It will be proven that $P_1 \leq P_0$. By definition of K_1, $L_o(P_0, K_1) = f(P_0)$, thus,

$$\begin{aligned}
P_0 - P_1 &= L_o(P_0,K_0) - L_o(P_1,K_1) \\
&= L_o(P_0,K_1) - L_o(P_1,K_1) + (K_0-K_1)[CP_0C^T + NN^T](K_0-K_1)^T, \\
&\quad \text{by Proposition 22.2.7.(b) \& (c),} \\
&= (A-K_1C)[P_0-P_1](A-K_1C)^T + (K_0-K_1)[CP_0C^T+NN^T](K_0-K_1)^T.
\end{aligned}$$

This, $\text{spec}(A-K_1C) \subset \text{D}_o$, and Theorem 22.1.2.(b) yield that $P_0 - P_1 \succeq 0$.
(*Step 4.*) The above procedure may be continued by defining for $t \in T$, $t \geq 2$,

$$\begin{aligned}
K(t) &= [AP(t-1)C^T + MN^T][CP(t-1)C^T + NN^T]^{-1}, \\
P(t) &= L_o(P(t), K(t)) \in \mathbb{R}_{pds}^{n_x \times n_x}.
\end{aligned}$$

By induction it may then be proven, as in (Step 2) and (Step 3), that
$\text{spec}(A-K(t)C) \subset \text{D}_o$, and $P(t) \succeq P(t+1)$ for all $t \in T$. Then
$P_0 \succeq P(t) \succeq P(t+1) \succeq 0$ imply by Proposition 22.2.10 that there exists a $Q \in \mathbb{R}_{pds}^{n_x \times n_x}$
such that $Q = \lim_{t\to\infty} P(t)$.
(*Step 5.*) It will be shown that $Q = \text{f}_{\text{FARE}}(Q)$. By step 4, $\lim_{t\to\infty} P(t) = Q$ in which the convergence is monotone, in fact decreasing. Let

$$\begin{aligned}
T(t) &= CP(t)C^T + NN^T,\ T = CQC^T + NN^T, \\
S(t) &= AP(t)C^T + MN^T,\ S = AQC^T + MN^T.
\end{aligned}$$

Then $\lim_{t\to\infty} T(t) = T \succ 0$, hence $\lim_{t\to\infty} T(t)^{-1} = E$ exists. Then also

$$I = \lim T(t)^{-1}T(t) = (\lim T(t)^{-1})(\lim T(t)) = ET,$$

and similarly $TE = I$, hence $E = T^{-1}$, $T^{-1} = \lim T(t)^{-1}$. Then
$\lim K(t+1) = \lim S(t)T(t)^{-1} = ST^{-1} = K$. Finally

$$\begin{aligned}
Q &= \lim_{t\to\infty} P(t+1) \\
&= \lim [A-K(t)C]P(t)[A-K(t)C]^T + [M-K(t)N][M-K(t)N]^T \\
&= [A-KC]Q[A-KC]^T + [M-KN][M-KN]^T \\
&= L_o(Q,K) = f(Q) = \text{f}_{\text{FARE}}(Q), \text{ by definition of } K.
\end{aligned}$$

(Step 1) through (Step 5) prove part (a).
(*Step 6.*) (A_f, M_f) a stabilizable pair, $Q = f(Q) = L_o(Q,K)$, and $Q = Q^T \succeq 0$, imply by Proposition 22.2.9.(a) that $\text{spec}(A-KC) \subset \text{D}_o$. This proves part (d).
(*Step 7.*) Suppose there exist two positive solutions Q_1, Q_2 of the algebraic Riccati equation. Then (A_f, M_f) a stabilizable pair, $Q_1 = f(Q_1) = L_o(Q_1, K_1)$, $Q_2 = f(Q_2)$, imply by Proposition 22.2.9.(b) that $Q_2 \leq Q_1$. Similarly $Q_2 \geq Q_1$, hence $Q_2 = Q_1$.
(Step 6) and (Step 7) prove part (b).
(*Step 8.*) Let $Q \in \mathbb{R}_{pds}^{n_x \times n_x}$ satisfy $Q = \text{f}_{\text{FARE}}(Q)$. By (Step 6) $\text{spec}(A-KC) \subset \text{D}_o$. Define the function,

$$Q_{ub}(t+1) = g(Q_{ub}(t), K),\ Q_{ub}(0) = Q_0,\ \ Q_{ub} : T \to \mathbb{R}^{n_x \times n_x}.$$

Then $\text{spec}(A-KC) \subset \text{D}_o$ and Theorem 22.1.2.(a) imply that $\lim_{t\to\infty} Q_{ub}(t) = Q_1$ exists, and $Q_1 = L_o(Q_1, K)$, $Q_1 \in \mathbb{R}^{n_x \times n_x}_{pds}$. Then,

$$\begin{aligned} Q_1 - Q &= L_o(Q_1,K) - f(Q) = L_o(Q_1,K) - L_o(Q,K), \text{ by Proposition 22.2.7.(b)} \\ &= (A-KC)[Q_1 - Q](A-KC)^T. \end{aligned}$$

Then Theorem 22.1.2.(b) and $\text{spec}(A-KC) \subset \text{D}_o$ imply that this equation has a unique solution, namely $Q_1 - Q = 0$. Hence $\lim Q_{ub}(t) = Q_1 = Q$.
(*Step 9.*) Define $Q_{lb} : T \to \mathbb{R}^{n_x \times n_x}$ by the recursion $Q_{lb}(0) = 0$ and $Q_{lb}(t+1) = \text{f}_{\text{FARE}}(Q_{lb}(t))$. It will be proven that for all $t \in T$, $Q_{lb}(t+1) \succeq Q_{lb}(t)$. Clearly $Q_{lb}(1) = f(Q_{lb}(0)) \succeq 0 = Q_{lb}(0)$ by Proposition 22.2.7.(a), while if $Q_{lb}(t) \succeq Q_{lb}(t-1)$ then by Proposition 22.2.7.(f) $QW_{lb}(t+1) = f(Q_{lb}(t)) \succeq f(Q_{lb}(t-1)) = Q_{lb}(t)$. Similarly it will be shown that for all $t \in T$, $Q_{lb}(t) \le Q_{ub}(t)$. Clearly $Q_{lb}(0) = 0 \le Q_{ub}(0)$. If $Q_{lb}(t) \le Q_{ub}(t)$ then

$$Q_{ub}(t+1) - Q_{lb}(t+1) = L_o(Q_{ub}(t),K) - f(Q_{lb}(t)) \succeq 0, \text{ by 22.2.7.(d).}$$

Now $Q_{lb} : T \to \mathbb{R}^{n_x \times n_x}$ is an increasing and bounded sequence, thus by Proposition 22.2.10 $\lim_{t\to\infty} Q_{lb}(t) = Q_2$ exists. As in (Step 5) it can be shown that $Q_2 = f(Q_2)$, $Q_2 = Q_2^T \ge 0$, and by part (b) $Q_2 = Q$.
(*Step 10.*) Finally let $Q : T \to \mathbb{R}^{n_x \times n_x}_{pds}$ $Q(t+1) = \text{f}_{\text{FARE}}(Q(t))$ and $Q(0) = Q_{x_0}$. It will be proven that for all $t \in T$ $Q_{ub}(t) \succeq Q(t) \succeq Q_{lb}(t)$. For $t = 0$, $Q_{ub}(0) = Q(0) \succeq 0 = Q_{lb}(0)$. Suppose it holds for $t \in T$. Then

$$Q_{ub}(t+1) - Q(t+1) = L_o(Q_{ub}(t),K) - f(Q(t)) \succeq 0,$$

by Proposition 22.2.7.(d), while from Proposition 22.2.7.(f) and $Q(t) \succeq Q_{lb}(t)$ follows that,

$$\begin{aligned} Q(t+1) &= \text{f}_{\text{FARE}}(Q(t)) \succeq \text{f}_{\text{FARE}}(Q_{lb}(t)) = Q_{lb}(t+1), \\ &\quad Q_{lb}(t) \preceq Q(t) \preceq Q_{ub}(t),\ \forall\, t \in T. \end{aligned}$$

If (A,C) is detectable and (A_f, M_f) is stabilizable then, because of (Step 8) and (Step 9), $\lim Q_{lb}(t) = Q = \lim Q(t) = \lim Q_{ub}(t)$, where Q is the unique positive definite solution of the Algebraic Riccati equation under consideration. Step 8, Step 9, and Step 10 prove part (c).
(*Step 11.*) Suppose there exist two solutions, say $Q_1, Q_2 \in \mathbb{R}^{n\times n}$ of the algebraic Riccati equation

$$\begin{aligned} Q_1 &= f(Q_1),\ Q_2 = f(Q_2), \\ K_1 &= [AQ_1C^T + MN^T][CQ_1C^T + NN^T]^{-1}C, \end{aligned}$$

with $\text{spec}(A-K_1C) \subset \text{D}_o$, and similarly K_2 and $\text{spec}(A-K_2C) \subset \text{D}_o$.
From Proposition 22.2.7.(e) follows that

$$\begin{aligned} Q_2 - Q_1 &= f(Q_2) - f(Q_1) = (A-K_2C)[Q_2 - Q_1](A-K_2C)^T \\ &\quad + (K_2 - K_1)[CQ_1C^T + NN^T](K_2 - K_1)^T. \end{aligned}$$

From Theorem 22.1.2 follows that $Q_2 - Q_1 \succeq 0$ or $Q_2 \succeq Q_1$. By symmetry $Q_1 \succeq Q_2$, hence $Q_2 = Q_1$. This proves part (e).
(*Step 12.*) Let $K_1(Q) = FQC^T[CQC^T + NQ_vN^T]^{-1}$. Then

$$\begin{aligned} Q &= f(Q) = FQF^T + M_fM_f^T - FQC^T[CQC^TNQ_vN^T]^{-1}CQF^T, \\ &\quad \text{by Proposition 22.2.8.(c),} \\ &= [F - K_1(Q)C]Q[F - K_1(Q)C]^T + M_fM_f^T + K_1(Q)NN^TK_1(Q)^T, \qquad (22.19) \\ &\quad \text{by Proposition 22.2.8.(d).} \end{aligned}$$

Let $JJ^T = M_fM_f^T + K_1(Q)NN^TK_1(Q)^T$. It follows from Proposition 21.2.12 that (A_f, M_f) a reachable pair implies that $(F - K_1(Q)C, J)$ is a reachable pair. From (d) follows that $\text{spec}(A - K(Q)C) \subset \text{D}_o$. From Proposition 22.2.8.(b) follows that,

$$\text{spec}(F - K_1(Q)C) = \text{spec}(A - K(Q)C) \subset \text{D}_o.$$

This, $(F - K_1(Q)C, J)$ a reachable pair, the equation (22.19), and Theorem 22.1.2.(d) imply that $Q \succ 0$. This proves (f).
(*Step 13.*) By Proposition 22.2.7.(e)

$$\begin{aligned} Q_1 - Q_2 &= f(Q_1) - f(Q_2) \qquad (22.20) \\ &= A(Q_1)[Q_2 - Q_1]A(Q_1)^T + \\ &\quad + [K(Q_1) - K(Q_2)](CQ_2C^T + NQ_vN^T)[K(Q_1) - K(Q_2)]^T. \end{aligned}$$

From Theorem 22.2.2.(d) follows that $\text{spec}(A(Q_1)) \subset \text{D}_o$. The equation (22.20), the assumption $CQ_2C^T + NN^T \succ 0$, $\text{spec}(A(Q_1)) \subset \text{D}_o$, and Theorem 22.1.2 imply that $Q_1 - Q_2 \succeq 0$. This proves (g). □

Dependence of the Solution on Matrices

Consider the filter algebraic Riccati equation,

$$\begin{aligned} Q &= \text{f}_{\text{FARE}}(Q) = AQA^T + MM^T \\ &\quad - [AQC^T + MN^T][CQC^T + NN^T]^{-1}[AQC^T + MN^T]^T, \qquad (22.21) \\ &Q \in \mathbb{R}_{pds}^{n_x \times n_x}. \qquad (22.22) \end{aligned}$$

If $(A,\ C)$ is a detectable pair and $(A_f,\ M_f)$ is a stabilizable pair then it follows from Theorem 22.2.2.(a) and (b) that (22.21) and (22.22) have a unique solution. If, in addition, $(A_f,\ M_f)$ is a reachable pair then it follows from Theorem 22.2.2.(f) that $Q \succ 0$ hence $Q \in \mathbb{R}_{spds}^{n_x \times n_x}$. Define

$$\text{GStocSP}_{s,d} = \left\{ \begin{array}{l} \{n_y, n_x, n_y, A, C, M, N\} \in \text{GStocSP}| \\ (A_f,\ M_f) \text{ a stabilizable pair, } (A,\ C) \text{ a detectable pair} \end{array} \right\}.$$

Define the map,

$$FARE : \text{GStocSP}_{s,d} \to \mathbb{R}_{pds}^{n_x \times n_x},\ FARE(\{n_y, n_x, n_y, A, C, M, N\}) = Q,$$

where Q is the solution of the equations (22.21,22.22).

Theorem 22.2.11. *The map FARE is continuous, in fact an analytic function and thus infinitely differentiable.*

Computation of the Solution

The application of a Kalman filter requires the computation of the solution of the filter algebraic Riccati equation. According to Theorem 22.2.2.(c) if,

$$Q(t+1) = \mathrm{f}_{\mathrm{FARE}}(Q(t)),\ Q(0) = Q_{x_0},$$

then $\lim_{t \to \infty} = Q$ and $Q \in \mathbb{R}_{pds}^{n_x \times n_x}$. Therefore a procedure for the solution of the filter algebraic Riccati equation is the above recursion.

Experience with this procedure indicates that in certain cases convergence is slow. This is the case if the eigenvalues of $A(Q) = A - K(Q)C$ are close to the stability boundary, the unit circle. This issue is related to the existence of an exponential-asymptotic convergence rate, see Def. 22.2.5.

Experience with the computation procedure sketched above also indicates that after several steps of the recursion the numerical round off in a computer is such that $Q(t)$ is no longer positive-definite or no longer symmetric.

To counter these defects square root algorithms have been developed, [3]. The Hamiltonian approach to the algebraic Riccati equation is a useful alternative to the recursive method.

The Hamiltonian Approach

Below will be presented a non-iterative procedure for the computation of the solution of the filter algebraic Riccati equation. It is based on the Hamiltonian approach to the discrete-time algebraic Riccati equation.

Assumption 22.2.12 *Consider the parameters of a Gaussian system and define the matrices,*

$$\begin{aligned}
&\{n_y, n_x, n_y, A, C, M, N\} \in \mathrm{GStocSP},\ NN^T \succ 0,\\
A_f &= A - MN^T(NN^T)^{-1}C \in \mathbb{R}^{n_x \times n_x},\\
M_f M_f^T &= MM^T - MN^T(NN^T)^{-1}NM^T,\ M_f \in \mathbb{R}^{n_x \times n_x},\\
K &= \begin{pmatrix} A_f^T & 0 \\ -M_f M_f^T & I \end{pmatrix} \in R^{2n_x \times 2n_x},\\
L &= \begin{pmatrix} I & C^T(NN^T)^{-1}C \\ 0 & A_f \end{pmatrix} = \begin{pmatrix} I & L_{12} \\ 0 & A_f \end{pmatrix} \in R^{2n_x \times 2n_x}.
\end{aligned}$$

Then $(K,\ L)$ *is a called a* Hamiltonian pair *of matrices. for the filter algebraic Riccati equation. Assume that* $(A_f,\ M_f)$ *is a stabilizable pair and that* $(A,\ C)$ *a detectable pair.*

It will be proven below that the solutions of the algebraic Riccati equation may be associated with invariant subspaces of the pair of Hamiltonian matrices $(K,\ L)$.

Proposition 22.2.13. *Assumption 22.2.12 holds and assume that* $\det(K-\lambda L)$ *is not identically zero as function of* $\lambda \in \mathbb{C}$. *Then no generalized eigenvalue of* (K,L) *lies on the unit circle.*

Proof. Suppose that $\lambda \in \mathbb{C}$ is a generalized eigenvalue of (K,L) and let $u \in \mathbb{C}^{2n_x}\backslash\{0\}$ be the associated generalized eigenvector, such that λ lies on the unit circle, $|\lambda| = 1$. Then,

$$\begin{aligned} Ku &= \lambda Lu \ \Leftrightarrow \\ \begin{pmatrix} A_f^T & 0 \\ -M_f M_f^T & I \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} &= \lambda \begin{pmatrix} I & L_{12} \\ 0 & F \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \ \Leftrightarrow \\ A_f^T u_1 &= \lambda u_1 + \lambda L_{12} u_2 \qquad (22.23) \\ -M_f M_f^T u_1 + u_2 &= \lambda A_f u_2. \qquad (22.24) \end{aligned}$$

Premultiplying (22.23) by $\lambda^* u_2^H$ and adding this to the complex conjugate of (22.24) times u_1 yields,

$$\begin{aligned} \lambda^* u_2^H A_f^T u_1 &= |\lambda|^2 u_2^H u_1 + |\lambda|^2 u_2^H L_{12} u_2 + u_1^H M_f M_f^T u_1 - u_2^H u_1 + \lambda^* u_2^H A_f^T u_1 \\ &\Rightarrow \text{(using that } |\lambda| = 1) \ u_2^H L_{12} u_2 + u_1^H M_f M_f^T u_1 = 0. \end{aligned}$$

Note that $M_f M_f^T \succeq 0$ and that $L_{12} = C^T (NN^T)^{-1} C \succeq 0$. Thus $L_{12} u_2 = 0$ and $M_f u_1 = 0$, hence $Cu_2 = 0$ and $u_1^H M_f = 0$. From (22.23) and (22.24) then follows that

$$A_f^T u_1 = \lambda u_1,\ A_f u_2 = \frac{1}{\lambda} u_2.$$

Then $A_f u_2 = u_2/\lambda$ and $Cu_2 = 0$ imply that $Au_2 = u_2/\lambda$. This, $|\lambda| = 1$, and $Cu_2 = 0$ imply by detectability that $u_2 = 0$. Then $|\lambda| = 1$, $u_1^H F = \lambda u_1^H$, and $u_1^H M_f = 0$ imply by stabilizability of (A_f, M_f) that $u_2 = 0$. Thus $u = 0$. This is a contradiction of the existence of an eigenvalue on the unit circle. □

Proposition 22.2.14. *Assumptiont 22.2.12 holds and assume that* $\det(K-\lambda L)$ *is not identically zero as a function of* $\lambda \in \mathbb{C}$. *If* $\lambda \in C$, $\lambda \neq 0$, *is a generalized eigenvalue of* K,L, *then so is* $1/\lambda^*$ *with the same multiplicity.*

Proof. Let

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \in \mathbb{R}^{2n_x \times 2n_x};\ \text{then } KJK^T = \begin{pmatrix} 0 & A_f^T \\ -A_f & 0 \end{pmatrix} = LJL^T.$$

Consider the generalized eigenvalue problems $Kx = \lambda Lx$, $L^T y = \mu K^T y$. Then $\det(L^T - \mu K^T) = \det(L - \mu K)$. Thus $\lambda \in \mathbb{C}$ satisfies $\det(K - \lambda L) = 0$ if and only if

$1/\lambda^*$ satisfies $\det(L^T - (1/\lambda^*)K^T) = 0$. Assume that $\lambda \in \mathbb{C}$ is a generalized eigenvalue of $Kx = \lambda Lx$ of multiplicity k. Then $\mu = 1/\lambda^*$ is a generalized eigenvalue of $L^T y = \mu K^T y$ also with multiplicity k. It will be shown that $\mu = 1/\lambda^*$ is a generalized eigenvalue of $Kx = \lambda Lx$ of multiplicity k. Because μ is a generalized eigenvalue of multiplicity k, there exists a chain of generalized eigenvectors associated with μ, say $x_1, \ldots, x_k$, such that,

$$\begin{aligned}
L^T x_1 &= \mu K^T x_1, (L^T - \mu K^T)x_i = K^T x_{i-1}, \quad i = 2,3,\ldots,k; \\
z_i &= JL^T x_i,\ i \in \mathbb{Z}_k \ \Rightarrow \\
Kz_1 &= KJL^T x_1 = \mu KJK^T x_1 = \mu LJL^T x_1 = \mu Lz_1, \\
(K - \mu L)z_i &= KJL^T x_i - \mu LJL^T x_i = KJL^T x_i - \mu KJK^T x_i \\
&= KJ[L^T - K^T \mu]x_i = KJK^T x_{i-1} = LJL^T x_{i-1} = Lz_{i-1}.
\end{aligned}$$

□

Proposition 22.2.15. *Assumption 22.2.12 holds and assume that* $\det(K - \lambda L)$ *is not identically zero. If* $\lambda = 0$ *is a generalized eigenvalue of* K, L *of multiplicity* $k \in \mathbb{Z}_+$*, then there are exactly* $2n_x - k$ *finite generalized eigenvalues. The* k *missing generalized eigenvalues will be called* infinite eigenvalues, *because by Proposition 22.2.14 they are reciprocals of* 0.

Proof. It can be proven that K, L may be transformed to upper triangular form without affecting the determinant $\det(K - \lambda L)$. Then

$$\det(K - \lambda L) = \prod_{i=1}^{2n} (\alpha_i - \lambda \beta_i),$$

with α_i, β_i respectively the diagonal elements of (K, L). If $\lambda = 0$ is a generalized eigenvalue of order k, then, without loss of generality, $\alpha_1 = \ldots = \alpha_k = 0$. As shown in the proof of Proposition 22.2.14, $\mu_1 = 0$ is a generalized eigenvalue of multiplicity k of (L^T, K^T). Then

$$\det(L^T - \mu K^T) = \det(L - \mu K) = \prod_{i=1}^{2n} (\beta_i - \mu \alpha_i).$$

Thus $\beta_{2n_x - k + 1} = \ldots = \beta_{2n_x} = 0$. Note that $\alpha_i = \beta_i = 0$ is not possible, because this implies that $\det(K - \mu L) = 0$ which is excluded by assumption. Therefore

$$\det(K - \lambda L) = \prod_{i=1}^{k} (-\lambda \beta_i) \prod_{i=k+1}^{2n_x - k} (\alpha_i - \lambda \beta_i) \prod_{i=2n_x-(k+1)}^{2n_x} \alpha_i.$$

□

Theorem 22.2.16. *Assumption 22.2.12 holds. The statements (a) and (b) below are equivalent.*

(a)There exists a matrix $Q \in \mathbb{R}^{n_x \times n_x}$ *such that*

1. $Q = Q^T \succeq 0$;
2. $Q = \mathrm{f}_{\mathrm{FARE}}(Q)$;
3. $\mathrm{spec}(A(Q)) = \mathrm{spec}(A - K(Q)C) \subset \mathrm{D}_o$;

(b)There exists $U_1,\ U_2,\ \Lambda \in \mathbb{C}^{n_x \times n_x}$ *such that,*

1. Λ *is in Jordan canonical form containing the generalized eigenvalues of the tuple of matrices of equation (22.25) below;*
2. $$\begin{pmatrix} A_f^T & 0 \\ -M_f M_f^T & I \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = \begin{pmatrix} I & C^T (NN^T)^{-1} C \\ 0 & A_f \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} \Lambda, \tag{22.25}$$
   where U_1, U_2 *are the generalized eigenvectors associated with the generalized eigenvalues of the matrix* Λ*;*
3. U_1 *is nonsingular;*
4. $\mathrm{spec}(\Lambda) \subset \mathrm{D}_o$;

(c)There always exists a $\Lambda \in \mathbb{C}^{n_x \times n_x}$ *such that (b.1), (b.2), and (b.4) hold.*

(d)If $U_1,\ U_2,\ \Lambda \in \mathbb{C}^{n_x \times n_x}$ *satisfy the conditions of (b), then,*

$$Q = U_2 U_1^{-1},\ \ A(Q) = A - K(Q)C = U_1^{-T} \Lambda^T U_1^T, \tag{22.26}$$

and Q *satisfies the conditions of (a).*
Conversely, if $Q \in \mathbb{R}^{n_x \times n_x}$ *satisfies the conditions of (a) then let* $A(Q) = U_1^{-T} \Lambda^T U_1^T$ *be the decomposition of* $A(Q)$ *in Jordan canonical form, and let* $U_2 = QU_1$*. Then* U_1, U_2, Λ *satisfy the conditions of (b).*

Theorem 22.2.16 may be used to compute the solution Q of the algebraic Riccati equation of Kalman filtering. In a numerical algorithm one would not compute the Jordan canonical form but rather a basis for the stable subspace associated with the Hamiltonian pair (K, L) and the set of generalized eigenvalues of Λ.

