Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

# System identification with information theoretic criteria

A.A. Stoorvogel and J.H. van Schuppen

Department of Operations Reasearch, Statistics, and System Theory

**BS-R9513 1995**

Report BS-R9513
ISSN 0924-0659

CWI
P.O. Box 94079
1090 GB  Amsterdam
The Netherlands

P.O. Box 94079, 1090 GB  Amsterdam (NL)
Kruislaan 413, 1098 SJ  Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

# System Identification with Information Theoretic Criteria

A.A. Stoorvogel[*]

*Department of Mathematics and Computing Science, Eindhoven University of Technology*
*P.O. Box 513, 5600 MB Eindhoven, The Netherlands*
wscoas@win.tue.nl

J.H. van Schuppen

*CWI*
*P.O. Box 94079, 1090 GB Amsterdam, The Netherlands*
J.H.van.Schuppen@cwi.nl

**Abstract**

Attention is focused in this paper on the approximation problem of system identification with information theoretic criteria. For a class of problems it is shown that the criterion of mutual information rate is identical to the criterion of exponential-of-quadratic cost and to $H_\infty$ entropy. In addition the relation between the likelihood function and divergence is explored. As a consequence of these relations a parameter estimator is derived by four methods for the approximation of a stationary Gaussian process by the output of a Gaussian system. The unified framework for approximation problems of system identification formulated in this paper seems extremely useful.

*AMS Subject Classification (1991):* 93B30, 93B36, 93E12, 93E20, 94A17.
*Keywords and Phrases:* System Identification, Information Theory, Parameter Estimation, Maximum Likelihood Method, Linear-Exponential-Quadratic-Gaussian Control, $H_\infty$ Optimal Control.
*Note*: The text of this report will appear in the proceedings of the NATO Advanced Study Institute 'From Identification to Learning' that was held in Como, Italy in August 1994.

## 1 Introduction

The purpose of this paper is to clarify and to explore the relationship between several approximation criteria in system identification.

The original motivation for the research reported on in this paper was to clarify the relation between LEQG optimal stochastic control and $H_\infty$ optimal control with an entropy criterion, and to explore the use of this relation in system identification. The investigation

[*] The research of dr. A.A. Stoorvogel has been made possible by a fellowship of the Royal Netherlands Academy of Sciences and Arts. His cooperation with the second author is supported in part by CWI.

lead to a study of information theoretic criteria. The unified framework that appeared seems quite useful for the theory of system identification.

A motivating problem of this paper is the choice of an approximation criterion for system identification. Up to recently the approximation criteria used mainly in system identification of stochastic processes were:

($i$) The likelihood function (maximum likelihood method).

($ii$) A quadratic cost function (least squares method).

For stationary Gaussian processes these criteria are identical. The resulting algorithms for parameter estimation with these criteria are well known.

The approximation criteria considered in this paper are:

($i$) Mutual information rate.

($ii$) A weighted likelihood function and divergence rate between the measure associated with a simple model and that associated with an element in the model class.

In the following sections these criteria are discussed in detail.

It will be shown that the criterion of mutual information rate is related to that of the LEQG optimal stochastic control problem and to that of an $H_\infty$ optimal control problem with an entropy criterion. Mutual information is related to the information theoretic concept of divergence. The divergence concept is the same as the Kullback-Leibler pseudo-distance, which in turn is related to the likelihood function.

The discussion of the approximation criteria is rather general. Thus the discussion is formulated in terms of stationary Gaussian processes but can easily be extended to other classes of processes.

In summary, in this paper several approximation criteria for system identification of stationary Gaussian processes will be introduced and their relationship will be discussed. Subsequently several parameter estimation problems are posed and solved.

The character of this paper is expository. The paper is primarily addressed to doctoral students in engineering, mathematics, and econometrics. The reader is assumed to be familiar with system and control theory, and with probability at a first year graduate level. The paper contains many definitions and elementary results that are known although not available in a concise and compact publication. The body of the paper contains the main results while the appendices contain details on concepts and theorems. A brief conference version of this paper appeared as [68].

A description of the contents by section follows. Section 2 contains a brief problem formulation. Parameter estimation by the approximation criterion of mutual information is the subject of Section 3 and by the approximation criterion of the likelihood function and divergence is the subject of Section 4 Section 5 presents concluding remarks. Appendix A introduces concepts from probability theory and stochastic processes and Appendix B concepts from system theory. Appendix C contains definitions from information theory, Appendix D formulas of information measures for Gaussian random variables, and Appendix E formulas of information measures for stationary Gaussian processes. Appendix F summarizes results for the LEQG optimal stochastic control problem and Appendix G the $H_\infty$ optimal control problem with an entropy criterion.

## 2 Problem formulation

System identification is a topic of the research area of systems and control. The problem of system identification is to obtain from data a mathematical model in the form of a dynamic system for a phenomenon. Examples of applications are system identification of a wind turbine, of a gas boiler, and of the flow of benzo-a-pyreen through the human body.

System identification of a phenomenon often proceeds by the following procedure:

(*i*) Physical modeling and selection of a model class of dynamic systems.

(*ii*) Experiment design, data collection, and preprocessing of the data.

(*iii*) Check on the identifiability of the parametrization of the model class.

(*iv*) Selection of an element in the model class that is an approximation of the data.

(*v*) Evaluation of the model determined and, possibly, redoing one or more of the previous steps.

In this paper attention is restricted to step (iv) of the above procedure, the selection of an element in the model class that is an approximation of the data.

In this paper attention is restricted to phenomena that, after preliminary physical or domain modeling, can be modeled by a stationary Gaussian process. See the appendices for concepts and terminology used in the body of the paper. It is well known from stochastic realization theory for stationary Gaussian processes that such a process can under certain conditions be represented as the output of a Gaussian system in state space form

$$\begin{aligned} x(t+1) &= Ax(t) + Kv(t), \\ y(t) &= Cx(t) + v(t), \end{aligned}$$

where $v : \Omega \times T \to \mathbb{R}^p$ is a Gaussian white noise process. Such a process can also be represented by an ARMA (Auto-Regressive-Moving-Average) representation of the form

$$\sum_{i=0}^{n} a_i y(t-i) = \sum_{j=0}^{m} b_j v(t-j).$$

Below attention is sometimes restricted to an AR (Auto-Regressive) representation of the form

$$y(t) + \sum_{i=1}^{n} a_i y(t-i) = v(t). \tag{2.1}$$

With the notation

$$\theta = (-a_1, \ldots, -a_n)^{\mathrm{T}}, \quad \phi(t) = (y(t-1), \ldots, y(t-n))^{\mathrm{T}},$$

this representation may be rewritten as

$$y(t) = \phi(t)^{\mathrm{T}} \theta + v(t). \tag{2.2}$$

**Problem 2.1** Parameter estimation problem for time-invariant Gaussian systems. *Consider observations of a stationary Gaussian process with values in $\mathbb{R}^p$. Consider the model class of time-invariant Gaussian systems*

$$\begin{aligned} x(t+1) &= Ax(t) + Kv(t), \\ y(t) \quad &= Cx(t) + v(t), \end{aligned} \tag{2.3}$$

*or, properly restricted, in AR representation,*

$$y(t) = \phi(t)^T\theta + v(t). \tag{2.4}$$

*Select an element in the model class that approximates the observations by one or more of the approximation criteria mentioned in the following sections.*

# 3 Approximation with mutual information

Successively this section presents the topics of mutual information, a parameter estimation problem, the relation between mutual information, LEQG, and $H_\infty$ entropy, parameter estimation by LEQG optimal stochastic control, and parameter estimation by the $H_\infty$ method.

## 3.1 Mutual information

There follows an introduction to the concept of mutual information. For a detailed treatment and references see the Appendices C, D, and E.

The mutual information of two discrete-valued random variables $x : \Omega \to X = \{a_1, \ldots, a_n\}$ and $y : \Omega \to Y = \{b_1, \ldots, b_m\}$, is defined by the formula

$$J(x,y) = \sum_{i=1}^{n}\sum_{j=1}^{m} p_{xy}(a_i,b_j) \ln\left(\frac{p_{xy}(a_i,b_j)}{p_x(a_i)p_y(b_j)}\right), \tag{3.1}$$

where

$$p_{xy}(a_i,b_j) = P(\{x = a_i, y = b_j\}), \quad p_x(a_i) = P(\{x = a_i\}), \quad p_y(b_j) = P(\{y = b_j\}).$$

The mutual information of two random variables describes the dependence between the associated probability measures. In general $0 \leq J(x,y) \leq +\infty$ and $J(x,y) = 0$ if and only if $x$ and $y$ are independent.

For random variables with continuous values the mutual information is defined as the limit of the mutual information of discrete valued random variables, in which the discrete random variables are required to approximate the continuous valued random variables. In case the random variables $x, y$ are jointly Gaussian random variables with a strictly positive definite variance, or $(x,y) \in G(0,Q)$ with $Q = Q^{\mathrm{T}} > 0$ where

$$Q = \begin{pmatrix} Q_x & Q_{xy} \\ Q_{xy}^{\mathrm{T}} & Q_y \end{pmatrix},$$

then

$$J(x,y) = -\frac{1}{2} \ln \left( \frac{\det Q}{\det Q_x \det Q_y} \right). \tag{3.2}$$

For stationary processes the corresponding concept is the mutual information rate. Let $x : \Omega \times T \to \mathbb{R}^m$, $y : \Omega \times T \to \mathbb{R}^p$, be stationary stochastic processes on $T = \mathbb{Z}$. The *mutual information rate* of $x$ and $y$ is defined by the formula

$$J(x,y) = \limsup_{n \to \infty} \frac{1}{2n+1} J\left( x|_{[-n,n]}, y|_{[-n,n]} \right), \tag{3.3}$$

where $x|_{[-n,n]}$ denotes the random variable defined by the restriction of the process $x$ to the interval $\{-n, \ldots, -1, 0, 1, \ldots, n\}$.

Let for $T = \mathbb{Z}$, $x : \Omega \times T \to \mathbb{R}^m$ and $y : \Omega \times T \to \mathbb{R}^p$ be two jointly stationary Gaussian processes that admit a spectral density. Denote their joint spectral density function by

$$\hat{W} : \mathbb{C} \to \mathbb{C}^{(m+p)\times(m+p)}, \quad \hat{W} = \begin{pmatrix} \hat{W}_x & \hat{W}_{xy} \\ \hat{W}_{xy} & \hat{W}_y \end{pmatrix}.$$

It follows from [24] that the mutual information of the processes $x, y$ is invariant under scaling by nonsingular linear systems. Let $S_x : \mathbb{C} \to \mathbb{C}^{m\times m}$ and $S_y : \mathbb{C} \to \mathbb{C}^{p\times p}$ be minimal and square spectral factors of respectively $\hat{W}_x$ and $\hat{W}_y$, hence

$$\hat{W}_x(z) = S_x(z) S_x(z^{-1})^{\mathrm{T}}, \quad \hat{W}_y(z) = S_y(z) S_y(z^{-1})^{\mathrm{T}}.$$

Define

$$S(z) = S_x(z)^{-1} \hat{W}_{xy}(z) S_y(z^{-1})^{-T}.$$

The mutual information rate of these processes is given by the formula

$$J(x,y) = \frac{-1}{4\pi i} \int_{\mathbb{D}} \frac{1}{z} \ln \det[I - S(z^{-1})^{\mathrm{T}} S(z)] dz. \tag{3.4}$$

Assume that the function $S$ admits a realization as a finite-dimensional linear system of the form

$$S(z) = C(zI - A)^{-1} B + D, \;\; sp(A) \subset \mathbb{C}^-, \tag{3.5}$$

and that the following set of relations admits a unique solution $Q \in \mathbb{R}^{n\times n}$

$$Q = Q^{\mathrm{T}} \geq 0, \tag{3.6}$$

$$Q = A^{\mathrm{T}} Q A + C^{\mathrm{T}} C + (A^{\mathrm{T}} Q B + C^{\mathrm{T}} D) N^{-1} (B^{\mathrm{T}} Q A + D^{\mathrm{T}} C), \tag{3.7}$$

$$N = I - D^{\mathrm{T}} D - B^{\mathrm{T}} Q B > 0, \tag{3.8}$$

$$sp(A + BN^{-1}(B^{\mathrm{T}} Q A + D^{\mathrm{T}} C)) \subset \mathbb{C}^-. \tag{3.9}$$

Then the mutual information rate of the processes $x, y$ is given by the formula

$$J(x,y) = -\frac{1}{2} \ln \det(I - B^{\mathrm{T}} Q B - D^{\mathrm{T}} D). \tag{3.10}$$

### 3.2 A parameter estimation problem

Consider the model class of Gaussian systems in AR representation as stated in Section 2,

$$\begin{aligned} \theta(t+1) &= \quad \theta(t) + v(t), \;\; \theta(0) = \theta_0, \\ y(t) \quad &= \phi(t)^{\mathrm{T}}\theta(t) + Nw(t), \end{aligned} \tag{3.11}$$

with $\theta_0 \in G(m_0, Q_0)$, $v(t) \in G(0, I)$, $w(t) \in G(0, I)$. The problem is to determine a linear recursive parameter estimator, say,

$$\hat{\theta}(t+1) = \hat{\theta}(t) + K(t)[y(t) - \phi(t)^{\mathrm{T}}\hat{\theta}(t)], \;\; \hat{\theta}(0) = m_0, \tag{3.12}$$

such that the estimation error

$$e : \Omega \times T \rightarrow \mathbb{R}^n, \;\; e(t) = \theta(t) - \hat{\theta}(t),$$

is small according to a criterion specified below. The recursion for the estimation error is then

$$e(t+1) = [I - K(t)\phi(t)^{\mathrm{T}}]e(t) + v(t) - K(t)Nw(t), \;\; e(0) = \theta_0 - \hat{\theta}_0.$$

Assume that $e$ is such that there exists a Gaussian random process $r$, which is independent of $e$, such that $z := e + r$ is a standard white noise process, i.e. $z \in G(0, I)$.

Our objective is to determine a parameter estimator such that the mutual information rate $J(e, z)$ between $e$ and $z$ is minimized. This implies that the error signal $e$ should be as small as possible. Clearly this measure is not very intuitive but it turns out to be closely related to LEQG and $H_\infty$ parameter estimators.

In particular, assume $e$ is asymptotically stationary with asymptotic density function $R$. The assumption regarding the existence of $r$ and $z$ implies

$$0 \leq R(z) \leq I.$$

It is then easy to show that the mutual information rate of the processes $z$ and $e$ is given by

$$J(z, e) = \frac{-1}{4\pi i}\int_{\mathbb{D}} \frac{1}{z} \ln \det[I - R(z)]dz = \frac{-1}{4\pi i}\int_{\mathbb{D}} \frac{1}{z} \ln \det[I - S(z^{-1})^{\mathrm{T}}S(z)]dz \tag{3.13}$$

where $S$ is a spectral factor of $R$. Expressing the mutual information in terms of the spectral factor $S$ will be useful later on.

**Problem 3.1** *Consider the parameter estimation setting defined above, with the observation equation and the parameter model as given in (3.11). Determine a parameter estimator of the form (3.12) that minimizes the mutual information rate (3.13) between the processes $z$ and $e$.*

Problem 3.1 will not be solved directly but via LEQG optimal stochastic control and via $H_\infty$ optimal control theory.

### 3.3 Relation of mutual information, $H_\infty$ entropy, and LEQG cost

An investigation of the authors has established that the criterion of mutual information of two stationary Gaussian processes is identical to the $H_\infty$ entropy criterion and to the limit of the cost of a stochastic control problem with an exponential-of-quadratic cost function. This result is first stated as a theorem and then explained.

**Theorem 3.2** *Consider a finite-dimensional linear system*

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{3.14}$$

*that is a minimal realization of its impulse response function and such that $sp(A) \subset \mathbb{C}^-$. Denote the transfer function of this system by $S$,*

$$S(z) = C(zI - A)^{-1}B + D. \tag{3.15}$$

*Assume that the set of relations (3.6)-(3.9) admits a unique solution $Q \in \mathbb{R}^{n \times n}$. Consider the expression*

$$-\frac{1}{2}\ln(\det(I - B^TQB - D^TD). \tag{3.16}$$

*a. If $e : \Omega \times T \to \mathbb{R}^m$ and $z : \Omega \times T \to \mathbb{R}^p$ are stationary Gaussian processes with spectral density*

$$\begin{pmatrix} I & S(z) \\ S(z^{-1})^T & I \end{pmatrix}, \tag{3.17}$$

*then the mutual information of $(e, z)$ is given by the expression (3.16).*

*b. The $H_\infty$ entropy of the system (3.14), defined by:*

$$J = \frac{-1}{4\pi i}\int_{\mathbb{D}} \frac{1}{z}\ln\det\left[I - S(z^{-1})^TS(z)\right] dz, \tag{3.18}$$

*is equal to the expression (3.16).*

*c. Let $u : \Omega \times T \to \mathbb{R}^m$ be a stationary Gaussian process in (3.14) with $u(t) \in G(0, I)$. Then*

$$\lim_{t\to\infty} \frac{1}{t}\ln E\left[\exp\left(\frac{1}{2}\sum_{s=0}^{t} y(s)^Ty(s)\right)\right]$$

*is equal to the expression (3.16).*

**Proof** a. See Theorem E.3. b. See Theorem G.1. c. See Proposition F.3. □

The discovery of the relation between the criterion of LEQG and the $H_\infty$ entropy has been published by K. Glover and J.C. Doyle [26]. That the $H_\infty$ entropy criterion is identical to mutual information is new as far as the authors know. The $H_\infty$ entropy criterion for continuous-time systems was introduced by K. Glover and D. Mustafa in [27] and has been used extensively since then in $H_\infty$ optimal control theory. In that paper the concept of $H_\infty$ entropy is referred to the paper [8] by D.Z. Arov and M.G. Krein. In the latter paper the formula (3.18) is presented with the explanation that this formula 'has a definite entropy sense' under reference to a publication by Kolmogorov, Gelfand, and Yaglom. In fact, the formula (3.18) is the mutual information of two stationary Gaussian processes that was apparently first analyzed in [24]. The expression of mutual information provided in [24] is in a geometric formulation from which (3.18) may be derived. The formula for the mutual information in the case of scalar processes is stated in [57]. The authors therefore conclude that $H_\infty$ entropy is actually mutual information.

