Håkan Hjalmarsson

Learning Dynamic Models
System Identification 20/20

Contents

Notations

$\delta(t)$ = Dirac's delta and Kronecker's delta

$\mathbf{1}_A(x)$ = the indicator function for the set A, equals 1 for $x \in A$ and 0 otherwise.

A^T = transpose of the matrix A

A^* = complex conjugate transpose of the matrix A

$\overline{z}$ = complex conjugate of z

$\det A$ = determinant of the matrix A

$\lfloor f, g \rfloor$ = the Gram matrix consisting of the inner-products between the elements of $f = \begin{bmatrix} f_1 & \dots & f_n \end{bmatrix}^T$ and $g = \begin{bmatrix} g_1 & \dots & g_m \end{bmatrix}^T$

$$\lfloor f, g \rfloor = \begin{bmatrix} \langle f_1, g_1 \rangle & \dots & \langle f_1, g_m \rangle \\ \vdots & \dots & \vdots \\ \langle f_n, g_1 \rangle & \dots & \langle f_n, g_m \end{bmatrix}$$

1
Signals and Systems

The intention with these lecture notes is to provide the reader with a thorough understanding of the many different facets of the problem of estimating models of dynamical systems using data. This topic is generally known as system identification and shares many aspects with other types of learning problems in, e.g., statistical and machine learning. In fact, the underlying principles are the same. To emphasize the kinship with these areas we have chosen the title to be dynamic model learning. While the notes heavily leans on general learning theory, the particular aspects of dynamical systems is emphasized and the purpose of this chapter is to introduce some general tools for signals and systems.

1.1 Signals

By a signal we mean a function of time or some other variables, or combinations thereof. For example, in a paper machine the thickness of the paper at one of the rolls in the machine can be viewed as a signal being a function both of time and the cross-directional position of the sheet.

1.1.1 Continuous time signals

It is useful to work with different classes of signals.

Definition 1.1.1. *The space $L_p(C)$, $0 < p < \infty$ consists of all measurable functions $F : C \to \mathbb{C}^{n\times m}$ on C for which*

$$\|F\|_p := \left(\int_C \|F(t)\|_F^p dt\right)^{1/p} < \infty$$

The class $L_\infty(C)$ consists of all measurable functions $F : C \to \mathbb{C}^{n\times m}$ on C for which

$$\|F\|_\infty := \operatorname*{ess\,sup}_{t\in C} \overline{\sigma}(F(t)) < \infty$$

where $\overline{\sigma}(A)$ denotes the largest singular value of the matrix A.

The essential supremum for a real-valued function f is defined as

$$\operatorname*{ess\,sup}_{t\in C} f(t) = \inf\{a : \ f(t) \le a \text{ almost everywhere in } C$$

where almost everywhere means except on a set that has Lesbegue measure zero in C. The L_p spaces are complete metric spaces (Banach spaces). The intersection $L_1(\mathbb{R}) \cap L_\infty(\mathbb{R})$ is a subset of all $L_p(\mathbb{R})$, $1 \le p \le \infty$ but otherwise there is no particular relation between the elements in these spaces. Another domain is the unit circle in the complex plane $\mathbb{T} = \{z : |z| = 1\}$ and here $\infty \ge p \ge q \ge 1 \Rightarrow L_p(\mathbb{T}) \subset L_q(\mathbb{T})$.

The Fourier transform of a signal $s(t)$ is defined as

$$S(i\omega) = \int_{-\infty}^{\infty} s(t)e^{-i\omega t}dt \tag{1.1}$$

and the inverse Fouriertransform as

$$\bar{s}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(i\omega)e^{i\omega t}d\omega \tag{1.2}$$

For certain signal classes S is well defined and $\bar{s}(t) = s(t)$ which gives (1.2) the interpretation as a decomposition of the signal s into sinusoidal components where $S(i\omega)/(2\pi)$ represents the contribution of different frequency components to s. We then say that s has a Fourier representation. All signals in L_1 do not have a Fourier representation as shown by Kolmogorov in a famous counterexample [1]. The following is known.

Theorem 1.1.1. *i) Suppose that $s \in L_1(\mathbb{R})$, then its Fourier transform S is uniformly continuous and vanishes at infinity.*

ii) Suppose that $s \in L_1(\mathbb{R})$ and that its Fourier transform $S \in L_1(\mathbb{R})$. Then

$$\bar{s}(t) = \int_{-\infty}^{\infty} S(i\omega)e^{i\omega t}d\omega$$

is continuous, vanishes at infinity and $\bar{s}(t) = s(t)$ almost everywhere[2].

iii) Suppose that $s \in L_p(\mathbb{R})$, $1 < p < \infty$, with Fourier transform S. Then

$$\lim_{R\to\infty} \int_{|\omega|\le R} S(i\omega)e^{i\omega t}d\omega = s(t) \quad \textit{almost everywhere}$$

Proof. For Part i) see Appendix B.1.1 in [3]. Part ii) is Theorem 9.11 in [4]. Part iii) was proven by Carleson for $p = 2$ [5] and Hunt [6]. □

1.1.2 *Discrete time signals*

Often data is available as sequences of discrete time signals $\{s(n)\}_{n=1}^N$. These are often continuous time signals $s(t)$ sampled with a certain sampling interval T resulting in $\{s(nT)\}_{n=1}^N$ but may also be actual discrete time signals, e.g. the number of transactions per day on the stock exchange. Discrete time signals are represented by sequences $\{s(t)\}_{t=-\infty}^{\infty}$, $s(t) \in \mathbb{C}^n$.

Definition 1.1.2. *The class ℓ_p, $0 < p < \infty$, consists of all sequences $\{s(t)\}$ for which*

$$\|s\|_p := \left(\sum_k |s(t)|^p\right)^{1/p} < \infty$$

[1] A. N. Kolmogorov. Une série de Fourier-Lebesque divergente presque partout. *Fund. Math.*, 4, 1923

[2] This means that the statement holds for all $t \in \mathbb{R}$ except for a set B which has Lesbegue measure zero, the latter loosely meaning that $\int_B dx = 0$.

[3] C.A. Desoer and M. Vidyasagar. *Feedback Systems: Input-Output Properties*. Academic Press, New York, 1975

[4] W. Rudin. *Real and Complex Analysis*. McGraw-Hill, London, 1986

[5] L. Carleson. On convergence and growth of partial sums of Fourier series. *Acta Math*, 116, 1966

[6] R.A. Hunt. On the convergence of Fourier series, orthogonal expansions and their continuous analogues. In *Proc. Conf., Edwardsville, Ill., 1967*, Southern Illinois Univ. Press, pages 235–255, Carbondale, Ill., 1968

The class ℓ_∞ consists of all sequences $\{s(t)\}$ for which

$$\|s\|_\infty := \sup_t |s(t)| < \infty$$

It holds that $\ell_p \subset \ell_q$ for $1 \leq p < q \leq \infty$. For a signal $s \in \ell_1$, the discrete time Fourier transform is defined as the Fourier series

$$S(e^{i\omega}) = \sum_{t=-\infty}^{\infty} s(t)e^{-i\omega t} \tag{1.3}$$

which is a 2π-periodic function. When $S \in L_1(\mathbb{T})$, the Fourier series coefficients

$$\bar{s}(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S(e^{i\omega})e^{i\omega t}$$

are well-defined and equal $s(t)$[7]

All ℓ_p spaces are complete and we can make ℓ_2 into a Hilbert space by introducing the inner product

$$\langle s, v\rangle = \sum_t \text{Trace}\{v^*(t)s(t)\}$$

Likewise $L_2(\mathbb{T})$ becomes a Hilbert space when equipped with the inner product

$$\langle S, V\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} \text{Trace}\left\{V^*(e^{i\omega})S(e^{i\omega})\right\} d\omega$$

For $L_2(\mathbb{T})$ the trigonometric functions $b_k(\omega) = e^{i\omega k}$, $k = 0, \pm 1, \pm, \ldots$ form a complete set of orthonormal functions and a direct consequence of this is the following theorem.

Theorem 1.1.2. *Any $S \in L_2(\mathbb{T})$ can be represented as the Fourier series* (1.3) *where*

$$s(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S(e^{i\omega})e^{i\omega t}$$

We recall from the theory of Hilbert spaces that $\|S\|_2 = 0$ does not imply that S is identically zero so convergence in norm does not mean point-wise convergence. Thus elements in $L_2(\mathbb{T})$ are grouped into equivalence classes where all pairwise differences between elements in one group have norm 0. Thus elements in a Hilbert space are distinguishable only up to these equivalence classes[8]. The representation in the previous theorem should therefore be interpreted in this sense.

The two spaces ℓ_2 and $L_2(\mathbb{T})$ are isomporphic meaning that there is a one-to-one relationship between the elements where the geometric properties represented by the inner product are preserved, i.e. it holds that

$$\langle S, V\rangle = \langle s, v\rangle$$

for all $S, V \in L_2(\mathbb{T})$, where $s \in \ell_2$ and $v \in \ell_2$ denote the Fourier coefficients of S and V, respectively. This is Parseval's theorem[9] In particular

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |S(e^{i\omega})|^2 d\omega = \|S\|_2^2 = \|s\|_2^2 = \sum_{t=-\infty}^{\infty} |s(t)|^2$$

Similar to Theorem 1.1.1 we have the following result.

[7] Section 9.4 .
W. Rudin. *Real and Complex Analysis.* McGraw-Hill, London, 1986

[8] 4.26 in
W. Rudin. *Real and Complex Analysis.* McGraw-Hill, London, 1986

[9] Section 4.26 in .
W. Rudin. *Real and Complex Analysis.* McGraw-Hill, London, 1986

Theorem 1.1.3. *i) Suppose that $s \in \ell_1$, then*

$$S(e^{i\omega}) = \lim_{N\to\infty} \sum_{t=-N}^{N} s(t)e^{-i\omega t} \tag{1.4}$$

exists and is continuous (since the convergence is uniform).

ii) Suppose that $s \in \ell_1$ and that $S \in L_1(\mathbb{T})$, where S is defined by (1.3). Then $\{s(t)\}$ are the Fourier series coefficients of S.

iii) Suppose that $S \in L_p(\mathbb{T})$, $1 < p < \infty$ with Fourier series coefficients $\{s(t)\}$. Then

$$\lim_{R\to\infty} \sum_{|t|\leq R} s(t)e^{i\omega\tau} = S(e^{i\omega}) \quad \textit{almost everywhere}$$

Proof. For Part i) and ii) see Section 9.4 in [10]. Part iii) was proven by Carleson for $p = 2$ [11] and Hunt [12]. □

A generalization of the Fourier transform is the two-sided z-transform

$$\tilde{S}(z) = \sum_{t=-\infty}^{\infty} s(t)z^{-t} \tag{1.5}$$

When $s(t) = 0$ for $t > 0$, this is a power series around $z = 0$, which defines a holomorphic (analytic) function in a disc of some radius R[13], called *region of convergence* (ROC). Recovering the sequence s from its z-transform requires the ROC to be known.

Example 1.1. *Let $\tilde{S}(z) = 1/(z - 0.5)$. This function has a singularity at $z = 0.5$ and we can expand it as*

$$\tilde{S}(z) = \frac{z}{z - 0.5} = \frac{1}{1 - 0.5/z} = \sum_{t=0}^{\infty} 0.5^t z^{-t}$$

which, comparing with (1.5), suggests that $s(t) = 0.5^t$ for $t \geq 0$ and $s(t) = 0$ for $t < 0$. However, we can also write

$$\tilde{S}(z) = \frac{z}{z - 0.5} = \frac{-2z}{1 - 2z} = -2z\sum_{t=0}^{\infty} 2^t z^t = \sum_{t=-\infty}^{0} -2^{-t-1}z^{-t-1} = \sum_{t=-\infty}^{-1} -2^{-t}z^{-t}$$

meaning that $s(t) = -2^{-t}$ for $t \leq 0$ and $s(t) = 0$ for $t > 0$.

The reason for the ambiguity is that the two expansions have different ROC. The first expansion is valid for $|z| > 0.5$, while the second is valid for $|z| < 0.5$.

When $s(t)$ is non-zero both for negative and positive t, (1.5) becomes a Laurent-series with the ROC being an annulus $\{z : |r| < |z| < R\}$. When the ROC includes the unit circle, we have that $\tilde{S}(e^{i\omega}) = S(e^{i\omega})$, i.e. the z-transform of s evaluated on the unit circle equals its Fourier transform. However, such signals have to decay exponentially fast as $|t| \to \infty$ and constitutes a limited class of signals, e.g. sinusoids are excluded[14] Thus the class of signals for which $\tilde{S}(z)$ is holomorphic in an annulus $r < |z| < R$ including the unit circle is quite restricted. What we can provide are larger classes of signals

[10] W. Rudin. *Real and Complex Analysis*. McGraw-Hill, London, 1986

[11] L. Carleson. On convergence and growth of partial sums of Fourier series. *Acta Math*, 116, 1966

[12] R.A. Hunt. On the convergence of Fourier series, orthogonal expansions and their continuous analogues. In *Proc. Conf., Edwardsville, Ill., 1967*, Southern Illinois Univ. Press, pages 235–255, Carbondale, Ill., 1968

[13] which may be 0

[14] Assume the ROC is $r < |z| < R$. Take $\delta > 0$ such that $R - \delta > 0$. Then the power series

$$\sum_{t=-\infty}^{-1} s(t)z^{-t}$$

converges for $z = R - \delta/2$ as this series belongs to the ROC. But then the series is absolutely convergent for every $|z| \leq R - \delta$, see Section 47 in , requiring which are holomorphic inside *or* outside the unit circle and well defined on the unit circle. For $s \in \ell_1$ for which $s(t) = 0$ for $t < 0$, $\tilde{S}(z)$ is holomorphic in $|z| > 1$ as the Laurent expansion converges in this set. $\tilde{S}(z)$ is also well defined on the unit circle, although we cannot claim that it is holomorphic there as the assumption $s \in \ell_1$ does not guarantee that $\tilde{S}(z)$ exists at any point outside the unit circle, and differentiation requires the function to be defined in an open neighborhood of the point where the derivative is computed. We also have that $\tilde{S}(e^{i\omega}) = S(e^{i\omega})$, i.e. the z-transform of s evaluated on the unit circle equals its Fourier transform. H_p spaces are spaces of functions holomorphic inside the unit circle for which the *radial limits*

$$\breve{S}(e^{i\omega}) := \lim_{r \to 1+} S(re^{i\omega})$$

exists. Above, 1+ indicates that the limit is taken from above.

Definition 1.1.3. *$H_p(\mathbb{T})$, $0 < p < \infty$ is the class of functions $F : \mathbb{T} \to \mathbb{C}^{n \times m}$ for which all elements are holomorphic in*[15] *$|z| > 1$ and for which there is an $M < \infty$ such that*

$$\int_{-\pi}^{\pi} \|F(re^{\omega})\|_F^p d\omega \leq M, \quad 1 < r < \infty$$

The class $H_p(\mathbb{T})$ is closely related to $L_p(\mathbb{T})$.

Theorem 1.1.4. *Let $1 < p < \infty$. Then $H_p(\mathbb{T})$ is the class of functions that can be written as*

$$S(z) = \sum_{t=0}^{\infty} \bar{s}(t) z^{-t}$$

where $\{\bar{s}(t)\}_{t=1}^{\infty}$ are the Fourier coefficiencts of some function in $L_p(\mathbb{T})$.

Proof. Exercise 25.d, Chapter 17 in [16]. Theorem 17.12 in [17] for $H_2(\mathbb{T})$, see Theorem B.2.1. □

For $S \in H_p(\mathbb{T})$, $p > 0$, its radial limit $\breve{S}$ exists and $\breve{S} \in L_2(\mathbb{T})$[18]. For the Fourier coefficients of the radial limit it follows that $\bar{s}(t) = 0$ for $t < 0$. This together with the previous theorem shows that for $1 < p < \infty$, $H_p(\mathbb{T})$ can be thought of as a subset of functions in $L_2(\mathbb{T})$ extended to $|z| \geq 1$.

For $H_2(\mathbb{T})$ this notion is exact as from the previous theorem we have that $H_2(\mathbb{T})$ is exactly characterized by functions holomorphic in $|z| > 1$ with series expansions

$$F(z) = \sum_{t=0}^{\infty} f(t) z^{-t}, \quad \text{for which } \{f(t)\} \in \ell_2 \tag{1.6}$$

and such functions are elements of $L_2(\mathbb{T})$ when seen as functions on $\mathbb{T}$ due to the isomorphism between ℓ_2 and $L_2(\mathbb{T})$.

1.2 Continuous time dynamic systems

Abstractly, a system is an entity that describes a set of relations between some signals. We will here use the somewhat simplistic notion

[15] We consider functions holomorphic outside the unit circle rather than inside as in Appendix B to conform with the standard used in signal processing and control theory. This only amounts to making the transformation $z \to 1/z$.

[16] W. Rudin. *Real and Complex Analysis.* McGraw-Hill, London, 1986

[17] W. Rudin. *Real and Complex Analysis.* McGraw-Hill, London, 1986

[18] Theorem B.2.2.

that a system maps trajectories of input signals to trajectories of output signals, i.e. a system is a map from a function space to another function space. In a dynamic system the output at a given time t depends not only of the values of the input at that point in time but also at other time points. If only the past influences the current output we say that the system is *causal*, conversely a system is *anti-causal* if only the future of the inputs affects the output. When both past and future play a role the system is said to be *non-causal*. We shall mainly deal with causal systems in this treatise.

Example 1.2. Modeling a Shock Absorber (contributed by Brett Ninness)

A simplified representation of a car shock-absorber as a parallel spring and damper is shown diagrammatically in Figure 1.2. Its purpose is to smooth the car height $y(t)$ compared to the road height $u(t)$. There are three forces acting on the car. [19] *The first one is due to gravity*

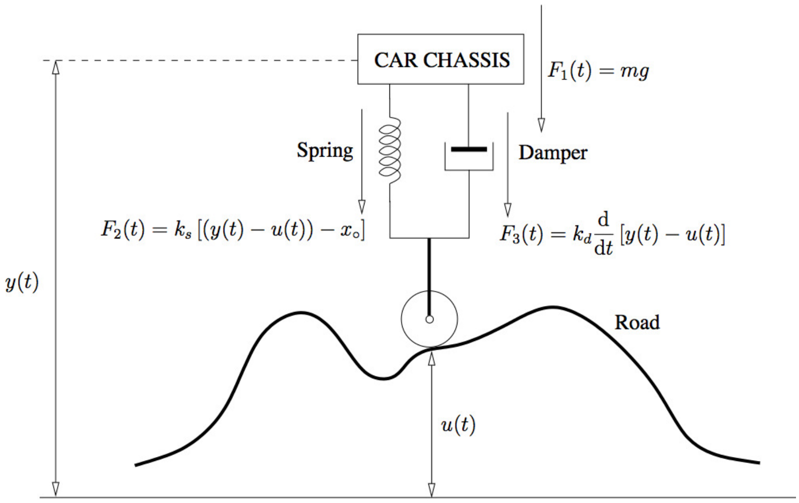
CAR CHASSIS
$F_1(t) = mg$
Spring
Damper
$F_2(t) = k_s\,[(y(t) - u(t)) - x_\circ]$
$F_3(t) = k_d \frac{\mathrm{d}}{\mathrm{d}t}\,[y(t) - u(t)]$
$y(t)$
Road
$u(t)$

Figure 1.1: Diagrammatic representation of shock absorber

$$F_1(t) = mg \tag{1.7}$$

where m is the mass of the car and g is acceleration due to gravity; 9.8 *ms*$^{-2}$. *The second force is due to the action of the spring*

$$F_2(t) = k_s[(y(t) - u(t)) - x_\circ] \tag{1.8}$$

where $x_\circ$ is the natural length of the spring with no force acting on it. The final force is due to the damper

$$F_3(t) = k_d \frac{\mathrm{d}}{\mathrm{d}t}[y(t) - u(t)]. \tag{1.9}$$

Newton's 2nd law states that the vector sum of forces $\mathbf{F}_1(T), \mathbf{F}_2(t), \cdots$ acting on a body must equal its mass m times its acceleration vector $\mathbf{a}(t)$

$$\text{(Newton's 2nd Law)} \qquad \sum_k \mathbf{F}_k(t) = \mathbf{a}(t). \tag{1.10}$$

[19] The direction of the force acting on the car is as shown by the arrows in figure 1.2)

Substituting the forces (1.7), (1.8) and (1.9) together into this law then gives a differential equation relationship [20] *between the road height $u(t)$ and the car height $y(t)$:*

$$mg + k_s[(y(t) - u(t)) - x_\circ] + k_d \frac{\mathrm{d}}{\mathrm{d}t}[y(t) - u(t)] = -m\frac{\mathrm{d}^2}{\mathrm{d}t^2}y(t). \quad (1.11)$$

This expression can be re-arranged into the slightly cleaner form

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}y(t) + \frac{k_d}{m}\frac{\mathrm{d}}{\mathrm{d}t}y(t) + \frac{k_s}{m}y(t) = \frac{k_d}{m}\frac{\mathrm{d}}{\mathrm{d}t}u(t) + \frac{k_s}{m}u(t) + \left(\frac{k_s}{m}x_\circ - g\right). \quad (1.12)$$

Note that at rest, when all first and higher order derivatives are zero, the model (1.12) reduces to [21]

$$y(t) - u(t) = x_\circ - \frac{mg}{k_s}. \quad (1.13)$$

Typically, we would want to model $y(t)$ with respect to this resting height being set at zero, in which case the model becomes

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}y(t) + \frac{k_d}{m}\frac{\mathrm{d}}{\mathrm{d}t}y(t) + \frac{k_s}{m}y(t) = \frac{k_d}{m}\frac{\mathrm{d}}{\mathrm{d}t}u(t) + \frac{k_s}{m}u(t) \quad (1.14)$$

■

As in Example 1.2, many models of dynamic systems are expressed in terms of ordinary differential equations (ODE) involving the output y and the input u

$$p(y(t), \dot{y}(t), \ldots, y^{(n)}(t), u(t), \dot{u}(t), \ldots, u^n(t)) = 0$$

1.2.1 *Linear time-invariant systems*

Finite dimensional systems. Finite dimensional Linear Time Invariant (LTI) systems is an important class of systems which can be described by linear time-invariant ODEs. In the case of scalar signals this means that

$$\sum_{k=0}^{n} a_k y^{(k)}(t) = \sum_{k=0}^{n} b_k u^{(k)}(t), \quad a_0 = 1, \quad (1.15)$$

where $y^{(k)}$ is the k'th derivative of y. Taking the one-sided Laplace transform of this expression and re-arranging terms gives

$$Y(s) = G(s)U(s) + Q(s) \quad (1.16)$$

where $Y(s)$ and $U(s)$ are the Laplace transforms of the output and input, respectively, where

$$G(s) = \frac{B(s)}{A(s)} := \frac{\sum_{k=0}^{n} b_k s^{n-k}}{\sum_{k=0}^{n} a_k s^{n-k}} \quad (1.17)$$

represents the system input-output behavior, and is known as the transfer function, and where[22]

$$Q(s) = \frac{B_Q(s)}{A(s)}$$

[20] Note that the right hand side of (1.11) involves a minus sign, since the vector sum on the left hand side is directed downwards, and hence the orientation of the dispacement vector differentiated on the right hand side of (1.11) must be consistent with this.

[21] That is, the resting height of the car above the road is the natural extension $x_\circ$ of the spring, minus the amount mg/k_s that the spring is compressed by the weight force due to the mass of the car.

[22]

$$B_Q(s) = \sum_{\tau=1}^{n-1} b^Q_\tau \, s^{n-\tau}$$

$$b^Q_\tau = \sum_{l=1}^{\tau}$$

$$(a_{\tau-l} y^{(l-1)}(0+) - b_{\tau-l} u^{(l-1)}(0+))$$

represents the transient behavior due to non-zero initial conditions.

Using (1.16) and properties of the Laplace transform, the input-output relationship in the time-domain can be written as

$$y(t) = \int_0^t g(\tau)u(t-\tau)d\tau + q(t) \tag{1.18}$$

where g is the inverse Laplace transform of $G(s)$, and known as the *impulse response* of the system, while $q(t) = \mathcal{L}^{-1}\{Q(s)\}$ is the *transient*. The name impulse response derives from that if $u(t) = \delta(t)$ (a Dirac impulse), then $u(t) = g(t)$.

Finite dimensional LTI systems can equivalently be described by linear time-invariant state-space models

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_o \tag{1.19}$$

$$y(t) = Cx(t) + Du(t) \tag{1.20}$$

where $x(t) \in \mathbb{R}^n$ is called the *state* at time t and represents the influence the past input has had on the system. For example, when the system is started at say time $t = 0$, the initial state (or condition) $x(0)$ needs to be specified. Often states can be given physical meaning such as velocity and acceleration.

Using the properties of the Laplace transform for (1.19) gives

$$\begin{aligned} G(s) &= C(sI - A)^{-1}B + D \\ Q(s) &= C(sI - A)^{-1}x_o \end{aligned}$$

We will use the operator p to denote differentiation $p = \frac{d}{dt}$ so that we, e.g., can write

$$y(t) = G(p)u(t)$$

to represent (1.15).

General LTI systems. Below we consider LTI systems with the input $u(t) \in \mathbb{R}^{n_u}$ and the output $y(t) \in \mathbb{R}^{n_y}$. Even though it from a practical point is most natural to use the description above that a system is started at some initial time with some initial conditions, for theoretical considerations it is of interest to study the behaviour when the input has been active over an infinite time interval. In general, the input-output relation for any LTI system is determined by its impulse response

$$y(t) = \int_{-\infty}^{\infty} g(\tau)u(t-\tau)d\tau \tag{1.21}$$

A LTI system is causal if and only if $g(\tau) = 0$, $\tau < 0$, which was the case for the systems we considered earlier.

The input-output behaviour depends obviously critically on the behaviour of the impulse response. For transfer functions that can be expressed as in (1.17), i.e. as rational functions, a partial fraction expansion gives

$$G(s) = \sum_{k=1}^{n} \frac{\alpha_k}{s - p_k}$$

where $\{p_k\}$ are the zeros of $A(s)$, called the *poles* of the system. As $1/(s-p)$ corresponds to e^{pt}

$$g(t) = \sum_{k=1}^{n} \alpha_k e^{p_k t}$$

We see that the impulse response grows exponentially if there is at least one pole in the right-half plane. Such a system is said to be *unstable*. For such systems the output (1.18) can be made to diverge with a bounded input as $t \to \infty$.

Definition 1.2.1. *A system G*

$$y = G(u)$$

is said to be bounded-input-bounded-output (BIBO) stable *if every bounded input results in a bounded output*

$$\forall u : |u(t)| \le M_u \ \forall t, \text{ for some } M_u < \infty \quad \Rightarrow \quad |y(t)| \le M_y \ \forall t, \text{ for some } M_y < \infty$$

Lemma 1.2.1. *A continuous time LTI system is BIBO-stable if and only if its impulse response $g \in L_1(\mathbb{R})$.*

Proof. Theorem 19, in Section 7.2, Chapter 4 in [23], which includes the time-varying case as well. □

Another notion of stability is strict stability.

Definition 1.2.2. *A continuous time LTI system with impulse response g is strictly stable if*

$$\int_{-\infty}^{\infty} |\tau| \ \|g(\tau)\|_F \, d\tau < \infty$$

BIBO-stability can be given an operator theoretic interpretation. With G BIBO stable, its impulse response belongs to $L_1(\mathbb{R}^{n_y \times n_u})$, but we can also see G as a map from $L_\infty(\mathbb{R}^{n_u})$ into $L_\infty(\mathbb{R}^{n_y})$ defined by (1.21). The norm of an operator is defined as

$$\|G\| = \sup_u \frac{\|G(u)\|}{\|u\|}$$

The if part of Lemma 1.2.1 is stated in operator form in the next lemma.

Lemma 1.2.2. *Suppose that G is defined by* (1.21) *with $g \in L_1(\mathbb{R}^{n_y \times n_u})$. Then $G : L_\infty(\mathbb{R}^{n_u}) \to L_\infty(\mathbb{R}^{n_y})$ with $\|G\| = \|g\|_1$.*

Proof. Theorem 3 in Section 6.2, Chapter 2 in [24]. □

One type of bounded-input signal is a sinusoid. Consider for simplicity a scalar system and let $u(t) = \cos(\omega t) = \mathrm{Re}\left\{e^{i\omega t}\right\}$. Then BIBO stability implies that

$$\begin{aligned} y(t) &= \int g(\tau) \mathrm{Re}\left\{e^{i\omega(t-\tau)}\right\} d\tau = \mathrm{Re}\left\{\int g(\tau) e^{-i\omega\tau} d\tau \, e^{i\omega t}\right\} \\ &= |G(i\omega)| \cos(\omega t + \arg G(i\omega)) \end{aligned}$$

[23] C.A. Desoer and M. Vidyasagar. *Feedback Systems: Input-Output Properties.* Academic Press, New York, 1975

[24] C.A. Desoer and M. Vidyasagar. *Feedback Systems: Input-Output Properties.* Academic Press, New York, 1975

where $G(i\omega)$ is the Fourier transform (1.1) of the impulse response. The reason why the Fourier transform is well defined is due to to the BIBO-stability, or, equivalently, that $g \in L_1(\mathbb{R})$.

Thus a BIBO-stable LTI system has the property that a sinusoidal input gives a sinusoidal output. The result carries over to multivariable LTI systems as well, with appropriate modifications.

It turns out that a BIBO-stable G also maps functions in $L_2(\mathbb{R}^{n_u})$ into $L_2(\mathbb{R}^{n_y})$ but the operator norm is different from the one in Lemma 1.2.2.

Theorem 1.2.1. *Suppose that G is defined by (1.21) with $g \in L_1(\mathbb{R}^{n_y \times n_u})$. Then $G : L_2(\mathbb{R}^{n_u}) \to L_2(\mathbb{R}^{n_y})$ with $\|G\| = \sup_\omega \|G(e^{i\omega})\|_2$.*

Proof. Theorem 7 in Section 6.2, Chapter 2 in [25]. □

1.2.2 *Non-linear state-space systems.*

Also non-linear ODEs can be represented on state-space from $L_2(\mathbb{R}^m)$ into $L_2(\mathbb{R}^m)$

$$\begin{aligned} \dot{x}(t) &= f(x(t), u(t)) \\ y(t) &= h(x(t), u(t)) \end{aligned}$$

1.3 *Discrete time systems*

For discrete time signals, difference equations form the equivalent of ODEs

$$g(y(t), y(t-1), \ldots, y(t-n), u(t), u(t-1), \ldots, u(t-n)) = 0$$

with linear time-invariant difference equations

$$\sum_{k=0}^{n} a_{n-k} y(t+k) = \sum_{k=0}^{n} b_{n-k} u(t+k) \tag{1.22}$$

corresponding to linear ODEs (1.15).

1.3.1 *LTI systems*

The developments for causal LTI discrete time systems parallels that of continuous time systems. For the difference equation (1.22), we can write the relation between the one-sided z-transforms of the input and output as

$$Y(z) = G(z)U(z) + Q(z)$$

where

$$G(z) = \frac{B(z)}{A(z)} = \frac{\sum_{k=0}^{n} b_k z^{n-k}}{\sum_{k=0}^{n} a_k z^{n-k}}$$

represents the system input-output behavior, and is known as the transfer function, and where[26]

[25] C.A. Desoer and M. Vidyasagar. *Feedback Systems: Input-Output Properties.* Academic Press, New York, 1975

[26] $B_Q(z) = \sum_{\tau=1}^{n-1} b_\tau^Q z^{n-\tau}$ where $b_\tau^Q = \sum_{l=1}^{\tau} (a_{\tau-l} y(1-l) - b_{\tau-l} u(1-l))$

$$Q(z) = \frac{B_Q(z)}{A(z)}$$

represents the transient behavior due to non-zero initial conditions.

The time-domain relationship is obtained from the inverse z-transform as

$$y(t) = \sum_{k=0}^{t-1} g(k)u(t-k) + q(t)$$

where the impulse pulse response g is the inverse z-transform of $G(z)$, while the transient $q(t)$ is the inverse z-transform of $Q(z)$. Stacking $y(1), \dots, y(N)$ into a vector, we can write

$$\mathbf{y} := \begin{bmatrix} y(1) \\ \vdots \\ y(N) \end{bmatrix} = T(\mathbf{u})\mathbf{g} +, \quad \mathbf{g} = \begin{bmatrix} g(0) \\ \vdots \\ g(N-1) \end{bmatrix}, = \begin{bmatrix} q(1) \\ \vdots \\ q(N) \end{bmatrix} \tag{1.23}$$

where $T(\mathbf{u})$ is a $N \times N$ lower Toeplitz matrix with $\mathbf{u} := \begin{bmatrix} u(1) & \dots & u(N) \end{bmatrix}^T$ as its first column. Due to the symmetry between g and u we can also write

$$\mathbf{y} = T(\mathbf{g})\boldsymbol{\Phi} \tag{1.24}$$

A state-space description for an LTI discrete time system is given by

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t), \quad x(0) = x_o \\ y(t) &= Cx(t) + Du(t) \end{aligned}$$

corresponding to

$$\begin{aligned} G(z) &= C(zI - A)^{-1}B + D \\ Q(z) &= C(zI - A)^{-1}x_o \end{aligned}$$

In continuous time we introduced the differentiation operator p. In discrete time the forward time shift operator q defined by $qy(t) = y(t+1)$ is convenient as we can express difference equations compactly, e.g. (1.22) can be written

$$A(q)y(t) = B(q)u(t), \quad A(q) = \sum_{k=0}^{n} a_k q^{n-k}, \; B(q) = \sum_{k=0}^{n} b_k q^{n-k} \tag{1.25}$$

Expressions like

$$y(t) = G(q)u(t)$$

where $G(q) = B(q)/A(q)$ should be interpreted as (1.25). Notice that $G(q) := G(z)|_{z=q}$, where $G(z)$ is the transfer function. We will therefore also call $G(q)$ the transfer function. A final notice on the use of the shift operator. One can equivalently express time shifts with the backward time shift operator q^{-1} defined by $q^{-1}y(t) = y(t-1)$. We will follow the convention in [27] and write difference equations such

[27] L. Ljung. *System identification, Theory for the user*. System sciences series. Prentice Hall, Upper Saddle River, NJ, USA, second edition, 1999

as (1.22) on the form

$$\sum_{k=0}^{n} a_k y(t-k) = \sum_{k=0}^{n} b_k u(t-k)$$

so that

$$\sum_{k=0}^{n} a_k q^{-k} y(t) = \sum_{k=0}^{n} b_k q^{-k} u(t)$$

which we write as $A(q)y(t) = B(q)u(t)$ with

$$A(q) = \sum_{k=0}^{n} a_k q^{-k}, \quad B(q) = \sum_{k=0}^{n} b_k q^{-k}$$

General LTI systems The equivalent description to the convolution formula (1.21) for a discrete time system G is that

$$y(t) = \sum_{\tau=-\infty}^{\infty} g(\tau)u(t-\tau) \tag{1.26}$$

Definition 1.3.1. *A discrete time LTI system with impulse response g is strictly stable if*

$$\sum_{\tau=-\infty}^{\infty} |\tau| \; \|g(\tau)\|_F < \infty$$

A BIBO-stable discrete time LTI system (1.26) responds to a sinusoid in a similar manner as a continuous time system

$$u(t) = \cos(\omega t) \quad \Rightarrow \quad y(t) = |G(i\omega)| \cos(\omega t + \arg G(i\omega))$$

where $G(i\omega)$ is the discrete time Fourier transform (1.3) of the impulse response g.