Proof. (a) ⇒ (b). Consider a Jordan canonical form of $A(Q)^T$, $A(Q)^T = U_1 \Lambda U_1^{-1}$, with $U_1, \Lambda \in \mathbb{C}^{n \times n}$. Let $U_2 = QU_1$. Then, by $Q = Q^T$, $U_2 U_1^{-1} = Q = Q^T = U_1^{-T} U_2^T$. Then condition (a.3) implies that $\mathrm{spec}(\Lambda) = \mathrm{spec}(A(Q)) \subset \mathrm{D}_o$, and by definition of the canonical form, the orthogonal matrix U_1 is nonsingular. Thus (b.3) and (b.4) hold.

By Proposition 22.2.8.(b) and (c),

$$
\begin{aligned}
A(Q) &= A-[AQC^T+MN^T][CQC^T+NN^T]^{-1}C \\
&= A_f-A_fQC^T[CQC^T+NN^T]^{-1}C, \\
f(Q) &= A_fQA_f^T+M_fM_f^T-A_fQC^T[CQC^T+NN^T]^{-1}CQA_f^T. \text{ Then,} \\
A(Q)QA_f^T &= A_fQA_f^T-A_fQC^T[CQC^T+NN^T]^{-1}CQA_f^T=Q-M_fM_f^T, \\
&\quad \text{by Condition (a.2), hence,} \\
&\quad U_1^{-T}\Lambda U_1^T U_1^{-T}U_2^T A_f^T=U_1^{-T}U_2^T-M_fM_f^T, \\
&\quad \Rightarrow \text{ (by transposition) } A_fU_2\Lambda=U_2-M_fM_f^TU_1. \text{ Note that,} \\
A(Q)QC^T &= A_fQC^T-A_fQC^T[CQC^T+NN^T]^{-1}CQC^T, \\
&\quad A_fQC^T[CQC^T+NN^T]^{-1}NN^T \\
&\quad = A_fQC^T[CQC^T+NN^T]^{-1}[CQC^T+NN^T-CQC^T] \\
&\quad = A_fQC^T-A_fQC^T[CQC^T+NN^T]^{-1}CQC^T \\
&\quad = A(Q)QC^T, \text{ hence,} \\
A_f-A(Q) &= A_fQC^T[CQC^T+NN^T]^{-1}C, \text{ by equation,} \\
&= A(Q)QC^T(NN^T)^{-1}C, \text{ or,} \\
U_1^TA_f-\Lambda^TU_1^T &= \Lambda^TU_1^TU_1^{-T}U_2^TC^T(NN^T)^{-1}C, \text{ or,} \\
A_f^TU_1-U_1\Lambda &= C^T(NN^T)^{-1}CU_2\Lambda.
\end{aligned}
$$

Thus (b.2) holds.

(b) $\Rightarrow$ (a). Let $Q=U_1^{-T}U_2^T$. From (b.2) follows that,

$$
\Lambda^TU_2^TA_f^T=U_2^T-U_1^TM_fM_f^T \tag{22.27}
$$

$$
\begin{aligned}
U_1^TA_f-\Lambda^TU_1^T &= \Lambda^TU_2^TC^T(NN^T)^{-1}C \\
\Rightarrow \quad A_f-U_1^{-T}\Lambda^TU_1^T &= U_1^{-T}\Lambda^TU_2^TC(NN^T)^{-1}C.
\end{aligned} \tag{22.28}
$$

Then,

$$
\begin{aligned}
&\mathrm{f}_{\mathrm{FARE}}(Q)-Q=A_fQA_f^T-Q-A_fQC^T[CQC^T+NN^T]^{-1}CQA_f^T+M_fM_f^T \\
&\quad \text{by 22.2.8.(c) and } A_f,\ M_f, \\
&= A_fU_1^{-T}U_2^TA_f^T-U_1^{-T}U_2^T+M_fM_f^T \\
&\quad -A_fU_1^{-T}U_2^TC^T[CU_1^{-T}U_2^TC+NN^T]^{-1}CU_1^{-T}U_2^TA_f^T, \text{ by equation (22.27),} \\
&= A_fU_1^{-T}U_2^TA_f^T-U_1^{-T}\Lambda^TU_2^TA_f^T \\
&\quad -A_fU_1^{-T}U_2^TC^T[CU_1^{-T}U_2^TC+NN^T]^{-1}CU_1^{-T}U_2^TA_f^T \\
&= [A_f-U_1^{-T}\Lambda^TU_1^T]U_1^{-T}U_2^TA_f^T \\
&\quad -A_fU_1^{-T}U_2^TC^T[CU_1^{-T}U_2^TC+NN^T]^{-1}CU_1^{-T}U_2^TA_f^T \\
&= U_1^{-T}\Lambda^TU_2^TC^T(NN^T)^{-1}CU_1^{-T}U_2^TA_f^T \\
&\quad -A_fU_1^{-T}U_2^TC^T[CU_1^{-T}U_2^TC+NN^T]^{-1}CU_1^{-T}U_2^TA_f^T, \text{ by equation (22.28),} \\
&= [U_1^{-T}\Lambda^TU_2^TC^T(NN^T)^{-1}(CU_1^{-T}U_2^TC^T+NN^T)-A_fU_1^{-T}U_2^TC^T] \\
&\quad \times[CU_1^{-T}U_2^TC+NN^T]^{-1}CU_1^{-T}U_2^TA_f^T=0.
\end{aligned}
$$

By Proposition 22.2.8.(c) and $Q = f(Q)$,

$$A_f - A_f QC^T [CQC^T + NN^T]^{-1} C = A - K(Q)C = A(Q).$$

But,

$$\begin{aligned}
A(Q) &= A_f - A_f QC^T [CQC^T + NN^T]^{-1} C \\
&= A_f - A_f U_1^{-T} U_2^T C^T [CU_1^{-T} u_2^T C^T + NN^T]^{-1} C \\
&= U_1^{-T} \Lambda^T U_1^T + U_1^{-T} \Lambda^T U_2^T C^T (NN^T)^{-1} C - A_f U_1^{-T} U_2^T C^T [\ldots]^{-1} C \\
&= U_1^{-T} \Lambda^T U_1^T \\
&\quad + \left[U_1^{-T} \Lambda^T U_2^T C^T (NN^T)^{-1} [CU_1^{-T} U_2^T C^T + NN^T] - A_f U_1^{-T} U_2^T C^T\right] \times \\
&\quad \times [\ldots]^{-1} C \\
&= U_1^{-T} \Lambda^T U_1^T + \left[U_1^{-T} \Lambda^T U_2^T C^T (NN^T)^{-1} C U_1^{-T} U_2^T C^T \right. \\
&\quad \left. + U_1^{-T} \Lambda^T U_2^T C^T - A_f U_1^{-T} U_2^T C^T\right] [\ldots]^{-1} C \\
&= U_1^{-T} \Lambda^T U_1^T \ \Rightarrow \ \text{spec}(\Lambda) \subset \mathrm{D}_o.
\end{aligned}$$

By Proposition 22.2.8.(d)

$$\begin{aligned}
Q &= [A_f - K_1(Q)C]Q[A_f - K_1(Q)C]^T + M_f M_f^T + K_1(Q) NN^T K_1(Q)^T \\
&= [A_f - K_1(Q)C]Q[A_f - K_1(Q)C]^T + JJ^T. \qquad (22.29)
\end{aligned}$$

Note that,

$$A_f - K_1(Q)C = A(Q) = U_1^{-T} \Lambda^T U_1^T, \ \ \text{spec}(A_f - K_1(Q)C) = \text{spec}(\Lambda) \subset \mathrm{D}_o.$$

Because $NN^T \succ 0$, $JJ^T \succeq 0$. The Lyapunov equation (22.29) has a unique solution that satisfies $Q = Q^T \succeq 0$. Then $Q = U_1^{-T} U_2^T = U_2 U_1^{-1}$.
(c) By Proposition 22.2.14 the generalized eigenvalues of (K, L) are

$$\text{spec}(K, L) = \{0, \ldots, 0, \lambda_1, \ldots, \lambda_{n_x - k}, \frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_{n_x - k}}\}$$

where it is supposed that the generalized eigenvalue 0 has multiplicity k. According to Proposition 22.2.13 there are no generalized eigenvalues on the unit circle. Therefore there are exactly n generalized eigenvalues of (K, L) in D_o. Thus there exists a $\Lambda \in \mathbb{C}^{n_x \times n_x}$ such that (b.4) holds, and $U_1, U_2 \in \mathbb{C}^{n_x \times n_x}$ such that (b.2) holds. □

22.3 Algebraic Riccati Equation of Gaussian Stochastic Realization

The weak Gaussian stochastic realization problem discussed in Chapter 6 and dissipativeness of a linear system leads to the following algebraic Riccati equations of stochastic realization (RARED),

$$
\begin{aligned}
& Q = \mathrm{f}_{\mathrm{RARED}}(Q) = FQF^T + [G - FQH^T][J + J^T - HQH^T]^{-1}[G - FQH^T]^T, \\
& \quad \{n_y, n_x, n_y, F, G, H, J\} \in \mathrm{LSP}.
\end{aligned}
$$

See the Chapters 23 and 24 for the derivation of this equation and see Chapter 6 for the formulation of the weak Gaussian stochastic realization problem. This algebraic Riccati equations differs from the filter algebraic Riccati equation of Kalman filtering. The equation is denoted by the label RARED for realization algebraic Riccati equation (RARE), and the D stands for the dual version in accordance to the notation of Chapter 23 and Chapter 24. In general the equation $Q = \mathrm{f}_{\mathrm{RARED}}(Q)$ does not admit a unique solution. As is shown in Chapter 24 the set of solutions,

$$
\partial \mathbf{Q}_{\mathbf{lsdp},\mathbf{r},\mathbf{s}} = \{Q \in \mathbb{R}^{n_x \times n_x} | 2J - HQH^T \succ 0,\ D(Q) = 0\}
$$

has a maximal and a minimal element $Q^-, Q^+ \in \partial \mathbf{Q}_{\mathbf{lsdp},\mathbf{r},\mathbf{s}}$.

The main *difference* between the solution set of the algebraic Riccati equation of Gaussian stochastic realization compared with the solution set of the filter algebraic Riccati equation is that the first set has only positive-definite matrices while the second set has matrices which are either positive-definite, negative-definite, or which are partly positive-definite and partly negative definite.

Existence and Characterization of the Solution

The properties of the algebraic Riccati equation of stochastic realization are analogous to, yet different from, those of the algebraic Riccati equation of Kalman filtering.

Assumption 22.3.1 *Let* $\{n_y, n_x, n_y, F, G, H, J\} \in \mathrm{LSP}$. *Assume that* $J + J^T \succ 0$. *Define the sets and functions,*

$$
\begin{aligned}
\mathbf{Q}_{ns} &= \mathbb{Q}_{nsng,pds} = \{Q \in \mathbb{R}^{n_x \times n_x}_{pds} |\ J + J^T - G^T QG \ \ nonsingular\}, \\
\mathbf{Q}_{nsd} &= \mathbb{Q}_{nsng,pds} = \{Q \in \mathbb{R}^{n_x \times n_x}_{pds} |\ J + J^T - HQH^T \ \ nonsingular\}; \\
& \mathrm{f}_{\mathrm{RARE}} : \mathbf{Q}_{ns} \to \mathbb{R}^{n_x \times n_x},\ \mathrm{f}_{\mathrm{RARED}} : \mathbf{Q}_{nsd} \to \mathbb{R}^{n_x \times n_x}, \\
\mathrm{f}_{\mathrm{RARE}}(Q) &= F^T QF + [H^T - F^T QG][J + J^T - G^T QG]^{-1}[H^T - F^T QG]^T, \\
\mathrm{f}_{\mathrm{RARED}}(Q) &= FQF^T + [G - FQH^T][J + J^T - HQH^T]^{-1}[G - FQH^T]^T.
\end{aligned}
$$

The algebraic Riccati equation of Gaussian stochastic realization *(RARE) and the* dual algebraic Riccati equation of Gaussian stochastic realization *(RARED), are defined respectively by the equations,*

$$
Q = \mathrm{f}_{\mathrm{RARE}}(Q),\ \ Q = \mathrm{f}_{\mathrm{RARED}}(Q).
$$

Below attention is restricted to the equation with the function $\mathrm{f}_{\mathrm{RARED}}$.

Define the functions,

$$\begin{aligned}
K_{rd}(Q) &= [G-FQH^T][J+J^T-HQH^T]^{-1},\ K_{rd}:\mathbf{Q}_{nsd}\to\mathbb{R}^{n_x\times n_y},\\
V_r(Q) &= [J+J^T-G^TQG]^{-1}[H^T-F^TQG]^T,\\
&\ \mathrm{L_o}:\mathbb{R}^{n_x\times n_x}_{pds}\times\mathbb{R}^{n_x\times n_y}\to\mathbb{R}^{n_x\times n_x},\\
\mathrm{L_o}(Q,L) &= [F-LH]Q[F-LH]^T+G(J+J^T)G^T+\\
&\ -[G(J+J^T)^{-1}-L](J+J^T)[G(J+J^T)^{-1}-L]^T.
\end{aligned}$$

Theorem 22.3.2. *Assumption 22.3.1 holds.*

(a)If (F,G) is a stabilizable pair, (F_1,H) is a detectable pair, then the sequence of matrices converges to the indicated limit,

$$\begin{aligned}
Q(t+1) &= f(Q(t)), Q(0)=0,\ Q:T\to\mathbb{R}^{n_x\times n_x};\\
\lim_{t\to\infty}Q(t) &= Q^*\in\mathbb{R}^{n_x\times n_x},\\
Q^* &= \mathrm{f_{RARE}}(Q^*),\ \ Q^*\in\mathbb{R}^{n_x\times n_x}_{pds},\ \ \mathrm{spec}(F-K_{rd}(Q^*)H)\subset\mathrm{D}_o.
\end{aligned}$$

It is essential for the convergence that $Q(0)=0$.

(b)If (F_1,G) is a stabilizable pair, (F,H) is a detectable pair,

$$\begin{aligned}
&Q_d:T\to\mathbb{R}^{n_x\times n_x},\ Q_d(t+1)=f_d(Q_d(t)),\ Q_d(0)=0;\\
&\lim_{t\to\infty}Q_d(t)=Q_d,\ Q_d\in\mathbb{R}^{n_x\times n_x}_{spds}\ \Rightarrow\ Q^+=(Q_d)^{-1},\\
Q^+ &= f(Q^+),\ \mathbb{R}^{n_x\times n_x}_{pds},\ \mathrm{spec}(F-GV_{rd}(Q^+)\subset((\mathrm{D}^c)_o)^c\\
(\mathrm{D}^c)_o &= \{c\in\mathbb{C}||c|>1\}.
\end{aligned}$$

(c)The equation for the matrix Q with the conditions,

$$Q=\mathrm{f_{RARE}}(Q),\ Q\in\mathbb{R}^{n_x\times n_x}_{pds},\ \mathrm{spec}(F-K_{rd}(Q)H)\subset\mathrm{D}_o,$$

has at most one solution $Q\in\mathbb{R}^{n_x\times n_x}$. Denote the solution, which exists by (a), by $Q^-\in\mathbb{R}^{n_x\times n_x}_{pds}$. Thus Q^- satisfies,

$$Q^-=f(Q^-),\ \ Q^-\in\mathbb{R}^{n_x\times n_x}_{pds},\ \mathrm{spec}(F-K_{rd}(Q^-)H)\subset\mathrm{D}_o.$$

(d)The equation for the matrix Q with the conditions,

$$Q=\mathrm{f_{RARE}}(Q),\ Q\in\mathbb{R}^{n_x\times n_x}_{pds},\ \mathrm{spec}(F-GV_r(Q))\subset(\mathrm{D}^c)_o,$$

has at most one solution $Q\in\mathbb{R}^{n_x\times n_x}$. Denote then by $Q^+=Q^{-1}\in\mathbb{R}^{n_x\times n_x}_{pds}$. Thus,

$$Q^+=f(Q^+),\ \ Q^+\in\mathbb{R}^{n_x\times n_x}_{pds},\ \mathrm{spec}(F-GV_r(Q^+))\subset(\mathrm{D}^c)_o.$$

(e)If Q is a solution of the algebraic Riccati equation of stochastic realization then,

$$Q=\mathrm{f_{RARE}}(Q),\ Q\in\mathbb{R}^{n_x\times n_x}\ \Rightarrow\ Q^-\preceq Q\preceq Q^+.$$

The proof of this theorem is based on several propositions.

Proposition 22.3.3. *Assumption 22.3.1 holds.*

(a)

$$\begin{aligned} \mathrm{f}_{\mathrm{RARED}}(Q) = {} & (F - K_{rd}(Q)H)\ Q\ (F - K_{rd}(Q)H)^T + G(J+J^T)^{-1}G^T + \\ & -(G(J+J^T)^{-1} - K_{rd}(Q))(J+J^T)(G(J+J^T)^{-1} - K_{rd}(Q))^T. \end{aligned}$$

(b)$\mathrm{f}_{\mathrm{RARED}}(Q) = \mathrm{L}_{\mathrm{o}}(Q, K_{rd}(Q))$.
(c)$\mathrm{L}_{\mathrm{o}}(Q,L) - \mathrm{f}_{\mathrm{RARED}}(Q) = -[K_{rd}(Q) - L][J + J^T - HQH^T][K_{rd}(Q) - L]^T$.
(d)Assume that $J + J^T - HQ_1H^T \succ 0$.

$$\begin{aligned} & \mathrm{L}_{\mathrm{o}}(Q_2, L) - \mathrm{f}_{\mathrm{RARED}}(Q_1) = [F - LH](Q_2 - Q_1)[F - LH]^T + \\ & -[K_{rd}(Q_1) - L][J + J^T - HQ_1H^T][K_{rd}(Q_1) - L]^T. \end{aligned}$$

(e)Assume that $J + J^T - HQ_1H^T \succ 0$ *and that* $Q_1,\ Q_2 \in \mathbb{R}_{pds}^{n_x \times n_x}$.

$$\begin{aligned} & \mathrm{f}_{\mathrm{RARED}}(Q_2) - \mathrm{f}_{\mathrm{RARED}}(Q_1) \\ & = [F - K_{rd}(Q_2)H](Q_2 - Q_1)[F - K_{rd}(Q_2)H]^T + \\ & \quad -[K_{rd}(Q_1) - K_{rd}(Q_2)][J + J^T - HQ_1H^T][K_{rd}(Q_1) - K_{rd}(Q_2)]^T. \end{aligned}$$

(f) Assume that $J + J^T - HQ_1H^T \succ 0$ *and that* $Q_1,\ Q_2 \in \mathbb{R}_{pds}^{n_x \times n_x}$. *If* $Q_2 \preceq Q_1$ *then* $\mathrm{f}_{\mathrm{RARED}}(Q_2) \preceq \mathrm{f}_{\mathrm{RARED}}(Q_1)$.

Proof.

$$\begin{aligned} K_{rd}(Q) &= [G - FQH^T][J + J^T - HQH^T]^{-1}; \\ G - FQH^T &= K_{rd}(Q)(J + J^T - HQH^T), \\ FQH^T &= G - K_{rd}(Q)(J + J^T - HQH^T), \\ FQH^T K_{rd}(Q)^T &= GK_{rd}(Q)^T - K_{rd}(Q)(J + J^T - HQH^T)K_{rd}(Q)^T. \end{aligned}$$

(a)

$$\begin{aligned} \mathrm{f}_{\mathrm{RARED}}(Q) &= F^T QF + [G - FQH^T][J + J^T - HQH^T]^{-1}[G - FQH^T]^T \\ &= [F - K_{rd}(Q)H]Q[F - K_{rd}(Q)H]^T \\ &\quad + FQH^T K_{rd}(Q)^T + K_{rd}(Q)F^T QH - K_{rd}(Q)HQH^T K_{rd}(Q) \\ &\quad + K_{rd}(Q)[J + J^T - HQH^T]K_{rd}(Q) \\ &= [F - K_{rd}(Q)H]Q[F - K_{rd}(Q)H]^T \\ &\quad + GK_{rd}(Q)^T - K_{rd}(Q)[J + J^T - HQH^T]K_{rd}(Q)^T + \\ &\quad + K_{rd}(Q)G^T - K_{rd}(Q)[J + J^T - HQH^T]K_{rd}(Q)^T + \\ &\quad + K_{rd}(Q)[J + J^T - HQH^T]K_{rd}(Q) - K_{rd}(Q)HQH^T K_{rd}(Q) \\ &\quad \text{using the formula above (a) in the proof,} \\ &= [F - K_{rd}(Q)H]Q[F - K_{rd}(Q)H]^T \\ &\quad + GK_{rd}(Q)^T + K_{rd}(Q)G^T - K_{rd}(Q)[J + J^T]K_{rd}(Q)^T \\ &= [F - K_{rd}(Q)H]Q[F - K_{rd}(Q)H]^T + G(J + J^T)^{-1}G^T + \\ &\quad -(G(J + J^T)^{-1} - K_{rd}(Q))(J + J^T)(G(J + J^T)^{-1} - K_{rd}(Q))^T. \end{aligned}$$

(b) This follows by definition of $\mathrm{L_o}(Q,L)$.

