# Predictive Information as a Criterion to Linear Dynamical Systems Reduction

M.Sc Thesis

by

Roi Weiss

Advisor: Prof. Naftali Tishby

The Racah Institute of Physics, The Hebrew University, Jerusalem, Israel

31 October 2007

# ABSTRACT

In this thesis we discuss the reduction of linear dynamical systems driven by stochastic input taking the criterion for reduction to be the information between the past and the future of the induced process. We first discuss the application of the information bottleneck method and characterize the information curves for the reduced system. Next we discuss a closely related model which is parametric and elegantly solved with the help of the celebrated cepstrum.

## ACKNOWLEDGMENT

I would like to thank Professor Naftali Tishby for deep insights and persistent encouragement.

# Contents

# Chapter 1

# Introduction

## 1.1 Motivation

The ability to predict the future is obviously an advantage for causal adapting systems. Let us take as an example a living organism which has to adapt to a stochastic environment. An organism will have an advantage over another one if he could predict the future to make better actions which result in raising its chances to survive and spread its genes. On the other hand the ability to predict the future is costlly and consume resources which are limited. It is suggested by Bialek, Steveninck and Tishby [1] that such an adaptive behavior can be a result of an optimization principle.

In formulating this principle we assume that there is some abstract utility function $U(A, W)$ (e.g evolutionary fitness or some reward function) which measure the quality of an action, $A$, given the current (or future) state of the world, $W$. We also assume that taking an action and gathering and processing the information needed to do the action is costlly and this cost is determined also by an abstract cost function $C(A, W)$. The goal of an organism is to maximize utility for a given cost or in other words to get the maximum of his efforts. This is obviously a trade-off problem and it defines a curve in the plane of "maximum utillity" vs. "cost" or in the biological context "adaptive value" vs. "resources".

This curve separate the achievable and unachievable by an organism. We sketch such a curve in the upper right quarter of figure 1.1.

In [1] it is shown that this problem can be mapped into an information theoretic problem as follows (see figure 1.1). The organism has only an internal representation $X_{int}$ of the world by whom it decide what its next action will be. By causality it is only a function of the past. With limited resources we have a limit of how well $X_{int}$ can represent the world's past (the channel capacity limit). The quality of this representaion is measured by the information $X_{int}$ has about the past, here denoted by $I(X_{past}; X_{int})$. In this state of affairs there is some curve which give the maximum information we can have about the past for a given limited resources (buttom right quarter in figure 1.1). This curve separate achievable and unachievable regions in the "information about the past"/"resources" plane.

On the other side of the chain, only representations that have some predictive information about the future can be used to raise the adaptive value; more stringently, there is a limit of the minimum information about the future possible in order to achieve a given adaptive value (this is the rate distortion limit). This draw a curve of the minimum information possible about the furture for a given adaptive value which separate achievable and unachievable regions in the "adaptive value"/"information about the future" plane (upper left quarter in figure 1.1). Polani et. al. (2001) [4] and (2006) [5], by taking notions from decision theory and information theory, gives a very beautiful way to quantify those notions (see also [2] and references there in for another similar possible direction).

Now, having limited representation of the past results in limited predictive information about the future. In other words $X_{int}$ is a bottleneck between the past and the future. Thus in order to get biological optimality , $X_{int}$ should capture the maximum information about the relevant future with a limited information about the past. In other words, there is curve in the "information about the future"/"information about the past" plane that give the maximum information about the future possible with a given information about the past; an organism that sit on the "maximum adaptive value for a given resources curve" must also sit on this optimized curve. We call this curve the "Information curve".

To summarise, optimallity in the biological curve is reduced to optimallity in the information curve (this is depicted by the arrows in figure 1.1) and thus we are left with an information theoretic problem of finding the maximum information about the future $I(X_{int}; X_{fututre})$ with a given information about the past $I(X_{past}; X_{int})$.

**Remark 1.1** *Notice that this problem is equivalent to the problem of finding the minimum of information about the past given the information about the future.*

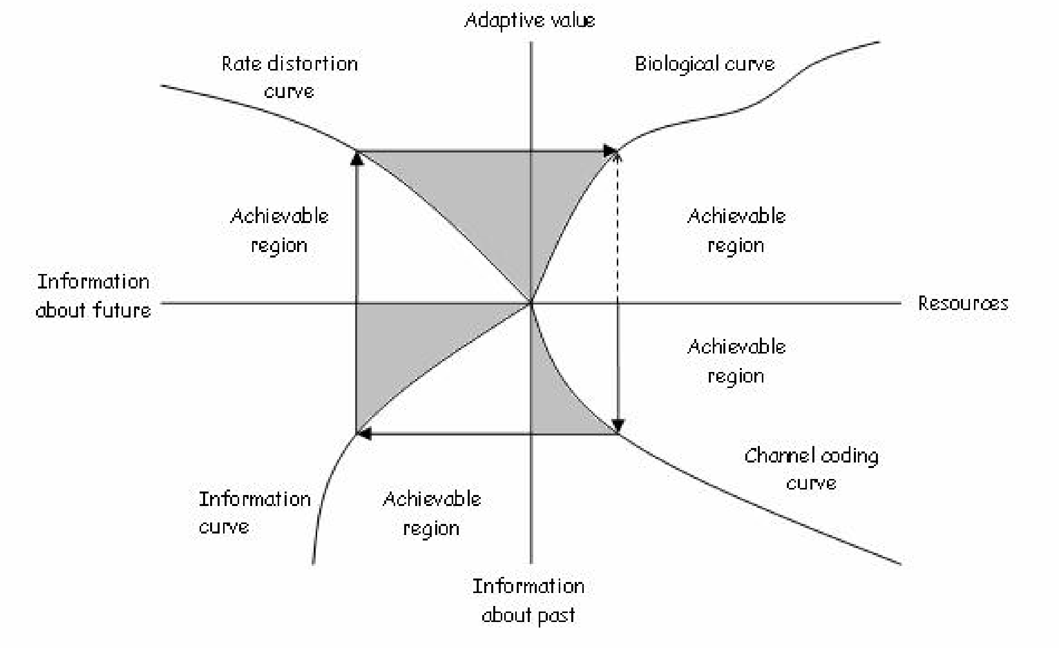

Figure 1.1: Optimal adaptation of an organism in limited resources (upper right quarter) reduced to an information theoretic optimization problem (buttom left quarter). Optimality in the biological curve induce optimality in the information curve. Taken from [1].

A crucial question in this context is how much information there is in the past about the relevant future? Assuming stationarity of the world this is only a property of the statistical structure of the data gathered by observation. Observing a window of duration T of the past $-T < t < 0$ (see figure 1.2) we need the entropy $S(T)$ to represent the data. This entropy is extensive with T. On the other hand the information this window have about the future, $I_{Pred}(T) = I_T(X_{past}; X_{future})$, is much smaller. This predictive information is shown to be subextensive with T [7]. This suggest that we can efficiently represent the past in a parsimonious internal representation while preserving the relevant information about the future. This kind of problem is addressed by the information bottleneck method presented by Tishby, Bialek et al. [3].

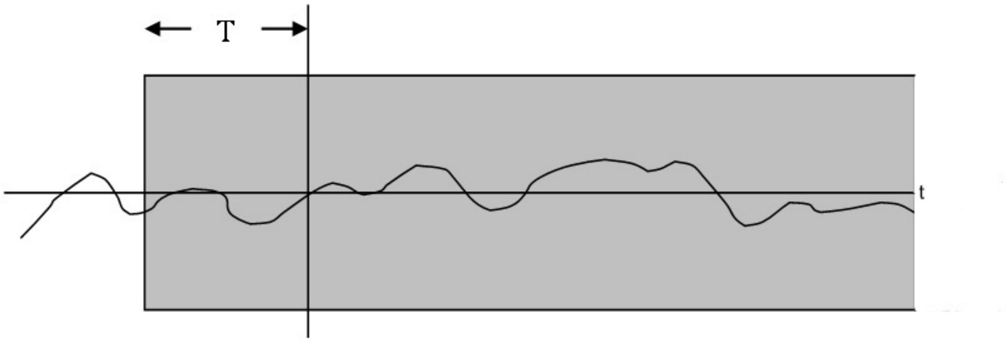

Figure 1.2: Time window.

In this thesis, as a small step, we want to characterize the above mentioned information curves to a restricted class. To do so we assume that the adaptive system (the organism) is represented by a dynamical system with internal representation about the outside stochastic world. We further restrict our dynamical system to be a discrete time linear time invariant system with stochastic input. we denote this system by $\hat{H}$. We also assume that the "world" can be approximated by another high dimensional linear dynamical system, denoted by $H$, also driven by a stationary stochastic input.

A linear dynamical system is characterized by a state space $\hat{x}(t)$ of finite dimension which we naturally relate to the internal representation discussed above. Indeed, for a linear state space system $\hat{H}$ with the stochastic input $u(t)$ and output $\hat{y}(t)$ the state $\hat{x}(\tau)$ at a given time $\tau$ is the only representation of the past input $u(t < \tau)$ that can affect the future state of the system $\hat{x}(t > \tau)$ or $\hat{y}(t > \tau)$; any data from the past not represented in $\hat{x}(\tau)$ is lost for future use. In information quantities this means

$$I(u_{past}; \hat{y}_{future}) = I(\hat{x}_{\tau}; \hat{y}_{future}). \tag{1.1}$$

**Remark 1.2** *As mentioned above we assume that the world can be approximated by a high dimensional linear system $H$ driven by the stochastic input $u$ with output $y$. In this state of affairs the maximum predictive information is $I(y_{past}; y_{future})$ and if we take the system $H$ to be constant (as we do here) this information is finite.*

The original problem is now mapped to the problem of finding an approximated system $\hat{H}_n$ with state space $\hat{x}(t)$ of dimension $n$ such that this system would capture as much relevant information about the future as possible (which "relevance" is determined by $H$) for a given information about the past.

**Remark 1.3** *The reduction of linear dynamical systems has a long history and very rich literature in variety of fields such as quantum and classical physics, control systems, signal processing, biology, economy and more. In later chapter we will mention some algorithms used in linear systems reduction.*

## 1.2 Formulating the Problem

In order to formulize the mapping discussed in the previous section we slightly generelize our problem and formulize it as an optimization problem under constrains

$$R(D) = \min_{\hat{H}_n : D(H, \hat{H}_n) \leq D} I(u_{past}; \hat{x}), \tag{1.2}$$

with a distortion function $D(H, \hat{H}_n)$ that will measure in some sense the "missed" relevant information.
So to say, the function $D(R)$ should measure the reduced system minimum loss of relevant information possible as a function of the infromation it capture about the past (i.e $R$).

By introducing the Lagrange muliplier $\beta$ we can write this optimization problem as

$$\min_{\hat{H}_n} \; I(u_{past}; \hat{x}) + \beta D(H, \hat{H}_n). \tag{1.3}$$

The most natural choice for the distortion function seems to be

$$D^{(1)}(H, \hat{H}_n) = I(x; y_{future}) - I(\hat{x}; y_{future}), \tag{1.4}$$

which obviously measure the lost of predictive information. Since $I(x; y_{future})$ is constant we are left with

$$\min_{\hat{H}_n} \; I(u_{past}; \hat{x}) - \beta I(\hat{x}; y_{future}). \tag{1.5}$$

This is just the information bottleneck problem where we compress the past $u_{past} \rightarrow \hat{x}$ with $y_{future}$ taken to be the relevant variable. We discuss this choice in chapter 4.
Another choice that we can think about is

$$D^{(2)}(H, \hat{H}_n) = -I(\hat{y}_{future}; y_{future}). \tag{1.6}$$

As we will see, since we take our model to be determenistic (although with stochastic input) this distortion function become infinte. Taking this distortion function with a stochastic system is treated in [9] and discussed briefly in chapter 4.

In chapter 6 we consider a distortion function that is closely related to both of those choices but finite and parametric. To define it we look at the scheme given in figure 1.3.

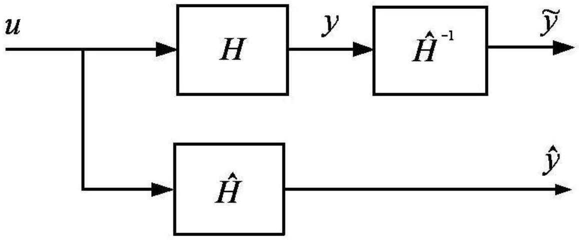

Figure 1.3: Definition of the residual process.

In this case we define our distortion function to be

$$D^{(3)}(H, \hat{H}_n) = I(u_{past}; \tilde{y}_{future}). \tag{1.7}$$

Filtering the original process $y$ with the inverse of the reduced system $\hat{H}_n^{-1}$ leave us with a residual process $\tilde{y}$. By solving the optimization problem with this distortion function this process will have only the most *irrelevant* information. Indeed, on the one extreme when the system $\hat{H}$ is null we get the maximum distortion $D(H, \hat{H}_n) = I(u_{past}; y_{future})$ and on the other extreme when the systems are equal we get $D(H, \hat{H}_n) = 0$. In between, the reduced system captures only part of the process which results in finite distortion.

Taking this choice to be our distortion function we will see later that this problem can be formulated elegantly via the poles and the zeroes of the systems with the help of the cepstrum norm. This will make it easy to solve the problem and plot the information curves.

**Remark 1.4** *As a final remark please notice that we can replace $I(u_{past}, \hat{x})$ by $I(u_{past}, \hat{y}_{future})$ and get the same effect. We will do that since later we will find that with our model, $I(u_{past}, \hat{x})$ become infinite. Thus our problem become*

$$R(D) = \min_{\hat{H}_n : d(H, \hat{H}_n) \leq D} I(u_{past}, \hat{y}_{future}). \tag{1.8}$$

We now outline the rest of the paper. In the next chapter we introduce the model we work with, discuss its properties and its equivalence to Auto-Regressive-Moving-Average (ARMA) models. In chapter 3 we introduce the concepts needed from information theory and canonical correlations and calculate the information between the past and the future for linear dynamical systems. Next we present the rate-distortion theory followed by the presentation of the information bottleneck method, its solution to gaussian variables and give a straight forward implementation for linear dynamical systems and discuss its properties. In chapter 5 we introduce the cepstrum and cepstrum norm in order to define a metric between ARMA models. In chapter 6 we use this metric to reformulate our problem as a parametric one. Next we present the results in chapter 7. Finally, in chapter 8 we summarize.

# Chapter 2

# Discrete Time Linear Systems and ARMA Models

In this chapter we discuss the dynamical system we work with. We consider a discrete time, finite dimensional linear system which we take to be time invariant and to have single input and single output (SISO) [24]. We drive the determenistic system with a sthocastic input making it's output also stochastic. We discuss both I/O and state-space descriptions, characterized by the impulse response and state equations respectively and define some related concepts. We also point out the equivalence to the ARMA model.

## 2.1 Introduction

### I/O description [27]

A (discrete time) system is (mathematically) defined as a unique transformation that maps the input sequence $u(n)$ into an output sequence $y(n)$ where $n$ represent the time which take integers values

$$y(n) = T[u(n)]. \tag{2.1}$$

The class of linear systems are constraind by the principal of superposition. If $y_1(n)$ and $y_2(n)$ are the outputs of the inputs $u_1(n)$ and $u_2(n)$ respectively, then the system is linear if and only if

$$T[au_1(n) + bu_2(n)] = aT[u_1(n)] + bT[u_2(n)] = ay_1(n) + by_2(n) \tag{2.2}$$

for any constants $a$ and $b$.
Since any sequence can be writen as

$$u(n) = \sum_{k=-\infty}^{\infty} u(k)\delta(n-k), \tag{2.3}$$

where

$$\delta(n) = \begin{cases} 0, & n \neq 0 \\ 1, & n = 0 \end{cases}, \tag{2.4}$$

a linear systems is completely characterized by $T[\delta(n-k)]$

$$y(n) = \sum_{k=-\infty}^{\infty} u(k)T[\delta(n-k)] = \sum_{k=-\infty}^{\infty} u(k)h_k(n). \tag{2.5}$$

We now restrict the system more by demending it will be time invariant which imply that if $y(n)$ is the response to $u(n)$ then $y(n-k)$ is the response to $u(n-k)$ to any $k$. This makes $h_k(n)$ to be a function only of the deiffernece $n-k$

$$y(n) = \sum_{k=-\infty}^{\infty} u(k)h(n-k). \tag{2.6}$$

$h(n)$ is called the impulse response of the system and the sum in equation (2.6) is called the convolution of $u(n)$ and $h(n)$ denoted by

$$y(n) = u(n) * h(n) = h(n) * u(n). \tag{2.7}$$

A very important aspect of linear time invariant (LTI) systems is that the impulse response of a cascade of two LTI is also a LTI with impulse response given by the convolution of the two impulse responses.

### Stability and Causality

A LTI system is called stable if for any bounded input its output is also bounded. It can be shown ([27]) that a LTI system is stable if and only if

$$\sum_{k=-\infty}^{\infty} |h(n)| < \infty. \tag{2.8}$$

A cuasal system is one for which the output at time $n_0$ is only a function of the past input $n < n_0$. A LTI system is causal if and only if the impulse response $h(n)$ is zero for $n < 0$.

### Linear difference equation

A subclass of LTI systems is a one represented by the pth-order linear difference equation

$$\sum_{k=0}^{p} a_k y(n-k) = \sum_{r=0}^{q} b_r u(n-r), \tag{2.9}$$

with $a_p$ and $b_q$ not zero and $b_0 = 1$ and $a_0 = 1$ by convention.

Notice that this equation does not uniquely specify the input output relation. We can add any component that is a solution to the homogeneous equation ($u(k)$=0) and in addition have a solution that is causal or acauasal. We relate a causal LTI system to the difference equation if its homogeneous solution correspond to the condition that if $u(n < n_0) = 0$ then $y(n < n_0) = 0$.

A causal interpertation of the difference equation is

$$y(n) = -\sum_{k=1}^{p} a_k y(n-k) + \sum_{r=0}^{q} b_r u(n-r). \tag{2.10}$$

### ARMA model

Taking in the above model the input $u(k)$ to be gaussian white noise, we get the Auto-Regressive Moving Average (ARMA) model and notate $ARMA(p, q)$. This is the model we work with in this work.

**Remark 2.1** *By setting $b_r = 0$ for all $r$ we get the $AR(p)$ process and by setting $a_k = 0$ for all $k$ we get the $MA(q)$ process.*

### State-space representaion

The state-space description of SISO discrete-time finite-dimensional dynamical systems is given by equations of the form

$$\begin{aligned} x(k+1) &= f(k, x(k), u(k)) \quad (2.11) \\ y(k) &= g(k, x(k), u(k)), \quad (2.12) \end{aligned}$$

where $x(k) \in R^n$ is the state-space , $u(k)$ is the input and $y(k)$ is the output, all evaluated at time $k$ which takes integers values. Also we have the functions $f : Z \times R^n \times R \to R^n$ and $g : Z \times R^n \times R \to R$. Equation (2.11) is called the state equation and (2.12) is called the output equation.

We now restrict ourselfs to the linear time-invariant system give by

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) \quad (2.13) \\ y(k) &= Cx(k) + Du(k), \quad (2.14) \end{aligned}$$

where $A \in R^{n \times n}$, $B \in R^n$, $C^T \in R^n$ and $D \in R$. With a minor loss of generality **we take $D = 1$ from now on**.

### State-space realization of LTI difference equation

We can make a realization of the difference equation (2.10) in the form of equations (2.13) and (2.14) [13]. Define

$$A = \begin{pmatrix} -a_1 & -a_2 & \dots & -a_{n-1} & -a_n \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix}, B = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{2.15}$$

and

$$C = \begin{pmatrix} b_1 - a_1 & b_2 - a_2 & \dots & b_n - a_n \end{pmatrix}, \tag{2.16}$$

This realization is called the controller form.

## 2.2 The Transfer Function and the $z$ Transform

The $z$-transform $X(z)$ of a sequence $x(n)$ is defined as

$$\mathscr{Z}[x(n)] = X(z) = \sum_{n=-\infty}^{\infty} x(n) z^{-n} \tag{2.17}$$

where z is a complex variable. Taking $z = e^{i\theta}$ we get the fourier transform of the sequence which is a $2\pi$ periodic function of $\theta$.

**z-transform of a convolution**

As we saw, the output $y(n)$ of a LTI system with input $u(n)$ is given by the convolution of the input with the impulse response $h(n)$ of the system

$$y(n) = u(n) * h(n). \tag{2.18}$$

It can be shown (e.g [27]) that if we take the z-transform of the convolution we find that it become a multiplication

$$Y(z) = U(z)H(z). \tag{2.19}$$

Where $H(z) = \mathscr{Z}[h(n)]$ is the $z$-transform of the impulse response. $H(z)$ is called the transfer function.