## 3.4 Parameter estimation with an exponential-of-quadratic cost

A short introduction to the LEQG optimal stochastic control problem follows. A detailed treatment may be found in Appendix F. Consider the Gaussian stochastic control system

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) + Mv(t), \\ y(t) &= C_1x(t) + D_1u(t) + Nv(t), \\ z(t) &= C_2x(t) + D_2u(t), \end{aligned} \tag{3.19}$$

where $v$ is Gaussian white noise with $v(t) \in G(0, V_1)$. A *control law* is a collection of maps $g = \{g_t, t \in T\}$, $g_t : Y^t \times U^t \rightarrow U$ such that the input $u(t)$ is given by

$$u(t) = g_t(y(0), \ldots, y(t-1), u(0), \ldots, u(t-1)).$$

Let $G$ be the set of control laws. Consider the cost function on a finite horizon $J_{t_1} : G \rightarrow \mathbb{R}$

$$J_{t_1}(g) = E\left[c \exp\left(\frac{1}{2}c\sum_{t=0}^{t_1} z(t)^{\mathrm{T}}z(t)\right)\right], \tag{3.20}$$

for $c \in \mathbb{R}$, $c \neq 0$. The LEQG optimal stochastic control problem is then to determine a control law $g^* \in G$ such that

$$J_{t_1}(g^*) = \inf_{g\in G} J_{t_1}(g). \tag{3.21}$$

Assume that we apply a controller $g$ to the stochastic system (3.19) such that the closed loop system is linear and time-invariant. We can then determine the limit of the cost function

$$\frac{1}{t}E\left[c \exp\left(\frac{1}{2}c\sum_{s=0}^{t} z(s)^{\mathrm{T}}z(s)\right)\right]. \tag{3.22}$$

as $t \rightarrow \infty$. In Appendix F it is proven that

$$\begin{aligned} &\lim_{t\rightarrow\infty} \frac{1}{t} \ln E\left[c \exp\left(\frac{1}{2}c\sum_{s=0}^{t} z(s)^{\mathrm{T}}z(s)\right)\right] \\ &\qquad = -\frac{1}{2}\ln\det(V_1^{-1} - cM^{\mathrm{T}}QM - cN^{\mathrm{T}}N) + \frac{1}{2}\ln\det V_1^{-1}. \end{aligned} \tag{3.23}$$

In (3.23) $Q$ is the solution of a set of relations similar to (3.6)-(3.9).

Next the parameter estimation problem is posed in terms of the LEQG problem.

**Problem 3.3** *Consider the Gaussian system*

$$\begin{aligned} \theta(t+1) &= \quad\;\; \theta(t) + v(t), \;\; \theta(0) = \theta_0, \\ y(t) \quad &= \phi(t)^T\theta(t) + w(t), \end{aligned} \tag{3.24}$$

*where $T = \{0, 1, \ldots, t_1\}$, $\theta_0 : \Omega \to \mathbb{R}^n$, $\theta_0 \in G(m_0, Q_0)$, $v : \Omega \to \mathbb{R}^n$ and $w : \Omega \to \mathbb{R}^p$ are Gaussian white noise processes with $v(t) \in G(0, Q_v)$, $w(t) \in G(0, Q_w)$ with $Q_w > 0$, $\theta_0, v, w$ are independent, and $\phi : \Omega \times T \to \mathbb{R}^{n \times p}$ is as specified in (2.2).*

*Determine the process $\hat{\theta} : \Omega \times T \to \mathbb{R}^n$ and a recursion for it such that $\hat{\theta}(t)$ is $F^y_{t-1}$ measurable for all $t \in T$, and such that the following expression is minimized*

$$E\left[c \exp\left(\frac{1}{2}c\sum_{s=0}^{t_1} z(s)^T z(s)\right)\right], \tag{3.25}$$

*where $z(t) = C[\theta(t) - \hat{\theta}(t)]$, $c \in \mathbb{R}$, $c \neq 0$.*

**Theorem 3.4** *Consider Problem 3.3. Define the functions $K : T \to \mathbb{R}^{n \times p}$, $Q : T \to \mathbb{R}^{n \times n}$ by the recursions:*

$$K(t) = Q(t)\begin{pmatrix}\phi(t) & C^T\end{pmatrix} S(t)^{-1}\begin{pmatrix} I \\ 0\end{pmatrix}, \tag{3.26}$$

$$Q(t+1) = Q(t) + Q_v - Q(t)\begin{pmatrix}\phi(t) & C^T\end{pmatrix} S(t)^{-1}\begin{pmatrix}\phi(t)^T \\ C\end{pmatrix} Q(t), \tag{3.27}$$

$$Q(0) = Q_0, \tag{3.28}$$

$$S(t) = \begin{pmatrix}\phi(t)^T \\ C\end{pmatrix} Q(t)\begin{pmatrix}\phi(t) & C^T\end{pmatrix} + \begin{pmatrix} Q_w & 0 \\ 0 & \frac{-1}{c}I\end{pmatrix}. \tag{3.29}$$

*Assume that for all $t \in T$ we have $Q(t) > 0$. The optimal LEQG parameter estimator is then specified by the recursion*

$$\hat{\theta}(t+1) = \hat{\theta}(t) + K(t)[y(t) - \phi(t)^T\hat{\theta}(t)]. \tag{3.30}$$

*If in addition $Q_v = 0$ then*

$$Q(t+1)^{-1} = Q(t)^{-1} + \phi(t)Q_w^{-1}\phi(t)^T - cC^TC, \tag{3.31}$$

$$K(t) = Q(t)\phi(t)\left(\phi(t)^T Q(t)\phi(t) + Q_w\right)^{-1}. \tag{3.32}$$

The details of the proof of Theorem 3.4 will not be provided in this paper because of space limitations. The problem is related to Problem F.1. It differs from that problem in that Problem 3.3 has output matrix, $\phi(t)$, that depends on the past observations. Therefore a conditional version of Problem F.1 is obtained. The result of Problem F.1 may be found in [32]. The second author of this paper is preparing a publication that contains the solution to the conditional version of Problem F.1. As pointed out in Appendix F, P. Whittle first solved the discrete time case of Problem F.1 but for a system representation slightly different from that used in this paper.

### 3.5 Parameter estimation with $H_\infty$ entropy

A short introduction to the $H_\infty$ optimal control problem follows. A detailed treatment may be found in Appendix G Consider the linear control system

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) + Mv(t), \\ y(t) &= C_1x(t) + D_1u(t) + Nv(t), \\ z(t) &= C_2x(t) + D_2u(t), \end{aligned} \tag{3.33}$$

where $v$ is this time an unknown *deterministic* signal. As before, a control law is a collection of maps $g = \{g_t, t \in T\}$, $g_t : Y^t \times U^t \to U$ such that the input $u(t)$ is given by

$$u(t) = g_t(y(0), \ldots, y(t-1), u(0), \ldots, u(t-1)).$$

Let $G$ be the set of control laws. Consider the cost function $J : G \to \mathbb{R}$

$$J(g) = \sup_{v,x_0} \frac{||z||^2_{2,t_1}}{||v||^2_{2,t_1} + x_0^{\mathrm{T}} R^{-1} x_0} \tag{3.34}$$

where

$$||v||^2_{2,t_1} := \sum_{s=0}^{t_1} v(t)^{\mathrm{T}} v(t). \tag{3.35}$$

Note that $R$ plays a role similar to the covariance of the initial condition for the LEQG problem. The larger $R$, the more uncertainty we have for the initial condition. If $t_1 = \infty$ then we must constrain $v$ to $\ell_2$, the class of signals for which the infinite sum (3.35) converges. The $H_\infty$ optimal control problem is then to determine a control law $g^* \in G$ such that

$$J(g^*) = \inf_{g \in G} J(g). \tag{3.36}$$

This problem turns out to be very hard. Hence instead we will look for a control law $g$ such that:

$$J(g) < c^{-1}.$$

This is equivalent to minimizing the following criterion

$$J_c(g) = \sup_{v,x_0} c||z||^2_{2,t_1} - ||v||^2_{2,t_1} - x_0^{\mathrm{T}} R^{-1} x_0.$$

As a matter of fact $J(g) < \infty$ if $J_c(g) < c^{-1}$. We will actually solve the problem to find $g^*$ such that

$$J_c(g^*) = \inf_{g \in G} J_c(g). \tag{3.37}$$

If $t_1 = \infty$, $x(0) = 0$ $(R = 0)$, and both the system and the controller are time-invariant, then this problem has another interpretation. Let us define

$$J_e(g) = \frac{-1}{4\pi i} \int_{\mathbb{D}} \frac{1}{z} \ln \det \left[ I - cS^{\mathrm{T}}(z^{-1}) S(z) \right] dz$$

where $S$ is the closed loop transfer matrix from $v$ to $z$. Then $g^*$ satisfies (3.37) if and only if

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

Next the parameter estimation problem is posed in terms of an $H_\infty$ control problem.

**Problem 3.5** *Consider the Gaussian system*

$$\begin{aligned} \theta(t+1) &= \quad\theta(t) + v(t), \;\; \theta(0) = \theta_0, \\ y(t) &= \phi(t)^T\theta(t) + w(t), \end{aligned} \tag{3.38}$$

*where $T = \{0, 1, \ldots, t_1\}$ and $\phi$ is as specified in (2.2).*

*Determine the process $\hat{\theta} : T \rightarrow \mathbb{R}^n$ and a recursion for it such that for all $t \in T$ we have that $\hat{\theta}(t)$ is a function of past measurements $y(0), \ldots, y(t-1)$ and such that the following expression is minimized*

$$\sup_{w,v,x_0} \left( -x_0^T Q_0^{-1} x_0 + \sum_{t=0}^{t_1} cz(t)^T z(t) - v(t)^T Q_v^{-1} v(t) - w(t)^T Q_w^{-1} w(t) \right) \tag{3.39}$$

*where $z(t) = C[\theta(t) - \hat{\theta}(t)]$, $c \in \mathbb{R}$, $c \neq 0$, $Q_v = Q_v^T > 0$, $Q_w = Q_w^T > 0$.*

**Proposition 3.6** *Consider Problem 3.5. There exists a $\hat{\theta}$ such that the supremum (3.39) is finite if and only if there exists a matrix $Q$ satisfying (3.27) and (3.28). Moreover, in that case the optimal estimator is given by (3.30).*

# 4 Approximation with likelihood and divergence

In this section the reader is presented successively with text on the approximation with likelihood, divergence, the relation of likelihood and divergence, approximation with divergence, and parameter estimation by divergence minimization.

## 4.1 Approximation with the likelihood function

In this subsection a particular recursive weighted maximum likelihood problem will be formulated and solved.

The maximum likelihood method for parameter estimation has been proposed and analyzed by Sir Ronald Fisher, see [21, 22]. In most lectures and most textbooks the likelihood function is defined by example only. In general the likelihood function may be defined as the Radon-Nikodym derivative of the measure of the observations with respect to the measure with respect to which the observations are independent in discrete-time or to the measure with respect to which the observation process is an independent increment process in continuous-time. The likelihood function for a Gaussian or ARMA representation is presented in [33, 63].

**Problem 4.1** *Consider the parameter estimation model of the first paragraph of Section 3, with the representation*

$$\begin{aligned} \theta(t+1) &= \quad \theta(t), \qquad\qquad \theta(0) = \theta, \\ y(t) \quad &= \phi(t)^T\theta(t) + w(t). \end{aligned} \tag{4.1}$$

*Let $c \in \mathbb{R}$, $c \neq 0$, $\theta \in G(0, Q_0)$, $Q_0 = Q_0^T > 0$. Suppose that at $t \in T$ the parameter estimates $\{\hat{\theta}(s), s = 0, \ldots, t\}$ have been determined. Choose the next parameter estimate $\hat{\theta}(t+1) \in \mathbb{R}^n$ as the solution of the optimization problem*

$$\sup_{\theta \in \mathbb{R}^n} \exp\left(\frac{1}{2}c\sum_{s=0}^{t} \|\theta - \hat{\theta}(s)\|^2\right) f(\bar{y}(0), \ldots, \bar{y}(t), \theta), \tag{4.2}$$

*where $f$ is the joint density function of the observation process $\{y(s), s = 1, \ldots, t\}$ and $\{\bar{y}(s), s = 1, \ldots, t\}$ are the numerical values of the observations. The density function $f$ depends on the parameter vector $\theta$.*

**Theorem 4.2** *Consider Problem 4.1 with $Q_w = I$. Assume that the model is such that $Q(t)$ is well-defined by*

$$Q(t+1)^{-1} = Q(t)^{-1} + \phi(t)Q_w^{-1}\phi(t)^T - cI. \tag{4.3}$$

*with $Q(0) = Q_0$ and $Q(t) > 0$ for all $t \in T$. Then the optimal estimate at $t \in T$ is given by the recursion*

$$\hat{\theta}(t+1) = \hat{\theta}(t) + K(t)[\bar{y}(t) - \phi(t)^T\hat{\theta}(t)], \ \hat{\theta}(0) = 0, \tag{4.4}$$

*where*

$$K(t) = Q(t)\phi(t)[\phi(t)^TQ(t)\phi(t) + 1]^{-1}. \tag{4.5}$$

*Note that the parameter estimator (4.4) is identical to that of Theorem 3.4.*

**Proof:** The logarithm of the criterion is, up to a term that does not depend on $\theta$, equal to

$$2h(\theta) = -\theta^{\mathrm{T}}Q_0^{-1}\theta - \sum_{s=0}^{t} \|\bar{y}(s) - \phi(s)^{\mathrm{T}}\theta\|^2 + c\sum_{s=0}^{t-1} \|\theta - \hat{\theta}(s)\|^2. \tag{4.6}$$

By assumption

$$2d^2h(\theta)/d\theta^2 = ctI - Q_0^{-1} - \sum_{s=0}^{t} \phi(s)\phi(s)^{\mathrm{T}} = -Q(t)^{-1} - cI < 0.$$

The first order conditions then yield:

$$Q(t)^{-1}\theta = \sum_{s=0}^{t} \bar{y}(s)\phi(s) - c\sum_{s=0}^{t-1} \hat{\theta}(s),$$

and after simple calculations one gets (4.4). □

## 4.2 Divergence

The Kullback-Leibler pseudo-distance or divergence is a function that measures the relation between probability measures. For system identification it is an important approximation criterion. There follows a brief introduction to divergence, the full details may be found in the Appendices C, D, and E

Consider the measurable space $(\Omega, F)$ and two probability measures $P_1, P_2$ defined on it. Let $Q$ be a $\sigma$-finite measure on $(\Omega, F)$ such that $P_1$ and $P_2$ are absolutely continuous with respect to $Q$, say with $r_1 = dP_1/dQ$ and $r_2 = dP_2/dQ$. Such a measure always exists. The divergence of $P_1, P_2$ is defined by the formula

$$D(P_1 \| P_2) = E_Q \left[ r_1 \ln(\frac{r_1}{r_2}) I_{(r_2>0)} \right]. \tag{4.7}$$

It may be shown that the definition of $D(P_1\|P_2)$ does not depend on the measure $Q$ chosen, that for all $P_1, P_2$ we have $D(P_1\|P_2) \geq 0$, and $D(P_1\|P_2) = 0$ iff $P_1 = P_2$. The triangle inequality is not satisfied by $D$ hence it is not a distance but a pseudo-distance. S. Kullback and R.A. Leibler in [46] introduced this concept which is therefore also known as the Kullback-Leibler pseudo-distance. The term divergence is used in information theory.

Divergence and mutual information for measures induced by two real valued random variables are related by

$$J(x, y) = D(P_{xy} \| P_x \times P_y).$$

Let $G(m_1, Q_1)$ and $G(m_2, Q_2)$ be two Gaussian measures on $\mathbb{R}^n$. Assume that $Q_1, Q_2 > 0$. The divergence between these measures is given by the formula (see D.5),

$$\begin{aligned} &2D\left(G(m_1, Q_1) \| G(m_2, Q_2)\right) = \\ &\quad - \ln\left(\frac{\det Q_1}{\det Q_2}\right) + \operatorname{tr}([Q_2^{-1} - Q_1^{-1}]Q_1) + (m_2 - m_1)^{\mathrm{T}} Q_2^{-1}(m_2 - m_1). \end{aligned} \tag{4.8}$$

Consider next two jointly stationary Gaussian processes $x_1, x_2 : \Omega \times T \to \mathbb{R}^n$ on $T = \mathbb{Z}$. Assume that they have a mean value function that is zero and that the joint processes admit a spectral density. Under additional conditions, see Theorem E.5, the divergence between the measures associated with these processes is given by the formula

$$D(P_1\|P_2) = -\frac{1}{2} \ln \det(B^{\mathrm{T}} Q B + D^{\mathrm{T}} D) + \frac{1}{2} \operatorname{tr}(B^{\mathrm{T}} R B + D^{\mathrm{T}} D - I), \tag{4.9}$$

where $Q \in \mathbb{R}^{n \times n}$ is the solution of an algebraic Riccati equation and $R \in \mathbb{R}^{n \times n}$ is the solution of a Lyapunov equation.