(c)

$$\begin{aligned}
&(F - K_{rd}(Q)H)QH^T - G + K_{rd}(Q)(J + J^T) \\
&= FQH^T - G + K_{rd}(Q)(J + J^T - HQH^T) \\
&= FQH^T - G + (G - FQH^T) = 0; \\
&\mathrm{L_o}(Q,L) - \mathrm{f_{RARED}}(Q) = \mathrm{L_o}(Q,L) - \mathrm{L_o}(Q,K_{rd}(Q)) \ \text{ by (b),} \\
&= [F - LH^T]Q[F - LH^T]^T + \\
&\quad +G(J + J^T)^{-1}G^T - (G(J + J^T)^{-1} - L)(J + J^T)(G(J + J^T)^{-1} - L)^T + \\
&\quad -[F - K_{rd}(Q)H^T]Q[F - K_{rd}(Q)H^T]^T - G(J + J^T)^{-1}G^T + \\
&\quad +(G(J + J^T)^{-1} - K_{rd}(Q))(J + J^T)(G(J + J^T)^{-1} - K_{rd}(Q)) \\
&= [F - K_{rd}(Q)H + (K_{rd}(Q) - L)H]Q[\ldots]^T + \\
&\quad +(G(J + J^T)^{-1} - K_{rd}(Q) + (K_{rd}(Q) - L))(J + J^T)(\ldots)^T \\
&\quad -[F - K_{rd}(Q)H]Q[F - K_{rd}(Q)H]^T + \\
&\quad +(G(J + J^T)^{-1} - K_{rd}(Q))(J + J^T)(G(J + J^T)^{-1} - K_{rd}(Q)) \\
&= (K_{rd}(Q) - L)HQH^T(K_{rd}(Q) - L)^T + \\
&\quad -(K_{rd}(Q) - L)(J + J^T)(K_{rd}(Q) - L) + \\
&\quad +(F - K_{rd}(Q)H)QH^T(K_{rd}(Q) - L)^T + \\
&\quad -(G(J + J^T)^{-1} - K_{rd}(Q))(J + J^T)(K_{rd}(Q) - L)^T + \\
&\quad +\text{two nonsymmetric terms} \\
&= -(K_{rd}(Q) - L)(J + J^T - HQH^T)(K_{rd}(Q) - L)^T.
\end{aligned}$$

(d)

$$\begin{aligned}
&\mathrm{L_o}(Q_2,L) - \mathrm{f_{RARED}}(Q_1) = \mathrm{L_o}(Q_2,L) - \mathrm{L_o}(Q_1,L) + \mathrm{L_o}(Q_1,L) - \mathrm{L_o}(Q_1,K_{rd}(Q_1)) \\
&= [F - LH](Q_2 - Q_1)[F - LH]^T \\
&\quad -[K_{rd}(Q_1) - L][J + J^T - H^TQ_1H][K_{rd}(Q_1) - L] \text{ by (c).}
\end{aligned}$$

(e)

$$\begin{aligned}
&\mathrm{f_{RARED}}(Q_2) - \mathrm{f_{RARED}}(Q_1) = \mathrm{L_o}(Q_2,K_{rd}(Q_2)) - \mathrm{f_{RARED}}(Q_1) \\
&= [F - K_{rd}(Q_2)H](Q_2 - Q_1)[F - K_{rd}(Q_2)H]^T + \\
&\quad -[K_{rd}(Q_1) - K_{rd}(Q_2)][J + J^T - HQ_1H^T][K_{rd}(Q_1) - K_{rd}(Q_2)]^T \ \text{ by (d).}
\end{aligned}$$

(f) This follows from (e). □

Proposition 22.3.4. *Assumption 22.3.1 holds. Define the matrices,*

$$\begin{aligned}
Q_{GJ} &= G(J + J^T)^{-1}G^T \in \mathbb{R}_{pds}^{n_x \times n_x}, \ Q_{JH} = H^T(J + J^T)^{-1}H \in \mathbb{R}_{pds}^{n_x \times n_x}, \\
F_1 &= F - G(J + J^T)^{-1}H \in \mathbb{R}^{n_x \times n_x}, \\
K_{rd1}(Q) &= K_{rd}(Q) - G(J + J^T)^{-1} \in \mathbb{R}^{n_x \times n_y}, \\
F(Q) &= F - K_{rd}(Q)H.
\end{aligned}$$

(a)

$$\begin{aligned} K_{rd1}(Q) &= -F_1 Q H^T (J + J^T - HQH^T)^{-1} = F(Q)QH(J+J^T)^{-1}, \\ F(Q) &= F - K_{rd}(Q)H = F_1 - K_{rd1}(Q)H = F_1 + F(Q)QQ_{JH}. \end{aligned}$$

(b)

$$\begin{aligned} \mathrm{f}_{\mathrm{RARED}}(Q) &= F_1 Q F_1^T + Q_{GJ} + F_1 Q H^T (J + J^T - HQH^T)^{-1} HQF_1^T \\ &= F(Q)QF(Q)^T + Q_{GJ} - F(Q)QQ_{JH}QF(Q)^T. \end{aligned}$$

Proof. (a)

$$\begin{aligned} & K_{rd1}(Q)[J + J^T - HQH^T] \\ &= [K_{rd}(Q) - G(J+J^T)^{-1}][J + J^T - HQH^T]^{-1} \\ &= [G - FQH^T] - G(J+J^T)^{-1}[J + J^T - HQH^T] \\ &= -[F - G(J+J^T)^{-1}H]QH^T = -F_1 QH^T; \\ K_{rd1}(Q) &= -F_1 QH^T [J + J^T - HQH^T]^{-1}, \\ & F_1 - K_{rd1}(Q)H \\ &= F - G(J+J^T)^{-1}H - [K_{rd}(Q) - G(J+J^T)^{-1}]H \\ &= F - K_{rd}(Q)H = F(Q); \\ & F(Q)QH^T(J+J^T)^{-1} \\ &= F_1 QH^T (J+J^T)^{-1} - K_{rd1}(Q)HQH^T(J+J^T)^{-1} \\ &= F_1 QH^T (J+J^T)^{-1} + F_1 QH^T [J + J^T - HQH^T]^{-1} HQH^T (J+J^T)^{-1} \\ &= F_1 QH^T (J+J^T)^{-1} + \\ & \quad + F_1 QH^T [\ldots]^{-1} [-(J + J^T - HQH^T) + (J+J^T)](J+J^T)^{-1} \\ &= F_1 QH^T (J+J^T)^{-1} - F_1 QH (J+J^T)^{-1} + F_1 QH^T [J + J^T - HQH^T]^{-1} \\ &= -K_{rd1}(Q) \ \Rightarrow \\ F(Q) &= F_1 - K_{rd1}(Q)H = F_1 + F(Q)QQ_{JH}. \end{aligned}$$

(b) From Proposition 22.3.3.(a) follows that,

$$\begin{aligned}
&\mathrm{f}_{\mathrm{RARED}}(Q)\\
&= (F - K_{rd}(Q)H)Q(F - K_{rd}(Q)H)^T + Q_{GJ} +\\
&\quad -[G(J+J^T)^{-1} - K_{rd}(Q)](J+J^T)[G(J+J^T)^{-1} - K_{rd}(Q)]^T\\
&\quad \text{by Proposition 22.3.3.(a).,}\\
&= (F_1 - K_{rd1}(Q)H)Q(F_1 - K_{rd1}(Q)H)^T + Q_{GJ} - K_{rd1}(Q)(J+J^T)K_{rd1}(Q))^T,\\
&\quad \text{by the formulas for } F_1 \text{ and } K_{rd1} \text{ of (a),}\\
&= F_1QF_1^T + Q_{GJ} - K_{rd1}(Q)HQF_1^T - F_1QH^TK_{rd1}(Q)^T\\
&\quad +K_{rd1}(Q)HQH^TK_{rd1}(Q)^T - K_{rd1}(Q)(J+J^T)K_{rd1}(Q)^T \text{ by (a),}\\
&= F_1QF_1^T + Q_{GJ} + F_1QH^T[J+J^T - HQH^T]^{-1}HQF_1^T,\\
&\quad \text{by } K_{rd1}(Q) = -F_1QH^T(J+J^T - HQH^T)^{-1};\\
&= (F_1 - K_{rd1}(Q)H)Q(F_1 - K_{rd1}(Q)H)^T + Q_{GJ} - K_{rd1}(Q)(J+J^T)K_{rd1}(Q))^T,\\
&= F(Q)QF(Q)^T + Q_{GJ} - F(Q)QH(J+J^T)^{-1}(J+J^T)(J+J^T)^{-1}H^TQF(Q)^T\\
&= F(Q)QF(Q)^T + Q_{GJ} - F(Q)QQ_{JH}QF(Q)^T.
\end{aligned}$$

□

Proposition 22.3.5. *Assume,*

$$\begin{aligned}
&0 \prec J + J^T - HQ_2H^T,\\
&Q_1 = \mathrm{f}_{\mathrm{RARED}}(Q_1),\ Q_2 = \mathrm{f}_{\mathrm{RARED}}(Q_2).
\end{aligned}$$

(a)If $\mathrm{spec}(F - K_{rd}(Q_1)H^T) \subset \mathrm{D}_o$ *then* $Q_2 \preceq Q_1$.
(b)If $\mathrm{spec}(F - K_{rd}(Q_2)H^T) \subset (\mathrm{D}^c)_o$ *then* $Q_2 \succeq Q_1$.

Proof. (a)

$$\begin{aligned}
Q_1 - Q_2 &= \mathrm{f}_{\mathrm{RARED}}(Q_1) - \mathrm{f}_{\mathrm{RARED}}(Q_2)\\
&= [F - GK(Q_1)]^T(Q_1 - Q_2)[F - GK(Q_1)]\\
&\quad -[K_{rd}(Q_2) - K_{rd}(Q_1)]^T[J + J^T - HQ_2H^T]^{-1}[K_{rd}(Q_2) - K_{rd}(Q_1)],
\end{aligned}$$

by Propostion 22.3.3.(e), hence

$$\begin{aligned}
Q_2 - Q_1 &= [F - GK(Q_1)]^T(Q_2 - Q_1)[F - GK(Q_1)]\\
&\quad +[K_{rd}(Q_2) - K_{rd}(Q_1)]^T[J + J^T - HQ_2H^T]^{-1}[K_{rd}(Q_2) - K_{rd}(Q_1)].
\end{aligned}$$

From Theorem 22.1.2, $\mathrm{spec}(F - K_{rd}(Q_1)H) \subset \mathrm{D}_o$, and from $J + J^T - HQ_2H^T \succ 0$ follows that $Q_2 - Q_1 \succeq 0$.
(b)

$$\begin{aligned}
Q_2 - Q_1 &= \mathrm{f}_{\mathrm{RARED}}(Q_2) - \mathrm{f}_{\mathrm{RARED}}(Q_1)\\
&= [F - GK_{rd}(Q_2)]^T(Q_2 - Q_1)[F - GK_{rd}(Q_2)]\\
&\quad -[K_{rd}(Q_1) - K_{rd}(Q_2)]^T[J + J^T - G^TQ_1G]^{-1}[K_{rd}(Q_1) - K_{rd}(Q_2)].
\end{aligned}$$

From the anti symmetric version of the discrete Lyapunov equation, from $\mathrm{spec}(F - K_{rd}(Q_2)H) \subset (\mathrm{D}^c)_o$, and from $J + J^T - HQ_1H^T \succ 0$, then follows that $Q_2 - Q_1 \succeq 0$. □

Proof. Proof of Theorem 22.3.2. (a) This follows along the lines of the proof of Theorem 22.2.2 for the filter algebraic Riccati equation. The existence also follows from Theorem 22.3.11 of the Hamiltonian approach to the Riccati equation.
(b) This is similar to (a) by duality.
(c) Suppose that $Q_1, Q_2 \in \mathbf{Q}_{ns}$ are such that,

$$\begin{aligned} Q_1 &= f(Q_1),\ Q_1^T \in \mathbb{R}_{pds}^{n_x \times n_x},\ \text{spec}(F - K_{rd}(Q_1)H) \subset \text{D}_o, \\ Q_2 &= f(Q_2),\ Q_2 \in \mathbb{R}_{pds}^{n_x \times n_x},\ \text{spec}(F - K_{rd}(Q_2)H) \subset \text{D}_o. \end{aligned}$$

From Proposition 22.3.5.(a) and (b) then follows that $Q_1 \preceq Q_2$ and that $Q_2 \succeq Q_1$, hence $Q_2 = Q_1$.
(d) This follows from Proposition 22.3.5.(b) as (c) from Proposition 22.3.5.(a).
(e) Let $Q \in \mathbf{Q}_{ns}$ be such that $Q = f(Q), Q = Q^T$. By (c) $\text{spec}(F - K_{rd}(Q^-))H)) \subset \text{D}_o$, $Q^- = \text{f}_{\text{RARED}}(Q^-)$. From Proposition 22.3.5.(b) then follows that $Q^- \preceq Q$. Similarly follows from Proposition 22.3.5.(a) that $Q \preceq Q^+$. □

An Iterative Algorithm for the Algebraic Riccati Equation of Stochastic Realization

Procedure 22.3.6 *An iterative algorithm for $Q^+ \in \partial \mathbf{Q}_{\mathbf{lsdp,r,s}}$.*
Declarations: $n_x, n_y \in \mathbb{Z}_+$, $F \in \mathbb{R}^{n_x \times n_x}$, $G \in \mathbb{R}^{n_x \times n_y}$, $H \in \mathbb{R}^{n_y \times n_x}$, $J \in \mathbb{R}^{n_y \times n_y}$ with $J = J^T$, Q, $Q^+ \in \mathbb{R}^{n_x \times n_x}$.
Input n_x, n_y, (F, G, H, J).
Output Q^+.

```
1. Set $Q(0) := 0$.
2. For $t = 1$ step $1$ to $\infty$ do
       $Q(t+1) = FQ(t)F^T +$
           $+[G - FQ(t)H^T][J + J^T - HQ(t)H^T]^{-1}[G - FQ(t)H^T]^T.$
3. Output $Q^- := Q(\infty)$.
```

Procedure 22.3.7 *An iterative algorithm for $Q^- \in \partial \mathbf{Q}_{\mathbf{lsp,r,s}}$.*
Declarations: $n_x, n_y \in \mathbb{Z}_+$, $F \in \mathbb{R}^{n_x \times n_x}$, $G \in \mathbb{R}^{n_x \times n_y}$, $H \in \mathbb{R}^{n_y \times n_x}$, $J \in \mathbb{R}^{n_y \times n_y}$ with $J = J^T$, Q, $Q^- \in \mathbb{R}^{n_x \times n_x}$.
Input n_x, n_y, (F, G, H, J).
Output Q^-.

```
1. Set $Q(0) = 0$.
2. For $t = 1$ step $1$ to $\infty$ do
       $Q(t+1) = F^T Q(t)F +$
           $+[H^T - F^T Q(t)G][J + J^T - G^T Q(t)G]^{-1}[H^T - F^T Q(t)G]^T.$
3. Output $Q^+ := Q(\infty)^{-1}$.
```

Two major questions for the above algorithms are: (1) What is the domain of attraction of Q^+, Q^-?; (2) What is the convergence speed of these algorithms?

The iterative procedures presented in Procedure 22.3.6 and in Procedure 22.3.7 clearly have disadvantages. Reports in the literature mention that in certain cases the algorithm produces estimates that converge slowly. Therefore there is an interest in a non-iterative algorithm.

The Hamiltonian Approach to the Algebraic Riccati Equation of Stochastic Realization

Below a non-iterative algorithm will be presented that is based on the Hamiltonian approach to the discrete-time algebraic Riccati equation.

Definition 22.3.8. Consider $\text{LSP} = \{n_y, n_x, n_y, F, G, H, J\} \in \text{LSP}_{min}$ with $0 \prec J + J^T$. Define the matrices,

$$\begin{aligned}
Q_{GJ} &= G(J+J^T)^{-1}G^T,\ Q_{JH} = H^T(J+J^T)^{-1}H, \\
F_1 &= F - G(J+J^T)^{-1}, \\
&\quad Q \in \mathbb{R}^{n_x \times n_x}_{pds} \text{ satisfies } 0 \prec (J + J^T - HQH^T), \\
K_{rd}(Q) &= [G - FQH^T](J + J^T - HQH^T)^{-1}, \\
F(Q) &= F - K_{rd}(Q)H = F - (G - FQH^T)[J + J^T - HQH^T]^{-1}H. \\
K_{rd1}(G) &= K_{rd}(Q) - G(J+J^T)^{-1}, \\
A_B &= \left(\begin{array}{c|c} F_1^T & 0 \\ \hline -Q_{GJ} & I_{n_x} \end{array}\right),\ A_F = \left(\begin{array}{c|c} I_{n_x} & -Q_{JH} \\ \hline 0 & F_1 \end{array}\right) \in \mathbb{R}^{2n_x \times 2n_x}.
\end{aligned}$$

It will be shown below that the solutions of the algebraic Riccati equation may be associated with invariant subspaces of the tuple (A_F, A_B).

Below use is made of the concept of generalized eigenvalues of (A_F, A_B) and of generalized eigenvectors. The motivation for the use of the generalized eigenvalues follows. It will be proven below that if V_1, $V_2, \Lambda \in \mathbb{C}^{n_x \times n_x}$ satisfy,

$$A_B \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} = A_F \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} \Lambda,$$

that then $Q = V_2 V_1^{-1}$ is a solution of the equation $Q = \text{f}_{\text{RARED}}(Q)$.

Proposition 22.3.9. *Consider the objects of Def. 22.3.8. Assume that* $\det(A_B - \lambda A_F)$ *as function of* λ *is not identically zero. If* $\lambda \in \mathbb{C}$, $\lambda \neq 0$, *is a generalized eigenvalue of* (A_F, A_B), *then so is* $1/\lambda^*$ *with the same multiplicity.*

Proof. This is identical to that of Proposition 22.2.14. □

Proposition 22.3.10. *Consider the objects of Def. 22.3.8. Assume that* $\det(A_B - \lambda A_F)$ *is not identically zero as a function of* λ. *If* $\lambda = 0$ *is a generalized eigenvalue of* (A_F, A_B) *of multiplicity k, then there are exactly* $2n_x - k$ *finite generalized* *eigenvalues. The k missing generalized eigenvalues will be called* infinite eigenvalues, *because by Proposition 22.3.9 they are reciprocals of* 0.

Proof. This is identical to the proof of Proposition 22.2.15. □

Theorem 22.3.11. *Consider* $lsp = \{n_y, n_x, n_y, F, G, H, J\} \in \mathrm{LSP}_{min}$ *and assume that* $0 \prec J + J^T$*. Recall the definition of the matrices of Def. 22.3.8.*

The statements (a) and (b) below are equivalent.

(a)There exists a $Q \in \mathbb{R}^{n_x \times n_x}$ *such that*

1. $Q \in \mathbb{R}^{n_x \times n_x}_{spds}$;
2. $0 \prec J + J^T - HQH^T$;
3. $\mathrm{spec}(F_{rd} - K_{rd}(Q)H) \subset \mathrm{D}_o$ *is such that:*
   (1) if $\lambda_1, \lambda_2 \in \mathrm{spec}(F(Q))$ *then* $\lambda_1 \lambda_2 \neq 1$*;*
   (2) if $\lambda \in \mathrm{spec}(F_{rd} - K_{rd}(Q)H)$ *then* $\lambda^* \in \mathrm{spec}(F_{rd} - K_{rd}(Q)H)$.
4. $Q = \mathrm{f}_{\mathrm{RARED}}(Q)$.

(b)There exists $V_1,\ V_2,\ \Lambda \in \mathbb{C}^{n_x \times n_x}$*, with*

1. Λ *in Jordan canonical form containing the generalized eigenvalues of (22.30) below, such that: (1) if* $\lambda_1, \lambda_2 \in \mathrm{spec}(\Lambda)$ *then* $\lambda_1 \lambda_2 \neq 1$*;*
   (2) if $\lambda \in \mathrm{spec}(\Lambda)$ *then* $\lambda^* \in \mathrm{spec}(\Lambda)$.
2.

$$A_B V = A_F V \Lambda, \tag{22.30}$$

$$\Leftrightarrow \begin{pmatrix} \begin{array}{c|c} F_1^T & 0 \\ \hline -Q_{GJ} & I_{n_x} \end{array} \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} = \begin{pmatrix} \begin{array}{c|c} I & -Q_{JH} \\ \hline 0 & F_1 \end{array} \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} \Lambda, \tag{22.31}$$

   where V_1, V_2 *are the two components of the matrix V of the generalized eigenvectors associated with the generalized eigenvalues of the matrix* Λ*;*
3. V_1 *is nonsingular;*
4. $0 \prec J + J^T - HV_2V_1^{-1}H^T$.

(c)If $V_1,\ V_2,\ \Lambda \in \mathbb{C}^{n_x \times n_x}$ *satisfy the conditions of (b), then* $Q = V_2 V_1^{-1}$ *satisfies the conditions of (a) and* $F - K_{rd}(Q)H = V_1 \Lambda V_1^{-1}$.
Conversely, if $Q \in \mathbb{R}^{n_x \times n_x}$ *satisfies the conditions of (a) then let* $F - K_{rd}(Q)H = V_1 \Lambda V_1^{-1}$ *be the decomposition of* $F_{rd} - K_{rd}(Q)H$ *in Jordan canonical form, and let* $V_2 = QV_1$*. Then* V_1, V_2, Λ *satisfy the conditions of (b).*

(d)If there are no generalized eigenvalues of (22.30) on the unit circle then there always exists a selection Λ *of generalized eigenvalues such that* $\mathrm{spec}(\Lambda) \subset \mathrm{D}_o$.

Proof. Note that equation (22.30) holds if and only if,

$$F_1^T V_1 = V_1 \Lambda - Q_{JH} V_2 \Lambda, \quad -Q_{GJ} V_1 + V_2 = F_1 V_2 \Lambda.$$

(1) Note that by 22.3.4.(b),

$$\begin{aligned}
K_{rd1}(Q) &= -F_1 Q H^T [J + J^T - H Q H^T]^{-1}, \\
F(Q) Q H^T (J + J^T)^{-1} &= -K_{rd1}(Q), \\
F(Q) &= F_1 - K_{rd1}(Q) H = F_1 + F(Q) Q Q_{JH}, \\
\mathrm{f}_{\mathrm{RARED}}(Q) &= F_1 Q F_1^T + Q_{GJ} + F_1 Q H^T [J + J^T - H Q H^T]^{-1} H Q F_1^T \\
&= F(Q) Q F(Q)^T + Q_{GJ} - F(Q) Q Q_{JH} Q F(Q)^T.
\end{aligned}$$

(2) (a) ⇒ (b). There exists a nonsingular matrix $V_1 \in \mathbb{C}^{n_x \times n_x}$ and a matrix $\Lambda \in \mathbb{C}^{n_x \times n_x}$ in Jordan canonical form such that $F(Q)^T = V_1 \Lambda V_1^{-1}$. Because $\mathrm{spec}(F(Q)) = \mathrm{spec}(\Lambda)$, Condition (a.3) implies (b.1). Let $V_2 = Q V_1$, hence $Q = V_2 V_1^{-1}$. By Step 2, $F(Q) = F_1 + F(Q) Q Q_{JH}$. Then,

$$\begin{aligned}
F_1^T &= F(Q)^T - Q_{JH} Q F(Q)^T = V_1 \Lambda V_1^{-1} - Q_{JH} V_2 V_1^{-1} V_1 \Lambda V_1^{-1} \Rightarrow \\
F_1^T V_1 &= V_1 \Lambda - Q_{JH} V_2 \Lambda.
\end{aligned}$$

If $Q = \mathrm{f}_{\mathrm{RARED}}(Q)$, then by (1) above,

$$\begin{aligned}
V_2 V_1^{-1} &= Q = F(Q) Q F(Q)^T + Q_{GJ} - F(Q) Q Q_{JH} Q F(Q)^T \\
&= Q_{GJ} + F_1 Q F(Q)^T = Q_{GJ} + F_1 V_2 V_1^{-1} V_1 \Lambda V_1^{-1}, \\
-Q_{GJ} V_1 + V_2 &= F_1 V_2 \Lambda.
\end{aligned}$$

Thus $(V_1,\, V_2,\, \Lambda)$ satisfy equation (22.30).
(3) (b) ⇒ (a). Condition (b.2) implies that,

$$F_1^T V_1 = V_1 \Lambda - Q_{JH} V_2 \Lambda, \quad -Q_{GJ} V_1 + V_2 = F_1 V_2 \Lambda.$$

By Condition (b.3), V_1 is nonsingular. Define $Q = V_2 V_1^{-1}$ and $F_3^T = V_1 \Lambda V_1^{-1}$. Note that it is not yet proven that the matrix Q is symmetric. It will be proven that $\mathrm{f}_{\mathrm{RARED}}(Q) = Q$.

Then,

$$F_1^T = F_3^T - Q_{JH} V_2 V_1^{-1} V_1 \Lambda V_1^{-1} = F_3^T - Q_{JH} Q F_3^T.$$

Define $K(Q) = -F_1 Q H^T [J + J^T - H Q H^T]^{-1}$. Then,

$$\begin{aligned}
& K(Q)[J + J^T - H Q H^T] = -F_1 Q H^T = -F_3 Q H^T + F_3 Q Q_{JH} Q H^T \\
&= -F_3 Q H^T + F_3 Q H (J + J^T)^{-1} H Q H^T \\
&= -F_3 Q H^T - F_3 Q H^T (J + J^T)^{-1} [J + J^T - H Q H^T] + \\
&\quad + F_3 Q H^T (J + J^T)^{-1} (J + J^T) \\
&= -F_3 Q H^T (J + J^T)^{-1} [J + J^T - H Q H^T] \Rightarrow \\
K(Q) H &= -F_3 Q H^T (J + J^T)^{-1} H = -F_3 Q Q_{JH}, \\
F_1 &= F_3 - F_3 Q Q_{JH} = F_3 + K(Q) H \Rightarrow F_3 = F_1 - K(Q) H.
\end{aligned}$$

Then further,

$$\begin{aligned}
& V_2 - Q_{GJ}V_1 = F_1 V_2 \Lambda \\
& \Leftrightarrow Q - Q_{GJ} = (V_2 - Q_{GJ}V_1)V_1^{-1} = F_1 V_2 V_1^{-1} V_1 \Lambda V_1^{-1} = F_1 Q F_3^T \\
& = [F_3 - F_3 Q Q_{JH}] Q F_3^T = F_3 Q F_3^T - F_3 Q Q_{JH} Q F_3^T, \\
& \Leftrightarrow Q = F_3 Q F_3^T + Q_{GJ} - F_3 Q Q_{JH} Q F_3^T \\
& = F_1 Q F_3^T + Q_{GJ} = F_1 Q (F_1 - K(Q)H)^T + Q_{GJ} \\
& = F_1 Q F_1^T + Q_{GJ} + F_1 Q H^T (J + J^T - H Q^T H^T)^{-1} H Q^T F_1^T . = \mathrm{f}_{\mathrm{RARED}}(Q),
\end{aligned}$$

where the latter equation follows from Proposition 22.3.4. Thus Q satisfies the equation $Q = \mathrm{f}_{\mathrm{RARED}}(Q)$. It follows from the latter equation and the assumptions Q is symmetric, that $Q \in \mathbb{R}_{pds}^{n_x \times n_x}$, and that the solution is unique because $\mathrm{spec}(F_3) = \mathrm{spec}(\Lambda) \subset \mathrm{D}_o$. □

Theorem 22.3.11 may be used to compute the minimal solution Q^- of the algebraic Riccati equation of stochastic realization. In a numerical procedure one would not compute the Jordan canonical form $Q_{v,1}, Q_{v,2}, \Lambda$ but rather a basis for the stable subspace associated with the Hamiltonian pair (K, L).

The possibility exists that the generalized eigenvalue problem (22.30) has generalized eigenvalues on the unit circle. Examples have been reported. In such a case there does not exist a Λ such that $\mathrm{spec}(\Lambda) \subset \mathrm{D}_o$.

It seems that by a proper choice of the generalized eigenvalues Λ one can generate all solutions of the algebraic Riccati equation, even all elements of $\mathbf{Q}_{\mathbf{lsp},\mathbf{r}}$.