### 2.2.1 Transfer Function for the ARMA Model

We take our system to be an ARMA process defined by the difference equation (2.10),

$$\sum_{k=0}^{p} a_k y(n-k) = \sum_{r=0}^{q} b_r u(n-r). \tag{2.20}$$

Taking the z-transform of this equation we get

$$H(z) = \frac{\sum_{r=0}^{q} b_r z^{-r}}{\sum_{k=0}^{p} a_k z^{-k}}, \tag{2.21}$$

or

$$H(z) = z^{p-q}\frac{b(z)}{a(z)} = z^{p-q}\frac{\prod_{i=1}^{q}(z-\beta_i)}{\prod_{i=1}^{p}(z-\alpha_i)}, \tag{2.22}$$

where $a(z) = z^p + a_1 z^{p-1} + \ldots + a_p$ and $b(z) = z^q + b_1 z^{q-1} + \ldots + b_q$. The poles $\alpha_1, \ldots, \alpha_p$ and the zeros $\beta_1, \ldots, \beta_q$ are the roots of the polynomials $a(z)$ and $b(z)$ respectively and since the polynomials are real they are all real or come in conjugate pairs. For the system to be stable the poles $\alpha_1, \ldots, \alpha_p$ should be inside the unit disk. Here we require also that the zeros $\beta_1, \ldots, \beta_q$ will be inside the unit disk making the inverse system to be stable as well. This condition is called minimum phase (see [27] for more detail).

**Remark 2.2** *By adding zeros at the origin we can absorb the $z^{p-q}$ into the definition of $b(z)$ thus making $a(z)$ and $b(z)$ of the same order. We will use both notations interchangely.*

### 2.2.2 Transfer Function for the State-Space Model

We can also write the transfer function of the system with the help of the state-space matrices $(A, B, C)$. Taking the z-transform of equations (2.13) and (2.14) and assuming zero initial conditions we get the transfer function

$$H(z) = C(zI - A)^{-1}B + 1. \tag{2.23}$$

It is straight forward to show (see e.g [24]) that $H(z)$ is the z-transform of the impulse response of the system. In addition, the poles $\alpha_1, \ldots, \alpha_p$ are the eigenvalues of $A$.

## 2.3 Stochastic Properties of the ARMA Model

**Stochastic Processes**

See e.g [25]. A discrete stochastic process is a countable series of random variables $\{X_t : t \subseteq Z\}$ where $t$ is called the time.

A stochastic process is called *stationary* if its statistics properties are invariant to a shift in the origin of $t$. That is, the processes $\{X_t\}$ and $\{X_{t+c}\}$ have the same statistics for any $c \in Z$.

**The Auto-Correlation Function**

The autocorrelation function is defined as the expected value of the product $X_{t_1}X_{t_2}$

$$R(t_1, t_2) = E\left[X_{t_1}X_{t_2}\right]. \tag{2.24}$$

For a stationary process the autocorrelation is only a function of the time difference $m = t_2 - t_1$ and we denote it as $R_{uu}(m)$. Note that $R_{uu}(0)$ is the average power.

Now consider a LTI system driven by the input $u(k)$ which we take to be a discrete-time zero-mean stationary gaussian stochastic process. We have

$$y(n) = \sum_{k=-\infty}^{\infty} h(n-k)u(k). \tag{2.25}$$

Since the input is gaussian it is fully characterized by its mean $m_u$ and the Auto-Correlation function $R_{uu}(m)$ (any other comulant vanish). The output's mean is obviously zero

$$m_y = m_u \sum_{k=-\infty}^{\infty} h(k) = 0. \tag{2.26}$$

The output Auto-Correlation is given by

$$\begin{aligned} R_{yy}(n, n+m) &= E[y(n)y(n+m)] \\ &= E\left[\sum_{k=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} h(k)h(r)u(n-k)u(n+m-r)\right] \\ &= \sum_{k=-\infty}^{\infty} h(k) \sum_{r=-\infty}^{\infty} h(r)E\left[u(n-k)u(n+m-r)\right] \\ &= \sum_{k=-\infty}^{\infty} h(k) \sum_{r=-\infty}^{\infty} h(r)R_{uu}(m+k-r) \\ &= \sum_{l=-\infty}^{\infty} R_{uu}(m-l)v(l), \end{aligned}$$

or simply

$$R_{yy}(m) = v(m) * R_{uu}(m), \tag{2.27}$$

where

$$v(l) = \sum_{k=-\infty}^{\infty} h(k)h(l+k) = h(l) * h(-l). \tag{2.28}$$

We see that $R_{yy}$ is only a function of $m$ showing that $y(n)$ is also stationary given $u(n)$ is stationary. Taking the z-transform of equation (2.27) we find

$$R_{yy}(z) = H(z)H(z^{-1})R_{uu}(z), \tag{2.29}$$

and going to the power density by puting $z = e^{i\theta}$ we have

$$P_{yy}(\theta) = |H(e^{i\theta})|^2 P_{uu}(\theta). \tag{2.30}$$

We can also regard the cross-correlation between the input and the output

$$\begin{aligned} R_{uy}(m) &= E\left[u(n)y(n+m)\right] && (2.31)\\ &= \sum_{k=-\infty}^{\infty} h(k)R_{uu}(m-k). && (2.32)\end{aligned}$$

Taking the z-transform we have

$$R_{uy}(z) = H(z)R_{uu}(z) \tag{2.33}$$

or in spectral density

$$P_{uy}(\theta) = H(e^{i\theta})P_{uu}(\theta). \tag{2.34}$$

In particular we take our input to be *white* gaussian process (i.e ARMA process), that is

$$\begin{aligned} m_u &= E[u(k)] = 0 && (2.35)\\ R_{uu}(m) &= E[u(k)u(k+m)] = \sigma_u^2\delta(m). && (2.36)\end{aligned}$$

We have $P_{uu}(\theta) = \sigma_u^2$ so

$$P_{yy}(\theta) = \sigma_u^2|H(e^{i\theta})|^2. \tag{2.37}$$

Notice also that the cross-correlation is proportional to the impulse response

$$R_{uy}(m) = \sigma_u^2 h(m) \tag{2.38}$$

and the cross power spectrum is proportional to the frequency response of the system

$$P_{uy}(\theta) = \sigma_u^2 H(e^{i\theta}). \tag{2.39}$$

## 2.4 Controllability, Observability and Realizations

In this section we define and discuss some usefull quantities regarding state-space realizations of LTI systems . All the proofs can be found at [24] unless stated otherwise.

Given an external discription of an LTI system by an impulse response, or equivalently by the transfer function, a triplet $(A, B, C)$ (with regard to the state-space model (2.13) and (2.14)) that have the desired transfer function is called a realization. It is easy to show that doing a nonsingular linear transformation $P$ to the state space $\hat{x} = Px$ of a realization does not change the input-output properties of the system (alhtough do change the computational properties such as round off errors etc.). Thus, there are infinitly many realizations to a particular external discription. On the other hand their internal response may be very different. We will use this property to choose a realization that will make calculations easier and/or will give us some more insight to the problem.

A minimum realization have the minimum dimension of the state-space with the desired transfer function. A $H(z)$ is realizable as a transfer function of SISI LTI system given by equation (2.13) and (2.14) if and only if it's a rational function of $z$ and

$$\lim_{z\to\infty} H(z) < \infty. \tag{2.40}$$

**Remark 2.3** *By setting $D = 1$ we actualy have in this work $\lim_{z\to\infty} H(z) = 1$.*

The minimum order of a SISO system with a transfer function $H(z)$ is determined by its number of poles (counting multiplicity).

**Remark 2.4** *(Poles multiplicity.) The poles are the eigenvalues of $A$ and therefore if A has an eigenvalue $\alpha_i$ with multiplicity $n > 1$ we get $(z - \alpha_i)^n$ in the denominator of the transfer function $H(z)$. Inspecting the Jordan form of $A$ we may conclude that there may be different possible realizations with the above* *multiplicity. For example for $n = 2$ there are 2 possibilities*

$$A^{(1)} = \begin{pmatrix} \alpha_i & 0 \\ 0 & \alpha_i \end{pmatrix} \quad and \quad A^{(2)} = \begin{pmatrix} \alpha_i & 1 \\ 0 & \alpha_i \end{pmatrix}. \tag{2.41}$$

*But for* SISO, *demending the realization to be of minimum order we find that only $A^{(2)}$ is possible. This is true for all $n$ meaning we need to take the largest Jordan block form.*

### Controllability

**Definition 2.1** *A state $x_0$ is called controllable if there is an input $u(n)$ that transfer the state from $x_0$ to the zero state in some finite number of steps.*

**Remark 2.5** *A state $x_0$ is called reachable if there is an input that take the system from the zero state to $x_0$ in a finite number of steps. Reachability imply controllability but vice versa only if the matrix $A$ is nonsingular. Here we work only with nonsingular $A$.*

**Definition 2.2** *A system is said to be controllable if any state in the state-space $x$ is controllable.*

**Theorem 2.1** *A system is controllable if and only if its* controllability matrix $\mathscr{C}$

$$\mathscr{C} \equiv \left[B,\ AB,\ A^2B,\ \ldots\right] \tag{2.42}$$

*has full raw rank $n$*

$$rank\ \mathscr{C} = n. \tag{2.43}$$

**Definition 2.3** *The* controllability Gramian *is defined as*

$$W^c = \sum_{i=1}^{\infty} A^i BB^T (A^T)^i = \mathscr{C}\mathscr{C}^T \in R^{n \times n}. \tag{2.44}$$

**Corollary 2.1** *A system is controllable if and only if its Controllabilty Gramian is of full rank $n$.*

### Observabillity

**Definition 2.4** *A state $x$ is called* unobservable *if the zero input response of the system to $x$ is zero for all $k \geq 0$,*

$$CA^k x = 0 \quad for\ every\ k \geq 0. \tag{2.45}$$

**Definition 2.5** *A system is said to be* observable *if the only state that is unobservable is the zero state $x = 0$.*

**Theorem 2.2** *A system is observable if and only if its* Observabillity matrix $\mathscr{O}$

$$\mathscr{O} \equiv \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \end{bmatrix} \tag{2.46}$$

*has full column rank $n$*

$$rank\ \mathscr{O} = n. \tag{2.47}$$

**Definition 2.6** *The* Observabillity Gramian *is defined as*

$$W^c = \sum_{i=1}^{\infty} (A^T)^i C^T C A^i = \mathscr{O}^T \mathscr{O} \in R^{n \times n}. \tag{2.48}$$

**Corollary 2.2** *A system is observable if and only if its Observabillity Gramian is of full rank $n$.*

**Theorem 2.3** *An n-dimensional realization $\{A, B, C\}$ of $H(z)$ is minimal if and only if it is both controllable and observable.*

**Remark 2.6 (Balanced realization and model reduction).** *Balanced realization was presented by Moore [23] and it correspondes to taking a realization such that both the Gramians $W^c$ and $W^o$ are diagonal and equal $W^c = W^o = \Sigma$. such a realization is always possible as shown by Moore. It has the property that it is controllable as observable. Moore suggeted a scheme for model reduction by removing the most uncontrollable, and consequently also the most unobservable, part of the system by truncating the state space in that realization resulting in an approximated lower order system although not in an optimal fashion. It is shown in [8] that it is also suited for stochastic model reduction where we want to take only the most informative part between the past and the future. Indeed, as was shown in [9] and will be discussed briefly in chapter 4, the information bottleneck method generelize this approach.*

**Definition 2.7** *We define the past of a process $y(k)$ to be $y_p = (y(-1), y(-2), \ldots)^T$ and future $y_f = (y(0), y(1), \ldots)^T$ and analogously for $u(k)$.*

**Definition 2.8** *The Hankel operator is given by*

$$H = \mathscr{O}\mathscr{C} \tag{2.49}$$

*or in terms of the system matrices*

$$H = \begin{pmatrix} CB & CAB & CA^2B & \ldots \\ CAB & CA^2B & CA^3B & \ldots \\ CA^2B & CA^3B & CA^4B & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{2.50}$$

If we truncate the input at time zero, the Hankel operator maps the input past into the output future. The Hankel operator play a major role in the theory of model reduction (see Moore [23]) since it is much easier to deal with.

We also define the matrix

$$\Delta = \begin{pmatrix} 1 & 0 & 0 & \dots \\ CB & 1 & 0 & \dots \\ CAB & CB & 1 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{2.51}$$

Using those definitions we have

$$y_f = Hu_p + \Delta u_f. \tag{2.52}$$

**The inverse system** The inverse model has the transfer function $H_{inverse}(z) = H^{-1}(z)$. Since the system is minimum phase also the inverse one is. The system matrices are related by

$$A_{inverse} = A - BC \tag{2.53}$$

$$B_{inverse} = B \tag{2.54}$$

$$C_{inverse} = -C. \tag{2.55}$$

It's easy to see that the observability matrix of the inverse model $\mathscr{O}_z$ is related to the observability matrix by

$$\mathscr{O}_z = -\Delta^{-1}\mathscr{O}. \tag{2.56}$$

We will also need the inverse model observability Gramian

$$W_z^o = \mathscr{O}_z^T \mathscr{O}_z. \tag{2.57}$$

# Chapter 3

# Information Theory and CCA

In this chapter we first define and discuss the information theoretic quantities needed; The Entropy of discrete random variables, differential entropy of continuos random variables, entropy of Gaussian variables, mutual information, relative entropy, chain rules and more. The standart references for those quantities are the book by Cover and Thomas [10] and the original paper by Shannon [11]. Reader who is familiar with information theory may skip directly to section 3.3. Next we discuss the notion of canonical correlations (cc) presented by Hotelling [12] which are closely related to the information between gaussian variables. As we will see the cc play a major role in formulating the problem in this work. Finally we discuss the information between the past and the future of a gaussian stochastic process.

## 3.1 Entropy

### 3.1.1 Entropy of a Discrete Random Variable

The entropy is a measure of uncertainty of a random variable and so is a measure of the amount of information required on the average to describe a random variable. Let $X$ be a discrete random variable with alphabet $\mathscr{X}$ and probability mass function $p(x) = Pr\{X = x\}, x \in \mathscr{X}$.

**Definition 3.1** *The entropy $H(X)$ of a discrete random variable $X$ is defined by*

$$H(X) = -\sum_{x \in \mathscr{X}} p(x) \log p(x). \tag{3.1}$$

The $\log$ is in base 2 and the entropy is measured in bits. We use the convention that $0 \log 0 = 0$.
We now extend the definition to a pair of random variables. (of course this is just a semantic change since we could regard the pair as a single random variable and use the definition above)

**Definition 3.2** *The joint entropy $H(X, Y)$ of a pair of discrete random variables $(X, Y)$ with joint distribution $p(x, y)$ is defined as*

$$H(X, Y) = -\sum_{x \in \mathscr{X}} \sum_{y \in \mathscr{Y}} p(x, y) \log p(x, y), \tag{3.2}$$

*which can also be expressed as*

$$H(X, Y) = -E[\log p(x, y)]. \tag{3.3}$$

We also define the conditional entropy of a random variable given another as the expected value of the entropies of the conditional distributions, averaged over the conditioning random variable.

**Definition 3.3** *Let $(X, Y)$ be random variables with mass distribution $p(x, y)$,* *we define the conditional entropy $H(Y|X)$ as*

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathscr{X}} p(x) H(Y|X = x) && (3.4) \\ &= -\sum_{x \in \mathscr{X}} p(x) \sum_{y \in \mathscr{Y}} p(y|x) \log p(y|x) && (3.5) \\ &= -\sum_{x \in \mathscr{X}} \sum_{y \in \mathscr{Y}} p(y,x) \log p(y|x) && (3.6) \\ &= -E_{p(x,y)}[\log p(Y|X)]. && (3.7) \end{aligned}$$

Using those definitions it's easy to prove the intuitive chain rule for entropy.

**Theorem 3.1 (Chain Rule for Entropy)**

$$H(X,Y) = H(X) + H(Y|X). \tag{3.8}$$

**Remark 3.1 (The entropy function as a natural choise for uncertainty).** *Shannon showed ([14] and [11]) that in order for the definition of uncertainty to be consistent we are forced to use the above definition of entropy (up to multiplicative constant)*

$$H(p_1, \ldots, p_n) = -\sum_i p_i \log p_i, \tag{3.9}$$

*where our variable takes the values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$ respectively. The consistency is represented by three properties we demand the uncertaity function to obey:*

1. *H should be a continuous function of the $p_i$.*
2. *If all $p_i$ are equal, the quantity $H(1/n, \ldots, 1/n)$ is a monotonic increasing function of $n$.*
3. *The composition law. For the definition to be consistent we must have that if we group events in different ways we still need to get the same*

*answer. Let us for example group the first $k$ events $x_1 \ldots, x_k$ into one event with probability $w_1 = p_1 + \ldots + p_k$, the next group as the $m$ events $x_{k+1}, \ldots, x_{k+m}$ with probability $w_2 = p_{k+1} + \ldots + p_{k+m}$ and so on such that we have $r$ groups in total. Given only the grouped events we have the uncertainty $H(w_1, \ldots, w_r)$. We now give the conditional probabilities $(p_1/w_1, \ldots, p_k/w_1)$ given that the first group has occurred and we do the same to all groups. In this state of affairs we should have the same amount of knowledge as if we were given $(p_1, \ldots, p_n)$ directly. Formally we should have*

$$\begin{aligned} H(p_1, \ldots, p_n) = H(w_1, \ldots, w_r) + w_1 H(p1/w_1, \ldots, p_k/w_1) \\ + w_2 H(p_{k+1}/w_2, \ldots, p_{k+m}/w_2) + \ldots . \end{aligned} \tag{3.10}$$

*The only function that obey those properties is the entropy function given above.*

### 3.1.2 Entropy for Continuous Variables

For continuous random variable we use the differential entropy. First we define what continuous random variable is. Next we define the differential entropy and finally we discuss it's relation to discrete entropy.

**Definition 3.4** *Let $X$ be a random variable with cumulative distribution function $F(x) = Pr(X \leq x)$. If $F(x)$ is continuous, the random variable is said to be continuous. Let $f(x) = F'(x)$ when the derivative is defined. If $\int_{-\infty}^{\infty} f(x) = 1$, then $f(x)$ is called the* probability density function *for $X$. The set where $f(x) > 0$ is called the* support set *of X.*

**Definition 3.5** *The* differential entropy *$h(x)$ of a continuous random variable $X$ with a density $f(x)$ is defined as*

$$h(x) = -\int_S f(x) \log f(x) \, dx \ , \tag{3.11}$$

*where $S$ is the support set of the random variable.*

To see the connection to discrete entropy let us consider the random variable with density f(x) (see figure 3.1). We divide the range of $X$ into bins of length $\Delta$. Let us assume that the density is continuous within the bins. By the mean value theorem, there is a value $x_i$ within each bin such that

$$f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)\ dx. \tag{3.12}$$

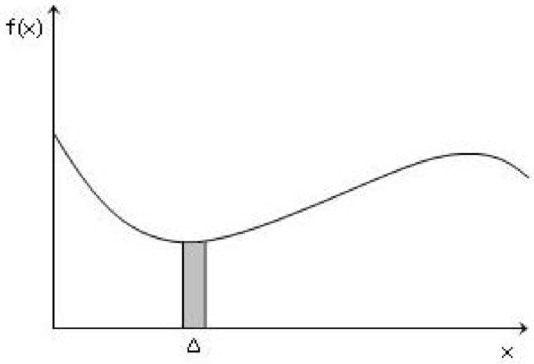

Figure 3.1: Quantization of a continuous random variable.

We now consider the discrete variable

$$X^{\Delta} = x_i, \quad if \quad i\Delta \leq X < (i+1)\Delta. \tag{3.13}$$

The probability mass function is

$$p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)\ dx. \tag{3.14}$$

The entropy of $X^{\Delta}$ is

$$
\begin{aligned}
H(X^{\Delta}) &= -\sum_{-\infty}^{\infty} p_i \log p_i && (3.15)\\
&= -\sum_{-\infty}^{\infty} f(x_i)\Delta \log(f(x_i)\Delta) && (3.16)\\
&= -\sum \Delta f(x_i) \log f(x_i) - \sum f(x_i)\Delta log\Delta && (3.17)\\
&= -\sum \Delta f(x_i) \log f(x_i) - log\Delta, && (3.18)
\end{aligned}
$$

where we used $\sum f(x_i)\Delta = \int f(x) = 1$. The first term just approach the differential entropy as $\Delta \to 0$ while the other term become infinite.
Thus we see that the actual entropy of a continuous variable is infinite.