Lemma 1.3.1. *i) A discrete time LTI system is BIBO-stable if and only if its impulse response $g \in \ell_1$.*

ii) G defined by (1.26) with $g \in \ell_1$, is an operator $G : \ell_\infty \to \ell_\infty$ with $\|G\| \leq \|g\|_1$, with equality for the scalar case.

iii) G as in ii) is also an operator $G : L_2(\mathbb{R}^{n_u}) \to L_2(\mathbb{R}^{n_y})$ with $\|G\| = \sup_\omega \|G(e^{i\omega})\|_2$.

Proof. i) is Theorem 14 in Section 7.1, Chapter 4 in [28], which covers the time-varying case as well. ii) can be found in Table 4.2 in [29]. □

We can split up the impulse response into a causal part g_c and an anti-causal part g_a according to

$$g_c(t) = \begin{cases} g(t) & t = 0, 1, \ldots \\ 0 & otherwise \end{cases}, \quad \text{and} \quad g_a(t) = \begin{cases} 0 & t = 0, 1, \ldots \\ g(t) & otherwise \end{cases}$$

For a BIBO stable systems, i.e. when $g \in \ell_1$, which in turn implies that $G(e^{i\omega}) \in L_1(\mathbb{T})$,

$$G_c(z) = \sum_{t=-\infty}^{\infty} g_c(t)z^{-t} = \sum_{t=0}^{\infty} g(t)z^{-t}$$

$$G_a(z) = \sum_{t=-\infty}^{\infty} g_n(t)z^{-t} = \sum_{t=-\infty}^{-1} g(t)z^{-t}$$

[28] C.A. Desoer and M. Vidyasagar. *Feedback Systems: Input-Output Properties.* Academic Press, New York, 1975

[29] K. Zhou, J. C. Doyle, and K. Glover. *Robust and Optimal Control.* Prentice-Hall, 1996

both belong to $L_1(\mathbb{T})$, and hence $G_c \in H_1(\mathbb{T})$, whereas G_a in $H_1^{\perp}(\mathbb{T})$, defined as the space of functions in $L_1(\mathbb{T})$ that are analytic in $|z| < 1$. Furthermore, $G(e^{i\omega}) = G_c(e^{i\omega}) + G_a(e^{i\omega})$. We interpret $G(z) = G_c(z) + G_a(z)$ as the transfer function, notice that formally this function may only be defined on $|z| = 1$.

For a rational transfer function, the ROC is an annulus as already pointed out, and this annulus must include the unit circle for the system to be BIBO stable, cf. the discussion after Example 1.1. As for signals, the ROC for a transfer function must be known in order to compute the impulse response.

1.4 Exercises

1.1. Consider a system $y(t) = G(q)u(t)$ with transfer function

$$G(q) = \frac{q}{q - 0.5}$$

a) Determine the possible impulse responses for G

b) Express the system as difference equation forward in time, i.e. on the form

$$y(t) = -\sum_{k=1}^{n} a_k y(t-k) + \sum_{k=0}^{n} b_k u(t-k)$$

c) Express the system as difference equation backward in time, i.e. on the form

$$y(t) = -\sum_{k=1}^{n} a_k y(t+k) + \sum_{k=0}^{n} b_k u(t+k)$$

d) Which of the two recursions in b) and c) are stable? Relate this to the ROC of $G(z)$ and the possible impulse responses of the system. What conclusions can you draw in regards to which direction one should simulate a given difference equation?

e) Suppose that

$$G(q) = \frac{q^2}{(q - 0.5)(q - 3)}$$

is BIBO stable. Suppose that $y(0) = y(N+1) = u(0) = u(N+1) =$ and propose two different ways to simulate the system for a given input trajectory over the interval $t = 1, \ldots, N$.

2
Principles of Learning

2.1 Introduction

The Cambridge Dictionary defines inference as

a guess that you make or an opinion that you form based on the information that you have.

This concept is formalized in decision theory which is the theory for making (optimal) decisions under uncertainty. In statistical inference observations of the object of interest, data, constitute the available information upon which decisions are to be made. Forming a model based on data is often an intermediate step in the decision making. With the object of interest being a dynamical system, this is precisely the problem we are interested in.

2.2 The approximative nature of modeling

It is important to keep in mind that models are just approximations of the real world phenomena that one is trying to model. In Example 1.2 for example, the spring and damper components used in that model are aggregated idealizations of the corresponding physical devices. Better, but still not perfect models, would be based on non-linear partial differential equations embodying the internal interactions in each component. Such models turn out to have an infinite number of states and are thus much more complex than the model in Example 1.2. A further, and extreme, refinement would be to use a quantum dynamical model. How detailed the model should be depends on the intended use of the model. If the shock absorber is to be modelled during normal driving conditions the simple model in Example 1.2 may be sufficient, but as the driving conditions become more demanding, e.g. a high speed pursuit on a very bumpy road, the non-linear behaviour of the device will become prominent. In general one tries to make the model no more detailed than necessary as overly complex models tend to obscure physical insights and renders the use of the model more difficult. As an example, suppose that the damping can be controlled by an electronic actuator and that the model is to be used to design a controller for the lateral motion of the vehicle[1]. For a linear model with a finite number of states there is a wide range of control design methods available whereas the de-

[1] Such controllers are used on high performance cars and motorbikes, see, e.g.,

sign becomes much more complicated when the model is non-linear and/or has an infinite number of states.

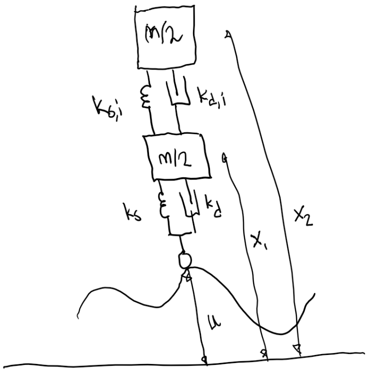

Figure 2.1: Extended model of shock absorber.

The fact that regardless of which model structure one uses, the true system cannot be captured perfectly is of fundamental importance in system identification. Without taking this into consideration, one may easily end up with a model that does not reflect the aspects of the system that one is interested in. To illustrate this let us again return to Example 1.2. As we already mentioned, the model in Figure 1.2 is an idealization. One of the simplifications that have been used is the lumping of the car chassis into a point mass. A more elaborate model is to consider the chassis as a flexible mechanical structure. In order to keep things (relatively) simple, let us consider the model in Figure 2.1 where the chassis is considered to consist of two masses (each having half of the total mass). To represent that the chassis is flexible, the two masses are interconnected by a spring and a damper. To model that the chassis is well described by a point mass at slow lateral motions, the spring should be very stiff and the damper offer some resistance to movements, i.e. the chassis parameter $k_{s,i}$ should be large[2].

Calculations based on the Laplace transform of (2.1) give that the transfer function from the input u to the position of the upper part of the chassis, i.e. x_2, is given by

$$G(s) = 4\,\frac{\left(\frac{k_d}{m}s + \frac{k_s}{m}\right)\left(\frac{k_{d,i}}{m}s + \frac{k_{s,i}}{m}\right)}{s^4 + \frac{4k_{d,i}+2k_d}{m}s^3 + \frac{4k_{s,i}+2k_s}{m}s^2 + 4\frac{k_s k_{d,i}+k_{s,i}k_d}{m^2}s + 4\frac{k_s k_{s,i}}{m^2}}$$

The Bode diagram of this transfer function is shown in Figure 2.2 for a $m = 1000$ kg heavy chassis with natural frequency $\omega_{o,i} = 11.2$ rad/s (1.8 Hz) and damping factor $\xi_i = 0.0045$. The shock absorber has natural frequency $\omega_o = 2.51$ rad/s and damping factor $\xi = 0.2$. The Bode diagram of the simplified model (1.14) is also shown.

Let us now make the *gedanken experiment* that the refined model in Figure 2.1 indeed represents the real system very accurately but that we *a priori do not know this*. Suppose now that we would like to

[2] Using Newton's 2nd law as in Example 1.2 gives

$$\begin{aligned} -k_{s,i}(x_2 - x_1) - k_{d,i}(\dot{x}_2 - \dot{x}_1) &= \frac{m}{2}\ddot{x}_2 \\ k_{s,i}(x_2 - x_1) + k_{d,i}(\dot{x}_2 - \dot{x}_1) & \\ -k_s(x_1 - u) - k_d(\dot{x}_1 - \dot{u}) &= \frac{m}{2}\ddot{x}_1 \end{aligned} \qquad (2.1)$$

Straightforward manipulations of the Laplace transform of the first equation in (2.1) give that the relationship between x_1 and x_2 is given by

$$X_2(s) = \frac{2\xi\omega_{o,i}s + \omega_{o,i}^2}{s^2 + 2\xi_i\omega_{o,i}s + \omega_{o,i}^2}X_1(s)$$

where the natural frequency is $\omega_{o,i} = \sqrt{2k_{s,i}/m}$ and the damping factor $\xi_i = \sqrt{k_{d,i}^2/(2k_{s,i}m)}$. This relationship has bandwidth approximately given by $\omega_{o,i}$. Thus if the variations of x_1 have a bandwidth well below $\omega_{o,i}$, x_2 will follow x_1 well. This means that we can replace $\ddot{x}_2$ in the first equation of (2.1) by $\ddot{x}_1$ and by substituting this in the second equation we obtain

$$-k_s(x_1 - u) - k_d(\dot{x}_1 - \dot{u}) = m\ddot{x}_1$$

identify a model for the absorber. To this end we carry out a test drive collecting measurements of the road height $u(t)$ and the car height $y(t)$. Suppose that the road profile is a sinusoid and that the driver, being very cautious, drives at low constant speed to avoid high speed lateral motions of the vehicle. The road height thus corresponds to a slowly varying sinusoid

$$u(t) = \sin(\omega t)$$

which for the assumed driving speed has frequency $\omega = \omega_1 := 2$ rad/s.

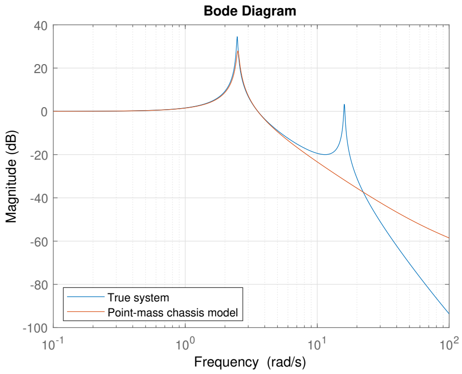

Figure 2.2: Bode diagrams of transfer function from road position (u) to position of upper part of the chassis (x_2). Included is also the simplified model (1.14) which for the given parameters capture the low frequency behavior including the main resonance peak.

This means that during the experiment the chassis can be very accurately approximated by a point mass and the model in Figure 1.2, with the same shock absorber parameters as those in the true system in Figure 2.1, can very accurately model the experimental data that were collected. Figure 2.3 shows the output of the model compared with the true system output.

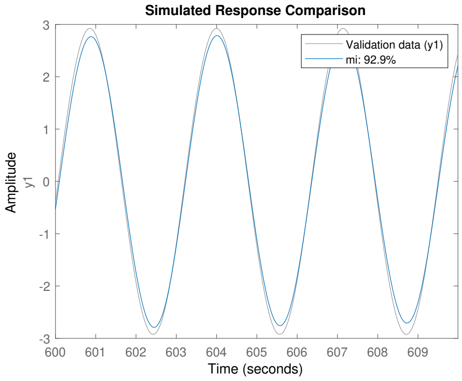

Figure 2.3: Comparison of the output of the model (1.14) and the true output.

The flexibility of the chassis will thus remain undetected, compare with the discrepancy between the Bode diagrams in Figure 2.2. While this is not necessarily a bad thing – the obtained model is indeed valid under cautious driving conditions – it points to a first tenet in system identification:

T1) *Identified models can only capture information available in the measurements*

This may seem a trivial observation but is nevertheless important to keep in mind when designing how experiments are to be carried out.

As a mental picture think in terms of the system's state trajectory, i.e. how the system state moves around in the state-space. There may be certain regions which the state never visits during the experiment. One can thus never learn about the system behaviour in these regions from the experiment. Compare with the states for the two masses of the chassis that moves in an almost identical fashion. One will only learn from data about the behavior in this direction.

Suppose now that a new experiment is conducted but now with a more aggressive driver behind the wheel who uses a much higher speed,[3] corresponding to the frequency $\omega_2 = 16.2$ rad/s of $u(t)$. With lateral motions being more rapid the flexibility of the chassis starts to become noticeable. However, it is still possible to find model parameters such that the point mass model in Figure 1.2 provides exactly the same car height as the real car for this experiment, see Figure 2.4.

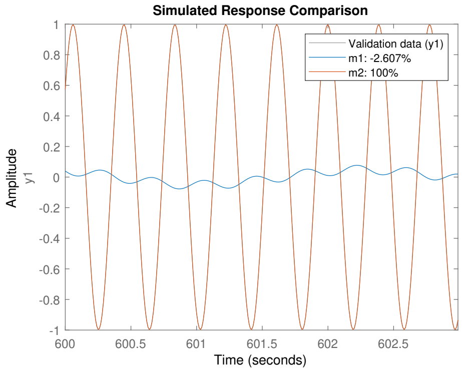

Figure 2.4: Comparison of the output of the model (1.14) and the true output for the more agressive driving condition corresponding to an input with frequency $\omega = \omega_2 := 16.2$ rad/s. With the original parameter settings ($m1$ in the figure) the model is very poor but $m2$ corresponds to another parameter setting which gives perfect fit.

This is again a manifestation of tenet T1. However, from an identification point, another very important phenomenon has occured. The new model parameters no longer correspond to the true values (i.e. those used in Figure 2.2). Thus, while the new simplified model describes the chassis movements under the more aggressive driving conditions, the model parameters have lost their physical interpretation as parameters of the shock absorber as now they are adopted to compensate for the effect of the flexibility of the chassis. Figure 2.5 shows that the new model has a Bode diagram that matches the true response at the frequency at which the system is excited but shows little resemblance to the true dynamics otherwise.

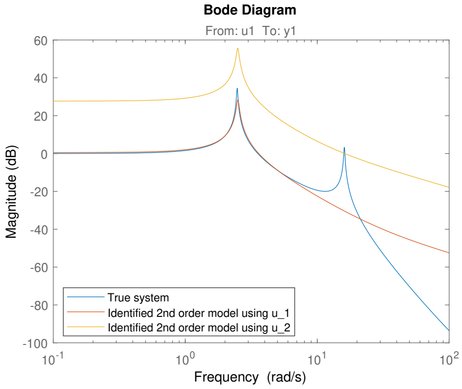

Figure 2.5: The new model parameters makes the model match the true response at $\omega = 16.2$, the frequency of the input.

[3] This is common practice in, e.g., the South African countryside to reduce the impact of pot-holes frequent at the roads.

We arrive at a second tenet in system identification:

T2) *The best model approximation depends on the experimental conditions.*

Generally speaking, this means that the experimental conditions should reflect those under which the model will be used.

However, there is yet one more lesson to be learned from this example. Suppose that yet another test drive is conducted but now the driver, while still driving fast, no longer drives at constant speed. It then turns out that there are no model parameters for which the simplified model can give a chassis height that exactly correspond to the true height. The behaviour of the system during this experiment is just too complex to be captured by the simplified model. The simplified model can thus be considered invalidated from this experiment[4]. This is illustrated in Figures 2.6 and Figures 2.7 which shows the results when the two sinusoids used above are used together to excite the system.

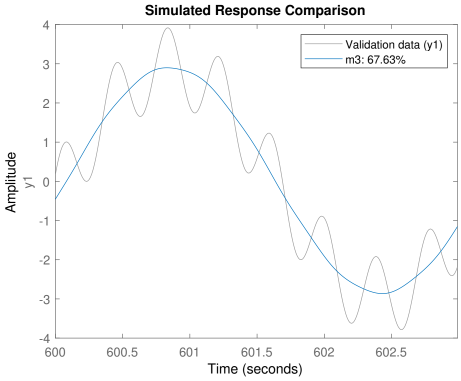

Figure 2.6: Simulated of best 2nd order model compared with true output when two sinusoids are used.

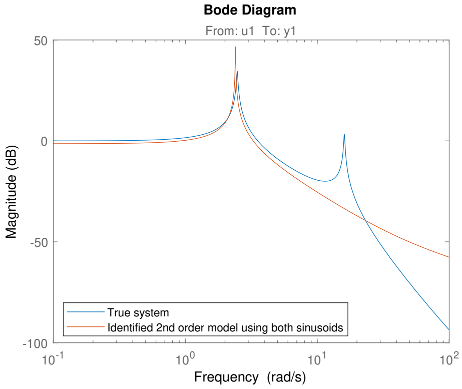

Figure 2.7: Bode diagram of the best model when the input is the sum of two sinusoids with frequencies $\omega_1 = 2$ and $\omega_2 = 16.2$.

We summarize this in a third tenet:

T3) *The capability of an experiment to invalidate a model depends on the "richness" of the input signal.*

While we will not at this point formally define what we mean by richness, we hope that the intuitive meaning is clear: Returning to our mental picture in the state-space, the input excitation must be such that parts of the state-space that are going to be used of the system in the application are visited during the experiment.

[4] Notice that an invalidated model may still be of use, cf. with the first model that gave a correct description at low speeds. This model would also be invalidated by the last experiment we discussed, however it is still valid under low speed conditions.

To summarize, we have seen that from a system identification perspective it is important to recognize that models are approximations of reality. This has important implications for how identification experiments should be designed. When we have developed suitable analysis tools we will return to a more formal treatment of this in Chapter **??**.

2.3 *Modeling disturbances and noise*

2.3.1 *Introduction*

In the previous section we saw that shortcomings of the dynamic model will lead to model errors. However, in practice it turns out that even if there exist a model that can recover the input/output behaviour of the true system, this model will never be recovered exactly using experimental data. One source of this problem is measuring errors. All signals from the device under test have to be captured by a measuring device - a sensor. These devices induce two sources of errors. Firstly, they typically have dynamics. For example a temperature sensor has a certain mass (even though small) that takes time to heat up or cool down to the ambient temperature. Thus the readings from this sensor will not be the instantaneous ambient temperature as desired. Secondly, sensors lack repeatability. This means that if exactly the same experiment is performed several times, the sensors will provide different readings.

In this book we will neglect the dynamics of the sensors. This can be done if their dynamics are known so that they can be compensated for in the final model. In practice this means that the sensors have to be calibrated before the experiment. This means that models for them have to be determined based on experimental data! Disregarding sensor dynamics is also motivated if the experiments are carried out such that these dynamics do not significantly perturb the measurements. To ensure this one needs to adapt the experiment to the technical specifications of the sensors.

A final source of error in an experiment is due to what is commonly called (unmeasurable) disturbances. These are external excitations that cannot be measured during the experiment, and therefore cannot be used as inputs. For the shock absorber in Example 1.2 one disturbance source is wind-gusts during the test-drive.

2.3.2 *Experimental set-ups*

A general experimental set-up of is shown in Figure 2.8. Using the sensor $\mathcal{S}$, measurements $\{z(kT)\}_{k=1}^{N}$, corresponding to the time-series in the diagrams, are collected from the system. From this data some decisions regarding the unknown system are to be deduced. Such decisions could simply be to deduce some properties of the system but could also be more involved, e.g. to deduce a feedback control policy where certain design objectives are satisfied. Aggravating the problem is that the sensor $\mathcal{S}$ may distort the system signals $x \in \mathbb{R}^{n_x}$, possibly in an (partially) unknown way, as well as be subject to measurement errors e, further corrupting the data. As nowadays all data are digitized, there is a sampling mechanism involved where, except for extracting signal samples, additional signal manipulations as well as noise corruption (v in the figure) occurs. In particular the measurements are quantized. The samples are usually collected with a fix sampling interval T but in industry it is not uncommon that this interval fluctuates over time. Different sampling intervals may also be used for different signals. A further complication is that the system may be subject to some disturbances w.

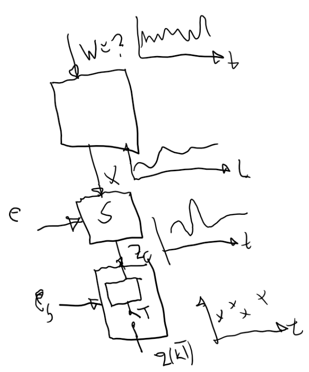

Figure 2.8: Learning consist of estimating the underlying mechanisms of a system from measurements $\{z(kT)\}_{k=1}^{N}$.

The causality of the system is sometimes known, i.e. it is known which signals "cause" the other signals. In this case the measurement $z(kT)$ can be split in two parts, the input $u(kT)$ representing samples of the cause, and the output $y(kT)$ representing the effect, see Figure 2.9.

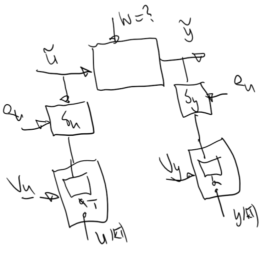

Figure 2.9: A causal system.

A common simplified setting is shown in Figure 2.10 where it is assumed that the input is known exactly whereas the output is subject to additive noise.

For a linear time-invariant system, using the linearity principle, the impact on the system of disturbances and measurement noise can be lumped to one point in the system, see the example in Figure

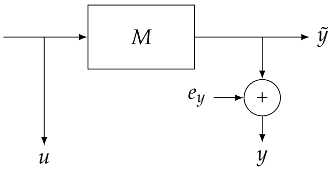

Figure 2.10: A simplified set-up for a causal system assuming that the input can be measured exactly.

2.11. We will therefore give a common treatment to these sources of errors, which we for simplicity will denote as noise.

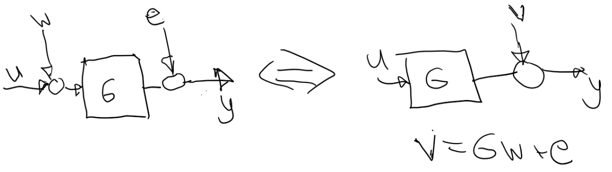

Figure 2.11: Disturbances and (measurement) noise can be merged in LTI models.

2.3.3 *First observations*

Adopting Figure 2.11 we have that the model provides a relation between measured inputs $u(t)$, the (unmeasured) noise $v(t)$ and the output $y(t)$:

$$y(t) = \mathrm{M}_{\boldsymbol{\theta}}[u](t) + v(t) \tag{2.2}$$

Here $\boldsymbol{\theta}$ indicates that the model depends on some unknown parameters $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$. Now, the noise $v(t)$ should be seen as part of the model, just as the model dynamics $\mathrm{M}_\theta[u]$. This means that for a given model structure, the modeling problem consists in determining both the model parameters θ and the noise $v(t)$. However, now a rather serious problem arises. Consider a given data set[5] $\mathcal{Z} = \{z(t)\}_{t=1}^{N}$, with $z(t) = \begin{bmatrix} y(tT) & u(tT) \end{bmatrix}^T$. Then for whatever model structure and model parameters $\boldsymbol{\theta}$ our model can perfectly reproduce the observed data by taking[6]

$$v(t) = v(t;\boldsymbol{\theta}) := y(t) - \mathrm{M}_{\boldsymbol{\theta}}[u](t), \quad t = 1, \ldots, N \tag{2.3}$$

So how can we discriminate between different models? In order to get a hint about this, let us return to the shock absorber in Example 1.2. Suppose that the error specifications of the lateral position sensor providing $y(t)$ is in the order of magnitude of mm, but that

[5] For simplicty we assume that uniform sampling, with sampling period 1.

[6] The particular noise signal $v(t;\theta)$ will be called the residuals, and represent the "left-overs" in the data that our model dynamics $\mathrm{M}_\theta[u]$ are not able to explain.

a particular model for this system gives noise terms $v(t;\theta)$ of the order of magnitude of meters. Clearly then either the sensor is faulty or that particular model has dynamics that is way off compared to the real system. This model can thus be discarded. In principle we can for each model assess whether the noise seems plausible or not, given what we know about the sensor characteristics and disturbances that act on the system. By this type of assessment, we can in principle sift all our candidate models and are left with those that appear plausible. Notice that when we are doing this we actually use our knowledge about the sensor characteristics and what is physically realistic. This points to a dilemma in system identification: Unless some a priori information is available, either about noise or the system dynamics, or both, no conclusions can be drawn from experimental data, no matter how long the data record. We arrive at another tenet.

T4) *A priori assumptions regarding noise and dynamics are necessary in order to discriminate among models.*

This puts noise and dynamics on an equal footing. Just as we restrict the class of models of the input-output dynamics by specifying a dynamic model[7], we need to restrict the behaviour of the noise. We will call these restrictions (or constraints) the noise model. As we will see it will be useful to also include dynamics in the noise model.

We now move on to provide the flavor of a few different types of noise models.

2.3.4 *Noise models*

We start with the following very simple example.

Example 2.1. *Suppose that the system is given by*

$$y(t) = \theta + v(t) \tag{2.4}$$

An experiment is conducted where two measurements $y(1) = 4$, $y(2) = -4$ are collected. The question is now: what can we say about the value of θ based on these observations?

As we have observed before we cannot say anything until we have specified a noise model. Assume therefore that it is known that $|v(t)| \le 6$, $t = 1, 2$. The two measurements then imply

$$\begin{aligned} 4 &= \theta + v(1), \ |v(1)| \le 6, \Leftrightarrow -2 \le \theta \le 10 \\ -4 &= \theta + v(2), \ |v(2)| \le 6, \Leftrightarrow -10 \le \theta \le 2 \end{aligned}$$

Taking two observations together we can conclude that

$$-2 \le \theta \le 2$$

Thus after the experiment we know for sure (provided our assumptions are correct) that θ belongs to the interval $[-2, 2]$. That is all we can say about θ.

In Figure 2.12 we illustrate geometrically what happens. The red striped

[7] Where some model parameters represent the degrees of freedom that are left.

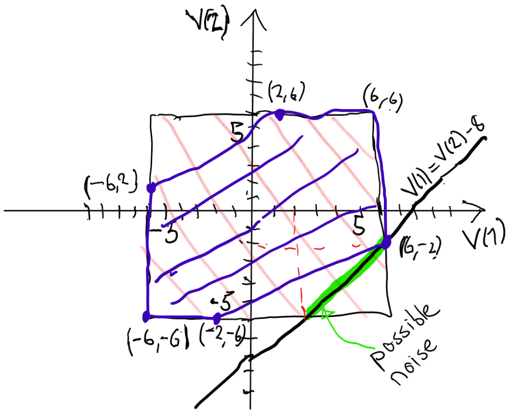

Figure 2.12: Red striped square indicates a priori set of possible noise terms. Solid black line the information the two measurements provide. The green segment is the set of noise terms consistent with both the prior and measurements in Example 2.1. Violet region indicates a priori set of possible noise terms in Example 2.3

square represents all possible disturbances, i.e. it is the region $|v(t)| \leq 6$, $t = 1,2$. Eliminating θ from (2.4)*, the set of noise pairs that are consistent with the observations $y(1), y(2)$ consists of the line*

$$v(2) = y(2) - y(1) + v(1)$$

This is the solid black line in the figure. The joint information of our prior knowledge about the noise and the measurements is the intersection of the red striped square with this line (the green segment in the figure). Thus after the experiment we know that the noise must be in the set

$$\{(v(1), v(2)) : \ 2 \leq v(1) \leq 6, \ v(2) = -8 + v(1)\} \tag{2.5}$$

Since $\theta = y(1) - v(1) = 4 - v(1)$ the set of θ consistent with these noise terms is the interval $[-2, 2]$.

One may suspect that the more prior information we have about the noise, the better we should be able to pin down the value of the unknown parameter θ after having taken our measurements. Let us illustrate this by returning to Example 2.1.

Example 2.2 (Example 2.1 continued)**.** *Suppose that in addition to $|v(t)| \leq 6$, we know that $v(2) = -v(1)$. The only point on the green segment in Figure 2.12 that satisfies this condition is $v(1) = 4$, $v(2) = -4$ corresponding to $\theta = 0$, i.e. now we get a perfect estimate of the parameter from the two observations.*

From the previous two examples we see that when constructing the noise model one should (as with the dynamical model) try to incorporate as much physical knowledge as possible. One ingredient that we will use is to incorporate dynamics in the noise models. Let us see how this could work.

Example 2.3 (Example 2.1 continued)**.** *Let us assume that the noise $v(t)$, $t = 1,2$, is generated by an underlying sequence $e(0), e(1), e(2)$ according to the dynamics*

$$v(t) = e(t) + 0.5e(t-1), \ t = 1,2; \quad |e(s)| \leq 4, \ s = 0,1,2 \tag{2.6}$$

Notice that then $|v(t)| \leq |e(t)| + 0.5|e(t-1)| \leq 4 + 0.5 \cdot 4 = 6$, *with equality when* $e(t) = e(t-1) = \pm 4$, *in accordance with our assumption in Example 2.1. However, the mechanism* (2.6) *is not able to generate all possible* $v(1), v(2)$ *in the red striped square in Figure 2.12. Some simple algebra gives that only those in the violet striped region in Figure 2.12 can be generated. Since this set is smaller than the red striped square in Figure 2.12, we thus have more prior knowledge about the noise than in Example 2.1.*

Let us now see if this information is useful to us. As in Example 2.1, the noise pairs $v(1), v(2)$ *consistent with the measurements* $y(1) = 4$, $y(2) = -4$ *(the same as in Example 2.1) is the solid line in Figure 2.12. Thus, the only noise pair consistent with the model* (2.4), *the noise model* (2.6) *and the observations is the single point* $v(1) = 6, v(2) = -2$ *which gives* $\theta = y(1) - v(1) = 4 - 6 = -2$ *as the only possible value of* θ. *The dynamic noise model has thus allowed us to reduce our uncertainty about* θ *from* $[-2, 2]$ *to complete certainty. All this, of course, predicated on the assumption that the dynamic noise model* (2.6) *correctly describes the noise.*

The noise model may take on many different formats depending on the context. From our examples above we arrive at another tenet:

T5) *An accurate noise model improves the estimate of the dynamic model.*

There are a number of generic noise models that have been developed and for which efficient estimation algorithms have been tailored. In this presentation we will stick to these, but the reader is advised to be careful if it is known that the noise has some very particular characteristics. Much can be gained by tailoring the noise model to these characteristics and the basics presented in this treatise are intended to guide you how to do this.

2.3.5 *Interaction between the dynamic model and the noise model*

The approach we will take in this book is, just as in Example 2.4, to model how the measured signals are built up by a dynamic model and disturbance and noise components, leading to an under-determined set of equations as in that example. This means that the dynamic model may try to capture disturbance/noise effects and vice-versa. There are two situations where this problem is accentuated. The first is when the dynamic model is unable to capture the true dynamics. In Section 2.2, the simplified model from Example 1.2 could not model the system in Figure 2.1 when the input was rich enough. This means that the estimated noise will contain dynamics from the true system. We will return to how to detect when this happens.

The second situation is when the model has more degrees of freedom than required to capture the true dynamics. Then there is a risk that the extra degrees of freedom will be used to model the disturbances and noise acting on the true system with the dynamic model. This is known as over-fitting and we will come back to how to avoid this from happening.

2.4 *Basic concepts*

2.4.1 *Models, model structures and the set of unfalsified models*

The learning problems we have seen examples of above all amount to solving an under-determined set of equations

$$\mathbf{z} = \mathbf{M}(\boldsymbol{\xi}) \tag{2.7}$$

where $\mathbf{z} \in \mathbb{R}^N$ are the measurements, where $\boldsymbol{\xi} \in \Xi \subseteq \mathbb{R}^{n_\xi}$ are the model parameters comprising noise, disturbance and input sequences as well as parameters in the dynamic model, and where $\mathbf{M}(\cdot)$ is a pre-specified function from its domain of definition Ξ to the image of Ξ in $\mathbb{R}^N$. With some abuse of language we will call $\boldsymbol{\xi}$ a *model*, Ξ a model set, and the function $\mathbf{M}(\cdot)$ a *model structure*.

An important feature is that the set of equations is under-determined, with more unknowns n_ξ than data N. We call the set of models $\boldsymbol{\xi}$ consistent with the observed data $\mathbf{z}$ the *set of unfalsified models*

$$\mathcal{U}(\mathbf{z}) := \{\boldsymbol{\xi} \in \Xi : \mathbf{M}(\boldsymbol{\xi}) = \mathbf{z}\} \tag{2.8}$$

Given that the model structure $\mathbf{M}(\cdot)$ is correct, $\mathcal{U}(\mathbf{z})$ is exactly the information contained in the data regarding the models $\boldsymbol{\xi}$. We will next turn to how select models from this set in a rational way.

2.4.2 *Identifiability and informative experiments*

Lack of observations may not be the only reason for why the model equation (2.7) is under-determined. The parametrization of $\mathbf{M}$ may also be inherently non-unique so that even if the number of free parameters are reduced below the number of observations $\mathbf{M}$ uniqueness does not hold. To formalize this let $\mathcal{S} \in \Xi$, be a subspace to $\mathbb{R}^{n_\xi}$. We then say that the model structure is *identifiable at the point* $\boldsymbol{\xi}^* \in \mathcal{S}$ in $\mathcal{S}$ if there is a $\tilde{\boldsymbol{\xi}} = \tilde{\boldsymbol{\xi}}(\boldsymbol{\xi}^*)$ in the orthogonal complement to $\mathcal{S}$ such that $\boldsymbol{\xi}^* + \tilde{\boldsymbol{\xi}} \in \Xi$ and

$$\mathbf{M}(\boldsymbol{\xi}^* + \tilde{\boldsymbol{\xi}}) = \mathbf{M}(\boldsymbol{\xi} + \tilde{\boldsymbol{\xi}}),\ \boldsymbol{\xi} \in \mathcal{S},\ \boldsymbol{\xi} + \tilde{\boldsymbol{\xi}} \in \Xi \ \Rightarrow \boldsymbol{\xi} = \boldsymbol{\xi}^*$$

The experiment generating $\mathbf{z} = \mathbf{M}(\boldsymbol{\xi}^* + \tilde{\boldsymbol{\xi}})$ is called *informative with respect to* $\boldsymbol{\xi}^* \in \mathcal{S}$.

The interpretation of identifiability at a point in $\mathcal{S}$ is that if we are given $\tilde{\boldsymbol{\xi}}$ together with the observation $\mathbf{z} = \mathbf{M}(\boldsymbol{\xi}^* + \tilde{\boldsymbol{\xi}})$ then $\boldsymbol{\xi}^* \in \mathcal{S}$ can be uniquely determined.

If all $\boldsymbol{\xi} \in \mathcal{S}$ are identifiable we say that *the model structure is identifiable in* $\mathcal{S}$. Notice that different experiments may be required for different $\boldsymbol{\xi} \in \mathcal{S}$. If the observation $\mathbf{z}$ can be used for all $\boldsymbol{\xi}\mathcal{S}$, we say that $\mathbf{z}$ is *informative with respect to* $\mathcal{S}$.

Example 2.4. *Consider the model*

$$z = \theta u + v$$

with $\boldsymbol{\xi} = \begin{bmatrix}\theta & u & v\end{bmatrix}^T$. *Let* $\mathcal{S}$ *be the subspace spanned by* $\begin{bmatrix}1 & 0 & 0\end{bmatrix}^T$. *Then taking* $\tilde{\boldsymbol{\xi}} = \begin{bmatrix}0 & \tilde{u} & \tilde{v}\end{bmatrix}^T$, *with* $\tilde{u} \neq 0$, *gives informative data with respect to* $\mathcal{S}$ *since*

$$\theta^* \tilde{u} + \tilde{v} = \theta\tilde{u} + \tilde{v}$$

implies that $\theta = \theta^*$. *Notice that* $\tilde{u} \neq 0$ *is required as otherwise the model is not identifiable.*

The model structure

$$z = \theta^2 u + v$$

is not identifiable in $\mathcal{S}$ *since z can only provide information about* θ^2, *the only point for which it is identifiable is* $\theta = 0$.