It remains to investigate the Hamiltonian approach in more detail. The invariant subspaces of the Hamiltonian respresentation (22.30) deserve further research.

22.4 Further Reading

Lyapunov equation. The discrete-time Lyapunov equation is covered in many books. Most computer packages for control include a subroutine for it. Theorem 22.1.3 is adjusted from [26]. Proposition 22.1.4 is adjusted from [22, 13.2].

Algebraic Riccati equations. History. The story of the origins of the scalar Riccati differential equation is told in [5]. The discrete-time Riccati equation is also covered in many books and papers. An early reference on the existence and uniqueness of a solution of the continuous-time algebraic Riccati equation is W.M. Wonham [36]. The approach to prove convergence is directly related to the approach to prove convergence of the value function in optimal control over an infinite horizon. Other references are [25, 27, 30, 31, 35]. The existence of a nonpositive solution of the algebraic Riccati equation is stated in [34]. The contributions of R.S. Bucy and J. Rodriguez-Canabal include [7, 8, 28]. Books on the algebraic Riccati equation include [3, 4, 21]. For a related equation see [1].

The geometry of the Riccati differential equation was formulated by J.C. Willems in the framework of dissipative linear systems, [33]. The geometry of the Riccati differential equation was investigated independently by J.M. Rodriguez-Canabal with his research advisor R.S. Bucy, [28, 8]. M.A. Shayman has published about the phase portrait of the matrix Riccati equation, [29, 30, 31]. Theorem 22.2.11 is proven in [26].

The reader is referred to Chapter 24 where much of the geometric framework of the Riccati difference equation is described.

The approach to the algebraic Riccati equation via the invariant subspaces of the Hamiltonian matrix may for example be found in [2] and [16, Ch. 7]. The Hamiltonian approach to the algebraic Riccati equation may also be found in [24]. See also, [17] and P.M. van Dooren [32].

Square-root algorithms for the algebraic Riccati equation were developed by G.J. Bierman, [3].

Numerical procedures for the solution of the algebraic Riccati equation are discussed in [18, 24, 32]. Numerical procedures for the Hamiltonian approach to the continuous-time algebraic Riccati equation are presented in [9, 11, 23, 32].

The dependence of the solution of the algebraic Riccati equation on the given matrices is discussed in [12, 26].

The algebraic Riccati equation for the stochastic realization problem is discussed in [14]. Procedure 22.3.6 and Procedure 22.3.7 are adjusted from [14, 15]. The Hamiltonian approach to the algebraic Riccati equation of stochastic realization is inspired by [24] but different.

References

1. W.N. Anderson Jr., T.D. Morley, and G.E. Trapp. Positive solutions to $X = A - BX^{-1}B^*$. *Linear Alg. Appl.*, 134:53–62, 1990. 849
2. H. Bart, I. Gohberg, and M.A. Kaashoek. *Minimal factorization of matrix and operator functions*. Birkhäuser, Basel, 1979. 850
3. G.J. Bierman. *Factorization methods for discrete sequential estimation*. Academic Press, New York, 1977. 289, 310, 833, 849, 850
4. D. Bini, B. Iannazzo, and B. Meini. *Numerical solution of algebriac Riccati equations*. SIAM, Philadelphia, 2011. 849
5. S. Bittanti, editor. *Count Riccati and the early days of the Riccati equation*. Pitagora Editrice Bologna, Bologna, 1989. 849
6. A. Browder. *Mathematical analysis - An introduction*. Undergraduate texts in mathematics. Springer-Verlag, New York, 1996. 30, 49, 424, 426, 475, 526, 635, 636, 677, 815
7. R.S. Bucy. The riccati equation and its bounds. *J. Comput. Syst. Sci.*, 6:343–353, 1972. 849, 867
8. R.S. Bucy and J.M. Rodriguez-Canabal. A negative definite equilibrium and its induced cone of global existence for the riccati equation. *SIAM J. Math. Anal.*, 3:644–646, 1972. 849, 867
9. A. Bunse-Gerstner and V. Mehrmann. A symplectic QR-like algorithm for the solution of the real algebraic Riccati equation. *IEEE Trans. Automatic Control*, 31:1104–1113, 1986. 850
10. S. Butterworth. On the theory of filter amplifiers. *Wireless Eng.*, 7:536–541, 1930. 823
11. R.P. Byers. A Hamiltonian QR algorithm. *SIAM J. Sci. Statist. Comput.*, 7:212–229, 1986. 850
12. D.F. Delchamps. Analytic stabilization and the algebraic riccati equation. In *Proceedings 22nd Conference on Decision and Control*, pages 1396–1401, New York, 1983. IEEE Press. 850

13. John C. Doyle. Guaranteed margins for LQG regulators. *IEEE Trans. Automatic Control*, 23:756–757, 1978. 440, 592, 596, 822, 824
14. P. Faurre, M. Clerget, and F. Germain. *Opérateurs rationnels positifs*. Dunod, Paris, 1979. 175, 180, 217, 275, 292, 310, 850, 865, 867, 877, 885
15. P. Faurre and J.P. Marmorat. Un algoritme de réalisation stochastique. *C.R. Acad. Sc. Paris*, 268:978–981, 1969. 850, 885
16. B.A. Francis. *A course in H_∞ control theory*. Number 88 in Lecture Notes in Control and Information Sciences. Springer-Verlag, New York, 1987. 560, 850
17. V. Ionescu, C. Oara, and M. Weiss. General matrix pencil techniques for the solution of algebraic riccati equations: A unified approach. *IEEE Trans. Automatic Control*, 42:1085–1097, 1997. 850
18. C. Kenney, A.J. Laub, and E.A. Jonckheere. Positive and negative solutions of dual Riccati equations by matrix sign function iteration. *Systems Control Lett.*, 13:109–116, 1989. 850
19. H. Kwakernaak. Asymptotic root loci of multivariable linear optimal regulators. *IEEE Trans. Automatic Control*, 21:378–382, 1976. 593, 822, 823
20. H. Kwakernaak and R. Sivan. *Linear optimal control systems*. Wiley-Interscience, New York, 1972. 120, 376, 410, 467, 489, 593, 822, 823
21. P. Lancaster and L. Rodman. *Algebraic Riccati equations*. Oxford Science Publications, Oxford, 1995. 849
22. P. Lancaster and M. Tismenetsky. *The theory of matrices - Second edition with applications*. Academic Press, San Diego, 1985. 636, 849
23. C. Paige and C.F. Van Loan. A Schur decomposition for Hamiltonian matrices. *Linear Algebra Appl.*, 41:11–32, 1981. 850
24. T. Pappas, A.J. Laub, and N.R. Sandell Jr. On the numerical solution of the discrete-time algebraic Riccati equation. *IEEE Trans. Automatic Control*, 25:631–641, 1980. 850
25. H.J. Payne and L. Silverman. On the discrete-time algebraic riccati equation. *IEEE Trans. Automatic Control*, 18:226–234, 1973. 849
26. J.W. Polderman. A note on the structure of two subsets of the parameter space in adaptive control problems. *Systems & Control Lett.*, 7:25–34, 1986. 849, 850
27. M.A. Poubelle, I.R. Petersen, M.R. Gevers, and R.R. Bitmead. A miscellany of results on an equation of count J.F. Riccati. *IEEE Trans. Automatic Control*, 31:651–654, 1986. 849
28. J.M. Rodriquez-Canabal. The geometry of the riccati equation. *Stochastics*, 1:129–149, 1973. 849, 867
29. M.A. Shayman. Inertia theorems for the periodic Lyapunov equation and periodic Riccati equation. *Syst. Control Lett.*, 4:27–32, 1984. 850
30. M.A. Shayman. On the phase portrait of the matrix Riccati equation arising from the periodic control problem. *SIAM J. Control Optim.*, 23:717–751, 1985. 849, 850
31. M.A. Shayman. Phase portrait of the matrix Riccati equation. *SIAM J. Control Optim.*, 24:1–65, 1986. 849, 850
32. P. Van Dooren. A generalized eigenvalue approach for solving Riccati equations. *SIAM J. Sci. Stat. Comput.*, 2:121–135, 1981. 850
33. J.C. Willems. Dissipative dynamical systems - Part i: General theory - Part ii: Linear systems with quadratic supply rates. *Arch. Rational Mech. Anal.*, 45:321–351; 352–393, 1972. 849, 853, 864
34. J.C. Willems. On the existence of a nonpositive solution to the Riccati equation. *IEEE Trans. Automatic Control*, 19, 1974. 849
35. H.K. Wimmer. Strong solutions of the discrete-time Riccati equation. *Systems & Control Lett.*, 13:455–457, 1989. 849
36. W.M. Wonham. On a matrix Riccati equation of stochastic control. *SIAM J. Control*, 6:681–697, 1968. 574, 824, 849

Chapter 23
Appendix G Covariance Functions and Dissipative Systems

Abstract The concept of a dissipative system is defined for a deterministic linear system and is satisfied if there exists a storage function and a supply rate such that the dissipation inequality holds. It is proven that a deterministic linear system is dissipative if and only if a related function is a positive-definite function. Any covariance function of a stochastic process is a positive-definite function. An equivalence condition is proven for a deterministic linear system to be dissipative in terms of an algebraic condition, a matrix has to satisfy a linear matrix inequality.

Key words: Dissipative system. Storage function. Covariance function.

The concepts of a dissipative system and that of a positive-definite function play a key role in the weak Gaussian stochastic realization problem. These concepts are also useful in control theory, circuit theory, thermodynamics, and in other areas of engineering and mathematics. The formulation of a dissipative system was proposed for continuous-time linear control systems by J.C. Willems, see [7].

The concept of dissipativity is also related to optimal stochastic control theory where it refers to the inequality between the value function and the cost-to-go of a non-optimal control law.

The contribution of J.C. Willems in [7], is to focus for a characterization of dissipativity of a system on the state-space realization of the corresponding system. Optimal control corresponds then to the determination of a realization for which a function of the state is infimized. This concept then generalizes to a nonlinear system and to a stochastic system.

In case the control system is nonlinear or the storage functions of the system are not quadratic, then the concepts of this chapter have to be adjusted accordingly. For each subset of systems, one has to formulate the appropriate concepts. The reader is referred to the literature for those cases.

The purpose of this chapter is to state definitions of these concepts for linear systems, to prove the relation between these concepts, and to present an algebraic condition for a linear system to be dissipative and for a particular function to be positive definite. In Appendix 24 the classification of dissipative systems and positive-definite functions is investigated further.

The reader who first learns about dissipative systems may first read the Definitions 23.1.1, 23.1.4, and 23.4.1, and the results 23.3.1 and 23.4.2.

23.1 Definitions

Definition 23.1.1. Consider a discrete-time finite-dimensional linear control system of Def. 21.1.1 with as many output componets as there are input components,

$$\begin{aligned} &\{n_y, n_x, n_u, F, G, H, J = J^T\} \in \text{LSP}, \;\; n_y = n_u, \\ &F \in \mathbb{R}^{n_x \times n_x}, \; G \in \mathbb{R}^{n_x \times n_y}, \; H \in \mathbb{R}^{n_y \times n_x}, \; J \in \mathbb{R}^{n_y \times n_y}, \\ x(t+1) &= Fx(t) + Gu(t), \; x(t_0) = x_0 \in \mathbb{R}^{n_x}, \\ y(t) &= Hx(t) + Ju(t). \end{aligned}$$

Define the *supply rate* as the function,

$$\begin{aligned} h(u(t), y(t)) &= u(t)^T y(t) = \frac{1}{2} \begin{pmatrix} u \\ y \end{pmatrix}^T J_s \begin{pmatrix} u \\ y \end{pmatrix}, \\ J_s &= \begin{pmatrix} 0 & I_{n_y} \\ I_{n_y} & 0 \end{pmatrix}, \;\; h : \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \to \mathbb{R}. \end{aligned}$$

The system will be called a *dissipative system* with supply rate h if there exists a *storage function* $S : \mathbb{R}^{n_x} \to \mathbb{R}_+$ such that for all $s, t \in T, \; s < t, x(s) \in \mathbb{R}^{n_x}$, and for all input functions $u \in \mathbf{U}$, the following inequality holds,

$$S(x(t)) - S(x(s)) - \sum_{\tau=s}^{t-1} h(u(\tau), y(\tau)) \leq 0. \tag{23.1}$$

The inequality (23.1) will be called the *dissipation inequality.*

The interpretation of the definition of a dissipative system is that over the time interval $[s, t]$ the change in storage from the initial time $s \in T$ to the terminal time $t \in T$, $S(x(t)) - S(x(s))$, minus the supply to the system of the form $\sum_{\tau=s}^{t-1} h(\tau)$, has been dissipated. The concept of a dissipative system is based on related concepts in thermodynamics. The energy dissipated could have been used to heat air which energy use is not accounted for in the model.

Problem 23.1.2. Consider the system of Definition 23.1.1 and the supply rate h.

(a)Determine whether or not this system is dissipative.
(b)If the system is dissipative, classify all storage functions.

Terminology and notation is introduced next.

Definition 23.1.3. Let $T \subseteq \mathbb{N} = \{0,1,2,\ldots\}$.

(a)

$$\begin{aligned}
&\forall\, u : T \to \mathbb{R}^{n_y}, \;\; \|u\|_2 = (\sum_{s \in T} u(s)^T u(s))^{1/2} \in \mathbb{R}_+ \cup \{\infty\}, \\
\mathbf{U} = {} & L^2(T, \mathbb{R}^{n_y}) = \{u : T \to \mathbb{R}^{n_y} |\; \|u\|_2 < \infty\}, \\
&\forall\, u_1,\; u_2 \in \mathbf{U}, \;\; (u_1, u_2) = \sum_{s \in T} u_1(s)^T u_2(s).
\end{aligned}$$

(b)For $W : T \times T \to \mathbb{R}^{n_y \times n_y}$ define the *operator* $\mathbf{W} : \mathbf{U} \to (\mathbb{R}^{n_y})^T$ as ,

$$(\mathbf{W}(u))(t) = \sum_{s=-\infty}^{t-1} W(t,s)u(s) + 1/2W(t,t)u(t), \;\; (\mathbb{R}^{n_y})^T = \{y : T \to \mathbb{R}^{n_y}\}.$$

(c)The function $W : T \times T \to \mathbb{R}^{n_y \times n_y}$ on $T = \mathbb{Z}$ is called *stationary* if $W(t,s) = W(t-s,0)$ for all $s,t \in T$; in that case introduce the notation $W_1 : T \to \mathbb{R}^{n_y \times n_y}$, for all $t \in \mathbb{N}$, $W_1(t) = W(t,0)$ and $W_1(-t) = W_1(t)$.

(d)The function $W : T \times T \to \mathbb{R}^{n_y \times n_y}$ is called *parasymmetric* if $W(t,s) = W(s,t)^T$ for all $s,t \in T$; if W is stationary this condition is equivalent with $W_1(t) = W_1(-t)^T$ for all $t \in T$, in particular $W(0) = W(0)^T$.

(e)The parasymmetric function $W : T \to \mathbb{R}^{n_y \times n_y}$ is called *finite-dimensional* if there exists a $\{n_y, n_x, n_y,\; F, G, H, J = J^T\} \in$ LSP such that,

$$W(t) = \begin{cases} HF^{t-1}G, & \text{if } t > 0, \\ J + J^T = 2J, & \text{if } t = 0, \\ G^T(F^T)^{-t-1}H^T, & \text{if } t < 0. \end{cases}$$

(f) The function $W : T \to \mathbb{R}^{n_y \times n_y}$ will be called *asymptotically stable* if $\lim_{t\to\infty} W(t) = 0$. If W is finite-dimensional having a minimal realization (see 21.8.9.(b)), then asymptotic stability is implied by $\text{spec}(F) \subset \text{D}_o$.

(g)The function $W : T \to \mathbb{R}^{n_y \times n_y}$ will be called *nonsingular* if, for all $t \in T$, $u \in \mathbb{R}^{n_y}$, $u \neq 0$, implies that $u^T W(t) u \neq 0$.

If $W : T \to \mathbb{R}^{n_y \times n_y}$ is the covariance function of a stationary Gaussian stochastic process then W is a stationary and a parasymmetric function, see Section 3.3 and Section 3.4.

Definition 23.1.4. Let $W : T \times T \to \mathbb{R}^{n_y \times n_y}$ and $\mathbf{W} : (\mathbb{R}^{n_y})^T \to (\mathbb{R}^{n_y})^T$ be the associated operator.

(a)The operator $\mathbf{W}$ is called a *positive-definite operator* and W is called a *positive-definite function* if, for any $u \in \mathbf{U}$,

$$u^T \mathbf{W} u = \sum_{s \in T} \sum_{t \in T} u(t)^T W(t,s) u(s) \geq 0.$$

(b)The operator $\mathbf{W}$ is called *strictly positive-definite* and W a *strictly-positive-definite function* if, for all $u \in \mathbf{U}$, $u \neq 0$ implies that $u^T \mathbf{W} u > 0$.

(c) The operator $\mathbf{W}$ is called *coercive* and W a *coercive function* if there exists a $c \in (0,\infty)$ such that, for any $u \in \mathbf{U}$, $u^T \mathbf{W} u \geq c\|u\|^2$.

Proposition 23.1.5. *Let* $W : T \times T \to \mathbb{R}^{n_y \times n_y}$ *be parasymmetric. Then* W *is a positive-definite function if and only if, for all* $u \in \mathbf{U}$, $(u, \mathbf{W}u) \geq 0$. *In fact the equality* $u^T \mathbf{W} u = 2(u, Wu)$ *holds for any* $u \in \mathbf{U}$.

Proof.

$$\begin{aligned}
u^T \mathbf{W} u &= \sum_{s \in T} u(s)^T \sum_{t=-\infty}^{s-1} W(s,t)u(t) + \sum_{s \in T} u(s)^T W(s,s)u(s) \\
&\quad + \sum_{s \in T} u(s)^T \sum_{t=s+1}^{\infty} W(s,t)u(t) \\
&= (u, \mathbf{W}u) + \sum_{t \in T} u(t)^T \sum_{s=-\infty}^{t-1} W(s,t)^T u(s) + \sum_{t \in T} 1/2 u(t)^T W(t,t)u(t) \\
&\quad \text{by interchanging the summation operations,} \\
&= (u, \mathbf{W}u) + \sum_{t \in T} u(t)^T \sum_{s=-\infty}^{t-1} W(t,s)u(s) + \sum_{t \in T} 1/2 u(t)^T W(t,t)u(t), \\
&\quad \text{because } W \text{ is parasymmetric,} \\
&= 2(u, \mathbf{W}u).
\end{aligned}$$

□

If $W : T \to \mathbb{R}^{n_y \times n_y}$ is the covariance function of a stationary Gaussian process y, then W is stationary, parasymmetric and a positive definite function: for all $u \in \mathbf{U}$,

$$\begin{aligned}
u^T \mathbf{W} u &= \sum_{s \in T} \sum_{t \in T} u(t)^T W(t-s)u(s) \\
&= \sum \sum u(t)^T E[(y(t) - E[y(t)])(y(s) - E[y(s)])^T] u(s) \\
&= E|\sum_{t \in T} u(t)^T (y(t) - E[y(t)])|^2 \geq 0.
\end{aligned}$$

23.2 Storage Functions

Definition 23.2.1. Consider the dynamic system defined in Definition 23.1.1 with supply rate h. The *available storage* is a function,

$$\begin{aligned}
S^- &: \mathbb{R}^{n_x} \to \overline{\mathbb{R}}, \\
S^-(x) &= \sup_{t>0,\; u \in \mathbf{U}(0,x,t,.)} [-\sum_{\tau=0}^{t-1} \frac{1}{2} \begin{pmatrix} u(\tau) \\ y(\tau) \end{pmatrix}^T J_s \begin{pmatrix} u(\tau) \\ y(\tau) \end{pmatrix}], \\
\mathbf{U}(t_0, x_0, t_1, x_1) &= \left\{ \begin{array}{l} u \in \mathbf{U} | \text{ if } x(t_0) = x_0, \text{ and} \\ u \text{ is applied as in Def. 23.1.1, then } x(t_1) = x_1 \end{array} \right\},
\end{aligned}$$

and y is the output function of the system due to starting the system at the time-state tuple (t_0,x_0), supplying the input u, and terminating the operation with the time-state tuple (t_1,x_1).

The interpretation of the available storage is that it is the maximal available energy that one can extract from the system in terms of the negative of the supply rate.

Proposition 23.2.2. *Consider the dynamic system of Definition 23.1.1 with supply rate h.*

(a)The system is dissipative if and only if the available storage is finite, or, for all $x \in \mathbb{R}^{n_x}$, $S^-(x) < \infty$.
(b)If the system is dissipative, then S^- *is a storage function.*
(c)If the system is dissipative and if S is a storage function, then $0 \leq S^- \leq S$.

Definition 23.2.3. Consider the system defined in Def. 23.1.1 with supply rate h. The *required supply* is the function $S^+ : \mathbb{R}^{n_x} \rightarrow \overline{\mathbb{R}}$,

$$S^+(x) = \inf_{t_0<0,u\in\mathbf{U}(t_0,.,0,x)} \sum_{\tau=t_0}^{-1} \frac{1}{2} \begin{pmatrix} u(\tau) \\ y(\tau) \end{pmatrix}^T J_s \begin{pmatrix} u(\tau) \\ y(\tau) \end{pmatrix} \in \mathbb{R}_+ \cup \{\infty\}.$$

The interpretation of the required supply is that it is the supply which is necessary to transfer the system from an initial tuple (t_0,x_0) to the terminal tuple $(t_1=0,x)$.

Proposition 23.2.4. *Consider the system of Def. 23.1.1 with supply rate h.*

(a)Assume that the linear system is controllable from a state $x_s \in \mathbb{R}^{n_x}$*. Then the system is dissipative if and only if there exists a* $c \in \mathbb{R}$ *such that for all* $x \in \mathbb{R}^{n_x}$

$$inf_{t<0,u\in\mathbf{U}((t,x_s),(0,x))} \sum_{\tau=t}^{-1} u(\tau)^T y(\tau) \geq c. \text{ Then,}$$

$$S_1(x) = S^-(x_s) + inf_{t<0,u\in\mathbf{U}((t,x_s),(0,x))} \sum_{\tau=t}^{-1} u(\tau)^T y(\tau), \;\; S_1 : \mathbb{R}^{n_x} \rightarrow \mathbb{R},$$

is a storage function.
(b)Assume that the system is dissipative with storage function S and that $S(0)=0$*. Then* $S^+(0)=0$ *and* $0 \leq S^- \leq S \leq S^+$.
(c)Assume that the system is dissipative and controllable from state 0*. Then* $S^+ < \infty$ *and* S^+ *is a storage function.*

Proposition 23.2.5. *Consider the system 23.1.1 with the supply rate given there. Assume that the system is dissipative.*

(a)The set of storage functions is convex, meaning that if S_1, S_2 *are storage function and* $c \in [0,1]$*, then* $S_c = cS_1 + (1-c)S_2$ *is a storage function.*
(b)If in addition the system is controllable from state $0 \in \mathbb{R}^{n_x}$ *then, for any* $c \in [0,1]$, S_c *is a storage function where,*

$$S_c = cS^- + (1-c)S^+, \;\; S_c : \mathbb{R}^{n_x} \rightarrow \mathbb{R}.$$

Proof. (a) This follows from the dissipation inequality.
(b) This follows from (a), Proposition 23.2.2, and 23.2.4. □

Definition 23.2.6. Consider the linear system of 23.1.1 with the supply rate given there. Assume that the system is dissipative. For any storage function S define the *dissipation rate* as the function $r : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \rightarrow \mathbb{R}$,

$$r(x(t),u(t)) = S(x(t+1)) - S(x(t)) - u(t)^T y(t), \; \forall\, t \in T; \; \Rightarrow$$
$$\sum_{\tau=s}^{t-1} r(x(\tau),u(\tau)) = S(x(t)) - S(x(s)) - \sum_{\tau=s}^{t-1} u(\tau)^T y(\tau), \; \forall\, s,t \in T, \; s < t.$$

23.3 Relations

The concept of a dissipative system is formulated for a state space representation. The concept of a positive-definite function is formulated in terms of the input-output relation of a system. Yet, the criterion for a function to be positive-definite will be phrased in terms of a state space representation. It therefore seems natural to place the concept of a dissipative system central in this discussion rather than the concept of a positive-definite function. Below the connection between dissipative systems and positive-definite functions is established.