**Remark 3.2** *Notice that $h(X)$ can be negetive while $H(X)$ is always non-negetive.*

An extension to more then one variable is straight forward. Also we can define the conditional differentail entropy:

**Definition 3.6** *If $X, Y$ have a joint density function $f(x, y)$, we can define the conditional differential entropy $h(X|Y)$ as*

$$h(X|Y) = -\int f(x, y) \log f(x|y) dxdy. \tag{3.19}$$

*The differential entropy obey the chain rule just like the discrete entropy (follows directly from the definitions).*

### 3.1.3 Entropy for Gaussian Variables

The probability density function of the Gaussian variables $X_1, X_2, \ldots, X_n$ (multivariae normal distribution) is of the form

$$f(\mathbf{x}) = \frac{1}{(\sqrt{2}\pi)^n |\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)} \ , \tag{3.20}$$

where $\mathbf{x} = (x_1, \ldots, x_n)^T$ , $\mu$ is the mean vector and $\Sigma$ is the symmetric covariance matrix:

$$\Sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)]. \tag{3.21}$$

$|\cdot|$ denotes determinant. We use $\mathscr{N}_n(\mu, \Sigma)$ to denote this distribution (see appendix B for details).

**Theorem 3.2 (Entropy of a multivariae normal distribution.)** *Let $X_1, X_2, \ldots, X_n$ have a multivariate normal distribution with mean $\mu$ and covariance matrix $\Sigma$. Then*

$$h(X_1, X_2, \ldots, X_n) = h(\mathscr{N}_n(\mu, \Sigma)) = \frac{1}{2}\log(2\pi e)^n|\Sigma|. \tag{3.22}$$

**Proof:** Using the definition of differential entropy

$$\begin{aligned}
h(f) &= -\int f(\mathbf{x})\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu) - \ln(\sqrt{2}\pi)^n|\Sigma|^{1/2}\right]d\mathbf{x} && (3.23)\\
&= \frac{1}{2}E\left[\sum_{i,j}(x_i-\mu_i)(\Sigma^{-1})_{ij}(x_j-\mu_j)\right] + \frac{1}{2}\ln(2\pi)^n|\Sigma| && (3.24)\\
&= \frac{1}{2}\sum_{i,j}E[(x_j-\mu_j)(x_i-\mu_i)](\Sigma^{-1})_{ij} + \frac{1}{2}\ln(2\pi)^n|\Sigma| && (3.25)\\
&= \frac{1}{2}\sum_i\sum_j\Sigma_{ji}(\Sigma^{-1})_{ij} + \frac{1}{2}\ln(2\pi)^n|\Sigma| && (3.26)\\
&= \frac{1}{2}\sum_j(\Sigma\Sigma^{-1})_{jj} + \frac{1}{2}\ln(2\pi)^n|\Sigma| && (3.27)\\
&= \frac{1}{2}\sum_j I_{jj} + \frac{1}{2}\ln(2\pi)^n|\Sigma| && (3.28)\\
&= \frac{n}{2} + \frac{1}{2}\ln(2\pi)^n|\Sigma| && (3.29)\\
&= \frac{1}{2}\ln(2\pi e)^n|\Sigma| \quad nats && (3.30)\\
&= \frac{1}{2}\log(2\pi e)^n|\Sigma| \quad bits. \quad \square && (3.31)
\end{aligned}$$

## 3.2 Mutual Information

### 3.2.1 Relative Entropy and Mutual Information

We start with a definition of *relative entropy* $D_{KL}[p\|q]$ which is a measure of the "distance" between two distributions $p$ and $q$. It can be interperted as the representation inefficiency by assuming that the distribution is $q$ when the true distribution is actually $p$. If we knew that the true distribution of a random variable is $p$ we could code the variable with an average length of $H(p)$. If we used the distribution $q$ instead, we would need $H(p) + D_{KL}[p\|q]$ bits on the average to describe the variable.

**Definition 3.7** *The relative entropy or Kullback Leilbler distance between two probability mass functions $p(x)$ and $q(x)$ is defined as*

$$D_{KL}[p\|q] = \sum_{x \in \mathscr{X}} p(x) \log \frac{p(x)}{q(x)}. \tag{3.32}$$

It can be shown that relative entropy is always non-negetive and is zero if and only if $p = q$. Notice however that it is not a true distance since it is not symmetric and do not satisfy the triangle inequality.

Next we define the mutual Information, which is a measure of the amount of information that one variable contains about another random variable. It is the reducion of the uncertainty of one variable due to knowledge of the other.

**Definition 3.8** *Let $X$ and $Y$ be two random variables with joint probability mass function $p(x, y)$ and the marginals $p(x)$ and $q(x)$. The mutual information $I(X;Y)$ is the relative entropy between the joint distribution $p(x, y)$ and the product of the marginals $p(x)p(y)$:*

$$I(X;Y) = \sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{3.33}$$

$$= D_{KL}[p(x,y)\|p(x)p(y)]. \tag{3.34}$$

It is a measure of how much the two variables are far from being independent. $X$ and $Y$ are independent if and only if $I(X;Y) = 0$. Notice that this quantity is symmetric and non-negetive.

### 3.2.2 Mutual Information and Entropy

We can write the definition of mutual information with the help of entropies:

$$I(X;Y) = H(X) - H(X|Y) \tag{3.35}$$

$$= H(Y) - H(Y|X) \tag{3.36}$$

$$= H(X) + H(Y) - H(X,Y). \tag{3.37}$$

**Corollary 3.1 (Conditioning reduce entropy).** *From the non-negetivity of the mutual information we get that conditioning reduce entropy*

$$H(X) \geq H(X|Y). \tag{3.38}$$

### 3.2.3 Chain Rules

We now extend the chain rule for entropy for more then two variables and present the chain rule for mutual information.

**Theorem 3.3 (Chain rule for entropy).** *Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, , x_n)$. Then*

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1). \tag{3.39}$$

**Definition 3.9** *The conditional mutual information between $X$ and $Y$ given $Z$ is defined as*

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z). \tag{3.40}$$

**Theorem 3.4** *(Chain rule for mutual information)*

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, X_{i-2}, \ldots, X_1). \tag{3.41}$$

### 3.2.4 Data Processing Inequality

We first define Markov chains and then present the data processing inequality which can be used to show that no manipulation of the data can improve the inferences that can be made from the data.

**Definition 3.10 (Markov chains).** *Random variables $X, Y, Z$ are said to form a Markov chain in that order (denoted by $X \rightarrow Y \rightarrow Z$)) if the conditional distribution of $Z$ depends only on $Y$ and is conditionally independent of $X$. That is*

$$p(x,y,z) = p(y)p(x|y)p(z|y). \tag{3.42}$$

**Remark 3.3** *Notice that if $X \rightarrow Y \rightarrow Z$ then also $Z \rightarrow Y \rightarrow X$.*

**Theorem 3.5 (Data processing inequality).** *If $X \rightarrow Y \rightarrow Z$, especially if $Z = f(Y)$, then*

$$I(Y;Z) \geq I(X;Z). \tag{3.43}$$

### 3.2.5 Sufficient Statistics

We now use the data processing inequality to present an important idea in statistics, suffecient statistics. Here we deal with a slightly different notion then the usual one used in parametric estimation problems. Suppuse we have a Markov chain $u \rightarrow x \rightarrow y$. By the data processing inequality we have

$$I(x;y) \geq I(u;y). \tag{3.44}$$

$x$ will be called suffcient statistic if an equality holds in the data processing inequality.

$$I(x;y) = I(u;y). \tag{3.45}$$

This is equivalent to the condition that given $x$, $u$ and $y$ are independent. We will use this notion to emphasize that the state space of a linear model is suffcient statistcs from the past about the future.

### 3.2.6 Mutual Information for Continuous Variables

We now extend the definitions of relative entropy and mutual information to probability densities.

**Definition 3.11** *The relative entropy (or Kullback Leibler distance) $D_{KL}[f\|g]$ between two densities $f$ and $g$ is defined by*

$$D_{KL}[f\|g] = \int f \log \frac{f}{g}. \tag{3.46}$$

Note that $D_{KL}[f\|g]$ is finite only if the support set of $f$ is contained in the support set of $g$.

**Definition 3.12** *The mutual information $I(X;Y)$ between two random variables with joint density $f(x,y)$ is defined as*

$$I(X;Y) = \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy. \tag{3.47}$$

We see that

$$I(X;Y) = h(X) - h(X|Y) \tag{3.48}$$

and

$$I(X;Y) = D_{KL}[f(x,y)\|f(x)f(y)]. \tag{3.49}$$

Those definitions have the same properties as for discrete variables. For example for the mutual information we have for the quantized version of $X$ and

$Y$

$$\begin{aligned} I(X^{\Delta}; Y^{\Delta}) &= H(X^{\Delta}) - H(X^{\Delta}; Y^{\Delta}) && (3.50)\\ &\approx h(X) - \log \Delta - (h(X|Y) - \log \Delta) && (3.51)\\ &= I(X;Y). && (3.52) \end{aligned}$$

It is also non-negetive and thus we have that conditioning reduce entropy $h(X) \geq h(X|Y)$.

### 3.2.7 Mutual Information for Gaussian Variables

**Theorem 3.6 (Mutual information between Gaussian variables.)** *Let $X$ be a $n_x$ dimensional gaussian variable with covariance matrix $\Sigma_X$, $Y$ be a $n_y$ dimensional gaussian variable with covariance matrix $\Sigma_Y$ and the cross covariance matrix $\Sigma_{XY}$ (The information is independent of the means). Then*

$$I(X;Y) = -\frac{1}{2} \log |I - \Sigma_X^{-1} \Sigma_{XY} \Sigma_Y^{-1} \Sigma_{YX}|. \qquad (3.53)$$

**Proof:** We know that (see 3.48)

$$I(X;Y) = h(X) - h(X|Y). \qquad (3.54)$$

For gaussian variables we get (see 3.22)

$$I(X;Y) = \frac{1}{2} \log |\Sigma_X| - \frac{1}{2} \log |\Sigma_{X|Y}|. \qquad (3.55)$$

Using the schur complement formula we get that the covariance $\Sigma_{X|Y}$ is (see appendix B)

$$\Sigma_{X|Y} = \Sigma_X - \Sigma_{XY} \Sigma_Y^{-1} \Sigma_{YX}. \qquad (3.56)$$

And thus we have

$$
\begin{aligned}
I(X;Y) &= h(X) - h(X|Y) && (3.57)\\
&= \frac{1}{2}\log|\Sigma_X| - \frac{1}{2}\log|\Sigma_{X|Y}| && (3.58)\\
&= -\frac{1}{2}\log|\Sigma_X|^{-1}|\Sigma_{X|Y}| && (3.59)\\
&= -\frac{1}{2}\log|\Sigma_X|^{-1}|\Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}| && (3.60)\\
&= -\frac{1}{2}\log|I - \Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}| \quad \square. && (3.61)
\end{aligned}
$$

Thus what we need to know are the eigenvalues of the matrix $\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}$. Those are called The caconical correlations between $X$ and $Y$ and are discussed in the next section.

## 3.3 Canonical Correlations Analysis

### 3.3.1 The Canonical Correlations

For two random variables $X_1$ and $X_2$, the correlation coeffecient is

$$
\rho_{12} \equiv cor(X_1, X_2) = \frac{\Sigma_{12}}{\sqrt{\Sigma_{11}}\sqrt{\Sigma_{22}}}, \qquad (3.62)
$$

where the $\Sigma_{ij}$ are the covariances between the variables, $E(x_i - \mu_i)(x_j - \mu_j)$, and $\mu$ are the averages. It can be shown that $-1 \leq \rho_{12} \leq 1$. The covariance between variables measure the linear dependence between those variables. If the two variables are independent then the correlation is zero but not vice versa. It can be that two random variables will be dependent while their correlation will be zero (e.g. when they have only quadradric dependence). It is well known, howerver, that for gaussian variables the correlations are zero if and only if the variables are independent.

Let $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_m\}$ be two sets of random variables. In the canonical correlation analysis [12] we want to maximaize the correlations between those two sets by introducing linear combinations $x^{(1)} = \sum_{i=1}^{n} a_i^{(1)} x_i$ and $y^{(1)} = \sum_{i=1}^{n} b_i^{(1)} y_i$ and finding:

$$(\hat{\mathbf{a}}^{(1)}, \hat{\mathbf{b}}^{(1)}) = arg \max_{\mathbf{a}^{(1)}, \mathbf{b}^{(1)}} cor(x^{(1)}, y^{(1)}). \tag{3.63}$$

$\hat{x}^{(1)}$ and $\hat{y}^{(1)}$ are the first pair of canoniacl variables. Next we find other linear combination coefficients $\hat{\mathbf{a}}^{(2)}$ and $\hat{\mathbf{b}}^{(2)}$ again maximazing the correlation with the constrain that the combinations are uncorrelated to the first pair. This goes on $\min\{n, m\}$ times. the optimal correlations

$$\rho_i = cor(\hat{x}^{(i)}, \hat{y}^{(i)}) \qquad i = 1, \ldots, \min\{n, m\}$$

are called the *canonical correlations*.

**Characterizing the canonical correlations.** Let us denote $\Sigma_{XY}$ the covariance matrix between $X$ and $Y$ and analogously $\Sigma_{XX}$ and $\Sigma_{YY}$. We have

$$cor(x, y) = \frac{\mathbf{a}^T \Sigma_{XY} \mathbf{b}}{\sqrt{\mathbf{a}^T \Sigma_{XX} \mathbf{a}} \sqrt{\mathbf{b}^T \Sigma_{YY} \mathbf{b}}}. \tag{3.64}$$

Notice that this quantity is invariant under rescaling of **a** and **b**. Thus maximazing $cor(x, y)$ is equivalent to maximazing $cov(x, y)$ under the constrains $\sqrt{\mathbf{a}^T \Sigma_{XX} \mathbf{a}} = 1$ and $\sqrt{\mathbf{b}^T \Sigma_{YY} \mathbf{b}} = 1$. Using Lagrange multipliers we maximazie the lagrangian

$$L(\rho_x, \rho_y, \mathbf{a}, \mathbf{b}) = \mathbf{a}^T \Sigma_{XY} \mathbf{b} - \frac{\rho_x}{2}(\mathbf{a}^T \Sigma_{XX} \mathbf{a} - 1) - \frac{\rho_y}{2}(\mathbf{b}^T \Sigma_{YY} \mathbf{b} - 1). \tag{3.65}$$

Taking derivatives with respect to **a** and **b** and equating to zero we get

$$\frac{\partial L}{\partial \mathbf{a}} = \Sigma_{XY}\mathbf{b} - \rho_x \Sigma_{XX}\mathbf{a} = 0 \tag{3.66}$$

$$\frac{\partial L}{\partial \mathbf{b}} = \Sigma_{YX}\mathbf{a} - \rho_y \Sigma_{YY}\mathbf{b} = 0. \tag{3.67}$$

multiplying the first equation by $\mathbf{a}^T$, the second by $\mathbf{b}^T$ and substracting we get the we must have $\rho_x = \rho_y \equiv \rho$. Substituing the second equation, $\mathbf{b} = \frac{\Sigma_{YY}^{-1}\Sigma_{YX}}{\rho}\mathbf{a}$, in the first we get thr eigenvalue problem for **a**

$$(\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} - \rho^2 I)\mathbf{a} = 0. \tag{3.68}$$

and similarly for **b**

$$(\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} - \rho^2 I)\mathbf{b} = 0. \tag{3.69}$$

Inspecting again the multiplied equations (3.66) and (3.67) and using the constrains we see that the eigenvalues are the squared canonical coefficients as the notation suggested. It can be shown that those eigenvalues are between $0$ and $1$. The eigenvectors are the coefficients in the linear combinations. Also, since the matrix $\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$ is symmetric it can be shown that the different linear combinations $x^{(j)} = \sum_{i=1}^{n} a_i^{(j)} x_i$ are uncorrelated (and also for $y$). The first canonial pair is choosen as the one with the greatest eigenvalue and the other canonical variables in decsending order.

Remembering the expresion for the mutual information between two set of gaussian vaiables, equation (3.53), we find that

**Theorem 3.7 (Mutual information for gaussian variables using the canonical correlations.)** *For $X_1$ and $X_2$ two sets of gaussian variables, the mutual information is given by*

$$I(X_1; X_2) = -\frac{1}{2}\log \prod_{i=1}^{n}(1-\rho_i^2), \tag{3.70}$$

*where $\rho_i$ are the canonical correlations.*

**Prrof:** It is a direct result of equation (3.53) using the well known result that a determinant of a symmetric matrix is the product of it's eigenvalues. □

**Definition 3.13** *We will notate the (squared) canonical correlations as eigenvalues*

$$cc^2(X,Y) = eig\left[\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\right]. \tag{3.71}$$

**Remark 3.4** *Notice that $cc(X,Y)$ is symmetric in $X$ and $Y$.*

### 3.3.2 Geometric View of CC

As mentioned above, for gaussian variables the correlations are zero if and only if the variables are independent. In addition, any linear combination of zero mean Gaussian variables is also a zero mean Gaussian variable (this additivity is also true for non zero mean Gaussian variables but here we concern only zero mean ones).

Thus it is very usefull to regard Gaussian variables as spaning a (possibly infinite) linear space (an Hilbert space). We can relate the standart orthogonormal base $\mathbf{e}_i$ in that space as independet gaussian variables $u_i$ with unit covariance. Hence we have the analogy

$$< \mathbf{e}_i, \mathbf{e}_j >= \delta_{ij} \Longleftrightarrow cov(u_i, u_j) = \delta_{ij}, \tag{3.72}$$

where $< \cdot, \cdot >$ is the inner product. Any Gaussian variable can be regarded as a vector in that space $X_1 = \sum_i a_i u_i$. We will use the covariance and the inner product interchanglly meaning the same thing: $< X_1, X_2 >= cov(X_1, X_2)$.

We see that the correlation coeffecient between two Gaussian variables is just the angle cosine between the two vector $X_1$ and $X_2$ in the above space.

$$\rho_{12} = \cos\theta_{12} = \frac{< X_1, X_2 >}{\sqrt{< X_1, X_1 >}\sqrt{< X_2, X_2 >}} \tag{3.73}$$

Thus the canonical correlations between two sets of Gaussian variables can be regarded as the angles between the subspaces spaned by the two sets of vectors $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_m\}$ (see e.g [17] for more details).

## 3.4 Past-Future Mutual Information

The main quantitiy we use in this thesis is the information between the past and the future of a process $I(y_p; y_f)$ (or $I(u_p; y_f)$) [32], [35], [36]. The information between the past input and the future output is given by

$$I(u_p, y_f) = h(y_f) - h(y_f|u_p), \tag{3.74}$$

where $h$ is the differential entropy. Since we deal with gaussian variables we have

$$h(y_f) = \frac{1}{2}\log 2\pi e|\Sigma_{y_f}| \tag{3.75}$$

and

$$h(y_f|u_p) = \frac{1}{2}\log 2\pi e|\Sigma_{y_f|u_p}| \tag{3.76}$$

where $|\cdot|$ denotes a determinant and the $\Sigma's$ are the covariance matrices. Using schur complement formula we have $\Sigma_{y_f|u_p} = \Sigma_{y_f} - \Sigma_{y_f u_p}\Sigma_{u_p}^{-1}\Sigma_{u_p y_f}$ and thus we get

$$I(u_p, y_f) = -\frac{1}{2}\log|I - \Sigma_{y_f}^{-1}\Sigma_{y_f u_p}\Sigma_{u_p}^{-1}\Sigma_{u_p y_f}|. \tag{3.77}$$

The (squared) canonical correlations between the **past input** and **future output** are given by the eigenvalues of

$$cc^2(u_p, y_f) = eig[\Sigma_{y_f}^{-1}\Sigma_{y_f u_p}\Sigma_{u_p}^{-1}\Sigma_{u_p y_f}]. \tag{3.78}$$

The (squared) canonical correlations between the **future input** and **future output** are given by the eigenvalues of

$$cc^2(u_f, y_f) = eig[\Sigma_{y_f}^{-1}\Sigma_{y_f u_f}\Sigma_{u_f}^{-1}\Sigma_{u_f y_f}]. \tag{3.79}$$

We define analogously $cc(y_p, y_f)$ and $cc(u_p, u_f)$. If the (minimal) system is of order $n$ there are $n$ nonzero canonical correlations $cc^2(u_p, y_f) = (\rho_1^2, \ldots, \rho_n^2)$. Thus we have

$$I(u_p; y_f) = -\frac{1}{2} log \prod_{i=1}^{n}(1 - \rho_i^2). \tag{3.80}$$

**Calculating the Canonical Correlations $\mathbf{cc(u_p, y_f)}$** (see also [33] and [34] or [18]). As we saw in chapter 2 we have

$$y_f = H u_p + \Delta u_f. \tag{3.81}$$

Define $\Sigma = E\left[x(0)x(0)^T\right]$ which is also equal to $\Sigma = \mathscr{C}\mathscr{C}^T = W^c$ (where $E\left[\ldots\right]$ denotes averaging over the inputs $u's$). We have

$$\begin{aligned}
\Sigma_{u_p} &= I, \\
\Sigma_{u_p y_f} &= H^T = (\mathscr{O}\mathscr{C})^T, \\
\Sigma_{y_f} &= HH^T + \Delta\Delta^T = \mathscr{O}\Sigma\mathscr{O}^T + \Delta\Delta^T, \\
\Sigma_{y_f u_p} &= H = \mathscr{O}\mathscr{C}.
\end{aligned} \tag{3.82}$$

Thus we need to find the eigenvalues of

$$H^T(HH^T + \Delta\Delta^T)^{-1}H = \mathscr{C}^T\mathscr{O}^T(\mathscr{O}\Sigma\mathscr{O}^T + \Delta\Delta^T)^{-1}\mathscr{O}\mathscr{C} \tag{3.83}$$

This matrix has rank $n$ thus it have infinitly many zero eigenvalues and only $n$ posssibly eigenvalues different from zero which are the eigenvalues of the matrix

$$\Sigma\mathscr{O}^T(\mathscr{O}\Sigma\mathscr{O}^T + \Delta\Delta^T)^{-1}\mathscr{O}. \tag{3.84}$$

From the definition it's easy to see that $\Delta$ is reversible and this become

$$\Sigma\mathscr{O}^T\Delta^{-T}(\Delta^{-1}\mathscr{O}\Sigma\mathscr{O}^T\Delta^{-T}+I)^{-1}\Delta^{-1}\mathscr{O}. \tag{3.85}$$

As we saw in chapter 2 we have $\mathscr{O}_z=-\Delta^{-1}\mathscr{O}$ and this can be written as

$$\Sigma\mathscr{O}_z^T(\mathscr{O}_z\Sigma\mathscr{O}_z^T+I)^{-1}\mathscr{O}_z. \tag{3.86}$$

Now we use the inversion formula.