## 4.3 Relation of likelihood function and divergence

The likelihood function as an approximation criterion is related to the approximation criterion of divergence discussed in the previous subsection. The relation between these criteria will be described for Gaussian random variables.

Consider $n$ real valued observations. The model class of the variable considered is the class of Gaussian random variables with values in $\mathbb{R}$, say $G(m, q)$, $m \in \mathbb{R}$, $q \in \mathbb{R}$, $q > 0$.

The observations are assumed to be independent. The measure of these observations and their representation are given by $y : \Omega \times T \to \mathbb{R}^n$, $y \in G(me, Q)$, where

$$e = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^n, \quad Q = qI.$$

The density function of $y$ with respect to Lebesgue measure is

$$p(v; me, Q) = \frac{1}{\sqrt{(2\pi)^n \det Q}} \exp\left(-\frac{1}{2}(v - me)^{\mathrm{T}} Q^{-1} (v - me)\right).$$

The likelihood function $L : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$ is therefore equal to

$$L(m, q; \bar{y}) = \frac{1}{\sqrt{(2\pi)^n \det Q}} \exp\left(-\frac{1}{2}(\bar{y} - me)^{\mathrm{T}} Q^{-1} (\bar{y} - me)\right),$$

where $\bar{y} \in \mathbb{R}^n$ is the vector with the numerical values of the observations. As is well known, the maximum likelihood estimate of the parameters $m, q$ are given by the formulas

$$\hat{m} = \frac{1}{n}\sum_{k=1}^{n} \bar{y}_k, \quad \hat{q} = \frac{1}{n}\sum_{k=1}^{n} (\bar{y}_k - \hat{m})^2.$$

The density function of a Gaussian measure with $\hat{m}, \hat{q}$ as parameters is denoted by $p(., \hat{m}, \hat{q})$.

Formula (4.8) for the divergence of two Gaussian measures on $\mathbb{R}$ reduces in this case to (assuming $q_1 > 0$ and $q_2 > 0$)

$$2D\left(p(.; m_1, q_1) \| p(., m_2, q_2)\right) = -\ln\left(\frac{q_1}{q_2}\right) - 1 + \frac{q_1}{q_2} + \frac{(m_1 - m_2)^2}{q_2}.$$

**Proposition 4.3** *Consider the likelihood function and the divergence for the Gaussian random variables defined above. Then*

$$L(m, q; \bar{y}) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(\frac{-n}{2}\left[2D\left(G(\hat{m}, \hat{q}) \| G(m, q)\right) + 1 + \ln \hat{q}\right]\right).$$

**Proof:** From the discussion above follows that

$$\ln L(m, q; \bar{y}) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}n[\ln(q) + \frac{q_1}{q}],$$
$$2D(G(\hat{m}, \hat{q}) \| G(m, q)) = -\ln(\frac{\hat{q}}{q}) - 1 + \frac{\hat{q}}{q} + \frac{(\hat{m} - m)^2}{q},$$

where

$$q_1 = \frac{1}{n}\sum_{k=1}^{n} (\bar{y}_k - m)^2.$$

Note that

$$q_1 = \frac{1}{n}\sum_{k=1}^{n} (\bar{y}_k - m)^2 = \frac{1}{n}\sum_{k=1}^{n} (\hat{y}_k - \hat{m} + \hat{m} - m)^2 = \hat{q} + (\hat{m} - m)^2,$$

hence the result. □

The conclusion from Proposition 4.3 is that maximizing the likelihood function is equivalent to minimizing the divergence between a measure in the model class and the measure associated with the maximum likelihood estimates. However, note that this fact cannot be used to find the maximum likelihood estimates since the maximum likelihood estimates $\hat{m}$ and $\hat{q}$ appear in the expression. Rather, it shows that the likelihood function is equivalent to the divergence criterion. In the example above both measures appearing in the divergence, $G(m, q)$ associated with the model class and $G(\hat{m}, \hat{q})$ associated with the maximum, are members of the model class. Another way to consider the relation between the likelihood function and divergence has been proposed by H. Akaike in [1].

Does the relation between the likelihood function and divergence extend to other classes of stochastic systems? In the maximum likelihood approach one maximizes the likelihood function with respect to the parameters of the model class. The divergence between the measure associated with the observations and the measure associated with the model class is a pseudo-distance. Therefore the maximum likelihood approach and the minimum divergence approach are related. The exponential transformation between both criteria as in Proposition 4.3 for the Gaussian case may not hold in general, although this relation does hold for a larger class of distributions than the Gaussian one.

S.-I. Amari has investigated the geometric structure of exponential families of probability distributions and explored the use of the likelihood function and of divergence for this family, see [4, 6, 5].

The minimum divergence method differs from the maximum likelihood method in that for the first method a measure must be chosen that represents the observations. In this point the maximum likelihood approach requires less.

## 4.4 Approximation with divergence

For system identification an approximation problem with the divergence criterion may be considered. The relationship between the likelihood function and divergence, see Proposition 4.3, suggests such an approach.

In system identification one is given observations. According to the procedure outlined in Section 2 one then selects a model class. In the case considered in this paper this is the class of stationary Gaussian processes which can be represented as a Gaussian system. With the observations one can associate a measure of a stationary Gaussian process. For example, one can take the measure associated with the mean value function of this process to be the zero function and the covariance function to be an estimate, say

$$\hat{W}(t) = \begin{cases} \frac{1}{t_1 - t} \sum_{s=0}^{t_1} y(s+t) y(s)^{\mathrm{T}}, & t = 0, \ldots, t_2, \\ 0, & \text{else.} \end{cases} \tag{4.10}$$

Another choice is a measure associated with an AR representation of high order.

**Problem 4.4** Approximation of a stationary Gaussian process by a time-invariant Gaussian system according to the divergence criterion. *Consider given a probability measure $P_1$ on the space of Gaussian processes with values in $\mathbb{R}^p$ and zero mean value function. Consider the model class $GS\Sigma(n)$ of Gaussian systems up to order $n$ and the associated* *set of probability measures $\{P(\theta), \theta \in GS\Sigma(n)\}$ on the output processes of such systems. Solve the approximation problem*

$$\inf_{\theta \in GS\Sigma(n)} D(P_1 \| P(\theta)). \tag{4.11}$$

A solution to Problem 4.4, if it exists, is a Gaussian system of which the probability measure on the output process minimizes the divergence to the probabiblity measure associated with the observations. Note that the probability measure associated with the observations may not be in the model class, for example because the covariance estimate is not a rational function. Problem 4.4 is therefore equivalent to the determination of an element in the model class that minimizes the divergence to the measure associated with the observations.

The principle of approximation by the divergence criterion is used in the literature. It has apparently first been proposed by H. Akaike in [1] who explored its use in [2, 3] and who derived the Akaike Information Criterion (AIC) from it. Other publications on this approach are those of I. Csiszár, [14, 15, 17] and, in the case of the ARMA model class, [20, 53].

The following model reduction problem for Gaussian random variables is of interest to system identification. In system identification a technique is used called *subspace methods* that is closely related to the following problem. It is expected that solution of this problem will be useful to subspace methods. The problem was first formulated in [69, Subsec. 3.8].

**Problem 4.5** Model reduction for a pair of Gaussian random variables. *Let $y_1 : \Omega \to \mathbb{R}^{p_1}$, $y_2 : \Omega \to \mathbb{R}^{p_2}$ with $(y_1, y_2) \in G(0, S)$,*

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{12}^T & S_{22} \end{pmatrix}, \quad \text{rank}(S_{12}) = n_1.$$

*Consider the model class, for $n_2 \in Z_+$, $n_2 < n_1$,*

$$\mathbf{G}(n_2) = \{G(0,Q) \text{ on } \mathbb{R}^{p_1+p_2} | \text{rank}(Q_{12}) = n_2, \}, \tag{4.12}$$

$$Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{12}^T & Q_{22} \end{pmatrix}, \tag{4.13}$$

*in which the components of $Q$ are compatible with the decomposition $(y_1, y_2)$. Solve for fixed $n_2 \in Z_+$, $n_2 < n_1$,*

$$\inf_{G(0,Q) \in \mathbf{G}(n_2)} D\left(G(0,S) \| G(0,Q)\right). \tag{4.14}$$

An interpretation of this problem follows. In stochastic realization theory of stochastic processes one uses the framework of past, future, and present or state. If one restricts attention to the case of Gaussian random variables, thus not of processes, then the past and the future are represented by Gaussian random variables, say $y_1, y_2$ as above. From observations one can estimate the measure of the past and future observations, say $G(0, S)$. The state space associated with the past and the future then has dimension $n_1 = \text{rank}(S_{12})$. In the model reduction problem one seeks a model or measure of the observations in which the dimension of the state space is less than that initially fixed, thus $\text{rank}(Q_{12}) = n_2 < n_1$.

The approximation criterion of divergence seems quite appropriate, also because of its relation with the likelihood function.

For initial results on this problem see [69, Subsec. 3.8]. The problem has a rich convex structure that remains to be explored.

## 4.5 Parameter estimation by divergence minimization

Consider Problem 2.1 of parameter estimation. Consider the time moment $t \in T$. Suppose that the observations $\{\bar{y}(s), s = 0, \ldots, t\}$ are available and that the estimates $\{\hat{\theta}(s), s = 0, \ldots, t-1\}$ have been chosen. The question is then how to choose $\hat{\theta}(t)$.

With the observations $\{\bar{y}(s), s = 0, \ldots, t\}$ we associate the probability measure $G(\bar{y}, I)$ and with the model class the probability measure $G(\Phi\theta, I)$ where

$$\Phi = \begin{pmatrix} \phi(0)^{\mathrm{T}} \\ \phi(1)^{\mathrm{T}} \\ \vdots \\ \phi(t)^{\mathrm{T}} \end{pmatrix}, \quad \bar{y} = \begin{pmatrix} \bar{y}(0) \\ \bar{y}(1) \\ \vdots \\ \bar{y}(t) \end{pmatrix}.$$

The unknown parameter $\theta$ has to be estimated. In this setting $\phi : T \rightarrow \mathbb{R}^{1\times n}$ is a deterministic function. Associate with the past parameter estimates the probability measure $G(\bar{\theta}, \bar{W})$ and with the model class the probability measure $G(\theta, c^{-1}I)$, where $c \in \mathbb{R}, c \neq 0$,

$$\bar{\theta} = \frac{1}{t}\sum_{s=0}^{t-1} \hat{\theta}(s), \quad \bar{W} = \frac{1}{t}\sum_{s=0}^{t-1} (\hat{\theta}(s) - \bar{\theta})(\hat{\theta}(s) - \bar{\theta})^{\mathrm{T}}.$$

**Problem 4.6** *Consider Problem 2.1 and the notation introduced above. For $t \in T$ choose the parameter estimate $\hat{\theta}(t)$ as the solution of the minimization problem*

$$\inf_{\theta \in \mathbb{R}^n} \left[ \frac{1}{t} D\left(G(\bar{y}, I) || G(\Phi\theta, I)\right) - D\left(G(\bar{\theta}, \bar{W}) || G(\theta, c^{-1}I)\right) \right]. \tag{4.15}$$

The problem formulated above is related to but different from that of the papers [34, 35].

**Theorem 4.7** *Consider Problem 4.6.*

a. *The divergence criterion of that problem is equivalent to the criterion*

$$\inf_{\theta \in \mathbb{R}^n} \frac{1}{t} \left[ ||\bar{y} - \Phi\theta||^2 - c||\theta - \bar{\theta}||^2 \right]. \tag{4.16}$$

b. *The parameter estimator that solves this problem is given by the formula*

$$\hat{\theta}(t+1) = \hat{\theta}(t) + K(t)[\bar{y}(t) - \phi(t)^T\hat{\theta}(t)], \quad \hat{\theta}(0) = 0, \tag{4.17}$$

*where the formula for the gain and the Riccati recursion are as stated in Theorem 4.2.*

*Note that the parameter estimator (4.17) is identical to the estimators of Theorem 3.4 and Theorem 4.2.*

**Proof a.** A simple calculation yields that

$$\begin{aligned}
2D\left(G(\bar{y},I)||G(\Phi\theta,I)\right) &= -\ln\left(\frac{\det I}{\det I}\right) - t + \mathrm{tr}(I^{-1}I) + ||\bar{y}-\Phi\theta||^2 \\
&= ||\bar{y}-\Phi\theta||^2, \\
2D\left(G(\bar{\theta},\bar{W})||G(\theta,c^{-1}I)\right) &= -\ln\left(\frac{\det \bar{W}}{\det(c^{-1}I)}\right) - t + \mathrm{tr}(c\bar{W}) + c||\theta-\bar{\theta}||^2, \\
\frac{2}{t}D\left(G(\bar{y},I)||G(\Phi\theta,I)\right) &- 2D(G(\bar{\theta},\bar{W})||G(\theta,c^{-1}I)) \\
&= \frac{1}{t}||\bar{y}-\Phi\theta||^2 - c||\theta-\bar{\theta}||^2 + \ln\det\bar{W} + t - c\,\mathrm{tr}\bar{W} + t\ln c.
\end{aligned}$$

Note that the last four terms depend on the observations and not on the parameter $\theta$. Hence they can be neglected in the criterion.

**b.** It will be shown that the criterion of part a of the theorem is equivalent to the criterion

$$||\bar{y}-\Phi\theta||^2 - c||\theta-\hat{\theta}||^2.$$

This will be established by showing that infimization of the functions

$$f(\theta) = ||\theta-\bar{\theta}||^2, \quad g(\theta) = \frac{1}{t}||\theta-\hat{\theta}||^2 = \frac{1}{t}\sum_{s=0}^{t-1}(\theta-\hat{\theta}(s))^{\mathrm{T}}(\theta-\hat{\theta}(s)),$$

is equivalent. Both functions are quadratic forms in $\theta$. The derivatives are

$$\begin{aligned}
\frac{df(\theta)}{d\theta} &= 2(\theta-\bar{\theta})^{\mathrm{T}}, \\
\frac{dg(\theta)}{d\theta} &= \frac{2}{t}\sum_{s=0}^{t-1}(\theta-\hat{\theta}(s))^{\mathrm{T}} = 2\theta - 2\frac{1}{t}\sum_{s=0}^{t-1}\hat{\theta}(s) = 2(\theta-\bar{\theta})^{\mathrm{T}}.
\end{aligned}$$

Hence the first order conditions are identical and therefore the optimal $\theta$ is the same for both optimization problems. This optimization problem is then given by:

$$\inf_{\theta\in\mathbb{R}^n}\frac{1}{t}\left[||\bar{y}-\Phi\theta||^2 - c||\theta-\hat{\theta}||^2\right].$$

The result then follows from Theorem 4.2. □

# 5 Concluding remarks

### What is the contribution of this paper to the theory of system identification?

The main contribution of the paper is the relation between several information theoretic criteria and their use in system identification. The concept of mutual information rate for stationary Gaussian processes is identical to $H_\infty$ entropy and to the exponential-of-quadratic cost in optimal stochastic control. The likelihood function is identical to divergence for Gaussian random variables. As a consequence of this relation a parameter estimator is presented that may be derived by four different approximation criteria. The problem concerns the approximation of observations from a stationary Gaussian process by the output of a Gaussian system of a specified order. Another product of the paper are formulas for information theoretic criteria of stationary Gaussian processes in case these processes are outputs of time-invariant Gaussian systems.

The relationship between the $H_\infty$ entropy for a finite-dimensional linear system and the exponential-of-quadratic cost for a Gaussian stochastic control system is understood at the level of the formulas. It is a question as to whether a deeper understanding is possible. By rephrasing the relation as a relation between mutual information and the exponential-of-quadratic cost, the interpretation becomes different but it is still not enlightening.

The parameter estimator derived in this paper should be tested on data. Its robustness properties require further study.

### Which problems of system identification theory require further attention?

The framework for system identification with information theoretic criteria should be explored further. For stationary Gaussian processes also the case of ARMA representations may be considered. Besides Gaussian processes also finite valued processes, with the hidden Markov model as stochastic system, and counting processes should be considered.

Theoretical problems of system identification and of realization may be considered using the framework of this paper. The relation between the likelihood function and divergence rate may be explored for model reduction of Gaussian systems. Possibly minimization of the divergence can be solved partly analytically. The approximation problem as such also requires further study.

System identification theory in general may benefit from studying other classes of dynamic systems in relation with practical problems. Examples of such classes are positive linear systems, particular classes of nonlinear systems, and the class of errors-in-variables systems. Approximation techniques for nonlinear systems, such as artificial neural nets, wavelets, and fuzzy modeling, may be explored.

## Acknowledgements

The authors thank S. Bittanti and G. Picci for the organization of the NATO Advanced Study Institute From Identification to Learning and for the invitation to the second named author to lecture on the topic of this paper. The meeting and its proceedings are very useful ways to transfer knowledge in the area of system identification.

The second named author of this paper also thanks L. Finesso of LADSEB, Padova, Italy, for a discussion on the subject of this paper and for useful references.

The research of this paper has been sponsored in part by the Commission of the European Communities through the SCIENCE Program by the project SC1*-CT92-0779.

# A Concepts from probability and the theory of stochastic processes

In this appendix concepts and notation of probability and of the theory of stochastic processes are introduced.