Proposition 23.3.1. *Consider the dynamic system defined in Definition 23.1.1 and the supply rate given there,*

$$x(t+1) = Fx(t) + Gu(t),$$
$$y(t) = Hx(t) + Ju(t), \; J = J^T; \; \textit{define, } W : T \rightarrow \mathbb{R}^{n_y \times n_y},$$
$$W(t) = \begin{cases} HF^{t-1}G, & \textit{if } t > 0, \\ J + J^T, & \textit{if } t = 0, \\ G^T (F^T)^{-t-1} H^T, & \textit{if } t < 0. \end{cases}$$

Assume that the system is controllable from state $0 \in \mathbb{R}^{n_x}$. *Then the system is dissipative if and only if* W *is a positive definite-function.*

Proof. By 23.1.5 W is a positive-definite function if and only if $s, \; t \in T, \; s < t$, $x(s) = 0$, $u \in \mathbf{U}$ imply that $\sum_{\tau=s}^{t-1} u(\tau)^T y(\tau) \geq 0$.
($\Rightarrow$) Suppose that W is not a positive definite function. Then there exists a $s,t \in T, s < t$, $x(s) = 0$, $u_1 \in \mathbf{U}$ such that $\sum_{\tau=s}^{t-1} u_1(\tau)^T y_1(\tau) < 0$. Let $c \in \mathbb{R}$ with $c \neq 0$. Apply instead of $u_1 \in \mathbf{U}$, $cu_1 \in \mathbf{U}$. The output is then cy_1, hence

$$\sum_{\tau=s}^{t-1} c^2 u_1(\tau)^T y_1(\tau) < 0,$$
$$S^-(0) = \sup_{t \geq 0, u \in \mathbf{U}(0,0;t,.)} (-\sum_{\tau=0}^{t-1} u(\tau)^T y(\tau)) = \infty.$$

By 23.2.2.(a) this implies that the system is not dissipative, which is a contradiction of the assumption.

($\Leftarrow$) Let $s, t \in T, s < 0 < t$, $x_0 \in \mathbb{R}^{n_x}$, $u_1 \in \mathbf{U}(s,0,0,x_0)$. Then W a positive-definite function implies that,

$$\sum_{\tau=s,\ u_1 \in \mathbf{U}(s,0;0,x_0)}^{-1} u_1(\tau)^T y_1(\tau) + \sum_{\tau=0,\ u_2 \in \mathbf{U}(0,x_0;t,.)}^{t-1} u_2(\tau)^T y_2(\tau) \geq 0; \quad \Rightarrow$$

$$- \sum_{\tau=0, u_2 \in \mathbf{U}(0,x_0;t,.)}^{t-1} u_2(\tau)^T y_2(\tau) \leq \sum_{\tau=s}^{-1} u_1(\tau)^T y_1(\tau),$$

$$S^-(x_0) = \sup_{t \geq 1, u_2 \in \mathbf{U}(0,x_0;t,.)} \left(- \sum_{\tau=0}^{t-1} u_2(\tau)^T y_2(\tau)\right) \ \leq \ \sum_{\tau=-s}^{-1} u_1(\tau)^T y_1(\tau) < \infty.$$

For s sufficiently large, $u_1 \in \mathbf{U}(s,0;0,x_0)$ because the system is controllable from state $0 \in \mathbb{R}^{n_x}$. Thus $S^- < \infty$, and by 23.2.2.(a) the system is dissipative. □

23.4 Algebraic Characterization of Dissipative Linear Systems

The problem of characterization of linear systems which are dissipative is reformulated in this section. As argued above, the storage functions of a realization of the system are quadratic functions of the state of the corresponding system, say of the form $S(x) = x^T Q x$. Such a function is characterized by its quadratic form, the symmetric square matrix $Q \in \mathbb{R}^{n_x \times n_x}$.

The problem is thus to describe the set of all matrices Q for which the function $S(x) = x^T Q x$ is a storage function of the corresponding system. It is shown below that the matrices Q which satisfy this condition is a set of symmetric square matrices satisfying a linear matrix inequality. The linear matrix inequality is regarded as an algebraic characterization of the set of dissipative linear systems. For subsets of systems which are not linear, one may expect different sets of functions of the state.

Definition 23.4.1. Let $\text{lsp} = \{n_y, n_x, n_y, F, G, H, J = J^T\} \in \text{LSP}$.
$\text{lsdp} = \{n_y, n_x, n_y, F^T, H^T, G^T, J^T = J\} \in \text{LSP}$ be the parameter of the dual system associated to $\text{lsp} \in \text{LSP}$. Assume that $0 \prec J + J^T$ and that $\text{spec}(F) \subset \text{D}_o$. Define,

$$\mathbf{Q}_{\mathbf{lsp}} = \left\{ Q \in \mathbb{R}_{pds}^{n_x \times n_x} \mid Q_{v,lsp}(Q) \succeq 0 \right\},$$

$$\mathbf{Q}_{\mathbf{lsdp}} = \left\{ Q \in \mathbb{R}_{pds}^{n_x \times n_x} \mid Q_{v,lsdp}(Q) \succeq 0 \right\};$$

$$Q_{v,lsp} : \mathbb{R}^{n_x \times n_x} \to \mathbb{R}^{(n_x+n_y) \times (n_x+n_y)},$$

$$Q_{v,lsdp} : \mathbb{R}^{n_x \times n_x} \to \mathbb{R}^{(n_x+n_y) \times (n_x+n_y)},$$

$$Q_{v,lsp}(Q) = \begin{pmatrix} Q - F^T Q F & H^T - F^T Q G \\ H - G^T Q F & J + J^T - G^T Q G \end{pmatrix}, \tag{23.2}$$

$$Q_{v,lsdp}(Q) = \begin{pmatrix} Q - F Q F^T & G - F Q H^T \\ G^T - H Q F^T & J + J^T - H Q H^T \end{pmatrix}. \tag{23.3}$$

As stated in Theorem 6.4.3.(d.3), the matrix $Q_{v,lsdp}(Q)$ is the variance matrix of the noise vector,

$$\begin{pmatrix} M \\ N \end{pmatrix} v(t).$$

Theorem 23.4.2. Characterizations of dissipativity of a system and of positive definitedness of a covariance function. *Consider the linear system of Definition 23.1.1 and recall that then* $0 \prec J + J^T$ *and* $\text{spec}(F) \subset \text{D}_o$.

$$\begin{aligned} x(t+1) &= Fx(t) + Gu(t), \\ y(t) &= Hx(t) + Ju(t), \\ h(u(t), y(t)) &= u(t)^T y(t), \quad W : T \to \mathbb{R}^{n_x \times n_x}, \\ W(t) &= \begin{cases} HF^{t-1}G, & \text{if } t > 0, \\ J + J^T, & \text{if } t = 0, \\ G^T (F^T)^{-t-1} H^T, & \text{if } t < 0. \end{cases} \end{aligned}$$

Assume that the system is a minimal realization of its associated impulse response function, see Th. 21.8.9.

(a)Then the following statments are equivalent:

(a.1)This dynamic system is dissipative;
(a.2)W is a positive definite function;
(a.3)There exists a matrix $Q \in \mathbf{Q}_{lsp}$.

(b)If the system is dissipative then there exist a minimal state-variance matrix $Q_d^- \in \mathbb{R}_{pds}^{n_x \times n_x} \cap \mathbf{Q_{lsdp}}$ which is a solution of the dual realization algebraic Riccati equation of the system with side conditions, see Def. 24.4.1, formulated for the matrix,

$$\begin{aligned} & Q_d \in \mathbb{R}_{pds}^{n_x \times n_x}, \; 0 \prec J + J^T - HQ_dH^T, \\ D_d(Q_d) &= Q_d - F^T Q_d F + \\ & \quad -(G - FQ_dH^T)(J + J^T - HQ_dH^T)^{-1}(G - FQ_dH^T)^T = 0, \\ & \quad \text{spec}(F - (G - FQ_dH^T)(J + J^T - HQ_dH^T)^{-1}H) \subset \text{D}_o; \\ Q_d^- &= Q_d. \end{aligned}$$

Because the system is a minimal realization, it is true that $0 \prec Q_d = Q_d^-$ *hence* $Q^- \in \mathbb{R}_{spds}^{n_x \times n_x}$.
In addition, if the system is dissipative then there exist a maximal state-variance matrix $Q_d^+ \in \mathbb{R}_{pds}^{n_x \times n_x} \cap \mathbf{Q_{lsdp}}$ which is also a solution of the dual realization algebraic Riccati equation of the system, see Def. 24.4.1. However, the equation for this maximal state variance matrix is mostly formulated as that of the inverse of the minimal state variance matrix of the nondual system thus of the form,

$$\begin{aligned} & Q \in \mathbb{R}_{pds}^{n_x \times n_x},\ 0 \prec J + J^T - G^T QG, \\ D(Q) &= Q - FQF^T + \\ & -(H^T - F^T QG)(J + J^T - G^T QG)^{-1}(H^T - F^T QG)^T = 0, \\ & \text{spec}(F - G(J + J^T - G^T QG)^{-1}(H^T - F^T QG)^T) \subset \mathrm{D}_o; \\ Q_d^+ &= Q^{-1}. \end{aligned}$$

Because the system is a minimal realization, it is true that $0 \prec Q$ *hence* $0 \prec Q_d^+ = Q^{-1}$ *and* $Q_d^+ \in \mathbb{R}_{spds}^{n_x \times n_x}$.
Denote for the nondual system the minimal and maximal elements by $(Q^-,\ Q^+)$ *and for the dual system the notation is* $(Q_d^-,\ Q_d^+)$.
Then the available storage and the required supply for the nondual system are given by,

$$S^-(x) = \frac{1}{2} x^T Q^- x,\ \ S^+(x) = \frac{1}{2} x^T Q^+ x,$$

(c)*The function* $S : \mathbb{R}^{n_x} \rightarrow \mathbb{R}$, $S(x) = \frac{1}{2} x^T Qx$, *with* $Q \in \mathbb{R}_{pds}^{n_x \times n_x}$ *is a storage function for the above dissipative system if and only if* $Q \in \mathbf{Q_{lsp}}$.
(d)*A consequence of the minimum and the maximum stated in (b) is that, for any* $Q \in \mathbf{Q_{lsp}}$, *the following inequality holds* $Q^- \preceq Q \preceq Q^+$ *and for any* $Q \in \mathbf{Q_{lsdp}}$, $Q_d^- \preceq Q \preceq Q_d^+$.

The proof of Theorem 23.4.2 requires the following proposition.

Proposition 23.4.3. *Consider the linear system of Theorem 23.4.2 and the supply function given there. Let* $Q \in \mathbb{R}^{n_x \times n_x}$, $Q = Q^T$. *For any* $s, t \in T, s < t$, *the following equality holds,*

$$\begin{aligned} & \frac{1}{2} x(t)^T Qx(t) - \frac{1}{2} x(s)^T Qx(s) - \sum_{\tau=s}^{t-1} \frac{1}{2} \begin{pmatrix} u(\tau) \\ y(\tau) \end{pmatrix}^T J_s \begin{pmatrix} u(\tau) \\ y(\tau) \end{pmatrix} \\ &= - \sum_{\tau=s}^{t-1} \frac{1}{2} \begin{pmatrix} x(\tau) \\ u(\tau) \end{pmatrix}^T Q_{v,lsp}(Q) \begin{pmatrix} x(\tau) \\ u(\tau) \end{pmatrix}, \end{aligned} \tag{23.4}$$

where $Q_{v,lsp}(Q)$ *is as defined in equation (23.2).*

Proof.

$$\begin{aligned}
&\frac{1}{2}x(t)^T Qx(t) - \frac{1}{2}x(s)^T Qx(s) - \sum_{\tau=s}^{t-1} u(\tau)^T y(\tau) \\
&= \sum_{\tau=s}^{t-1}[\frac{1}{2}x(\tau+1)^T Qx(\tau+1) - \frac{1}{2}x(\tau)^T Qx(\tau) - u(\tau)^T y(\tau)] \\
&= \frac{1}{2}\sum_{\tau=s}^{t-1}[(Fx(\tau)+Gu(\tau))^T Q(Fx(\tau)+Gu(\tau)) - x(\tau)^T Qx(\tau) \\
&\quad -2u(\tau)^T(Hx(\tau)+Ju(\tau))] \\
&= -\frac{1}{2}\sum_{\tau=s}^{t-1} x(\tau)^T(Q-F^T QF)x(\tau) + 2u(\tau)^T(H-G^T QF)x(\tau) \\
&\quad +u(\tau)^T(2J-G^T QG)u(\tau) \\
&= -\frac{1}{2}\sum_{\tau=s}^{t-1}\begin{pmatrix} x(\tau) \\ u(\tau)\end{pmatrix}^T Q_{v,lsp}(Q)\begin{pmatrix} x(\tau) \\ u(\tau)\end{pmatrix}.
\end{aligned}$$

□

Proof. Of Theorem 23.4.2. (1) If there exists a $Q \in \mathbf{Q_{lsp}}$ then $Q \in \mathbb{R}_{pds}^{n_x \times n_x}$ and $Q_{v,lsp}(Q) \succeq 0$. Thus, by Proposition 23.4.3,

$$\begin{aligned}
&\frac{1}{2}x(t)^T Qx(t) - \frac{1}{2}x(s)^T Qx(s) - \sum_{\tau=s}^{t-1} u(\tau)^T y(\tau) \\
&= -\frac{1}{2}\sum_{\tau=s}^{t-1}\begin{pmatrix} x(\tau) \\ u(\tau)\end{pmatrix}^T Q_{v,lsp}(Q)\begin{pmatrix} x(\tau) \\ u(\tau)\end{pmatrix} \leq 0.
\end{aligned}$$

and $S(x) = 1/2x^T Qx$ is a storage function and the system dissipative.
(2) Assume that the system is dissipative. Let $x \in \mathbb{R}^{n_x}$ and let

$$\mathbf{U}(-\infty,0,0,x) = \{u \in \mathbf{U} | x = \sum_{s=-\infty}^{-1} F^{-s-1}Gu(s)\}.$$

Consider the optimal control problem

$$\begin{aligned}
x(t+1) &= Fx(t) + Gu(t), x(-\infty) = 0, \\
S^+(x) &= \inf_{t<0, u \in \mathbf{U}(t,x(t),0,x)} \sum_{\tau=t}^{-1} u(\tau)^T y(\tau) \\
&= \inf_{u \in \mathbf{U}(-\infty,0,0,x)} \sum_{\tau=-\infty}^{-1} u(\tau)^T y(\tau) \\
&= \inf_{u \in \mathbf{U}(-\infty,0,0,x)} \sum_{\tau=-\infty}^{-1} [u(\tau)^T Hx(\tau) + u(\tau)^T Ju(\tau)].
\end{aligned}$$

Because the system is a minimal realization, it is controllable, hence $\mathbf{U}(-\infty,0,0,x) \neq \emptyset$. By assumption $0 \prec J+J^T$. Then it follows from the results of the realization algebraic Riccati equation, see Theorem 22.3.2, for the nondual system that there exists a matrix $Q^- \in \mathbb{R}_{pds}^{n_x \times n_x}$ such that, which is a solution of the realization algebraic Riccati equation, formulated below for the matrix,

$$\begin{aligned} & Q \in \mathbb{R}_{pds}^{n_x \times n_x}, \\ D(Q) &= Q - F^T Q F + \\ & -(H^T - F^T Q G)(J + J^T - G^T Q G)^{-1}(H^T - F^T Q G)^T = 0, \\ & \text{spec}(F - (J + J^T - G^T Q G)^{-1}(H^T - F^T Q G)^T) \subset \mathrm{D}_o; \\ & Q^- = Q; \ \text{ moreover,} \\ u(t) &= -[J + J^T - G^T Q^- G]^{-1}[H^T - F^T Q G]^T x(t), \\ S^-(x) &= \frac{1}{2} x^T Q^- x. \end{aligned}$$

However, in Section 22.3 these results are stated only for the dual realization algebraic Riccati equation. The other statements follow from the theory of Chapter 24.
(3) Assume that the system is dissipative with storage function $S(x) = \frac{1}{2} x^T Q x$. Take $t \in T$, $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^{n_y}$. Then the dissipation inequality and 23.4.3 imply that,

$$S(x(t+1)) - S(x(t)) - u(t)^T y(t) = -\frac{1}{2} \begin{pmatrix} x(t) \\ u(t) \end{pmatrix}^T Q_{v,lsp}(Q) \begin{pmatrix} x(t) \\ u(t) \end{pmatrix} \leq 0.$$

Since $x(t), u(t)$ are arbitrary, $Q_{v,lsp}(Q) \succeq 0$, hence $Q \in \mathbf{Q_{lsp}}$.
(4) The proof of (a) now follows from Proposition 23.4.3 and from (1) for the direction ($\Leftarrow$), while for the other direction it follows from Proposition 23.2.4.(c), (2) and (4). The proof of (b) follows from (2) and (3), while that of (c) follows from (1) and (4).
(5) If $Q \in \mathbf{Q_{lsp}}$ then by (c) $S(x) = 1/2 x^T Q x$ is a storage function. From Proposition 23.2.4.(b) then follows that,

$$\begin{aligned} & 0 \preceq S^- \preceq S \preceq S^+, \ 0 \leq \frac{1}{2} x^T Q^- x \ \leq \frac{1}{2} x^T Q x \leq \frac{1}{2} x^T Q^+ x, \ \forall x \in \mathbb{R}^{n_x}, \\ & \Rightarrow Q^- \preceq Q \preceq Q^+. \end{aligned}$$

□

Definition 23.4.4. Consider the system in LS defined in Definition 23.1.1

$$\begin{aligned} x(t+1) &= F x(t) + G u(t), \\ y(t) &= H x(t) + J u(t), \\ & \text{lsp} = \{n_y, n_x, n_y, F, G, H, J\} \in \text{LSP}, \ J = J^T. \end{aligned}$$

The *dual system* associated with the system is

$$\begin{aligned} z(t+1) &= F^T z(t) + H^T u(t), \\ y(t) &= G^T x(t) + J u(t), \\ & \text{lsdp} = \{p, n, p, F^T, H^T, G^T, J\} \in \text{LSP}, \ J = J^T. \end{aligned}$$

Note that lsp $\in \text{LSP}_{min}$ if and only if lsdp $\in \text{LSP}_{min}$.

Definition 23.4.5. For a function $W : T \to \mathbb{R}^{n_y \times n_y}$ its associated *time-reversed* function $W_d : T \to \mathbb{R}^{n_y \times n_y}$, reversed at $t = 0$, is defined by $W_d(t) = W(-t)$.

If $W : T \to \mathbb{R}^{n_y \times n_y}$ is parasymmetric, then its time-reversed function satisfies $W_d(t) = W(-t) = W(t)^T$. If W is parasymmetric and finite-dimensional, say

$$W(t) = \begin{cases} HF^{t-1}G, & \text{if } t > 0, \\ 2J, & \text{if } t = 0, \\ G^T(F^T)^{-t-1}H^T, & \text{if } t < 0. \end{cases} \text{, then,}$$

$$W_d(t) = \begin{cases} G^T(F^T)^{t-1}H^T, & \text{if } t > 0, \\ 2J, & \text{if } t = 0, \\ HF^{-t-1}G, & \text{if } t < 0. \end{cases}$$

Proposition 23.4.6. *Consider the linear system of Definition 23.1.1 with the supply function given there. Associate with this system the function W as in Proposition 23.3.1, with W the time-reversed function W_d, and with W_d the system* LSPS *which is the dual system of* LSP. *Assume that $\sigma \in$ LS is minimal and that $J = J^T \succeq 0$. The following statements are equivalent:*

(a)there exists a $Q \in \mathbf{Q_{lsp}}$ with LSP $= \{n_y, n_x, n_y, F, G, H, J\} \in \text{LSP}_{min}$*;*
(b)The system is a dissipative system;
(c)W is a positive definite function;
(d)W_d is a positive definite function;
(e)$\sigma_d \in$ LS is a dissipative system;
(f) there exists a $Q_d \in \mathbf{Q_{lsdp}}$, with LSPS $= \{n_y, n_x, n_y, F^T, H^T, G^T, J\} \in \text{LSP}_{min}$.

Proof. The equivalence of (a), (b), and (c) follows from Theorem 23.4.2, and similarly that of (d), (e), and (f). The equivalence from (c) and (d) follows from

$$\begin{aligned} \sum_{t \in T} \sum_{s \in T} u(t)^T W(t-s) u(s) &= \sum_{t \in T} \sum_{s \in T} u(s)^T W(t-s)^T u(t) \\ &= \sum_{t \in T} \sum_{s \in T} u(s)^T W_d(t-s) u(t). \end{aligned}$$

□

The proof that a linear system is dissipative, or the corresponding output operator positive definite, if and only if there exists a $Q \in \mathbf{Q_{lsp}}$ has been given via optimal control theory. The case of the characterization of dissipative system in case J is not strictly positive definite may be treated via the solution of the singular optimal control problem.

23.5 Further Reading

The main reference on dissipative systems is the paper by J.C. Willems [7] which deals with dissipativity of continuous-time linear systems.

Dissipativity of discrete-time linear systems was developed by P. Faurre and colleagues and is described in the French-language book [2, Ch. 3]. The presentation of this chapter is reformulated with respect to that reference.

An alternative proof of Theorem 23.4.2 can be formulated in the frequency domain and in terms of spectral density matrices. An early paper is that of B. McMillan, [4, 5], which deals with the characterization of complex-valued matrices which can be realized as open-circuit impedance matrices of finite passive networks. Other papers are those of D.C. Youla, [8], and V.A. Jakubovic, [3].

Recent books on dissipative nonlinear systems include [6].

The linear matrix inequality of a dissipative system is discussed from the linear algebra view point in [1].

References

1. S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear matrix inequalities in system and control theory*. Number 15 in SIAM Studies in Applied Mathematics. SIAM, Philadelphia, 1994. 865
2. P. Faurre, M. Clerget, and F. Germain. *Opérateurs rationnels positifs*. Dunod, Paris, 1979. 175, 180, 217, 275, 292, 310, 850, 865, 867, 877, 885
3. V.A. Jakubovič. The solution of certain matrix inequalities in automatic control. *Soviet Math.*, 3:620–623, 1962. 865
4. B. McMillan. Introduction to formal realizability theory I. *Bell System Techn. J.*, 31:217–279, 1952. 174, 807, 865
5. B. McMillan. Introduction to formal realizability theory II. *Bell System Techn. J.*, 31:541–600, 1952. 174, 807, 865
6. A. van der Schaft. *L_2-Gain and passivity techniques in nonlinear control*. Number 218 in LNCIS. Springer, Berlin, 1996. 865
7. J.C. Willems. Dissipative dynamical systems - Part i: General theory - Part ii: Linear systems with quadratic supply rates. *Arch. Rational Mech. Anal.*, 45:321–351; 352–393, 1972. 849, 853, 864
8. D.C. Youla. On the factorization of rational matrices. *IEEE Trans. Information Theory*, 7:172–189, 1961. 217, 865

Chapter 24
Appendix H State-Variance Matrices

Abstract Concepts and results of the geometric structure of the set of state variance matrices of a time-invariant Gaussian system are provided in this chapter. With respect to a condition, the set is convex with a minimal and a maximal element. In case the noise variance matrix satisfies a nonsingularity condition, the matrix inequality is equivalent to an inequality of Riccati type. The singular boundary matrices of the set of state variances play a particular role. Finally the classification of all elements of the set of state variances can be described in terms of an increment above the minimal element or below the maximal element, which increments satisfy a Lyapunov equation.

Key words: Linear matrix inequality. Geometric structure.

In Appendix 23 the concepts of a dissipative dynamic system and a positive definite function have been introduced. There it has been shown that a covariance realization is positive-definite if and only if the associated system is dissipative if and only if there exists a matrix in the set of state variance matrices $\mathbf{Q_{lsp}}$. In this appendix the structure of the set $\mathbf{Q_{lsp}}$ is described.

The main concepts of this appendix are those given in Definition 24.1.1, 24.4.1, 24.5.1, and 24.6.1. The structure of the set $\mathbf{Q_{lsp}}$ and that of its dual set $\mathbf{Q_{lsdp}}$ are primarily characterized by the results of Theorem 24.3.1, Proposition 24.4.2, Proposition 24.6.2, Proposition 24.6.3, and Theorem 24.7.3.

The detailed framework for the classification of the set of state variance matrices for the continuous-time case is due to P. Faurre and his co-workers M. Clerget and F. Germain, see their French language book [3].

However, the framework of the set of variance matrices was formulated in the paper of J.C. Willems on dissipative linear systems, see [7]. Related publications are due to J. Rodriguez-Canabal and R.S. Bucy, [1, 2, 6]. M. Shayman has investigated the phase-portraint of the continuus-time Riccati equation.