**Lemma 3.1** *(***The Matrix Inversion Formula.***) For non-singular matrices $A\in\mathbb{R}^{q\times q}$ and $R\in\mathbb{R}^{p\times p}$ we have*

$$(A+BRC)^{-1}=A^{-1}-A^{-1}B(R^{-1}+CA^{-1}B)^{-1}CA^{-1}.$$

Taking in the above formula $A=I$, $R=\Sigma$, $B=\mathscr{O}_z$ and $C=\mathscr{O}_z^T$ and remembering that $W_z^o=\mathscr{O}_z^T\mathscr{O}_z\equiv\mathscr{G}_z$ and $W^c=\mathscr{C}\mathscr{C}^T=\Sigma$ we get

$$cc^2(u_p,y_f)=eig[\mathscr{G}_z\Sigma(I+\mathscr{G}_z\Sigma)^{-1}],0,0,\ldots \tag{3.87}$$

Doing exactly the same steps for $cc(u_f,y_f)$ we get

$$cc^2(u_f,y_f)=eig[(I+\mathscr{G}_z\Sigma)^{-1}],1,1,\ldots \tag{3.88}$$

It can be shown that there is a realization such that $W_z^o=\mathscr{G}_z$ and $W^c=\Sigma$ are diagonal and equal.

**Theorem 3.8 (Properites of the canonical corelations.)**

- *The $cc(u_p,u_f)$ are all zeros.*
- *The $cc(u_p,y_p)$ are all ones.*
- $cc(u_p,y_f)$=$cc(y_p,y_f)$.

- $cc^2(u_f, y_f) = 1 - cc^2(u_p, y_f)$.

**Proof:** The $cc(u_p, u_f)$ are all zeros since $u_p$ and $u_f$ are independent. The $cc(u_p, y_p)$ are all ones since all $y_p$ are functions only of $u_p$. $cc(u_p, y_f)$=$cc(y_p, y_f)$ since the space spaned by $u_p$ and $y_p$ is equal. The last property is a very important evident to be used when formulating the problem. It's evident by considering eq(3.87) and eq(3.88). □

**Remark 3.5 (A property of the canonical correlations between the outputs of two ARMA models.)** *Let us consider the canonical correlation between the future outputs, $y_f^{(1)}$and $y_f^{(2)}$, of two AR models, $H^{(1)}$ and $H^{(2)}$, of order $n_1$ and $n_2$ respectively, driven by the same white Gaussian input $u(k)$. With a little abuse of notation those are denoted by $cc(H^{(1)}, H^{(2)})$, where it is understood that we concern the future outputs of those systems.*
*Now, it can be shown (see [19]) that if we feed both the systems with the Gaussian white noise $u(k)$ filterd by an AR model $H^{(3)}$, then the canonical correaltions of the outputs does not change. That is*

$$cc(H^{(1)}, H^{(2)}) = cc(H^{(3)}H^{(1)}, H^{(3)}H^{(2)}). \tag{3.89}$$

*In addition, it is well known that any MA model can be approximated arbitrarly well by an $m$ dimensional AR process, with $m$ large enough. Thus taking $H^{(3)} = (H^{(1)})^{-1}$ we have*

$$cc(H^{(1)}, H^{(2)}) = cc(1, (H^{(1)})^{-1}H^{(2)}). \tag{3.90}$$

*But we know that $cc(1, (H^{(1)})^{-1}H^{(2)}) = cc(u_f, \tilde{y}_f)$ where $\tilde{y}_f$ is the future output of the system $\tilde{H} = (H^{(1)})^{-1}H^{(2)}$ driven by white Gaussian noise. So what we have is*

$$cc(y_f^{(1)}, y_f^{(2)}) = cc(u_f, \tilde{y}_f). \tag{3.91}$$

*We will use this evidence in chapter 6.*

## 3.5 Past Future MI for ARMA Process In Terms of the Poles and Zeros

**Theorem 3.9** *Given* $H(z) = z^{p-q}\frac{b(z)}{a(z)}$ *stable minimum phase ARMA model with poles* $\alpha_1, \ldots, \alpha_p$ *and zeros* $\beta_1, \ldots, \beta_q$ *that is driven by white noise with variance* $\sigma^2 = 1$. *We have*

$$I(y_p; y_f) = \frac{1}{2} log \frac{\prod_{i,j=1}^{p,q} \left|1 - \alpha_i \overline{\beta}_j\right|^2}{\prod_{i,j=1}^{p}(1 - \alpha_i \overline{\alpha}_j) \prod_{i,j=1}^{q}(1 - \beta_i \overline{\beta}_j)}. \tag{3.92}$$

**Proof:** See Appendix C.

This result can be formulated as distances between ARMA models. For this we need the cepstrum norm discussed in chapter 5.

# Chapter 4

# Rate-Distortion and Information Bottleneck Method

In this chapter we first review the rate-distortion theory, mainly taken from [10]. Next we describe the information bottleneck method (IB) presented in [3] and its application to gaussian variables [15]. Finally we give a straight forward application of the IB to linear dynamical systems and discuss its properties.

## 4.1 Rate-Distortion Theory

Rate-distortion theory, as its name suggest, concern the trade-off between rate and distortion. Given a random variable $X$, we introduce the random variable $\hat{X}$ as a lossy compressed version of $X$. Given a distortion measure between $X$ and $\hat{X}$, we would like to compress $X$ as much as possible while keeping the expected distortion below a given bound.

More formaly, we define a distortion measure

**Definition 4.1** *A distortion measure is a mapping*

$$d : \mathscr{X} \times \hat{\mathscr{X}} \to R^+. \tag{4.1}$$

*such that $d(x, \hat{x})$ measure the cost of representing $x$ by $\hat{x}$.*

On the other hand the goodness of compression, or rate, is measured by the mutual information $I(X; \hat{X})$.

In rate distortion theory we wish to minimize both the rate, $I(X; \hat{X})$, and the expected distortion $D(X, \hat{X}) \equiv E_{p(x,\hat{x})}\left[d(x, \hat{x})\right]$ with respect to the distirbution $p(\hat{x}|x)$. This is obviously a trade-off problem since, with a reasonable distortion function, taking the rate to be too small will inevitably result in a raise of the distortion.
This problem is formulized as

**Problem 1** *Given a random variable $X$ with a given distribution $p(x)$ and a distortion function $d(x, \hat{x})$, the rate distortion function $R(D)$ is given by*

$$R(D) = \min_{p(\hat{x}|x):\; E[d(x,\hat{x})] \leq D} I(X; \hat{X}), \tag{4.2}$$

*where the minimization is over $p(\hat{x}|x)$ with the constrain $E_{p(x,\hat{x})}\left[d(x, \hat{x})\right] \leq D$.*

$R(D)$ measures the maximum compression possible of $X$ at a given maximum allowed distortion $D$. Equivalently, $D(R)$ measures the minimum distortion possible with a given rate $R$. The function $R(D)$ draw a curve in the $(R, D)$ plane separating the achievable and unachievable ragions of $(R, D)$ pairs. We now give an example.

**Example 4.1 (Rate distortion function for a gaussian variable.)** *Taken from [10]. The rate distortion function for a $\mathscr{N}(0, \sigma^2)$ source with distortion function $d(x, \hat{x}) = (x - \hat{x})^2$ is*

$$R(D) = \begin{cases} \frac{1}{2} \log \frac{\sigma^2}{D}, & 0 \leq D \leq \sigma^2, \\ 0, & D > \sigma^2. \end{cases} \tag{4.3}$$

*The rate distortion curve for this example is plotted in figure 4.1. On the limits, when $D \to 0$ we know that the rate we need is the entropy of the variable which* *for a continuous variable is infinite. On the other hand when $D > \sigma^2$ we can take $\hat{x} = 0$ for all $x$ resulting in an* expected *distortion $\sigma^2$ which is less then $D$ and have $R = 0$.*

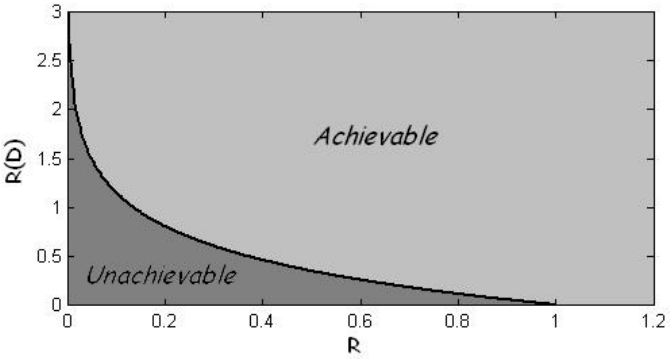

Figure 4.1: Rate distortion function for a Gaussian source with $\sigma^2 = 1$.

**Remark 4.1** *Shannon proved a rate distortion theorem that states that if we regard a series of i.i.d sampling of a source $X$ and try to compress it to a code with rate $R$, in the limit of large blocks, the above rate distortion function is the infimum of all achievable rates R with the distortion constrain $D$ (see [10]).*

Problem 1 can be formulated by introducing a Lagrange multiplier $\beta$ and minimizing the Lagrangian

$$\mathscr{L}\left[p(\hat{x}|x)\right] = I(X;\hat{X}) + \beta D(X,\hat{X}). \tag{4.4}$$

With the additional constrain that $p(\hat{x}|x)$ is a distribution. The solution to this problem is given by the self-consistent distribution (see [10])

$$p(\hat{x}|x) = \frac{p(\hat{x})}{Z(x,\beta)} e^{-\beta d(\hat{x},x)}. \tag{4.5}$$

In addition, $\beta > 0$, and formaly we have

$$\frac{\delta D}{\delta R} = -\frac{1}{\beta}. \tag{4.6}$$

A crucial step in formulating the rate distortion problem for a specific application is to find a distortion function $D(X, \hat{X})$ that will represent well our needs.

## 4.2 IB Method

The arbitrariness of the distortion function is a drawback of the theory if we want to deal with semantics. For a given source $X$, taking two different distortion function will result in two different rate distortion functions not necesseraly representing the relevant features in $X$.

**Remark 4.2** *In fact, this is only one example how the original Shannon's theory of information [11] disregard all connection to semantics giving its universality and strength. Until recently it was unclear how semantics can be incoporated into this theory (see Polani, 2001 [4] and Bar-Hillel, 1964 [6] for a discussion). the information bottleneck principle introduced in this section is one such possible approach.*

The information bottleneck principle suggested by Thisby, Pereira and Bialek [3] (and discussed in detail in [16]) introduce a new external variable $Y$ that represent the relevant features in $X$ that we want to preserve and uses it to define a distortion function.

More formaly, $\hat{X}$ being a compressed version of $X$ and $Y$ being a relevance variable about $X$, we have the markov chain

$$\hat{X} \leftrightarrow X \leftrightarrow Y. \tag{4.7}$$

Using the data processing inequality we have $I(\hat{X};Y) \leq I(X;Y)$. That is, we can't have more information in $\hat{X}$ about $Y$ then the original variable $X$ have. We now define the expected distortion function to be

$$D(\hat{X}, X) = I(X;Y) - I(\hat{X};Y), \tag{4.8}$$

which measure the lost information about $Y$ when considering $\hat{X}$ instead of $X$. By the data processing inequality it is non-negetive.

Since $I(X, Y)$ is constant we are left with the problem

**Problem 2** *Minimize the Lagrangian*

$$\mathscr{L}[p(\hat{x}|x)] = I(X;\hat{X}) - \beta I(\hat{X};Y) \tag{4.9}$$

*with respect to* $p(\hat{x}|x)$.

Notice that as $\beta \to \infty$, $\hat{X}$ will capture more and more information about $X$ but only if it is represented in $Y$.

The solution to this problem is also given by equation (4.5) with the "emergent" distortion function

$$d(\hat{x}, x) = D_{KL}\left[p(y|x)\|p(y|\hat{x})\right] \tag{4.10}$$

emphasizing the role $y$ playes as the relevance variable. It is show in [22] that this chioce of $d(x, \hat{x})$ is natural and in some sense unique.

## 4.3 IB for Gaussian Variables

Let $(X, Y)$ be two multivariate gaussian variables with dimensions $n_x$ and $n_y$ with covariances matrices $\Sigma_X$, $\Sigma_Y$ and $\Sigma_{XY}$. As an application of the IB method, we now want to compress $X$ as much as possible while preserving the information about $Y$. It can be shown that the $\hat{X}$ should be gaussian ([15]) and therefore can be reprsented by a noisy linear projection $\hat{X} = CX + \xi$, where $\xi \propto \mathscr{N}(0, \Sigma_\xi)$ is independent of $X$. Notice that $\Sigma_{\hat{X}} = C\Sigma_X C^T + \Sigma_\xi$ and that we do not restrict the dimension of $\hat{X}$. Thus the problem to solve is

**Problem 3** *Minimize the Lagrangian*

$$\mathscr{L}[C, \Sigma_\xi] = I(X; \hat{X}) - \beta I(\hat{X}; Y) \tag{4.11}$$

*with respect to $C$ and $\Sigma_\xi$ where $\hat{X} = CX + \xi$.*

**Theorem 4.1** *[15] The solution to problem 3 for a given value of $\beta$ is given by $\Sigma_\xi = I_{n_x}$ and*

$$C = \begin{cases} [\,\mathbf{0}^T \;\; ; \;\; \mathbf{0}^T \;\; ; \;\; \ldots \;\; ; \;\; \mathbf{0}^T] & 0 \leq \beta \leq \beta_1^c \\ [\alpha_1 \mathbf{v}_1^T \;\; ; \;\; \mathbf{0}^T \;\; ; \;\; \ldots \;\; ; \;\; \mathbf{0}^T] & \beta_1^c \leq \beta \leq \beta_2^c \\ [\alpha_1 \mathbf{v}_1^T \;\; ; \;\; \alpha_2 \mathbf{v}_2^T \;\; ; \;\; \ldots \;\; ; \;\; \mathbf{0}^T] & \beta_2^c \leq \beta \leq \beta_3^c \\ \qquad\qquad\vdots & \end{cases} \tag{4.12}$$

*where $\left\{\mathbf{v}_1^T, \mathbf{v}_2^T, \ldots, \mathbf{v}_{n_x}^T\right\}$ are the left eigenvectors of $\Sigma_{x|y}\Sigma_x^{-1}$, sorted by the corresponding asceding eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{n_x}$. $\beta_i^c = \frac{1}{1-\lambda_i}$ are critical $\beta$ values, $\alpha_i$ are the coefficients defined by $\alpha_i \equiv \left(\frac{\beta(1-\lambda_i)-1}{\lambda_i r_i}\right)_+$ with $r_i \equiv \mathbf{v}_i^T \Sigma_x \mathbf{v}_i$, $\mathbf{0}^T$ is a $n_x$ dimensional zero raw vector and semicolons seperates raws in the matrix $C$. $(\ldots)_+$ denotes the positive part.*

**Proof:** Due lack of space we omit the proof. see [15].

The solution asserts that as $\beta$ increases from zero, the dimension of the projected variable $\hat{X}$ increase in a series of critical $\beta_i^c$. For $0 \leq \beta \leq \beta_1^c$, the compression takes it all and $\hat{X}$ is just a white noise correspond to $C = 0$. As $\beta$ is increased above some $\beta_i^c$ we add one more eigenvector of $\Sigma_{x|y}\Sigma_x^{-1}$ to $C$ that $X$ is projected on. This is done such that the most informative eigenvectors, corresponding to lower $\lambda$, are added first. This can be seen by noticing that the eigenvalues $\lambda$'s are related to the canonical correlations between $X$ and $Y$ by $\lambda_i = 1 - \rho_i^2$ (see chapter 3). Note also that the transitions are smooth since the $\alpha_i$'s are zero at $\beta_i^c$.

**The information curve.** As in the rate distortion theory we can characterize the trade-off between compression and information preserving by ploting an information curve. This curve is defined as the maximum value of information preseving $I(\hat{X}; Y)$ as a function of the representation complexity of $X$ given by $I(\hat{X}; X)$.

For Gaussian IB this curve can be written in a closed form. By substituting the optimal projection matrix $A(\beta)$ into the definitions of $I(\hat{X}; X)$ and $I(\hat{X}; Y)$, eliminating $\beta$ and isolating $I(\hat{X}; Y)$ as a function of $I(\hat{X}; X)$ we get

$$I(\hat{X}; Y) = I(\hat{X}; X) - \frac{n_I}{2} \log \left( \prod_{i=1}^{n_I} (1 - \lambda_i)^{\frac{1}{n_I}} + e^{\frac{2I(\hat{X};X)}{n_I}} \prod_{i=1}^{n_I} \lambda_i^{\frac{1}{n_I}} \right) \tag{4.13}$$

where the products are over the first $n_I$ eigenvalues obeying the critical $\beta$ condition which can be written as $c_{n_I} \leq I(\hat{X}; X) \leq c_{n_I + 1}$, where $c_{n_I} = \sum_{i=1}^{n_I - 1} \frac{\lambda_{n_I}}{\lambda_i} \frac{1-\lambda_i}{1-\lambda_{n_I}}$. One such curve is given in Figure 4.2.

The derivative of this curve is given by $\beta^{-1}$ and it can be shown that this curve is concave. In IB, since we have the data processing inequality, we get that only for $\beta \geq 1$ we will have a solution $\hat{X}$ that is not null. In the GIB we even have a tighter bound and we get a nontrivial solution (i.e $C \neq 0$) only if $\beta \geq \frac{1}{1-\lambda_1} \geq 1$. This gives a bound on the derivative at the origin.

**Remark 4.3 (The role of $\xi$.)** *It can be shown (see along the proof of theorem 4.1) that the information curve (equation (4.13)) is invariant to the variance of*

*$\xi$. $\xi$ can be seen as a reference noise that all the solution is build around it; notice for example that as $\beta \to \infty$ the elements in the matrix $C$ become infinite to overcome this reference noise so to have the maximum information possible with $Y$. Notice however, that taking the limit $\xi \to 0$ is a different problem from regarding $\xi = 0$ from the start, where all the quantities become infinte and the problem lose any sense. In the next section we concern an IB method problem applied to linear dynamical systems and introduce a noise term $\xi$ as above. In chapter 6 we concern a twist of this problem that allow us to disregard this reference noise.*

## 4.4 IB for Linear Dynamical Systems

We are now in a position to introduce the main problem we want to address in this thesis. Given a complex process $Y(k)$, we want to find a parsimonious as possible approximated version of this process such that at the current time will have the maximum predictive information about the future. This is a trade off problem since by using more information from the past we will be able to predicte more about the future but inevitably result in a more complex process.

We now discuss the most straight forward formulation of this problem using the GIB.

We assume that $Y(k)$ can be well approximataed by a finite space-state model with large enough dimension $m$. In the GIB language we take the state-space process $\hat{y}$ at time $k = 0$ to be a compressed version of $u_p$ (Or more precisely $x(k = 0)$) with the relevant variable being the future $y_f$. To this end we consider a slightly modified version of the model given in chapter 2 following the discussion in remark 4.3. Consider the SISO LTI model

$$\begin{aligned} x(k+1) = &\quad Ax(k) + Bu(k), \\ y(k) = &\quad Cx(k) + \xi(k), \end{aligned} \tag{4.14}$$

where $x(k)$ is a state spcae vector of dimension $n$ and $\{A, B, C\}$ are with the appropriate dimensions. The input $u(k)$ is a gaussian white noise with variance $\sigma_u^2 = 1$ and the output noise $\xi(k)$ is a new gaussian white noise, independent of $u$, with variance $\sigma_\xi^2 = 1$. As shown in chapter 2, Given that $u(k)$ and $\xi(k)$ are stationary, $y(k)$ is also stationary.