General mathematics notation follows. The set of integers is denoted by $\mathbb{Z}$ and the set of positive integers by $\mathbb{Z}_+$. The set of real numbers is denoted by $\mathbb{R}$ and the set of the positive real numbers by $\mathbb{R}_+ = [0, \infty)$. The $n$-dimensional vector space over $\mathbb{R}$ is denoted by $\mathbb{R}^n$ and the set of $n \times m$ matrices over this vector space by $\mathbb{R}^{n \times m}$. The transpose of a matrix $A \in \mathbb{R}^{n \times n}$ is denoted by $A^T$. The matrix $A \in \mathbb{R}^{n \times n}$ is said to be positive definite if $x^T A x \geq 0$ for all $x \in \mathbb{R}^n$ and strictly positive definite if $x^T A x > 0$ for all $x \in \mathbb{R}^n, x \neq 0$. The set of complex numbers is denoted by $\mathbb{C}$ and

$$\begin{aligned} \mathbb{C}^- &:= \{c \in \mathbb{C} \mid |c| < 1\}, \\ \mathbb{D} &:= \{c \in \mathbb{C} \mid |c| = 1\}. \end{aligned}$$

For $A \in \mathbb{R}^{n \times n}$, $sp(A) \subset \mathbb{C}$ denotes the spectrum of the matrix; in other words the set of eigenvalues.

## A.1 Probability concepts

A measurable space, denoted by $(\Omega, F)$, is defined as a set $\Omega$ and a $\sigma$-algebra $F$. A probability space is defined as a triple $(\Omega, F, P)$ where $(\Omega, F)$ is a measurable space and $P : F \to \mathbb{R}$ is a probability measure.

Let $I_A : \Omega \to \mathbb{R}$ be the indicator function of the event $A \in F$ defined by $I_A(\omega) = 1$, if $\omega \in A$ and $I_A(\omega) = 0$ otherwise. For any random variable $x$ let $F^x$ denote the smallest $\sigma$-algebra on which the random variable $x$ is measurable.

Let $P, Q$ be two probability measures on a measurable space $(\Omega, F)$. Then $P$ is said to be absolutely continuous with respect to $Q$, denoted by $P \ll Q$, if $P(A) = 0$ is implied by $Q(A) = 0$. If $P \ll Q$ then it follows from the Radon-Nikodym theorem that there exists a random variable $r : \Omega \to \mathbb{R}_+$ such that

$$P(A) = E_Q[r I_A] = \int r(\omega) I_A(\omega) dQ.$$

## A.2 Gaussian random variables

The Gaussian probability distribution function with parameters $m \in \mathbb{R}^n$ and $Q \in \mathbb{R}^{n \times n}$, with $Q$ strictly positive definite, is defined by the probability density function

$$p(v) = \frac{1}{\sqrt{(2\pi)^n \det Q}} \exp\left( -\frac{1}{2}(v - m)^T Q^{-1} (v - m) \right). \tag{A.1}$$

A random variable $x : \Omega \to \mathbb{R}^n$ is said to be a Gaussian random variable with parameters $m \in \mathbb{R}^n$ and $Q \in \mathbb{R}^{n \times n}$, $Q$ positive definite, if for all $u \in \mathbb{R}^n$

$$E[\exp(iu^T x)] = \exp\left( iu^T m - \frac{1}{2} u^T Q u \right). \tag{A.2}$$

The notation $x \in G(m,Q)$ will be used in this case. Moreover, $(x_1,\ldots,x_n) \in G(m,Q)$ denotes that, with $x = (x_1,\ldots,x_n)^{\mathrm{T}}$, $x \in G(m,Q)$. In this case $x_1,\ldots,x_n$ are said to be jointly Gaussian random variables.

A random variable with Gaussian probability distribution function is a Gaussian random variable. A Gaussian random variable does not necessarily have a Gaussian probability density function, only so in the case its variance is strictly positive definite.

In the following the geometric approach to Gaussian random variables is introduced. In this approach one considers the space that the random variables generate rather than the random variables themselves.

**Proposition A.1** *Let $x : \Omega \to \mathbb{R}^n$ $x \in G(m,Q)$, $S \in \mathbb{R}^{n_1 \times n}$.*

*a. Then $F^{Sx} \subset F^x$.*

*b. $F^{Sx} = F^x$ iff ker $Q$ = ker $SQ$.*

The above result motivates the following definition.

**Definition A.2** *Let $x : \Omega \to \mathbb{R}^n$, $x \in G$ and consider the $\sigma$-algebra $F^x$.*

*a. A basis for $F^x$ is a triple $(n_1, m_1, Q_1) \in \mathbb{N} \times \mathbb{R}^{n_1} \times \mathbb{R}^{n_1 \times n_1}$ such that there exists a $x_1 : \Omega \to \mathbb{R}^{n_1}$ satisfying $x_1 \in G(m_1, Q_1)$, $Q_1 = Q_1^T \geq 0$, and $F^x = F^{x_1}$.*

*b. A minimal basis for $F^x$ is a basis $(n_1, m_1, Q_1)$ such that* $\mathrm{rank}(Q_1) = n_1$.

*c. A basis transformation of $F^x$ is a map $x \mapsto Sx$, with $S \in \mathbb{R}^{n_1 \times n}$ such that $F^{Sx} = F^x$.*

By use of linear algebra one can, given a basis $(n_1, m_1, Q_1)$, always construct a minimal basis for $F^x$.

**Proposition A.3** *Let $x_1 : \Omega \to \mathbb{R}^{n_1}$, $x_1 \in G(0, Q_1)$, and $Q_1 = Q_1^T \geq 0$. Then there exists a $n_2 \in \mathbb{N}$ and a basis transformation $S \in \mathbb{R}^{n_2 \times n_1}$ such that, if $x_2 : \Omega \to \mathbb{R}^{n_2}$ $x_2 = Sx_1$, then $x_2 \in G(0, Q_2)$, $Q_2 = Q_2^T > 0$, and $F^{x_2} = F^{x_1}$.*

**Problem A.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 $S = block - diag(S_1, S_2)$ with $S_1, S_2$ nonsingular, leaves the spaces $F^{y_1}$, $F^{y_2}$ invariant. Therefore this operation introduces an equivalence relation on the spaces $F^{y_1}, F^{y_2}$, hence one can speak of a canonical form.

**Definition A.5** *Let $y_1 : \Omega \to \mathbb{R}^{k_1}$ and $y_2 : \Omega \to \mathbb{R}^{k_2}$ be jointly Gaussian random variables with $(y_1, y_2) \in G(0,Q)$. Then $(y_1, y_2)$ are said to be in* canonical variable form *if*

$$Q = \begin{pmatrix} I & 0 & \Lambda & 0 \\ 0 & I & 0 & 0 \\ \Lambda & 0 & I & 0 \\ 0 & 0 & 0 & I \end{pmatrix} \in \mathbb{R}^{(k_1+k_2) \times (k_1+k_2)}$$

*where $\Lambda \in \mathbb{R}^{k_{12} \times k_{12}}$, $k_{12} \in \mathbb{N}_+$, $\Lambda = diag(\lambda_1, \ldots, \lambda_{k_{12}})$, $1 \geq \lambda_1 \geq \ldots \geq \lambda_{k_{12}} > 0$. One then says that $(y_{11}, \ldots, y_{1k_1})$, $(y_{21}, \ldots, y_{2k_2})$ are the* canonical variables *and $(\lambda_1, \ldots, \lambda_{k_{12}})$ the* canonical correlation coefficients.

**Theorem A.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 $(y_1, y_2) \in G(0, Q)$. Then there exists a basis transformation $S = block - diag(S_1, S_2)$ such that with respect to the new basis $(S_1 y_1, S_2 y_2) \in G(0, Q_1)$ has the canonical variable form.*

The transformation to canonical variable form is not unique in general. The remaining invariance of the canonical variable form will not be stated here because of space limitation. For additional results on canonical variables the reader is referred to [7, 25].

### A.3 Concepts from the theory of stochastic processes

A real valued *stochastic process* $x : \Omega \times T \to \mathbb{R}$ on a probability space $(\Omega, F, P)$ and a time index set $T$ is defined to be a function such that for all $t \in T$ the map $x(., t) : \Omega \to \mathbb{R}$ is a random variable.

A stochastic process $x : \Omega \times T \to \mathbb{R}^n$ is said to be a *Gaussian process* if every finite dimensional distribution is Gaussian. Thus, if for all $m \in \mathbb{Z}_+$ and $(t_1, \ldots, t_m) \in T$ the collection of random variables $x(t_1), \ldots, x(t_m)$ is jointly Gaussian. Such a process is completely specified by its mean value function $m : T \to \mathbb{R}^n$, $m(t) = E[x(t)]$ and its covariance function $W : T \times T \to \mathbb{R}^{n \times n}$, $W(t,s) = E[(x(t) - m(t))(x(s) - m(s))^{\mathrm{T}}]$. A (discrete time) *Gaussian white noise* process with values in $\mathbb{R}^k$ and intensity $V : T \to \mathbb{R}^{k \times k}$, $V(t) = V(t)^{\mathrm{T}} \geq 0$, is a stochastic process such that (1) $\{v(t), t \in T\}$ is a collection of independent Gaussian random variables; (2) $v$ is a Gaussian process with $v(t) \in G(0, V(t))$.

A stochastic process $x : \Omega \times T \to \mathbb{R}^n$ on $T = \mathbb{Z}$ or $T = \mathbb{R}$ is said to be *stationary* if for all $m \in \mathbb{Z}_+$, $s \in T$, and $t_1, \ldots, t_m \in T$ the joint distribution of $(x(t_1), \ldots, x(t_m))$ equals the joint distribution of $(x(t_1 + s), \ldots, x(t_m + s))$. If for all $m \in \mathbb{Z}_+$ and $t_1, \ldots, t_m \in T$ the distribution function of $(x(t_1 + s), \ldots, x(t_m + s))$ converges to a limit as $s \to \infty$ then we call the process asymptotically stationary. A Gaussian process with mean value function $m : T \to \mathbb{R}^n$ and covariance function $W$ is stationary iff $m(t) = m(0)\ \forall t \in T$ and $W(t,s) = W(t+u, s+u)$, $\forall t, s, u \in T$. In this case the function $W_1 : T \to \mathbb{R}^{n \times n}$, $W_1(t) = W(t, 0)$ is also called the covariance function.

The concept of canonical correlation process has been defined for stationary Gaussian processes in analogy with the canonical variable form and canonical correlation coefficients for Gaussian random variables. That this may be done follows from the fact that many properties of a stationary Gaussian process depend only on the spaces generated by such a process. For references on this topic see [28, 42, 43, 54].

## B Concepts from system theory

This appendix contains several definitions of concepts from system theory that are needed at several places in the paper.

**Definition B.1** *A* discrete-time finite-dimensional linear dynamic system *or, by way of abbreviation, a* linear system, *is a dynamic system*

$$\sigma = \{T, \mathbb{R}^p, \mathbf{Y}, \mathbb{R}^n, \mathbb{R}^m, \mathbf{U}, \phi, r\}$$

*in which the state transition map $\phi$ and the read-out map $r$ are specified by*

$$\begin{aligned} \phi : x(t+1) &= A(t)x(t) + B(t)u(t), x(0) = x_0, \\ r : y(t) &= C(t)x(t) + D(t)u(t). \end{aligned} \tag{B.1}$$

*Here $T \subset \mathbb{Z}, m, n, p \in \mathbb{Z}_+$, $u : T \rightarrow \mathbb{R}^m, u \in \mathbf{U}$, $x_0 \in \mathbb{R}^n$, $A : T \rightarrow \mathbb{R}^{n \times n}$, $B : T \rightarrow \mathbb{R}^{n \times m}$, $C : T \rightarrow \mathbb{R}^{p \times n}$, $D : T \rightarrow \mathbb{R}^{p \times m}$, and $x : T \rightarrow \mathbb{R}^n, y : T \rightarrow \mathbb{R}^p$ are determined by the above recursions and are called respectively the* state function *and the* output function. *Such a system is called* time-invariant *if for all $t \in T$ $A(t) = A(0), B(t) = B(0), C(t) = C(0), D(t) = D(0)$. The parameters of such a system are denoted by $\{n, m, p, A, B, C, D\}$.*

**Definition B.2** *A Gaussian stochastic control system is defined by the representation*

$$\begin{aligned} x(t+1) &= A(t)x(t) + B(t)u(t) + M(t)v(t), \quad x(0) = x_0, \\ y(t) &= C(t)x(t) + D(t)u(t) + N(t)v(t), \end{aligned} \tag{B.2}$$

*where $T = \{0, 1, \ldots, t_1\}$, $t_1 \in \mathbb{Z}_+$, $X = \mathbb{R}^n$, $U = \mathbb{R}^m$, $Y = \mathbb{R}^p$, $x_0 : \Omega \rightarrow X$, $x_0 \in G(m_0, Q_0)$, $v : \Omega \times T \rightarrow \mathbb{R}^r$ is a Gaussian white noise process with for all $t \in T$, $v(t) \in G(0, V(t))$, $V : T \rightarrow \mathbb{R}^{r \times r}$, $V(t) = V(t)^T \geq 0$, the $\sigma$-algebras $F^{x_0}$ and $F^v_{t_1}$ are independent, $A : T \rightarrow \mathbb{R}^{n \times n}$, $B : T \rightarrow \mathbb{R}^{n \times m}$, $C : T \rightarrow \mathbb{R}^{p \times n}$, $D : T \rightarrow \mathbb{R}^{p \times m}$, $M : T \rightarrow \mathbb{R}^{n \times r}$, $N : T \rightarrow \mathbb{R}^{p \times r}$, $x : \Omega \times T \rightarrow X$, and $y : \Omega \times T \rightarrow Y$.*

### Stochastic realization

The problem of obtaining a representation of a stochatic process as the output of a stochastic system is called the *stochastic realization problem.* Below the weak stochastic realization problem for stationary Gaussian processes is mentioned. For references on this problem see [19] and the paper by G. Picci elsewhere in this volume.

Consider a stationary Gaussian process with mean value function equal to zero and covariance function $W : T \rightarrow \mathbb{R}^{p \times p}$. The problem is whether there exists a time-invariant Gaussian system, say of the form

$$\begin{aligned} x(t+1) &= Ax(t) + Mv(t), \\ y(t) &= Cx(t) + Nv(t), \end{aligned} \tag{B.3}$$

with $sp(A) \subset \mathbb{C}^-$ such that the covariance function of the output process $y$ equals the given covariance function. If so, then this system is said to be a *stochastic realization* of the given process. There is a necessary and sufficient condition for the existence of such a realization. When a realization exists there may be many realizations. Attention is then restricted to minimal realizations for which the dimension of the state space is as small as possible. There is a classification of all minimal stochastic realizations. In addition, there is an algorithm to construct a minimal realization.

## C Concepts from information theory

The purpose of this appendix is to describe for the reader the main concepts of information theory. In Appendix D formulas are presented for information measures of Gaussian random variables and in Appendix E formulas for information measures of stationary Gaussian processes.

Information theory originated with the work of C.E. Shannon [61, 62]. The Russian mathematicians A.N. Kolmogorov, I.M. Gelfand, and A.M. Yaglom partly developed the measure theoretic formulation of information theory, see [24]. The book [57] by M.S. Pinsker partly summarizes information theory as developed by the Russian school. The publications of A. Pérez, [55, 56], offer another measure theoretic formulation of information theory in which martingale theory is used as a technique to establish convergence results. Of more recent contributions to information theory we mention the work of I. Csiszár [16].

A recent textbook on information theory is the one by T.M. Cover and J.A. Thomas [13]. Information theory of continuous stochastic processes is treated in the book by S. Ihara [37]. Other books are [30, 45].

**Definition C.1** *Let $x : \Omega \to X$ with $X = \{a_1, a_2, \ldots, a_n\}$ be a finite valued random variable and let $p_x : X \to \mathbb{R}$ be its frequency function, $p_x(a_i) = P(\{x = a_i\})$. Define the* entropy of the finite valued random variable $x$ *as*

$$H_C(x) = \sum_{i=1}^{n} p_x(a_i) \ln \left( \frac{1}{p_x(a_i)} \right). \tag{C.1}$$

The concept of entropy of a finite valued random variable should be regarded as entropy with respect to a counting measure. Entropy is a property of a probability measure, not of a random variable. Nevertheless, the terminology of information theory is followed in referring to entropy of a random variable.

**Definition C.2** *Let $x : \Omega \to \mathbb{R}^n$ be a random variable whose probability distribution function is absolutely continuous with respect to Lebesgue measure with density $p_x : \mathbb{R}^n \to \mathbb{R}_+$. Define the* entropy *with respect to Lebesgue measure of such a random variable by*

$$H_L(x) = \int_{\mathbb{R}^n \cap S} p_x(v) \ln \left( \frac{1}{p_x(v)} \right) dv, \tag{C.2}$$

*where $S$ is the support of $p_x$.*

The entropy defined above depends on Lebesgue measure, or in general on the measure with respect to which it is defined.

**Definition C.3** *[57, p. 19] Let $(\Omega, F)$ be a measurable space consisting of a set $\Omega$ and a $\sigma$-algebra $F$. Let $P_1, P_2$ be two probability measures defined on $(\Omega, F)$. Define the* entropy *of $P_1$ with respect to $P_2$ by the formula*

$$H(P_1, P_2) = \sup_{\{E_1, \ldots, E_n\}} \sum_{i=1}^{n} P_1(E_i) \ln \left( \frac{P_1(E_i)}{P_2(E_i)} \right), \tag{C.3}$$

*where the supremum is taken over all finite measurable partitions of $\Omega$.*

Besides entropy there is the concept of mutual information.