There follows a brief sketch of the geometric framework of the set of state variance matrices. There exists a minimal and a maximal element of that set denoted by $Q^-,\ Q^+ \in \mathbf{Q_{lsp}}$ as described in the previous chapter. One defines the cones of positive-definite symmetric matrices, denoted by $Q \in \mathbb{R}_{pds}^{n_x \times n_x}$, by the formulas $Q^- \prec Q$ and $Q \prec Q^+$. Then $\mathbf{Q_{lsp}}$ equals the intersection of these two cones. It is possible to formulate a Lyapunov-like inequality to describe the elements of these two cones of positive-definite symmetric matrices. The theory of stochastic realization of a σ-algebras or of a σ-algebra family, as described in the Sections 7.3 and 7.4, is closely related to the above formulation of the intersection of two cones.

This chapter treats the set of state variance matrices of discrete-time Gaussian system and the theory is inspired by the continuous-time case. The formulations and the proofs are therefore different from those of the quoted reference.

24.1 Definition and Problem Formulation

Definition 24.1.1. Let $lsp = \{n_y, n_x, n_y, F, G, H, J = J^T\} \in$ LSP with $J \in \mathbb{R}_{pds}^{n_y \times n_y}$ and $lsdp = \{n_y, n_x, n_y, F^T, H^T, G^T, J = J^T\} \in$ LSP.

Define respectively the *state-variance matrix* Q, the *set of state variance matrices* $\mathbf{Q_{lsp}}$, the *forward noise variance matrix* $Q_{v,lsp}(Q)$, the *set of state variance matrices of the dual system* $\mathbf{Q_{lsdp}}$, and the *backward noise variance matrix* $Q_{v,lspd}(Q)$, by the formulas,

$$\begin{aligned}
& Q \in \mathbb{R}_{pds}^{n_x \times n_x}, \\
\mathbf{Q_{lsp}} &= \{Q \in \mathbb{R}_{pds}^{n_x \times n_x} | \; Q_{v,lsp}(Q) \succeq 0\}, \\
Q_{v,lsp}(Q) &= \begin{pmatrix} Q - F^T Q F & H^T - F^T Q G \\ H - G^T Q F & J + J^T - G^T Q G \end{pmatrix}, \\
\mathbf{Q_{lsdp}} &= \{Q \in \mathbb{R}_{pds}^{n_x \times n_x} | \; Q_{v_d,lsdp}(Q) \succeq 0\}, \\
Q_{v_d,lsdp}(Q) &= \begin{pmatrix} Q - F Q F^T & G - F Q H^T \\ G^T - H Q F^T & J + J^T - H Q H^T \end{pmatrix}.
\end{aligned}$$

The inequality $Q_{v,lsp}(Q) \succeq 0$ for the matrix Q is called an *affine matrix inequality*. In the literature this also called a linear matrix inequality.

For the reader it is recalled that of a time-invariant forward Gaussian system representation and the associated backward Gaussian system representation, it follows from Theorem 4.4.5 and from Theorem 4.4.7 that the system matrices satisfy the following relations,

$$\begin{aligned}
Q_v(Q) &= \begin{pmatrix} Q_x - A Q_x A^T & Q_{x^+,y} - A Q_x C^T \\ Q_{x^+,y}^T - C Q_x A^T & Q_y - C Q_x C^T \end{pmatrix} = \begin{pmatrix} M M^T & M N^T \\ N M^T & N N^T \end{pmatrix} \succeq 0, \\
Q_{v_d}(Q) &= \begin{pmatrix} Q_x - A_b Q_x A_b^T & Q_{x,y} - A_b Q_x C_b^T \\ Q_{x,y}^T - C_b Q_x A_b^T & Q_y - C_b Q_x C_b^T \end{pmatrix} = \begin{pmatrix} M_b M_b^T & M_b N_b^T \\ N_b M_b^T & N_b N_b^T \end{pmatrix} \succeq 0.
\end{aligned}$$

Thus the asymptotic state variance matrix Q_x of a Gaussian system with the parameters $(n_y, n_x, n_y, A, Q_{x^+,y}, C^T, Q_y)$ satisfies that $Q_x \in \mathbf{Q_{lsdp}}$

Problem 24.1.2. Given $lsp \in$ LSP and $\mathbf{Q_{lsp}}$.

(a) Classify all elements of $\mathbf{Q_{lsp}}$.
(b) Determine a procedure for the computation of elements of $\mathbf{Q_{lsp}}$.

24.2 Transformations

Proposition 24.2.1. *Consider the parameters of a linear system* $\{n_y, n_x, n_y, F, G, H, J = J^T\} \in \text{LSP}$, $Q_1,\ Q_2 \in \mathbf{Q_{lsp}}$, *and* $c \in [0,1]$. *Then,*

$$Q_v(cQ_1 + (1-c)Q_2) = cQ_v(Q_1) + (1-c)Q_v(Q_2).$$

Consequently, $\mathbf{Q_{lsp}}$ *is a convex set. For* $\mathbf{Q_{lsdp}}$ *a corresponding conclusion holds.*

Proof. From the affine dependence of the matrix $Q_v(Q)$ on the state variance matrix Q follows that,

$$\begin{aligned} & Q_1,\ Q_2 \in \mathbf{Q_{lsp}} \ \Rightarrow\ Q_v(Q) \succeq 0,\ Q_v(Q_2) \succeq 0, \\ \Rightarrow\ & Q_v(cQ_1 + (1-c)Q_2) = cQ_v(Q_1) + (1-c)Q_v(Q_2) \succeq 0 \\ \Rightarrow\ & Q = cQ_1 + (1-c)Q_2 \in \mathbf{Q_{lsp}}. \end{aligned}$$

□

Definition 24.2.2. *Congruence*. Let $lsp_1 = \{n_y, n_x, n_y, F_1, G_1, H_1, J_1\}$, $lsp_2 = \{n_y, n_x, n_y, F_2, G_2, H_2, J_2\} \in \text{LSP}$. Then $\mathbf{Q}_{lsp_1}$ is called *congruent* to $\mathbf{Q}_{lsp_2}$ if there exists a matrix $L_x \in \mathbb{R}_{nsng}^{n_x \times n_x}$ nonsingular such that, if $Q_1 \in \mathbf{Q}_{lsp_1}$, then $L_x^{-T} Q_1 L_x^{-1} \in \mathbf{Q}_{lsp_2}$ and, if $Q_2 \in \mathbf{Q}_{lsp_2}$, then $L_x^T Q_2 L_x \in \mathbf{Q}_{lsp_1}$. Congruence of these sets is denoted by $\mathbf{Q}_{lsp_1} = L_x^T \mathbf{Q}_{lsp_2} L_x$.

One can prove that congruence is an equivalence relation.

Proposition 24.2.3. Characterization of congruence as a system similarity. *Let* $lsp_1 = \{n_y, n_x, n_y, F_1, G_1, H_1, J_1 = J_1^T\}$, $lsp_2 = \{n_y, n_x, n_y, F_2, G_2, H_2, J_2 = J_2^T\} \in \text{LSP}$, *be related by the nonsingular matrices* $L_x \in \mathbb{R}_{nsng}^{n_x \times n_x}$, $L_y \in \mathbb{R}_{nsng}^{n_y \times n_y}$, *via,*

$$F_2 = L_x F_1 L_x^{-1},\ G_2 = L_x G_1 L_y,\ H_2 = L_y^T H_1 L_x^{-1},\ J_2 = L_y^T J_1 L_y. \tag{24.1}$$

Then $\mathbf{Q}_{lsp_1}, \mathbf{Q}_{lsp_2}$ *are congruent, in fact* $\mathbf{Q}_{lsp_1} = L_x^T \mathbf{Q}_{lsp_2} L_x$. *Moreover,* $\mathbf{Q}_{lsdp_1} = L_x^{-1} \mathbf{Q}_{lsdp_2} L_x^{-T}$.

Proof. Let $Q_1 \in \mathbf{Q}_{lsp_1}$. Then $Q_1 = Q_1^T$ and $Q_{v1,lsp}(Q_1) \succeq 0$. Furthermore,

$$\begin{aligned} & L_x^{-T} Q_v(Q_1) L_x^{-1} = (L_x^{-T} Q_v(Q_1) L_x^{-1})^T \succeq 0, \text{by } Q_1 \in \mathbf{Q}_{lsp_1}, \\ & Q_{v2}(L_x^{-T} Q_1 L_x^{-1}) \\ & = \left(\begin{array}{c|c} L_x^{-T} Q_1 L_x^{-1} - F_2^T L_x^{-T} Q_1 L_x^{-1} F_2 & H_2^T - F_2^T L_x^{-T} Q_1 L_x^{-1} G_2 \\ \hline H_2 - G_2^T L_x^{-T} Q_1 L_x^{-1} F_2 & 2J_2 - G_2^T L_x^{-1} Q_1 L_x^{-1} G_2 \end{array} \right) \\ & = \begin{pmatrix} L_x^{-T} & 0 \\ 0 & L_y^{-T} \end{pmatrix} \left(\begin{array}{c|c} Q_1 - F_1^T Q_1 F_1 & H_1^T - F_1^T Q_1 G_1 \\ \hline H_1 - G_1^T Q_1 F_1 & J_1 + J_1^T - G_1^T Q_1 G_1 \end{array} \right) \begin{pmatrix} L_x^{-1} & 0 \\ 0 & L_y^{-1} \end{pmatrix} \succeq 0, \end{aligned}$$

because $Q_1 \in \mathbf{Q}_{lsp_1}$. Thus $L_x^{-T} Q_1 L_x^{-1} \in \mathbf{Q}_{lsp_2}$. By symmetry the conclusion follows. □

Proposition 24.2.4. Relation of the state variance matrices of a system and its dual system. *Let* $lsp = \{n_y, n_x, n_y, F, G, H, J\} \in \text{LSP}$ *and assume that* $J + J^T \prec 0$. *Then* $Q \in \mathbf{Q_{lsdp}}$ *and* $Q \succ 0$ *if and only if* $Q^{-1} \in \mathbf{Q_{lsp}}$ *and* $Q^{-1} \succ 0$. *If either condition holds then,*

$$\begin{pmatrix} Q & 0 \\ 0 & I \end{pmatrix} Q_v(Q^{-1}) \begin{pmatrix} Q & 0 \\ 0 & I \end{pmatrix} = Q_{v_d}(Q).$$

Proof. (⇒) Let $Q \in \mathbf{Q_{lsdp}}$, hence

$$Q_{v_d}(Q) = \begin{pmatrix} Q - FQF^T & G - FQH^T \\ G^T - HQF^T & J + J^T - HQH^T \end{pmatrix} \succeq 0.$$

By Theorem 6.4.3 there exists a weak Gaussian stochastic realization of a covariance function of the form,

$$W(t) = \begin{cases} HF^{t-1}G, & \text{if } t > 0, \\ J + J^T, & \text{if } t = 0, \\ G^T (F^T)^{-t-1} H^T, & \text{if } t < 0, \end{cases}$$

in which the variance of the state process is Q. By Theorem 6.4.3 this stochastic system has a backward representation. The relation between the parameters of the forward and the backward representation are specified in Theorem 4.5.2.

$$\begin{aligned}
A^b &= Q(A^f)^T Q^{-1} = QF^T Q^{-1}, \\
C^b &= C^f Q (A^f)^T Q^{-1} + N Q_{v_d}(Q) M^T Q^{-1} \\
&= HQF^T Q^{-1} + (G^T - HQF^T) Q^{-1} = G^T Q^{-1}, \\
H &= C^f = C^b A^f + NM^T Q^{-1} = G^T Q^{-1} F + NM^T Q^{-1}, \\
& \quad NQ_{v_d}(Q) M^T = (H - G^T Q^{-1} F) Q = HQ - G^T Q^{-1} FQ.
\end{aligned}$$

Thus,

$$\begin{aligned}
& \begin{pmatrix} Q & 0 \\ 0 & I \end{pmatrix} Q_v(Q^{-1}) \begin{pmatrix} Q & 0 \\ 0 & I \end{pmatrix} \\
&= \begin{pmatrix} Q & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} Q^{-1} - F^T Q^{-1} F & H^T - F^T Q^{-1} G \\ H - G^T Q^{-1} F & J + J^T - G^T Q^{-1} G \end{pmatrix} \begin{pmatrix} Q & 0 \\ 0 & I \end{pmatrix} \\
&= \begin{pmatrix} Q - QF^T Q^{-1} FQ & QH^T - QF^T Q^{-1} G \\ HQ - G^T Q^{-1} FQ & J + J^T - G^T Q^{-1} G \end{pmatrix} \\
&= \left(\begin{array}{c|c} Q - QF^T Q^{-1} Q Q^{-1} FQ & QH^T - QF^T Q^{-1} G \\ \hline HQ - G^T Q^{-1} FQ & J + J^T - G^T Q^{-1} Q Q^{-1} G \end{array} \right) \\
&= \begin{pmatrix} Q - A_b Q (A_b)^T & M_b Q_v(Q) N_b^T \\ N_b Q_v(Q) M_b^T & J + J^T - C_b Q (C_b)^T \end{pmatrix} \\
&= \begin{pmatrix} Q - FQF^T & G - FQH^T \\ (G - FQH^T)^T & J + J^T - HQH^T \end{pmatrix} = Q_{v_d}(Q) \succeq 0.
\end{aligned}$$

hence $Q_v(Q^{-1}) \succeq 0$ and $Q^{-1} \in \mathbf{Q_{lsp}}$.
($\Leftarrow$) This follows by symmetry from the above direction. □

Proposition 24.2.5. Transformation to independent noise components. *Let* $lsp = \{n_y, n_x, n_y, F, G, H, J = J^T\} \in \text{LSP}$ *and consider* $\mathbf{Q_{lsp}}$. *Assume that* $J \succ 0$. *Then* $Q \in \mathbf{Q_{lsp}}$ *if and only if*

$$lsp2 = \{n_y, n_x, n_y, F_1, G, H, J = J^T\} \in \text{LSP}, \tag{24.2}$$

$$F_1 = F - G(J+J^T)^{-1}H, \; Q \in \mathbb{R}_{pds}^{n_x \times n_x}, \tag{24.3}$$

$$Q_{v2,lsp2}(Q) = \left(\begin{array}{c|c} Q - F_1^T Q F_1 - H^T(J+J^T)^{-1}H & -F_1^T QG \\ \hline -G^T Q F_1 & J + J^T - G^T QG \end{array}\right) \succeq 0.$$

$$Q_{v2,lsp2}(Q) = L_e Q_{v,lsp} v(Q) L_e^T, \; with, \tag{24.4}$$

$$L_e = \left(\begin{array}{c|c} I & -H^T(J+J^T)^{-1} \\ \hline 0 & I \end{array}\right) \in \mathbb{R}_{nsng}^{(n_x+n_y)\times(n_x+n_y)}. \tag{24.5}$$

Proof. That (24.4) holds is a calculation, and the result then follows from the definition of $\mathbf{Q_{lsp}}$. □

Proposition 24.2.6. Monotonicity of the set of state variances.
Let $lsp_1 = \{n_y, n_x, n_y, F_1, G_1, H_1, J_1 = J_1^T\}$, $lsp_2 = \{n_y, n_x, n_y, F_2, G_2, H_2, J_2 = J_2^T\} \in$ LSP, *and* $0 \prec J_1 \preceq J_2$. *Then* $\mathbf{Q}_{lsp_1} \subseteq \mathbf{Q}_{lsp_2}$.

Proof. Let $Q \in \mathbf{Q}_{lsp_1}$. From the assumption follows that $J_2 \succeq J_1$ hence $J_2 - J_1 \succeq 0$. Then $Q \in \mathbb{R}_{pds}^{n_x \times n_x}$ and

$$\begin{aligned} Q_{v2}(Q) &= \begin{pmatrix} Q - F_1^T Q F_1 & H_1^T - F_1^T Q G_1 \\ H_1 - G_1^T Q F_1 & J_2 + J_2^T - G_1^T Q G_1 \end{pmatrix} \\ &= Q_{v1}(Q) + \begin{pmatrix} 0 & 0 \\ 0 & (J_2 - J_1) + (J_2 - J_1)^T \end{pmatrix} \succeq 0; \quad \Rightarrow \; Q \in \mathbf{Q}_{lsp_2}. \end{aligned}$$

□

24.3 The Geometric Structure

Theorem 24.3.1. Geometric structure of the set of state variances.
Consider $lsp = \{n_y, n_x, n_y, F, G, H, J = J^T\} \in \text{LSP}_{min}$ *and* $\mathbf{Q_{lsp}}$. *Assume that* $\mathbf{Q_{lsp}} \neq \emptyset$, *and that* $J \succ 0$. *Then* $\mathbf{Q_{lsp}}$ *is a convex, closed, and bounded set, and there exists a* $Q^-, Q^+ \in \mathbf{Q_{lsp}}$ *respectively the minimal and the maximal element of the set of state variances* $\mathbf{Q_{lsp}}$, *hence, for any* $Q \in \mathbf{Q_{lsp}}$, $Q^- \preceq Q \preceq Q^+$.

Proof. That $\mathbf{Q_{lsp}}$ is convex follows from Proposition 24.2.1. That it is closed follows directly from the definition. The existence of $Q^-, Q^+ \in \mathbf{Q_{lsp}}$ with the indicated inequality follows from Theorem 23.4.2.(d). This then implies that $\mathbf{Q_{lsp}}$ is bounded. □

Definition 24.3.2. *Strictly-positive-definite state variance matrices.*
A set $\mathbf{Q} \subseteq \{Q \in \mathbb{R}_{pds}^{n_x \times n_x}\}$ will be called *strictly positive-definite* if for any $Q \in \mathbf{Q}$ the inequality $Q \succ 0$ holds. The notation $\mathbf{Q} \succ 0$ denotes this property.

Further results on the geometric structure follow in Section 24.7 after the introduction of additional concepts.

24.4 Regularity

The concept of regularity refers to the nonsingularity of a submatrix of the noise variance matrix. This turns out to be useful for the subsequent theory.

Definition 24.4.1. *Regular state variance matrices*. Consider the parameters of a linear system, $\text{LSP} = \{n_y, n_x, n_y, F, G, H.J = J^T\} \in \text{LSP}$.

(a) The *regular part of* $\mathbf{Q_{lsp}}$ is defined as the subset

$$\mathbf{Q_{lsp,r}} = \{Q \in \mathbf{Q_{lsp}} | J + J^T - G^T Q G \succ 0\}. \tag{24.6}$$

The set $\mathbf{Q_{lsp}}$ will be called *regular* if $\mathbf{Q_{lsp}} = \mathbf{Q_{lsp,r}}$. The *regular part of* $\mathbf{Q_{lsdp}}$ is defined as the subset

$$\mathbf{Q_{lsdp,r}} = \{Q \in \mathbf{Q_{lsdp}} | J + J^T - H Q H^T \succ 0\}. \tag{24.7}$$

The set $\mathbf{Q_{lsdp}}$ will be called *regular* if $\mathbf{Q_{lsdp}} = \mathbf{Q_{lsdp,r}}$.

(b) Define the *realization algebraic Riccati function* (RARE),

$$D : \mathbf{Q_{lsp,r}} \to \mathbb{R}_{pds}^{n_x \times n_x},$$

$$D(Q) = Q - F^T Q F - [H^T - F^T Q G][J + J^T - G^T Q G]^{-1}[H^T - F^T Q G]^T \tag{24.8}$$

The equation $D(Q) = 0$ for $Q \in \mathbb{R}_{spd}^{n_x \times n_x}$ will be called an *algebraic Riccati equation of stochastic realization*.

(c) Correspondingly define the *dual algebraic Riccati function*,

$$D : \mathbf{Q_{lsdp,r}} \to \mathbb{R}_{pds}^{n_x \times n_x},$$

$$D_d(Q) = Q - F Q F^T - [G - F Q H^T][J + J^T - H Q H^T]^{-1}[G - F Q H^T]^T,$$

and $\mathbf{Q_{lsdp}}$ is regular if $\mathbf{Q_{lsdp}} = \mathbf{Q_{lsdp,r}}$. The equation $D_d(Q) = 0$ for $Q \in \mathbb{R}_{pds}^{n_x \times n_x}$ is seen also to be an algebraic Riccati equation of stochastic realization. The equation $D_d(Q) = 0$ will be called the *dual algebraic Riccati equation of stochastic realization* of the state-variance matrix Q.

Proposition 24.4.2. Characterization of regularity by the maximal state variance. *Let* $lsp \in \text{LSP}$*, and assume that there exists a matrix* $Q^+ \in \mathbf{Q_{lsp}}$ *such that for all* $Q \in \mathbf{Q_{lsp}}$*,* $Q \preceq Q^+$*. Then* $\mathbf{Q_{lsp}}$ *is regular if and only if* $0 \prec J + J^T - G^T Q^+ G$*.*

Proof. ($\Rightarrow$) This follows from the definition of $\mathbf{Q_{lsp}}$ being regular and $Q^+ \in \mathbf{Q_{lsp}}$. ($\Leftarrow$) Let $Q \in \mathbf{Q_{lsp}}$. It follows from Theorem 24.3.1 and from the definition of Q^+ that $Q \preceq Q^+$. Thus $0 \prec J + J^T - G^T Q^+ G \preceq J + J^T - G^T QG$, $Q \in \mathbf{Q_{lsp,r}}, \mathbf{Q_{lsp}} = \mathbf{Q_{lsp,r}}$, and $\mathbf{Q_{lsp}}$ is regular. □

Proposition 24.4.3. Diagonalization of the noise variance matrix.
Let $lsp = \{n_y, n_x, n_y, F, G, H, J = J^T\} \in \text{LSP}$. *Consider a* $Q \in \mathbb{R}^{n_x \times n_x}_{pds}$. *Assume that* $0 \prec J + J^T - G^T QG$.

(a)Define the matrix L_e *and then,*

$$L_e = \left(\begin{array}{c|c} I & 0 \\ \hline -[J+J^T-G^TQG]^{-1}[H-G^TQF] & I \end{array}\right) \in \mathbb{R}^{(n_x+n_y)\times(n_x+n_y)};$$

$$L_e^T Q_v(Q) L_e = \left(\begin{array}{c|c} D(Q) & 0 \\ \hline 0 & J+J^T-G^TQG \end{array}\right).$$

(b)Then $0 \preceq Q_v(Q)$ *if and only if* $0 \preceq D(Q)$. *Also* $0 \prec Q_v(Q)$ *if and only if* $0 \prec D(Q)$. *In fact,* $rank(Q_v(Q)) = rank(D(Q)) + n_y$.
(c)

$$\mathbf{Q}_{lsp,r} = \{Q \in \mathbb{R}^{n_x \times n_x}_{pds} \mid 0 \prec J + J^T - G^T QG,\ 0 \preceq D(Q)\}.$$

Proof. (a) This is a calculation.
(b) This follows from the assumption, from (a), and from the nonsingularity of L_e.
(c) This is a consequence of (b). □

Proposition 24.4.4. Characterization of noise variance matrices in case of independent noise components. *Let* $lsp = \{n_y, n_x, n_y, F, G, H, J = J^T\} \in \text{LSP}$ *and consider* $\mathbf{Q_{lsp}}$. *Assume that* $J \succ 0$. *Define,*

$$F_1 = F - G(J+J^T)^{-1}H,\ \ Q \in \mathbb{R}^{n_x \times n_x}_{spd},\ 0 \succ J + J^T - G^T QG,$$

$$D_1(Q) = Q - F_1^T QF_1 - H^T(J+J^T)^{-1}H - F_1^T QG(J+J^T-G^TQG)^{-1}G^TQF_1.$$

(a)Define,

$$L_e = \left(\begin{array}{c|c} I & 0 \\ \hline (J+J^T-G^TQG)^{-1}G^TQF_1 & I \end{array}\right) \in \mathbb{R}^{(n_x+n_y)\times(n_x+n_y)}_{nsng}. \text{ Then,}$$

$$L_e^T Q_{v2}(Q) L_e = \left(\begin{array}{c|c} D_1(Q) & 0 \\ \hline 0 & J+J^T-G^TQG \end{array}\right).$$

Here $Q_{v2}(Q)$ *is as defined in 24.2.5.*
(b)Assume that $0 \prec J + J^T - G^T QG$.
Then $0 \preceq Q_v(Q)$ *if and only if* $0 \preceq D_1(Q)$; $0 \prec Q_v(Q)$ *if and only if* $0 \prec D_1(Q)$; *and* $rank(Q_v(Q)) = rank(D_1(Q)) + n_y$.
(c)

$$\mathbf{Q_{lsp,r}} = \{Q \in \mathbb{R}^{n_x \times n_x}_{pds} \mid 0 \prec J + J^T - G^T QG \text{ and } 0 \preceq D_1(Q)\}.$$

Proof. (a) This follows from 24.2.5 as 24.4.3 from 24.1.1.
(b) This follows from (a) and Proposition 24.2.5.
(c) This follows from (b). □

24.5 The Boundary of the Set of State-Variance Matrices

The following notation will be used in the sequel

$$\|Q\|_2 = \left(\sup_{x \in R^n, x \neq 0} \frac{x^T Q^T Q x}{x^T x} \right)^{1/2}, \quad B(Q, \varepsilon) = \{ \overline{Q} \in \mathbb{R}_{spd}^{n_x \times n_x} | \ \|\overline{Q} - Q\|_2 \leq \varepsilon \},$$

called the L^2-ball with center the matrix Q and as radius the value ε.