We truncate the input $u(k)$ at $k = 0$ and let the system evolve. As in the GIB we define

$$\hat{y}(k) = \hat{C}x(k) + \hat{\xi}(k), \tag{4.15}$$

where $\hat{C}$ is the "projection" matrix and $\hat{\xi}(k)$ is an independent of $u$ and $\xi$ white gaussian noise with variance $\sigma_{\hat{\xi}}^2 = 1$. Now consider the IB problem (we notate $x(0) = x_0$ and $y(0) = y_0$)

**Problem 4** *Minimize the Lagrangian*

$$\mathscr{L}[\hat{C}] = I(x_0; \hat{y}_0) - \beta I(\hat{y}_0; y_f) \tag{4.16}$$

*with respect to $\hat{C}$ where $\hat{y}(k) = \hat{C}x(k) + \hat{\xi}(k)$ .*

$\hat{y}$ is a compressed version of the original state space. Notice that we do not restrict the dimension of $\hat{y}(k)$, so we have a new LTI system with the same structure as the original one, with output $\hat{y}(k)$, only now it is a single input *multiple* output (SIMO) system. Notice also that we have the markov chain $\hat{y}_0 \to x_0 \to y_f$.

We now recall (see chapter 2) the observability matrix $\mathscr{O}$, the controllability matrix $\mathscr{C}$ and the Gramians $W^o = \mathscr{O}^T\mathscr{O} \equiv \mathscr{G}$ and $W^c = \mathscr{C}\mathscr{C}^T \equiv \Sigma$. The current state space is given by

$$x_0 = \mathscr{C}u_p, \tag{4.17}$$

and the future output by

$$y_f = \mathscr{O}x_0 + \xi. \tag{4.18}$$

Following the solution for the GIB we need to find the eigenvalues and the left eigenvectors of the matrix

$$\Sigma_{x|y_f}\Sigma_x^{-1} = I - \Sigma_{xy_f}\Sigma_{y_f}^{-1}\Sigma_{y_f x}\Sigma_x^{-1}. \tag{4.19}$$

By the above we have

$$\begin{aligned} \Sigma_x &= \Sigma, \\ \Sigma_{xy_f} &= \Sigma\mathscr{O}^T, \\ \Sigma_{y_f} &= \mathscr{O}\Sigma\mathscr{O}^T + I, \\ \Sigma_{y_f x} &= \mathscr{O}\Sigma. \end{aligned} \tag{4.20}$$

Putting this in equation (4.19)

$$I - \Sigma\mathscr{O}^T\left(\mathscr{O}\Sigma\mathscr{O}^T + I\right)^{-1}\mathscr{O} \tag{4.21}$$

and using the matrix inversion formula (lemma 3.1) we get

$$\Sigma_{x|y_f}\Sigma_x^{-1} = (I + \Sigma\mathscr{G})^{-1}. \tag{4.22}$$

Taking the realization to be balanced, that is $\Sigma = \mathscr{G} = diag(\sigma_1, \dots, \sigma_m)$,where $\sigma_i$ are the squared singular values (see appendix A) of the hankel operator, we have immidiatly the eigenvalues and the eigenvectors; those are $\lambda_i = (1 + \sigma_i^2)^{-1}$ and the standart base $(0, \dots, 0, \underbrace{1}_{ith\ place}, 0, \dots, 0)$ respectively. Thus in the balanced realization the process $\hat{y}$ is given by the LTI system:

$$\begin{aligned} \hat{x}(k+1) &= \hat{A}\hat{x}(k) + \hat{B}u(k), \\ \hat{y}(k) &= \hat{C}x(k) + \hat{\xi}(k), \end{aligned} \tag{4.23}$$

with $\hat{A} = A$ and $\hat{B} = B$ and

$$\hat{C} = diag(\alpha_1, \dots, \alpha_m) \tag{4.24}$$

where

$$\alpha_i = \left(\frac{\beta(1-\lambda_i)-1}{\lambda_i r_i}\right)_+ = \left(\frac{\sigma_i^2(\beta-1)-1}{\sigma_i}\right)_+ . \tag{4.25}$$

The $\alpha's$ become positive in critical $\beta^c$'s given by

$$\beta_i^c = \frac{1}{1-\lambda_i} = \frac{1+\sigma_i^2}{\sigma_i^2}. \tag{4.26}$$

We plot the information curve given by equation 4.13 in figure 4.2. It is concave and although it is build by segments, it is continuous and smooth. Notice the critical points where new eigenvectors are added to $\hat{C}$. The suboptimal curves correspond to only the few first eigenvectors.

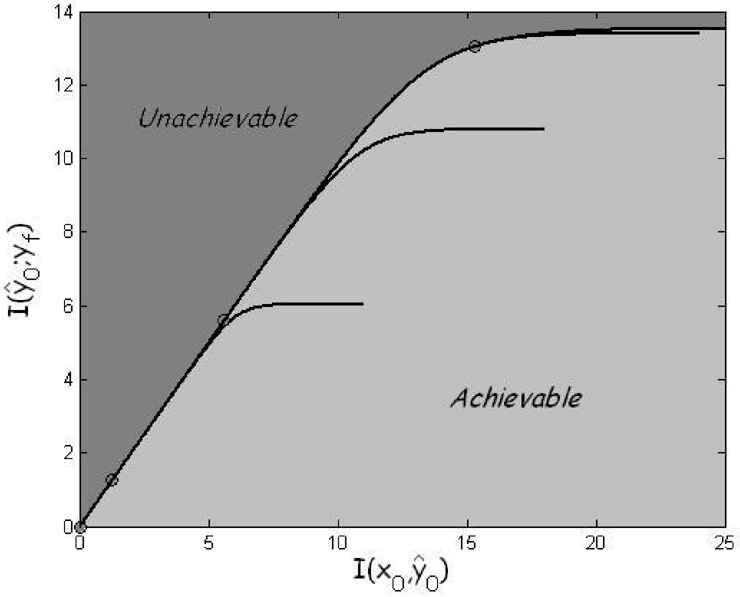

Figure 4.2: Information curves for the system $H(z) = \frac{z^4}{(z-0.8)^4}$. The circles denotes the "critical" points where additional eigenvectors were added. The suboptimal curves correspond to only the first eigenvectors.

**Remark 4.4** *We could also solve this problem without truncating the input at time $k = 0$. Taking the original system $H$ to be of the form of equations 2.13* *and 2.14 and the approximating system $\hat{H}$ as in here, all the changes we have to do is to replace $\Sigma_{y_f} = \mathscr{O}\Sigma\mathscr{O}^T + I$ with $\Sigma_{y_f} = \mathscr{O}\Sigma\mathscr{O}^T + \Delta\Delta^T$. As was shown in chapter 3 all we need to do now is to find the eigenvalues and left eigenvectors of*

$$\Sigma_{x|y_f}\Sigma_x^{-1} = (I + \Sigma\mathscr{G}_z)^{-1}. \tag{4.27}$$

*It can be shown that there is a realization such that $\Sigma$ and $\mathscr{G}_z$ are diagonal and equal thus we can solve it in the same route as above.*

**Discussion.** The solution asserts that the components in the output $\hat{y}$ are uncorrelated with each other and they are sorted (and scaled) such that the more informative parts are first uncoverd. Since the dimension of the output decrease with decreasing $\beta$ it is natural to ask if this imply also a reduction of the order of the approximated system state-space. As theorem 2.3 asserts, this will happen only if the new system is unobservable (It is surely controlloable since we haven't changed $A$ and $B$). Since $\hat{C}$ is diagonal, this in turn imply that in the *balanced realization* $A$ should have a block lower triangolar form

$$A_{balanced} = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}. \tag{4.28}$$

On the other hand, since the original system is observable and $A$ being in the above form we must have $C_{balanced} = [C_1 \ C_2]$ where $C_2 \neq 0$. Taking this into account and since the observability Gramian is diagonal we must have $A_{21} \neq 0$. All those are only necessery conditions. Those conditions in general will not be met but it can be shown that the set of SISO systems having $A_{balanced}$ in the above form with $A_{21} \neq 0$ and with diagonal and equal Gramians is not empty.

Alternatively, instead of compressing the original state space, we could try to work direcetly on $u_p$. This is done in [9]. When truncating the input $u(k)$ at $k = 0$ we have

$$y_f = Hu_p + \xi, \tag{4.29}$$

where $H$ is the Hankel operator (see definition 2.8 and the discussion below it).

We define

$$\hat{y}_f = H(\beta)u_p + \hat{\xi}, \tag{4.30}$$

And solve the IB problem

**Problem 5** *Minimize the Lagrangian*

$$\mathscr{L}[H(\beta)] = I(u_p; \hat{y}_f) - \beta I(\hat{y}_f; y_f) \tag{4.31}$$

*with respect to $H(\beta)$ where $\hat{y}_f = H(\beta)u_p + \hat{\xi}$ and $y_f = Hu_p + \xi$.*

**Remark 4.5** *Notice that if we do not truncate the input at $k = 0$ we will not have the Markov chain $\hat{y}_f \to u_p \to y_f$ since given $u_p$, $\hat{y}_f$ and $y_f$ are not independent anymore.*

We write the singular value decomposition (see appendix A) of the Hankel operator $H$ as

$$H = U^T \Sigma V, \tag{4.32}$$

with $\Sigma = diag(\sigma_1, \ldots, \sigma_n)$ where $\sigma_i$ is the $ith$ singular value. It is shown in [9] that the solution to problem 5 is given by

$$H(\beta) = U^T \Sigma(\beta) V, \tag{4.33}$$

with $\Sigma(\beta) = diag(\alpha_1(\beta), \ldots, \alpha_n(\beta))$ and $\alpha_i^2(\beta) = (\sigma_i^2(\beta - 1) - 1)_+$.

Unfortunately, $H(\beta)$ is not an Hankel operator any more. Ignoring this and taking a realization from $H(\beta)$ by using the Kalman and Ho algorithm (see [31]) we will generically get a system that is very close to the GIB information curve. However we can find situations where this realization is quite far from this curve. Notice that taking the Kalman and Ho realization with $H(\beta)$ we take only some of the singular directions in $V$ (or $U$) but it may be that those vectors consist a combination from all poles. Indeed the conditions for $H(\beta)$ to be an Hankel operator are essentially the same as the conditions for the system in the solution to problem 4 to be unobservable.

**Conclusions** Those two approaches are essentialy the same. We conclude that those approaches, although that gave some great insights, are not fully satisfactory since we want to have more control on the approximation and on the order of the system. In particular when regarding the second problem we would like to know that our approximated system can reach the information curve. In chapter 6 we give an alternative solution to a closely related problem which overcome those shorts, but in order to introduce this problem elegantly we need first the cepstrum norm and its relation to ARMA models.

# Chapter 5

# The Cepstrum Norm for ARMA models

In this chapter we introduce the cepstrum of a stochastic process and calculate it for an ARMA model in terms of the process poles and zeros. Next we follow Martin [20] and DeCock [19] and define a norm for, and a distance between, ARMA processes using the cepstrum. Finally, as done in [19], we relate this distance to the mutual information between the past and the future for ARMA processes.

The cepstrum was introduced by Bogert, Healy and Tukey in their paper [29] where it used to detect echos in seismological data. In parallel, the theory of homomorphic systems was developed by Oppenheim [26] which generelize the class of systems obeying the superposition principle (i.e linear systems) with the adventage that those are easy to analyze and are well studied. An introduction to homomorphic systems and its application to the feilds of image and audio processing can be found in chapter 10 of the book [27]. Here we only give a brief review.

It turned out that the cepstral processing is a special case of homomorphic processing for systems obeying convolution. A review on the history of the cepstrum, written by Oppenheim and Schafer, is given in [28].

## 5.1 The Power Cepstrum

**Definition 5.1** *The power cepstrum $c_k$ of a stationary scalar stochastic process $y(k)$ is the inverse Fourier trasform of the logarithm of the power spectrum of the process:*

$$c_k = \frac{1}{2\pi} \int_0^{2\pi} log(S(e^{i\theta}))e^{ik\theta} d\theta \tag{5.1}$$

*with the power spectrum given by*

$$S(e^{i\theta}) = \sum_{k=-\infty}^{\infty} R_{yy}(k) e^{-ik\theta}, \tag{5.2}$$

*where $R_{yy}(k)$ is the autocorrelation function and we assume $S(e^{i\theta}) \neq 0$ for all $\theta$.*

The power cepstrum is even and real.

## 5.2 Homomorphic Signal Processing

We now give a brief review of homomorphic signal processing and its application to convolution systems like the models we work with (see chapter 2). We do this only to give some insight to the role of the cepstrum. This review is taken mainly from [27].

The princple of superposition require that the system transformation $T$ acted upon the two inputs $u_1(n)$ and $u_2(n)$ obey

$$T\left[u_1(n) + u_2(n)\right] = T\left[u_1(n)\right] + T\left[u_2(n)\right] \tag{5.3}$$

and for any scalar $c$

$$T\left[cu_1(n)\right] = cT\left[u_1(n)\right]. \tag{5.4}$$

We generalize this principle by introducing an operation $\Box$ to combine between inputs and by $:$ to combine input with scalar. Similarly $\diamond$ To combine between outputs and $\perp$ to combine output with a scalar. We rquire

$$H\left[u_1(n)\ \square\ u_2(n)\right] = H\left[u_1(n)\right] \diamond H\left[u_2(n)\right] \tag{5.5}$$

and

$$H\left[c : u_1(n)\right] = c \perp H\left[u_1(n)\right]. \tag{5.6}$$

The linear systems are a special case with $\square$ and $\diamond$ taken as addition and : and $\perp$ as multipliction.

If we wish to use this principle with the use of linear vector spaces (one for the input and one for the output and $H$ as a transformation between them) we also have to demand that the inputs and outputs must satisfy the algebric poatulates of vector addition and scalar multiplication. Indeed it can be shown (see [27]) that if the system inputs constitute a vector space with $\square$ and : as vector addition and scalar multiplication respectively and system outputs constitute a vector space with $\diamond$ and $\perp$ as vector addition and scalar multiplication respectively, then those systems can be represented by the cascade scheme given in figure 5.1.

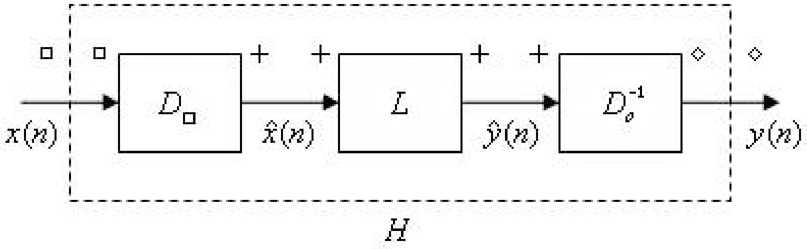

Figure 5.1: Homomorphic system.

$D_\square$ and $D_\diamond$ are systems obeying the generalized superposition principle with the input operator being $\square$ or $\diamond$ and the output operation is $+$ and most importantly L is a conventional linear system.

If we now take the input operation $\square$ and the output opretaion $\diamond$ to be the convolution operator $*$ of two signals the characaristic function $D_*$ can be found as follows.

As we saw in chapter 2 the Z-transform of two convolved signal is the multipliction of their Z-transforms

$$U(z) = U_1(z) \cdot U_2(z). \tag{5.7}$$

Taking the complex logarithm

$$\hat{U}(z) = \log U_1(z) + \log U_2(z). \tag{5.8}$$

and taking the inverse Z-transform we get the series $\hat{u}(n) = \hat{u}_1(n) + \hat{u}_2(n)$ which is called the complex cepstrum (although it is real for real $u(n)$). It can be shown that for minimum phase ARMA systems, $\hat{u}(n)$ is causal and it is equal to the $k \geq 0$ part of the power cepstrum.

**Remark 5.1** *We have omitted the delicate subject of the definition of the complex logarithm. For details see [27].*

This give us some insight as to why the cepstrum (and more generally with all homomorphic systems) work well in some applications. When in the cepstrum a well separated behavior of the two (or more) signals is apparent, by linear filtering the cepstrum (liftering) we can remove the contribution of one of the terms. What we actually did is a deconvolution of the original signals by linear filtering; this is much harder to do by working directly on the original signal.

A very beautiful application of the cepstrum as a deconvolver is in the theory of speech analysis. Human speech is supposed to be composed of a series of phonemes which can crudely be separated to voiced (using the vocal strings) and unvoiced ones. Those phonemes can be modeled by the convolution of the response of the vocal tract and a pulse train in the case of voiced sounds and a random noise in the unvoiced sounds. The response of the vocal tract can further be modeled as a series of LTI systems changing slowly in time. By deconvolution we can find the pitch and the vocal tract response which can be further used in speech simulations and recognition applications. For (much) more details see [30].

## 5.3 The Cepstrum of a Minimum Phase ARMA Model

**Theorem 5.1 (The Cepstrum of a Minimum Phase ARMA model.)** *Let $y(k)$ be the output of a stable minimum phase ARMA(p,q) model with poles $\alpha_1, \ldots, \alpha_p$ and zeros $\beta_1, \ldots, \beta_q$ that is driven by white noise with variance $\sigma^2$. then it's cepstrum is given by:*

$$c_k = \begin{cases} \log \sigma^2 & k = 0, \\ \sum_{i=1}^{p} \frac{\alpha_i^{|k|}}{|k|} - \sum_{i=1}^{q} \frac{\beta_i^{|k|}}{|k|} & k \neq 0. \end{cases} \tag{5.9}$$

**Proof:** The power spectrum of an ARMA model is given by (see equations (2.30 and (2.22)))

$$S(e^{i\theta}) = \sigma^2 \frac{\prod_{i=1}^{q} |e^{i\theta} - \beta_i|^2}{\prod_{i=1}^{p} |e^{i\theta} - \alpha_i|^2} = \sigma^2 \frac{\prod_{i=1}^{q} |1 - \beta_i e^{-i\theta}|^2}{\prod_{i=1}^{p} |1 - \alpha_i e^{-i\theta}|^2}. \tag{5.10}$$

Taking the logarithm

$$\log S(e^{i\theta}) = \log \sigma^2 + \sum_{i=1}^{q} \left[\log(1 - \beta_i e^{-i\theta}) + \log(1 - \bar{\beta}_i e^{i\theta})\right] \\ - \sum_{i=1}^{p} \left[\log(1 - \alpha_i e^{-i\theta}) + \log(1 - \bar{\alpha}_i e^{i\theta})\right]. \tag{5.11}$$

Since the system is assumed to be minimum phase all the poles and the zeros are inside the unit circle and we can use the power expansion $\log(1-x) = -\sum_{k=1}^{\infty} \frac{x^k}{k}$

for $|x| < 1$ to get

$$\log S(e^{i\theta}) = \log \sigma^2 + \sum_{i=1}^{p} \left( \sum_{k=1}^{\infty} \frac{\alpha_i^k}{k} e^{-ik\theta} + \frac{\bar{\alpha}_i^k}{k} e^{ik\theta} \right) - \sum_{i=1}^{q} \left( \sum_{k=1}^{\infty} \frac{\beta_i^k}{k} e^{-ik\theta} + \frac{\bar{\beta}_i^k}{k} e^{ik\theta} \right). \quad (5.12)$$

Equating coefficients to the definition of the cepstrum as a fourier transform

$$\log S(e^{i\theta}) = \sum_{k=-\infty}^{\infty} c_k e^{-ik\theta} \quad (5.13)$$

and using the fact that the poles and zeros comes in conjugate pairs we get the theorem, equation (5.9). □

Notice that every contribution from a pole or a zero is additive in the cepstrum.

## 5.4 The Cepstrum Norm as a Distance Between ARMA Models

We can use the cepstrum to define a distance between stochastic processes. Given two stochastic processes $y^{(1)}(k)$ and $y^{(2)}(k)$ with cepstrums $c_k^{(1)}$ and $c_k^{(2)}$, one such family is given by

$$\Delta^2(y^{(1)}, y^{(2)}) = \sum_{k=0}^{\infty} w_k \left( c_k^{(2)} - c_k^{(1)} \right)^2 \quad (5.14)$$

where $w_k$ are some positive weights. Note that since the cepstrum is an even function we only need to sum for $k \geq 0$. Following martin [20] we now define the distance between two minimum phase ARMA processes.