**Definition C.4** *[24, p. 200] Let $x : \Omega \to X$, $y : \Omega \to Y$ be random variables taking values in the finite sets $X = \{x_1, \ldots, x_n\}$, $Y = \{y_1, \ldots, y_m\}$ respectively. The* mutual information of $x$ and $y$ *is defined by the formula*

$$J(x, y) = \sum_{i=1}^{n} \sum_{j=1}^{m} p_{xy}(i, j) \ln \left( \frac{p_{xy}(i, j)}{p_x(i) p_y(j)} \right), \tag{C.4}$$

*where*

$$p_{xy}(i, j) = P(\{x = x_i, y = y_j\}), \quad p_x(i) = P(\{x = x_i\}), \quad p_y(j) = P(\{y = y_j\}).$$

Mutual information of arbitrary real valued random variables may be defined by a limiting argument whereby the entropy is defined as the limit of the entropy of a discrete approximation of the real valued random variables, see [24].

**Proposition C.5** *[24, pp. 209-210]. Let $x : \Omega \to \mathbb{R}^n$, $y : \Omega \to \mathbb{R}^p$ be random variables. Assume that the probability distribution functions associated with the probability measures of $P_{xy}, P_x, P_y$ are absolutely continuous with respect to Lebesgue measure with densities denoted by $p_{xy}, p_x, p_y$. Then the mutual information, if it is finite, is given by the formula*

$$J(x, y) = \int \int p_{xy}(u, v) \ln \left( \frac{p_{xy}(u, v)}{p_x(u) p_y(v)} \right) dudv, \tag{C.5}$$

*where the integral is a Lebesgue integral.*

Mutual information of two random variables satisfies $0 \leq J(x, y) \leq +\infty$. Moreover, $J(x, y) < \infty$ if the probability measure $P_{xy}$ is absolutely continuous with respect to the product measure $P_x \times P_y$. Also, $J(x, y) = 0$ iff $x, y$ are independent random variables.

The third concept from information theory introduced here is divergence.

**Definition C.6** *Given a measurable space $(\Omega, F)$. Let*

$$\begin{aligned} F_{2s} &= \{f : \mathbb{R}_+ \to \mathbb{R} \mid ; f \textit{strictly convex and } f(1) = 0\}, \\ \mathbf{P} &= \{P : F \to \mathbb{R}_+ \mid P \textit{ is probability measure } \}. \end{aligned}$$

*For $f \in F_{2s}$ define the* pseudo-distance $d_f$ *on $\mathbf{P}$ as $d_f : \mathbf{P} \times \mathbf{P} \to \mathbb{R}$*

$$d_f(P_1, P_2) = E_Q[f(\frac{r_1}{r_2})r_2] = E_{P_2}[f(\frac{r_1}{r_2})], \tag{C.6}$$

*where $Q$ is a $\sigma$-finite measure on $(\Omega, F)$ such that*

$$P_1 \ll Q, \; \frac{dP_1}{dQ} = r_1, \qquad P_2 \ll Q, \; \frac{dP_2}{dQ} = r_2.$$

A $\sigma$-finite measure $Q$ such as used above always exists, $Q = P_1 + P_2$ will do. The definition of $d_f$ does not depend on the choice of $Q$. It can then be shown that for all $P_1, P_2$, we have $d_f(P_1, P_2) \geq 0$ and $d_f(P_1, P_2) = 0$ if and only if $P_1 = P_2$. In general $d_f$ does not satisfy the triangle inequality hence it is not a distance.

**Definition C.7** *The* divergence *or the* Kullback-Leibler pseudo-distance *on* **P** *is defined as a special case of the pseudo-distance defined above with*

$$f : \mathbb{R}_+ \to \mathbb{R},\ f(x) = \begin{cases} x \ln x, & \textit{if } x > 0, \\ 0, & x = 0, \end{cases} \tag{C.7}$$

$$D(P_1 \| P_2) = d_f(P_1, P_2) = E_Q \left[ f(\frac{r_1}{r_2}) r_2 \right] = E_{P_2} \left[ f(\frac{r_1}{r_2}) \right] = E_Q \left[ r_1 \ln(\frac{r_1}{r_2}) I_{(r_2 > 0)} \right]. \tag{C.8}$$

In general $D(P_1 \| P_2) \neq D(P_2 \| P_1)$.

Divergence is a special case of a pseudo-distance. The choice of the function $f$ specified by (C.7) seems arbitrary. Information and communication theory provide ample evidence that the choice of this function and the concepts of entropy, mutual information, and divergence derived from it are useful for engineering.

Mutual information of real-valued random variables is related to divergence of the probability measures associated with the same variables by the formula

$$J(x, y) = D(P_{xy} \| P_x \times P_y). \tag{C.9}$$

**Theorem C.8** *[57, Th. 2.4.2., p. 20] Let $(\Omega, F)$ be a measurable space with two probability measures $P_1, P_2$ defined on it. If the entropy $H(P_1, P_2)$ is finite then $P_1$ is absolutely continuous with respect to $P_2$ and*

$$H(P_1, P_2) = D(P_1 \| P_2). \tag{C.10}$$

The last stated theorem and the remark above it illustrate the relationship between entropy, mutual information, and divergence. Divergence of probability measures is the basic concept, mutual information may be derived from it, and so may entropy.

# D Information measures of Gaussian random variables

**Proposition D.1** *[13, Th. 9.4.1] Let $y : \Omega \to \mathbb{R}^k$, $y \in G(m, Q)$. Assume that $Q = Q^T > 0$. The entropy of the random variable $y$, or of the measure $G(m, Q)$, with respect to Lebesgue measure is given by the formula*

$$H_L(G(m, Q)) = \frac{1}{2} \ln \left( (2\pi e)^k \det Q \right). \tag{D.1}$$

Note that the entropy does not depend on the mean of the Gaussian distribution.

**Proposition D.2** *[24, Th. 1.2., p.209] Let $x : \Omega \to \mathbb{R}^n$, $y : \Omega \to \mathbb{R}^k$ be Gaussian random variables and $S \in \mathbb{R}^{n \times n}$. Then $J(Sx, y) \leq J(x, y)$. If in addition the matrix $S$ is nonsingular then $J(Sx, y) = J(x, y)$.*

**Proposition D.3** *[24, Th. 2.1] Let $x : \Omega \to \mathbb{R}^n$, $y : \Omega \to \mathbb{R}^k$, $(x, y) \in G(0, Q)$, $Q = Q^T > 0$, and*

$$Q = \begin{pmatrix} Q_x & Q_{xy} \\ Q_{xy}^T & Q_y \end{pmatrix}.$$

a. *The mutual information of the random variables $x, y$ is given by the formula*

$$J(x,y) = -\frac{1}{2} \ln \left( \frac{\det Q}{\det Q_x \det Q_y} \right). \tag{D.2}$$

b. *Let $\lambda_1, \ldots, \lambda_r \in (0,1)$ be the nonzero canonical correlations of the random variables $x, y$. Then*

$$J(x,y) = -\frac{1}{2} \sum_{i=1}^{r} \ln(1 - \lambda_i^2). \tag{D.3}$$

If in Proposition D.3

$$Q = \begin{pmatrix} I & Q_{xy} \\ Q_{xy}^{\mathrm{T}} & I \end{pmatrix},$$

then the mutual information is given by the formula

$$J(x,y) = -\frac{1}{2} \ln \det(I - Q_{xy}^{\mathrm{T}} Q_{xy}). \tag{D.4}$$

Because of Proposition D.2 mutual information is invariant under scaling of the individual random variables. Hence mutual information of Gaussian random variables can be expressed in terms of canonical correlations as in Proposition D.3. Such a representation was first derived in [24].

**Proposition D.4** *Let $G(m_1, Q_1)$ and $G(m_2, Q_2)$ be two Gaussian measures on $\mathbb{R}^n$. Assume that $Q_1 > 0$ and $Q_2 > 0$. The divergence between these measures is given by the formula*

$$\begin{aligned} &2D\left(G(m_1, Q_1) \| G(m_2, Q_2)\right) \\ &\quad = -\ln \left( \frac{\det Q_1}{\det Q_2} \right) + \mathrm{tr}([Q_2^{-1} - Q_1^{-1}]Q_1) + (m_1 - m_2)^T Q_2^{-1} (m_1 - m_2) &\text{(D.5)} \\ &\quad = \sum_{i=1}^{n} [\lambda_i(Q_1, Q_2) - \ln \lambda_i(Q_1, Q_2) - 1] + (m_1 - m_2)^T Q_2^{-1} (m_1 - m_2), &\text{(D.6)} \end{aligned}$$

*where $\{\lambda_i(Q_1, Q_2), i \in 1, \ldots, n\}$ are the generalized eigenvalues of $Q_1$ with respect to $Q_2$, or the solutions of the polynomial equation in $\lambda$*

$$\det(Q_2 \lambda - Q_1) = 0. \tag{D.7}$$

The authors have not found this result in the literature although it is an elementary calculation. Note the well known convex function $f(x) = x - \ln x - 1$ in (D.6).

**Proof of D.4** Both measures are absolutely continuous with respect to Lebesgue measure. Thus

$$D(G(m_1, Q_1) \| G(m_2, Q_2)) = E_L \left[ r_1 \ln \left( \frac{r_1}{r_2} \right) \right],$$

where

$$r_1(v) = \frac{1}{\sqrt{(2\pi)^n \det Q_1}} \exp\left( -\frac{1}{2}(v - m_1)^{\mathrm{T}} Q_1^{-1}(v - m_1) \right),$$

$$2 \ln \left( \frac{r_1(v)}{r_2(v)} \right) = -\ln\left( \frac{\det Q_1}{\det Q_2} \right) - (v - m_1)^{\mathrm{T}} Q_1^{-1}(v - m_1).$$

and hence we find

$$\begin{aligned} & 2D(G(m_1, Q_1) \| G(m_2, Q_2)) \\ & = 2 \int r_1(v) \ln \left( \frac{r_1(v)}{r_2(v)} \right) dv \\ & = -\ln \left( \frac{\det Q_1}{\det Q_2} \right) - \mathrm{tr}(Q_1^{-1} Q_1) + \mathrm{tr}(Q_2^{-1} Q_1) + (m_1 - m_2)^{\mathrm{T}} Q_2^{-1}(m_1 - m_2). \end{aligned}$$

The expression in terms of generalized eigenvalues is then easy to derive. □

**Proposition D.5** *Let* $x : \Omega \to \mathbb{R}^n$, $z : \Omega \to \mathbb{R}^p$, $x \in G(m_1, V_1)$, $m_1 \in \mathbb{R}^n$, $V_1 \in \mathbb{R}^{n \times n}$, $V_1 = V_1^T > 0$, $u \in \mathbb{R}^m$, $C \in \mathbb{R}^{p \times n}$, $D \in \mathbb{R}^{p \times m}$, $c \in \mathbb{R}$,

$$z = Cx + Du. \tag{D.8}$$

*Assume that*

$$V_2^{-1} := V_1^{-1} - cC^TC > 0. \tag{D.9}$$

*Then*

$$E[c \exp(\frac{1}{2} cz^T z)] = c \left( \frac{\det V_2}{\det V_1} \right)^{1/2} \exp \left[ \frac{1}{2} c \begin{pmatrix} m_1 \\ u \end{pmatrix}^T M \begin{pmatrix} m_1 \\ u \end{pmatrix} \right], \tag{D.10}$$

*where*

$$M = \begin{pmatrix} \frac{1}{c}[V_1^{-1} V_2 V_1^{-1} - V_1^{-1}] & V_1^{-1} V_2 C^T D \\ D^T C V_2 V_1^{-1} & D^T D + cD^T C V_2 C^T D \end{pmatrix}. \tag{D.11}$$

**Proof:** The proof is a tedious calculation. Let

$$r = -\left[V_1^{-1} - cC^{\mathrm{T}} C\right]^{-1} \begin{pmatrix} -V_1^{-1} & -cC^{\mathrm{T}} D \end{pmatrix} \begin{pmatrix} m_1 \\ u \end{pmatrix}. \tag{D.12}$$

Then

$$E\left[ \exp\left( \frac{1}{2} cz^{\mathrm{T}} z \right) \right] = \int \frac{1}{\sqrt{(2\pi)^n \det V_1}} \exp\left( \frac{1}{2} cz^{\mathrm{T}} z - \frac{1}{2}(v - m_1)^{\mathrm{T}} V_1^{-1} (v - m_1) \right) dv.$$

The exponent can be computed as

$$\begin{aligned} & -cz^{\mathrm{T}} z + (v - m_1)^{\mathrm{T}} V_1^{-1} (v - m_1) \\ & \quad = \begin{pmatrix} v \\ u \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} -cC^{\mathrm{T}} C & -cC^{\mathrm{T}} D \\ -cD^{\mathrm{T}} C & -cD^{\mathrm{T}} D \end{pmatrix} \begin{pmatrix} v \\ u \end{pmatrix} + \begin{pmatrix} v \\ m_1 \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} V_1^{-1} & -V_1^{-1} \\ -V_1^{-1} & V_1^{-1} \end{pmatrix} \begin{pmatrix} v \\ m_1 \end{pmatrix} \\ & \quad = (v - r)^{\mathrm{T}} (V_1^{-1} - cC^{\mathrm{T}} C)(v - r) - c \begin{pmatrix} m_1 \\ u \end{pmatrix}^{\mathrm{T}} M \begin{pmatrix} m_1 \\ u \end{pmatrix}. \end{aligned}$$

Then:

$$\begin{aligned}
&E\left[\exp\left(\tfrac{1}{2}cz^{\mathrm{T}}z\right)\right] \\
&= \int \frac{1}{\sqrt{(2\pi)^n \det V_1}} \exp\left(-\frac{1}{2}(v-r)^{\mathrm{T}}(V_1^{-1} - cC^{\mathrm{T}}C)(v-r)\right) \\
&\quad \times \exp\left(\frac{1}{2}c\begin{pmatrix} m_1 \\ u \end{pmatrix}^{\mathrm{T}} M \begin{pmatrix} m_1 \\ u \end{pmatrix}\right) dv \\
&= \left(\frac{\det(V_1^{-1} - cC^{\mathrm{T}}C)^{-1}}{\det V_1}\right)^{\frac{1}{2}} \exp\left(\frac{1}{2}c\begin{pmatrix} m_1 \\ u \end{pmatrix}^{\mathrm{T}} M \begin{pmatrix} m_1 \\ u \end{pmatrix}\right). \qquad \square
\end{aligned}$$

Suppose that in Proposition D.5 there holds $m_1 = 0$, $u = 0$, $c > 0$, and $V_1 = I$. Condition (D.9) is then equivalent to $I - cC^TC > 0$ and

$$\ln E[c^{1/2}\exp(\frac{1}{2}cz^Tz)] = -\frac{1}{2}\ln\det(I - cC^TC) + \frac{1}{2}\ln c = -\frac{1}{2}\ln\det(\frac{1}{c}I - C^TC). \tag{D.13}$$

Note the analogy of equation (D.13) with (D.4).

# E Information measures of stationary Gaussian processes

**Definition E.1** *[13, p. 63] Let $x : \Omega \to X$ be a stochastic process on $T = \mathbb{Z}$. The* entropy rate *of this process is defined by the formula*

$$H(x) = \lim_{n\to\infty} \frac{1}{n} H(x_1, x_2, \ldots, x_n), \tag{E.1}$$

*if the limit exists.*

Let $x : \Omega \times \mathbb{Z} \to \mathbb{R}$ be a stationary Gaussian process with spectral density $S$. Then the entropy rate of the process $x$ is given by

$$h(x) = \frac{1}{2}\ln(2\pi e) + \frac{1}{4\pi}\int_{-\pi}^{\pi} \ln S(e^{i\lambda})d\lambda. \tag{E.2}$$

This result is due to A.N. Kolmogorov [44].

**Definition E.2** *[29, p. 86; p.135], [30], [37, 2.1.5]. Let $x : \Omega \times T \to X$, $y : \Omega \times T \to Y$, be stochastic processes on $T = \mathbb{Z}$. The* mutual information rate *of $x$ and $y$ is defined by the formula*

$$J(x, y) = \limsup_{n\to\infty} \frac{1}{2n+1} J(x|_{[-n,n]}, y|_{[-n,n]}), \tag{E.3}$$

*where $J(x|_{[-n,n]}, y|_{[-n,n]})$ is the mutual information of the processes $x, y$ restricted to the interval $\{-n, -n+1, \ldots, -1, 0, 1, \ldots, n\}$.*

Let for $T = \mathbb{Z}$, $x : \Omega \times T \to \mathbb{R}^m$ and $y : \Omega \times T \to \mathbb{R}^p$ be two jointly stationary and Gaussian processes that admit a joint spectral density. Denote their joint spectral density function by

$$\hat{W} : \mathbb{C} \to \mathbb{C}^{(m+p)\times(m+p)}, \quad \hat{W} = \begin{pmatrix} \hat{W}_x & \hat{W}_{xy} \\ \hat{W}_{xy}^{\mathrm{T}} & \hat{W}_y \end{pmatrix}.$$

Assume that the spectral density is nonsingular. It follows from [24] that the mutual information of two stationary Gaussian processes is invariant under scaling by nonsingular finite-dimensional linear systems. Therefore transformation to the following canonical form is useful. Let $S_x : \mathbb{C} \to \mathbb{C}^{m\times m}$ and $S_y : \mathbb{C} \to \mathbb{C}^{p\times p}$ be minimal square spectral factors of respectively $\hat{W}_x$ and $\hat{W}_y$, or

$$\hat{W}_x(z) = S_x(z)S_x(z^{-1})^{\mathrm{T}}, \quad \hat{W}_y(z) = S_y(z)S_y(z^{-1})^{\mathrm{T}}.$$

Define

$$S(z) = S_x(z)^{-1}\hat{W}_{xy}(z)S_y(z^{-1})^{-T}.$$

The spectral density of the transformed process is then

$$\begin{pmatrix} I & S(z) \\ S(z^{-1})^{\mathrm{T}} & I \end{pmatrix}.$$

From the fact that $\hat{W}$ is a spectral density follows that $S(z)S(z^{-1})^{\mathrm{T}} \leq I$.