Definition 24.5.1. The *boundary and the interior of the set of state variance matrices*. Let $lsp \in$ LSP and consider $\mathbf{Q_{lsp}}$. Define respectively the *boundary* and the *interior* of the set $\mathbf{Q_{lsp}}$ as the sets,

$$\begin{aligned} \partial(\mathbf{Q_{lsp}}) &= \left\{ \begin{array}{l} Q \in \mathbf{Q_{lsp}} | \forall \varepsilon \in (0, \infty), \ \exists Q_o \in B(Q, \varepsilon) \\ \text{such that } Q_o \notin \mathbf{Q_{lsp}} \end{array} \right\}, \\ \text{int}(\mathbf{Q_{lsp}}) &= \mathbf{Q_{lsp}} \cap (\partial \mathbf{Q_{lsp}})^c; \text{ then,} \\ \mathbf{Q_{lsp}} &= \partial(\mathbf{Q_{lsp}}) \cup \text{int}(\mathbf{Q_{lsp}}). \end{aligned}$$

Proposition 24.5.2. Characterization of the boundary and of the interior of the set of state variance matrices. *Let* $lsp = \{n_y, n_x, n_y, F, G, H, J = J^T\} \in$ LSP.

(a)$Q \in \partial \mathbf{Q_{lsp}}$ *if and only if* $Q_v(Q)$ *is singular.* $Q \in int(\mathbf{Q_{lsp}})$ *if and only if* $Q_v(Q) \succ 0$.
(b)Assume that $\mathbf{Q_{lsp}}$ *is regular. Then* $Q \in \partial \mathbf{Q_{lsp}}$ *if and only if* $D(Q)$ *is singular; and* $Q \in int(\mathbf{Q_{lsp}})$ *if and only if* $D(Q) \succ 0$.

Proof. (a) ($\Leftarrow$) Let $Q \in \mathbf{Q_{lsp}}$ and $Q_v(Q)$ singular. Let $z \in \mathbb{R}^{n_x+n_y}, z \neq 0$, be such that $z^T Q_v(Q) z = 0$. Suppose first that there exists a $Q_1 \in \mathbf{Q_{lsp}}$ such that $Q_v(Q_1) \succ 0$. For $\lambda \in \mathbb{R}$, let $Q_\lambda = \lambda Q_1 + (1 - \lambda) Q$. Then, for $\lambda < 0$,

$$z^T Q_v(Q_\lambda) z = \lambda z^T Q_v(Q_1) z + (1 - \lambda) z^T Q_v(Q) z < 0,$$

hence $Q_\lambda \notin \mathbf{Q_{lsp}}$. By definition of $\partial \mathbf{Q_{lsp}}$, $Q \in \partial \mathbf{Q_{lsp}}$. In the case that there does not exist a $Q_1 \in \mathbf{Q_{lsp}}$ such that $Q_v(Q_1) > 0$, it will be shown in Theorem 24.7.4 that then always $Q \in \partial \mathbf{Q_{lsp}}$.
($\Rightarrow$) Let $Q \in \partial \mathbf{Q_{lsp}}$. Suppose that $Q_v(Q)$ is nonsingular. Then there exists a $\varepsilon > 0$ such that for all $x \in \mathbb{R}^{n_x+n_y}$, $x^T Q_v(Q) x \geq \varepsilon x^T x$. Let

$$c = \| \begin{pmatrix} F^T & F^T \\ G^T & G^T \end{pmatrix} \|_2, \ \varepsilon_1 = \varepsilon / (2(1 + c^2)).$$

It will be shown that $B(Q, \varepsilon_1) \subset \mathbf{Q_{lsp}}$, which is a contradiction of the assumption that $Q \in \partial \mathbf{Q_{lsp}}$. Let $Q_1 \in B(Q, \varepsilon_1)$. Then,

$$\begin{aligned}
&Q_v(Q_1) - Q_v(Q)\\
&= \begin{pmatrix} Q_1 - Q - F^T(Q_1 - Q)F & -F^T(Q_1 - Q)G \\ -G^T(Q_1 - Q)F & -G^T(Q_1 - Q)G \end{pmatrix}\\
&= \begin{pmatrix} Q_1 - Q & 0 \\ 0 & 0 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} F & G \\ F & G \end{pmatrix}^T \begin{pmatrix} Q_1 - Q & 0 \\ 0 & Q_1 - Q \end{pmatrix} \begin{pmatrix} F & G \\ F & G \end{pmatrix},\\
&\|Q_v(Q_1) - Q_v(Q)\|_2\\
&\leq \|Q_1 - Q\|_2 + \frac{1}{2}c^2\|Q_1 - Q\|_2^2 \leq \varepsilon_1 + \varepsilon_1 c^2 = \varepsilon_1(1 + c^2) = \frac{1}{2}\varepsilon.
\end{aligned}$$

The definition of the norm implies then that, for $x \in \mathbb{R}^{n_x+n_y}$,

$$\begin{aligned}
&x^T(Q_v(Q_1) - Q_v(Q))^2 x \leq (\frac{1}{2}\varepsilon)^2 x^T x, \text{ hence,}\\
&x^T(Q_v(Q_1) - Q_v(Q))x \geq -1/2\ \varepsilon\ x^T x. \text{ Finally,}\\
&x^T Q_v(Q_1)x = x^T Q_v(Q)x + x^T(Q_v(Q_1) - Q_v(Q))x\\
&\geq \varepsilon x^T x - \frac{1}{2}\varepsilon x^T x = \frac{1}{2}\varepsilon x^T x \geq 0,
\end{aligned}$$

hence $Q_v(Q_1) \succeq 0$ and $Q_1 \in \mathbf{Q_{lsp}}$. The second statement follows from the first and the definition of int$(\mathbf{Q_{lsp}})$.
(b) This follows from (a) and Proposition 24.4.3.(a). □

Corollary 24.5.3. *Let $lsp = \{n_y, n_x, n_y, F, G, H, J = J^T\} \in$ LSP. Assume that $J \succ 0$. Let $F_1 = F - G(J + J^T)^{-1}H$.*

(a)$Q \in \partial\mathbf{Q_{lsp}}$ if and only if $Q_{v,lsp1}(Q)$ is singular, and $Q \in int(\partial\mathbf{Q_{lsp}})$ if and only if $Q_{v,lsp1}(Q) \succ 0$. Here $Q_{v,lsp1}(Q)$ is as defined in 24.2.5.
(b)Assume that $\mathbf{Q_{lsp}}$ is regular. Then $Q \in \partial\mathbf{Q_{lsp}}$ if and only if $D_1(Q)$ is singular; and $Q \in int(\mathbf{Q_{lsp}})$ if and only if $0 \prec D_1(Q)$. Here $D_1(Q)$ is as defined in 24.4.4.

Proof. (a) This follows from 24.5.2.(a) and Proposition 24.2.5.
(b) This follows from 24.5.2.(a) and Proposition 24.4.4. □

Proposition 24.5.4. *Let $lsp = \{n_y, n_x, n_y, F, G, H, J = J^T\} \in$ LSP and $\mathbf{Q_{lsp}} \neq \emptyset$.*

(a)Then $Q \in \partial\mathbf{Q_{lsp}}$ and $0 \prec Q$ if and only if $Q^{-1} \in \partial\mathbf{Q_{lsdp}}$ and $0 \prec Q^{-1}$.
(b)Then $Q \in int(\mathbf{Q_{lsp}})$ and $0 \prec Q$ if and only if $Q^{-1} \in int(\mathbf{Q_{lsdp}})$ and $0 \prec Q^{-1}$.

Proof. (a) If either condition holds then it follows from Propostion 24.2.4 that,

$$\begin{aligned}
&\begin{pmatrix} Q & 0 \\ 0 & I \end{pmatrix} Q_v(Q^{-1}) \begin{pmatrix} Q & 0 \\ 0 & I \end{pmatrix} = Q_{v_d}(Q); \text{ note that then,}\\
&Q \in \partial\mathbf{Q_{lsp}} \text{ and } 0 \prec Q\\
&\Leftrightarrow Q_v(Q) \text{ singular, and } 0 \prec Q \text{ by Proposition 24.5.2.(a),}\\
&\Leftrightarrow Q_{v_d}(Q^{-1}) \text{ singular and } 0 \prec Q^{-1} \ \Leftrightarrow\ Q^{-1} \in \mathbf{Q_{lsdp}}.
\end{aligned}$$

(b)

$$\begin{aligned}
& Q \in \text{int}(\mathbf{Q_{lsp}}) \text{ and } 0 \prec Q \\
\Leftrightarrow\ & 0 \prec Q_v(Q) \text{ and } 0 \prec Q \text{ by Proposition 24.5.2.(a),} \\
\Leftrightarrow\ & 0 \prec Q_{v_d}(Q^{-1}) \text{ and } 0 \prec Q^{-1} \ \Leftrightarrow\ Q^{-1} \in \text{int}(\mathbf{Q_{lsdp}}).
\end{aligned}$$

□

24.6 Singular Boundary Matrices

The boundary of the set of state variances contains several elements which play a role in the theory. The reader should recall that the set of state variances is a convex set.

Definition 24.6.1. *Singular elements of the boundary of the set of state variance matrices.* Let $lsp = \{n_y, n_x, n_y, F, G, H, J = J^T\} \in \text{LSP}$ and consider $\mathbf{Q_{lsp}}$.

(a)The set of *singular boundary matrices of* $\mathbf{Q_{lsp}}$ is defined as

$$\partial\mathbf{Q}_{\mathbf{lsp},\mathbf{s}} = \{Q \in \partial\mathbf{Q_{lsp}} | \text{rank}(Q_v(Q)) = \text{rank}(J + J^T - G^T QG)\}. \tag{24.9}$$

(b)The set of *singular boundary matrices of the regular part of* $\mathbf{Q_{lsp}}$ is defined as

$$\partial\mathbf{Q}_{\mathbf{lsp},\mathbf{r},\mathbf{s}} = \{Q \in \mathbf{Q}_{\mathbf{lsp},\mathbf{r}} \cap \partial\mathbf{Q_{lsp}} | \text{rank}(Q_v(Q)) = \text{rank}(J + J^T - G^T QG) = n_y\}. \tag{24.10}$$

Proposition 24.6.2. Characterization of singular elements of the boundary of the set of state variance matrices. *Let* $lsp = \{n_y, n_x, n_y, F, G, H, J = J^T\} \in \text{LSP}$.

(a)$\partial\mathbf{Q}_{\mathbf{lsp},\mathbf{r},\mathbf{s}} = \{Q \in \mathbb{R}_{pds}^{n_x \times n_x} | \ J + J^T - G^T QG \succ 0, D(Q) = 0\}$.
(b)Assume in addition that $J + J^T \succ 0$. *Then,*
$\partial\mathbf{Q}_{\mathbf{lsp},\mathbf{r},\mathbf{s}} = \{Q \in \mathbb{R}_{pds}^{n_x \times n_x} | \ J + J^T - G^T QG \succ 0, \ \ D_1(Q) = 0\}$.

Proof. (a) By Proposition 24.4.3, $\text{rank}(Q_v(Q)) = \text{rank}(D(Q)) + n_y$, and $Q_v(Q) \succeq 0$ if and only if $D(Q) \succeq 0$. The result then follows.
(b) This follows from Proposition 24.4.4.(b). □

Proposition 24.6.3. *Let* $lsp = \{n_y, n_x, n_y, F, G, H, J = J^T\} \in \text{LSP}_{min}$.
Assume that $\mathbf{Q_{lsp}} \neq \emptyset$*, and that* $J \succ 0$.
Then $Q^-, Q^+ \in \partial\mathbf{Q}_{\mathbf{lsp},\mathbf{r},\mathbf{s}}$ *hence* $D(Q^-) = 0 = D(Q^+)$.

Proof. This follows from Theorem 23.4.2. □

This result shows also that in general the algebraic Riccati equation $D(Q) = 0$ does not have a unique solution because $D(Q^-) = D(Q^+)$ and from the assumption follows that $Q^- \prec Q^+$.

Proposition 24.6.4. *Let* $lsp = \{n_y, n_x, n_y, F, G, H, J = J^T\} \in \text{LSP}$ *and assume that* $\mathbf{Q_{lsp}} \neq \emptyset$. *Then,*

$$Q \in \partial \mathbf{Q}_{\mathbf{lsp},\mathbf{r},\mathbf{s}},\ 0 \prec Q \ \Leftrightarrow\ Q^{-1} \in \partial \mathbf{Q}_{\mathbf{lsdp},\mathbf{r},\mathbf{s}},\ 0 \prec Q^{-1}.$$

If either condition holds then $J + J^T - G^T Q G = J + J^T - H Q H^T$.

Proof.

$$\begin{aligned}
& Q \in \partial \mathbf{Q}_{\mathbf{lsdp},\mathbf{r},\mathbf{s}}\ 0 \prec Q \\
\Leftrightarrow\ & \text{rank}(Q_v(Q)) = \text{rank}(J + J^T - G^T Q G) = n_y,\ 0 \prec Q \\
\Leftrightarrow\ & \text{rank}(Q_{v_d}(Q)) = \text{rank}(J + J^T - H Q^{-1} H^T) = n_y,\ 0 \prec Q \\
& \text{because of Proposition 24.2.4, and because} \\
& J + J^T - G^T Q G = J + J^T - G^T Q Q^{-1} Q G = J + J^T - H Q^{-1} H^T, \\
\Leftrightarrow\ & Q \in \partial \mathbf{Q}_{\mathbf{lsdp},\mathbf{r},\mathbf{s}} \text{ and } 0 \prec Q^{-1}.
\end{aligned}$$

□

Definition 24.6.5. *Extremal elements of the boundary of the set of state variance matrices*. Let $lsp = \{n_y, n_x, n_y, F, G, H, J = J^T\} \in \text{LSP}$. Then $Q \in \mathbf{Q_{lsp}}$ will be called an *extremal element of the set of state-variances* $\mathbf{Q_{lsp}}$, if it is so in the sense of convex analysis; equivalently,
if $Q_1, Q_2 \in \mathbf{Q_{lsp}}$, $\lambda \in (0,1)$, and $Q = \lambda Q_1 + (1-\lambda) Q_2 \in \mathbf{Q_{lsp}}$ imply that $Q_1 = Q_2$.

Problem 24.6.6. (a)Determine all extremal elements of $\mathbf{Q_{lsp}}$.
(b)Determine the structure of the extremal elements of $\mathbf{Q_{lsp}}$.

Problem 24.6.7. Let $lsp = \{n_y, n_x, n_y, F, G, H, J = J^T\} \in \text{LSP}_{min}$.
If $Q \in \partial \mathbf{Q}_{\mathbf{lsp},\mathbf{r},\mathbf{s}}$ then Q is an extremal element of the set $\partial \mathbf{Q}_{\mathbf{lsp},\mathbf{r},\mathbf{s}}$.

The above conjecture is true for the associated continuous-time problem, see [3, Prop. 4.11, p. 85]. For the continuous-time case, the converse is apparently not true. For the discrete-time case, the solution to Problem 24.6.7 is not known to the author. The following result is a partial answer.

Proposition 24.6.8. The line connecting two regular and singular boundary state-variance matrices belongs to the set of regular state-variance matrices.
Let $lsp = \{n_y, n_x, n_y, F, G, H, J = J^T\} \in \text{LSP}$, *and assume that* $n_y < n_x$. *Assume in addition that* $\mathbf{Q_{lsp}} \neq \emptyset$. *Let* $Q_1, Q_2 \in \partial \mathbf{Q}_{\mathbf{lsp},\mathbf{r},\mathbf{s}}$, $\lambda \in (0,1)$, $Q_\lambda = \lambda Q_1 + (1-\lambda) Q_2$. *Then* $Q_\lambda \in \mathbf{Q}_{\mathbf{lsp},\mathbf{r}}$.

Proof. Because $Q_1, Q_2 \in \partial \mathbf{Q}_{\mathbf{lsp},\mathbf{r},\mathbf{s}}$, $\text{rank}(Q_v(Q_1)) = n_y = \text{rank}(Q_v(Q_2))$. Then $\lambda \in (0,1)$ and $Q_v(Q_\lambda) = \lambda Q_v(Q_1) + (1-\lambda) Q_v(Q_2)$, imply that $\text{rank}(Q_v(Q_\lambda)) \leq 2 n_y < n_x + n_y$ by $n_y < n_x$, hence $Q_v(Q_\lambda)$ is singular and from Proposition 24.5.2.(a) follows that $Q_\lambda \in \partial \mathbf{Q_{lsp}}$.

Because $Q_1,\ Q_2 \in \partial \mathbf{Q}_{\mathbf{lsp},\mathbf{r},\mathbf{s}}$, $0 \prec J + J^T - G^T Q_1 G$ and $0 \prec J + J^T - G^T Q_2 G$,

$$\begin{aligned}
& J + J^T - G^T Q_\lambda G = \lambda [J + J^T - G^T Q_1 G] + (1-\lambda)[J + J^T - G^T Q_2 G] \succ 0, \\
\Rightarrow\ & Q_\lambda \in \mathbf{Q}_{\mathbf{lsp},\mathbf{r}}.
\end{aligned}$$

□

Proposition 24.6.9. Sufficient condition for strictly-positive-definite state-variance matrices. *Consider* $lsp = \{n_y, n_x, n_y, F, G, H, J = J^T\} \in \text{LSP}_{min}$ *and* $\mathbf{Q_{lsp}}$. *Assume that: (1)* $\mathbf{Q_{lsp}} \neq \emptyset$*, (2)* $J \succ 0$*, and (3) that* $(F - G(J+J^T)^{-1}H, (J+J^T)^{-\frac{1}{2}}H)$ *is an observable pair.*

Then $\mathbf{Q_{lsp}}$ *is strictly positive definite, or, equivalently, for any* $Q \in \mathbf{Q_{lsp}}$*,* $Q \succ 0$.

Proof. Let $F_1 = F - G(J+J^T)^{-1}H$, $H_1 = (J+J^T)^{-\frac{1}{2}}H$. By Proposition 24.6.3 and Proposition 24.6.2.a,

$$0 \prec J + J^T - G^T Q^- G,$$
$$0 = Q^- - F^T Q^- F - (H^T - F^T Q^- G)(J + J^T - G^T Q^- G)^{-1}(H^T - F^T Q^- G)^T.$$

From 24.6.2.b then follows that,

$$Q^- - F_1^T Q^- F_1 - H_1^T H_1 - F_1^T Q^- G(J + J^T - G^T Q^- G)^{-1} G^T Q^- F_1 = 0.$$

Let $L_x \in \mathbb{R}_{nsng}^{n_x \times n_x}$ be such that

$$L_x^T L_x = H_1^T H_1 + F_1^T Q^- G(J + J^T - G^T Q^- G)^{-1} G^T Q^{-1} F_1 \succ 0.$$

The assumption (F_1, H_1) an observable pair, the definition of L_x, and Proposition 21.2.12 imply then that (F_1, L_x) is an observable pair. This and,

$$\exists\, Q^- \in \mathbb{R}_{pds}^{n_x \times n_x}, \;\; Q^- = F_1^T Q^- F_1 + L_x^T L_x, \tag{24.11}$$

imply by Theorem 22.1.2.d that $\text{spec}(F_1) \subset \text{D}_o$. Then $\text{spec}(F_1) \subset \text{D}_o$, (F_1, S) an observable pair, and $Q^- \in \mathbb{R}_{pds}^{n_x \times n_x}$ a solution of (24.11) implies by Theorem 22.1.2.(d) that $0 \prec Q^-$. Let $Q \in \mathbf{Q_{lsp}}$. By Theorem 24.3.1 $0 \prec Q^- \preceq Q$ and $\mathbf{Q_{lsp}}$ is strictly positive definite. □

24.7 The Classification of State-Variance Matrices

In this section a classification is provided of all state variance matrices of the set $\mathbf{Q_{lsp}}$ and its dual $\mathbf{Q_{lsdp}}$.

The structure of the set of matrices is simple. One starts from both the minimal matrix Q^- and the maximal matrix Q^+ of the set of state variance matrices. For any matrix $Q \in \mathbf{Q_{lsp}}$, define then the increment matrices,

$$\Delta_+(Q) = Q - Q^-, \;\; \Delta_-(Q) = Q^+ - Q.$$

Then the increments Δ_+ and Δ_- are characterized not by algebraic Riccati equations but by Lyapunov type matrix inequalities. Define then the *forward subset of state-variance matrices* and the *backward subset of state-variance matrices* by the expressions,

$$\begin{aligned}
{\mathbf{Q_{lsp}}}^{+}(Q_{-}) &= \left\{ \begin{array}{l} Q_{-} + \Delta_{+} \in \mathbf{Q_{lsp}} | \; \Delta_{+} \in \mathbb{R}^{n_x \times n_x}_{spds}, \\ \Delta_{+} \text{ satisfies conditions of Prop. 24.7.1} \end{array} \right\}, \\
{\mathbf{Q_{lsp}}}^{-}(Q_{+}) &= \left\{ \begin{array}{l} Q_{+} - \Delta_{-} \in \mathbf{Q_{lsp}} | \; \Delta_{-} \in \mathbb{R}^{n_x \times n_x}_{spds}, \; \Delta_{-} \text{ satisfies} \\ \text{conditions similar to those of Prop. 24.7.1} \end{array} \right\}; \text{ then,} \\
\mathbf{Q_{lsp}} &= {\mathbf{Q_{lsp}}}^{+}(Q_{-}) \cap {\mathbf{Q_{lsp}}}^{+}(Q_{+}).
\end{aligned}$$

The parametrization of the set of stochastic realizations described above is related to the case of stochastic realization of σ-algebras and of stochastic realization of a σ-algebra family, see Section 7.3 and Section 7.4. In the classification of those stochastic realizations a similar characterizations in terms of monotone sets are used.