**Definition 5.2** *Given two minimum phase ARMA processes $y^{(1)}(k)$ and $y^{(2)}(k)$ with transfer functions $H^{(1)}(z)$ and $H^{(2)}(z)$ we define their distance by taking $w_k = k$ in the above family*

$$\Delta^2(\log H^{(1)}, \log H^{(2)}) = \sum_{k=0}^{\infty} k \left( c_k^{(2)} - c_k^{(1)} \right)^2 , \tag{5.15}$$

**Remark 5.2** *It seems that this is not a true metric since we dont have $\Delta^2(y^{(1)}, y^{(2)}) = 0 \Leftrightarrow y^{(1)} = y^{(2)}$. This is because we might have $c_0^{(1)} \neq c_0^{(2)}$. But by taking the input signal to have the same variance we solve the problem.*

We can also define a norm (DeCock [19])

**Definition 5.3** *Given a transfer function $H(z)$ with cepstrum $c_k$ the cepstrum norm is defined as:*

$$\|logH(z)\|^2 = \sum_{k=1}^{\infty} k c_k^2. \tag{5.16}$$

Now consider the process with the transfer function $\frac{H^{(1)}(z)}{H^{(2)}(z)}$. Since the cepstrum is additive we have that its cepstrum is given by $c_k^{(1)} - c_k^{(2)}$ and thus its norm is given by

$$\|logH^{(1)} - \log H^{(2)}\|^2 = \Delta^2(\log H^{(1)}, \log H^{(2)}). \tag{5.17}$$

That is, the cepstrum norm of $\frac{H^{(1)}(z)}{H^{(2)}(z)}$ is the cepstral distance between $H^{(1)}$ and $H^{(2)}$.

**Remark 5.3** *Taking $y^{(2)}$ to be a white noise we have $H^{(2)} = 1$ and thus we find that the cepstral norm of an ARMA process is its distance from being white noise.*

**Remark 5.4** *The norm of an $ARMA(p, q)$ process with transfer function $H(z) = z^{p-q}\frac{b(z)}{a(z)}$ (see (2.22)), can be seen as a distance between two AR processes with transfer functions $H^{(1)}(z) = \frac{z^p}{a(z)}$ and $H^{(2)}(z) = \frac{z^q}{b(z)}$.*

Notice also that we have

$$\|logH(z)\|^2 = \|logH^{-1}(z)\|^2. \tag{5.18}$$

The cepstrum norm of an ARMA process can be expressed in terms of the poles and zeros of the system (e.g [20])

**Theorem 5.2 (The Cepstrum Norm for ARMA model in terms of poles and zeros.)** *Given $H(z) = z^{p-q}\frac{b(z)}{a(z)}$ a stable minimum phase ARMA(p,q) model with poles $\alpha_1, \ldots, \alpha_p$ and zeros $\beta_1, \ldots, \beta_q$ that is driven by white noise with variance $\sigma_u^2 = 1$. Then*

$$\|logH(z)\|^2 = log\frac{\prod_{i=1}^{p}\prod_{j=1}^{q}\left|1-\alpha_i\overline{\beta}_j\right|^2}{\prod_{i,j=1}^{p}(1-\alpha_i\overline{\alpha}_j)\prod_{i,j=1}^{q}(1-\beta_i\overline{\beta}_j)}. \tag{5.19}$$

**Proof:** We have

$$\begin{aligned}
\|logH(z)\|^2 &= \sum_{k=0}^{\infty} kc_k^2 \\
&= \sum_{k=0}^{\infty} k\left(\sum_{i=1}^{p}\frac{\alpha_i^k}{k} - \sum_{i=1}^{q}\frac{\beta_i^k}{k}\right)^2 \\
&= \sum_{k=0}^{\infty}\left(\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\alpha_i^k\alpha_j^k}{k} + \sum_{i=1}^{q}\sum_{j=1}^{q}\frac{\beta_i^k\beta_j^k}{k} - 2\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\alpha_i^k\beta_j^k}{k}\right).
\end{aligned}$$

Inserting the sum on $k$ into the bruckets and using the expansion $-log(1-x) = \sum_{k=1}^{\infty}\frac{x^k}{k}$ for $|x| \leq 1$ we get:

$$\begin{aligned}
\|logH(z)\|^2 = -\sum_{i=1,j=1}^{p,p}\log(1-\alpha_i\alpha_j) - \sum_{i=1,j=1}^{q,q}\log(1-\beta_i\beta_j)+ \\
+2\sum_{i=1,j=1}^{p,q}log(1-\alpha_i\beta_j).
\end{aligned} \tag{5.20}$$

To emphesise that this quantity is real we reorder the sums in conjugate pairs, we also put all the terms together under one log function

$$\|logH(z)\|^2 = log\frac{\prod_{i=1}^{p}\prod_{j=1}^{q}\left|1-\alpha_i\overline{\beta}_j\right|^2}{\prod_{i,j=1}^{p}(1-\alpha_i\overline{\alpha}_j)\prod_{i,j=1}^{q}(1-\beta_i\overline{\beta}_j)}, \tag{5.21}$$

proving the theorem. □

**Theorem 5.3** *(DeCock [19]) Inspecting theorems 3.9 and 5.2 we find that (for the conditions given in the theorem 5.2 above) we have*

$$\|logH\|^2 = 2I(u_p; y_f) \tag{5.22}$$

An alternative proof of theorem 5.3 is as follows. In [20] it is shown, using what is called the resultant, that the cepstrum norm can be written as

$$\|\log H\|^2 = \sum_{k=1}^{n} k\ln(1-|r_k|^2) \tag{5.23}$$

where $n$ is the order of the system and $r_k$ are the partial autocorreletions (or reflection coefficients) defined by [21]

$$r_k = E[y_0 y_k | y_1, y_2, \ldots, y_{k-1}]. \tag{5.24}$$

Using the chain rule for mutual information we have

$$\begin{aligned}
I(y_p; y_f) &= I(\ldots, y_{-2}, y_{-1}; y_0, y_1, y_2, \ldots) && (5.25)\\
&= \sum_{k=1}^{\infty} I(y_{-k}; y_0, y_1, y_2, \ldots | y_{-(k-1)} \ldots, y_{-2}, y_{-1}) && (5.26)\\
&= \sum_{k=1,i=1}^{\infty,\infty} I(y_{-k}; y_i | y_{-(k-1)} \ldots, y_{-2}, y_{-1}, y_0, y_1, \ldots, y_{i-1}) = && (5.27)\\
&= \sum_{k=1}^{\infty} kI(y_0; y_{0+k} | y_1, y_2, \ldots, y_{k-1}) && (5.28)
\end{aligned}$$

where in the last step we used stationarity. Now recall that since our variables are Gaussian and that given the state space the future is independent of the past we have

$$I(y_t; y_{t+k}|y_{t+1}, y_{t+2}, \ldots, y_{t+k-1}) = \begin{cases} \frac{1}{2}log(1-|r_k|^2) & 1 \leq k \leq n \\ 0 & n < k \end{cases} . \quad (5.29)$$

Thus proving the theorem.

We now have all the ingredients needed to formulize the new problem.

# Chapter 6

# Formulation of the (new) Problem

Let us consider again the IB problem 5 (see equation (4.31)). Given a process $y$ induced by the $m$ dimensional system $H$ driven by a stochastic input, find an approximating process $\hat{y}$ given by the $n$ dimensional system $\hat{H}$ driven by the same input and in the same class of $H$, such that while preserving the information $I(\hat{y}_f; y_f)$ we minimize the information $I(u_p; \hat{y}_f)$. Let us take $y$ to be of the form of equations (2.13) and (2.14) (instead of (4.14) in problem 5) given here again for reference

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) && (6.1) \\ y(k) &= Cx(k) + u(k). && (6.2) \end{aligned}$$

with $u(k)$ being a white gaussian noise. Notice that the system is deterministic. The problem is

**Problem 6** *Minimize the Lagrangian*

$$\mathscr{L}[\hat{H}] = I(u_p; \hat{y}_f) - \beta I(\hat{y}_f; y_f). \tag{6.3}$$

*with respect to $\hat{H}$.*

As was shown in chapter 3 this can be written using the canonical correlations:

$$I(u_p;\hat{y}_f) - \beta I(\hat{y}_f;y_f) = -\frac{1}{2}log\prod_{i=1}^{n}(1-\rho_i^2) + \frac{\beta}{2}log\prod_{i=1}^{\infty}(1-\tau_i^2), \qquad (6.4)$$

where $\rho_i^2$ are the $cc^2(u_p,\hat{y}_f)$ and $\tau_i^2$ are the $cc^2(\hat{y}_f,y_f)$.

Unfortunately the second term is $\infty$. This is because the $cc(\hat{y}_f,y_f)$ (there are infinitely many) are all 1 except for maximum $n+m$ canonical correlations that are smaller than one (these are the intersting ones because they quantify the difference of information taken from the past by the two processes).
Putting only the smallest $n+m$ canonical correlation is in no help. This is because that by a wise choice of the state-space we can always make one of the canonical correlations $\tau_i^2$ to be 1 causing the second expression to become $\infty$ again, while the first expression is always finite, therefore minimizing $\mathscr{L}$ and killing the trade-off.
Also replacing the second expression with $I(\hat{y}_p;y_f)$ which is always finite will not rescue us since without output noise we have $I(\hat{y}_p;y_f) = I(u_p;y_f)$ (recall that all $cc(u_p,\hat{y}_p)$ are 1, see theorem 3.8) thus making the problem trivial .

As a saver let us consider the transformation

$$\frac{\beta}{2}log\prod_{i=1}^{\infty}(1-\tau_i^2) \longrightarrow -\frac{\beta}{2}log\prod_{i=1}^{\infty}\tau_i^2 \qquad (6.5)$$

This new term is always finite. By minimizing $-\frac{\beta}{2}log\prod_{i=1}^{n}\tau_i^2$ we want to make the canonical correlations $\tau_i^2$ closer to 1 thus increasing the information between $y_f$ and $\hat{y}_f$ the same as before. Notice however that the set of systems obeying $\prod_{i=1}^{n}(1-\tau_i^2) = constant_1$ is not the same set as the systems obeying $\prod_{i=1}^{n}\tau_i^2 = constant_2$; Obviously those are not the same problem.

As discussed in remark 3.5 we have

$$cc^2(\hat{y}_f, y_f) = cc^2(u_f, \tilde{y}_f), \tag{6.6}$$

that is, $\tau_i^2$ can be seen as $cc(u_f, \tilde{y}_f)$ where $u_f$ is the future input and $\tilde{y}$ is the process given by the transfer function $H\hat{H}^{-1}$.
By using theorem 3.8 $(cc^2(u_f, y_f) = 1 - cc^2(u_p, y_f))$ we have:

$$-\frac{\gamma}{2} log \prod_{i=1}^{n} \tau_i^2 = -\frac{\gamma}{2} log \prod_{i=1}^{n} (1 - \mu_i^2) \tag{6.7}$$

where $\mu_i^2$ are the $cc(u_p, \tilde{y}_f)$. Thus we replaced problem 6 with the new problem

**Problem 7** *Minimize the Lagrangian*

$$\mathscr{L}[\hat{H}_n] = I(u_p; \hat{y}_f) + \beta I(u_p; \tilde{y}_f), \tag{6.8}$$

*with respect to $\hat{H}_n$, where $\hat{H}_n$ is of the form of equations (6.1) and (6.2) of order $n$.*

Refering to figure 6.1 the process $\tilde{y}$ induced by the system $\tilde{H} = H\hat{H}^{-1}$ when driven by stochastic input can be viewed as a residual process having only the information $\hat{H}$ missed about $H$. Indeed, when $\hat{H} = 1$ we have that $\tilde{H} = H$ and all the information is lost. On the other hand when $\hat{H} = H$ we have $\tilde{H} = 1$ and all the information is represented.

In the spirit of the introduction this problem can be stated also as

$$R_n(D) = \min_{\hat{H}_n : D(H, \hat{H}_n) \leq D} I(u_{past}, \hat{y}_{future}) \tag{6.9}$$

with the distortion function $D^{(3)}(H, \hat{H}_n) = I(u_{past}, \tilde{y}_{future})$.

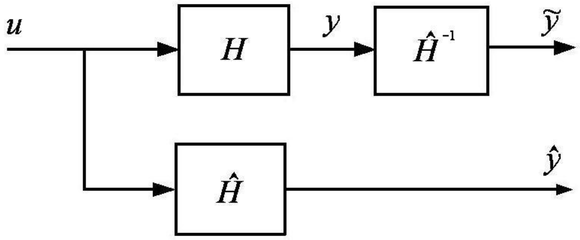

Figure 6.1: Definition of the residual system $\tilde{H}$

$R_n(D)$ measures the minimum possible representation (in the information theoretic sense) of $u_p$ in a state-space of order $n$ at a particular time $k$ such that we wont loss more then $D$ nuts of information from the past about the future. Equivalently $D_n(R)$ measures the minimum possible loss of information between the past and the future with a given amount of representation of $u_p$ in a state-space of order $n$.

As we saw, the first expression can be writen as a cepstral norm $\left\|log\hat{H}_n\right\|^2$ where $\hat{H}_n$ is the transfer function of $\hat{y}$. Now also the new term can be written as a cepstral norm and we can formulate the problem as:

$$\min_{\hat{H}_n} \left\|log\hat{H}_n\right\|^2 + \beta\left\|log\hat{H}_n - logH\right\|^2. \tag{6.10}$$

Making use of the cepstrum norm in terms of the poles and zeroes (theorem 5.2) we can take our optimaization parameters to be the poles $\hat{\alpha}_1, \ldots, \hat{\alpha}_n$ and zeroes $\hat{\beta}_1, \ldots, \hat{\beta}_n$ of the approximating system $\hat{H}_n$.

To conclude, since our formulation is constructive we now have a problem that the order of the solution is in our hands and in addition we can be sure that the approximated system will reach the information curve. In the next chapter we inspect the solution in detail.

Finally we observe that since our system is determenistic, an important question of robustness to noise arise. Let us for example assume that we wanted to solve the seemingly natural IB problem of working directly on the state-space with the deteriminstic system (2.13) and (2.14) (see also (6.1) and (6.2)). As you can recall, this problem was solved in chapter 4 for stochastic system with noise in the output and truncating the input in $k = 0$. That is, solve

$$\min_{\hat{H}} I(u_p; \hat{x}) - \beta I(\hat{x}; y_f). \tag{6.11}$$

But since our system is deterministic we have that $I(u_p; \hat{x}) = \infty$ to any non-trivial $\hat{x}$ and thus we are forced to change this term to $I(\hat{x}; \hat{y}_f)$ which is finite and also represent in some sense the information $\hat{x}$ has about the past. Thus we solve

$$\min_{\hat{H}} I(\hat{x}; \hat{y}_f) - \beta I(\hat{x}; y_f). \tag{6.12}$$

We see that $I(\hat{x}; \hat{y}_f) \ngeq I(\hat{x}; y_f)$ and thus we pay the price by sacrificing the information processing inequality. This results in the observation that we can take $\hat{x}$ to have infinitesimal representation of the past, and thus have infinitesimal $I(\hat{x}; \hat{y}_f)$, but have finite information with $y_f$. This can be seen by recalling that in the geometric view (section 3.3.2) the canonical correlations between $\hat{x}$ and $y_f$ are determined by the angles between the subspaces rather then by the magnitudes of the vectors. Thus even that $\hat{x}$ has infinitesimal small magnitude we can take it to have a non trivial angle with the subspace spaned by $y_f$ resulting in a finite $I(\hat{x}; y_f)$. This results in a discontinuity in the solution for $\beta \to 0$ since for $\beta = 0$ we obviously have that the solution is $\hat{H} = 1$ with $I(\hat{x}; y_f) = 0$ while for $\beta \to 0^+$ we have $I(\hat{x}; y_f) \to c$ with $c \neq 0$. This behavior is non-robustic since adding infinitesimal noise to the stae-space will alter the solution totally; with the presence of such a state-space noise, having infinitesimal information about the past will result also in an infinitesimal information about $y_f$.

In our problem this non-robustness (or discontinuity) will not happen since we work with the canonical correlations between the output's future. To see this we observe that in the null case we already have a non-trivial angles between those subspaces (i.e in the null case we have $cc(\hat{y}_f, y_f) \to cc(u_f, y_f)$), thus changing the angles a bit by introducing an infinitesimal information about the past cannot change those cc much thus not altering the solution drastically. In other words we have that $I(u_p; \hat{y}_f) \rightarrow 0$ inevitebly result in $I(u_p; \tilde{y}_f) \rightarrow I(u_p; y_f)$ which is also the solution for $\beta = 0$. Indeed, in the next chapter we find that the solution is continuous in the limit $\beta \rightarrow 0$.

A noise can also be present in the outputs of the systems but we can use the same arguments and conclude that adding infinitesimal noise in the output will result in a continuous change in the $cc(\hat{y}_f; y_f)$ and thus also in the information $I(u_p; \tilde{y}_f)$ making the solution robust also to this kind of noise.

Thus we conclude that although that our model system is determenistic, the solution to the problem is robust against making our system stochastic by adding infintesimal noise. This result can be verified by numerical calculations. This robustness is of course a good thing making our analysis applicable.

# Chapter 7

# Results

As was shown in the previous chapter, given the system $H(z) = \frac{b(z)}{a(z)}$ with poles $\alpha_1, \ldots, \alpha_m$ and zeros $\beta_1, \ldots, \beta_m$ we need to find

$$\min_{\hat{H}_n} \mathscr{L} : \mathscr{L}[\hat{H}_n] = \left\| log\hat{H}_n \right\|^2 + \beta \left\| log\hat{H}_n - logH \right\|^2, \tag{7.1}$$

or minimize

$$\mathscr{L}[\hat{\alpha}_1, \ldots, \hat{\alpha}_n, \hat{\beta}_1, \ldots, \hat{\beta}_n] = \left\| log\frac{\hat{b}(z)}{\hat{a}(z)} \right\|^2 + \beta \left\| log\frac{b(z)\hat{a}(z)}{a(z)\hat{b}(z)} \right\|^2, \tag{7.2}$$

w.r.t. $\hat{\alpha}_1, \ldots, \hat{\alpha}_n$ and $\hat{\beta}_1, \ldots, \hat{\beta}_n$, the poles and zeros of $\hat{H}_n(z) = \frac{\hat{b}(z)}{\hat{a}(z)}$ respectively. Indeed, using theorem 5.2 we have a closed form of $\mathscr{L}[\hat{\alpha}_1, \ldots, \hat{\alpha}_n, \hat{\beta}_1, \ldots, \hat{\beta}_n]$ in terms of the poles and zeros of both systems and we can work directly on them.

To inspect this parametric optimization problem we give a series of examples which the generic one is a mixture of those behaviours. Since AR models capture all the complexity of the problem we start with examples from this class and later extend to ARMA models.

## 7.1 Finding the Optimal System

Here we consider the AR model with transfer function $H(z) = \frac{z^m}{(z-\alpha_1)...(z-\alpha_m)}$ of order $m$ and its approximate version $\hat{H}(z) = \frac{z^n}{(z-\hat{\alpha}_1)...(z-\hat{\alpha}_n)}$ of order $n$.

The Lagrangian becomes:

$$\mathscr{L}[\hat{\alpha}_1, \ldots, \hat{\alpha}_n] = -\ln \prod_{i,j=1}^{n,n} (1-\hat{\alpha}_i\hat{\alpha}_j^*) - \beta \ln \frac{\prod_{i,j=1}^{n,n}(1-\hat{\alpha}_i\hat{\alpha}_j^*)\prod_{i,j=1}^{m,m}(1-\alpha_i\alpha_j^*)}{\prod_{i,j=1}^{n,m}|1-\hat{\alpha}_i\alpha_j^*|^2} \tag{7.3}$$

which we want to minimize w.r.t the $\hat{\alpha}_i' s$ (and treating the $\alpha' s$ as constants). Define

$$\kappa = \frac{\beta}{\beta+1}.$$

Taking derivatives $\frac{\partial L}{\partial \hat{\alpha}_i}$ and equating to zero (recall that the poles comes in conjugate pairs) we get:

$$\sum_{j=1}^{n} \frac{\hat{\alpha}_j}{1-\hat{\alpha}_i\hat{\alpha}_j^*} = \kappa \sum_{j=1}^{m} \frac{\alpha_j}{1-\hat{\alpha}_i\alpha_j^*}. \tag{7.4}$$

Using $\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$ for $|x| < 1$ and defining $\alpha^{(s)} = \sum_{i=1}^{m} \alpha_i^s$ and $\hat{\alpha}^{(s)} = \sum_{i=1}^{n} \hat{\alpha}_i^s$ (notice that this definition is just the cepstrum coeffinients $c_s$ and $\hat{c}_s$ for $H$ and $\hat{H}$ respectively), after suming over $i$ we get

$$\sum_{s=0}^{\infty} \hat{\alpha}^{(s)}(\hat{\alpha}^{(s+1)} - \kappa\alpha^{(s+1)}) = 0. \tag{7.5}$$

Since the problem is (semi-) convex there is always a (stable) solution, and it can be found with the help of numerical tools.