**Theorem E.3** *a. The mutual information rate of the stationary Gaussian processes defined above is given by the formula*

$$J(x,y) = \frac{-1}{4\pi i}\int_{\mathbb{D}} \frac{1}{z} \ln \det[I - S(z^{-1})^T S(z)]dz. \tag{E.4}$$

*b. Assume that the function $S$ admits a realization as a finite-dimensional linear system with*

$$S(z) = C(zI - A)^{-1}B + D, \quad sp(A) \subset \mathbb{C}^-. \tag{E.5}$$

*Assume that the following set of relations admits an unique solution $Q \in \mathbb{R}^{n\times n}$*

$$Q = Q^T \geq 0, \tag{E.6}$$
$$Q = A^TQA + C^TC + (A^TQB + C^TD)N^{-1}(B^TQA + D^TC), \tag{E.7}$$
$$N = I - B^TQB - D^TD > 0, \tag{E.8}$$
$$sp(A + BN^{-1}(B^TQA + D^TC)) \subset \mathbb{C}^-. \tag{E.9}$$

*Then*

$$J(x,y) = -\frac{1}{2}\ln \det(I - B^TQB - D^TD). \tag{E.10}$$

I.M. Gelfand and A.M. Yaglom in [24] derived a formula for the mutual information rate in a geometric formulation. In [57] a formula analogous to (E.4) is presented for the scalar case. See also [37, Ch. 5]. In the case the stationary Gaussian processes are generated by finite dimensional Gaussian systems the result of Theorem E.3 seems new.

**Proof:** a. A proof of part a. can be given along the same line as the proof of theorem E.5. However, it also automatically follows from theorem E.5 by using the relation between divergence and mutual information as given by (C.9).

b. If $\|S\|_\infty = 1$ then $J(x,y) = +\infty$. If $\|S\|_\infty < 1$ then it follows from the bounded real lemma that there exists a $Q \in \mathbb{R}^{n\times n}$ that satisfies the relations of part b. of the theorem. A realization $\tilde{\Sigma}$ of the transfer function $I - S^{\mathrm{T}}(z^{-1})S(z)$ is given by

$$\begin{aligned} \sigma^{-1}x &= A^{\mathrm{T}}x + C^{\mathrm{T}}Cp + C^{\mathrm{T}}Du, \\ \sigma p &= Ap + Bu, \\ z &= -B^{\mathrm{T}}x - D^{\mathrm{T}}Cp + (I - D^{\mathrm{T}}D)u. \end{aligned}$$

Let

$$A_x = B^{\mathrm{T}}QA + D^{\mathrm{T}}C, \quad x_n = x - QAp - QBu.$$

Then a realization of $\tilde{\Sigma}$ is given by

$$\begin{aligned} \sigma^{-1}x_n &= A^{\mathrm{T}}x_n - A_x^{\mathrm{T}}N^{-1}A_xp + A_x^{\mathrm{T}}u, \\ \sigma p &= Ap + Bu, \\ z &= -B^{\mathrm{T}}x_n - A_xp + Nu. \end{aligned}$$

It is then easy to check that the transfer matrix of $\tilde{\Sigma}$ is equal to $H^{\mathrm{T}}(z^{-1})H(z)$ where

$$H(z) = -N^{-1/2}A_x(zI - A)^{-1}B + N^{1/2}.$$

Note that $H(z)$ and $H(z)^{-1}$ are, as a consequence of (E.5) and (E.9), both analytic outside the unit disc. Then

$$\begin{aligned} J(x,y) &= \frac{-1}{4\pi i}\int_{\mathbb{D}} \frac{1}{z} \ln \det \left[I - S^{\mathrm{T}}(z^{-1})S(z)\right] dz \\ &= \frac{-1}{4\pi i}\int_{\mathbb{D}} \frac{1}{z} \ln \det \left[H^{\mathrm{T}}(z^{-1})H(z)\right] dz \\ &= \frac{-1}{4\pi i}\int_{\mathbb{D}} \frac{1}{z} \ln |\det H^{\mathrm{T}}(z^{-1})|^2 dz \\ &= \frac{-1}{4\pi}\int_{-\pi}^{\pi} \ln |\det H^{\mathrm{T}}(e^{-i\theta})|^2 d\theta \\ &= Re\frac{-1}{4\pi}\int_{-\pi}^{\pi} 2\ln \det H^{\mathrm{T}}(e^{-i\theta}) d\theta \\ &= Re\frac{-1}{2\pi i}\int_{\mathbb{D}} \frac{1}{z} \ln \det H^{\mathrm{T}}(z^{-1}) dz \\ &= Re\frac{-1}{2\pi i} 2\pi i \ln \det H^{\mathrm{T}}(\infty) \\ &= -\frac{1}{2}\ln \det(I - B^{\mathrm{T}}QB - D^{\mathrm{T}}D). \end{aligned}$$

□

**Definition E.4** *[37, 2.1.6] Let $x_1, x_2 : \Omega \times T \to \mathbb{R}^p$ be two stationary processes on $T = \mathbb{Z}$. Denote by $P_1, P_2$ the measures induced by $x_1, x_2$ respectively on $(\mathbb{R}^n)^T$. The* divergence rate *between $P_1, P_2$ is defined by the formula*

$$D(P_1 \| P_2) = \lim_{n \to \infty} \frac{1}{2n+1} D(P_1|_{[-n,n]} \| P_2|_{[-n,n]}), \tag{E.11}$$

*if the limit exists, where $P_1|_{[-n,n]}, P_2|_{[-n,n]}$ denote the restrictions of $P_1, P_2$ respectively to the time index set $\{-n, \ldots, -1, 0, 1, \ldots, n\}$.*

**Theorem E.5** *Let $y_1, y_2 : \Omega \times T \to \mathbb{R}^p$ be two jointly stationary Gaussian processes on $T = \mathbb{Z}$ with values in $\mathbb{R}^p$. Assume that they have a mean value function that is zero, that their covariance functions are denoted by $W_1, W_2 : T \to \mathbb{R}^{p \times p}$, and that they admit spectral densities, say $\hat{W}_1, \hat{W}_2 : \mathbb{C} \to \mathbb{C}^{p \times p}$ respectively.*

*a. The divergence rate between the measures induced by these processes exists and is given by the formula*

$$\begin{aligned} & D(P_1 \| P_2) \\ & = \frac{1}{4\pi} \int_{-\pi}^{\pi} \ln \left( \frac{\det \hat{W}_2(e^{i\lambda})}{\det \hat{W}_1(e^{i\lambda})} \right) + \operatorname{tr} \left( \hat{W}_2(e^{i\lambda})^{-1} [\hat{W}_1(e^{i\lambda}) - \hat{W}_2(e^{i\lambda})] \right) d\lambda, && \text{(E.12)} \\ & = \frac{1}{4\pi} \int_{-\pi}^{\pi} \operatorname{tr} \left( S^T(e^{-i\lambda}) S(e^{i\lambda}) - I \right) - \ln[\det S(e^{i\lambda})]^2 d\lambda, && \text{(E.13)} \end{aligned}$$

*where*

$$\begin{aligned} \hat{W}_1(z) &= G_1^T(z^{-1}) G_1(z), && \text{(E.14)} \\ \hat{W}_2(z) &= G_2^T(z^{-1}) G_2(z), && \text{(E.15)} \\ S(z) &= G_1(z) G_2(z)^{-1}. && \text{(E.16)} \end{aligned}$$

*b. Assume in addition that the transfer function $S$ admits a realization as a finite-dimensional linear system with a minimal realization parametrized by*

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{E.17}$$

*where $(A, B, C, D) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times p} \times \mathbb{R}^{p \times n} \times \mathbb{R}^{p \times p}$ and $sp(A) \subset \mathbb{C}^-$.*

*Assume further that there exists a matrix $Q \in \mathbb{R}^{n \times n}$ such that the following relations are satisfied*

$$\begin{aligned} & Q = Q^T \geq 0, && \text{(E.18)} \\ & Q = A^T Q A + C^T C - [A^T Q B + C^T D][D^T D + B^T Q B]^{-1} [A^T Q B + C^T D]^T, && \text{(E.19)} \\ & N = D^T D + B^T Q B > 0, && \text{(E.20)} \\ & sp(A - BN^{-1}(B^T Q A + D^T C)) \subset \mathbb{C}^-, && \text{(E.21)} \end{aligned}$$

*and that this set of relations admits an unique solution. Let $R \in \mathbb{R}^{n\times n}$ be the unique solution to the Lyapunov equation*

$$R = A^T R A + C^T C. \tag{E.22}$$

*Then the divergence rate between the measures induced by the two processes is given by the formula*

$$D(P_1\|P_2) = -\frac{1}{2}\ln\det(B^T Q B + D^T D) + \frac{1}{2}\mathrm{tr}(B^T R B + D^T D - I). \tag{E.23}$$

**Proof:** a. Consider the processes restricted to the interval $\{-m,\dots,-1,0,1,\dots,m\}$. Denote the measures associated with the processes $y_1, y_2$ by $P_1, P_2$ respectively and their restrictions to the finite interval by $P_1|_{[-m,m]}, P_2|_{[-m,m]}$. Define for $i = 1,2$ the block-Toeplitz matrix

$$R_i^m = [W_i(j-k)]_{j,k=-m,\dots,0,\dots,m}$$

From Proposition D.4 follows that

$$D(P_1|_{[-m,m]}\|P_2|_{[-m,m]}) = -\frac{1}{2}\ln\left(\frac{\det R_1^m}{\det R_2^m}\right) + \frac{1}{2}\mathrm{tr}\left([(R_2^m)^{-1} - (R_1^m)^{-1}]R_1^m\right).$$

From Szegö's limit theorem for block-Toeplitz matrices (see e.g. [58]) follows that

$$\lim_{m\to\infty}[\det R_i^m]^{1/2m} = \exp\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln\det\hat{W}_i(e^{i\lambda})d\lambda.$$

Hence

$$\lim_{m\to\infty}\frac{1}{2m}\ln\left(\frac{\det R_2^m}{\det R_1^m}\right) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\ln\left(\frac{\det\hat{W}_2(e^{i\lambda})}{\det\hat{W}_1(e^{i\lambda})}\right)d\lambda.$$

Let

$$S^m(t) = [(1-t)(R_1^m)^{-1} + t(R_2^m)^{-1}]^{-1}[(R_2^m)^{-1} - (R_1^m)^{-1}].$$

As a consequence of Szegö's limit theorem, we know that the spectrum of $R_i^m$ is contained in a region $[\bar{m}_i, \bar{M}_i]$ with $\bar{m}_i > 0$ where $\bar{m}_i, \bar{M}_i \in \mathbb{R}$ are independent of $m$. This implies that the spectrum of $S^m(t)$ is also contained in some region $[\bar{m}, \bar{M}]$ where $\bar{m}, \bar{M} \in \mathbb{R}$ are independent of $m, t$. Hence for a neighborhood of $t = 0$, $\frac{1}{m}\mathrm{tr}S^m(t)$ is a uniformly continuous function in $m$ and $t$. Therefore, making use of the formula

$$\frac{d}{dt}\log\det K(t) = \mathrm{tr}\left(K^{-1}(t)\frac{d}{dt}K(t)\right),$$

we get

$$\begin{aligned}
&\lim_{m\to\infty} \frac{1}{2m}\mathrm{tr}(R_1^m[(R_2^m)^{-1} - (R_1^m)^{-1}]) \\
&= \lim \frac{1}{2m}\mathrm{tr}S^m(0) \\
&= \lim \frac{d}{dt}\frac{1}{2m} \ln\det((1-t)(R_1^m)^{-1} + t(R_2^m)^{-1})\big|_{t=0} \\
&= \frac{d}{dt}\left(\lim \frac{1}{2m} \ln\det((1-t)(R_1^m)^{-1} + t(R_2^m)^{-1})\right)\bigg|_{t=0} \\
&= \frac{d}{dt}\Big[\lim \Big(-\ln(\det R_1^m)^{1/2m} - \ln(\det R_2^m)^{1/2m}\Big) \\
&\qquad\qquad + \ln\det((1-t)R_2^m + tR_1^m)^{1/2m}\Big)\Big]\Big|_{t=0} \\
&= \frac{d}{dt}\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\det[(1-t)\hat{W}_2(e^{i\lambda}) + t\hat{W}_1(e^{i\lambda}) \right. \\
&\qquad\qquad \left. - \ln\det\hat{W}_1(e^{i\lambda}) - \ln\det\hat{W}_2(e^{i\lambda})]d\lambda\right]\bigg|_{t=0} \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{tr}\left(\hat{W}_2^{-1}[\hat{W}_1(e^{i\lambda}) - \hat{W}_2(e^{i\lambda})]\right) d\lambda.
\end{aligned}$$

Thus

$$\begin{aligned}
&D(P_1\|P_2) \\
&\quad= \lim_{m\to\infty} \frac{1}{2m+1} D(P_1|_{[-m,m]}\|P_2|_{[-m,m]}) \\
&\quad= \frac{1}{4\pi}\int_{-\pi}^{\pi} \ln\left(\frac{\det\hat{W}_2(e^{i\lambda})}{\det\hat{W}_1(e^{i\lambda})}\right) d\lambda + \frac{1}{4\pi}\int_{-\pi}^{\pi} \mathrm{tr}\left(\hat{W}_2^{-1}[\hat{W}_1(e^{i\lambda}) - \hat{W}_2(e^{i\lambda})]\right) d\lambda \\
&\quad= \frac{1}{4\pi}\int_{-\pi}^{\pi} -\ln\det(S(e^{i\lambda})^2)d\lambda + \frac{1}{4\pi}\int_{-\pi}^{\pi} \mathrm{tr}\left(S^{\mathrm{T}}(e^{-i\lambda})S(e^{i\lambda}) - I\right) d\lambda.
\end{aligned}$$

b. A realization $\tilde{\Sigma}$ of $S^{\mathrm{T}}(z^{-1})S(z)$ is provided by

$$\begin{aligned}
\sigma^{-1}x &= A^{\mathrm{T}}x + C^{\mathrm{T}}Cp + C^{\mathrm{T}}Du, \\
\sigma p &= Ap + Bu, \\
z &= B^{\mathrm{T}}x + D^{\mathrm{T}}Cp + D^{\mathrm{T}}Du.
\end{aligned}$$

Let

$$A_x = B^{\mathrm{T}}QA + D^{\mathrm{T}}C, \quad x_n = x - QAp - QBu.$$

Then the realization of $\tilde{\Sigma}$ may be written as

$$\begin{aligned}
\sigma^{-1}x_n &= A^{\mathrm{T}}x_n + A_x^{\mathrm{T}}N^{-1}A_xp + A_x^{\mathrm{T}}u, \\
\sigma p &= Ap + Bu, \\
z &= B^{\mathrm{T}}x_n + A_xp + Nu,
\end{aligned}$$

and it is then easy to check that the transfer matrix of $\tilde{\Sigma}$ is equal to $H(z^{-1})^{\mathrm{T}} H(z)$ where

$$H(z) = N^{-1/2} A_x (zI - A)^{-1} B + N^{1/2}.$$

Note that $H(z)$ and $H(z)^{-1}$ are both analytic outside the unit disc. Then, analogously to the proof of Theorem E.3,

$$\begin{aligned} D(P_1 \| P_2) &= \frac{1}{4\pi i} \int_{\mathbb{D}} \frac{1}{z} \left( \mathrm{tr} \left( S^{\mathrm{T}}(z^{-1}) S(z) - I \right) - \ln |\det(H^{\mathrm{T}}(z^{-1})|^2 \right) dz \\ &= -\mathrm{Re}\ \ln \det H(\infty) + \frac{1}{2} \mathrm{tr}(B^{\mathrm{T}} R B + D^{\mathrm{T}} D - I) \\ &= -\frac{1}{2} \ln \det(D^{\mathrm{T}} D + B^{\mathrm{T}} Q B) + \frac{1}{2} \mathrm{tr}(B^{\mathrm{T}} R B + D^{\mathrm{T}} D - I). \end{aligned}$$

□

# F LEQG optimal stochastic control

The purpose of this appendix is to introduce the reader to the discrete-time optimal stochastic control problem with a Gaussian stochastic control problem and an exponential-of-quadratic cost function (LEQG). In addition an expression is derived for the cost in case of an infinite horizon.

The problem formulation consists of a discrete-time partially observed stochastic control system driven by Gaussian noise processes. The cost function over a finite-horizon is the expected value of the exponent of an additive form that is quadratic in the state and in the input process. The class of control laws consists of all measurable nonlinear functions of the past observations and inputs.