Proposition 24.7.1. Characterization of forward and backward increment state variances. *Let* $lsp = \{n_y, n_x, n_y, F, G, H, J = J^T\} \in \text{LSP}$, *assume that* $\mathbf{Q_{lsp}} \neq \emptyset$ *and that it is regular.*

(a)Assume that

$$\begin{aligned}
& Q_0 \in \partial \mathbf{Q_{lsp,r,s}}, \text{ and define} \\
F_0 &= F - G[J + J^T - G^T Q_0 G]^{-1} [H^T - F^T Q_0 G]^T. \text{ Then} \\
& (Q_0 + \Delta Q) \in \mathbf{Q_{lsp}} \text{ and } \Delta Q \succ 0, \\
& \Leftrightarrow \Delta Q \in \mathbb{R}^{n_x \times n_x}_{spds}, \\
& (\Delta Q)^{-1} - F_0 (\Delta Q)^{-1} F_0^T - G[J + J^T - G^T Q_0 G]^{-1} G^T \succeq 0. \\
& \text{Moreover, } (Q_0 + \Delta Q) \in \partial \mathbf{Q_{lsp,r,s}} \\
& \Leftrightarrow (\Delta Q)^{-1} - F_0 (\Delta Q)^{-1} F_0^T - G[J + J^T - G^T Q_0 G]^{-1} G^T = 0.
\end{aligned}$$

(b)Assume that

$$\begin{aligned}
& Q_1 \in \partial \mathbf{Q_{lsp,r,s}} \text{ and let} \\
F_1 &= F - G[J + J^T - G^T Q_1 G]^{-1} [H^T - F^T Q_1 G]^T. \\
& \text{Then } (Q_1 - \Delta Q) \in \mathbf{Q_{lsp}} \text{ and } \Delta Q \succ 0 \\
& \Leftrightarrow \Delta Q \in \mathbb{R}^{n_x \times n_x}_{spds}, \\
& (\Delta Q)^{-1} - F_1 (\Delta Q)^{-1} F_1^T + G[J + J^T - G^T Q_1 G]^{-1} G^T \preceq 0. \\
& \text{Moreover, } (Q_1 - \Delta Q) \in \partial \mathbf{Q_{lsp,r,s}} \\
& \Leftrightarrow (\Delta Q)^{-1} - F_1 (\Delta Q)^{-1} F_1^T + G[J + J^T - G^T Q_1 G]^{-1} G^T = 0.
\end{aligned}$$

Proof. (a) Let

$$\begin{aligned}
L_e &= \begin{pmatrix} I & 0 \\ -[J+J^T-G^TQ_0G]^{-1}[H^T-F^TQ_0G]^T & I \end{pmatrix} \in \mathbb{R}^{(n_x+n_y)\times(n_x+n_y)}. \\
Q_1 &= Q_0+\Delta Q \in \mathbf{Q}_{\mathbf{lsp}},\ \Delta Q \succ 0, \\
&\Leftrightarrow Q_1 = Q_1^T \succeq 0,\ Q_v(Q_1) \succeq 0,\ \Delta Q \succ 0 \\
&\Leftrightarrow \Delta Q = (\Delta Q)^T \succ 0, \\
0 &\leq L_e^T Q_v(Q_1) L_e = L_e^T Q_v(Q_0) L_e + L_e^T (Q_v(Q_1)-Q_v(Q_0)) L_e \\
&= \begin{pmatrix} 0 & 0 \\ 0 & J+J^T-G^TQ_0G \end{pmatrix} + \begin{pmatrix} \Delta Q - F_0^T\Delta Q F_0 & -F_0^T\Delta Q G \\ -G^T\Delta Q F_0 & -G^T\Delta Q G \end{pmatrix}, \\
&\quad \text{by Proposition 24.6.2 and } Q_0 \in \partial\mathbf{Q}_{\mathbf{lsp},\mathbf{r},\mathbf{s}}, \\
&= \begin{pmatrix} \Delta Q - F_0^T\Delta Q F_0 & -F_0^T\Delta Q G \\ -G^T\Delta Q F_0 & [J+J^T-G^TQ_0G]-G^T\Delta Q G \end{pmatrix}, \\
&\Leftrightarrow \Delta Q \in \mathbf{Q}_{lsp_1}, \\
&\quad lsp_1 = \{p,n,p,F_0,G,0,J+J^T-G^TQ_0G\} \in \text{LSP}, \Delta Q \succ 0, \\
&\Leftrightarrow (\Delta Q)^{-1} \in \mathbf{Q}_{lsdp_1},\ \text{by Proposition 24.6.4}, \\
&\quad (\Delta Q)^{-1} \succ 0,\ \text{where}, \\
&\quad lsdp_1 = \{p,n,p,F_0^T,0,G^T,J+J^T-G^TQ_0G\} \in \text{LSP}, \\
&\quad \text{by Proposition 24.2.4},
\end{aligned}$$

$$\begin{aligned}
&\Leftrightarrow (\Delta Q)^{-1} = ((\Delta Q)^{-1})^T \succ 0, \\
&\quad \begin{pmatrix} (\Delta Q)^{-1} - F_0(\Delta Q)^{-1}F_0^T & G \\ G^T & J+J^T-G^TQ_0G \end{pmatrix} \succ 0, \\
&\Leftrightarrow (\Delta Q)^{-1} = ((\Delta Q)^{-1})^T \succ 0, \\
&\quad (\Delta Q)^{-1} - F_0(\Delta Q)^{-1}F_0^T - G[J+J^T-G^TQ_0G]G^T \succeq 0, \\
&\quad \text{by Proposition 24.4.3.} \\
&\quad \text{From the above follows that,} \\
Q_v(Q_1) &= L_e^{-T} \begin{pmatrix} \Delta Q - F_0^T\Delta Q F_0 & -F_0^T\Delta Q G \\ -G^T\Delta Q F_0 & [J+J^T-G^TQ_0G]-G^T\Delta Q G \end{pmatrix} L_e^{-1}, \\
&\quad \text{where } L_e \text{ is nonsingular. Then } Q_1 \in \partial\mathbf{Q}_{\mathbf{lsp},\mathbf{r},\mathbf{s}},\ \Delta Q \succ 0, \\
&\Leftrightarrow\ \operatorname{rank}(Q_v(Q_1)) = n_y,\ \Delta Q \succ 0,\ \text{by Def. 24.6.1}, \\
&\Leftrightarrow \Delta Q \in \partial\mathbf{Q}_{\mathbf{lsp},\mathbf{r},\mathbf{s}},\ \Delta Q \succ 0,\ \text{by the above expression}, \\
&\quad [J+J^T-G^TQ_0G]-G^T\Delta Q G = J+J^T-G^TQ_1G \succ 0, \\
&\quad \text{and by Def. 24.6.1}, \\
&\Leftrightarrow\ (\Delta Q)^{-1} \in \partial\mathbf{Q}_{\mathbf{lsp},\mathbf{r},\mathbf{s}}, \Delta Q \succ 0,\ \text{by Proposition 24.6.4}, \\
&\Leftrightarrow (\Delta Q)^{-1} = ((\Delta Q)^{-1})^T \succ 0, \\
&\quad (\Delta Q)^{-1} - F_0(\Delta Q)^{-1}F_0^T - G[J+J^T-G^TQ_0G]^{-1}G^T = 0,
\end{aligned}$$

by Def. 24.6.1 and $Q_0 \in \partial\mathbf{Q}_{\mathbf{lsp},\mathbf{r},\mathbf{s}}$.
(b) Let

$$
\begin{aligned}
L_{e2} &= \begin{pmatrix} I & 0 \\ -[J+J^T-G^TQ_1G]^{-1}[H^T-F^TQ_1G]^T & I \end{pmatrix} \in \mathbb{R}^{(n_x+n_y)\times(n_x+n_y)}.
\end{aligned}
$$

Then,

$$
\begin{aligned}
Q_0 &= Q_1 - \Delta Q \in \mathbf{Q_{lsp}},\ \Delta Q \succ 0, \\
&\Leftrightarrow Q_0 = Q_0^T \geq 0,\ V(Q_0) \geq 0,\ \Delta Q \succ 0, \\
&\Leftrightarrow \Delta Q = (\Delta Q)^T \succ 0, \\
0 &\succeq L_{e2}^T V(Q_0) L_{e2} = L_{e2}^T Q_v(Q_1) L_{e2} + L_{e2}^T (Q_v(Q_0) - Q_v(Q_1)) L_{e2} \\
&= -\begin{pmatrix} \Delta Q - F_1^T \Delta Q F_1 & -F_1^T \Delta Q G \\ -G \Delta Q F_1 & -[J+J^T-G^TQ_1G] - G^T \Delta Q G \end{pmatrix},
\end{aligned}
$$

by Proposition 24.6.2 and $Q_1 \in \partial \mathbf{Q_{lsp,r,s}}$,

$$
\begin{aligned}
&\Leftrightarrow \Delta Q \in \mathbf{Q}_{lsp_1},\ \ \Delta Q \succ 0,\ \text{where} \\
&\quad lsp_1 = \{p, n, p, F_1, G, 0, -[J+J^T-G^TQ_1G]\}, \\
&\Leftrightarrow (\Delta Q)^{-1} \in \mathbf{Q}_{lsdp_1},\ \Delta Q \succ 0,\ \text{where} \\
lsdp_1 &= \{n_y, n_x, n_y, F_1^T, 0, G^T, -[J+J^T-G^TQ_1G]\},
\end{aligned}
$$

by the analogon of 24.2.4,

$$
\begin{aligned}
&\Leftrightarrow (\Delta Q)^{-1} = ((\Delta Q)^{-1})^T \succ 0, \\
0 &\succeq \begin{pmatrix} (\Delta Q)^{-1} - F_1 (\Delta Q)^{-1} F_1^T & G \\ G^T & -[J+J^T-G^TQ_1G] \end{pmatrix}, \\
&\Leftrightarrow (\Delta Q)^{-1} = ((\Delta Q)^{-1})^T \succ 0, \\
&\quad 0 \succeq (\Delta Q)^{-1} - F_1 (\Delta Q)^{-1} F_1^T + G[J+J^T-G^TQ_1G]^{-1} G^T,
\end{aligned}
$$

by Proposition 24.5.2 and a formula analogous to (24.8). The second statement of (b) follows analogously to that of (a). □

Proposition 24.7.2. Characterization of the width of the set of state variances. *Let* $lsp = \{n_y, n_x, n_y, F, G, H, J = J^T\} \in$ LSP. *Assume that there exists a maximal and a minimal element* $Q^-, Q^+ \in \partial \mathbf{Q_{lsp,r,s}}$, *that* $\Delta Q = Q^+ - Q^- \succ 0$, *and that* $\mathbf{Q_{lsp}}$ *is regular. The matrix* ΔQ *is called the* width *of the set of state-variances* $\mathbf{Q_{lsp}}$.

(a)Then $\Delta(Q)$ *is a solution of the two Lyapunov equations,*

$$
\begin{aligned}
0 &= (\Delta Q)^{-1} - F^- (\Delta Q)^{-1} (F^-)^T - G[J+J^T-G^TQ^-G]^{-1} G^T, \\
0 &= (\Delta Q)^{-1} - F^+ (\Delta Q)^{-1} (F^+)^T + G[J+J^T-G^TQ^+G]^{-1} G^T;\ \textit{where,} \\
F^- &= F - G[J+J^T-G^TQ^-G]^{-1}[H^T - FQ^-G]^T, \\
F^+ &= F - G[J+J^T-G^TQ^+G]^{-1}[H^T - FQ^+G]^T.
\end{aligned}
$$

This can be used to compute, from the minimal matrix Q^-, *the maximal matrix by* $Q^+ = Q^- + \Delta\ Q$. *Or, from the maximal matrix, the minimal matrix by* $Q^- = Q^+ - \Delta Q$.

(b)Assume in addition that $\mathbf{Q_{lsp}}$ *is strictly positive definite. Let*

$$\begin{aligned}
\Delta_d Q &= (Q^-)^{-1} - (Q^+)^{-1}. \textit{ Then } \Delta_d Q \succ 0,\\
0 &= (\Delta_d Q)^{-1} - F_d^-(\Delta_d Q)^{-1}(F_d^-)^T - H^T[J+J^T-H(Q^-)^{-1}H^T]^{-1}H,\\
0 &= (\Delta_d Q)^{-1} - F_d^+(\Delta_d Q)^{-1}(F_d^+)^T + H^T[J+J^T-H(Q^+)^{-1}H^T]^{-1}H;\\
F_d^- &= F - [G - F(Q^+)^{-1}H^T][J+J^T-H(Q^+)^{-1}H^T]^{-1}H,\\
F_d^+ &= F - [G - F(Q^-)^{-1}H^T][J+J^T-H(Q^-)^{-1}H^T]^{-1}H.
\end{aligned}$$

Proof. (a) With $Q^+ = Q^- + \Delta Q \in \partial \mathbf{Q}_{\mathbf{lsp},\mathbf{r},\mathbf{s}}$ and $\Delta Q \succ 0$, the conclusion follows from Proposition 24.7.1.(a), while the second conclusion follows from
$Q^- = Q^+ - \Delta Q$, $\Delta Q \succ 0$ and Proposition 24.7.1.(b).
(b) Let $lsdp = \{p, n, p, F^T, H^T, G^T, J\} \in$ LSP. By Proposition 24.5.4
$Q_d^+ = (Q^-)^{-1} \in \partial \mathbf{Q}_{\mathbf{lsdp},\mathbf{r},\mathbf{s}}$, $Q_d^- = (Q^+)^{-1} \in \partial \mathbf{Q}_{\mathbf{lsdp},\mathbf{r},\mathbf{s}}$, and
$\Delta_d Q = (Q^-)^{-1} - (Q^+)^{-1} = Q_d^+ - Q_d^- \succ 0$. The conclusion then follows from (a).
□

Theorem 24.7.3. Characterization of the elements of $\mathbf{Q}_{\mathbf{lsp}}$ with respect to the minimal state variance. *Let* $lsp = \{n_y, n_x, n_y F, G, H, J = J^T\} \in \text{LSP}_{min}$. *Assume that* $\mathbf{Q}_{\mathbf{lsp}} \neq \emptyset$ *and that it is regular. Consider* $Q^- \in \mathbf{Q}_{\mathbf{lsp}}$ *and* $F^-(Q^-)$ *as defined in Proposition 24.7.2.*

Then $Q^- + \Delta Q \in \mathbf{Q}_{\mathbf{lsp}}$ *and* $\Delta Q \succ 0$ *if and only if the following conditions all hold:*

1. $\Delta Q \in \mathbb{R}_{spds}^{n_x \times n_x}$*;*
2. there exists a matrix $Q_1 \in \mathbb{R}_{pds}^{n_x \times n_x}$ *such that*
$(\Delta Q)^{-1} - F^-(\Delta Q)^{-1}(F^-)^T - G[J + J^T - G^T Q^- G]^{-1} G^T - Q_1 = 0$*;*
3. $\text{spec}(F^-) \subset \text{D}_o$.

Proof. From $lsp \in \text{LSP}_{min}$ follows that (F, G) is a controllable pair. Then so is,

$$\begin{aligned}
&(F^-, G) = (F - GK, G),\\
F - GK &= F - G[J + J^T - G^T Q^- G]^{-1}[H^T - F^T Q^- G]^T,\\
&(F^-, G[J + J^T - G^T Q^- G]^{-\frac{1}{2}}); \text{ let,}\\
LL^T &= G[J + J^T - G^T Q^- G]^{-1} G^T + Q_o, \;\; L \in \mathbb{R}^{n_x \times n_x}.
\end{aligned}$$

Using Proposition 21.2.12 one proves that (F^-, L) is a controllable pair.
($\Rightarrow$) By Proposition 24.6.3 $Q^- \in \partial \mathbf{Q}_{\mathbf{lsp},\mathbf{r},\mathbf{s}}$. The equation for $(\Delta Q)^{-1}$ then follows from Proposition 24.7.2. Then (F^-, L) a reachable pair and $(\Delta Q)^{-1}$ satisfying

$$(\Delta Q)^{-1} = F^-(\Delta Q)^{-1}(F^-)^T + LL^T, (\Delta Q)^{-1} = ((\Delta Q)^{-1})^T,$$

imply by Theorem 22.1.2 that $\text{spec}(F^-) \subset \text{D}_o$.
($\Leftarrow$) $\text{spec}(F^-) \subset \text{D}_o$ and (F^-, L) a reachable pair imply by Theorem 22.1.2 that there exists a $(\Delta Q)^{-1}$ such that

$$(\Delta Q)^{-1} = F^-(\Delta Q)^{-1}(F^-)^T + LL^T, \; (\Delta Q)^{-1} = ((\Delta Q)^{-1})^T \succ 0.$$

The conclusion then follows from 24.7.1.(a). □

Theorem 24.7.4. *Let $lsp = \{n_y, n_x, n_y, F, G, H, J = J^T\} \in \mathrm{LSP}_{min}$. Assume that $\mathbf{Q_{lsp}} \neq \emptyset$ and that $J \succ 0$. The following statements are equivalent:*

(a)The following covariance function is uniformly strictly positive-definite, see Def. 3.4.2,

$$W(t) = \begin{cases} HF^{t-1}G, & t > 0, \\ J + J^T, & t = 0, \\ G^T (F^T)^{-t-1} H^T, & t < 0. \end{cases} \tag{24.12}$$

(b)$\mathbf{Q_{lsp,r}} \neq \emptyset$ *and* $\Delta Q = Q^+ - Q^- \succ 0$*;*
(c)$\mathbf{Q_{lsp,r}} \neq \emptyset$ *and* $\mathrm{spec}(F^-) \subset \mathrm{D}_o$*;*
(d)int$(\mathbf{Q_{lsp}}) \neq \emptyset$*, or, equivalently, there exists a* $Q \in \mathbf{Q_{lsp}}$ *such that* $Q_v(Q) \succ 0$.

Proof. (a) ⇒ (b). If W is uniformly strictly positive-definite then there exists an $\varepsilon \in (0, \infty)$ such that $W - \varepsilon I$ is positive definite. By 24.5.1 there then exists a $Q \in \mathbf{Q}_{lsp_1}$ with,

$$lsp_1 = \{n_y, n_x, n_y, F, G, H, J - \frac{1}{2}\varepsilon I\} \in \mathrm{LSP}_{min}.$$

Then $(J + J^T) \succeq (J + J^T - \varepsilon I)$ and Proposition 24.2.6 imply that $\mathbf{Q}_{lsp_1} \subset \mathbf{Q_{lsp}}$. Take any $Q_1 \in \mathbf{Q}_{lsp_1}$. Then $Q_1 \succeq Q^-$, by $\mathbf{Q}_{lsp_1} \subset \mathbf{Q_{lsp}}$, where $Q^- \in \mathbf{Q_{lsp}}$. Hence $Q^+ - Q^- \succeq Q^+ - Q_1$. Let $u \in \mathbf{U}$ such that,

$$\frac{1}{2} x^T Q^+ x = \inf_{u_1 \in \mathbf{U}} \sum_{\tau=-\infty}^{-1} u_1(\tau)^T y_1(\tau) = \sum_{\tau=-\infty}^{-1} u(\tau) y(\tau).$$

Then by Def. 24.5.1

$$\begin{aligned} \frac{1}{2} x^T Q^+ x = \sum_{\tau=-\infty}^{-1} u(\tau)^T y(\tau) &= \frac{1}{2} x^T Q_1 x + \sum \begin{pmatrix} x(\tau) \\ u(\tau) \end{pmatrix}^T Q_v(Q_1) \begin{pmatrix} x(\tau) \\ u(\tau) \end{pmatrix} \\ &= \frac{1}{2} x^T Q_1 x + \varepsilon \sum_{\tau=-\infty}^{-1} u(\tau)^T u(\tau) + \\ &+ \sum \begin{pmatrix} x(\tau) \\ u(\tau) \end{pmatrix}^T \begin{pmatrix} Q_1 - F^T Q_1 F & H^T - F^T Q_1 G \\ H - G^T Q_1 F & J + J^T - G^T Q_1 G - \varepsilon I \end{pmatrix} \begin{pmatrix} x(\tau) \\ u(\tau) \end{pmatrix} \\ &\geq \frac{1}{2} x^T Q_1 x + \varepsilon \sum u(\tau)^T u(\tau), \\ &\Rightarrow \frac{1}{2} x^T (Q^+ - Q_1) x \geq \varepsilon \sum u(\tau)^T u(\tau) > 0, \end{aligned}$$

hence $Q^+ - Q^- \succeq Q^+ - Q_1 \succ 0$. Because $W - \varepsilon I$ is positive definite, by Def. 24.5.1 there exists a $Q \in \mathbf{Q}_{lsp_1}$. Then

$$J + J^T - \varepsilon I - G^T Q G \succeq 0 \Rightarrow J + J^T - G^T Q G \succeq \varepsilon I \succ 0.$$

This and $Q \in \mathbf{Q}_{lsp_1} \subset \mathbf{Q}_{\mathbf{lsp}}$ imply that $Q \in \mathbf{Q}_{\mathbf{lsp},\mathbf{r}}$ and $\mathbf{Q}_{\mathbf{lsp},\mathbf{r}} \neq \emptyset$.
(b) ⇒ (c). By Proposition 24.7.2 $\Delta Q = Q^+ - Q^- \succ 0$ satisfies

$$(\Delta Q)^{-1} = F^-(\Delta Q)^{-1}(F^-)^T + G[J + J^T - G^T Q^- G]^{-1} G^T,$$
$$(\Delta Q)^{-1} = ((\Delta Q)^{-1})^T.$$

(F,G) a controllable pair implies that $(F^-,G) = (F - GK, G)$ is a controllable pair, and so is $(F^-, G[J + J^T - G^T Q^- G]^{-1})$. This and Theorem 22.1.2 imply that $\text{spec}(F^-) \subset \text{D}_o$.
(c) ⇒ (d). Take a $Q \in \mathbf{Q}_{\mathbf{lsp},\mathbf{r}}$. Then $Q \geq Q^-$ and,

$$J + J^T - G^T Q^- G \geq J + J^T - G^T Q G \succ 0.$$

Take a matrix $Q_o \in \mathbb{R}^{n_x \times n_x}$, $\text{rank}(Q_o) = n$ such that (F^-, Q_o) is a controllable pair. Let $L \in \mathbb{R}^{n_x \times n_x}$ be such that,

$$LL^T = Q_o + G[J + J^T - G^T Q^- G]^{-1} G^T.$$

From Proposition 21.2.12 follows that (F^-, L) is a controllable pair. Then there exists an unique solution
$(\Delta Q)^{-1} \in \mathbb{R}^{n_x \times n_x}$ of the equation,

$$(\Delta Q)^{-1} = F^-(\Delta Q)^{-1}(F^-)^T + LL^T, \ (\Delta Q)^{-1} = ((\Delta Q)^{-1})^T \succ 0,$$

by Theorem 22.1.2. Because $\text{rank}(Q_o) = n$,

$$\begin{aligned} D_1((\Delta Q)^{-1}) &= (\Delta Q)^{-1} - F^-(\Delta Q)^{-1}(F^-)^T - G[J + J^T - G^T Q^- G]^{-1} G^T \\ &= Q_o \succ 0, \end{aligned}$$

hence by Proposition 24.5.2.(b) $(\Delta Q)^{-1} \in \text{int}(\mathbf{Q}_{lsdp_2})$,

$$lsdp_2 = \{p, n, p, F^T, 0, G^T, J + J^T - G^T Q^- G\}.$$

Then, by Proposition 24.7.1.(a), $(Q^- + \Delta Q) \in \text{int}(\mathbf{Q}_{\mathbf{lsp}})$.
(d) ⇒ (a). Let $Q \in \mathbf{Q}_{\mathbf{lsp}}$ be such that $Q_v(Q) \succ 0$. Then there exists an $\varepsilon \in (0,\infty)$ such that $Q_v(Q) - \varepsilon I \succ 0$. Then $Q = Q^T \succeq 0$,

$$\begin{pmatrix} Q - F^T Q F & H^T - F^T Q G \\ H - G^T Q F & J + J^T - \varepsilon I - G^T Q G \end{pmatrix} = (Q_v(Q) - \varepsilon I) + \begin{pmatrix} \varepsilon I & 0 \\ 0 & 0 \end{pmatrix} \succeq 0,$$

and $Q \in \mathbf{Q}_{lsp_1}$, $lsp_1 = \{p, n, p, F, G, H, J - \frac{1}{2}\varepsilon I\} \in \text{LSP}$. Hence $W - \varepsilon I$ is positive-definite by Def. 24.5.1 and W is uniformly strictly positive-definite. □

Problem 24.7.5. 1. Can a more explicit classification of the elements of $\mathbf{Q}_{\mathbf{lsp}}$ be given?
2. Characterize the extremal elements of $\mathbf{Q}_{\mathbf{lsp}}$.
3. Investigate the connection of the singular boundary matrices of $\mathbf{Q}_{\mathbf{lsp}}$ with the convergence properties of the Riccati equation.

The geometric structure for the continuous-time dissipation matrix inequality is explored in [7].

The computation of the elements of the set $\mathbf{Q_{lsp}}$ may be setup as follows. Two types of procedures may be distinguished:

1. procedures that determine elements in $\partial \mathbf{Q_{lsp,r,s}}$, in particular Q^-, Q^+;
2. procedures that determine elements in $\mathbf{Q_{lsp}} \cap (\partial \mathbf{Q_{lsp,r,s}})^c$.

Procedures for the first type may be found in Section 22.3. Procedures of the second type are based on Theorem 24.7.3.

24.8 Further Reading

The results of this appendix are primarily adapted from [3, Ch. 4]. In that reference the matrix inequality for continuous-time linear systems is treated while in this appendix attention is restricted to discrete-time linear systems. Additional information on the matrix inequality may be found in [4, 7]. For the relation between extremal matrices and the solution of the Riccati equation see also [5].

References

1. R.S. Bucy. The Riccati equation and its bounds. *J. Comput. Syst. Sci.*, 6:343–353, 1972. 849, 867
2. R.S. Bucy and J.M. Rodriguez-Canabal. A negative definite equilibrium and its induced cone of global existence for the Riccati equation. *SIAM J. Math. Anal.*, 3:644–646, 1972. 849, 867
3. P. Faurre, M. Clerget, and F. Germain. *Opérateurs rationnels positifs*. Dunod, Paris, 1979. 175, 180, 217, 275, 292, 310, 850, 865, 867, 877, 885
4. P. Faurre and J.P. Marmorat. Un algoritme de réalisation stochastique. *C.R. Acad. Sc. Paris*, 268:978–981, 1969. 850, 885
5. A. Lindquist, C. Martin, and G. Picci. Extreme points of Riccati equations. *IEEE Trans. Automatic Control*, 29:1034, 1984. 885
6. J.M. Rodriquez-Canabal. The geometry of the Riccati equation. *Stochastics*, 1:129–149, 1973. 849, 867
7. J.C. Willems. Least squares stationary optimal control and the algebraic Riccati equation. *IEEE Trans. Automatic Control*, 16:621–634, 1971. 525, 867, 885