Identifiability is often important when the model parameters have physical interpretations and are used for decision making.

2.4.3 Model selection

Let us now look at how to pick a single model $\boldsymbol{\xi}$ from the set of unfalsified models. This means constructing a map from every set of unfalsified models $\mathcal{U}(\mathbf{z})$ to one model $\boldsymbol{\xi}$: $\boldsymbol{\xi}(\mathcal{U}(\mathbf{z})) \in \mathbb{R}^{n_\xi}$. Now, $\mathcal{U}(\mathbf{z})$, which is a set-valued function from $\mathbb{R}^N$ to sets in Ξ, is injective since two different observations cannot have the same parameters. We can thus simplify the indexation of the map by using $\mathbf{z}$: $\boldsymbol{\xi}(\mathbf{z})$. For the time being we call such a function a *model selection function*. Next we provide an example of such a function.

The Chebyshev center. If we without further information are asked to select a model in the set of unfalsified models it may seem "safe" to pick one in the midst of the set. There is a rich literature based on the Chebyshev center, defined as

$$\boldsymbol{\xi}_c(\mathbf{z}) := \arg\min_{\tilde{\boldsymbol{\xi}}} \max_{\boldsymbol{\xi}\in\mathcal{U}(\mathbf{z})} |\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi}| \tag{2.9}$$

However, we can of course come up with other "alibis" for constructing model selection functions. Let us see how we could take the intended model use into account.

Decision making under uncertainty. As we have point out above, often the model does not have any interest per se. Instead it is used in some application involving the true system. In process industry, a process model could be used to design a feedback controller for the system, and in telecommunications, a model of the fading could be used to design an equalizer to reduce the intersymbol interference. To formalize notions, let $\boldsymbol{\rho}$ denote the policy that is to be designed [8] which when applied to the true system, which we assume corresponds to the model $\boldsymbol{\xi}_o$, results in a "reward"[9] $R(\boldsymbol{\rho}, \boldsymbol{\xi}_o)$. Denoting the set of allowed policies by $\mathcal{F}$, the optimal policy is

$$\boldsymbol{\rho}^*(\boldsymbol{\xi}_o) := \arg\max_{\boldsymbol{\rho}\in\mathcal{F}} R(\boldsymbol{\rho}, \boldsymbol{\xi}_o)$$

[8] We do not want to enter into too much mathematical formalism at this point so the policy can be thought of as the device that is to be constructed, mathematically it could be a function $\boldsymbol{\rho} = \boldsymbol{\rho}(\cdot)$.

[9] The reward is a mathematical quantification of the achieved performance.

To align with the machine learning literature we introduce the regret for the policy $\boldsymbol{\rho}$ for the model $\boldsymbol{\xi}_o$

$$L(\boldsymbol{\rho},\boldsymbol{\xi}_o) = R(\boldsymbol{\rho}^*(\boldsymbol{\xi}_o),\boldsymbol{\xi}_o) - R(\boldsymbol{\rho},\boldsymbol{\xi}_o) \geq 0 \tag{2.10}$$

In principle, we could find the optimal policy by applying different policies to the true system and observing the resulting rewards through observations, finally selecting the best one. Practical implementations of this approach are often called *direct adaptive algorithms*. The term *direct* alluding to that the policy is designed directly from the data $\mathbf{z}$ without any use of an intermediate model. Such methods require that the system is available for experimentation which may not be possible due to economic-, time- or other constraints. When given an observation $\mathbf{z} \in \mathbb{R}^N$ from the system and a model structure, one may instead use a worst-case approach over the set of unfalsified models. The worst-case regret for a policy $\boldsymbol{\rho}$ when the true system is known to be in the set of unfalsified models is defined as

$$L(\boldsymbol{\rho},\mathcal{U}(\mathbf{z})) := \max_{\boldsymbol{\xi}_o \in \mathcal{U}(\mathbf{z})} L(\boldsymbol{\rho},\boldsymbol{\xi}_o)$$

The worst-case optimal policy is then given by

$$\boldsymbol{\rho}^*(L,\mathcal{U}(\mathbf{z})) := \arg\min_{\boldsymbol{\rho} \in \mathcal{F}} L(\boldsymbol{\rho},\mathcal{U}(\mathbf{z}))$$

It may be very challenging to solve this (robust) functional minimization problem. In model based design, the policy is a function of a model, i.e. we can index $\boldsymbol{\rho}$ as $\boldsymbol{\rho}(\boldsymbol{\xi})$. One could, e.g., choose the policy $\boldsymbol{\rho}(\boldsymbol{\xi}) = \boldsymbol{\rho}^*(\boldsymbol{\xi})$ such that zero regret is obtained if $\boldsymbol{\xi}_o = \boldsymbol{\xi}$, cf. optimal control. This policy is called the *certainty equivalence principle.* The worst-case optimal model is then defined as[10,11]

$$\boldsymbol{\xi}^*(L,\mathcal{U}(\mathbf{z})) := \arg\min_{\boldsymbol{\xi} \in \Xi} L(\boldsymbol{\rho}^*(\boldsymbol{\xi}),\mathcal{U}(\mathbf{z}))$$

This model thus results in a policy $\boldsymbol{\rho}(\boldsymbol{\xi}^*(L,\mathcal{U}(\mathbf{z})))$ that has smallest worst-case regret among the set of considered policies $\{\mathcal{F}_{\boldsymbol{\xi}} : \boldsymbol{\xi} \in \Xi\}$. Thus, by accounting for the intended use of the model, we have been able to select *one*[12] model from the set of unfalsified models in a rational way, i.e. we have constructed a meaningful model selection function.

Above, we can see $-L(\boldsymbol{\rho}(\boldsymbol{\xi}),\mathcal{U}(\mathbf{z}))$ as a function ranking the different models such that a larger value means a more preferrable model. Ranking has turned out to be a very useful concept so next we turn to some general considerations for how to rank models.

2.5 *Ranking models*

2.5.1 *Top ranked models*

At first glance, it may seem like a roundabout way to select a model by first ranking the different models $\boldsymbol{\xi} \in \Xi$ and then pick the one inside the set of unfalsified models with the largest ranking. However,

[10] Note that it may happen that the worst-case optimal model is not in $\mathcal{U}(\mathbf{z})$.

[11] A more elaborate scheme is obtained by also optimizing over ρ_{ξ}, i.e. we pick the best policy in a set $\mathcal{F}_{\xi}$ depending on the model ξ, according to

$$\rho^*(\xi,\mathcal{U}(\mathbf{z})) = \arg\min_{\rho \in \mathcal{F}_{\xi}} L(\rho,\mathcal{U}(\mathbf{z}))$$

The optimal model to use from a worst-case perspective would then be

$$\begin{aligned}&\xi(L,\mathcal{U}(\mathbf{z}))\\ &:= \arg\min_{\xi \in \Xi} L(\rho^*(\xi,\mathcal{U}(\mathbf{z})),\mathcal{U}(\mathbf{z}))\end{aligned}$$

[12] Of course there may be several global minima of $J(\rho_{\xi},\mathcal{U}(\mathbf{z}))$ but we are concerned with principles rather than details here.

we have already seen that the worst-case approach in the previous section can be interpreted this way and in Section 2.5.6 we will come back to the rationales. With[13]

$$p(\boldsymbol{\xi}) \quad (\geq 0) \tag{2.11}$$

denoting the ranking of model $\boldsymbol{\xi}$, the top ranked model would then be

$$\hat{\boldsymbol{\xi}}(\mathbf{z}) = \arg\max_{\boldsymbol{\xi}\in\mathcal{U}(\mathbf{z})} p(\boldsymbol{\xi}) \tag{2.12}$$

For practical reasons we can turn this into an unconstrained problem. As preparation we introduce a generalization of Dirac's delta function $\delta(x)$. Recall that loosely speaking this function is defined $\int f(t)\delta(t)dt = f(0)$. Now, for a vector $\mathbf{x} = \begin{bmatrix} x(1) & \dots & x(n) \end{bmatrix}^T \in \mathbb{R}^n$, we define

$$\delta(\mathbf{x}) := \prod_{k=1}^{n} \delta(x(k))$$

The joint ranking of model parameters $\boldsymbol{\xi}$ and observations $\mathbf{z}$ is now defined as

$$p(\boldsymbol{\xi}, \mathbf{z}) := p(\boldsymbol{\xi})\delta(\mathbf{z} - \mathbf{M}(\boldsymbol{\xi})), \tag{2.13}$$

and we can write

$$\hat{\boldsymbol{\xi}}(\mathbf{z}) = \arg\max_{\boldsymbol{\xi}} p(\boldsymbol{\xi}, \mathbf{z})$$

where we use the convention that $a\delta(0) < b\delta(0) \Leftrightarrow a < b$.

Let us see how this could work in the simple setting of Example 2.1.

Example 2.5. *Consider the model* (2.4). *Let us to begin with assume that we have only one measurement*[14] $\mathbf{z} \in \mathbb{R}$ *in which case we can write*

$$\mathbf{M}(\boldsymbol{\xi}) = \boldsymbol{\theta} + \mathbf{v}, \quad \boldsymbol{\xi} := \begin{bmatrix} \boldsymbol{\theta} & \mathbf{v} \end{bmatrix}^T$$

The set of unfalsified models is given by

$$\mathcal{U}(\mathbf{z}) = \{\begin{bmatrix} \boldsymbol{\theta} & \mathbf{v} \end{bmatrix}^T : \boldsymbol{\theta} \in \mathbb{R}, \mathbf{v} = \mathbf{z} - \boldsymbol{\theta})\}$$

One possible ranking is

$$p(\boldsymbol{\theta}, \mathbf{v}) = \mathcal{N}(\mathbf{v}; 0, \lambda), \quad \lambda = 0.1$$

i.e. we prefer models with small noise $\mathbf{v}$ *but put no preference over different* $\boldsymbol{\theta}$*'s is given. Let us stress that we use the probability density function (pdf) of the normal distribution only because it is a convenient positive function with well known properties - what we are engaged in does not have anything to do with probability theory.*

The function to maximize is

$$p(\boldsymbol{\theta}, \mathbf{v}, \mathbf{z}) = \mathcal{N}(\mathbf{v}; 0, \lambda)\delta(\mathbf{z} - \boldsymbol{\theta} - \mathbf{v})$$

which clearly is minimized by taking $\mathbf{v} = \hat{\mathbf{v}}(\mathbf{z}) := 0$ *and* $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(\mathbf{z}) := \mathbf{z}$. *The top ranked model is thus* $(\hat{\boldsymbol{\theta}}(\mathbf{z}), \hat{\mathbf{v}}(\mathbf{z})) = (\mathbf{z}, 0)$.

[13] There is no restriction to assume that the rankings are non-negative.

[14] We use the symbol $\mathbf{z}$ for the observations to be consistent with the notation in this section.

Building on the previous example, we consider a case with multiple measurements.

Example 2.6. *Suppose now that we have N measurements* $\mathbf{z} \in \mathbb{R}^N$ *and that we use the model*

$$\mathbf{M}(\boldsymbol{\xi}) = \mathbf{T}\boldsymbol{\theta} + \mathbf{v}, \quad \boldsymbol{\xi} := \begin{bmatrix} \boldsymbol{\theta}^T & \mathbf{v}^T \end{bmatrix}^T$$

where $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$, $n_\theta \leq N$, *and where* $\mathbf{T} \in \mathbb{R}^{N \times n_\theta}$, *leading to the set of unfalsified models*

$$\mathcal{U}(y) = \{\begin{bmatrix} \boldsymbol{\theta}^T & \mathbf{v}^T \end{bmatrix}^T : \boldsymbol{\theta} \in \mathbb{R}^{n_\theta}, \ \mathbf{v} = \mathbf{z} - \mathbf{T}\boldsymbol{\theta})\}$$

An extension of the ranking function used in the previous example could be

$$p(\boldsymbol{\theta}, \mathbf{v}) = \mathcal{N}(\mathbf{v}; 0, \lambda \mathbf{I}), \quad \lambda = 0.1$$

so that we should maximize

$$p(\boldsymbol{\theta}, \mathbf{v}, \mathbf{z}) = \mathcal{N}(\mathbf{v}; 0, \lambda \mathbf{I}) \delta(\mathbf{z} - \mathbf{T}\boldsymbol{\theta} - \mathbf{v}) \tag{2.14}$$

Eliminating $\mathbf{v}$, *which has to be* $\mathbf{v} = \mathbf{z} - \mathbf{T}\boldsymbol{\theta}$, *gives*

$$\begin{aligned}\hat{\boldsymbol{\theta}}(\mathbf{z}) &= \arg\max_{\boldsymbol{\theta}} \mathcal{N}(\mathbf{z} - \mathbf{T}\boldsymbol{\theta}; 0, \lambda \mathbf{I}) = \arg\min_{\theta} \frac{1}{\lambda} |\mathbf{z} - \mathbf{T}\boldsymbol{\theta}|^2 + N \log \lambda \\ &= \arg\min_{\theta} |\mathbf{z} - \mathbf{T}\boldsymbol{\theta}|^2 \end{aligned} \tag{2.15}$$

Thus $\hat{\boldsymbol{\theta}}(\mathbf{z})$ *is obtained as the solution to the above least-squares problem. When* $\mathbf{T}$ *has full (column) rank the solution is*

$$\hat{\boldsymbol{\theta}}(\mathbf{z}) = (\mathbf{T}^T \mathbf{T})^{-1} \mathbf{T}^T \mathbf{z}, \tag{2.16}$$

and $\hat{\mathbf{v}}(\mathbf{z}) = \mathbf{z} - \hat{\boldsymbol{\theta}}(\mathbf{z})$. *These choices of* $\boldsymbol{\theta}$ *and* $\mathbf{v}$ *correspond to the top ranked model.*

2.5.2 *Sets of top ranked models*

Obviously it may be difficult to come up with a suitable ranking function so it may appear risky to just pick the model that has the highest ranking in the set of unfalsified models. Instead, one may opt to select a subset $\mathcal{U}_{0.95}(\mathbf{z}) \subset \mathcal{U}(\mathbf{z})$ corresponding to the, say 95%, highest ranked models, i.e. $\mathcal{U}_{0.95}(\mathbf{z})$ is such that

$$\frac{\int_{\mathcal{U}_{0.95}(\mathbf{z})} p(\boldsymbol{\xi}, \mathbf{z}) d\boldsymbol{\xi}}{\int_{\Xi} p(\boldsymbol{\xi}, \mathbf{z}) d\boldsymbol{\xi}} = 0.95$$

where all models in $\mathcal{U}(\mathbf{z})$ that do not belong to $\mathcal{U}_{0.95}(\mathbf{z})$ have lower ranking than those in $\mathcal{U}_{0.95}(\mathbf{z})$.

A slightly simpler expression can be obtained if we introduce the *total rankings for* $\mathbf{z}$

$$p(\mathbf{z}) := \int_{\Xi} p(\boldsymbol{\xi}, \mathbf{z}) d\boldsymbol{\xi} \tag{2.17}$$

and the normalized ranking of ξ given $\mathbf{z}$

$$p(\xi|\mathbf{z}) := \frac{p(\xi,\mathbf{z})}{p(\mathbf{z})} \tag{2.18}$$

which satisfies

$$\int p(\xi|\mathbf{z})d\xi = 1 \tag{2.19}$$

Then $\mathcal{U}_{0.95}(\mathbf{z})$ should satisfy

$$\begin{aligned}&\int_{\mathcal{U}_{0.95}(\mathbf{z})} p(\xi|\mathbf{z})d\xi = 0.95\\ &\bar{\xi} \in \mathcal{U}(\mathbf{z}) \setminus \mathcal{U}_{0.95}(\mathbf{z}),\ \xi \in \mathcal{U}_{0.95}(\mathbf{z}) \ \Rightarrow\ p(\bar{\xi}) \le p(\xi)\end{aligned}$$

Let us see how this could work in the previous example.

Example 2.7 (Example 2.6 continued.)**.** *The ranking function is given by* (2.14)

$$p(\boldsymbol{\theta},\mathbf{v},\mathbf{z}) = \mathcal{N}(\mathbf{v};0,\lambda\mathbf{I})\delta(\mathbf{z}-\mathbf{T}\boldsymbol{\theta}-\mathbf{v}), \quad \lambda = 0.1$$

and all models in $\mathcal{U}(\mathbf{z})$ can be parametrized as

$$\xi = \begin{bmatrix}\boldsymbol{\theta}\\ \mathbf{z}-\mathbf{T}\boldsymbol{\theta}\end{bmatrix}$$

Any subset in $\mathcal{U}(\mathbf{z})$ can thus be written as

$$\mathcal{U}_{\boldsymbol{\Theta}}(\mathbf{z}) := \left\{\begin{bmatrix}\boldsymbol{\theta}\\ \mathbf{z}-\mathbf{T}\boldsymbol{\theta}\end{bmatrix} :\ \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^n\right\}$$

The problem is thus to determine $\boldsymbol{\Theta}$. In order to have 95% of the rankings we should select $\boldsymbol{\Theta}$ such that

$$0.95 = \frac{\int_{\mathcal{U}_{\boldsymbol{\Theta}}(\mathbf{z})} p(\boldsymbol{\theta},\mathbf{v},\mathbf{z})d\mathbf{v}d\boldsymbol{\theta}}{p(\mathbf{z})} = \frac{\int_{\boldsymbol{\Theta}} \mathcal{N}(\mathbf{z}-\mathbf{T}\boldsymbol{\theta};0,\lambda\mathbf{I})d\boldsymbol{\theta}}{\int \mathcal{N}(\mathbf{z}-\mathbf{T}\boldsymbol{\theta};0,\lambda\mathbf{I})d\boldsymbol{\theta}} \tag{2.20}$$

Now, a model in $\mathcal{U}_{\boldsymbol{\Theta}}(\mathbf{z})$ has weighting

$$\mathcal{N}(\mathbf{z}-\mathbf{T}\boldsymbol{\theta};0,\lambda\mathbf{I}) = \frac{1}{(2\pi)^{N/2}\lambda^{N/2}}e^{-\frac{1}{2\lambda}|\mathbf{z}-\mathbf{T}\boldsymbol{\theta}|^2}$$

In order to obtain the highest ranked models we should thus select $\boldsymbol{\Theta}$ as those $\boldsymbol{\theta}$ for which $\mathbf{T}\boldsymbol{\theta}$ is closest to $\mathbf{z}$. From (A.2) *in Lemma A.5.1 we obtain that*

$$\frac{|\mathbf{z}-\mathbf{T}\boldsymbol{\theta}|^2}{\lambda} = (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}(\mathbf{z}))^T\frac{\mathbf{T}^T\mathbf{T}}{\lambda}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}(\mathbf{z})) + \frac{|\mathbf{z}-\mathbf{T}\hat{\boldsymbol{\theta}}(\mathbf{z})|^2}{\lambda} \tag{2.21}$$

where $\hat{\boldsymbol{\theta}}(\mathbf{z})$ is the solution to (2.15) *which is given by* (2.16)*. The second term on the right is independent of $\boldsymbol{\theta}$ whereas the first term has level curves that are ellipsoids. All models on such a level curve have the same ranking. We should thus take*

$$\boldsymbol{\Theta} = \boldsymbol{\Theta}(\mathbf{z}) := \{\boldsymbol{\theta} :\ (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}(\mathbf{z}))^T\frac{\mathbf{T}^T\mathbf{T}}{\lambda}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}(\mathbf{z})) \le c\} \tag{2.22}$$

for some constant c which should be adjusted such that (2.20) *is satisfied. This condition can be expressed as*[15]

$$\int_{\Theta} \frac{e^{-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}(\mathbf{z}))^T \frac{\mathbf{T}^T\mathbf{T}}{\lambda}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}(\mathbf{z}))}}{(2\pi)^{n/2}\sqrt{\det \lambda(\mathbf{T}^T\mathbf{T})^{-1}}} d\boldsymbol{\theta} = 0.95 \tag{2.25}$$

We recognize that this is the probability that $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *assuming that* $\boldsymbol{\theta} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}(\mathbf{z}), \lambda(\mathbf{T}^T\mathbf{T})^{-1})$. *For the set* (2.22)*, this is the same probability as that a* $\chi^2(n_\theta)$ *distributed variable is less than c . We thus arrive at the conclusion that the set of* 95% *top-ranked models is given by*

$$\mathcal{U}_{0.95}(\mathbf{z}) = \left\{ \begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{z} - \mathbf{T}\boldsymbol{\theta} \end{bmatrix} : \; (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\mathbf{z}))^T \frac{\mathbf{T}^T\mathbf{T}}{\lambda} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\mathbf{z})) \le F^{-1}_{\chi^2(n_\theta)}(0.95) \right\} \tag{2.26}$$

where F_D^{-1} *denotes the inverse of the distribution function of the distribution D.*

It is also worth pointing out that, with the notation $\mathbf{v}(\boldsymbol{\theta}) = \mathbf{z} - T\boldsymbol{\theta}$ *and using* $T\boldsymbol{\theta} - T\hat{\boldsymbol{\theta}} = \mathbf{z} - T\hat{\boldsymbol{\theta}} - (\mathbf{z} - T\boldsymbol{\theta}) = \mathbf{v}(\hat{\boldsymbol{\theta}}) - \mathbf{v}(\boldsymbol{\theta})$*, we can express the set of top-ranked models in terms of* $v(\boldsymbol{\theta})$ *as*

$$\mathcal{U}_{0.95}(\mathbf{z}) = \left\{ \begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{z} - \mathbf{T}\boldsymbol{\theta} \end{bmatrix} : \; \frac{\frac{1}{N}|\mathbf{v}(\boldsymbol{\theta}) - \mathbf{v}(\hat{\boldsymbol{\theta}})|^2}{\lambda} \le \frac{F^{-1}_{\chi^2(n_\theta)}(0.95)}{N} \right\}$$

We see that $\frac{1}{N}|\mathbf{v}(\boldsymbol{\theta}) - \mathbf{v}(\hat{\boldsymbol{\theta}})|^2$ *should not deviate too much from* λ *which represents the width of the ranking function.*

Another characterization of the set of top ranked models is obtained by using (A.3) *in* (2.26)

$$\mathcal{U}_{0.95}(\mathbf{z}) = \left\{ \begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{z} - \mathbf{T}\boldsymbol{\theta} \end{bmatrix} : \; |\mathbf{z} - \mathbf{T}\boldsymbol{\theta}|^2 \le |\mathbf{z} - \mathbf{T}\hat{\boldsymbol{\theta}}|^2 + \lambda F^{-1}_{\chi^2(n_\theta)}(0.95) \right\} \tag{2.27}$$

Thus the top ranked models are characterized by that they fit $\mathbf{T}\boldsymbol{\theta}$ *to the observation* $\mathbf{z}$ *to within a margin of the least-squares estimate* $T\hat{\boldsymbol{\theta}}$ *given by the second term on the right-hand side of the inequality in* (2.27).

Now, the sample correlations between the columns of $\mathbf{T}$ *and* $\mathbf{v}(\boldsymbol{\theta})$ *are obtained by projecting* $\mathbf{v}(\boldsymbol{\theta})$ *onto the columns of* $\mathbf{T}$ *and then normalizing with the norm of* $\mathbf{v}(\boldsymbol{\theta})$

$$\frac{\mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{v}(\boldsymbol{\theta})}{|\mathbf{v}(\boldsymbol{\theta})|}$$

We can re-write the cross-correlations as

$$N\,\frac{\mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}C_{\mathbf{T},\mathbf{v}(\boldsymbol{\theta})}}{|\mathbf{v}(\boldsymbol{\theta})|}$$

by introducing the corresponding vector of sample cross-covariances

$$C_{\mathbf{T},\mathbf{v}(\boldsymbol{\theta})} := \frac{1}{N}\mathbf{T}^T\mathbf{v}(\boldsymbol{\theta})$$

The cross-correlations are in the interval $[-1, 1]$ *with large absolute values when there are strong correlations. Thus taking the sum of the squares of the correlations will be a number between* 0 *and* n_θ

$$0 \le N^2\,\frac{C^T_{\mathbf{T},\mathbf{v}(\boldsymbol{\theta})}(\mathbf{T}^T\mathbf{T})^{-1}C_{\mathbf{T},\mathbf{v}(\boldsymbol{\theta})}}{|\mathbf{v}(\boldsymbol{\theta})|^2} = \frac{C^T_{\mathbf{T},\mathbf{v}(\boldsymbol{\theta})}\left(\frac{\mathbf{T}^T\mathbf{T}}{N}\right)^{-1}C_{\mathbf{T},\mathbf{v}(\boldsymbol{\theta})}}{|\mathbf{v}(\boldsymbol{\theta})|^2/N} \le n_\theta$$

[15] We have that

$$\begin{aligned} p(\mathbf{z}) &= \int \frac{e^{-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}(\mathbf{z}))^T \frac{\mathbf{T}^T\mathbf{T}}{\lambda}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}(\mathbf{z})) + \frac{|\mathbf{z}-\mathbf{T}\hat{\theta}|^2}{\lambda}}}{(2\pi)^{N/2}\lambda^{N/2}} d\boldsymbol{\theta} \\ &= \frac{e^{-\frac{1}{2}\frac{|\mathbf{z}-\mathbf{T}\hat{\boldsymbol{\theta}}(\mathbf{z})|^2}{\lambda}}}{(2\pi)^{N/2}\lambda^{N/2}} \int e^{-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}(\mathbf{z}))^T \frac{\mathbf{T}^T\mathbf{T}}{\lambda}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}(\mathbf{z}))} d\boldsymbol{\theta} \\ &= \frac{e^{-\frac{1}{2}\frac{|\mathbf{z}-\mathbf{T}\boldsymbol{\theta}|^2}{\lambda}}}{(2\pi)^{(N-n_\theta)/2}\lambda^{(N-n_\theta)/2}\sqrt{\det \mathbf{T}^T\mathbf{T}}} \\ &\quad \times \int \frac{e^{-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}(\mathbf{z}))^T \frac{\mathbf{T}^T\mathbf{T}}{\lambda}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}(\mathbf{z}))}}{(2\pi)^{n_\theta/2}\sqrt{\det \lambda(\mathbf{T}^T\mathbf{T})^{-1}}} d\boldsymbol{\theta} \\ &= \frac{e^{-\frac{1}{2}\frac{|\mathbf{z}-\mathbf{T}\hat{\boldsymbol{\theta}}(\mathbf{z})|^2}{\lambda}}}{(2\pi)^{(N-n_\theta)/2}\lambda^{(N-n_\theta)/2}\sqrt{\det \mathbf{T}^T\mathbf{T}}} \end{aligned} \tag{2.23}$$

$$\begin{aligned} &\int_{\mathcal{U}_{\boldsymbol{\Theta}}(\mathbf{z})} p(\boldsymbol{\theta}, \mathbf{v}, \mathbf{z}) d\boldsymbol{\theta} \\ &= \frac{e^{-\frac{1}{2}\frac{|\mathbf{z}-\mathbf{T}\hat{\boldsymbol{\theta}}(\mathbf{z})|^2}{\lambda}}}{(2\pi)^{(N-n_\theta)/2}\lambda^{(N-n_\theta)/2}\sqrt{\det \mathbf{T}^T\mathbf{T}}} \\ &\quad \times \int_{\mathcal{U}_{\boldsymbol{\Theta}}(\mathbf{z})} \frac{e^{-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}(\mathbf{z}))^T \frac{\mathbf{T}^T\mathbf{T}}{\lambda}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}(\mathbf{z}))}}{(2\pi)^{n_\theta/2}\sqrt{\det \lambda(\mathbf{T}^T\mathbf{T})^{-1}}} d\boldsymbol{\theta} \end{aligned} \tag{2.24}$$

Now we notice that[16]

$$NC_{\mathbf{T},\mathbf{v}(\boldsymbol{\theta})} = \mathbf{T}^T(\mathbf{z}-\mathbf{T}\boldsymbol{\theta}) = \mathbf{T}^T(\mathbf{z}-\mathbf{T}\hat{\boldsymbol{\theta}}(\mathbf{z})+\mathbf{T}\hat{\boldsymbol{\theta}}(\mathbf{z})-\mathbf{T}\boldsymbol{\theta}) = \mathbf{T}^T\mathbf{T}(\hat{\boldsymbol{\theta}}(\mathbf{z})-\boldsymbol{\theta})$$

so that

$$(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}(\mathbf{z}))^T\mathbf{T}^T\mathbf{T}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}(\mathbf{z})) = N^2C^T_{\mathbf{T},\mathbf{v}(\boldsymbol{\theta})}\left(\mathbf{T}^T\mathbf{T}\right)^{-1}C_{\mathbf{T},\mathbf{v}(\boldsymbol{\theta})}$$

and, hence, (2.26) *can be written*

$$\mathcal{U}_{0.95}(\mathbf{z}) = \left\{\begin{bmatrix}\boldsymbol{\theta}\\ \mathbf{z}-\mathbf{T}\boldsymbol{\theta}\end{bmatrix} : \frac{C^T_{\mathbf{T},\mathbf{v}(\boldsymbol{\theta})}\left(\frac{\mathbf{T}^T\mathbf{T}}{N}\right)^{-1}C_{\mathbf{T},\mathbf{v}(\boldsymbol{\theta})}}{|\mathbf{v}(\boldsymbol{\theta})|^2/N} \leq \frac{1}{N}\,\frac{\lambda}{|\mathbf{v}(\boldsymbol{\theta})|^2/N}\,F^{-1}_{\chi^2(n_\theta)}(0.95)\right\} \tag{2.28}$$

Thus the set of top ranked consists of models for which the cross-correlations between the noise and the columns of $\mathbf{T}$ *are sufficiently small. We notice that the condition becomes more severe for models for which the ratio* $\frac{|\mathbf{v}(\boldsymbol{\theta})|^2/N}{\lambda}$ *is large, i.e. if the norm of the noise is larger than what we expect for a high ranked model*[17]*, there has to be compelling evidence in terms of that little of the noise can be modelled by* $\mathbf{T}\boldsymbol{\theta}$*.*

2.5.3 *Tuning the ranking function*

The ranking function should encode our prior beliefs about how the system behaves. Doing this is a non-trivial task and in Section 2.5.6 we shall see how we can map our ranking philosophy to a constructive approach.

However for now an important observation is that given a specific observation $\mathbf{z}$ we only need to provide a ranking for the models in $\mathcal{U}(\mathbf{z})$, there is no need to waste effort on ranking other models. This can greatly simplify our task. One way to go about doing this is to start with a parametrized ranking function $p(\boldsymbol{\xi};\boldsymbol{\eta})$. These parameters, here denoted $\boldsymbol{\eta}$, will be called *hyperparameters* as they do not directly contribute to the model $\mathbf{M}(\boldsymbol{\xi})$, instead they control which model is selected in the set of unfalsified models, e.g. we will have that the top ranked model depends on the hyperparameters: $\hat{\boldsymbol{\xi}} = \hat{\boldsymbol{\xi}}(\boldsymbol{\eta})$.

So how should we select these new parameters? Well, an idea that quickly comes to mind is that one should pick the hyperparameters such that the models in the set of unfalsified models are highly ranked. After all, these are the models that are consistent with the data. There are, of course, different ways to measure if these models are highly ranked. We could for example maximize the ranking of the top ranked model $\hat{\boldsymbol{\xi}}(\boldsymbol{\eta})$

$$\hat{\boldsymbol{\eta}}(\mathbf{z}) := \arg\max_{\boldsymbol{\eta}} p(\hat{\boldsymbol{\xi}}(\boldsymbol{\eta}),\mathbf{z};\boldsymbol{\eta}) \tag{2.29}$$

which is the same as solving

$$\left(\hat{\boldsymbol{\xi}}(\mathbf{z}),\hat{\boldsymbol{\eta}}(\mathbf{z})\right) := \arg\max_{\boldsymbol{\xi}\in\Xi,\,\boldsymbol{\eta}} p(\boldsymbol{\xi},\mathbf{z};\boldsymbol{\eta}) \tag{2.30}$$

[16] Why?

[17] λ specifies the width of the ranking function. Recall that we have specified that $\lambda = 0.1$.

An alternative could be to maximize the total rankings[18]

$$\hat{\boldsymbol{\eta}}(\mathbf{z}) := \arg\max_{\boldsymbol{\eta}} p(\mathbf{z};\boldsymbol{\eta}) \tag{2.31}$$

which, if we view $p(\mathbf{z};\boldsymbol{\eta})$ as a ranking of the hyperparameters, we also can interpret as picking the top ranked hyperparameter.

Let us see what these two approaches give in Example 2.6.

Example 2.8 (Example 2.6 continued)**.** *A parameter that does not influence* $\mathbf{z}(\boldsymbol{\theta},\mathbf{v}) = \mathbf{T}\boldsymbol{\theta}+\mathbf{v}$ *is* λ *that appears in the ranking function* $p(\boldsymbol{\theta},\mathbf{v}) = \mathcal{N}(\mathbf{v};0,\lambda\mathbf{I})$*. Thus* λ *is a hyperparameter. We first observe that the top ranked model* (2.16) *does not depend on this hyperparameter. In this example,* (2.29) *corresponds to maximizing* (2.15) *first with respect to* $\boldsymbol{\theta}$ *and then to* λ*, or as the problem is well-behaved to simultaneously maximize the function on the right in* (2.15) *with respect to both* $\boldsymbol{\theta}$ *and* λ*. Now, the minimum with respect to* $\boldsymbol{\theta}$ *is independent of* λ *and given by the least-squares solution* (2.16)*. This means that* (2.29) *becomes*

$$\hat{\lambda}(\mathbf{z}) = \arg\min_{\lambda} \frac{|\mathbf{z}-\mathbf{T}\hat{\boldsymbol{\theta}}(\mathbf{z})|^2}{\lambda} + N\log\lambda$$

Setting the derivative of the objective function gives

$$-\frac{|\mathbf{z}-\mathbf{T}\hat{\boldsymbol{\theta}}(\mathbf{z})|^2}{\lambda^2} + \frac{N}{\lambda} = 0$$

which gives the solution

$$\hat{\lambda}(\mathbf{z}) = \frac{1}{N}|\mathbf{z}-\mathbf{T}\hat{\boldsymbol{\theta}}(\mathbf{z})|^2$$

i.e the average of the minimum possible squared errors.

For (2.31) *we can take the logarithm of* (2.23) *and eliminate* λ*-independent terms, giving*

$$\hat{\lambda}(\mathbf{z}) = \arg\min_{\lambda} \frac{|\mathbf{z}-\mathbf{T}\hat{\boldsymbol{\theta}}(\mathbf{z})|^2}{\lambda} + (N-n_\theta)\log\lambda$$

which has solution

$$\hat{\lambda}(\mathbf{z}) = \frac{1}{N-n_\theta}|\mathbf{z}-\mathbf{T}\hat{\boldsymbol{\theta}}(\mathbf{z})|^2$$

We see that there is a slight difference compared to the previous estimate in that the normalization is $1/(N-n_\theta)$ *rather than* $1/N$*.*

We can also determine the hyperparameters as their conditional average given the observation $\mathbf{z}$

$$\bar{\boldsymbol{\eta}}(\mathbf{z}) = \int \boldsymbol{\eta}\, p(\boldsymbol{\eta}|\mathbf{z})d\boldsymbol{\eta} \tag{2.32}$$

where, similar to (2.19), $p(\boldsymbol{\eta}|\mathbf{z}) = p(\boldsymbol{\eta};\mathbf{z})/p(\mathbf{z})$ with

$$p(\mathbf{z}) = \int p(\mathbf{z};\boldsymbol{\eta})d\boldsymbol{\eta}, \tag{2.33}$$

see Exercise 2.5.