The history of the LEQG optimal stochastic control problem is summarized below. D.H. Jacobson formulated and solved the discrete-time and the continuous-time complete observations case of the problem in [38]. Complete observations refers to the case in which the input may depend on the current and the past state. Various special cases of the partial observations case were solved [65, 64, 47, 48] but the general case was considered as not to admit a solution in the form of a finite-dimensional control law. P. Whittle [70] solved the discrete-time partial observations LEQG optimal stochastic control problem and established the existence of a finite-dimensional control law. Additional publications on this and on related problems by Whittle are [71, Ch. 19] and [76, 72, 73, 74, 75]. The solution to the continuous-time partial observations LEQG optimal stochastic control problem was presented in [11]. Related publications on the LEQG problem are [32, 39, 59, 60]. Recently a rigorous and elegant approach to the LEQG optimal stochastic control problem has been proposed by M.R. James, see [12, 40, 41].

The LEQG optimal stochastic control problem is of interest for several reasons. Of interest to the development of stochastic control theory is the fact that a partially observed optimal stochastic control problem admits a control law with a finite-dimensional representation. It may point the way to other optimal stochastic control problems with the same property. The solution of the LEQG problem has been shown to be equivalent to the solution of a $H_\infty$-optimal control problem with the entropy criterion. Therefore the solution to the LEQG problem has certain robustness properties that are of interest in control engineering.

**Problem F.1** The discrete-time linear-exponential-quadratic-Gaussian (LEQG) optimal stochastic control problem with partial observations. *Consider the Gaussian stochastic control system*

$$\begin{aligned} x(t+1) &= A_1(t)x(t) + \ B_1(t)u(t) + M(t)v(t), \quad x(0) = x_0, \\ y(t) \quad &= C_1(t)x(t) + \ D_1(t)u(t) + N(t)v(t), \\ z(t) \quad &= C_2(t)x(t) + D_2(t)u(t), \end{aligned} \tag{F.1}$$

*where $T = \{0, 1, \ldots, t_1\}$, $T_1 = \{0, 1, \ldots, t_1 - 1\}$, $t_1 \in \mathbb{Z}_+$. See Appendix B for the full details of the definition of a Gaussian stochastic control system. Assume that for all $t \in T$, $N(t)V(t)N(t)^T > 0$. Define the class $G$ of control laws by $g \in G$, $g = \{g_0, g_1, \ldots, g_{t_1-1}\}$, $g_0 \in U$, and for $t = 1, \ldots, t_1-1$, $g_t : Y^t \times U^t \to U$, where $g_t$ are measurable maps. For the definition of the cost function let $c \in \mathbb{R}$, $c \neq 0$, Assume that for all $t \in T_1$ $D_2(t)^T D_2(t) > 0$ and that $L_{1e} = L_{1e}^T \geq 0$. Define the cost function $J : G \to \mathbb{R}$*

$$J(g) = E[c \exp\left(\frac{1}{2}c\left[x^g(t_1)^T L_{1e} x^g(t_1) + \sum_{s=0}^{t_1-1} z(s)^T z(s)\right]\right)]. \tag{F.2}$$

*The problem is to determine*

$$\inf_{g \in G} J(g), \tag{F.3}$$

*and to determine a control law $g^* \in G$ such that*

$$J(g^*) = \inf_{g \in G} J(g). \tag{F.4}$$

The representation of the stochastic control system (F.1) differs from that considered by P. Whittle in [70] given by

$$\begin{aligned} x(t+1) &= A(t)x(t) + B(t)u(t) + v(t), \\ y(t+1) &= C(t)x(t) \qquad\qquad\quad + w(t). \end{aligned} \tag{F.5}$$

It is well known in system theory for discrete-time systems that the solutions to the control problems for the representations (F.1) and (F.5) differ and these solutions cannot be directly transformed into each other.

Consider the Gaussian stochastic control system (F.1). Define the class $G$ of control laws by $g \in G$, $g = \{g_0, g_1, \ldots, g_{t_1-1}\}$, $g_0 \in U$, and for $t = 1, \ldots, t_1 - 1$, $g_t : Y^t \times U^t \to U$, where $g_t$ are measurable maps. For any $g \in G$ the closed-loop control system is specified by the relations

$$\begin{aligned} x^g(t+1) &= A_1(t)x^g(t) + B_1(t)g_t + M(t)v(t), \quad x^g(0) = x_0, \\ y^g(t) \quad &= C_1(t)x^g(t) + D_1(t)g_t + N(t)v(t). \end{aligned} \tag{F.6}$$

where $g_t = g_t(y^g(0), \ldots, y^g(t-1), u^g(0), \ldots, u^g(t-1))$. In the sequel the superindex $g$ on $x^g, y^g$ will be omitted.

The formulation of a closed-loop stochastic control system includes the case of a dynamic compensator. The solution to Problem F.1 above will not be presented here. Instead an expression will be presented for the cost of an LEQG problem. This expression is needed in the body of the paper.

**Proposition F.2** *Consider a Gaussian system*

$$\begin{aligned} x(t+1) &= A(t)x(t) + M(t)v(t), \quad x(t_0) = x_0, \\ z(t) &= C(t)x(t) + N(t)v(t), \end{aligned} \tag{F.7}$$

*with* $x_0 \in G(m_0, Q_0)$ *and* $v : \Omega \times T \to R^r$ *Gaussian white noise with* $v(t) \in G(0, V_1)$*. Let* $c \in \mathbb{R}$*. Assume that*

(*i*) $V_1 = V_1^T > 0$;

(*ii*) $Q_0 = Q_0^T > 0$;

(*iii*) *for all* $t \in T$, $V_1^{-1} - c[N^T(t)N(t) + M^T(t)Q_1(t+1)M(t)] > 0$ ;

(*iv*) $Q_0^{-1} - cQ_1(0) > 0$.

*Define* $Q_1 : T \to \mathbb{R}^{n\times n}$, $V_2 : T \to \mathbb{R}^{r\times r}$, $r : T \to \mathbb{R}$ *recursively by*

$$Q_1(t_1+1) = 0, \tag{F.8}$$

$$r(t_1+1) = c, \tag{F.9}$$

$$V_2(t)^{-1} = V_1^{-1} - c[N(t)^T N(t) + M(t)^T Q_1(t+1)M(t)], \tag{F.10}$$

$$\begin{aligned} Q_1(t) = {} & A(t)^T Q_1(t+1)A(t) + c[C(t)^T N(t) + A(t)^T Q_1(t+1)M(t)] \\ & \times V_2(t)[C(t)^T N(t) + A(t)^T Q_1(t+1)M(t)]^T + C(t)^T C(t), \end{aligned} \tag{F.11}$$

$$r(t) = r(t+1)\left(\frac{\det V_2(t)}{\det V_1}\right)^{1/2}, \tag{F.12}$$

$$Q_2 = Q_0^{-1}[Q_0^{-1} - cQ_1(0)]^{-1}Q_0^{-1} - Q_0^{-1}, \tag{F.13}$$

$$Q_3^{-1} = Q_0^{-1} - cQ_1(0). \tag{F.14}$$

*Then*

$$E\left[c\exp\left(\frac{1}{2}c\sum_{s=0}^{t_1} z(s)^T z(s)\right)\right] = r(0)\left(\frac{\det Q_3}{\det Q_0}\right)^{1/2}\exp\left(\frac{1}{2}m_0^T Q_2 m_0\right). \tag{F.15}$$

**Proof** 1. Let $W : T \times X \to \mathbb{R}$

$$W(t, x(t)) = E\left[c\exp\left(\frac{1}{2}c\sum_{s=t}^{t_1} z(s)^{\mathrm{T}} z(s)\right)\middle| F_t^x \vee F_{t-1}^v\right]. \tag{F.16}$$

This function is well defined because $x$ is a Markov process and $v$ is Gaussian white noise. It will be shown that $W$ satisfies the recursion

$$W(t, x(t)) = E\left[\exp\left(\frac{1}{2}cz(t)^{\mathrm{T}}z(t)\right)W(t+1, x(t+1))\middle| F_t^x \vee F_{t-1}^v\right],$$

with $W(t_1+1, x(t_1+1)) = c$. By definition, the formula at the terminal time holds. Let

$t \in T$ $t < t_1$ and assume that the recursion formula holds for $s = t+1, \ldots, t_1$. Then

$$\begin{aligned}
&W(t, x(t)) \\
&= E\left[ c \exp\left( \frac{1}{2} c \sum_{s=t}^{t_1} z(s)^{\mathrm{T}} z(s) \right) \middle| F_t^x \vee F_{t-1}^v \right] \\
&= E\left[ \exp\left( \frac{1}{2} c z(t)^{\mathrm{T}} z(t) \right) E\left\{ c \exp\left( \frac{1}{2} \sum_{s=t+1}^{t_1} z(s)^{\mathrm{T}} z(s) \right) \middle| F_{t+1}^x \vee F_t^v \right\} \middle| F_t^x \vee F_{t-1}^v \right] \\
&= E\left[ \exp\left( \frac{1}{2} c z(t)^{\mathrm{T}} z(t) \right) W(t+1, x(t+1)) \middle| F_t^x \vee F_{t-1}^v \right].
\end{aligned}$$

where the second equality follows by reconditioning and because $z(t)$ is measurable on $F_{t+1}^x \vee F_t^v$ The result then follows by induction.

2. The proof proceeds as in dynamic programming, except that there is no optimization in each step. Below use is made of Proposition D.5. Thus

$$\begin{aligned}
W(t_1, x(t_1)) &= E\left[ c \exp\left( \frac{1}{2} c z(t_1)^{\mathrm{T}} z(t_1) \right) \middle| F_{t_1}^x \vee F_{t_1-1}^v \right] \\
&= E\left[ c \exp\left( \frac{1}{2} c \begin{pmatrix} v(t_1) \\ x(t_1) \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} N^{\mathrm{T}} N & N^{\mathrm{T}} C \\ C^{\mathrm{T}} N & C^{\mathrm{T}} C \end{pmatrix} \begin{pmatrix} v(t_1) \\ x(t_1) \end{pmatrix} \right) \middle| F_{t_1}^x \vee F_{t_1-1}^v \right] \\
&= c \exp\left( \frac{1}{2} c x(t_1)^{\mathrm{T}} Q_1(t_1) x(t_1) \right) \left( \frac{\det V_2(t_1)}{\det V_1} \right)^{1/2} \\
&= r(t_1) \exp\left( \frac{1}{2} c x(t_1)^{\mathrm{T}} Q_1(t_1) x(t_1) \right).
\end{aligned}$$

Suppose that for $s = t+1, t+2, \ldots, t_1$

$$W(s, x(s)) = r(s) \exp\left( \frac{1}{2} c x(s)^{\mathrm{T}} Q_1(s) x(s) \right).$$

It will be shown that this formula holds for $s = t$.

$$\begin{aligned}
&W(t, x(t)) \\
&= E\left[ \exp\left( \frac{1}{2} c z(t)^{\mathrm{T}} z(t) \right) W(t+1, x(t+1)) \middle| F_t^x \vee F_{t-1}^v \right] \\
&= E\left[ r(t+1) \exp\left( \frac{1}{2} c \left[ z(t)^{\mathrm{T}} z(t) + x(t+1)^{\mathrm{T}} Q_1(t+1) x(t+1) \right] \right) \middle| F_t^x \vee F_{t-1}^v \right] \\
&= E\left[ r(t+1) \exp\left( \frac{1}{2} c \begin{pmatrix} v(t) \\ x(t) \end{pmatrix}^{\mathrm{T}} Z \begin{pmatrix} v(t) \\ x(t) \end{pmatrix} \right) \middle| F_t^x \vee F_{t-1}^v \right] \\
&= r(t) \exp\left( \frac{1}{2} c x(t)^{\mathrm{T}} Q_1(t) x(t) \right)
\end{aligned}$$

where

$$Z = \begin{pmatrix} N^{\mathrm{T}} N + M^{\mathrm{T}} Q_1(t+1) M & N^{\mathrm{T}} C + M^{\mathrm{T}} Q_1(t+1) A \\ C^{\mathrm{T}} N + A^{\mathrm{T}} Q_1(t+1) M & C^{\mathrm{T}} C + A^{\mathrm{T}} Q_1(t+1) A \end{pmatrix}.$$

The first equality follows from step 1 of the proof. The second equality is a consequence of the induction step. For the last steps we used Proposition D.5 and the formulas for $V_2(t)^{-1}$, $Q_1(t)$, and $r(t)$. Finally

$$
\begin{aligned}
E & \left[ c \exp\left( \frac{1}{2} c \sum_{s=0}^{t_1} z(s)^{\mathrm{T}} z(s) \right) \right] \\
& = E\left[ E\left[ c\exp\left( \frac{1}{2} c \sum_{s=0}^{t_1} z(s)^{\mathrm{T}} z(s) \right) \middle| F_0^x \vee F_{-1}^v \right] \right] \\
& = E\left[ W(0, x(0)) \right] = E\left[ r(0) \exp\left( \frac{1}{2} c x(0)^{\mathrm{T}} Q_1(0) x(0) \right) \right] \\
& = r(0) \left( \frac{\det Q_3}{\det Q_0} \right)^{1/2} \exp\left( \frac{1}{2} m_0^{\mathrm{T}} Q_2 m_0 \right).
\end{aligned}
$$

□

**Proposition F.3** *Consider the time-invariant Gaussian system*

$$
\begin{aligned}
x(t+1) &= Ax(t) + Mv(t), \quad x(t_0) = x_0, \\
z(t) &= Cx(t) + Nv(t),
\end{aligned}
\tag{F.17}
$$

*with* $x_0 \in G(m_0, Q_0)$ *and* $v : \Omega \times T \to \mathbb{R}^r$ *Gaussian white noise with* $v(t) \in G(0, V_1)$. *Let* $c \in \mathbb{R}$, $c > 0$. *Define* $Q_1 : T \to \mathbb{R}^{n \times n}$ *as in Proposition F.2. Assume that:*

(*i*) $V_1 = V_1^T > 0$;

(*ii*) $Q_0 = Q_0^T > 0$;

(*iii*) *for all* $t \in T$, $V_1^{-1} - c[N^T N + M^T Q_1(t+1) M] > 0$;

(*iv*) $Q_0^{-1} - cQ_1(0) > 0$;

(*v*) *the following set of relations admits an unique solution* $Q \in \mathbb{R}^{n \times n}$

$$
Q = A^T Q A + C^T C + c[A^T Q M + C^T N] V_2 [A^T Q M + C^T N]^T, \tag{F.18}
$$

$$
V_2^{-1} = V_1^{-1} - cM^T Q M - cN^T N > 0, \tag{F.19}
$$

$$
Q_0^{-1} - cQ > 0; \tag{F.20}
$$

(*vi*) *For all* $t \in T$, $\lim_{T \to \infty} Q_{1,T}(t) = Q$ *where* $Q_{1,T}(t)$ *is the solution of (F.11) with terminal condition* $Q_{1,T}(T+1) = 0$.

*Then*

$$
\begin{aligned}
\lim_{t \to \infty} \frac{1}{t} \ln E & \left[ c \exp\left( \frac{1}{2} c \sum_{s=0}^{t} z(s)^T z(s) \right) \right] \\
& = -\frac{1}{2} \ln \det(V_1^{-1} - cM^T Q M - cN^T N) + \frac{1}{2} \ln \det V_1^{-1}.
\end{aligned}
\tag{F.21}
$$

**Proof:** Consider the horizon $\{0, 1, \ldots, T\}$ with $x_0 \in G(0, Q_0)$. By (F.10) and assumption (vi)

$$\begin{aligned}
\lim_{T\to\infty} V_2(t)^{-1} &= \lim_{T\to\infty} [V_1^{-1} - cM^{\mathrm{T}} Q_{1,T}(t) M - cN^{\mathrm{T}} N] = V_1^{-1} - cM^{\mathrm{T}} Q M - cN^{\mathrm{T}} N, \\
\lim_{T\to\infty} Q_3^{-1} &= \lim_{T\to\infty} [Q_0^{-1} - cQ_{1,T}(0)] = Q_0^{-1} - cQ > 0, \quad \text{and finite}, \\
\lim_{T\to\infty} Q_2 &< \infty.
\end{aligned}$$

Then

$$\begin{aligned}
&\lim_{t\to\infty} \frac{1}{t} \ln E \left[ c \exp \left( \frac{1}{2} c \sum_{s=0}^{t} z(s)^{\mathrm{T}} z(s) \right) \right] \\
&= \lim \frac{1}{t} \ln \left( c \left( \frac{\det Q_3}{\det Q_0} \right)^{1/2} \prod_{s=0}^{t} \left( \frac{\det V_2(s)}{\det V_1} \right)^{1/2} \right) \\
&\text{(by Proposition F.2)} \\
&= \lim \left[ \frac{1}{t} \ln c + \frac{1}{2t} \ln \left( \frac{\det Q_3}{\det Q_0} \right) + \frac{1}{2t} \sum_{s=0}^{t} \ln \left( \frac{\det V_2(s)}{\det V_1} \right) \right] \\
&= \frac{1}{2} \ln \left( \frac{\det V_2}{\det V_1} \right) \\
&= -\frac{1}{2} \ln \det(V_1^{-1} - cM^{\mathrm{T}} Q M - cN^{\mathrm{T}} N) + \frac{1}{2} \ln \det(V_1^{-1}). \qquad \square
\end{aligned}$$

# G H-infinity control with an entropy criterion

The purpose of this section is to give the reader a very brief introduction to the discrete time $H_\infty$ control problem. The $H_\infty$ control problem was formulated by G. Zames in [77]. However its roots go back much further to work on differential games. One of the first publications directly related to $H_\infty$ was [51]. A first solution of the $H_\infty$ control problem was based on frequency domain techniques and for an overview we refer to [23]. A breakthrough was the paper [18] which presented a complete solution of the $H_\infty$ control problem based on time-domain techniques. This approach was based on Riccati equations and hence could be implemented easily. A first solution of the discrete time $H_\infty$ problem can be found in [66, 49, 9]. Currently several good books are available [10, 67, 50, 31].