**Remark 7.1** *Notice that all we did is differentiating the Lagrangian, written* *with the help of the cepstrum, using the chain rule*

$$\sum_{s=0}^{\infty} s \frac{\partial \hat{c}_s}{\partial \hat{\alpha}_i} (\hat{c}_s - \kappa c_s) = 0. \tag{7.6}$$

**Remark 7.2** *It is tempting to set $\hat{\alpha}^{(s+1)} - \kappa \alpha^{(s+1)} = 0$ (or $\hat{c}_s = \kappa c_s$) for all $s$ but unless $n = \infty$ this cannot be done. Below we suggest that this approach can be used to do an approximation.*

**Example 7.1** *Taking the original system to be of order $m = 4$ with the poles $[\alpha_1, \alpha_2, \alpha_3, \alpha_4] = [0.3, 0.5, 0.7, 0.9]$*

$$H(z) = \frac{z^4}{(z-0.3)(z-0.5)(z-0.7)(z-0.9)} \tag{7.7}$$

*and taking the approximate system to be with the same order $n = 4$, we get the optimal poles $\hat{\alpha}'s$ as function of $\beta$, shown in Figure 7.1. The solution is obtained by optimizing the Lagrangian numericaly (using matlab's function fminsearch.m). Notice the bifurcations taking place in critical $\beta_c$. For $\beta$ small from a $\beta_c$ the poles concide having the same value. Those bifurcations happen also if we take $n < m$.* ◇

Obviously adding poles does not by itself increase the information $I(u_p; \hat{y}_f)$ and can actually reduce it. This is because the sets of systems of order k is a subset of the set of systems with higer order $n > k$ (we have $\{H_k\} \subset \{H_n\}$) and thus by increasing the system order we can only do better. By inspecting Figure 7.1 we find that what is really costly is to distinguish between the poles. By inceasing $\beta$ we reveal the fine structure of the original process by distinguishing between the poles which result in increasing the complexity of the approximating system.

From the above discussion it seems that the route to the limit $\beta \to 0$ with $\beta > 0$ for different approximating system dimension $n$ is not the same (although the limit itself is the same; for the point $\beta = 0$ the solutions are all the same for all $n$: reaching $\hat{H}_n = 1$ in a continuous fashion). In other words, unlike in the GIB but similarlly to the solution to problem 5 in chapter 4, the order of the system does not reduce with $\beta$. But as we are going to see in the next section although the order of the system is not exactly reduced the exact solution imply that we can reduce it in a controlled fashion without lossing the relevant information.

**Remark 7.3** *The property that by adding poles we can only do better is the main reason for the obstacles raised in the solution of problem 5. As in here the solution to problem 5 "wants" to use all the poles of the system and thus we wont get a reduced order system. But since here we have the control on the dimension of our approximating system we can quantify what we miss by truncating the dimesion (see next section).*

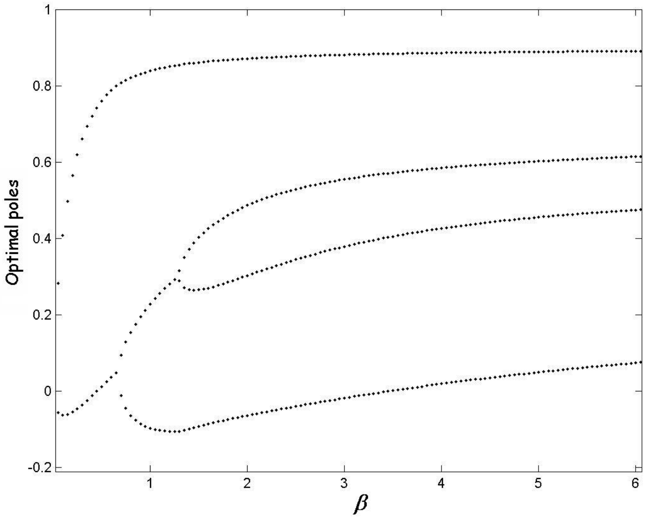

Figure 7.1: Optimal poles as a function of $\beta$ for example 7.1. The original system poles are 0.3, 0.5, 0.7 and 0.9.

**Remark 7.4** *Bifurcations are generic for such optimization problems. Our Lagrangian (7.2) has a permutation symmetry of the poles and zeros ($S_n$ symme-*

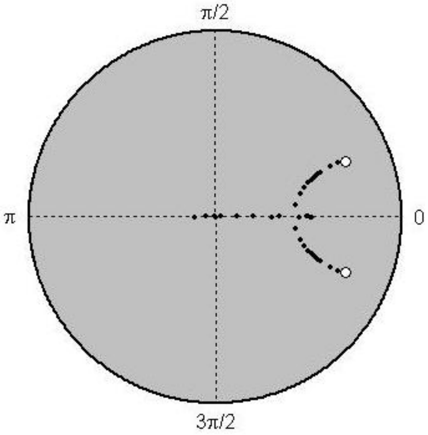

Figure 7.2: Optimal poles in the unit (complex) disc for a sequence of $\beta's$ for example 7.2. ∘ denotes the original poles.

*try). In [37] it is shown that for such optimization problems with the increase of* $\beta$ *we get a cascade of* $n$ *symmetry breaking happning in critical* $\beta$*s.*

**Example 7.2** *Taking the original system to be of order* $m = 2$ *with poles* $[\alpha_1, \alpha_2] = [0.7 + i0.3\ ,\ 0.7 - i0.3]$

$$H(z) = \frac{z^2}{(z - 0.7 - i0.3)(z - 0.7 + i0.3)} \tag{7.8}$$

*and the approximate system with the same order* $n = 2$ *we find that for* $\beta < \beta_c$ *the* $\hat{\alpha}'s$ *are on the real axis and only for* $\beta > \beta_c$ *they start to develope an imaginary part as shown in Figure 7.2.* ⋄

**Remark 7.5 Approximating scheme.** *Although that the condition to optimality (7.5) can be solved by numerical tools we can approximate the solution to get some analytic results. Since the terms in the sum of (7.5) decrease rapidly with* $s$ *we can set only the first* $n$ *terms to zero while the others are negligible.*

*Thus, given the order of the approximating system* $\hat{H}$ *is* $n$ *we solve*

$$\hat{\alpha}^{(s)} = \kappa \alpha^{(s)} \qquad s = 1, \ldots, n. \tag{7.9}$$

*For example lets take $n = 2$. this means*

$$\hat{\alpha}^{(1)} \equiv \hat{\alpha}_1 + \hat{\alpha}_2 = \kappa\alpha^{(1)} \tag{7.10}$$

$$\hat{\alpha}^{(2)} \equiv \hat{\alpha}_1^2 + \hat{\alpha}_2^2 = \kappa\alpha^{(2)} \tag{7.11}$$

*which results in*

$$\hat{\alpha}_{1,2} = \frac{\kappa\alpha^{(1)}}{2} \pm \frac{1}{2}\sqrt{2\kappa\alpha^{(2)} - \kappa^2(\alpha^{(1)})^2}. \tag{7.12}$$

*This approximated solution have the same behavior as in figure 7.2 with the "critical" $\beta_c$ given by equating the discriminant to zero*

$$\kappa_c = \frac{2\alpha^{(2)}}{(\alpha^{(1)})^2} \tag{7.13}$$

*which agree with the numerical calculation only for poles close to the origin.*

*Notice that this approximation makes sense for all $\beta$ only if $0 < |\alpha^{(1)}| < 1$. but this (the upper bound) is consistent with what we are doing since truncating the sum of (7.5) in the $n$th term is equivalent to taking only the first $n$ terms in the expansion of (7.4) in powers of $\alpha$ and $\hat{\alpha}$, i.e regarding them small so $|\alpha^{(1)}| < 1$ is feasible. This approximation is more accurate as we take $\alpha \to 0$. On the other hand this approximation become accurate as $\kappa \to 0$ (or equivalently $\beta \to 0$) for all $\alpha^{(1)}$.*

*Since we can solve the problem numerically to any accuracy we dont discuss this approximation further.*

**Example 7.3 Symmetric case.** *An interesting case is when the original system is symmetric by having opposite poles $\alpha_1 = -\alpha_2 = \alpha$. We can solve (7.4) analytically for both $n = 1$ and $n = 2$. Taking $n = 1$ we get that the stable solution is given by (see figure 7.3)*

$$\hat{\alpha}^2 = \begin{cases} \frac{2\alpha^2\kappa - 1}{\alpha^2(2\kappa - 1)} & \kappa > \frac{1}{2\alpha^2} \\ 0 & \kappa \leq \frac{1}{2\alpha^2} \end{cases} \tag{7.14}$$

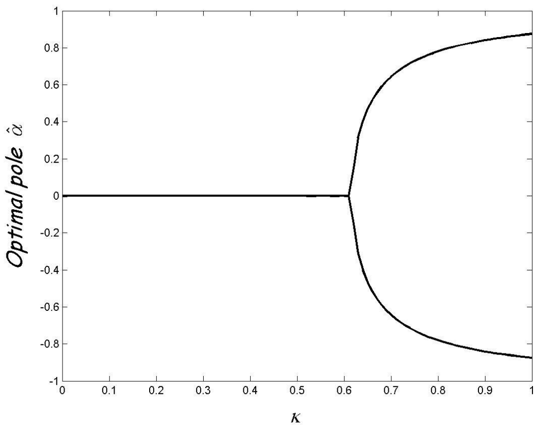

Figure 7.3: Optimal pole $\hat{\alpha}$ as a function of $\kappa$ for the symmetric case $\alpha_1 = -\alpha_2 = 0.9$ (see example 7.3).

*We see that if $\alpha > \frac{1}{\sqrt{2}}$ there is a critical $\kappa_c = \frac{1}{2\alpha^2}$ ($\beta_c = \frac{1}{2\alpha^2-1}$) such that for $\kappa < \kappa_c$ the solution is $\hat{\alpha} = 0$ but for $\kappa > \kappa_c$ a spontaneous symmetry breaking occurs and we have $\hat{\alpha} \neq 0$ (see figure 7.3). In this examlpe we can also calculate analytically for $n = 2$ and we find that the poles goes smoothly with $\kappa$ in a symmetric fashion from the origin toward the original symmetric pair. Those results agree with numerical calculations.* ◇

## 7.2 Characterizing the Information Curves

We now want to characterize the analog in our case for the information curves.

Suppose we are given an AR model $H$. we write again our problem as

$$R_n(D) = \min_{\hat{H}_n : D(H,\hat{H}_n) \leq D} I(u_{past}, \hat{y}_{future}) \tag{7.15}$$

with the distortion function $D(H, \hat{H}_n) = I(u_{past}, \tilde{y}_{future})$ where $\tilde{y}$ is the residual process.

In words, for a given distortion limit $D$ (i.e for a given $\beta$), what is the best that we can do with an AR model of order $n$? That is, what is the minimum information between past and future needed so that we wont get a distortion larger then $D$ by using an AR model of order $n$?

In particular, what is the best that we can do with an AR model of any order? To answer this question we need to observe that by not limiting the structure of the cepstrum coefficients to that of a finite order AR model we get the desired answer. This is because we get the largest set of processes possible. As can easily be seen this is equivalent to the tempting condition we stated earlier:

$$\hat{\alpha}^{(s)} = \kappa\alpha^{(s)} \Leftrightarrow \hat{c}_s = \kappa c_s \qquad \forall s. \tag{7.16}$$

Using this condition we can calculate the AR optimal information curve. Denoting $I \equiv I(u_{past}, y_{future})$ as the predictive information we have

$$R \equiv I^*(u_{past}, \hat{y}_{future}) = \sum s\hat{c}_s^2 = \kappa^2 \sum sc_s^2 = \kappa^2 I.$$

and

$$D \equiv I^*(u_{past}, \tilde{y}_{future}) = \sum s(\hat{c}_s - c_s)^2 = (\kappa - 1)^2 \sum sc_s^2 = (\kappa - 1)^2 I.$$

Eliminating $\kappa$ we get

$$D^{(\infty)}(R) = (\sqrt{R} - \sqrt{I})^2. \tag{7.17}$$

This curve is plotted in Figure 7.4 as the curve separeting the "Unachievable" part from the "Achievable". Notice that this curve is independent of the details of the original system and thus universal in some sense.

Another result can be obtained when the order of $\hat{H}$ is $n = 1$. In this case we have

$$R = -\log(1 - \hat{\alpha}^2)$$

or

$$\hat{\alpha} = \pm\sqrt{1 - e^{-R}}$$

and we have

$$D^{(1)}(R) = I + R + 2\log\prod_j(1 - \sqrt{1 - e^{-R}}\alpha_j) \tag{7.18}$$

We can also get approximated information curves calculated from the approximation discussed in the previous section. Although we can't write them analyticlly we have a parametric solution where the parameter is $\beta \in [0, \infty)$. Those are not very interesting since we can plot the real curves very accuratlly with numerical tools as the next example show.

**Example 7.4** *Taking the same system as in example 7.1, by minimizing the Lagrangian (7.2) we compute the information curves for different order of the approximating system $\hat{H}$. Those curves are shown in figure 7.4.*
*The dashed lines are solutions to the problem*

$$R(D) = \max_{\hat{H}: D(H, \hat{H}) \leq D} I(u_{past}, \hat{y}_{future}). \qquad \diamond \tag{7.19}$$

Taking the limit $\beta \to \infty$ we actually solve the problem of finding the closest system of order $n$ to the original system (of order $m$) where the distance is measured with the cepstrum norm (or the residual information between the past and the future)

$$D^{(n)}(\beta \to \infty) = \min_{\hat{H}_n} \left\| \log \hat{H}_n - \log H \right\|. \tag{7.20}$$

Indeed, in example 7.3, while for a low order $\hat{H}$ $(n = 0, 1, 2)$ the system cannot take the distortion to zero, from the figure it can be seen that there is't any

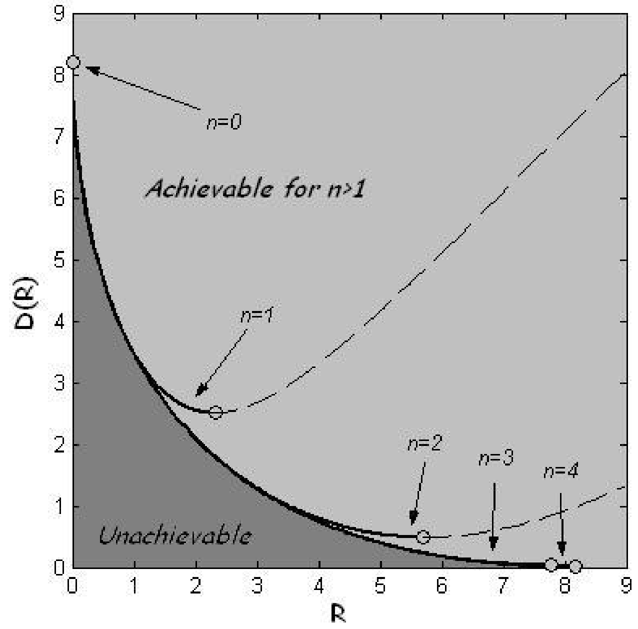

Figure 7.4: Information curves for the system of example 7.1 with different $\hat{H}$ order. For low $n$ there is a distortion that cannot be achieved. The dashed lines are solutions to a dual problem (see text).

practical difference between order $n = 3$ and $n = 4$.

As the next example illustrates, also for larger systems usually a much lower order system capture most of the relevant information.

**Example 7.5** *We take 50 systems of order 10 with random poles (with a prior uniform on the unit disk) and plot the averaged residual information* $\left\langle \frac{I(u_p;\tilde{y}_f)}{I(u_p;y_f)} \right\rangle$ *of the closest system as a function of its order* $n$*. See figure 7.5. Regarding the discussion below example 7.1 we conclude that this curve is nonincreasing. It is apparent that essentially a system of order* $n \approx 6$ *take the residual information very close to zero.* $n \approx 6$ *is typical for even larger systems.* $\diamond$

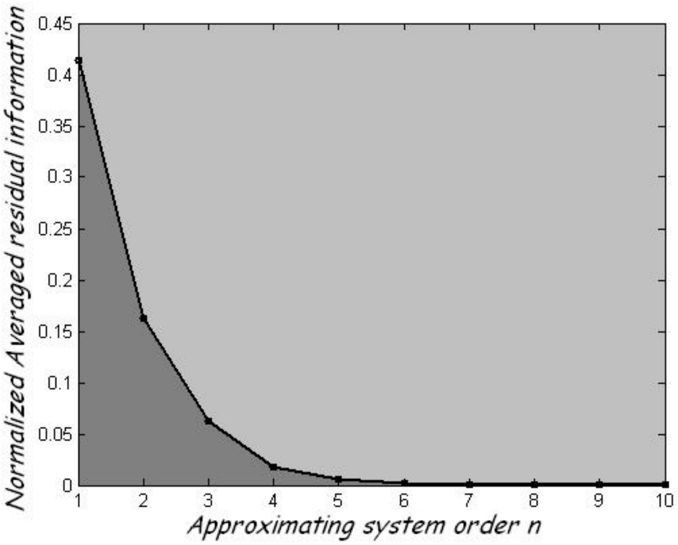

Figure 7.5: The averaged and normalized residual information $\left\langle \frac{I(u_p;\tilde{y}_f)}{I(u_p;y_f)} \right\rangle$ for $\beta \to \infty$ as a function of the approximating system dimension $n$ (see example 7.5).

## 7.3 Extending to ARMA Models.

AR models capture most of the complexity of this problem. Extending to ARMA models does not introduce any new kind of behavior. However ARMA models can approximate much better as the next example suggest.

**Example 7.6** *We take the system*

$$H(z) = \frac{(z-0.2)(z+0.2)}{(z-0.3)(z-0.5)}. \tag{7.21}$$

*and minimize (7.2). The information curves are plotted in figure 7.6. Notice that we capture all the relevant information by an approximating system with one pole and one zero while taking AR model with only one pole result in distortion for* $\beta \to \infty$*.*

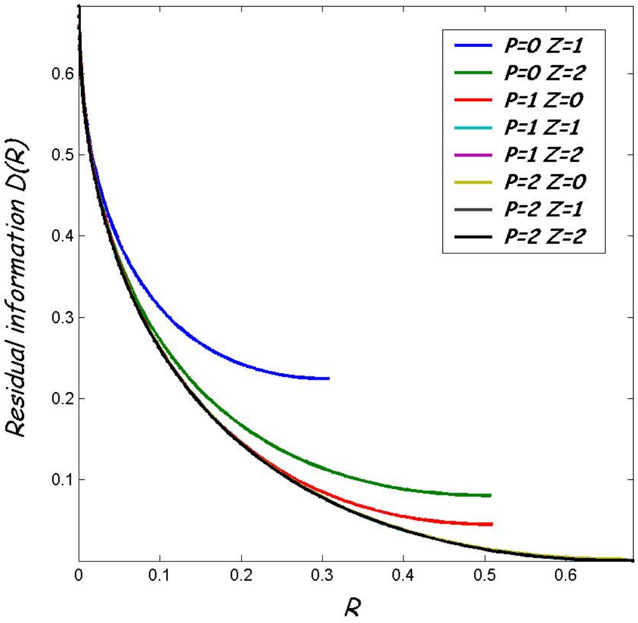

Figure 7.6: Information curves for example 7.6. The unobserved curves concide with the black one. P and Z state the number of poles and zeros allowed to be out of the origin respectivelly. Notice that by P=1 and Z=1 we capture all the relevant information while with P=1 and Z=0 (i.e AR model) we have finite distortion.

## 7.4 Discussion

As discussed below example 7.1, unlike the GIB, as $\beta \longrightarrow 0$ the systems does not reduce in order, or in other words, the approximating system uses all its available poles all the way to $\beta \longrightarrow 0$. But inspecting example 7.4 (and 7.6) we see that although not exact, for most of the systems the curves almost concide up to some $\beta$ (exception is e.g the symmetric case discussed in example 7.5). This suggest that the lower order systems does do almost the same as the higher dimensional ones as we tight the trade off. Thus we conclude that we actually can do a reduction of the system by inspecting the information curve and this algorithm gives the reduced system to any level of distortion one wishes. On the extreme, taking the limit $\beta \to \infty$, as in example 7.5, we get the best a system can do to approximate the original system with the criterion of having the same predictive information. We find that we can reduce the order of a large system to $n \approx 6$ and mimic the original system well.

# Chapter 8

# Summary

Considering the predictive information as the relevant quantifier of information in a stochastic process we discussed the application of the information bottleneck method to linear dynamical systems. We have found that the order of the new system does not reduce as we tight the constrain of limited information about the past although in most cases truncating the system order arbitrarily results in a system sitting very close to the information curve (which is optimal). Later we introduced a new model which is closely related to the information bottleneck method which is parametric and elegantly written with the help of the cepstrum. We characterized the information curves for this model and found that we can reduce the order of the system in a controlled fashion quantifying the loss of relevant information.