[18] Recall (2.23) , i.e. that $p(\mathbf{z};\boldsymbol{\eta})$ is the total of the rankings in the set of unfalsified models.

By the above it should be clear that selecting hyperparameters using data can be done using the same principles as for model parameters. We can also parametrize p as $p(\boldsymbol{\xi};\boldsymbol{\eta})p(\boldsymbol{\eta})$, i.e. the ranking of the model parameters is parametrized as before but a separate ranking is provided to the hyperparameters. We may then parametrize the ranking function for the hyperparameters by new parameters, which in turn may be given a separate ranking. This procedure may be continued in as many steps as desired leading to the hierarchically structured ranking function

$$p(\boldsymbol{\xi};\boldsymbol{\eta}_0)\prod_{i=1}^{m} p(\boldsymbol{\eta}_{i-1};\boldsymbol{\eta}_i)$$

Before moving on, we remark that through the use of hyperparameters the same ranking function can be achieved in different ways. We illustrate this in an example.

Example 2.9 (Example 2.5 continued)**.** *In Example 2.5 we used the ranking function $p(\boldsymbol{\theta},\mathbf{v}) = \mathcal{N}(\mathbf{v};0,0.1)$, i.e. we did not include $\boldsymbol{\theta}$ in the ranking function. The same can be achieved with the ranking function $p(\boldsymbol{\theta},v,\eta) = \mathcal{N}(\mathbf{v};0,0.1)\delta(\boldsymbol{\theta}-\boldsymbol{\eta})$, where here $\boldsymbol{\eta}$ is a hyperparameter to be determined from data. This ranking function forces $\boldsymbol{\theta} = \boldsymbol{\eta}$ and thus the set of unfalsified models is defined by $\mathbf{z} = \boldsymbol{\theta}+\mathbf{v} = \boldsymbol{\eta}+\mathbf{v}$, and on this set $p(\boldsymbol{\theta},\mathbf{v},\mathbf{z},\boldsymbol{\eta}) = \mathcal{N}(\mathbf{z}-\boldsymbol{\eta};0,0.1)$, i.e. $\boldsymbol{\eta}$ plays the same role as $\boldsymbol{\theta}$ in Example 2.5, and whatever procedure we have come up with to select $\boldsymbol{\theta}$, the same can be used for $\boldsymbol{\eta}$.*

Notice that when we parametrize the ranking function as in Example 2.9, we in a sense violate the notion that hyperparameters should not directly enter in the model $\mathbf{M}$ in that we can view $\boldsymbol{\eta}$, which is a hyperparameter, as being part of $\mathbf{M}$. However, formally this is not formally the case since $\boldsymbol{\theta}$ is used as a proxy for $\boldsymbol{\eta}$.

To further accentuate that there is little difference between model parameters and hyperparameters we return to Example 2.8.

Example 2.10 (Example 2.8 continued)**.** *The model*

$$\mathbf{z}(\boldsymbol{\theta},\mathbf{v}) = \mathbf{T}\boldsymbol{\theta} + \sqrt{\lambda}\mathbf{v}$$

with ranking $p(\boldsymbol{\theta},\mathbf{v}) = \mathcal{N}(\mathbf{v};0,I)$ give rise to exactly the same ranking as the set-up in Example 2.8. However, here λ is considered a model parameter rather than a hyperparameter.

2.5.4 *Alternative model selection functions*

As it is typically difficult for a user to justify the choice of ranking function there may seem to be no particular reason for why the top-ranked model should be selected. There are of course many alternative strategies to pick a model from the set of unfalsified models. Each one having its own justification. Here we will give a few examples.

The conditional average ranking model and the median model Consider the situation depicted in Figure 2.13 where the ranking function has two peaks far apart close in height. Instead of picking the top ranked model one could argue that it would be more robust to pick a point more in the center of $\mathcal{U}(\mathbf{z})$, e.g. the center point ξ_c or the point $\bar{\bar{\xi}}$ which is the average of the rankings in $\mathcal{U}(\mathbf{z})$

$$\bar{\bar{\xi}}(\mathbf{z}) = \int_{\mathcal{U}(\mathbf{z})} \xi p(\xi|\mathbf{z}) d\xi, \tag{2.34}$$

We call this model the *conditional average ranking model*.

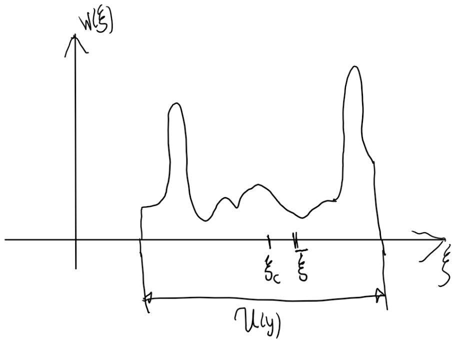

Figure 2.13: The center point ξ_c of the set of unfalsified models and the conditional average ranking model $\bar{\bar{\xi}}$.

These choices seem motivated also for the situation in Figure 2.14. However, here one may be more inclined to pick the point $\tilde{\xi}$ which has half the "mass" of the rankings on either side, we call this the median model.

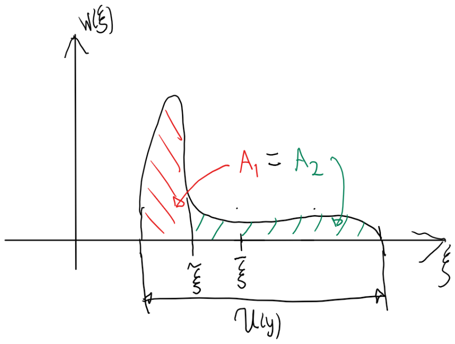

Figure 2.14: The median model and the conditional average ranking model.

Let us compute the conditional average ranking model for Example 2.6.

Example 2.11 (Example 2.6 continued)**.** *The conditional average ranking* *model of $\boldsymbol{\theta}$ can be written as*

$$\bar{\boldsymbol{\theta}}(\mathbf{z}) = \int \boldsymbol{\theta} p(\boldsymbol{\theta}, \mathbf{v}|\mathbf{z}) d\mathbf{v} d\boldsymbol{\theta} = \frac{\int \boldsymbol{\theta} p(\boldsymbol{\theta}, \mathbf{v}, \mathbf{z}) d\mathbf{v} d\boldsymbol{\theta}}{p(\mathbf{z})}$$

Comparing with (2.20) *and* (2.25) *gives that this is the integral over all θ with the same integrand as in* (2.25) *multiplied with $\boldsymbol{\theta}$. But since the integrand in* (2.25) *is $\mathcal{N}(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}(\mathbf{z}), \lambda(\mathbf{T}^T\mathbf{T})^{-1})$, $\bar{\theta}(\mathbf{z})$ is the mean of this distribution*

$$\bar{\boldsymbol{\theta}}(\mathbf{z}) = \hat{\boldsymbol{\theta}}(\mathbf{z})$$

Thus in this simple case the conditional average ranking model and the top ranked model are the same.

The conditional average ranking model has an interesting optimality property. For a random vector $\mathbf{x} \in \mathbb{R}^n$ and any constant vector in $\mathbf{m} \in \mathbb{R}^n$ we have[19]

$$\begin{aligned} \mathbb{E}\left[(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T\right] &= \mathbb{E}\left[(\mathbf{x}-\mathbb{E}\left[\mathbf{x}\right])(\mathbf{x}-\mathbb{E}\left[\mathbf{x}\right])^T\right] + (\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T \\ &\geq \mathbb{E}\left[(\mathbf{x}-\mathbb{E}\left[\mathbf{x}\right])(\mathbf{x}-\mathbb{E}\left[\mathbf{x}\right])^T\right] \end{aligned}$$

with equality if and only if $\mathbf{m} = \mathbb{E}\left[\mathbf{x}\right]$. Taking the trace gives

$$\mathbb{E}\left[|\mathbf{x}-\mathbf{m}|^2\right] \geq \mathbb{E}\left[|\mathbf{x}-\mathbb{E}\left[\mathbf{x}\right]|^2\right]$$

Now, recall that $p(\boldsymbol{\xi}|\mathbf{z})$ is normalized (2.19) and therefore can be regarded as a pdf with domain $\mathcal{U}(\mathbf{z})$. Thus the result above implies that for any model selection function $\boldsymbol{\xi}^*(\mathbf{z})$ that only depends on $\mathbf{z}$,

$$\int_{\mathcal{U}(\mathbf{z})} |\boldsymbol{\xi}^*(\mathbf{z}) - \boldsymbol{\xi}|^2 p(\boldsymbol{\xi}|\mathbf{z}) d\boldsymbol{\xi} \geq \int_{\mathcal{U}(\mathbf{z})} |\bar{\boldsymbol{\xi}}(\mathbf{z}) - \boldsymbol{\xi}|^2 p(\boldsymbol{\xi}|\mathbf{z}) d\boldsymbol{\xi} \qquad (2.35)$$

where $\bar{\boldsymbol{\xi}}(\mathbf{z})$ is the conditional average ranking model (3.2). We can interpret this inequality in decision theoretic terms. Suppose that we have a family of possible data generating mechanisms, represented by $\mathcal{U}(\mathbf{z})$, of which one has generated our data. Our task is to decide on one of the possible mechanisms, i.e. choosing $\boldsymbol{\xi}^*(\mathbf{z})$. The penalty for choosing $\boldsymbol{\xi}^*(\mathbf{z})$ when the true mechanism is $\boldsymbol{\xi}$ is taken as the squared distance in Ξ, weighted by the ranking function $p(\boldsymbol{\xi})$[20]. As all $\boldsymbol{\xi} \in \mathcal{U}(\mathbf{z})$ are possible what we would like to minimize is then the total penalty as $\boldsymbol{\xi}$ ranges over all possibilites dictated by the observation $\mathbf{z}$. This penalty is represented by the integral

$$\int_{\mathcal{U}(\mathbf{z})} |\boldsymbol{\xi}^*(\mathbf{z}) - \boldsymbol{\xi}|^2 p(\boldsymbol{\xi}|\mathbf{z}) d\boldsymbol{\xi}$$

According to (2.35), the conditional average ranking model solves this decision problem.

Now, we can multiply the above inequality with $p(\mathbf{z})$. Since $p(\mathbf{z}) \geq 0$, the inequality is maintained and further, since per definition $p(\boldsymbol{\xi}|\mathbf{z})p(\mathbf{z}) = p(\boldsymbol{\xi}, \mathbf{z})$, we obtain after integrating over $\mathbf{z}$

$$\int \int_{\mathcal{U}(\mathbf{z})} |\boldsymbol{\xi}^*(\mathbf{z}) - \boldsymbol{\xi}|^2 p(\boldsymbol{\xi}, \mathbf{z}) d\boldsymbol{\xi} d\mathbf{z} \geq \int \int_{\mathcal{U}(\mathbf{z})} |\bar{\boldsymbol{\xi}}(\mathbf{z}) - \boldsymbol{\xi}|^2 p(\boldsymbol{\xi}, \mathbf{z}) d\boldsymbol{\xi} d\mathbf{z}$$

[19]

$$\begin{aligned} &\mathbb{E}\left[(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T\right] \\ =&\mathbb{E}\left[(\mathbf{x}-\mathbb{E}\left[\mathbf{x}\right]+\mathbb{E}\left[\mathbf{x}\right]-\mathbf{m})(\mathbf{x}-\mathbb{E}\left[\mathbf{x}\right]+\mathbb{E}\left[\mathbf{x}\right]-\mathbf{m})^T\right] \\ =&\mathbb{E}\left[(\mathbf{x}-\mathbb{E}\left[\mathbf{x}\right])(\mathbf{x}-\mathbb{E}\left[\mathbf{x}\right])^T\right] + 2\mathbb{E}\left[(\mathbf{x}-\mathbb{E}\left[\mathbf{x}\right])(\mathbb{E}\left[\mathbf{x}\right]-\mathbf{m})^T\right] \\ &+ (\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T \\ =&\mathbb{E}\left[(\mathbf{x}-\mathbb{E}\left[\mathbf{x}\right])(\mathbf{x}-\mathbb{E}\left[\mathbf{x}\right])^T\right] + (\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T \end{aligned}$$

[20] Recall that $p(\boldsymbol{\xi}|\mathbf{z}) = p(\boldsymbol{\xi})/p(\mathbf{z})$, so weighting with $p(\boldsymbol{\xi}|\mathbf{z})$ or $p(\boldsymbol{\xi})$ only differ by a constant.

Since for fix $\mathbf{z}$, $p(\xi,\mathbf{z})$ vanishes outside $\mathcal{U}(\mathbf{z})$ we can extend the domain of integration for ξ to Ξ

$$\int\int_{\Xi}|\xi^*(\mathbf{z})-\xi|^2p(\xi,\mathbf{z})d\xi d\mathbf{z}\geq\int\int_{\Xi}|\bar{\xi}(\mathbf{z})-\xi|^2p(\xi,\mathbf{z})d\xi d\mathbf{z}$$

Further, using that $p(\xi,\mathbf{z})=p(\xi)\delta(\mathbf{z}(\xi)-\mathbf{z})$ in the above inequality and reversing the order of integration gives

$$\int_{\Xi}|\xi^*(\mathbf{z}(\xi))-\xi|^2p(\xi)d\xi\geq\int_{\Xi}|\bar{\xi}(\mathbf{z}(\xi))-\xi|^2p(\xi)d\xi \tag{2.36}$$

Again we can make a decision theoretic interpretation. Using the same penalty as above, the conclusion is that the conditional average ranking model minimizes the total penalty as θ ranges over all possible data generating mechanisms that we are considering. Formally, $\bar{\xi}(\cdot)$ solves the function minimization problem[21]

$$\arg\min_{f:\mathbb{R}^N\to\Xi}\int_{\Xi}|f(\mathbf{z}(\xi))-\xi|^2p(\xi)d\xi \tag{2.37}$$

which at first sight looks highly non-trivial. If we review what we have done, the key to solving the problem above is to split up the integral and solve the sub-problem where $\mathbf{z}(\xi)$ is fix; this is (2.35).

Using the total ranking for subsets of model parameters One may also use the total ranking in the set of unfalsified models for a subset of the model parameters to select a model. Let $\xi=\begin{bmatrix}\xi_1^T & \xi_2^T\end{bmatrix}^T$, we may then form the total rankings for ξ_1

$$p(\xi_1;\mathbf{z})=\int_{\Xi}p(\xi;\mathbf{z})d\xi_2$$

and then use this in place of $p(\xi;\mathbf{z})$ to construct a model selection function for ξ_1, e.g. any one of those we have discussed above. One may do the same for ξ_2, or one may split up the elements of ξ in more subsets. In particular, each element may be treated separately.

Low dimensional models. Often there may be reasons to favor a model which resides in a lower dimensional subspace of $\mathbb{R}^n$ (to which ξ belongs). A straightforward way of achieving this is to consider the corresponding model structure and use some model selection function to select a model. However, it frequently happens that the user is fairly confident that the present model structure is relevant and also that the ranking function makes sense. One would thus like to use this knowledge to see if the model structure can be simplified. One way to approach this problem would be to try to find a lower dimensional model in the set of (95%) top ranked models. Let us use Example 2.6 to see how this could look like.

Example 2.12 (Example continued). *In Example 2.7 we saw that the set of top ranked models is characterized by models where $\mathbf{T}\theta$ fits the observation $\mathbf{z}$ well, cf.* (2.27). *Suppose now that we are interested in simplifying the model so that only a subset of the elements of θ are non-zero. Let $\tilde{\mathbf{T}}\tilde{\theta}$ be the reduced model, i.e. $\tilde{\mathbf{T}}\tilde{\theta}=\mathbf{T}\theta$ when θ has the desired zeros. Then the*

[21] Replacing $|\cdot|^2$ with some other penalty will result in other conditional ranking models, see, e.g., Corollary 1.2 in Chapter 4 of

E. L. Lehmann and G. Casella. *Theory of Point Estimation*. John Wiley & Sons, New York, second edition edition, 1998 *model of this type with best chance to belong to the set of top ranked models is the model corresponding to the least squares estimate* $\hat{\hat{\boldsymbol{\theta}}} := (\tilde{\mathbf{T}}^T\tilde{\mathbf{T}})^{-1}\tilde{\mathbf{T}}^T\mathbf{z}$*. One can thus test all possible selections of columns of* $\mathbf{T}$ *in* $\tilde{\mathbf{T}}$ *and take the model structure with smallest number of columns satisfying*

$$|\mathbf{z} - \tilde{\mathbf{T}}\hat{\hat{\boldsymbol{\theta}}}|^2 \leq |\mathbf{z} - \mathbf{T}\hat{\boldsymbol{\theta}}|^2 + \lambda F^{-1}_{\chi^2(n)}(0.95)$$

This criterion can be given several interesting interpretations.

2.5.5 *Functions of model parameters*

A very common situation is that we are not directly interested in the model but some derived quantity. For example, the intended use of the model may be to design a controller. With $\gamma(\cdot) : \Xi \to \mathbb{R}^m$ denoting the specific control design algorithm, mapping a model to the parameters of the controller, our interest is $\gamma(\xi)$. For this we may apply $\gamma(\cdot)$ to whatever model from the set of unfalsified models we have judged appropriate, e.g. the top ranked model or the conditional average ranked model, giving $\gamma(\hat{\xi})$ and $\gamma(\bar{\xi})$, respectively.

An alternative is to reparametrize the ranking function in terms of $\gamma = \gamma(\xi)$. Notice that typically the dimension of γ is lower than the dimension of ξ which means that the dimension of the model set Ξ is reduced. Formally, we can achieve this by introducing $\gamma \in \mathbb{R}^m$ as a fictitious observation together with its "model" $\gamma = \gamma(\xi)$ so that the complete model is described by

$$\begin{bmatrix} \gamma \\ \mathbf{z} \end{bmatrix} = \begin{bmatrix} \gamma(\xi) \\ \mathbf{z}(\xi) \end{bmatrix}$$

We can then define the ranking function for all our parameters and observations (including the fictitious γ)

$$p(\xi, \gamma, \mathbf{z}) := p(\xi, \mathbf{z})\delta(\gamma - \gamma(\xi)) = p(\xi)\delta(\mathbf{z} - \mathbf{z}(\xi))\delta(\gamma - \gamma(\xi))$$

and define the total rankings corresponding to $\mathbf{z}$ and γ by integrating over ξ

$$p(\gamma, \mathbf{z}) := \int_{\Xi} p(\xi, \gamma, \mathbf{z}) d\xi$$

Recall that we can view $p(\xi, \mathbf{z})$ as the ranking of ξ taking into account the set of feasible models, similarly $p(\xi, \gamma)$ can now be seen as the ranking of γ taking into account the set of feasible models. We can therefore use this ranking for selecting a γ consistent with the observation $\mathbf{z}$ in *exactly* the same way as we have done when selecting a model ξ using the ranking $p(\xi, \mathbf{z})$. We can thus compute, e.g., the top ranked γ

$$\hat{\gamma}(\mathbf{z}) = \arg\max_{\gamma} p(\gamma, \mathbf{z})$$

or the conditional average rank model

$$\bar{\gamma}(\mathbf{z}) = \int \gamma p(\gamma, \mathbf{z}) d\gamma$$

Notice that, in general, neither $\hat{\gamma}(\mathbf{z}) = \gamma(\hat{\xi}(\mathbf{z}))$ (see Exercise 3.1.a), nor $\bar{\gamma}(\mathbf{z}) = \gamma(\bar{\xi}(\mathbf{z}))$.

2.5.6 Choosing the ranking function

We have now come to the core problem: How should we choose the ranking function? Well, a first observation is that the true data generating mechanism should be given high rank; ideally it should be ranked infinitely high so that no matter what reasonable selection mechanism $\xi(\mathbf{z})$ we use, this model is selected. Now this is perhaps a silly observation since we wouldn't be in the need of learning if this information was available to us. However it provides us with an important guideline for how to encode prior knowledge into rankings:

Models that are consistent with our prior knowledge should be given high rankings.

The better we are able to do this, the more precise we will be able to model the observation. It's as simple as that. Yet, it may seem like a formidable task to assign rankings to all possible models, even if we, as we have seen, can restrict attention to the set of unfalsified models. To deal with this we will now embark on a path that we will pursue throughout the lecture notes. We will use a constructive approach which has proven to be very successful, able to incorporate a rich variety of qualitative prior information.

The basic idea is to introduce some additional assumptions on the model parameters ξ and then construct a ranking function which favors models for which these assumptions are satisfied.

Example 2.13. *Let us consider the model*

$$y(t) = \sum_{k=1}^{n} \theta_k u(t-k) + v(t)$$

where $\{u(t)\}$ is the input and $\{v(t)\}$ is the noise.

A natural assumption may be that $u(t)$ should not depend on the present and the future of the noise sequence $v(t+k)$, $k = 0, 1, \ldots$. Now for finite data, the model above can be written as in Example 2.7

$$\mathbf{z} = \mathbf{T}\boldsymbol{\theta} + \mathbf{v}$$

where $\mathbf{T}$ is a Toeplitz matrix[22]

$$\mathbf{T} = \begin{bmatrix} u(0) & u(-1) & \ldots & u(1-n) \\ u(1) & u(0) & \ldots & u(2-n) \\ \vdots & \vdots & \ddots & \vdots \\ u(N-1) & u(N-2) & \ldots & u(N-n) \end{bmatrix}$$

A (partial) way to capture that the input and the disturbance are independent is to introduce the assumption that the covariance between the two quantities is close to zero. Assuming the sample mean of the input is zero, this means

$$\left|\frac{1}{N}\mathbf{u}^T\mathbf{v}\right| \text{ "small"}, \quad \mathbf{u} = \begin{bmatrix} u(1) & \ldots & u(N) \end{bmatrix}^T$$

To incorporate this in the ranking approach, we could for example choose the ranking function to measure how small the magnitude of the covariance

[22] A Toeplitz matrix has the same elements along all its subdiagonals.

is, e.g. by using $p(\boldsymbol{\theta},\mathbf{v}) = -|\mathbf{u}^T\mathbf{v}|$. Maximizing this ranking function on the set of unfalsified models gives

$$\min_{\boldsymbol{\theta}} \left|\frac{1}{N}\mathbf{u}^T(\mathbf{z}-\mathbf{T}\boldsymbol{\theta})\right|$$

However, setting $\mathbf{u}^T(\mathbf{z}-\mathbf{T}\boldsymbol{\theta})$ to zero gives a linear equation so unless $\boldsymbol{\theta}$ is a scalar, typically zero covariance can be achieved by multiple solutions. To remove the ambiguity, additional covariances may be used. For this we notice that the Toeplitz structure of $\mathbf{T}$ gives that $\frac{1}{N}\mathbf{T}^T\mathbf{v}$ contains sample covariances for lags $\tau = 1,\ldots,n_\theta$. This gives the ranking function

$$p(\boldsymbol{\theta},\mathbf{z}-\mathbf{T}\boldsymbol{\theta}) = \left|\mathbf{T}^T(\mathbf{z}-\mathbf{T}\boldsymbol{\theta})\right|$$

Here the number of equations and unknowns are equal giving the unique solution

$$\boldsymbol{\theta} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{z}$$

when $\mathbf{T}$ has full column rank, i.e. we arrive at the least-squares estimate which also appeared in Example 2.6.

The more elaborate assumptions that are introduced, the more information we will be able to squeeze out of the observations, if the assumptions are correct that is. Care thus has to be exercised when introducing prejudices.

2.6 *Essential Aspects*

2.6.1 *A water-bed effect and overfitting*

As we now have repeatedly have seen, the fundamental problem in learning is that the data-model relationship (2.7) is under-determined, i.e. the number of model parameters n_ζ exceeds the number of observations N. We will call the number of excess parameters, $n_\zeta - N$ the *model degrees of freedom (mdf)*[23]. As a consequence of this ambiguity, measurement noise and disturbances in data may be attributed to model parameters describing the dynamics and it may seem intuitive that this effect is exacerbated with increasing degrees of freedom.

Now, the constraint

$$\mathbf{z} = \mathbf{M}(\boldsymbol{\zeta})$$

give rise to a *water-bed effect*. Incorrectly attributing a part of the observations to certain model parameters means that the same parameters can be used to describe less of what they actually should explain in the observations.

We once more return to Example 2.6 to illustrate this phenomenon.

Example 2.14 (Example 2.6 continued). *Recall the model*

$$\mathbf{M}(\boldsymbol{\zeta}) = \mathbf{T}\boldsymbol{\theta} + \mathbf{v}, \quad \boldsymbol{\zeta} := \begin{bmatrix}\boldsymbol{\theta}^T & \mathbf{v}^T\end{bmatrix}^T$$

[23] Later we will define the degrees of freedom in a statistically meaningful way.

where $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$, $n_\theta \leq N$, and where $\mathbf{T} \in \mathbb{R}^{N \times n_\theta}$ is full (column) rank, which with the ranking $p(\boldsymbol{\theta}, \mathbf{v}) = \mathcal{N}(\mathbf{v}; 0, \lambda)$ gives the top ranked model

$$\hat{\boldsymbol{\theta}} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{z}$$

Assuming that the true system is $\mathbf{z} = \mathbf{T}\boldsymbol{\theta}_o + \mathbf{v}_o$, we can write this as

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_o + (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{v}_o$$

Thus, we have

$$\mathbf{T}\hat{\boldsymbol{\theta}} = \mathbf{T}\boldsymbol{\theta}_o + \boldsymbol{\Delta}_\mathbf{v}, \quad \boldsymbol{\Delta}_\mathbf{v} := \mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{v}_o$$

Here we recognize $\mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T$ as the orthogonal projection onto the subspace of $\mathbb{R}^N$ spanned by the columns of $\mathbf{T}$. Now the number of parameters in our model is $\dim \boldsymbol{\theta} + \dim \mathbf{v} = n_\theta + N$ while the number of observations is $\dim \mathbf{z} = N$ and hence the model degrees of freedom equals $n_{\boldsymbol{\theta}} + N - N = n_\theta$. But this is exactly the dimension of this subspace (since $\mathbf{T}$ is assumed full rank). Thus, the error consist of the true noise projected onto a subspace whose dimension equals the model degrees of freedom. The corresponding model of the noise is

$$\begin{aligned}\hat{\mathbf{v}} :&= \mathbf{z} - \mathbf{T}\hat{\boldsymbol{\theta}} = \mathbf{T}\boldsymbol{\theta}_o + \mathbf{v}_o - \mathbf{T}\boldsymbol{\theta}_o - \boldsymbol{\Delta}_\mathbf{v} \\ &= \mathbf{v}_o - \mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{v}_o = (\mathbf{I} - \mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T)\mathbf{v}_o\end{aligned}$$

and here we recognize $\mathbf{I} - \mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T$ as the orthogonal projection on the subspace orthogonal to the span of the columns of $\mathbf{T}$. This space has dimension $N - n_\theta$. Notice that since $\boldsymbol{\Delta}_\mathbf{v}$ and $\hat{\mathbf{v}}$ are orthogonal projections of the vector $\mathbf{v}_o$ on complementary subspaces, it holds that

$$|\mathbf{v}_o|^2 = |\hat{\mathbf{v}}|^2 + |\boldsymbol{\Delta}_\mathbf{v}|^2 \tag{2.38}$$

As a consequence, $|\hat{\mathbf{v}}| < |\mathbf{v}_o|$ which means that the model underestimates *the amount of noise in the data. This effect can be significant when the ratio n_θ/N is non-negligible. This is the water-bed effect in force: the part $\boldsymbol{\Delta}_\mathbf{v}$ of $\mathbf{v}_o$ has been used to explain a part of the contribution from the term $\mathbf{T}\boldsymbol{\theta}_o$ and hence cannot be used to explain $\mathbf{v}_o$ since the constraint $\mathbf{z} = \mathbf{T}\boldsymbol{\theta} + \mathbf{v}$ has to be satisfied. Notice also that $|\hat{\mathbf{v}}|^2$ is the minimum of the cost function on the right in* (2.15).

Consider now that we augment our model with an additional regressor $\mathbf{t}$, giving the model

$$\mathbf{z} = \mathbf{T}\boldsymbol{\theta} + \mathbf{t}\alpha + \mathbf{v}$$

where α is another unknown parameter. Assuming that the observations are still given by $\mathbf{z} = \mathbf{T}\boldsymbol{\theta}_o + \mathbf{v}_o$ this means that we have a surplus model parameter α to determine. However, the derivations above apply also for this model which gives that the error between $\mathbf{T}\boldsymbol{\theta}_o$ and our top ranked augmented model is given by the orthogonal projection of $\mathbf{v}_o$ on the subspace spanned by the columns of $\mathbf{T}$ and the vector $\mathbf{t}$, i.e. a larger subspace than we had before. Thus the augmented model has a larger norm of the error than the original model and consequently a smaller norm of $\hat{\mathbf{v}}$ (since it is the projection on the orthogonal complement of that subspace, which instead has shrunk in dimension).

An alternative view of what happens when we augment a model with more parameters than necessary is that the added degrees of freedom (α in the example) allows the minimum of (2.15) to be decreased, implying that a smaller fraction of the noise $\mathbf{v}_0$ is part of the noise model $\hat{\mathbf{v}}$. However, in view of (2.38) the noise on the model of $\mathbf{T}\boldsymbol{\theta}_0$ must then have increased since the sum of the two noise terms is constant. For this reason, the increase in the model error is said to be due to *overfitting*. Overfitting in a much more general setting than Example 2.14 is discussed in Exercise 3.7.

Overfitting has a serious impact on the overall learning problem. Typically it is beforehand not known exactly which model structure $\mathbf{M}$ to use. A simple remedy to this would be to use a very flexible model structure, i.e. one that has many model parameters; the reader may think of a universal model structure able to accomodate any system behavior. However, overfitting effectively prevents the use of such model structures as then most of the disturbances and noise in the data will be attributed to the system model. Thus model structure selection and how to constrain flexible model structures so that overfitting does not occur are two of the most challenging problems in learning. The latter problem is known as *regularization*.

2.6.2 The role of the excitation

For a given model structure $\mathbf{M}(\cdot)$, it is the observations $\mathbf{z}$ that define the set of unfalsified models (2.8) but since the model equation (2.7) is under-determined the set of unfalsified models will always be unbounded. However, the information in the observations in regards to the different model parameters will influence how the set of unfalsified models is shaped. In particular the information contents will influence how the highest ranked models in the set of unfalsified models cluster. Notice that we are here talking about information in the observations *as specified by the used model structure* and not the actual information that is present in the observations. Only when the true system is described within the model structure do these notions coincide.

In Example 2.7 the set of unfalsified $\boldsymbol{\theta}$ does not shrink as more measurements are collected; all possible values of $\boldsymbol{\theta}$ are feasible since all possible $\mathbf{v}$ are allowed. However, the 95% top ranked models belong to an ellipsoid. Let us analyse this ellipsoid for a specific case.

Example 2.15 (Example 2.7 continued)**.** *Suppose that the model structure is*

$$\mathbf{z}(t) = \boldsymbol{\theta}\mathbf{u}(t) + \mathbf{v}(t),\ t = 1, \ldots, N$$

where $\mathbf{u}(t)$ is the input to the system. We can write this in vector form as

$$\mathbf{z} = \mathbf{T}(\mathbf{u})\boldsymbol{\theta} + \mathbf{v}, \quad \mathbf{T}(\mathbf{u}) = \begin{bmatrix} \mathbf{u}(1) \\ \vdots \\ \mathbf{u}(N) \end{bmatrix}$$

This model structure thus corresponds to the model structure used in Examples 2.6-2.7 with a specific regressor matrix $\mathbf{T} = \mathbf{T}(\mathbf{u})$.

Suppose now that the input is constant $\mathbf{u}(1) = \ldots = \mathbf{u}(N) = A$. *This means that we make repeated observations of the scalar* $A\boldsymbol{\theta}$ *in noise. Since* $\mathbf{T}^T(\mathbf{u})\mathbf{T}(\mathbf{u}) = A\,N$, (2.26) *corresponds to* $\boldsymbol{\theta}$ *satisfying*

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^2 \leq \frac{\lambda F^{-1}_{\chi^2(1)}(0.95)}{NA^2}$$

i.e. the width of the set consisting of the top 95% ranked model parameters $\boldsymbol{\theta}$ *is of the order* $\mathcal{O}(1/\sqrt{NA^2})$. *Thus the width can be reduced arbitrarily by taking the product* NA^2 *large enough. Interpreting the squared amplitude* A^2 *as the power of the input, the product* NA^2 *can be seen as the energy in the input signal. Thus when the input energy grows to infinity, the set of top-ranked models shrinks to a single point - the top ranked model. As we will see this is no coincidence. Notice that this do not imply that the top ranked model corresponds to how the data has been generated.*

The behavior of the size of the top-ranked models seen in the previous example is quite typical, but, of course, the specifics depend on the way we have chosen to rank our models. However, it also depends critically on the information contents in the observations as specified by the model structure. In Example 2.15 the model structure specified that each observation contained the same information about $\boldsymbol{\theta}$. Le us now see what happens if we change this specification.

Example 2.16 (Example 2.15)**.** *continued] Suppose that instead* $u(t) = e^{-(t-1)}$. *Then*

$$\mathbf{T}^T(\mathbf{u})\mathbf{T}(\mathbf{u}) = \sum_{t=0}^{N-1} e^{-2t} = \frac{1 - e^{-2N}}{1 - e^{-2}}$$

and hence (2.26) *corresponds to* θ *satisfying the constraint*

$$(\theta - \hat{\theta})^2 \leq \lambda F^{-1}_{\chi^2(1)}(0.95)\,\frac{1 - e^{-2N}}{1 - e^{-2}}$$

i.e. regardless of how many measurements N we use, the width of the set of the 95% top ranked θ*:s never shrinks below* $\sqrt{\lambda F^{-1}_{\chi^2(1)}(0.95)/(1 - e^{-2})}$.

The feature in the previous example is that $\boldsymbol{\theta}$ becomes (exponentially) less visible in $\mathbf{z}(t)$ for increasing time index t. This makes the distribution of the rankings in the set of unfalsified models much less peaked than in Example 2.15. The information contents in the observations as specified by the model structure is thus instrumental for the size of the set of top ranked models. At the same time, the two preceding examples illustrate that the information contents can (often) be controlled with the external excitation if there are inputs that can be manipulated by the user and, typically, the energy of the excitation determines the size of the set of top ranked models.

Not only the size but also the shape of the set of top-ranked models can be controlled by the input excitation.