The main reason for studying the $H_\infty$ control problem is model uncertainty. Using $H_\infty$, it is possible to suppress the effect of model uncertainty on the behaviour of our plant. Also dependence on our knowledge of noise characteristics can be handled via $H_\infty$ control.

We consider the following control system:

$$\Sigma: \begin{array}{ll}
x(t+1) = A(t)x(t) + B_1(t)u(t) + M(t)v(t), & x(0) = x_0, \\
y(t) \quad\;\; = C_1(t)x(t) + D_1(t)u(t) + N(t)v(t), & \\
z(t) \quad\;\; = C_2(t)x(t) + D_2(t)u(t). &
\end{array} \tag{G.1}$$

Here $x \in \mathbb{R}^n$ is a state vector in which we are interested. $y \in \mathbb{R}^p$ is a measurement vector and $v \in \mathbb{R}^r$ is a disturbance affecting both the state evolution as well as our measurement. Note that $v$ might consist of two independent components; one affecting the state evolution and one the measurement. In contrast with LEQG we do not assume that $v$ is white noise but an arbitrary $\ell_2$ signal. Finally, $u$ denotes our control input and $z$ is an output we want to make small.

Define the class $G$ of control laws by $g \in G$, $g = \{g_0, g_1, \dots, g_{t_1-1}\}$, $g_0 \in U$, and for $t = 1, \dots, t_1 - 1$, $g_t : Y^t \times U^t \to U$, where $g_t$ are measurable maps. We define the following cost function:

$$J(g) = \sup_{v, x_0} \frac{\|z\|_{2,t_1}^2}{\|v\|_{2,t_1}^2 + x_0^{\mathrm{T}} R^{-1} x_0} \tag{G.2}$$

where

$$\|v\|_{2,t_1}^2 := \sum_{s=0}^{t_1} v(t)^{\mathrm{T}} v(t). \tag{G.3}$$

If $t_1 = \infty$ then we must constrain $v$ to $\ell_2$, the class of signals for which the infinite sum (G.3) converges. The problem is then to determine

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

This problem turns out to be very hard but the related suboptimal problem does have an elegant solution. For each $c \in \mathbb{R}$ we are looking for a controller $g \in G$ such that $J(g) < c^{-1}$. In the literature, necessary and sufficient conditions for the solvability of this suboptimal problem have been derived. Moreover, if it exists, one particular suboptimal controller is given. This controller is suboptimal for the original optimization problem. However, this controller minimizes the following criterion:

$$J_c(g) = \sup_{v, x_0} \; c\|z\|_{2,t_1}^2 - \|v\|_{2,t_1}^2 - x_0^{\mathrm{T}} R^{-1} x_0.$$

If $t_1 = \infty$, $x(0) = 0$, and both the system and the controller are time-invariant, then this particular controller has another interpretation.

First note that in this case

$$J(g) = \|S\|_\infty^2 := \sup_{\lambda \in [-\pi, \pi]} \|S(e^{i\lambda})\|^2$$

where $S$ denotes the closed loop transfer matrix from $v$ to $z$. As shown in [52] for the continuous time and later in [36] for the discrete time, the particular controller mentioned above is optimal for the following optimization problem:

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

with

$$J_e(g) = \frac{-1}{4\pi i} \int_{\mathbb{D}} \frac{1}{z} \ln \det \left( I - cS^{\mathrm{T}}(z^{-1}) S(z) \right) dz. \tag{G.4}$$

Clearly this interpretation is only valid for a time-invariant system. Note that this expression is the same as the expression for the mutual information given in Theorem E.3. Hence given a controller and hence the closed loop transfer matrix then (G.4) is equal to an expression of the form (E.10). For the sake of completeness we formulate this in the following theorem.

**Theorem G.1** *Assume that the transfer matrix $S$ admits a realization as a finite dimensional linear system with*

$$S(z) = C(zI - A)^{-1}B + D, \quad sp(A) \subset \mathbb{C}^{-}. \tag{G.5}$$

*The following set of relations admits an unique solution $Q \in \mathbb{R}^{n\times n}$*

$$Q = Q^T \geq 0, \tag{G.6}$$
$$Q = A^TQA + C^TC + (A^TQB + C^TD)N^{-1}(B^TQA + D^TC), \tag{G.7}$$
$$N = c^{-1}I - B^TQB - D^TD > 0, \tag{G.8}$$
$$sp(A + BN^{-1}(B^TQA + D^TC)) \subset \mathbb{C}^{-}, \tag{G.9}$$

*if and only if $\|S\|_\infty < c^{-1/2}$. Moreover, in that case*

$$\frac{-1}{4\pi i}\int_{\mathbb{D}} \frac{1}{z}\ln\det\left(I - cS^T(z^{-1})S(z)\right)\,dz = -\frac{1}{2}\ln\det(I - cB^TQB - cD^TD). \tag{G.10}$$

# References

[1] H. Akaike, “Information theory and an extension of the maximum likelihood principle”, in Proceedings 2nd International Symposium Information Theory, B.N. Petrov and F. Csaki, eds., Akademia Kiado, Budapest, 1973, pp. 267–281.

[2] ——, “Canonical correlation analysis of time series and the use of an information criterion”, in System identification - Advances and case studies, R.K. Mehra and D.G. Lainiotis, eds., Academic Press, New York, 1976, pp. 27–96.

[3] ——, “On entropy maximization principle”, in Applications of statistics, P.R. Krishnaiah, ed., North-Holland Publ. Co., Amsterdam, 1977, pp. 27–41.

[4] S.I. Amari, “Differential geometry of curved exponential families - Curvatures and information loss”, Ann. Statist., 10 (1982), pp. 357–385.

[5] ——, “Differential geometry of a parametric family of invertible linear systems - Riemannian metric, dual affine connections, and divergence”, Math. Systems Th., 20 (1987), pp. 53–82.

[6] Shun-Ichi Amari, *Differential-geometrical methods in statistics*, vol. 28 of Lecture Notes in Statistics, Springer-Verlag, Berlin, 1985.

[7] T.W. Anderson, *An introduction to multivariate statistical analysis*, Wiley, New York, 1958.

[8] D.Z. Arov and M.G. Krein, “Problem of search of the minimum of entropy in indeterminate extension problem”, Functional Analysis and its Applications, 15 (1981), pp. 123–126.

[9] T. Başar, “A dynamic games approach to controller design: disturbance rejection in discrete time”, in Proc. CDC, Tampa, 1989, pp. 407–414.

[10] T. Başar and P. Bernhard, *$H_\infty$-optimal control and related minimax design problems: a dynamic game approach*, Birkhäuser, Boston, 1991.

[11] 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 (1985), pp. 599–613.

[12] M.C. Campi and M.R. James, "Nonlinear discrete-time risk-sensitive optimal control", Preprint X, Department of Systems Engineering, Australian National University, Canberra, 1993.

[13] T.M. Cover and J.A. Thomas, *Elements of information theory*, John Wiley & Sons, New York, 1991.

[14] I. Csiszár, "Information type measures of difference of probability distributions and indirect observations", Studia Sci. Math. Hungar., 2 (1967), pp. 229–318.

[15] ——, "I-divergence geometry of probability distributions and minimization problems", Ann. Probab, 3 (1975), pp. 146–158.

[16] I. Csiszár and J. Körner, *Information Theory*, Academic Press, New York, 1981.

[17] I. Csiszár and G. Tusnady, "Information geometry and alternating minimization procedures", in Statistics and Decisions, Supplement issue no.1, X, ed., R. Oldenbourg Verlag, München, 1984, pp. 205–237.

[18] J.C. Doyle, K. Glover, P.P. Khargonekar, and B.A. Francis, "State space solutions to standard $H_2$ and $H_\infty$ control problems", IEEE Trans. Aut. Contr., 34 (1989), pp. 831–847.

[19] P. Faurre, M. Clerget, and F. Germain, *Opérateurs rationnels positifs*, Dunod, Paris, 1979.

[20] D.F. Findley, "On the unbiasedness property of AIC for exact or approximating linear stochastic time series models", J. Time Series Analysis, 6 (1985), pp. 229–252.

[21] R.A. Fisher, "On an absolute criterion for fitting frequency curves", Mess. of Math., 41 (1912), pp. 155–160.

[22] ——, "Theory of statistical estimation", Proc. Cambridge Philos. Soc., 22 (1925), pp. 700–725.

[23] B.A. Francis, *A course in $H_\infty$ control theory*, vol. 88 of Lecture notes in control and information sciences, Springer-Verlag, 1987.

[24] I.M. Gelfand and A.M. Yaglom, "Calculation of the amount of information about a random function contained in another such function", American Mathematical Society Translations, Series 2, 12 (1959), pp. 199–246.

[25] R.D. Gittens, *Canonical analysis - A review with applications in ecology*, Springer-Verlag, Berlin, 1985.

[26] K. Glover and J.C. Doyle, "State-space formulae for all stabilizing controllers that satisfy an $H_\infty$-norm bound and relations with risk sensitivity", Systems & Control Lett., 11 (1988), pp. 167–172.

[27] K. Glover and D. Mustafa, "Derivation of the maximum entropy $H_\infty$-controller and a state-space formula for its entropy", Int. J. Control, 50 (1989), pp. 899–916.

[28] A. Gombani and M. Pavon, "On the Hankel-norm approximation of linear stochastic systems", Systems & Control Lett., 5 (1985), pp. 283–288.

[29] R.M. Gray, *Probability, random processes, and ergodic properties*, Springer, Berlin, 1988.

[30] R.M. Gray and L.D. Davisson, *Ergodic and information theory*, Dowden, Hutchinson, Ross, X, 1977.

[31] M. Green and D.J.N. Limebeer, *Linear robust control*, Information and System Sciences Series, Prentice Hall, 1994.

[32] Chih hai Fan, J.L. Speyer, and C.R. Jaensch, "Centralized and decentralized solutions of the linear-exponential-Gaussian problem", IEEE Trans. Automatic Control, 39 (1994), pp. 1986–2003.

[33] E.J. Hannan and M. Deistler, *The statistical theory of linear systems*, John Wiley & Sons, New York, 1988.

[34] B. Hassibi, A.H. Sayed, and T. Kailath, "Recursive linear estimation in Krein spaces - Part I Theory", in Proc. 32nd IEEE Conference on Decision and Control, New York, 1993, IEEE, pp. 3489–3494.

[35] ——, "Recursive linear estimation in Krein spaces - Part II Applications", in Proc. 32nd IEEE Conference on Decision and Control, New York, 1993, IEEE, pp. 3495–3500.

[36] P.A. Iglesias, D. Mustafa, and K. Glover, "Discrete time $H_\infty$ controllers satisfying a minimum entropy criterion", Syst. & Contr. Letters, 14 (1990), pp. 275–286.

[37] S. Ihara, *Information theory for continuous systems*, World Scientific, Singapore, 1993.

[38] D.H. Jacobson, "Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games", IEEE Trans. Automatic Control, 18 (1973), pp. 124–131.

[39] C.R. Jaensch and J.L. Speyer, "Centralized and decentralized stochastic control problems with an exponential cost criterion", in Proceedings of the 27th Conference on Decision and Control, New York, 1988, IEEE, pp. 286–291.

[40] M.R. James and J.S. Baras, "Robust output feedback control for discrete-time nonlinear systems", Preprint X, Department of Systems Engineering, Australian National University, Canberra, 1993.

[41] M.R. James, J.S. Baras, and R.J. Elliott, "Risk-sensitive control and dynamic games for partially observed discrete-time nonlinear systems", Preprint X, Department of Systems Engineering, Australian National University, Canberra, 1993.

[42] N.P. Jewell and P. Bloomfield, "Canonical correlations of past and future for time series: Definitions and theory", Ann. Statist., 11 (1983), pp. 837–847.

[43] N.P. Jewell, P. Bloomfield, and F.C. Bartmann, "Canonical correlations of past and future for time series: Bounds and computation", Ann. Statist., 11 (1983), pp. 848–855.

[44] A.N. Kolmogorov, "On the Shannon theory of information in the case of continuous signals", IRE. Trans. Inform. Theory, 2 (1956), pp. 102–108.

[45] S. Kullback, J.C. Keegel, and J.H. Kullback, *Topics in statistical information theory*, vol. 42 of Lecture Notes in Statistics, Springer-Verlag, Berlin, 1987.

[46] S. Kullback and R.A. Leibler, "On information and sufficiency", Ann, Math. Statist., 22 (1951), pp. 79–86.

[47] P.R. Kumar, *Stochastic optimal control and stochastic differential games*, PhD thesis, Department of Systems Science and Mathematics, Washington University, St. Louis, MO, U.S.A., 1977.

[48] 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 (1981), pp. 312–332.

[49] D.J.N. Limebeer, M. Green, and D. Walker, "Discrete time $H_\infty$ control", in Proc. CDC, Tampa, 1989, pp. 392–396.

[50] J.M. Maciejowski, *Multivariable feedback design*, Addison-Wesley, Reading, MA, 1989.

[51] J. Medanic, "Bounds on the performance index and the Riccati equation in differential games", IEEE Trans. Aut. Contr., 34 (1967), pp. 613–614.

[52] D. Mustafa and K. Glover, *Minimum entropy $H_\infty$ control*, vol. 146 of Lecture notes in control and information sciences, Springer Verlag, Berlin, 1990.

[53] R. Nishii, "Maximum likelihood principle and model selection when the true model is unspecified", J. Multivar. Anal., 27 (1988), pp. 392–403.

[54] M. Pavon, "Canonical correlations of past inputs and future outputs for linear stochastic systems", Syst. Control Lett., 4 (1984), pp. 209–215.

[55] A. Perez, "Notions generalisees d' incertitude d' entropie et d' information du point de vue de la theorie de martingales", in Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes, Prague, 1957, Publishing House of the Czechoslovak Academy of Sciences, pp. 183–208.

[56] ——, "Sur la theorie de l' information dans le cas d'un alphabet abstrait", in Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes, Prague, 1957, Publishing House of the Czechoslovak Academy of Sciences, pp. 209–243.

[57] M.S. Pinsker, *Information and information stability of random variables*, Holden-Day, San Francisco, 1964.

[58] M. Rosenblatt, "Asymptotic distribution of eigenvalues of block Toeplitz matrices", Bull. Am. Math. Soc., 66 (1960), pp. 320–321.

[59] 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, New York, 1990, IEEE Press, pp. 1018–1023.

[60] ——, "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, New York, 1991, IEEE Press, pp. 935–936.

[61] C.E. Shannon, "A mathematical theory of communication", Bell System Techn. Journal, 27 (1948), pp. 379–423, 623–656.

[62] C.E. Shannon and W.W. Weaver, *The mathematical theory of communication*, University of Illinois, Urbana, 1949.

[63] V. Solo, "The exact likelihood for a multivariate ARMA model", J. Multivar. Anal., 15 (1984), pp. 164–173.

[64] J.L. Speyer, "An adaptive terminal guidance scheme based on an exponential cost criterion with application to homing missile guidance", IEEE Trans. Automatic Control, 21 (1976), pp. 371–375.

[65] J. Speyer, J. Deyst, and D.H. Jacobson, "Optimization of stochastic linear systems with additive measurement and process noise using exponential performance criteria", IEEE Trans. Automatic Control, 19 (1974), pp. 358–366.

[66] A.A. Stoorvogel, "The discrete time $H_\infty$ control problem with measurement feedback", SIAM J. Contr. & Opt., 30 (1992), pp. 182–202.

[67] ——, *The $H_\infty$ control problem: a state space approach*, Prentice-Hall, Englewood Cliffs, 1992.

[68] A.A. Stoorvogel and J.H. van Schuppen, “An $H_\infty$-parameter estimator and its interpretation”, in Proc. 10th IFAC Symposium System Identification, vol. 3, Pergamon Press, 1994, pp. 267–270.

[69] J.H. van Schuppen, “Stochastic realization problems”, in Three decades of mathematical system theory, J.M. Schumacher H. Nijmeijer, ed., vol. 135 of Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, 1989, pp. 480–523.

[70] P. Whittle, “Risk-sensitive Linear/Quadratic/Gaussian control”, Adv. Appl. Prob., 13 (1981), pp. 764–777.

[71] ——, *Optimization over time*, John Wiley, New York, 1982.

[72] ——, “Scheduling and characterization problems for stochastic networks”, J.R. Statist. Soc. B., 47 (1985), pp. 407–415.

[73] ——, “A risk-sensitive maximum principle”, Systems & Control Lett., 15 (1990), pp. 183–192.

[74] ——, *Risk-sensitive optimal control*, Wiley, Chichester, 1990.

[75] ——, “A risk-sensitive maximum principle: The case of imperfect state observations”, IEEE Trans. Automatic Control, 36 (1991), pp. 793–801.

[76] P. Whittle and J. Kuhn, “A Hamiltonian formulation of risk-sensitive Linear/Quadratic/Gaussian control”, Int. J. Control, 43 (1986), pp. 1–12.

[77] G. Zames, “Feedback and optimal sensitivity: model reference transformations, multiplicative seminorms, and approximate inverses”, IEEE Trans. Aut. Contr., 26 (1981), pp. 301–320.

# Contents