# Appendix A

# Singular Value Decomposition

The Singular Value Decomposition (SVD) is a matrix factorization widely used. This factorization is given in the next theorem.

**Theorem A.1** *Any complex $m \times n$ matrix $A$ can be factorized as († denotes complex conjugate transpose)*

$$A = U \Sigma V^{\dagger}, \tag{A.1}$$

*where $U \in \mathbb{C}^{m \times m}$ and $V \in \mathbb{C}^{n \times n}$ are unitary. The $\Sigma \in \mathbb{R}^{m \times n}$ matrix is real and is of the form*

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix} \tag{A.2}$$

*with $\Sigma_1 = diag(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r \times r}$, $r = Rank(A)$ and $\sigma_1 \geq \ldots \geq \sigma_r > 0$. This factorization is depicted in figure A.1.*
*The first $min(m, n)$ elements on the diagonal of $\Sigma$ are called the* singular values*. Since $Rank(A) = r$ there are $r$ non-zero singular values. The columns of $U$ are called the left singular vectors and are the eigenvectors of the symmetric semi-positive-definite matrix $AA^{\dagger}$ while the columns of $V$ are called the right singular vectors and are the eigenvectors of the symmetric semi-positive-definite matrix $A^{\dagger}A$. Notice that $\sigma_i^2$ is the i-th eigenvalue of $AA^{\dagger}$. If $A$ is real it can be shown*

*that we can take $U$, $V$ and $\Sigma$ to be also real.*

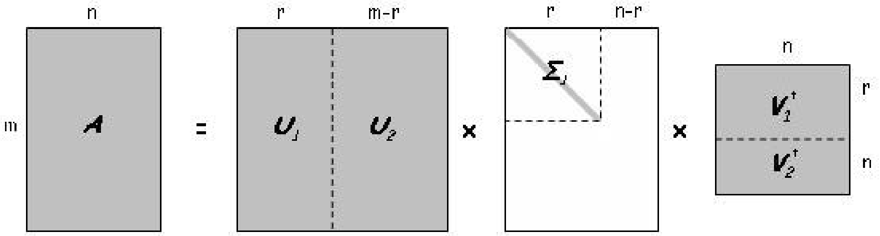

Figure A.1: Singular value decomposition.

*We can give an alternative factorization using the matrices depicted in figure A.1:*

$$A = U_1 \Sigma_1 V_1^{\dagger}. \tag{A.3}$$

*where $U_1 \in \mathbb{C}^{m\times r}$, $\Sigma_1 \in \mathbb{R}^{r\times r}$ and $V_1 \in \mathbb{C}^{n\times r}$ and $U_1 U_1^{\dagger} = V_1^{\dagger} V_1 = I_r$. This factorization is depicted in figure A.2.*

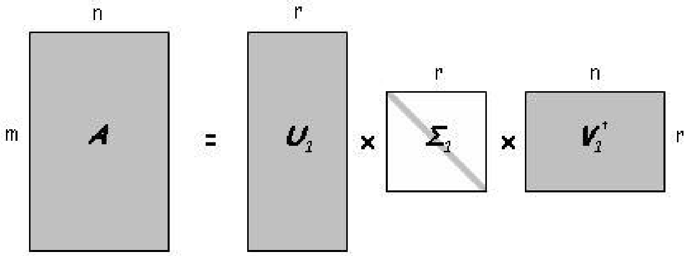

Figure A.2: Reduced singular value decomposition.

# Appendix B

# The Gaussian distribution

## B.1 The Distribution

The probability density function (multivariate normal distribution) of the Gaussian variables $X_1, X_2, \ldots, X_n$ is of the form

$$f(\mathbf{x}) = \frac{1}{(\sqrt{2}\pi)^n |\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)} \ , \tag{B.1}$$

where $\mathbf{x} = (x_1, \ldots, x_n)^T$ , $\mu$ is the mean vector and $\Sigma$ is the symmetric covariance matrix:

$$\Sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)]. \tag{B.2}$$

$|\cdot|$ denotes determinant. We use $\mathscr{N}_n(\mu, \Sigma)$ to denote this distribution.

## B.2 Conditioning

We have $X_1$ and $X_2$ two multivariate normal random variables with joint distribution

$$f(\mathbf{x}_1, \mathbf{x}_2) = \tag{B.3}$$

$$\frac{1}{(2\pi)^{(n_1+n_2)/2}|\Sigma|^{1/2}} \cdot exp\left(-\frac{1}{2}\begin{bmatrix} \mathbf{x}_1 - \mu_1 \\ \mathbf{x}_2 - \mu_2 \end{bmatrix}^T \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{x}_1 - \mu_1 \\ \mathbf{x}_2 - \mu_2 \end{bmatrix}\right),$$

where $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ and $\Sigma_{21} = \Sigma_{12}^T$ are the cross covariances.
First we have $x_1 \propto \mathcal{N}(\mu_1, \Sigma_{11})$ and $x_2 \propto \mathcal{N}(\mu_2, \Sigma_{22})$. Now we want to know what is the distribution of $f(\mathbf{x}_1|\mathbf{x}_2)$.

**Theorem B.1 (The distribution of $f(\mathbf{x}_1|\mathbf{x}_2)$).**

$$\mathbf{x}_1|\mathbf{x}_2 \propto \mathcal{N}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right) \tag{B.4}$$

# Appendix C

# PFMI via Poles and Zeros - Proof

In this appendix we prove theorem 3.9. The proof is taken from [19] and [17].

**Theorem C.1** *Given $H(z) = \frac{b(z)}{a(z)}$ stable minimum phase ARMA model with poles $\alpha_1, \ldots, \alpha_n$ and zeros $\beta_1, \ldots, \beta_n$ that is driven by white noise with variance $\sigma^2 = 1$. We have*

$$I(u_p; y_f) = \frac{1}{2} log \frac{\prod_{i,j=1}^{n} \left|1 - \alpha_i \overline{\beta}_j\right|^2}{\prod_{i,j=1}^{n}(1 - \alpha_i \overline{\alpha}_j) \prod_{i,j=1}^{n}(1 - \beta_i \overline{\beta}_j)}. \tag{C.1}$$

For simplicity we will prove this theorem only for distinct poles and zeros. Let us define $H^{(1)} = \frac{z^n}{a(z)}$ and $H^{(2)} = \frac{z^n}{b(z)}$, We have $H = H^{(1)}(H^{(2)})^{-1}$. Recalling the discussion in remark 3.5 we have $cc^2(u_f, y_f) = cc^2(y_f^{(1)}, y_f^{(2)})$ and using the fact that $cc^2(u_p, y_f) = 1 - cc^2(u_f, y_f)$ (see theorem 3.8) we find that

$$I(u_p, y_f) = \sum_{i=1}^{n} \log \tau_i^2, \tag{C.2}$$

where $\tau_i$ is the $ith$ canonical correlation $cc(y_f^{(1)}, y_f^{(2)})$. We now observe that $y_f^{(1)}$ is spaned by $(x_n^{(1)}, u_f(k \geq n))$ and similarly $y_f^{(2)}$ is spaned by $(x_n^{(2)}, u_f(k \geq n))$.

Thus we conclude that the $n$ canonical correlations that are different from 1 are the canonical correlations between $x_n^{(1)}$ and $x_n^{(2)}$, $cc(x_n^{(1)}, x_n^{(2)})$. Since we have stationarity we can replace $n \rightarrow 0$ and we have

$$\begin{aligned} x_0^{(1)} =& \quad \mathscr{C}^{(1)} u_p && \text{(C.3)} \\ x_0^{(2)} =& \quad \mathscr{C}^{(2)} u_p. && \text{(C.4)} \end{aligned}$$

Taking expectations we have

$$\Sigma_{ij} \equiv E\left[x_0^{(i)} (x_0^{(j)})^\dagger\right] = \mathscr{C}^{(i)} (\mathscr{C}^{(j)})^\dagger. \tag{C.5}$$

Since the $\tau_i^2$ are the eigenvalues of $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ we have

$$I(u_p, y_f) = \log \prod_{i=1}^{n} \tau_i^2 = \log |\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}| = \log \frac{|\Sigma_{12}||\Sigma_{21}|}{|\Sigma_{11}||\Sigma_{22}|}. \tag{C.6}$$

To calculate $|\Sigma_{ij}|$ we first need to discuss diagonal realizations.

As discussed in section 2.4 since $I(u_p, y_f)$ is an I/O propetry we can choose any realization we wish to calculate $|\Sigma_{ij}|$. We will find that using the diagonal realization make the calculation of $|\Sigma_{ij}|$ easy. To get the diagonal realization [24] of a system $H$ with distinct poles $\lambda_1, \ldots, \lambda_n$ we expand the stricly proper transfer function $\bar{H}(z) = H(z) - \lim_{z\rightarrow\infty} H(z)$ in terms of partial fractions

$$\bar{H}(z) = \sum_{i=1}^{n} \frac{1}{z - \lambda_i} R_i, \tag{C.7}$$

where $R_i$ is found by the relation

$$R_i = \lim_{z\rightarrow\lambda_i} (z - \lambda_i)\bar{H}(z). \tag{C.8}$$

We write

$$R_i = C_i B_i, \quad i = 1, \ldots, n. \tag{C.9}$$

It can be shown [24] that

$$A = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ B_2 \\ \vdots \\ B_n \end{pmatrix}, \tag{C.10}$$

and

$$C = \begin{pmatrix} C_1 & C_2 & \dots & C_n \end{pmatrix} \tag{C.11}$$

is a minimal realization of $H(z)$. Notice that we can take $B_i = 1 \; \forall i$. In this case the controllability matrix is taking the form

$$\mathscr{C} = \begin{bmatrix} B & AB & A^2B & \dots \end{bmatrix} = \begin{bmatrix} 1 & \lambda_1 & \lambda_1^2 & \dots \\ 1 & \lambda_2 & \lambda_2^2 & \dots \\ \vdots & \vdots & \vdots & \ddots \\ 1 & \lambda_n & \lambda_n^2 & \dots \end{bmatrix},$$

Going back to our problem we have

$$\mathscr{C}^{(1)} = \begin{bmatrix} 1 & \alpha_1 & \alpha_1^2 & \dots \\ 1 & \alpha_2 & \alpha_2^2 & \dots \\ \vdots & \vdots & \vdots & \ddots \\ 1 & \alpha_n & \alpha_n^2 & \dots \end{bmatrix}, \quad \mathscr{C}^{(2)} = \begin{bmatrix} 1 & \beta_1 & \beta_1^2 & \dots \\ 1 & \beta_2 & \beta_2^2 & \dots \\ \vdots & \vdots & \vdots & \ddots \\ 1 & \beta_n & \beta_n^2 & \dots \end{bmatrix}. \tag{C.12}$$

Using the geometric series $\sum_{i=0}^{\infty} r^i = \frac{1}{1-r}$ for $|r| < 1$ we find

$$\Sigma_{12} = \mathscr{C}^{(1)} (\mathscr{C}^{(2)})^{\dagger} = \begin{bmatrix} \frac{1}{1-\alpha_1\bar{\beta}_1} & \frac{1}{1-\alpha_1\bar{\beta}_2} & \cdots & \frac{1}{1-\alpha_1\bar{\beta}_n} \\ \frac{1}{1-\alpha_1\bar{\beta}_1} & \frac{1}{1-\alpha_1\bar{\beta}_2} & \cdots & \frac{1}{1-\alpha_1\bar{\beta}_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{1-\alpha_n\bar{\beta}_1} & \frac{1}{1-\alpha_n\bar{\beta}_2} & \cdots & \frac{1}{1-\alpha_n\bar{\beta}_n} \end{bmatrix}, \tag{C.13}$$

and similarly to $|\Sigma_{11}|$,$|\Sigma_{22}|$ and $|\Sigma_{21}|$. This matrix has the form of a Cauchy matrix which has the determinant (proven by induction)

$$|\Sigma_{12}| = \frac{\prod_{i<j}^{n}(\alpha_i - \alpha_j)(\bar{\beta}_i - \bar{\beta}_j)}{\prod_{i,j}^{n}(1 - \alpha_i \bar{\beta}_j)}. \tag{C.14}$$

similarly we have $|\Sigma_{11}| = \frac{\prod_{i<j}^{n}(\alpha_i - \alpha_j)(\bar{\alpha}_i - \bar{\alpha}_j)}{\prod_{i,j}^{n}(1 - \alpha_i \bar{\alpha}_j)}$ and $|\Sigma_{22}| = \frac{\prod_{i<j}^{n}(\beta_i - \beta_j)(\bar{\beta}_i - \bar{\beta}_j)}{\prod_{i,j}^{n}(1 - \beta_i \bar{\beta}_j)}$. Putting it all together we get

$$I(u_p; y_f) = \frac{1}{2} log \frac{\prod_{i,j=1}^{n} \left|1 - \alpha_i \overline{\beta}_j\right|^2}{\prod_{i,j=1}^{n}(1 - \alpha_i \overline{\alpha}_j) \prod_{i,j=1}^{n}(1 - \beta_i \overline{\beta}_j)}. \qquad \square \tag{C.15}$$

**Remark C.1** *We have proven the theorem for distinct poles and zeros but it is also true for arbitrary multiplicity of poles and zeros (including in the origin). All we have to do is to use the Jordan form of the matrix $A$ (instead of the diagonal used here). It can be shown that a pole with multiplicety $m > 1$ contribute the following block to the contollability matrix*

$$\mathscr{C}^{(m)} = \begin{bmatrix} \binom{0}{0} & \binom{1}{0}\alpha & \binom{2}{0}\alpha^2 & \cdots & \binom{m-1}{0}\alpha^{m-1} & \binom{m}{0}\alpha^{m} & \cdots \\ 0 & \binom{1}{1} & \binom{2}{1}\alpha & \cdots & \binom{m-1}{1}\alpha^{m-1} & \binom{m}{1}\alpha^{m-1} & \cdots \\ 0 & 0 & \binom{2}{2} & \cdots & \binom{m-1}{2}\alpha^{m-3} & \binom{m}{2}\alpha^{m-2} & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots \\ 0 & 0 & 0 & \cdots & \binom{m-1}{m-1} & \binom{m}{m-1}\alpha & \cdots \end{bmatrix},$$

*where $\binom{j}{i} = \frac{j!}{i!(j-i)!}$ is the binomial coefficient. Using those modified controllabilty matrices result in the same solution form, equation (C.1).*

# Bibliography

[1] William Bialek, R. R. de Ruyter van Steveninck and Naftali Tishby. Optimization principles for organisms. Preprint (2007).

[2] Carl T. Bergstrom and Michael Lachmann. The fitness value of information. Working paper, arXiv physics.soc-ph/0707.0609v1.

[3] Naftali Tishby, Fernando Pereira and William Bialek. The information bottleneck method. Proc. 37th Allerton Conference on Communication and Computation, (1999).

[4] D. Polani, T. Martinetz, J.T. Kim. An Information-Theoretic Approach for the Quantification of Relevance. Lecture Notes In Computer Science; Vol. 2159, Proceedings of the 6th European Conference on Advances in Artificial Life (2001).

[5] D. Polani, C. Nehaniv, T. Martinetz and J.T. Kim. Relevant Information in Optimized Persistence vs. Progeny Strategies. In M.Rocha, L., Bedau, M., Floreano, D., Goldstone, R., Vespignani, A., and Yaeger, L., editors, (2006). Proc. Artificial Life X (2006).

[6] Y. Bar-Hillel and R. Carnap. Semantic Information. The British Journal for the Philosophy of Science, Vol. 4, No. 14., pp. 147-157 (1953).

[7] W Bialek, I Nemenman and N Tishby. Predictability complexity and learning. Neural Comp 13, 24092463 , physics/0007070, (2001).

[8] U.B. Desai, D. Pal. A Realization Approach to Stochastic Model Reduction and Balanced Stochastic Realizations. Decision and Control, 21st IEEE Conference on, (1982).

[9] Creutzig Felix. The Past-Future Information Bottleneck in Dynamical Systems. In preparation (2007).

[10] Thomas M. Cover, Joy A. Thomas. Elements of Information Theory. Wiley-Interscience Publications (1991).

[11] C.E Shannon. A Mathematical Theory of Communication. Reprinted with corrections from The Bell System Technical Journal, Vol. 27, pp. 379423, 623656 (1948).

[12] H. Hotelling. Relations Between Two Sets of Variates. Biometrika, Vol. 28, No. 3/4. pp. 321-377 (1936).

[13] Piet de Jong, Jeremy Penzer. The ARMA model in state space form. Statistics and Probability Letters 70 119125 (2004).

[14] E. T. Jaynes. Information Theory and Statistical Mechanics. Phys. Rev. 106, 620 - 630 (1957).

[15] Gal Chechik, Amir Globerson, Naftali Tishby, Yair Weiss. Information Bottleneck for Gaussian Variables. The Journal of Machine Learning Research Volume 6 Pages: 165 - 188 (2005).

[16] Noam Slonim. The Information Bottleneck: Theory and Applications. Phd Thesis, HUJI (2002).

[17] De Cock, K. De Moor, B. Subspace angles between linear stochastic models. Proceedings of the 39th IEEE Conference on Decision and Control, 2000, vol.2 1561-1566 (2000).

[18] Katrien De Cock, Bart De Moor, "Canonical correlations between input and output processes of linear stochastic models", in Proceedings of the International Symposium on the Mathematical Theory of Networks and Systems (MTNS 2002), University of Notre Dame, USA, August 2002.

[19] Katrien De Cock. Phd thesis - Principal angles in system theory, information theory and signal processing (2002).

[20] Richard J. Martin. A Metric for ARMA Processes, IEEE Trans. Signal Process. 48 (4) (April 2000) 1164-1170.

[21] A. J. Lawrance. Partial and Multiple Correlation for Time Series. The American Statistician, Vol. 33, No. 3. , pp. 127-130 (1979).

[22] Peter Harremos, Naftali Tishby. The Information Bottleneck Revisited or How to Choose a Good Distortion Measure. Submitted to ISIT (2007).

[23] Bruce C. Moore. Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction. IEEE TRANSACTIONS ON AUTOMATIC CONTROL. VOL. AC-26, NO. 1. (1981).

[24] P.J. Antsaklis, A.N. Michel. Linear Systems. McGraw-Hill (1997).

[25] Papoulis A. Probability, random variables, and stochastic processes. Fourth edition. New York: McGraw Hill, (2002).

[26] A. V. Oppenheim. Superposition in a class of nonlinear systems. Tech- ards nical Report 432, Research Laboratory of Electronics, MIT, Cambridge, erence USA, March (1965).

[27] A.V. Oppenheim, R.W. Schafer. Digital Signal Processing. Prentice-Hall (1975).

[28] A.V. Oppenheim, R.W. Schafer. From frequency to quefrency: a history of the cepstrum. Signal Processing Magazine, IEEE Volume 21, Issue 5, Sept. Page(s):95 - 106 (2004).

[29] B.P. Bogert, M.J.R. Healy, and J.W. Tukey. The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking. Time Series Analysis, M. Rosenblatt, Ed. , ch. 15, pp. 209-243 ( 1963).

[30] L.R. Rabiner, R.W. Schafer. Digital Processing of Speech Signals. Prentice-Hall, Signal Processing Series (1978).

[31] B.L. Ho, R.E. Kalman. Effective construction of linear state-variable models from input /output functions (Algorithm for minimal realization of linear finite-dimensional dynamical system displayed by Markov parameters) REGELUNGSTECHNIK. Vol. 14, no. 12, pp. 545-548 (1966).

[32] I. A. Ibragimov, and Y. A. Rozanov. Gaussian Random Processes. New York, Heidelberg, Berlin: Springer-Verlag (1978).

[33] Jewell, N. P. and Bloomfield, P. Canonical correlations of past and future for time series: definitions and theory. Annals of Statistics 11, 83747 (1983).

[34] Nicholas P. Jewell; Peter Bloomfield; Flavio C. Bartmann. Canonical Correlations of Past and Future for Time Series: Bounds and

Computation. The Annals of Statistics, Vol. 11, No. 3., pp. 848-855 (1983).

[35] Li, L. M. and Xie, Z. Model selection and order determination for time series by information between the past and the future. Journal of Time Series Analysis 17, 6584 (1996).

[36] Li, Lei M. Some Notes on Mutual Information Between Past and Future. Journal of Time Series Analysis, Vol. 27, No. 2, pp. 309-322, March (2006).

[37] Albert E. Parker, Tom Gedeon. Bifurcation Structure of a Class of SN-invariant Constrained Optimization Problems. Journal of Dynamics and Differential Equations, Volume 16, Number 3, pp. 629-678(50) (2004).