Example 2.17. *Suppose that the model structure is*

$$\mathbf{z}(t) = \boldsymbol{\theta}_1\mathbf{u}(t) + \theta_2\mathbf{u}(t-1) + \mathbf{v}(t),\ t = 1,\dots,N$$

where $\mathbf{u}(t)$ *is the input to the system. We can write this in vector form as*

$$\mathbf{z} = \mathbf{T}(\mathbf{u})\boldsymbol{\theta} + \mathbf{v}, \quad \mathbf{T}(\mathbf{u}) = \begin{bmatrix} \mathbf{u}(1) & \mathbf{u}(0) \\ \vdots & \vdots \\ \mathbf{u}(N) & \mathbf{u}(N-1) \end{bmatrix}$$

Then

$$\mathbf{T}^T(\mathbf{u})\mathbf{T}(\mathbf{u}) = \begin{bmatrix} \mathbf{u}(1) & \dots & \mathbf{u}(N) \\ \mathbf{u}(0) & \dots & \mathbf{u}(N-1) \end{bmatrix} \begin{bmatrix} \mathbf{u}(1) & \mathbf{u}(0) \\ \vdots & \vdots \\ \mathbf{u}(N) & \mathbf{u}(N-1) \end{bmatrix}$$
$$= \sum_{t=1}^{N} \begin{bmatrix} \mathbf{u}^2(t) & \mathbf{u}(t-1)\mathbf{u}(t) \\ \mathbf{u}(t-1)\mathbf{u}(t) & \mathbf{u}^2(t-1) \end{bmatrix}$$

implying that the sample covariances of the input determine the shape of (2.26) *(which in this two-dimensional case is an ellipse).*

We can in fact choose any orientation and axis lengths of the ellipse by appropriate choice of the input sequence. With

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{bmatrix}^T \begin{bmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{bmatrix} = \begin{bmatrix} r_{11}^2 & r_{11}r_{12} \\ 0 & r_{12}^2 + r_{22}^2 \end{bmatrix} > 0$$

denoting the desired $\mathbf{T}^T(\mathbf{u})\mathbf{T}(\mathbf{u})$*, and taking* $N = 2$ *with*

$$u(0) = \frac{r_{11}r_{22} - r_{12}^2 - r_{22}^2}{d}$$
$$u(1) = \frac{-r_{11}r_{12}}{d}$$
$$u(2) = \frac{r_{11}r_{22} - r_{11}^2}{d}$$
$$d = \sqrt{r_{12}^2 + (r_{11} - r_{22})^2}$$

will give the desired matrix[24]*. The set of top-ranked models enclosed by the blue curve in Figure 2.15 is a disk, meaning that the rankings are distributed uniformly with respect to angular direction. The corresponding matrix satisfies* $\mathbf{T}^T(\mathbf{u})\mathbf{T}(\mathbf{u}) = \mathbf{I}$*, which was obtained using*

$$u(0) = 1,\ u(1) = 0,\ u(2) = 1$$

More generally $\mathbf{T}^T(\mathbf{u})\mathbf{T}(\mathbf{u}) = A^2\mathbf{I}$*, meaning that the circle will have radius* $1/A$*, is obtained by the input sequence*

$$u(0) = A,\ u(1) = 0,\ u(2) = A$$

Thus, as in the preceeding examples, the inverse of the input amplitude determines the width (in this case the radius) of the set of top-ranked models. Now, the allowed input amplitude of an input is typically constrained. Then, as before, one can decrease the width by performing a longer experiment. There is, however, an alternative approach, indicated by the ellipse with the red border in Figure 2.15.

[24] When $d = 0$ the left-hand sides should be seen as limits.

The top-ranked set enclosed by the red curve has a direction where the top-ranked models are more dense and therefore the semi-minor axis is shorter than the other axis for this set. In this direction the width of the set is smaller than the set with the blue circle as border and this is achieved with an input having an amplitude smaller than for the set corresponding to the blue curve. Thus, we have here traded off an increase in one direction for a decrease in another direction using less input power.

The ellipse has

$$\mathbf{R} = A\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

with $\rho = 0.99$ *used in the figure, which gives the input sequence*

$$\begin{aligned} u(0) &= -A\frac{\sqrt{1-\sqrt{1-\rho^2}}}{\sqrt{2}} \\ u(1) &= -A\frac{\rho}{d(\rho)} \\ u(2) &= u(0) \\ d(\rho) &= \sqrt{2}\sqrt{1-\sqrt{1-\rho^2}} \end{aligned}$$

and thus $|u(0)| = |u(2)| \leq A/\sqrt{2}$. *Furthermore,* $\rho/d(\rho)$ *is a monotonically decreasing function on the interval* $[0,1]$, *taking the values* 1 *and* $1/\sqrt{2}$ *at the end-points. This function is less than* c *when* $\rho > 2c/(1+c^2)$. *Thus* $|u(1)| \leq A$ *regardless of* ρ.

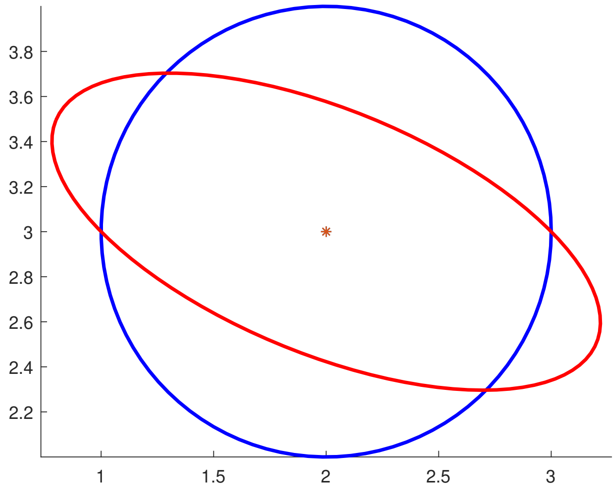

Figure 2.15: Examples of two different ellipses that can be obtained in Example 2.17.

Tweaking the input to achieve certain properties of the obtained model is called *experiment design*. Often, as we did in Example 2.17, it concerns achieving a certain shape of the set of top ranked models. Now Example 2.17 was an extremely simple case where we designed the input samples of an experiment that was only three samples long. The set of top-ranked models also had a simple characterization as an ellipsoid centered at the top-ranked model but with a shape that did not depend on this model. In general this is not the case, the set can have a complex structure and may also depend on the top ranked model which only is available after the experiment. We will see in Chapter **??** how these issues can be handled.

2.6.3 Linking model selection to a "true" system

So far we have not touched on the subject which model selection method to choose. Is it better to use the top ranked model or the conditional average ranked model or maybe some other model selection method is to be preferred? In order to perform such an analysis we need to specify how the data actually has been generated. Here one typically assumes that the true data generating mechanism belongs to the model set Ξ, but also analysing the case where this is not the case can be highly relevant in order to understand how modeling imperfections can influence the obtained model. Let us for now stay with the former case.

The analysis typically consists of considering that the true data generating mechanism belongs to a subset of the model set and analysing how the models that are selected compare with the ground truth.

Worst-case analysis. In *worst-case analysis* all alternatives are considered and the worst-case is singled out. We illustrate this with a simple example.

Example 2.18. *Let our model for the observation* $\mathbf{z} \in \mathbb{R}^N$ *be*

$$\mathbf{z} = \theta\,\mathbf{1} + \mathbf{v} \tag{2.39}$$

where $\mathbf{1} = \begin{bmatrix} 1 & \dots & 1 \end{bmatrix}^T$. *Let the model set be*

$$\Xi = \{(\theta, \mathbf{v}) : |v(t)| \leq C,\ t = 1, \dots, N\}$$

This type of specification is known as unknown-but-bounded noise.

Then the set of unfalsified models is given by

$$\mathcal{U}(\mathbf{z}) = \{(\theta, \mathbf{v}) : |y(t) - \theta| \leq C,\ t = 1, \dots, N,\ \mathbf{v} = \mathbf{z} - \theta\,\mathbf{1}\}$$

Suppose now that we would like to analyze the Chebyshev center (2.9)

$$\theta_c(\mathbf{z}) = \arg\min_{\theta} \max_{\tilde{\theta} \in \mathcal{U}(\mathbf{z})} |\theta - \tilde{\theta}|$$

from a worst-case perspective, assuming that the data actually has been generated according to (2.39) *for some* θ.

We are thus only interested in the error in θ. *A moments reflection gives that*

$$\theta_c(\mathbf{z}) = \frac{\min_{1 \leq t \leq N} z(t) + \max_{1 \leq t \leq N} z(t)}{2}$$

which, if $\mathbf{z}$ *was generated by a model in the model set, can be written*

$$\begin{aligned} \theta_c(\mathbf{z}) &= \frac{\min_{1 \leq t \leq N} \theta + v(t) + \max_{1 \leq t \leq N} \theta + v(t)}{2} \\ &= \theta + \frac{\min_{1 \leq t \leq N} v(t) + \max_{1 \leq t \leq N} v(t)}{2} \end{aligned} \tag{2.40}$$

The error for a particular model $(\theta, \mathbf{v})$ is thus given by

$$\theta_c(\mathbf{z}) - \theta = \frac{\min_{1 \le t \le N} v(t) + \max_{1 \le t \le N} v(t)}{2}$$

Maximizing the right-hand side over $\xi \in \mathcal{U}(\mathbf{z})$ gives the worst-case error given the observation $\mathbf{z}$ as

$$J_{wc}(\mathbf{z}) := \max_{(\theta,\mathbf{v}) \in \mathcal{U}(\mathbf{z})} |\theta_c(\mathbf{z}) - \theta| = \max_{(\theta,\mathbf{v}) \in \mathcal{U}(\mathbf{z})} \left| \frac{\min_{1 \le t \le N} v(t) + \max_{1 \le t \le N} v(t)}{2} \right|$$

We can proceed and also consider the worst-case that could happen for all possible $\mathbf{z}$ consistent with the model set

$$J_{wc}(\Xi) = \max_{(\theta,\mathbf{v}) \in \Xi} J_{wc}(\mathbf{z}(\theta, \mathbf{v}))$$

This problem has the solution

$$J_{wc}(\Xi) = \max_{|v(t)| \le C,\, t=1,\ldots,N} \left| \frac{\min_{1 \le t \le N} v(t) + \max_{1 \le t \le N} v(t)}{2} \right| = C$$

The worst-case occurs when the sequence $\{v(t)\}_{t=1}^{N}$ is constant and equal to one of its extreme values: $v(t) = C$, $t = 1, \ldots, N$ or $v(t) = -C$, $t = 1, \ldots, N$. As the solution does not depend on the number of observations, there is a priori no guarantee that using many observations will give more information than a single one. This does not mean that a given observation $\mathbf{z}$ will allow us to reduce the error for θ, $J_{wc}(\mathbf{z})$ may very well be less than C, and even 0 if we are in the lucky situation that some elements of $\mathbf{v}$ attain the extremes $\pm C$.

Randomized analysis. We observe that the worst-case error does not depend on the number of observations in the previous example. This is a common phenomenon and inherent in worst-case studies; after all we have an under-determined problem of equations. A more refined analysis can be obtained by considering that the ground truth belongs to an appropriate subset of Ξ, but what is more common is that one conducts an analysis where random draws from Ξ are considered and one then computes the statistics of the errors that one obtains. Let us now analyze the set-up in Example 2.18 in this way.

Example 2.19 (Example 2.18 continued)**.** *Let θ be a fix number and let us assume that we perform M experiments, where in each experiment, indexed by i, we let each element of $\mathbf{v}_i$ be a draw from a random number generator which has a binary distribution $\pm C$, each with probability $1/2$. For each experiment we form the observation*

$$\mathbf{z}_i := \theta\, \mathbf{1} + \mathbf{v}_i$$

for which we compute the Chebyshev center $\theta_{c,i}$. We then form the sample average of the squared errors

$$J_{MSE,N,M} = \frac{1}{M} \sum_{i=1}^{M} |\theta_{c,i}(\mathbf{z}_i) - \theta|^2$$

Since the experiments are independent, or more precisely the $\{\mathbf{v}_i\}$ are independent, $J_{MSE,N,M}$ will converge to the mean-squared error *(MSE)*

$$J_{MSE,N} := \mathbb{E}\left[|\theta_c(\mathbf{z}) - \theta|^2\right]$$

where $\mathbf{z} = \theta\,\mathbf{1} + \mathbf{v}$, where $\mathbf{v}$ is a random vector with the entries independent each binary distributed ±C.

With some skills in probability theory it is now possible to express the MSE as[25]

$$J_{MSE,N} = \frac{C}{2^{N-1}}$$

We see that the MSE quickly becomes very small as the number of observations N grows. The intuitive reason is that for a sequence of identically binary distributed random variables it is very likely that both possible outcomes are observed, which gives an exact estimate.

This is in stark contrast to the worst-case analysis we did above; the worst case error remained C regardless of the number of observations. We must, however, remember that we are now studying another type of error. The MSE is the squared error we can expect if $\mathbf{v}$ from the true data generating mechanism is a realization from a binary distribution. This distribution only occupies the vertices of the model set Ξ.

We may instead consider what happens if the elements of $\mathbf{v}$ have a uniform distribution on $[-C, C]$. In [26] *it is shown that then $J_{MSE,N}$ is $o(1/N^{2-\delta})$ for every $0 < \delta < 2$. Considerably slower than for the binary distribution but still a very fast decay in terms of N. Again the intuition is that the extreme values of a sequence of independent random variables having a pdf which is rather large at the end-points cluster closely to the end-points $-C$ and C, respectively, as the number of observations grows.*

A third distribution with support $[-C, C]$ is

$$p(v) = ce^{-\frac{\tan(\pi|v|}{C}}, \quad |v| \le C$$

where c normalizes p to have unit integral between $[-C, C]$. The pdf is shown in Figure 2.16. Characteristic to this distribution is that it is very thin at the end-points. This prevents the extreme-values of the elements of $\mathbf{v}$ to accumulate at the end-points; instead they end up somewhere on the flat part near the end-points. Formally, it can be shown that the MSE is $\mathcal{O}(1/(\log N)^2)$ which tends to zero much much slower than for the uniform distribution.

The example above is an example of "mind-games" one can play. For a given model selection function $\hat{\xi}(\cdot)$ one can derive its probability distribution assuming that the observations are outcomes of a true data generation mechanism $\mathbf{M}(\xi)$ governed by a certain probability distribution. This distribution can then be analyzed with respect to how concentrated it is, typically by computing the mean $\mathbb{E}\left[\hat{\xi}(\mathbf{z})\right]$, the covariance matrix

$$\mathbf{P}_{\hat{\xi}} := \mathbb{E}\left[(\hat{\xi}(\mathbf{z}) - \mathbb{E}\left[\hat{\xi}(\mathbf{z})\right])(\hat{\xi}(\mathbf{z}) - \mathbb{E}\left[\hat{\xi}(\mathbf{z})\right])^T\right] \qquad (2.41)$$

[25] From (2.40) we see that we should study the distribution of $(\min_{1\le t\le N} v(t), \max_{1\le t\le N} v(t))$ which has the three outcomes $(-C,-C)$, $-C,C)$ and (C,C). For the first and the last all elements of $\mathbf{v}$ need to take on one and the same value. The probability of this is $(1/2)^N$. We thus have

$$\mathbf{P}\left((\min_{1\le t\le N} v(t), \max_{1\le t\le N} v(t)) = (-C,-C)\right)$$

$$= \mathbf{P}\left((\min_{1\le t\le N} v(t), \max_{1\le t\le N} v(t)) = (C,C)\right) = \frac{1}{2^N}$$

$$\mathbf{P}\left((\min_{1\le t\le N} v(t), \max_{1\le t\le N} v(t)) = (-C,C)\right) = 1 - \frac{1}{2^{N-1}}$$

For the first two outcomes we have $|\theta_c - \theta| = C$, while for the middle $|\theta_c - \theta| = 0$. This thus gives us the MSE as

$$J_{MSE,N} = C\frac{1}{2^N} + 0\left(1 - \frac{1}{2^{N-1}}\right) + C\frac{1}{2^N} = \frac{C}{2^{N-1}}$$

[26] H. Cramér. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, 1946

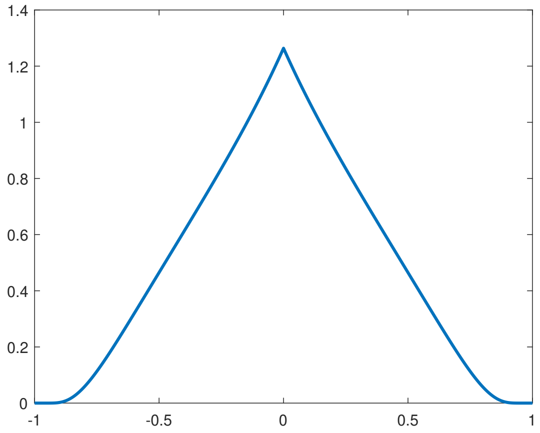

Figure 2.16: Pdf of the distribution used at the end of Example 2.19..

and the MSE

$$\begin{aligned}\mathbf{MSE}_{\hat{\xi}} &:= \mathbb{E}\left[(\hat{\xi}(\mathbf{z})-\xi)(\hat{\xi}(\mathbf{z})-\xi)^T\right] \\ &= \mathbf{P}_{\hat{\xi}} + \left(\mathbb{E}\left[\hat{\xi}(\mathbf{z})\right]-\mathbb{E}\left[\xi\right]\right)\left(\mathbb{E}\left[\hat{\xi}(\mathbf{z})\right]-\mathbb{E}\left[\xi\right]\right)^T \qquad (2.42)\end{aligned}$$

Especially, making such studies for different distributions can build up confidence or detect short-comings in a particular model selection method, knowledge which in turn can be used to select an appropriate model selection method.

Before moving on we remark that the water-bed effect discussed above manifests itself in that for every model selection function $\hat{\boldsymbol{\theta}}(\cdot)$, the moments of the model $\mathbf{M}(\hat{\xi}(\mathbf{z}))$ matches the moments of the observations, e.g.

$$\begin{aligned}\mathbb{E}\left[\mathbf{M}(\hat{\xi}(\mathbf{z}))\right] &= \mathbb{E}\left[\mathbf{z}\right] \\ \mathbb{E}\left[\mathbf{M}(\hat{\xi}(\mathbf{z}))\mathbf{M}(\hat{\xi}(\mathbf{z}))^T\right] &= \mathbb{E}\left[\mathbf{z}\mathbf{z}^T\right]\end{aligned}$$

2.6.4 Generalization

Up to now we have only studied inference of model parameters related to a specific data set $\mathbf{z}$. Now, as we have touched upon earlier, even if we try to maintain identical experimental conditions we will get different responses for different experiments. This effect we typically attribute to noise and disturbances, but can also be due to changes in the dynamics of the system. We thus need a mechanism that can model such effects so that

we can make inference from one data set in regards to properties of another data set

It is also of interest to make inference about what will happen when we extend the length of an experiment.

To this end the use of probability theory has turned out to be very useful. By modelling model parameters as *realizations of random variables* one can use statistical inference to make statements about what can happen in the future - of course predicated on that the true data generating mechanism is governed by the assumed probability distribution. This is thus in line with the randomized analysis of model selection functions discussed in the previous section. Now, this means that we not only need to provide a model structure $\mathbf{M}(\cdot)$ and a ranking function but also we need to specify a probability distribution for the model parameters $\boldsymbol{\xi}$, adding significantly to the burden of the user. However, this issue can be mitigated in several ways.

Firstly, we can connect this distribution to the ranking function. Recall that the ranking function should specify our preference of models. Now a distribution specifies how likely different realizations are supposed to be. Thus it should be possible to relate the ranking function to a distribution function. In fact, in some of the examples we have already made use of pdfs when constructing the ranking functions. In Example 2.5 we used the ranking function $p(\theta, v) = \mathcal{N}(v; 0, 0.1)$. This can be interpreted as a pdf for v. In that example, though, it is important to realize that $p(\theta, v)$ is not a pdf since

$$\int\int p(\theta, v) dv d\theta = +\infty$$

The reason for this is that p is not depending on θ at all, i.e. we are not providing any preference for different values of θ. This means that when we study generalization properties of our model, θ will be kept constant. We conclude that through a (partial) probabilistic description of the model parameters we obtain a ranking function at the same time. Thus the ranking function becomes a part of the model structure. We will call this a *probabilistic model structure*, a concept that will be formalized in the next chapter. We also note that by specifying the model parameters as random variables, we also impose relations between the observations $\mathbf{z}$ in one experiments. We can use this feature to construct ranking functions by way of measuring how well different models can predict the observed relations.

Secondly, recall that we in Section 2.5.3 discussed how to tune the ranking function to the data. The way this was accomplished was to parametrize the ranking function with hyperparameters and then to determine these using observations. In a probabilistic setting this means that we parametrize the distribution of the model parameters and then determine these hyperparameters using the observations.

Thirdly, if we select a model selection function that only use certain properties of the ranking function, then the exact specification of the distribution is not so important. An example is the use of the least-squares method, which can be motivated for many distributions. Formally, this is the problem of robust estimation.

Decision making under uncertainty - Expected regret. Probabilistic models of our observations allow us to further develop the concept of decision making under uncertainty. Returning to Section 2.4.3, $\rho^*(\boldsymbol{\xi})$ was defined as the optimal policy when the observations are generated by the model $\boldsymbol{\xi}$. With $\boldsymbol{\xi}$ unknown and instead using the certainty equivalence policy of replacing the unknown $\boldsymbol{\xi}$ by the output from a model selection function $\hat{\boldsymbol{\xi}}(\mathbf{z})$, the regret $L(\rho^*(\hat{\boldsymbol{\xi}}(\mathbf{z}), \boldsymbol{\xi})$ becomes a random variable. We can now ask the hypotetical question what the average regret would be if we indefinitely repeat the process of performing an experiment where we collect new observations which are subsequently used to compute the model based policy $\rho^*(\hat{\boldsymbol{\xi}}(\mathbf{z}))$, which is then applied to the system. This is the *expected regret*

$$\bar{L}(\boldsymbol{\rho}^*(\hat{\boldsymbol{\xi}})) := \mathbb{E}\left[L(\boldsymbol{\rho}^*(\hat{\boldsymbol{\xi}}(\mathbf{z}), \boldsymbol{\xi})\right]$$

where the expectation is over the probability distribution of $\boldsymbol{\xi}$, i.e.

$$\bar{L}(\boldsymbol{\rho}^*(\hat{\boldsymbol{\xi}})) = \int L(\boldsymbol{\rho}^*(\hat{\boldsymbol{\xi}}(\mathbf{M}(\boldsymbol{\xi})), \boldsymbol{\xi})p(\boldsymbol{\xi})d\boldsymbol{\xi}$$

We can thus compare different model structure selection functions with respect to the expected regret. But we can also define optimality of a model selection function with respect to the expected regret $\bar{L}(\rho^*(\hat{\boldsymbol{\xi}}))$. However, there is no particular reason for restricting the family of policies over which we optimize to optimal policies ρ^* when $\boldsymbol{\xi}$ is known. Instead we can consider policies $\boldsymbol{\rho}$ that are functions directly of data instead

$$\begin{aligned} \bar{\boldsymbol{\rho}}(\cdot) &:= \arg\min_{\rho(\cdot)} \; \bar{L}(\boldsymbol{\rho}(\cdot)) \\ \bar{L}(\boldsymbol{\rho}(\cdot)) &:= \mathbb{E}\left[L(\rho(\mathbf{z}), \boldsymbol{\xi})\right] \end{aligned}$$

This is a functional minimization problem which can be difficult to solve. However, it can be solved pointwise, i.e. $\bar{\rho}(\mathbf{z})$ can be solved for each $\mathbf{z}$, and for certain certain regrets an analytic solution exists for this problem. The reader is encouraged to relate (2.37) to the above problem.

Above we have (implicitly) assumed that all model parameters are integrated out when $\bar{L}$ is computed, otherwise we end up with a policy that we cannot use since it is then a function of unknown parameters. This means that two scenarios are covered by the approach above:

i) all model parameters are considered to vary from experiment to experiment according some probability distribution $p(\boldsymbol{\xi})$, or

ii) parameters that do not vary from experiment to experiment are known in advance.

The first scenario is relevant when considering populations of systems sharing characteristics but having individual variations, e.g. a fleet of similar vehicles or a population of animals. It can also apply if the device of interest changes over time. However, often decision making concerns an individual device which may not change over time. Scenario ii) is very limiting for decision making for such settings. Thus the analysis needs to be extended so that it can handle model parameters that are fixed but unknown. The worst-case approach of Section 2.4.3 can then be used but we can also use the expected regret. Let $\bar{L}(\boldsymbol{\rho}, \boldsymbol{\theta})$ denote the expected regret for the system with $\boldsymbol{\theta}$ as fixed model parameters, and with the policy $\boldsymbol{\rho}$. The optimal policy with respect to the expected regret is given by

$$\boldsymbol{\rho}^*(\boldsymbol{\theta}) := \arg\min_{\rho} \bar{L}(\boldsymbol{\rho}, \boldsymbol{\theta})$$

Now we let $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{z})$ be a model selection function for the fix elements of $\boldsymbol{\xi}$. Using the policy $\hat{\boldsymbol{\rho}} := \rho^*(\boldsymbol{\theta}(\mathbf{z}))$ gives the regret $\bar{L}(\hat{\boldsymbol{\rho}}, \boldsymbol{\theta})$ which depends on the particular observation $\mathbf{z}$ we have used.

Example 2.20. *Consider a non-linear state space model*

$$\begin{aligned}\mathbf{x}(t+1) &= f(\mathbf{x}(t), \mathbf{u}(t), \mathbf{w}(t), \boldsymbol{\theta})\\ \mathbf{y}(t) &= h(\mathbf{x}(t), \mathbf{e}(t), \boldsymbol{\theta})\end{aligned}$$

where the disturbance $\{\mathbf{w}(t)\}$ and the noise $\{\mathbf{e}(t)\}$ are modeled as independent random variables with known pdfs, and where $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$ is unknown. The application is to design a feedback controller

$$\mathbf{u}(t) = g(\mathbf{y}^t, \boldsymbol{\rho}) \tag{2.43}$$

where g is a given function parametrized by $\boldsymbol{\rho} \in \mathbb{R}^{n_\rho}$, which is to be chosen. The signals $\mathbf{y}$ and $\mathbf{u}$ thus both depend on both $\boldsymbol{\rho}$ and $\boldsymbol{\theta}$: $\mathbf{y}(t) = \mathbf{y}(t; \boldsymbol{\rho}, \boldsymbol{\theta})$ and $\mathbf{u}(t) = \mathbf{u}(t; \boldsymbol{\rho}, \boldsymbol{\theta})$. The reward is

$$R(\rho, \boldsymbol{\theta}) = \lim_{N\to\infty} -\sum_{t=1}^{N} \left(|\mathbf{y}(t; \boldsymbol{\rho}, \boldsymbol{\theta})|^2 + r|\mathbf{u}(t; \boldsymbol{\rho}, \boldsymbol{\theta})|^2\right), \quad r > 0$$

for which we denote the optimal policy $\rho(\boldsymbol{\theta})$.

Under some (stationarity) assumptions on the disturbance and noise, the reward equals the expected reward

$$\bar{R}(\rho, \boldsymbol{\theta}) = -\mathbb{E}\left[|\mathbf{y}(t; \boldsymbol{\rho}, \boldsymbol{\theta})|^2\right] - r\mathbb{E}\left[|\mathbf{u}(t; \boldsymbol{\rho}, \boldsymbol{\theta})|^2\right]$$

giving the expected regret

$$\bar{L}(\rho, \boldsymbol{\theta}) = \bar{R}(\rho(\boldsymbol{\theta}), \boldsymbol{\theta}) - \bar{R}(\rho, \boldsymbol{\theta})$$

Using $\rho^*(\hat{\boldsymbol{\theta}}(\mathbf{z}))$, where $\hat{\boldsymbol{\theta}}$ is a model selection function in the controller, gives the expected regret $\bar{L}(\rho^*(\hat{\boldsymbol{\theta}}(\mathbf{z})), \boldsymbol{\theta})$, which is the loss in performance obtained by collecting data $\mathbf{z}$ from one experiment, then computing the controller parameter $\rho(\hat{\boldsymbol{\theta}}(\mathbf{z}))$ and applying this controller in a new *infinitely long experiment.*

To get some further insight into how using a data dependent policy affects the regret, let us make the second order expansion

$$\bar{L}(\boldsymbol{\rho}^*(\hat{\boldsymbol{\theta}}), \boldsymbol{\theta}) = \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \rho_{\boldsymbol{\theta}}^T(\gamma) \bar{L}_{\rho\rho}(\gamma, \boldsymbol{\theta}) \rho_{\boldsymbol{\theta}}(\gamma)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \tag{2.44}$$

for some γ between $\hat{\boldsymbol{\rho}}$ and $\rho^*(\boldsymbol{\theta})$. Here we have made use of that $\bar{L}(\boldsymbol{\rho}^*(\boldsymbol{\theta}), \boldsymbol{\theta}) = 0$ and that

$$\bar{L}_\rho(\boldsymbol{\rho}(\boldsymbol{\theta}), \boldsymbol{\theta}) = 0$$

The sensitivity $\boldsymbol{\rho}_{\boldsymbol{\theta}}$ can be obtained from the previous expression as it implies

$$\bar{L}_{\rho\rho}(\boldsymbol{\rho}(\boldsymbol{\theta}), \boldsymbol{\theta})\boldsymbol{\rho}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + \bar{L}_{\rho\boldsymbol{\theta}}(\boldsymbol{\rho}(\boldsymbol{\theta}), \boldsymbol{\theta}) = 0$$

giving

$$\boldsymbol{\rho}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = -\bar{L}_{\rho\rho}^{-1}(\boldsymbol{\rho}(\boldsymbol{\theta}),\boldsymbol{\theta})\bar{L}_{\rho\theta}(\boldsymbol{\rho}(\boldsymbol{\theta}),\boldsymbol{\theta}) \tag{2.45}$$

which inserted in (2.44) and assuming that the error $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ is small, gives that the regret approximately is given by

$$\begin{aligned}&\bar{L}(\boldsymbol{\rho}^*(\hat{\boldsymbol{\theta}}),\boldsymbol{\theta}) \approx \\ &\frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \bar{L}_{\theta\rho}(\boldsymbol{\rho}^*(\boldsymbol{\theta}),\boldsymbol{\theta})\bar{L}_{\rho\rho}^{-1}(\boldsymbol{\rho}^*(\boldsymbol{\theta}),\boldsymbol{\theta})\bar{L}_{\rho\theta}(\boldsymbol{\rho}^*(\boldsymbol{\theta}),\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\end{aligned} \tag{2.46}$$

This expression is not very transparent but we can overbound this expression using Schur complement (see Appendix A.4). Assuming $J_{\rho,\rho} > 0$,

$$0 \le \begin{bmatrix} J_{\rho,\rho} & J_{\rho,\theta} \\ J_{\theta,\rho} & J_{\theta,\theta} \end{bmatrix} \Leftrightarrow J_{\theta,\theta} \ge J_{\theta,\rho} J_{\rho,\rho}^{-1} J_{\rho,\theta}$$

which then implies that

$$\bar{L}(\boldsymbol{\rho}^*(\hat{\boldsymbol{\theta}}),\boldsymbol{\theta}) \le \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \bar{L}_{\theta\theta}(\boldsymbol{\rho}^*(\boldsymbol{\theta}),\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \tag{2.47}$$

This shows the natural result that the regret will be small if the error $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ is small in the directions where the regret is sensitive to $\boldsymbol{\theta}$, and this irrespective of how the policy $\boldsymbol{\rho}$ depends on $\boldsymbol{\theta}$.

When performing a new experiment on the same system (where $\boldsymbol{\theta}$ remains the same but the other model parameters are generated from the underlying probability distribution) a new observation $\mathbf{z}$ will be obtained and then a new expected regret $\bar{L}(\boldsymbol{\rho}^*(\hat{\boldsymbol{\theta}}(\mathbf{z})),\boldsymbol{\theta})$ will be obtained as the policy $\boldsymbol{\rho}^*(\hat{\boldsymbol{\theta}}(\mathbf{z}))$ is a function of the observation $\mathbf{z}$. We can then also take the expectation with respect to the observation $\mathbf{z}$ used in the policy in order to quantify the average regret one would experience by repeating the procedure of experiments that generate observations $\mathbf{z}$ used by the policy $\boldsymbol{\rho}^*(\hat{\boldsymbol{\theta}}(\mathbf{z}))$, and using this policy in the application and measuring the expected regret $\bar{L}(\boldsymbol{\rho}^*(\hat{\boldsymbol{\theta}}(\mathbf{z})),\boldsymbol{\theta})$. The approximations above give that

$$\begin{aligned}&\mathbb{E}\left[\bar{L}(\boldsymbol{\rho}^*(\hat{\boldsymbol{\theta}}(\mathbf{z})),\boldsymbol{\theta})\right] \approx \\ &\frac{1}{2}\text{Trace}\left\{\bar{L}_{\theta\rho}(\boldsymbol{\rho}^*(\boldsymbol{\theta}),\boldsymbol{\theta})\bar{L}_{\rho\rho}^{-1}(\boldsymbol{\rho}^*(\boldsymbol{\theta}),\boldsymbol{\theta})\bar{L}_{\rho\theta}(\boldsymbol{\rho}^*(\boldsymbol{\theta}),\boldsymbol{\theta})\mathbf{MSE}_{\hat{\theta}}\right\} \\ &\le \frac{1}{2}\text{Trace}\left\{\bar{L}_{\theta\theta}(\boldsymbol{\rho}^*(\boldsymbol{\theta}),\boldsymbol{\theta})\mathbf{MSE}_{\hat{\theta}}\right\}\end{aligned} \tag{2.48}$$

where $\mathbf{MSE}_{\hat{\theta}}$ is the MSE of $\hat{\boldsymbol{\theta}}$, see (2.42),

$$\mathbf{MSE}_{\theta} = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\right]$$

2.7 *Inspiring Pitfalls*

At this point the reader should have a general idea of methods that can be used for system identification, although apart from simple examples, we have not given so much details. As often is the case, to get things to work the devil is in the details so as a motivation for continuing studying the theory of system identification, we in this section illustrate some of the, perhaps unexpected, problems that one may encounter. Later on we will see how the problems can be resolved. In all examples the models are estimated by adjusting the model parameters such that the simulated model output $\hat{y}(t)$ fits the output measurements $y(t)$ as well as possible according to the quadratic error criterion

$$\sum_{t=1}^{N}(y(t)-\hat{y}(t))^2$$

This is thus known as a prediction error method.

Model simulation. We begin with a discrete time example where it is known that the true system is of first order.

Example 2.21. *Suppose that we have input-output samples $u(t)$ and $y(t)$, $t = 1,\dots,N$ from a first order discrete time system $G(q) = bq^{-1}/(1+fq^{-1}$. The parameters b and f may then be determined by minimizing*

$$\sum_{t=1}^{N}\left(y(t)-\frac{bq^{-1}}{1+fq^{-1}}u(t)\right)^2 \tag{2.49}$$

with respect to these. Using this approach for the data in Figure 2.17, gives a model whose frequency response well matches the true one as shown in Figure 2.18.

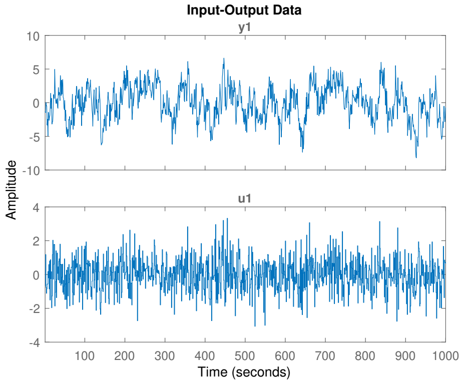

Figure 2.17: First data set in Example 2.21.

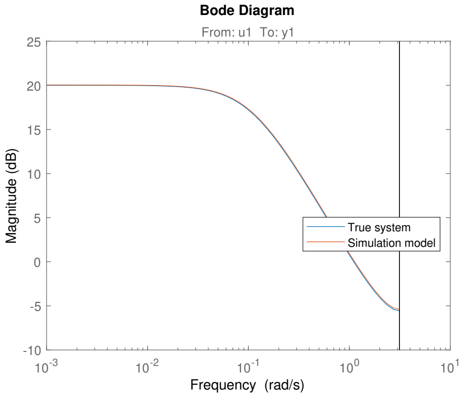

Figure 2.18: Bode diagram of model and true system.

Now, a downside with this approach is that the criterion is a high order multinomial in b and f and hence highly nonlinear in the decision variables. This, since the simulated output

$$\hat{y}(t;b,f) := \frac{bq^{-1}}{1+fq^{-1}}u(t)$$

is obtained by iterating the difference equation

$$(1+fq^{-1})\hat{y}(t;b,f) = bq^{-1}u(t),$$

that is

$$\hat{y}(t;b,f) = -f\hat{y}(t-1,b,f) + bu(t-1) \tag{2.50}$$

for $t = 1,2,\ldots,N$. Assuming $y(0) = u(0) = 0$ this gives

$$\begin{aligned}
\hat{y}(1;b,f) &= 0\\
\hat{y}(2;b,f) &= bu(1)\\
\hat{y}(3;b,f) &= -fbu(1) + bu(2)\\
\hat{y}(4;b,f) &= f^2bu(1) - fbu(2) + bu(3)\\
\hat{y}(5;b,f) &= -f^3bu(1) + f^2bu(2) - fbu(3) + bu(4)\\
&\vdots
\end{aligned}$$

An alternative would be to write the model

$$y(t) = \frac{bq^{-1}}{1+fq^{-1}}u(t)$$

as

$$(1+fq^{-1})y(t) = bq^{-1}u(t)$$

that is

$$y(t) = -fy(t-1) + bu(t-1)$$

Notice that, different from (2.50)*, here $y(t-1)$ does not come from the previous simulation step but is a data sample. Based on this model description we could then minimize the so-called equation error*

$$\sum_{t=1}^{N}(y(t) - fy(t-1) - bu(t-1))^2$$

Notice that this is a least-squares problem and thus there is an explicit expression for the minimizer. Trying this on the data in Figure 2.17 results in a model whose frequency response is given in Figure 2.19. Clearly, this approach, albeit much easier numerically, gives a much worse estimate.

Now, another data set generated by the same system with the same input is shown in Figure 2.20. Comparing with Figure 2.17 we see that the output has a slightly different characteristic. The reason for this is that there is some measurement errors on the output measurements and for this second data set these errors have a little bit different characteristics.

Now, repeating the exercise above of estimating a model both by minimizing the simulation error and the equation error results in models with frequency responses shown in Figure 2.21 and Figure 2.22. Comparing the two diagrams we see that for this data-set the equation error based model seems to give a more accurate model.

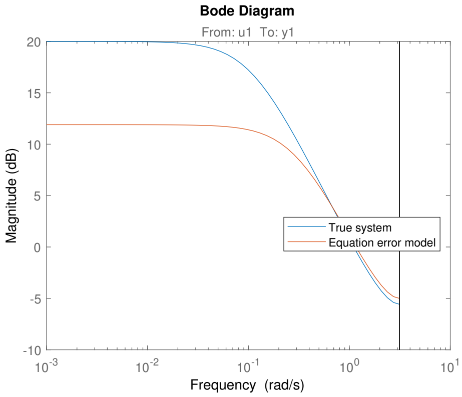

Figure 2.19: Bode diagram of model and true system.

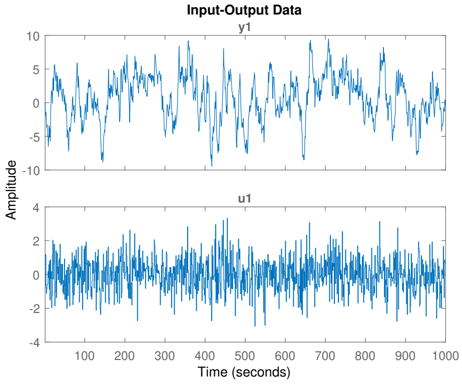

Figure 2.20: Second data set in Example 2.21.

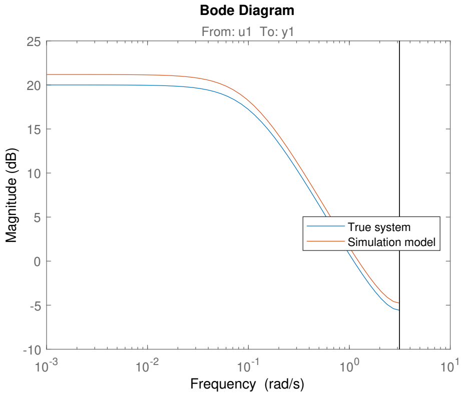

Figure 2.21: Bode diagram of model and true system.

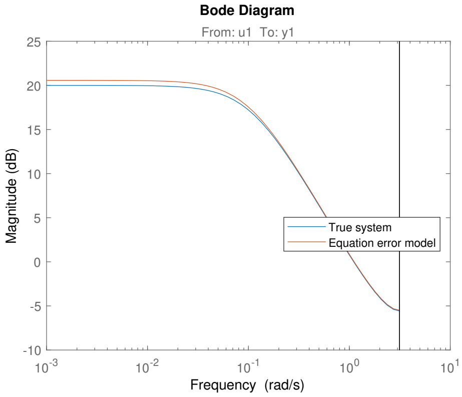

Figure 2.22: Bode diagram of model and true system.

Example 2.21 shows that "implementations" of the same model seem to give very different results. Is there a systematic approach to address this?

Closed loop experiments. Next we take a look at an identification problem where data is generated in closed loop.

Example 2.22. *An experiment configured as in Figure 2.23 is used to generate data which are used to identify the time discrete first order system $G_o(q) = 0.1q^{-1}/(1-0.9q^{-1})$, which has static gain 1. The signal $v(t)$ represents the measurement errors which is generated by white noise filtered through the low-pass filter H_o.*

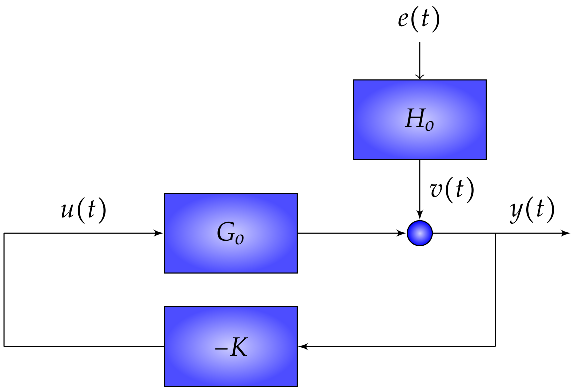

Figure 2.23: Experimental configuration in Example 2.22.

The data are shown in Figure 2.24.

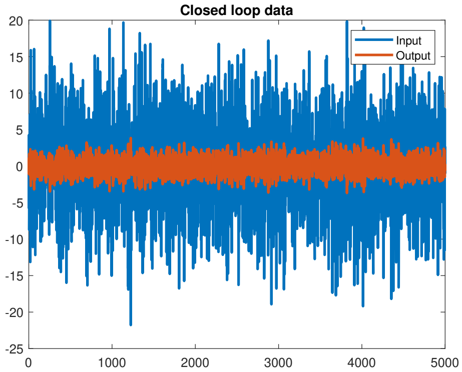

Figure 2.24: Input-output data in Example 2.22.

The parameters of the first order model $G(q) = bq^{-1}/(1+fq^{-1})$ are estimated by minimizing (2.49). The Bode diagram of the resulting model is shown in Figure 2.25, also including the true frequency response. Despite a lot of data the model is very poor.

In a second attempt the experiment is set-up as in Figure 2.26. This is thus an open-loop experiment. As input exactly the input that was used in the previous experiment is used. This results in the data shown in Figure 2.27.

The Bode diagram of the first order model identified using this data set is shown in Figure 2.28.

Clearly the model obtained from this experiment is much better.

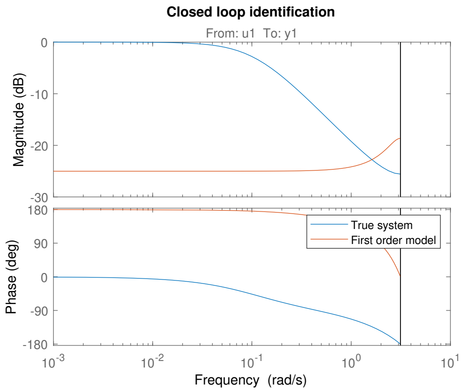

Figure 2.25: Bode diagram of estimated first order model compared with the true response.

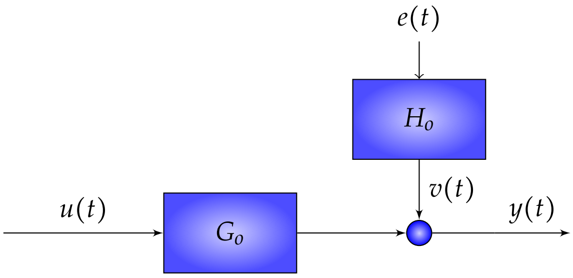

Figure 2.26: Second experimental setup in Example 2.22.

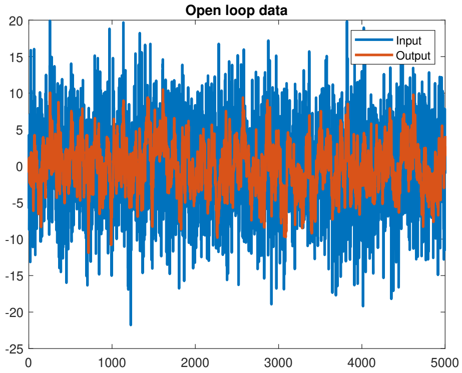

Figure 2.27: Input-output data in the open-loop experiment in Example 2.22.

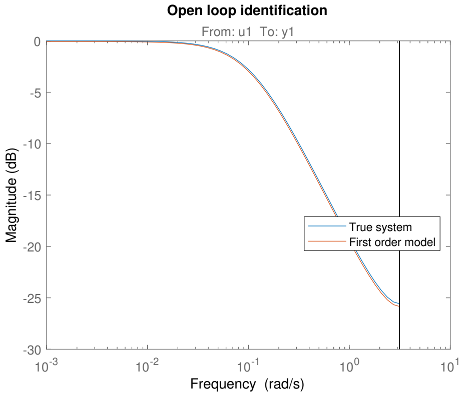

Figure 2.28: Bode diagram of first order model estimated from open loop data compared with the true response.

Example 2.22 illustrates that closed-loop identification can be intricate. A hint for why this is the case can be obtained by studying the close-up of the input-output signals given in Figure 2.29. At a first glance the output seems to react in the opposite direction of the input which suggests that the system has negative gain which is obviously the not the case as $G_o(q) = 0.1q^{-1}/(1-0.9q^{-1})$. However, from Figure 2.25 we see that also the model believes that the system has negative gain as the phase is 180^o. However, a closer look at Figure 2.29 reveals that the it is the output that causes the input to react in the opposite direction. It is thus the effect of the controller that is seen and apparently the model has the same problem as our eyes (at least the author's) to distinguish cause and effect. So the question arises if there are means to help distinguish the effects caused by the systems from those caused by the controller.

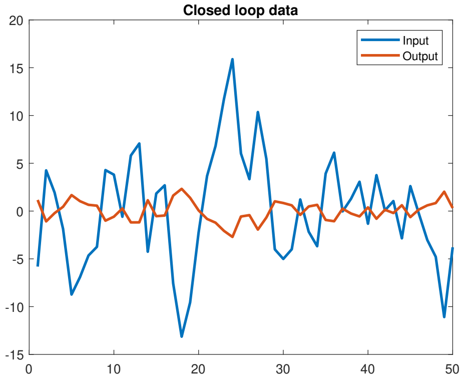

Figure 2.29: Close-up of input-output data in Example 2.22.

Sampling effects. Most systems are operating in continuous time, e.g. physical systems such as robots or vehicles. However, nowadays it is very rare to collect continuous measurements, instead digital recording devices equipped with ADCs (analog-to-digital converters) are used. This means that only sampled data is available. We shall be mostly concerned with the case where a fix sampling interval T is used and input-output samples $u(nT)$, $y(nT)$, $n = 1,\dots,N$ are collected. Sampling means loss of information and it is important to understand the effects of this. Another aspect is that sampling also means that data is quantized, i.e. only a finite number of amplitudes can be recorded. We will, however, disregard this aspect.

Example 2.23. *The system*

$$G(s) = \frac{1}{s+1} \tag{2.51}$$

is excited with the input in Figure 2.30, which results in the output given in Figure 2.31.

The sampled data are used to estimate the parameters $b_1,\dots,b_n$ and $f_1,\dots,f_n$ in models of the type

$$y(nT) = \frac{\sum_{k=1}^{n} b_k q^{-k}}{1+\sum_{k=1}^{n} f_k q^{-k}} u(nT)$$

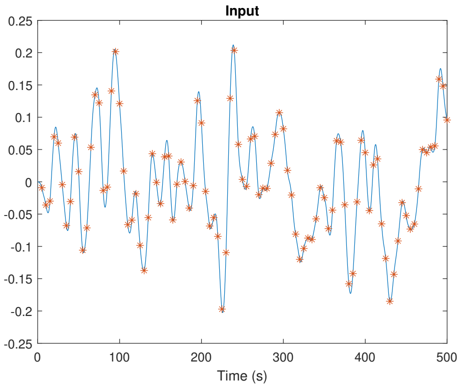

Figure 2.30: Example of a signal sampled with sampling interval $T = 5$ s.

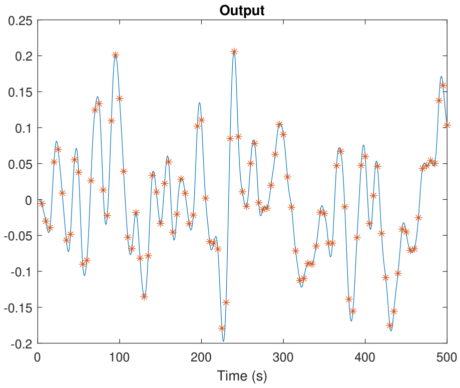

Figure 2.31: The output of (2.51) with the input given in Figure 2.30.

by minimizing

$$\sum_{n=1}^{N}\left(y(nT) - \frac{\sum_{k=1}^{n} b_k q^{-k}}{1+\sum_{k=1}^{n} f_k q^{-k}} u(nT)\right)^2$$

with respect to the parameters.

The Bode diagrams of the resulting models for $n = 1$ and $n = 3$ are shown in Figure 2.32.

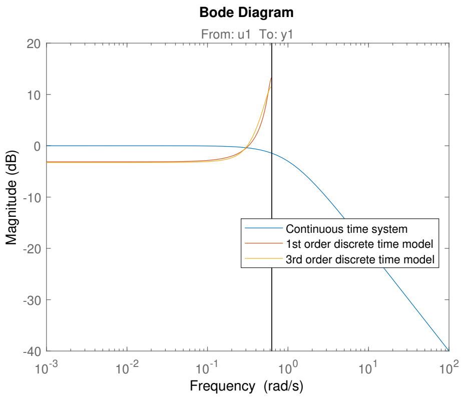

Figure 2.32: The frequency responses of the estimated models of order 1 and 3 compared with that of the true system.

Clearly, the discrete time system has a frequency response very different from the continuous time system despite that the sampling frequency to the eye looks reasonable.

The simulated output of the model is compared with the sampled output from the true system in Figure 2.33. As can be seen the identified model does a fair job of representing the system behaviour at the sampling points.

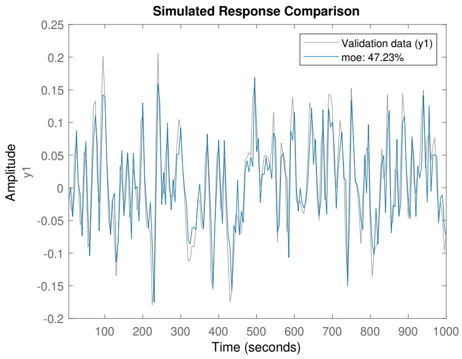

Figure 2.33: A comparison between the simulated output of the identified model and the true output.

Another experiment is set-up where the input instead is piecewise constant with identical samples as in the previous experiment. The input-output data are shown in Figures 2.34–2.35.

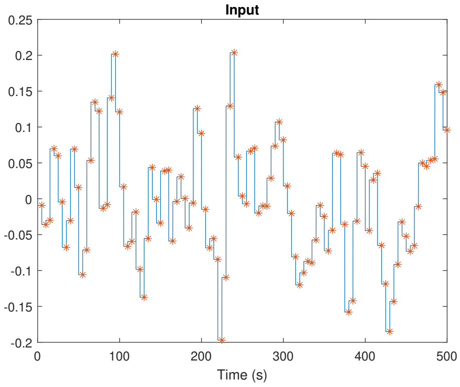

Figure 2.34: Piece-wise constant input.

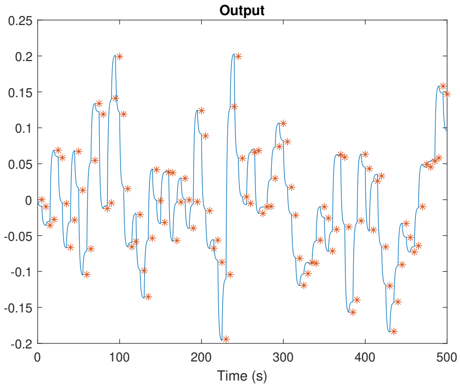

Figure 2.35: The output of (2.51) with the input given in Figure 2.34.

The Bode diagram of the first order model identified with the sampled data from this experiment is given in Figure 2.36. Clearly it shows more resemblance to the continuous time system than the previous model. But even more interesting is that when the model is simulated the output matches the sampled output exactly as shown in Figure 2.37.

Example 2.23 prompts a number of questions such as: i) Can the sampled data behaviour when the input is piece-wise constant always be represented exactly by a discrete time model, and if so, what is the relationship between these models. ii) Why is the first model so poor despite that the sampling frequency seems reasonable? iii)

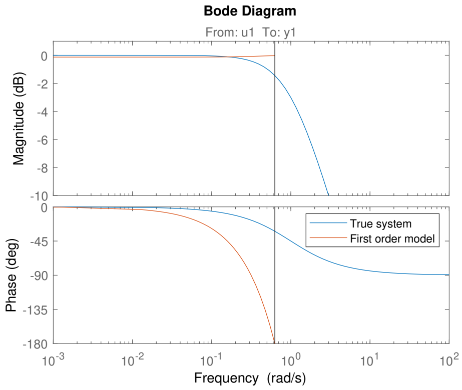

Figure 2.36: Identified first order model when the input is piece-wise constant.

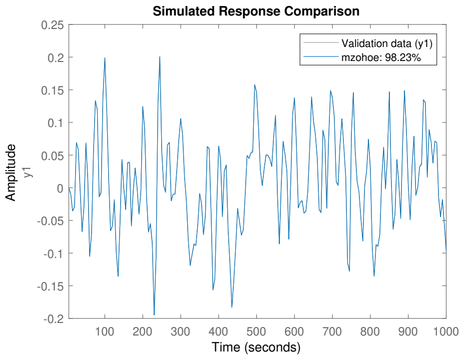

Figure 2.37: A perfect match between the simulated output and the sampled output when the input is piece-wise constant.

How should the sampling frequency be chosen? iv) How can the sampled data be used to recover the original continuous time system in the two cases that were studied? For example, is there a 1-1 mapping between the continuous time system and the discrete time model?

Measurement errors.

Example 2.24. *Suppose that one would like to identify the first-order LTI block G_2 in Figure 2.24.*

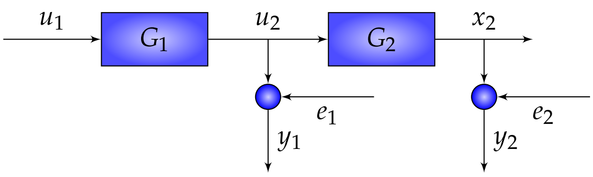

Figure 2.38: System configuration in Example 2.24.

A big data set is available consisting of 100.000 samples of the input $u(t)$, and the two outputs $y_1(t)$. The first 100 samples are plotted in Figure 2.24. The output measurements are corrupted by measurement noise e_1 and e_2.

Also the LTI block G_1 is unknown and hence the standard approach would be to identify both G_1 and G_2 from the data. However, G_1 is of high order which makes this approach somewhat difficult from a computational point of view since many parameters have to be identified.

Now, if in addition u_2 would be available, one could simply use this signal and y_2 to identify a first-order model for G_2. The Bode diagram of

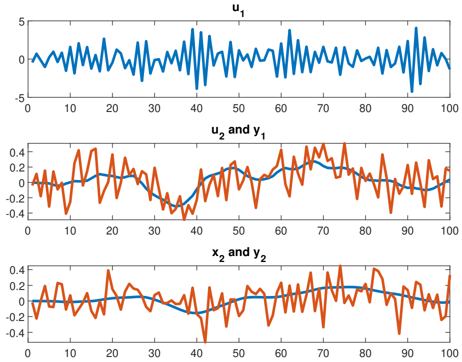

Figure 2.39: Data in Example 2.24.

the resulting model is shown in Figure 2.24. Since the data set is big the model matches the true system almost perfectly.

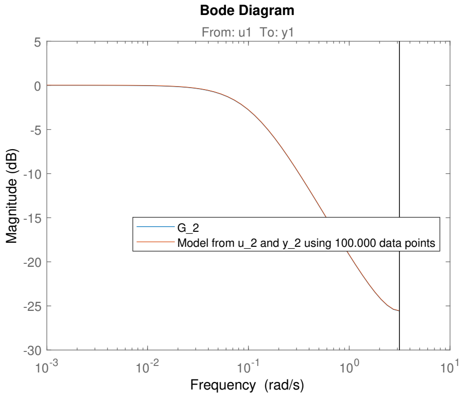

Figure 2.40: Bode diagram of identified models in Example 2.24.

In the case where u_2 is not accessible we observe that y_1 is a measurement of u_2 and as seen in the middle plot of Figure 2.24 this signal is quite representative of u_2. Thus it is tempting to replace u_2 with y_1 as input when identifying G_2. Trying this results in a model with the Bode diagram in Figure 2.24. We see that despite the big data set there is a rather substantial error. Doubling the number of samples to 200.000 does not diminish the error as also shown in the figure.

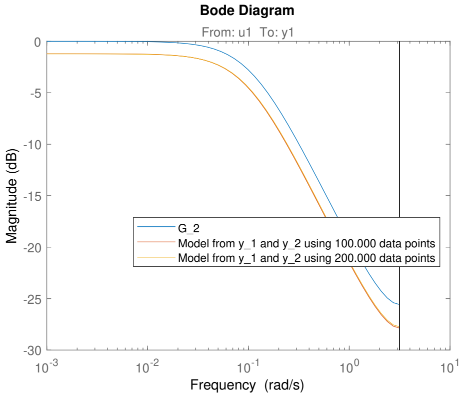

Figure 2.41: Bode diagram of identified models in Example 2.24.

Example 2.24 shows that there seems to be a significant difference between having measurement errors on the outputs or on the inputs. When only the output was corrupted a very good model could be achieved but this was not possible when the input was corrupted, despite increasing the sample size even further. A relevant question is if this is a fundamental problem or if it can be remedied in some way?

Complex models. In our final example the system is of relatively high order.

Example 2.25. *A system with known order of 25 is to be identified and the relatively large data set in Figure 2.25 has been collected. Included is also the noise free output and as can be seen the noise level is low.*

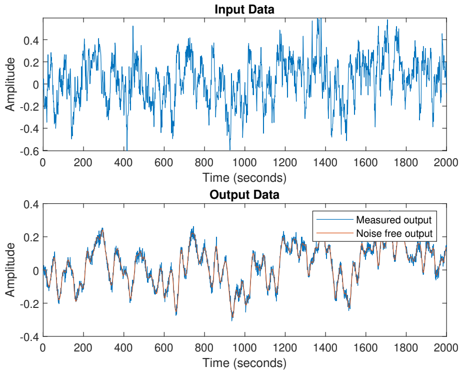

Figure 2.42: Data used in Example 2.25.

Identification of a 25th order model using a state-of-the art algorithm results in a model with Bode diagram given in Figure 2.25. Comparing with the true system response we see that the model is very poor despite apparently data of good quality.

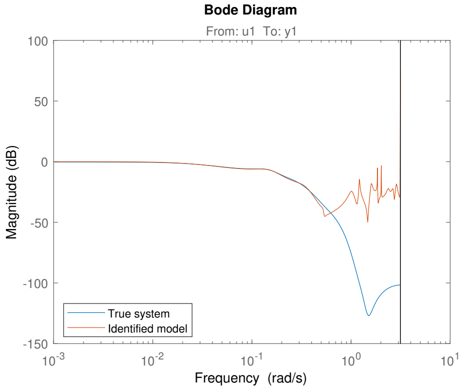

Figure 2.43: Bode diagram for the identified model in Example 2.25.

A potential problem in Example 2.25 is the high order of the model. An order of 25 means that 50 parameters (25 each in the numerator and denominator polynomials) need to be identified. With the optimization problem being non-convex, cf. Example 2.21, there is then a large risk that a local search algorithm ends up in a local minimum. Determining good initial parameters is thus an important problem in identification. To examine if this can be the case in this example, we modify the identification algorithm so that it is initialized at the *true value* of the parameters. The Bode diagram of the resulting model is shown in Figure 2.7. While we see some improvement compared to the model in Figure 2.25, there is still some problem. Thus the issue in this example is not only that it is difficult to find good initial values.

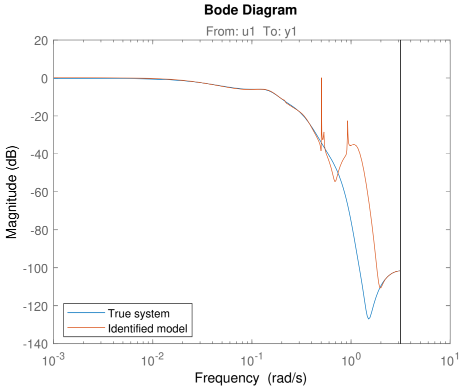

Figure 2.44: Bode diagram for the identified model in Example 2.25.

2.8 *Summary*

2.9 *Exercises*

2.1. Consider the setting in Example 2.6. Suppose that the ranking function is instead chosen as $p(\boldsymbol{\theta}, \mathbf{v}) = \mathcal{N}(\boldsymbol{\theta}; 0, \mathbf{P})\mathcal{N}(\mathbf{v}; 0, \lambda \mathbf{I})$.

a) Determine a closed form expression for the top ranked model. Compare with the top ranked estimate obtained in Example 2.6.

b) Determine a closed form expression for the conditional average ranked model.

2.2. Consider the model

$$\mathbf{z} = \boldsymbol{\varphi}_1 \theta_1 + \boldsymbol{\varphi}_2 \theta_2 + \mathbf{v}, \ \theta_1 \in \mathbb{R}, \ \theta_2 \in \mathbb{R}$$

where we use the ranking $p(\boldsymbol{\theta}, \mathbf{v}) = \mathcal{N}(\mathbf{v}; 0, \lambda \mathbf{I})$, where $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix}^T$. Suppose that our interest is in determining $\gamma(\boldsymbol{\theta}) = \theta_1 + \theta_2$ given $\mathbf{z} \in \mathbb{R}^N$. Compute $\gamma(\hat{\boldsymbol{\theta}}(\mathbf{z}))$, where $\hat{\boldsymbol{\theta}}(\mathbf{z})$ is the top ranked model of $\boldsymbol{\theta}$, and the top ranked γ given by $\hat{\gamma}(\mathbf{z}) = \arg\max_\gamma p(\mathbf{z}, \gamma)$. Are they equal?

2.3. UBB with input.

2.4. Consider the state-space model

$$\begin{aligned} x(t+1) &= \theta x(t) + w(t), \ x(0) = 0 \\ y(t) &= x(t) + v(t) \end{aligned}$$

a) Show that $\mathbf{z} = \begin{bmatrix} y(1) & \dots & y(N) \end{bmatrix}^T$ can be written as

$$\mathbf{z} = \mathbf{F}(\theta)\mathbf{w} + \mathbf{v}, \ \mathbf{w} = \begin{bmatrix} w(0) & \dots & w(N-1) \end{bmatrix}^T, \ \mathbf{v} = \begin{bmatrix} v(1) & \dots & v(N) \end{bmatrix}^T$$

for a suitably chosen matrix $\mathbf{F}(\theta)$.

b) The model parameters are $\boldsymbol{\xi} = \begin{bmatrix} \theta & \mathbf{w}^T & \mathbf{v}^T \end{bmatrix}^T$. Consider the ranking $p(\boldsymbol{\xi}) = \mathcal{N}(\mathbf{v}; 0,)\mathcal{N}(\mathbf{w}; 0,)$. Derive closed form expressions the top-ranked models for $\boldsymbol{\xi}$ and θ, respectively, where the latter is considered as a function of $\boldsymbol{\xi}$. Compare the element in the top-ranked model of $\boldsymbol{\xi}$ corresponding to θ, and compare it with the expression you have obtained for the top ranked model for θ. Are they the same?

2.5. Consider the setting in Example 2.6. Consider λ as an hyperparameter and estimate this parameter by the conditional average (2.32).

Hint: Show first that $p(\lambda|\mathbf{z})$ is an inverse gamma distribution. The pdf of an inverse gamma distribution with parameters α and β is given by

$$\text{Inv-Gamma}(x; \alpha, \beta)\frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{x^{\alpha+1}} e^{-\frac{\beta}{x}}$$

3
Probabilistic Models for Dynamic Systems

In this chapter we will introduce a formalism for probabilistic models of dynamic systems. We will use the nomenclature from probability theory and statistics and thus rename some of the concepts we introduced in the previous chapter[1]. However, we emphasize that what we will be doing is in essence playing a mind-game to obtain alibis for different ranking functions.

3.1 Models and model structures

As before we will denote by $\mathbf{z}(t) \in \mathbb{R}^{n_z}$ the vector of measurements that are obtained at time t. A standard set-up is that $\mathbf{z}(t)$ consist of both a vector of output measurements $\mathbf{y}(t)$ and a vector of input measurements $\mathbf{u}(t)$: $\mathbf{z}(t) = \begin{bmatrix}\mathbf{y}^T(t) & \mathbf{u}^T(t)\end{bmatrix}^T$. But other settings are possible as well, e.g. in blind identification only the outputs are measured and in missing data problems the dimension of $\mathbf{z}(t)$ may vary.

We will need to keep track of time and sample size so we introduce the notation

$$\mathbf{z}^t = \begin{bmatrix}\mathbf{z}(1)\\ \vdots \\ \mathbf{z}(t)\end{bmatrix} \in \mathbb{R}^{n_z t}$$

As in the previous section, a model will be associated with a set of model parameters. However, to account for arbitrary long experiments it will be a sequence rather than a vector. We will also add a probabilistic description to allow for inference of prolonged and new experiments.

Definition 3.1.1. *A model parameter $\boldsymbol{\xi} \in \Xi := \Xi(0) \times \Xi(1) \times \ldots$, is a sequence of the form $\boldsymbol{\xi} = \{\boldsymbol{\xi}(t)\}_{t=0}^{\infty}$, where $\boldsymbol{\xi}(t) \in \Xi(t) \subseteq \mathbb{R}^{n_{\xi_t}}$.*

For a model parameter $\boldsymbol{\xi}$, define[2]

$$\boldsymbol{\xi}^t = \begin{bmatrix}\boldsymbol{\xi}(0)\\ \vdots \\ \boldsymbol{\xi}(t)\end{bmatrix} \in \Xi^t \subseteq \mathbb{R}^{n_{\xi^t}},\ n_{\xi^t} := \sum_{k=0}^{t} n_{\xi_t}$$

where $\Xi^t \subseteq \mathbb{R}^{n_{\xi^t}}$ is defined in the obvious way that the kth block of element belongs to $\Xi(k-1)$.

[1] Sorry about this!

[2] For a model parameter the definition of $\boldsymbol{\xi}^t$ thus differ from the convention in that $\xi(0)$ is included as the first element.

A model structure $\mathcal{M}(\mathbf{M}.,\Xi)$ is a sequence of measurable functions defined on Ξ^t with the same range as the observations: $\mathcal{M}(\mathbf{M}.,\Xi) = \{\mathbf{M}_t : \Xi^t \to \mathbb{R}^{n_z}\}_{t=1}^{\infty}$.

For a model structure $\mathcal{M}(\mathbf{M}.,\Xi)$,

$$\mathbf{z}(t) = M_t(\boldsymbol{\xi}^t),\ t = 1,2,\dots \tag{3.1}$$

is a model of the observations $\{\mathbf{z}(t)\}_{t=1}^{\infty}$ corresponding to the model parameter $\boldsymbol{\xi}$.

The set of sequences

$$\left\{\{M_t(\boldsymbol{\xi}^t)\}_{t=1}^{\infty} :\ \boldsymbol{\xi}(t) \in \Xi(t)\right\}$$

is called the model set corresponding to the model structure $\mathcal{M}(\mathbf{M}.,\Xi)$.

Let $\{p_t : \Xi^t \to [0,\infty)\}$ be a sequence of pdfs determining a probability distribution for $\boldsymbol{\xi}^t$, related by that

$$\int_{\Xi(t)} p_t(\boldsymbol{\xi}^t) d\boldsymbol{\xi}(t) = p_{t-1}(\boldsymbol{\xi}^{t-1}),\ t = 1,2,\dots$$

When $\boldsymbol{\xi}$ is a realization from the probability distribution defined by $\{p_t\}_{t=1}^{\infty}$,

$$\mathbf{z}(t) = M_t(\boldsymbol{\xi}^t),\ t = 1,2,\dots$$

is called a realization of the observed signals for the model structure $\mathcal{M}(M.,p.)$.

With $\{M_t\}$, $\{p_t\}$ and $\{\Xi_t\}$ as above, $\mathcal{M} = \mathcal{M}(M.,\Xi.,p.)$ is called a probabilistic model structure.

A number of remarks are warranted at this point:

- The definitions above incorporate that we are dealing with dynamic systems, meaning that the present response depends on the entire past history, and also that the probability distributions for the past does not change when we advance in time.

- The first element in a model parameter, $\xi(0)$, may be used for initial conditions, for example. It may also be used for parameters that are constant over time.

- All pdfs p_t, $t = 1,2,\dots$, may be parametrized by a common hyperparameter $\boldsymbol{\eta} \in \mathbb{R}^{n_\eta}$: $p_t = p_t(\boldsymbol{\xi}^t,\boldsymbol{\eta})$. For realizations of $\boldsymbol{\eta}^N$ we indicate the dependence on hyperparameters by writing $\boldsymbol{\xi}(\boldsymbol{\eta})$.

- Notice that different from what we assumed in the previous chapter, p_t is a bona fide pdf. To exclude certain model parameters from the pdf, the method based on hyperparameters briefly discussed in Example 2.9 can be used.

- For a probabilistic model, the split between model parameters and hyperparameters is not unique. The model $M(\theta,v) = \theta v$ with probability distribution $p(\theta,v) = \mathcal{N}(v;0,1)\delta(\theta - \eta)$ is equivalent to the model $M(w) = w$ with the probability distribution $p(w) = \mathcal{N}(w;0,\eta^2)$.

- We will continue to call ξ a model and Ξ a model set.

The set of unfalsified models is defined similarly as in the previous chapter. We use the notation

$$M^N(\boldsymbol{\xi}^N) = \left[M_1^T\boldsymbol{\xi}^1) \quad \dots M_N^T(\boldsymbol{\xi}^N)\right]^T$$

Definition 3.1.2. *Given data $\mathbf{z}^N$, the set of unfalsified models for the model structure $\mathcal{M}(M., p.)$ is defined as*

$$\mathcal{U}(\mathbf{z}^N) = \left\{\boldsymbol{\xi} : \ M^N(\boldsymbol{\xi}^N) = \mathbf{z}^N\right\}$$

We use an example to illustrate how exact measurements can be modelled.

Example 3.1. *Consider the model*

$$y(t) = \theta u(t) + v(t)$$

where $\mathbf{z}(t) = \begin{bmatrix} y(t) & u(t) \end{bmatrix}^T$ are our observations. We can then take $\xi(0) = \theta$, $\xi(t) = \begin{bmatrix} v(t) & u(t) \end{bmatrix}^T$, $t = 1, 2, \dots$ and use the model structure

$$M_t(\xi^t) = \begin{bmatrix} \theta\xi(t) + v(t) \\ u(t) \end{bmatrix}$$

Then $u(t)$ will be equal to the second element of $\mathbf{z}(t)$ for all models in the set of unfalsified models.

3.2 *Probabilistic interpretations of ranking functions*

Now we will identify the ranking functions we discussed in the previous section with their probabilistic interpretations.

When the ranking function (2.11) is normalized so that it integrates to 1, it corresponds to the prior probability $p_N(\boldsymbol{\xi}^N)$ for the model parameters and the joint ranking function for $\boldsymbol{\xi}$ and the observation $\mathbf{z}$ (2.13) corresponds to the joint pdf for $\boldsymbol{\xi}^N$ and $\mathbf{z}^N$ given by

$$p_N(\boldsymbol{\xi}^N, \mathbf{z}^N) := p_N(\boldsymbol{\xi}^N) \prod_{t=1}^{N} \delta(\mathbf{z}(t) - M_t(\boldsymbol{\xi}(t)))$$

In Section 2.5.5 we defined the total rankings corresponding to a function γ of the model parameters $\boldsymbol{\xi}^N$ by integrating over the model parameters for which the function is constant[3]. Here this procedure results in the joint pdf for $\gamma(\boldsymbol{\xi}^N) = \gamma$ and $\mathbf{z}$:

$$p_N(\gamma, \mathbf{z}^N) := \int_{\Xi^N} p_N(\boldsymbol{\xi}^N, \mathbf{z}) \delta(\gamma - \gamma(\boldsymbol{\xi}^N)) d\boldsymbol{\xi}^N$$

Typically γ is a subset of $\boldsymbol{\xi}^N$. When all variables $\boldsymbol{\xi}^N$ are marginalized out we write $p_N(\mathbf{z}^N)$, and this corresponds to the total rankings for $\mathbf{z}^N$ as in (2.23).

The normalized ranking of $\boldsymbol{\xi}$ conditioned on $\mathbf{z}$ was defined in (2.18). This corresponds to the posterior pdf for $\boldsymbol{\xi}^N$ given $\mathbf{z}^N$

$$p_N(\boldsymbol{\xi}^N | \mathbf{z}^N) := \frac{p_N(\boldsymbol{\xi}^N, \mathbf{z}^N)}{p_N(\mathbf{z}^N)}$$

[3] Integrating out certain variables from a pdf is called marginalization.

When the pdfs $\{p_t\}$ are parametrized by hyperparameters $\boldsymbol{\eta}$, the pdfs above become functions of these as well and we indicate this by indexing the pdfs as $p_N(\boldsymbol{\xi}^N;\boldsymbol{\eta})$, $p_N(\boldsymbol{\xi}^N,\mathbf{z}^N;\boldsymbol{\eta})$, $p_N(\boldsymbol{\xi}^N|\mathbf{z}^N;\boldsymbol{\eta})$, and $p_N(\mathbf{z}^{;}\boldsymbol{\eta})$. We can also integrate out the hyperparameters from $p_N(\mathbf{z}^{;}\boldsymbol{\eta})$ giving

$$p_N(\mathbf{z}^N) = \int p_N(\mathbf{z}^N;\boldsymbol{\eta})d\boldsymbol{\eta}$$

and when this quantity is finite we can then define

$$\begin{aligned} p_N(\boldsymbol{\xi}^N,\boldsymbol{\eta}|\mathbf{z}^N) &:= \frac{p_N(\boldsymbol{\xi}^N,\mathbf{z}^N;\boldsymbol{\eta})}{p_N(\mathbf{z}^N)} \\ p_N(\boldsymbol{\eta}|\mathbf{z}^N) &:= \frac{p_N(\mathbf{z}^N;\boldsymbol{\eta})}{p_N(\mathbf{z}^N)} \end{aligned}$$

Due to the normalization these quantities are bona fide pdfs. Tough, this does not mean that we interpret $\boldsymbol{\eta}$ as a random vector. Nevertheless, in the spirit of interpreting the ps as ranking functions, we may use these pdfs to obtain estimates of both model parameters and hyperparameters. We may of course also marginalize $p_N(\boldsymbol{\xi}^N,\boldsymbol{\eta}|\mathbf{z}^N)$ and $p_N(\boldsymbol{\eta}|\mathbf{z}^N)$ with respect to some of the parameters.

3.2.1 *Non-informative priors*

It often happens that the user has very vague prior knowledge of certain model parameters. It is then tempting to use a flat prior, i.e. a pdf which is (almost) constant over a large region. However, this also injects some prejudice into the ranking and there is nothing like a non-informative ranking. Ignorance can be obtained by assuming that a parameter is unknown without specifying a pdf.

Example 3.2. *Consider the model*

$$\mathbf{z} = \boldsymbol{\theta}^3 + \mathbf{v}$$

where it is generally known that $\mathbf{v} \sim \mathcal{N}(0,1)$ *and that* $\boldsymbol{\theta} \in [-1,1]$.

To reflect the latter, User A assigns a uniform distribution on the interval $[-1,1]$ *to* $\boldsymbol{\theta}$.

However, User B thinks it is simpler to work with the model

$$\mathbf{z} = \boldsymbol{\rho} + \mathbf{v}$$

as there is a one-to-one relationship between $\boldsymbol{\rho}$ *and* $\boldsymbol{\theta}$. *Now as the only information about* $\boldsymbol{\rho}$ *is that it belongs to the interval* $[-1,1]$, *this user assigns a uniform distribution on the interval* $[-1,1]$ *to* $\boldsymbol{\rho}$.

Clearly, even though the two users try to encode the prior knowledge in an impartial way, the two users arrive at completely different priors just because they have chosen different parametrizations. The consequence of this is that if the two users use the same ranking function for model selection they will get different results, despite that what they both have done is simply to try to be impartial.

3.3 Estimators

We will from now on replace the use of the term model selection function with the commonly used term *estimator*.

Definition 3.3.1. *Given a model structure $\mathcal{M}(M., p., \Xi.)$, an estimator is a sequence of functions $\{\hat{\boldsymbol{\xi}}^t\}_{t=1}^{\infty}$*

$$\hat{\boldsymbol{\xi}}^t : \mathbb{R}^{n_{z_t}} \to \Xi^t \subseteq \mathbb{R}^{n_{\xi_t}}$$

We will frequently only be interested in certain elements of an estimator, and will then use the notation $\hat{\boldsymbol{\theta}}_t$

All model selection functions that we have discussed in Section 3.3.1 are meaningful estimators. In fact, many of them are of such interest that they have been baptized in the statistics literature.

3.3.1 Ranking based estimators

The model selection functions presented in Section 3.3.1 are all based on the ranking function.

Top ranked models. We introduced the top ranked model in (2.12), which we here can write

$$\hat{\boldsymbol{\xi}}^N(\mathbf{z}^N) = \arg\max_{\boldsymbol{\xi}^N \in \Xi^N} p_N(\boldsymbol{\xi}^N, \mathbf{z}^N)$$

This estimator is called the *Maximum A Posteriori* (MAP) estimator of $\boldsymbol{\xi}^N$ and we denote it by $\hat{\boldsymbol{\xi}}^N_{MAP}(\mathbf{z}^N)$. The name derives from that

$$p_N(\boldsymbol{\xi}^N, \mathbf{z}^N) = p_N(\boldsymbol{\xi}^N | \mathbf{z}^N) p_N(\mathbf{z}^N)$$

so that

$$\hat{\boldsymbol{\xi}}^N_{MAP}(\mathbf{z}^N) = \arg\max_{\boldsymbol{\xi}^N \in \Xi^N} p_N(\boldsymbol{\xi}^N | \mathbf{z}^N)$$

i.e. it is the point in Ξ^N having the highest posterior probability[4].

We can also compute MAP estimators of functions of the model parameters $\gamma(\boldsymbol{\xi}^N)$[5]

$$\hat{\gamma}_{MAP}(\mathbf{z}^N) = \arg\max_{\gamma} p_N(\gamma, \mathbf{z}^N)$$

which in general differs from $\gamma(\hat{\boldsymbol{\xi}}_{MAP}(\mathbf{z}^N))$, see Exercise 3.1. Notice that marginalization with respect to some parameters is obtained by taking γ to be the orthogonal projection onto the the variables that one would like to keep. For example, marginalization of $\boldsymbol{\xi}^N$ with respect to $\boldsymbol{\xi}(N)$ is obtained by taking $\gamma(\boldsymbol{\xi}^N) = \boldsymbol{\xi}^{N-1}$. MAP estimators based on marginalized pdfs will be called marginalized MAP (MMAP) estimators.

When the pdfs $\{p_t\}$ are parametrized by a hyperparameter $\boldsymbol{\eta}$ we indicate this in an estimator by indexing it by $\boldsymbol{\eta}$, e.g. $\hat{\boldsymbol{\xi}}^N_{MAP}(\mathbf{z}^N; \boldsymbol{\eta})$. We also need to provide estimators for the hyperparameters. It is

[4] In the sense of having the largest value of the pdf.

[5] Recall from Section 2.5.5 that $p_N(\gamma, \mathbf{z}^N) = \int_{\Xi^N} p_N(\xi^N, \mathbf{z}^N)\delta(\gamma - \gamma(\xi^N))d\xi^N$.

common to first eliminate the model parameters by marginalization so that an estimator for the hyperparameters $\hat{\boldsymbol{\eta}}(\mathbf{z}^N)$ can be obtained separately. As estimator for the model parameters one can then take $\hat{\boldsymbol{\xi}}^N(\mathbf{z}^N; \hat{\boldsymbol{\eta}}(\mathbf{z}^N))$.

In (2.31) we suggested the use of the hyperparameter that maximizes the total rankings (corresponding to $\mathbf{z}$). This function, $p_N(\mathbf{z}^N; \boldsymbol{\eta})$, is called the likelihood function and therefore this estimator is called the *maximum likelihood* (ML) estimator

$$\hat{\boldsymbol{\eta}}_{ML}(\mathbf{z}^N) := \arg\max_{\boldsymbol{\eta}} p_N(\mathbf{z}^N; \boldsymbol{\eta})$$

as it makes the pdf for the observations have its maximum at the point $\mathbf{z}$ where the observation was obtained. We can also here marginalize over certain of the parameters which results in marginalized ML (MML) estimators. For numerical reasons it turns out convenient to work with the *negative log likelihood* function $-\log p_N(\mathbf{z}^N; \boldsymbol{\eta})$, for which the ML-estimate is defined by the minimization problem[6]

$$\hat{\boldsymbol{\eta}}_{ML}(\mathbf{z}^N) := \arg\min_{\boldsymbol{\eta}} -\log p_N(\mathbf{z}^N; \boldsymbol{\eta})$$

For ease of notation, we will frequently rescale and drop constants in the negative log-likelihood function without commenting on this as these operations do not change the optimum.

The estimator (2.30)

$$\left(\hat{\boldsymbol{\xi}}^N(\mathbf{z}^N), \hat{\boldsymbol{\eta}}(\mathbf{z}^N)\right) := \arg\max_{\boldsymbol{\xi}^N \in \Xi^N, \boldsymbol{\eta}} p_N(\boldsymbol{\xi}^N, \mathbf{z}^N; \boldsymbol{\eta})$$

is called the *joint* MAP/ML estimator of $\boldsymbol{\xi}^N$ and $\boldsymbol{\eta}$. Also here, of course, one can first marginalize one or both of the parameters obtaining joint MMAP/ML, MAP/MML and MMAP/MML estimators.

The conditional average ranking model In (3.2) we introduced the conditional average ranking model, which we here can write

$$\hat{\boldsymbol{\xi}}_{PM}^N(\mathbf{z}^N) = \int_{\mathcal{U}(\mathbf{z}^N)} \boldsymbol{\xi}^N p_N(\boldsymbol{\xi}^N | \mathbf{z}^N) d\boldsymbol{\xi}^N, \tag{3.2}$$

This is the conditional mean of $\boldsymbol{\xi}^N$ given the observation $\mathbf{z}^N$, which we will refer to as the *posterior mean* (PM) estimator. As was discussed in Section 2.5.4, this estimator has the optimality property (2.36). Using that now p_t is a pdf, we can write this using expectations over $\boldsymbol{\xi}^N$, obtaining that for any estimator $\tilde{}^N$

$$\mathbb{E}\left[|\tilde{\boldsymbol{\xi}}^N(M^N(\boldsymbol{\xi}^N)) - \boldsymbol{\xi}^N|^2\right] \geq \mathbb{E}\left[|\hat{\boldsymbol{\xi}}_{PM}^N(M^N(\boldsymbol{\xi}^N)) - \boldsymbol{\xi}^N|^2\right]$$

Realizing that when the model is correct, $M^N(\boldsymbol{\xi}^N)$ is the observation $\mathbf{z}^N$ when the model parameters are $\boldsymbol{\xi}^N$, we can write this as

$$\mathbb{E}\left[|\tilde{\boldsymbol{\xi}}^N(\mathbf{z}^N) - \boldsymbol{\xi}^N|^2\right] \geq \mathbb{E}\left[|\hat{\boldsymbol{\xi}}_{PM}^N(\mathbf{z}^N) - \boldsymbol{\xi}^N|^2\right]$$

[6] Recall that the logarithm is a monotone function and thus does not change the location of optima.

We can interpret this inequality as that the posterior mean is the estimator that minimizes the *mean-squared error* (MSE) when we repeat the process of drawing a model parameter $\boldsymbol{\xi}^N$ from the distribution p_N and then estimate $\boldsymbol{\xi}^N$ from the observation $\mathbf{z}^N = M^N(\boldsymbol{\xi}^N)$. From (2.35) we also have that

$$\mathbb{E}\left[|\tilde{\boldsymbol{\xi}}^N(\mathbf{z}^N) - \boldsymbol{\xi}^N|^2|\mathbf{z}^N\right] \geq \mathbb{E}\left[|\hat{\boldsymbol{\xi}}^N_{PM}(\mathbf{z}^N) - \boldsymbol{\xi}^N|^2|\mathbf{z}^N\right]$$

meaning that $\bar{\boldsymbol{\xi}}^N_{PM}(\mathbf{z}^N)$ is the constant vector which minimizes the MSE if we restrict the draws of $\boldsymbol{\xi}^N$ to those that generate the observation $\mathbf{z}^N$. An important remark is that this interpretation is lost when estimated hyperparameters are used as $p(\boldsymbol{\xi}^N|\mathbf{z}^N;\hat{\boldsymbol{\eta}}(\mathbf{z}^N))$ cannot be interpreted as a posterior pdf of the model parameters. This does not prevent $\hat{\boldsymbol{\xi}}^N_{PM}(\mathbf{z}^N;\hat{\boldsymbol{\eta}}(\mathbf{z}^N))$ from potentially being a useful estimator.

The posterior mean of a function of the model parameters $\gamma(\boldsymbol{\xi}^N)$ is given by

$$\hat{\gamma}_{PM}(\mathbf{z}^N) = \int_{\mathcal{U}(\mathbf{z}^N)} \gamma(\boldsymbol{\xi}^N) p_N(\boldsymbol{\xi}^N|\mathbf{z}^N) d\boldsymbol{\xi}^N. \tag{3.3}$$

See also Exercise 3.1.b.

For an hyperparameter $\boldsymbol{\eta}$, the mean of $p_N(\boldsymbol{\eta}|\mathbf{z}^N)$ can be taken as estimator

$$\hat{\boldsymbol{\eta}}_{PM}(\mathbf{z}^N) = \mathbb{E}\left[\boldsymbol{\eta}|\mathbf{z}^N\right]$$

However, we stress again that this does not mean that we interpret $\boldsymbol{\eta}$ as a random vector, the expectation is just a compact way to write the integral

$$\hat{\boldsymbol{\eta}}_{PM}(\mathbf{z}^N) = \int \boldsymbol{\eta} p(\boldsymbol{\eta}|\mathbf{z}^N) d\boldsymbol{\eta}$$

3.3.2 *Predictive estimators*

In Section 3.3.1 we had a brief encounter with estimation theory. We saw that the mean of a random variable is the estimator of the random variable that minimizes the MSE when no observations are available. We also saw that the mean is optimal (with respect to the MSE) also when (indirect) observations are available, but then one should use the posterior mean, i.e. the conditional mean with respect to the observation.

Now, with a probabilistic model we can use this theory to construct estimators of model parameters as well as hyperparameters. The basic idea is to first choose some functions $f(\cdot)$ with the same domain as our observed data vector $\mathbf{z}^N$ lives in and then pick the model that is best able to *predict* (estimate) the function value $\mathbf{s} = f(\mathbf{z}^N)$ for the observation at hand $\mathbf{z}^N$. We will work with measurable functions, meaning that $\mathbf{s} = f(M^N(\boldsymbol{\xi}^N))$ becomes a random variable as it is a (measurable) function of the random variable $\boldsymbol{\xi}^N$. The output $\mathbf{s}$ is then called a *statistic*.

While this seems may seem attractive, two major questions arise: Which functions should be chosen, and how should we measure "best"? But perhaps the first question is, why do we need to use a function at all? Would it not be better to see which model is best able to predict the entire data set? Well, this we have already discussed. It leads to the set of unfalsified models and then one need to make a decision rule based on user preferences, e.g. by ranking, leading to methods such as ML and MAP. By using statistics and estimation theory we can verify how good the probabilistic description for a model is. Which functions and criteria to use are issues strongly linked to the resulting computational complexity, the accuracy that can be obtained and the robustness of the method[7]. One can also view the use of statistics to build estimators as a data compression step and a natural question is if and how this can be done without information loss. We will return to this in Section 5.5.

Below we will briefly discuss three common families of predictive estimators.

Moment estimators. Assuming for simplicity $\mathbf{z}(t) \in \mathbb{R}$, non-central sample moments such as

$$m_k(\mathbf{z}^N) = \frac{1}{N}\sum_{t=1}^{N} \mathbf{z}^k(t),\ k = 1, 2, \ldots$$

are common statistics. For a model $M^N(\boldsymbol{\xi}^N)$, the optimal estimator of $m_k(\mathbf{z}^N)$ is the corresponding non-central moment

$$m_k(\boldsymbol{\eta}) = \frac{1}{N}\sum_{t=1}^{N} \mathbb{E}\left[M_t^k(\boldsymbol{\xi}^t(\boldsymbol{\eta}))\right]$$

These moments will only depend on the hyperparameters and not the model parameters[8] as we take the mean with respect to $\boldsymbol{\xi}^t$. When as many moments as the number of hyperparameters n_η are used, setting the optimal estimators equal to the sample moments results in a set of equations

$$m_k(\mathbf{z}^N) = m_k(\boldsymbol{\eta}), \quad k = 1, \ldots, n_\eta$$

from which an estimator of $\boldsymbol{\eta}$ can be obtained. This approach is known as the *Method of Moments*.

One can also use a larger number of moments in which case an $\boldsymbol{\eta}$ cannot be found so that all moments match the sample moments. However, given that the moments are optimal in the MSE sense it may make sense to take the minimizer of

$$V(\boldsymbol{\eta}) = \begin{bmatrix} m_1(\mathbf{z}^N) - m_1(\boldsymbol{\eta}) \\ \vdots \\ m_K(\mathbf{z}^N) - m_K(\boldsymbol{\eta}) \end{bmatrix}^T \begin{bmatrix} m_1(\mathbf{z}^N) - m_1(\boldsymbol{\eta}) \\ \vdots \\ m_K(\mathbf{z}^N) - m_K(\boldsymbol{\eta}) \end{bmatrix}$$

as estimator of $\boldsymbol{\eta}$. It may also make sense to include a weighting matrix $\mathbf{W}$ to account for that the error in the different sample moments

[7] Typically, better accuracy or robustness cost more in terms of computations - one of the many facets of the no free lunch theorem

[8] recall, though, that constant model parameters are included in the hyperparameters as well (with a Dirac's delta function linking such model parameters and the corresponding hyperparameters in the pdf.

are different in size

$$V(\boldsymbol{\eta}) = \begin{bmatrix} m_1(\mathbf{z}^N) - m_1(\boldsymbol{\eta}) \\ \vdots \\ m_K(\mathbf{z}^N) - m_K(\boldsymbol{\eta}) \end{bmatrix}^T \mathbf{W} \begin{bmatrix} m_1(\mathbf{z}^N) - m_1(\boldsymbol{\eta}) \\ \vdots \\ m_K(\mathbf{z}^N) - m_K(\boldsymbol{\eta}) \end{bmatrix} \tag{3.4}$$

This approach is known as the *Generalized Method of Moments*. We will return to the choice of weighting matrix in Section 5.9.

Indirect inference. The generalized method of moments can also be derived in the following way. Let us start with the following unstructured model

$$\mathbf{z}(t) = \mathbf{v}(t) \tag{3.5}$$

where $\{\mathbf{v}(t)\}$ is a sequence of i.i.d. random variables where the first K non-central moments are hyperparameters $\tilde{\boldsymbol{\eta}}_k$, $k = 1, \ldots, K$. These hyperparameters are then estimated by the sample moments

$$\hat{\tilde{\boldsymbol{\eta}}}_k(\mathbf{z}^N) = m_k(\mathbf{z}^N)$$

It is not hard to come up with alibis for these estimators. The idea is now that if the model that we actually are interested in is correct then if we apply the same estimator as above to observations from the model we should obtain a similar estimate as when observations from the true system are used. Thus we should have

$$\hat{\tilde{\boldsymbol{\eta}}}_k(\mathbf{z}^N) \approx \hat{\tilde{\boldsymbol{\eta}}}_k(M^N(\boldsymbol{\xi}^{(}\beta))))$$

if the hyperparameter $\boldsymbol{\eta}$ in our model is correct. Now this is the same as

$$m_k(\mathbf{z}^N) \approx m_k(M^N(\boldsymbol{\xi}^N(\boldsymbol{\eta}))),\ k = 1, \ldots, K$$

We then make the further observation that $m_k(M^N(\boldsymbol{\xi}^N(\boldsymbol{\eta})))$ contain random variations that have nothing to do with the real system but are just caused by our random number generator. These we can remove by taking the expectation, giving that we should choose $\boldsymbol{\eta}$ such that

$$m_k(\mathbf{z}^N) \approx \mathbb{E}\left[m_k(M^N(\boldsymbol{\xi}^N(\boldsymbol{\eta})))\right] = \frac{1}{N}\sum_{t=1}^{N} \mathbb{E}\left[M_t^k(\boldsymbol{\xi}^t(\boldsymbol{\eta}))\right] = m_k(\boldsymbol{\eta})$$

for $k = 1, \ldots, K$. Choosing (3.4) as measure of "distance" between the sample moments and the model moments now gives the Generalized Method of Moments.

The key point of the above procedure is that an intermediate, simplified, model is used (above represented by (3.5)) and that the estimates of the hyperparameters using observations and fictitious observations from the model are matched. It is thus the estimates of the hyperparameters of the simplified model that serve as our statistic and we find the model which is able to best match this statistic.

By using another intermediate model than (3.4) we can obtain other estimators of the form

$$\hat{\boldsymbol{\eta}}(\mathbf{z}^N) := \arg\min_{\boldsymbol{\eta}} V_{wse}(\boldsymbol{\eta}, \mathbf{z}^N) \tag{3.6}$$

where

$$\begin{aligned} &V_{wse}(\boldsymbol{\eta}, \mathbf{z}^N) := \\ &\left(\hat{\hat{\boldsymbol{\eta}}}(\mathbf{z}^N) - \mathbb{E}\left[\hat{\hat{\boldsymbol{\eta}}}(M^N(\boldsymbol{\xi}^N(\boldsymbol{\eta})))\right]\right)^T W \left(\hat{\hat{\boldsymbol{\eta}}}(\mathbf{z}^N) - \mathbb{E}\left[\hat{\hat{\boldsymbol{\eta}}}(M^N(\boldsymbol{\xi}^N(\boldsymbol{\eta})))\right]\right) \end{aligned} \tag{3.7}$$

This approach is known as *indirect inference*. Other cost functions than $V_{wse}(\boldsymbol{\eta}, \mathbf{z}^N)$ can of course be used. If a likelihood function $p(\mathbf{z}^N; \tilde{\beta})$ has been specified for the intermediate model, one possibility is

$$V_{lh}(\boldsymbol{\eta}, \mathbf{z}^N) := \mathbb{E}\left[p(M^N(\boldsymbol{\xi}^N(\boldsymbol{\eta}), \hat{\hat{\eta}}_{ML}(\mathbf{z}^N))\right]$$

i.e. we try to find a model that generates data that matches a given ML-estimate as measured by the likelihood function used to estimate this ML-estimate.

When the expectation over $\boldsymbol{\xi}^N(\boldsymbol{\eta})$ is difficult to compute, Monte Carlo simulations may be used instead. Defining

$$\hat{\mathbb{E}}^Q\left[g(\boldsymbol{\xi}^N(\boldsymbol{\eta}))\right] := \frac{1}{M}\sum_{k=1}^{Q} g(\boldsymbol{\xi}_k^N(\boldsymbol{\eta}))$$

where $\{\boldsymbol{\xi}_k^N(\boldsymbol{\eta})\}_{k=1}^Q$ are Q independent of realizations of $\boldsymbol{\xi}^N(\boldsymbol{\eta})$, one can use

$$\hat{V}_{lh}^Q(\boldsymbol{\eta}, \mathbf{z}^N) := \hat{\mathbb{E}}^Q\left[p(M^N(\boldsymbol{\xi}^N(\boldsymbol{\eta}), \hat{\hat{\eta}}_{ML}(\mathbf{z}^N))\right]$$

as criterion. The expectation in V_{wse} may also be replaced by a Monte Carlo estimate.

Prediction error methods The methods above use the model to predict certain statistics. We also observed that it does not make sense to try to predict the entire observation $\mathbf{z}^N$ as this just leads to the set of unfalsified models. However, we could use the model to construct an estimator of subsets of $\mathbf{z}^N$ given other subsets of $\mathbf{z}^N$. Such estimators are called *predictors*. We could then determine model- and hyperparameters by minimizing the errors between the observations we try to predict and the corresponding predictors. These errors are called *prediction errors*, and therefore this class of methods is called *Prediction Error Methods* (PEM). For dynamic systems, it is natural to consider k-step ahead predictors of the output. Let us for simplicity of argument suppose that $\mathbf{z}(t) = \begin{bmatrix}\mathbf{y}^T(t) & \mathbf{u}^T(t)\end{bmatrix}^T$ and that the model is

$$\mathbf{y}(t) = f_t(\mathbf{u}^t, \mathbf{v}^t; \boldsymbol{\theta}), \; t = 1, 2, \ldots$$

where the sequence $\{\mathbf{v}^t\}$ is characterized by the pdfs $\{p_t\}$. A k-step ahead predictor is then a sequence of functions $\hat{f}_{t+k|t}$ which define the k-step ahead predictor as

$$\hat{\mathbf{y}}(t+k|t;\boldsymbol{\theta}) := \hat{f}_{t+k|t}(\mathbf{u}^{t+k}, \mathbf{y}^t; \boldsymbol{\theta})$$

Notice that the k-step predictor is only allowed to use the output history with a lag of k time steps. We would then choose $\boldsymbol{\theta}$ such that the prediction errors

$$\boldsymbol{\varepsilon}(t+k|t;\boldsymbol{\theta}) = \mathbf{y}(t+k) - \hat{\mathbf{y}}(t+k|t;\boldsymbol{\theta}),\ t = 1,\ldots,N-k$$

are small in some sense. We may for example use the quadratic criterion (3.7)

$$V_{pe,k}(\boldsymbol{\theta}, \mathbf{z}^N) := \begin{bmatrix} \boldsymbol{\varepsilon}(1+k|1;\boldsymbol{\theta}) \\ \vdots \\ \boldsymbol{\varepsilon}(N|N-k;\boldsymbol{\theta}) \end{bmatrix}^T W \begin{bmatrix} \boldsymbol{\varepsilon}(1+k|1;\boldsymbol{\theta}) \\ \vdots \\ \boldsymbol{\varepsilon}(N|N-k;\boldsymbol{\theta}) \end{bmatrix}$$

3.3.3 Ranking statistics

An alternative way to come up with an estimator from a statistic $\mathbf{s}$ is to compute the pdf $p(\mathbf{s};\boldsymbol{\eta})$ and then use any of the previously proposed "ranking" methods to construct an estimator of $\boldsymbol{\eta}$.

Example 3.3. *Suppose that*

$$\mathbf{z}(t) = \mathbf{v}(t)$$

where $\{\mathbf{v}(t)\}$ is a sequence of independent $\mathcal{N}(0,\lambda)$-distributed random variables. The neg-loglikelihood is given by

$$\frac{1}{\lambda}\sum_{t=1}^{N}\mathbf{z}^2(t) + N\log\lambda$$

The sample second order moment is given by

$$\mathbf{s} = \frac{1}{N}\sum_{t=1}^{N}\mathbf{z}^2(t)$$

We have that $N\mathbf{s}/\lambda$ is $\chi^2(N)$ distributed, and hence $\mathbf{s}$ has pdf

$$\begin{aligned} p(\mathbf{s}) &= p\left(\frac{\lambda}{N}\frac{N\mathbf{s}}{\lambda}\right) = \frac{N}{\lambda}p\left(\frac{N\mathbf{s}}{\lambda}\right) = \frac{N}{\lambda}\frac{1}{2^{N/2}\Gamma(N/2)}\left(\frac{N\mathbf{s}}{\lambda}\right)^{N/2-1} e^{-\frac{N\mathbf{s}}{2\lambda}} \\ &= \frac{N^{N/2}}{2^{N/2}\Gamma(N/2)}\mathbf{s}^{N/2-1}\frac{1}{\lambda^{N/2}}e^{-\frac{N\mathbf{s}}{2\lambda}} \end{aligned}$$

This gives the following neg-loglikelihood

$$\frac{N}{\lambda}\mathbf{s} + N\log\lambda = \frac{1}{\lambda}\sum_{t=1}^{N}\mathbf{z}^2(t) + N\log\lambda$$

which is the same as for $\mathbf{z}^N$ itself as seen above.

Example 3.3 showed that the neg-loglikelihood may remain the same even when we compress data. This is not always the case as illustrated in the next example. In Section 5.5 we will return to the question of when this is possible.

Example 3.4. *Suppose that*

$$\mathbf{z}(t) = \mathbf{v}(t)$$

where $\{\mathbf{v}(t)\}$ *is a sequence of uniformly distributed random variables on* $[\boldsymbol{\theta}, \boldsymbol{\theta}+1]$. *Let us for simplicity consider the case when we have* $N = 2$ *observations. The joint density for* $(\mathbf{z}_1, \mathbf{z}_2)$ *is 1 on the square* $[\boldsymbol{\theta}, \boldsymbol{\theta}+1] \times [\boldsymbol{\theta}, \boldsymbol{\theta}+1]$, *see Figure 3.1*

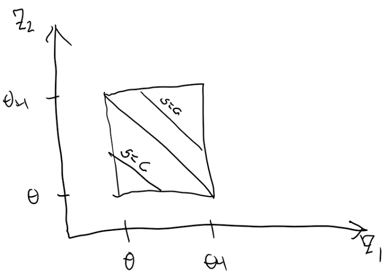

Figure 3.1: The pdf is 1 within the square.

Thus any $\boldsymbol{\theta}$ *for which the observation* $(\mathbf{z}_1, \mathbf{z}_2)$ *belongs to this square is a ML-estimate. This is equivalent to that*

$$\boldsymbol{\theta} \le \min(\mathbf{z}_1, \mathbf{z}_2) \le \max(\mathbf{z}_1, \mathbf{z}_2) \le \boldsymbol{\theta} + 1$$

giving that all $\boldsymbol{\theta} \in [max(\mathbf{z}_1, \mathbf{z}_2) - 1, min(\mathbf{z}_1, \mathbf{z}_2)]$ *are ML-estimates.*

Now consider the statistic $\mathbf{s} = \mathbf{z}1 + \mathbf{z}2$. *The first moment of the corresponding model* $\mathbf{v}1 + \mathbf{v}2$ *is* $2\boldsymbol{\theta} + 1$. *Thus we could take* $(\mathbf{s} - 1)/2$ *as estimator of* $\boldsymbol{\theta}$.

We can also derive the pdf for $\mathbf{s}$. *We need to integrate over the lines in Figure 3.1 where* $\mathbf{s} = \mathbf{z}_1 + \mathbf{z}_2$ *is constant. This gives the pdf in Figure 3.2. The ML estimate is thus to choose* $\boldsymbol{\theta}$ *such that*

$$\mathbf{s} = 2\boldsymbol{\theta} + 1$$

i.e. the ML estimate is $(\mathbf{s} - 1)/2$. *i.e. the same as we obtained using the optimal estimator for the sum of the observations, but different from the ML-estimate using* $\mathbf{z}^2$.

A relation between ML estimation and PEM. Returning to prediction error minimization, we can for a specific choice of cost function interpret this method as consisting of a transformation of data followed by ML-estimation. Let p be a generic notation for a pdf, where the

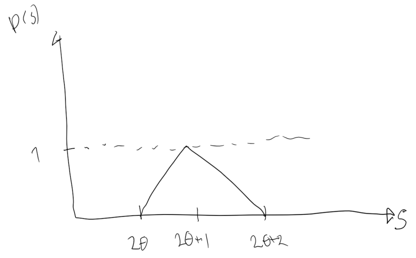

Figure 3.2: The pdf for **s**.

argument makes it clear for which quantity it is a pdf. Let us also hide the input dependence of a pdf. The likelihood $p(\mathbf{y}^N;\boldsymbol{\theta})$ can be factorized as

$$p(\mathbf{y}^N;\boldsymbol{\theta}) = p(\mathbf{y}(N)|\mathbf{y}^{N-1};\boldsymbol{\theta})p(\mathbf{y}^{N-1};\boldsymbol{\theta}) = \prod_{t=1}^{N} p(\mathbf{y}(t)|\mathbf{y}^{t-1};\boldsymbol{\theta})\ p(\mathbf{y}(0);\boldsymbol{\theta})$$

Now, since $\hat{\mathbf{y}}(t+1|t;\boldsymbol{\theta})$ is a function of $\mathbf{y}^{t-1}$, $p(\mathbf{y}(t)|\mathbf{y}^{t-1};\boldsymbol{\theta}) = p(\mathbf{y}(t) - \hat{\mathbf{y}}(t+1|t;\boldsymbol{\theta})|\mathbf{y}^{t-1};\boldsymbol{\theta}) = p(\varepsilon(t;\boldsymbol{\theta})|\mathbf{y}^{t-1})$ so we can write

$$p(\mathbf{y}^N;\boldsymbol{\theta}) = \prod_{t=1}^{N} p(\varepsilon(t|t-1;\boldsymbol{\theta})|\mathbf{y}^{t-1};\boldsymbol{\theta})\ p(\mathbf{y}(0);\boldsymbol{\theta})$$

We can thus express the neg-loglikelihood in terms of the prediction errors as

$$-\sum_{t=1}^{N} \log p(\varepsilon(t|t-1;\boldsymbol{\theta})|\mathbf{y}^{t-1};\boldsymbol{\theta})\ -\log p(\mathbf{y}(0);\boldsymbol{\theta}) \tag{3.8}$$

Thus the prediction error method using the one-step ahead predictor with the above cost function is equivalent to ML-estimation of $\boldsymbol{\theta}$. Thus we can interpret the prediction error method as first making a transformation of the observation $\mathbf{z}^N$ into the prediction errors $\boldsymbol{\varepsilon}^N(\boldsymbol{\theta})$ and then maximizing the likelihood (3.8) for this new data set.

Notice that above we have made no assumptions on the form of the predictor. In the case when the predictor is such that

$$\mathbf{y}(t;\boldsymbol{\theta}) = \hat{\mathbf{y}}(t|t-1;\boldsymbol{\theta}) + \mathbf{e}(t) \tag{3.9}$$

where $\mathbf{e}(t)$ is independent of the past $\mathbf{y}^{t-1}$, then

$$p(\varepsilon(t|t-1;\boldsymbol{\theta})|\mathbf{y}^{t-1};\boldsymbol{\theta}) = p_{\mathbf{e}(t)}(\varepsilon(t|t-1;\boldsymbol{\theta}))$$

and hence the likelihood becomes

$$-\sum_{t=1}^{N} \log p_{e(t)}(\varepsilon(t|t-1;\boldsymbol{\theta}))\ -\log p(\mathbf{y}(0);\boldsymbol{\theta})$$

When (3.9) holds, the predictor $\hat{\mathbf{y}}(t|t-1;\boldsymbol{\theta})$ is optimal in the MSE sense, see Exercise 3.5. Hence, $\hat{\mathbf{y}}(t|t-1;\boldsymbol{\theta})$ must be the posterior mean. Notice that the reverse does not hold, i.e. $\mathbf{y}(t) - \mathbb{E}\left[y(t)|\mathbf{y}^{t-1}\right]$ is not necessarily independent of the past, see Exercise 3.6.

Parameter dependent statistics. What may appear bizarre in the interpretation of the prediction error method above is that our transformed "observation" $\varepsilon^N(\boldsymbol{\theta})$ is parameter dependent. Let us see what effect this may have.

Example 3.5 (Example 3.3 continued)**.** *It may be tempting to use $N\mathbf{s}/\lambda$ as "observation" instead of* $\mathbf{s}$ *since we directly know that this is a $\chi^2(N)$-distributed random variable. The pdf for this parameter dependent statistic is*

$$p\left(\frac{N\mathbf{s}}{\lambda}\right) = \frac{N}{\lambda}\,\frac{1}{2^{N/2}\Gamma(N/2)}\left(\frac{N\mathbf{s}}{\lambda}\right)^{N/2-1} e^{-\frac{N\mathbf{s}}{2\lambda}}$$

with neg-loglikelihood

$$\frac{N}{\lambda}\mathbf{s} + (N-2)\log\lambda = \frac{1}{\lambda}\sum_{t=1}^{N}\mathbf{z}^2(t) + (N-2)\log\lambda$$

which differs from the neg-loglikelihood for $\mathbf{z}$*. We will thus obtain a different estimate of λ.*

Example 3.5 shows that the likelihood function may change if the transformation is parameter dependent. We will need to better understand what goes on here. Lemma D.1.1 tells us how the pdf is transformed when a 1-1 mapping f is applied. With $\mathbf{z}_f^N := f(\mathbf{z}^N)$

$$p(\mathbf{z}_f^N) = \frac{p(\mathbf{z})}{|\det f'(\mathbf{z})|}$$

Thus the neg-loglikelihood for $\mathbf{z}_f^N$ is given by

$$-\log p(\mathbf{z}_f^N) = -\log p(\mathbf{z}) + \log|\det f'(\mathbf{z})| \tag{3.10}$$

Thus if $|\det f'(\mathbf{z})|$ is independent of the hyperparameters, the ML-estimators using $\mathbf{z}_f^N = f(\mathbf{z}^N)$ and $\mathbf{z}^N$ will be identical. However, when $|\det f'(\mathbf{z})|$ is a function of the hyperparameters the two ML-estimators may very well be different even if in this case there is no data-compression and the function is one-to-one so that $\mathbf{z}^N$ can be recovered from $\mathbf{z}_f^N$ (by applying f^{-1}). Thus care has to be exercised when using parameter dependent transformations of data.

In view of this, let us return to prediction error minimization where the transformation f is given by

$$\begin{bmatrix} \mathbf{y}(1) \\ \mathbf{y}(2) \\ \vdots \\ \mathbf{y}(N-1) \\ \mathbf{y}(N) \end{bmatrix} \to \begin{bmatrix} \mathbf{y}(1) - \hat{\mathbf{y}}(1|\mathbf{y}^0;\boldsymbol{\theta}) \\ \mathbf{y}(2) - \hat{\mathbf{y}}(2|\mathbf{y}^1;\boldsymbol{\theta}) \\ \vdots \\ \mathbf{y}(N-1) - \hat{\mathbf{y}}(N-1|\mathbf{y}^{N-2};\boldsymbol{\theta}) \\ \mathbf{y}(N) - \hat{\mathbf{y}}(N|\mathbf{y}^{N-1};\boldsymbol{\theta}) \end{bmatrix}$$

Here $\mathbf{y}(0)$ is not an observation so it must either be assumed known or being part of the unknown $\boldsymbol{\theta}$. The structure of the map above means that the Jacobian $f'(\mathbf{y})$ is upper triangular and with 1's on the diagonal. But such a matrix has determinant one and is hence independent of $\boldsymbol{\theta}$. Using this in (3.10) shows that the parameter dependent parts of the neg-loglikelihoods for $\mathbf{y}$ and $\varepsilon^N(\boldsymbol{\theta})$ will be the same.

3.4 A Probabilistic Toolshed

In this section we will provide some tools for probabilistic modeling. There are different objects we would like to model. Disturbances and noise in sampled data systems can be modelled by a sequence of random variables. This corresponds to discrete time stochastic processes $\{s(t)\}_{t\in T}$, where the domain T is the integers $T = \mathbb{Z}$ or the natural numbers $T = \mathbb{N}$. A nonlinear function $f : \mathbb{R} \to \mathbb{R}$ can also be modelled as a stochastic process but then the domain is $T = \mathbb{R}$, or perhaps a subset of $\mathbb{R}$. Thus we will need to treat different domains.

3.4.1 Basic concepts

In mathematical terms probability theory is concerned with functions that map sets to real numbers between 0 and 1. We start with a set Ω, called the sample space, whose elements ω, the sample points, can be thought of representing all possible outcomes that can occur. An event is simply a subset of Ω. A probability measure $\mathbf{P}$ assigns a probability to every event, i.e. a number between 0 and 1. It has to hold that $\mathbf{P}(\Omega) = 1$ and for a disjoint family of sets $\{A_k\}_{k=1}^{\infty}$, that $\mathbf{P}(\cup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} \mathbf{P}(A_k)$. We can express the probability of an event in integral form as

$$\mathbf{P}(A) = \int_A d\mathbf{P} = \int_A d\mathbf{P}(\omega)$$

where a generalization of the Riemann's definition is needed for the integral. The reader not familiar with measure theory may think of $d\mathbf{P}$ as a non-negative weight function, weighting different outcomes ω.

We would of course like to assign a probability to all possible events, i.e. to all possible subsets of Ω. However, the sets of sets is humongous and leads to inconsistencies.

Example 3.6 (Banach and Tarski)**.** *Let Ω be the unit sphere $\mathcal{S}^2$ in $\mathbb{R}^3$ and define the probability measure for a set F as the area of F divided with the area of $\mathcal{S}^2$. Then it can be shown that there exists an $F \in \mathcal{S}^2$ and disjoint rotations $\{F_{i,k}\}_{i=1}^{k}$ of F such that $\mathbf{P}(F_{i,j}) = \mathbf{P}(F)$ and $S^2 = \cup_{i=1}^{k} F_{i,k}$ for $k \geq 3$. However, since the sets are disjoint this implies that*

$$1 = \mathbf{P}(\mathcal{S}^2) = \mathbf{P}(\cup_{i=1}^{k} F_{i,k}) = \sum_{i=1}^{k} \mathbf{P}(F_{i,k}) = k\mathbf{P}(F),\ k = 3, 4, \dots$$

i.e. it would appear that the area of F is not unique.

Thus, if we want to have probabilities defined in a meaningful and consistent way the family of allowed sets must be restricted. We will denote such a family of sets by $\mathcal{F}$. Natural requirements are that i) $\Omega \in \mathcal{F}$, ii) if $A \in \mathcal{F}$ then its complement $A^c \in \mathcal{F}$, and iii) the union of two sets belonging to $\mathcal{F}$ should be in $\mathcal{F}$ as well. However, these requirements are not enough to be able to answer pertient questions in estimation theory. Let us jump ahead and study a typical estimation problem.

Example 3.7. *Let $\hat{\theta}_N$ be an estimator of a scalar quantity θ based on the random vector $Y_N \in \mathbb{R}^N$ and let us assume that we would like to examine the properties of $\hat{\theta}_N$ when N becomes large. For example we might be interested in the probability that $\hat{\theta}_N$ eventually remains within a distance $\varepsilon > 0$ from θ. This event can be expressed as*

$$F = \{\omega : \limsup_{N\to\infty} |\hat{\theta}_N - \theta| \leq \varepsilon\} = \{\omega : |\hat{\theta}_N - \theta| \leq \varepsilon \textit{ for N sufficiently large}\}$$

Defining $F_N = \{\omega : |\hat{\theta}_N - \theta| \leq \varepsilon\}$ we can write

$$F = \cup_{m=1}^{\infty} \cap_{n=m}^{\infty} F_n$$

For sets of the type F in the example to belong to $\mathcal{F}$ given that $F_k \in \mathcal{F}$, $k = 1, 2, \ldots$, it turns out that we have to require

$$\text{iv)} \quad F_k \in \mathcal{F},\ k = 1, 2, \ldots \ \Rightarrow \cup_{k=1}^{\infty} F_k \in \mathcal{F}$$

A family $\mathcal{F}$ of sets satisfying i)–iv) is called a *σ-algebra*, the pair $(\Omega, \mathcal{F})$ a *measurable space* and the triplet $(\Omega, \mathcal{F}, \mathbf{P})$ a *probability space*.

Starting from a family $\mathcal{C}$ of subsets on Ω, $\sigma(\mathcal{C})$ is the smallest σ-algebra that contains all sets in $\mathcal{C}$. An example is the Borel σ-algebra $\mathcal{B}$ which is the smallest σ-algebra containing the open sets on the real axis. The sets in this σ-algebra are called Borel sets. The concept of Borel algebra extends to $\mathbb{R}^N$, $N < \infty$.

3.4.2 *Random Variables*

Given a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, a real valued random variable is a function from the sample space to the real axis: $\Omega \to \mathbb{R}$. A typical event for a random variable Y that we might be interested in is

$$\{\omega : Y(\omega) < c\} \tag{3.11}$$

for some constant c (this leads to the distribution function). For us to be able to assign probabilities to such events, they have to be measurable, i.e. they have to belong $\mathcal{F}$. Real valued functions f from the sample space for which sets of the type $\{\omega : f(\omega) \in B\}$, where B is any Borel set, are measurable, i.e. belong to $\mathcal{F}$, are called measurable functions. Sets of the type (3.11) are of this type and by requiring random variables to be measurable functions, the probability for the event (3.11) is well defined.

A random variable X is characterized by its probability distribution function $\mathbf{P}_X(B) := \mathbf{P}(X \in B)$ for all $B \in \mathcal{B}$, where $X \in B = \{\omega : X(\omega) \in B\}$. We can see $\mathbf{P}_X(B)$ as a probability measure on the measurable space $(\mathbb{R}, \mathcal{B})$. The distribution function is defined as $F_X(x) := \mathbf{P}_X((-\infty, x])$, and since $\mathbf{P}$ is a measurable function so is F_X meaning that we can write

$$F_X(y) = \int_{-\infty}^{y} dF_X(x)$$

If $\mathbf{P}_X$ is absolutely continuous with respect to the Lesbegue measure, there is a measurable function $p_X : \mathbb{R} \to \mathbb{R}$ called the *probability density function* (pdf) such that

$$\mathbf{P}_X(B) = \int_B p_X(x) dx \tag{3.12}$$

Here absolutely continuous means that[9].

$$\int_B dx = 0 \quad \Rightarrow \quad \mathbf{P}_X(B) = 0$$

When $\mathbf{P}_X$ is absolutely continuous, it follows from (3.12) that the distribution function for a random variable X can be written as

$$F_X(x) = \mathbf{P}_X((-\infty, x)) = \int_{-\infty}^{x} p_X(x)dx$$

and hence F_X is absolutely continuous. Furthermore, by the fundamental law of calculus $\frac{d}{dx}F_X(x) = p_X(x)$. In measure theoretic terms we thus have $dF_X(x) = p_X(x)dx$.

Absolutely continuous distribution functions cannot represent distributions of random variables for which there is a non-zero probability of events like $X = x$, c.f. with the probabilities of the outcomes from throwing a dice. For this discrete distribution functions are needed. Such functions are piecewise constant, right continuous, with at most a countable number of positive jumps

$$F_X(x) = \sum_{k=0}^{\infty} p_k \sigma(x - x_k), \quad p_k \geq 0, \ \sum_k p_k = 1$$

where

$$\sigma(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

The x_k are simply the outcomes of X that can occur, and p_k their associated probabilities

A discrete distribution function can have at most a countable points of discontinuity. Also a discrete distribution function is a measurable function and here

$$dF_X(x) = \sum_{k=0}^{\infty} p_k \delta(x - x_k)$$

where δ is Dirac's delta function.

With X and Y being independent and having a discrete and absolutely continuous distrbution function, respectively, $X + Y$ will have a combination of the two as distribution function.

Theorem 3.4.1 (Theorem 1.3.2 in [10]). *Every distribution function can be uniquely decomposed into a convex combination of a discrete, an absolutely continuous, and a continuous singular distribution function.*

A distribution function is right-continuous at every point of its domain. A distribution function is *singular* if it is not identically zero and its derivative exists and equals zero almost everywhere. Thus a continuous singular distribution function can only increase on a set of measure zero, on which the derivative does not exist but where the function is still continuous. We refer to p.12 in [11] for an example of how such a peculiar function can be constructed.

[9] Theorem 7.18 in
W. Rudin. *Real and Complex Analysis*. McGraw-Hill, London, 1986

[11] K.L. Chung. *A Course in Probability Theory*. Academic Press, Orlando, Florida 32887, 1974

3.4.3 Expectation

A random variable X, defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, has expectation

$$\mathbb{E}[X] = \int_\Omega X d\mathbf{P} := \int_\Omega X(\omega)\mathbf{P}(d\omega)$$

when the integral is defined. However, we can also express the expectation in terms of the probability distribution function $\mathbf{P}_X$, or the distribution function

$$\mathbb{E}[X] = \int_\Omega X(\omega)\mathbf{P}_\mathbf{X}(X(d\omega)) = \int_{-\infty}^{\infty} x\mathbf{P}_X(dx) = \int_{-\infty}^{\infty} xdF_X(x) \quad (3.13)$$

When the sample space is discrete, so that the distribution function is discrete, it follows that

$$\mathbb{E}[X] = \sum_k x_k \mathbf{P}_X(x_k) = \sum_k x_k p_k$$

and when the pdf exists

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x p_X(x) dx$$

With f being a measurable function, $f(X)$ is a random variable and its expectation is given by

$$\mathbb{E}[f(X)] = \int_{-\infty}^{\infty} f(x) p_X(x) dx$$

when the pdf of X exists. This is a non-trivial result.

3.4.4 Random vectors

A *random vector* $\mathbf{X} = \begin{bmatrix} X(1) & \dots & X(n) \end{bmatrix}^T$ is a vector where all elements are random variables. One can then proceed as for a random variable and define a probability distribution function $\mathbf{P}_\mathbf{X}(B)$ for events $B = B_1 \times \dots \times B_n \in \mathcal{B}^n = \mathcal{B} \times \dots \times \mathcal{B}$ such that

$$\mathbf{P}_\mathbf{X}(B) = \mathbf{P}(\mathbf{X} \in B)$$

as well as the distribution function

$$F_\mathbf{X}(x_1, \dots, x_n) = \mathbf{P}((X(1) < x_1) \cap \dots \cap (X(n) < x_n))$$

The distribution function uniquely defines the probability distribution (Theorem 2 in \$ 3, Chapter II [12]). Furthermore, when $\mathbf{P}_\mathbf{X}$ is absolutely continuous the pdf $p_\mathbf{X}(x_1, \dots, x_n)$ is defined by

$$\mathbf{P}_\mathbf{X}(B) = \int_{B_1} \dots \int_{B_n} p_\mathbf{X}(x_1, \dots, x_n) dx_1 \dots dx_n$$

which, with $\mathbf{x} = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix}^T$, we write compactly as

$$\mathbf{P}_\mathbf{X}(B) = \int_B p_\mathbf{X}(\mathbf{x}) d\mathbf{x}$$

A different definition of a random vector could be that $\mathbf{X}$ is a random vector in $\mathbb{R}^n$ if for all Borel sets B in $\mathbb{R}^n$, $\{\omega : \mathbf{X}(\omega) \in B\}$ is measurable.

[12] A.N. Shiryaev. *Probability*. Springer, 2nd edition, 1989

However, this definition is equivalent to the one above. This is a subtle but non-trivial result which follows from that the set of Borel sets in $\mathbb{R}^n$ is the smallest σ-algebra containing the sets $B_1 \times \ldots \times B_n$, where the B_i are Borel sets in $\mathbb{R}$.

Notice that, with $\bar{\mathbf{X}}$ being identical to $\mathbf{X}$ save that $X(k)$ has been removed,

$$p_{\bar{\mathbf{X}}}(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n) = \int_{-\infty}^{\infty} p_{\mathbf{X}}(x_1, \ldots, x_n) dx_k$$

The procedure of integrating out certain variables is called *marginalization*. We can also define expectation for random vectors and functions of random vectors

$$\begin{aligned} \mathbb{E}[\mathbf{X}] &= \int_{-\infty}^{\infty} \mathbf{x} p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\ \mathbb{E}[f(\mathbf{X})] &= \int_{-\infty}^{\infty} f(\mathbf{x}) p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \end{aligned}$$

Here f can also be matrix-valued, in particular the *covariance* between two random vectors $\mathbf{X}$ and $\mathbf{Y}$ is defined as

$$C_{\mathbf{X},\mathbf{Y}} := \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T\right]$$

We use $C_{\mathbf{X}}$ or even $\text{Cov}\{\mathbf{X}\}$ for short of $C_{\mathbf{X},\mathbf{X}}$.

3.4.5 *Stochastic Processes*

A stochastic process is a generalization of a random vector to an infinite number of random variables $X(t) \in \mathbb{R}$, $t \in T$ defined on a common probability space $(\Omega, \mathsf{P}, \mathcal{F})$. The index set T can be either countable, e.g. the natural numbers $T = \mathbb{N}$ or the integers $T = \mathbb{Z}$, but T can also be uncountable, e.g. the set of reals $T = \mathbb{R}$. Unless important for the treatment we will not specify T.

For a given finite collection of these random variables $\mathbf{X} = \begin{bmatrix} X(t_1) & \ldots & X(t_n) \end{bmatrix}^T$, a finite dimensional probability distribution $\mathbf{P}_{\mathbf{X}}$ is induced on $(\mathbb{R}^n, \mathcal{B}^n)$ as for random vectors. These in turn define a family of finite dimensional distribution functions

$$F_{t_1,\ldots,t_n}(x_1, \ldots, x_n) := \mathbf{P}_{\mathbf{X}}(X(t_1) \le x_1, \ldots, X(t_n) \le x_n),\ t_1 < \ldots < t_n \tag{3.14}$$

and when these measures are absolutely continuous, the corresponding pdfs $p_{t_1,\ldots,t_n}(x_1, \ldots, x_n)$ exists.

A stochastic process can also be vector valued, i.e. $\mathbf{X}(t) \in \mathbb{R}^n$, with obvious modifications of the definitions of the probability measures and pdfs above.

For two stochastic processes $\mathbf{X}(t)$ and $\mathbf{Y}(t)$ $t \in T$, defined on the same probability space, we can define moment functions, such as the mean, cross-correlation and cross-covariance functions

$$\begin{aligned} m_{\mathbf{X}}(t) &:= \mathbb{E}[\mathbf{X}(t)] \\ R_{\mathbf{X},\mathbf{Y}}(t,s) &:= \mathbb{E}\left[\mathbf{X}(t)\mathbf{Y}^T(s)\right] \\ C_{\mathbf{X},\mathbf{Y}}(t,s) &:= \mathbb{E}\left[(\mathbf{X}(t) - m_{\mathbf{X}}(t))(\mathbf{Y}(s) - m_{\mathbf{Y}}(s))^T\right] \end{aligned}$$

We call $R_{\mathbf{X},\mathbf{X}}(t,s)$ and $C_{\mathbf{X},\mathbf{X}}(t,s)$ the *auto-correlation* (akf) and *covariance* function, respectively.

The characterization of a stochastic process given above has the shortcoming that it only allows us to compute probabilities for a process $\mathbf{X}(t)$ which can be described by the process at a finite number of time instances. What we need is to adopt the alternate view we briefly discussed for random vectors where we saw $\mathbf{X}(\omega)$ as an outcome of $\mathbf{X} \in \mathbb{R}^n$ rather than the elements being individual random variables. Similarly, for random processes we can think of $X(\cdot,\omega)$ as an element in $\mathbb{R}^T$, the space of real-valued functions defined on T. The functions $\mathbf{X}(\cdot,\omega)$ are called *realizations* of $\mathbf{X}$, or *sample paths* or *trajectories*.

Example 3.8 (Example 3.1 in [13]). *Let η be a random variable uniformly distributed on $[0,1]$ and define $X(t) = \delta(t-\eta)$, where $\delta(x)$ is Kronecker's delta function, for $t \in [0,1]$. Then*

$$\mathbf{P}_{\mathbf{X}}(X(t) \in B) = \begin{cases} 1 & 0 \in B \\ 0 & otherwise \end{cases}$$

since $\eta = t$ with probability 0, and for the same reason

$$\mathbf{P}_{\mathbf{X}}(X(t_1) \in B_1, \dots, X(t_n) \in B_n) = \begin{cases} 1 & 0 \in \cap_{k=1}^{n} B_k \\ 0 & otherwise \end{cases}$$

The question now is for which sets B of functions in $\mathbb{R}^T$ a probability is induced by $\mathbf{P}$? Now for each $t \in T$, $\mathbf{X}(t)$ is a random variable so for sets of the type

$$B_t(B) = \{\mathbf{X}(\cdot,\omega) \in \mathbb{R}^T : \mathbf{X}(t) \in B\}, \quad B \in \mathcal{B}$$

we can define a probability $\mathbf{P}_{\mathbf{X}}(B_t(B)) = \mathbf{P}(B)$. As for random vectors we can take the intersection of such sets, i.e.

$$B_{t_1,\dots,t_n}(B_1 \times \dots \times B_n) = \{\mathbf{X}(\cdot,\omega) \in \mathbb{R}^T : X(t_1) \in B_n, \dots, \mathbf{X}(t_n) \in B_n\},$$

where $B_1, \dots, B_n \in \mathcal{B}$. It then turns out that $\mathbf{P}$ induce probabilities for sets in $\mathcal{B}(\mathbb{R}^T)$, the smallest σ-algebra containing all sets $B_{t_1,\dots,t_n}(B_1 \times \dots \times B_n)$ (Theorem 4 \$3, Chapter II in [14]).

Seen from a modeling perspective, e.g. think of the problem of defining a stochastic process which models a disturbance, it seems natural to first define a suitable probability measure and then define a stochastic process obeying this probability measure. When constructing a probability measure $\mathsf{P}_{\mathbf{X}}$ one must ensure that it is consistent, meaning that for all sets $s = \{s_1, \dots, s_k\}$ and $t = \{t_1, \dots, t_n\}$, $s \subseteq t$,

$$\begin{aligned} &\mathsf{P}_{\mathbf{X}}\left(\{\mathbf{X}(s_1), \dots, \mathbf{X}(s_k)\} : \{\mathbf{X}(s_1), \dots, \mathbf{X}(s_k)\} \in B\right) \\ &= \mathsf{P}_{\mathbf{X}}\left(\{\mathbf{X}(t_1), \dots, \mathbf{X}(t_n)\} : \{\mathbf{X}(s_1), \dots, \mathbf{X}(s_k)\} \in B\right) \end{aligned}$$

c.f. marginalization. However, at this point it is not clear if for a given probability measure $\mathsf{P}_{\mathbf{X}}$ on $(\mathbb{R}^T, \mathcal{B}(\mathbb{R}^T))$ there exists a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ and a stochastic process $\mathbf{X}(t)$ that has $\mathsf{P}_{\mathbf{X}}$ as

[14] A.N. Shiryaev. *Probability*. Springer, 2nd edition, 1989

probability distribution function. Fortunately, it has been shown that this can always be done and that it suffices to specify the finite dimensional distribution functions (3.14) (Kolmogorov's extension theorem, Theorem 1, §9, Chapter II in [15]). It also holds that $\mathbf{P}$ is unique and one can thus say that the finite dimensional distribution functions completely specify a stochastic process.

However, it should be noted that different stochastic processes may have the same probability distribution functions but different realizations, and hence there are restrictions in the sets for which probabilities can be computed using the finite dimensional probability distributions, especially when T is uncountable.

Example 3.9 (Example 3.8 continued)**.** *Let $Y(t) = 0 \cdot \eta$ for $t \in [0,1]$. Then Y has the same finite dimensional probability distributions as $X(t)$ in Example 3.8. However,*

$$\mathbf{P}(\sup_{t\in[0,1]} Y(t) = 0) = \mathbf{P}(\sup_{t\in[0,1]} X(t) = 1) = 1$$

so with probability 1 the sample paths of the two processes do not coincide.

The problem in the previous example is that the functional $h(f) := \sup_{t\in[0,1]} f(t)$ is not measurable in $\mathbb{R}^T$, $T = [0,1]$, i.e. $\{f : h(f) \in B\}$, where $B \in \mathcal{B}$, does not belong to $\mathcal{B}(\mathbb{R}^T)$.

Modeling considerations. Even if it suffices to specify the finite dimensional distribution functions of a stochastic process, the degrees of freedom in defining a stochastic process are overwhelming. A partial specification is given by the mean function and the akf. There is a one-to-one correspondance between akfs and positive definite functions (p. 132 of volume 2 [16]), where the latter are functions $K : T \times T \to \mathbb{R}^{n\times n}$ such that K is non-negative in the sense that

$$\sum_{i=1}^{m}\sum_{j=1}^{m} a^*(i)K(t_i,t_j)a(j) \geq 0, \quad \forall a(i) \in \mathbb{C}^n,\ t_i \in T,\ m \in \mathbb{N} \tag{3.15}$$

The condition (3.15) implies that K is symmetric in the sense $K(t,s) = K^T(s,t)$ [17], see Exercise 3.9.

For any function $f(\tau) : T \to \mathbb{R}^m$, $R(t,s) = f(t)f^*(s)$, is a positive definite function as well as a sum of different such products[18]. This leads to a simple way to parametrize an akf. For some suitable (basis) functions $\varphi_k : T \to \mathbb{R}^m$ form[19]

$$R(t,s) = \sum_{k=1}^{\infty} \lambda_k \varphi_k(t)\varphi_k^T(s), \quad \infty > \lambda_1 \geq \lambda_2 \geq \ldots \geq 0, \tag{3.16}$$

We may of course take only a finite number of basis functions to further simplify the parametrization.

An abstract version of this approach is to consider maps $\Phi : T \to \mathcal{H}^n$, where $\mathcal{H}^n$ is used to denote that $\Phi(t)$ is a vector where each element belongs to the Hilbert space $\mathcal{H}$, and define $\langle\Phi(t),\Phi(s)\rangle$ as the matrix with ijth element $\langle\Phi_i(t),\Phi_j(s)\rangle$. Taking

$$R(t,s) = \lfloor \Phi(t),\Phi(s) \rfloor$$

[15] A.N. Shiryaev. *Probability*. Springer, 2nd edition, 1989

[16]

[17] An equivalent definition of a positive definite function is that (3.15) holds for real vectors, and that K is symmetric.

[18] A further generalization is that if $\bar{R}$ is a positive definite function, then so is $f(t)\bar{R}(t,s)f^*(s)$

[19] That R is symmetric is trival and that it is non-negative follows from

$$\begin{aligned}&\sum_{i=1}^{m}\sum_{j=1}^{m} a^T(i)R(t_i,t_j)a(j)\\ &= \sum_{k=1}^{\infty}\lambda_k \sum_{i=1}^{m}\sum_{j=1}^{m} a^T(i)\varphi_k(t_i)\varphi_k^T(t_j)a(j)\\ &= \sum_{k=1}^{\infty}\lambda_k \left|\sum_{i=1}^{m} a^T(i)\varphi_k(t_i)\right|^2 \geq 0\end{aligned}$$

where $\lfloor\Phi(t),\Phi(s)\rfloor$ is the matrix with $\langle\Phi_i(t).\Phi_j(s)\rangle$ as element ij, results in a positive definite function since

$$\begin{aligned}\sum_{i,j} a^T(i)R(t_i,t_j)a(j) &= \sum_{i,j} a^T(i)\langle\Phi(t_i),\Phi(t_j)\rangle a(j)\\ = \sum_{i,j}\left\langle\sum_k a_k(i)\Phi_k(t_i), \sum_k a_k(j)\Phi(t_j)\right\rangle &= \left\|\sum_i a^T(i)\Phi(t_i)\right\| \geq 0\end{aligned}$$

where $\|\cdot\|$ denotes the norm in $\mathcal{H}$.

Some theory for positive definite functions.* For background material we refer to Appendix C. We will consider a particular choice of basis functions in (3.16). As preparations for this we introduce the following class of functions.

Definition 3.4.1. *$L_2^m(T)$ is the space of measurable functions, $f: T \to \mathbb{C}^m$ that are square integrable*

$$\int_T |f(t)|^2 dt < \infty$$

We now T to be a compact set[20] in $\mathbb{R}^n$. Equipped with the inner product

$$\langle f,g\rangle = \int_T g^*(t)f(t)dt \tag{3.17}$$

$L_2^m(T)$ is a separable Hilbert space. Let us now in (3.16) take $\{\varphi_k\}_{k=1}^\infty$ to be a complete orthonormal basis for $L_2^m(T)$ and $\{\lambda_k\}$ satisfying $\sum_{k=1}^\infty \lambda_k < \infty$. Let us also ensure that $R(t,s)$ is bounded by enforcing $\{\varphi_k\}_{k=1}^\infty$ to be uniformly bounded

$$|\varphi_k(t)| \leq C, \quad k = 1,2,\dots,\forall t \in T$$

Then I_R, defined as

$$I_R(f)(t) = \int_T R(t,s)f(s)ds,$$

is called an integral operator with kernel R. Notice that R is linear and since R is bounded it is well defined for functions f in $L_2(T)$. Now[21]

$$\begin{aligned}\|I_R(f)\|_2^2 = \langle I_R(f), I_R(f)\rangle &= \int_T \left|\int_T \overline{R(s,t)}f(t)ds\right|^2 dsdt\\ &\leq \int_T \int_T \|R(s,t)\|_F^2 ds \int_T |f(r)|^2 dr\, dt\\ &= \int_T \int_T |R(s,t)|_F^2 dsdt\, \|f\|_2^2\end{aligned}$$

where[22]

$$\begin{aligned}\int_T \int_T |R(s,t)|_F^2 dsdt &= \sum_{k=1}^\infty \sum_{l=1}^\infty \lambda_k\lambda_l \int_T \int_T \varphi_k^*(s)\varphi_l(s)\varphi_l^*(t)\varphi_k(s)dsdt\\ &= \sum_{k=1}^\infty \lambda_k^2 < \infty\end{aligned}$$

[20] Think about a closed interval $[a,b]$, $-\infty < a < b < \infty$.

[21] See Exercise 3.8 for the inequality step

[22] The inequality follows from our assumption that $\sum_{k=1}^\infty \lambda_k < \infty$.

so that

$$\frac{\|I_R(f)\|_2}{\|f\|_2} \leq \sum_{k=1}^{\infty} \lambda_k^2$$

This means that seen as a function, $I_R(f)$ belongs to $L_2(T)$ which in turn means that I_R, as an operator, defines a map from the functions in $L_2(T)$ to $L_2(T)$ which has finite induced norm[23] $\|I_R\|_2 \leq \sum_{k=1}^{\infty} \lambda_k^2$.

Using the same type of derivation as above, we also obtain

$$\langle I_R(f), f\rangle = \int_T \int_T f^*(t)R(t,s)f(s)dtds = \sum_{k=1}^{\infty} \lambda_k |\langle <\varphi_k, f\rangle|^2, \quad \forall f \in L_2(T)$$

This shows that

$$\int_T \int_T f^*(t)R(t,s)f(s)dtds \geq 0, \quad \forall f \in L_2(T) \tag{3.18}$$

A kernel with this property is called positive definite. In regards to the operator I_R, the above shows

$$\langle I_R(f), f\rangle = \langle f, I_R(f)\rangle \geq 0 \quad \forall f \in L_2(T)$$

Operators satisfying the first equality are said to be *self-adjoint*. A self-adjoint operator satisfying the inequality above is said to be positive. Self-adjoint positive operators can be seen as generalizations of linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^n$ corresponding to positive semi-definite Hermitian matrices $A \in \mathbb{C}^{n \times n}$, $A = A^* \geq 0$, for which it is well known that they have eigen-decomposition

$$A = \sum_{k=1}^{n} \lambda_k \boldsymbol{\varphi}_k \boldsymbol{\varphi}_k^*, \quad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$$

In fact, using the orthonormality of the basis functions,

$$I_R(\boldsymbol{\varphi}_l)(t) = \int_T \sum_{k=1}^{\infty} \lambda_k \boldsymbol{\varphi}_k(t) \boldsymbol{\varphi}_l^T(s) \boldsymbol{\varphi}_l(s) ds = \lambda_l \boldsymbol{\varphi}_l(t)$$

To summarize, the parametrization (3.16) where the basis functions form a basis in $L_2(T)$ lead to that apart from $R(t,s)$ being a positive definite function, it is also a positive definite kernel, which defines a positive integral operator which has $\{\lambda_k\}$ as eigenvalues and $\{\boldsymbol{\varphi}_k\}$ as eigenfunctions.

Let us now turn the reasoning around and start with a bounded kernel satisfying (3.18) but not necessarily of the form (3.16)[24]. The integral operator I_R is then still well-defined and positive for which we can define eigenvalues and eigenfunctions. An important result, known as Mercer's theorem [25], states that we can always decompose R as in (3.16). We state this result in the general setting of a finite measure space (of which a compact T together with the Lesbegue measure is an example).

Theorem 3.4.2 (Mercer's theorem. Theorem 3.a.1 in [26]). *Let (Ω, μ) be a finite measure space and $R_\infty(\Omega^2, \mu^2)$ be a positive definite kernel*[27].

[23] A linear operator with finite operator norm is called *bounded* which is equivalent to that the operator is continuous.

[24] This type of operator is a special case of what is known as Hilbert-Schmidt operator (add: Fukumizu).

[25] J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society, London*, 209:415–446, 1909; and H. K *Eigenvalue Distribution of Compact Operators*

[27] This is equivalent to that $I_R : L_2(\Omega, \mu) \to L_2(\Omega, \mu)$ is positive.

Then the eigenvalues λ_k of I_R are absolutely summable and the corresponding normalized eigenfunctions $\boldsymbol{\varphi}_k \in Ł_2(\Omega, \mu)$ form an orthonormal set and belong to $L_\infty(\Omega, \mu)$ with $\sup_k |\varphi_k|_\infty < \infty$ and

$$R(t,s) = \sum_{k=1}^{\infty} \lambda_k \boldsymbol{\varphi}_k(t) \boldsymbol{\varphi}_k^*(s), \text{ holds } \mu \text{ almost everywhere}$$

where the series converges absolutely and uniformly almost everywhere.

From a modeling perspective, Mercer's theorem tells us that the family of bounded positive definite kernels $T \times T \to \mathbb{R}^m$, with T compact is completely parametrized by the expansion (3.16) with $\{\boldsymbol{\varphi}_k\}$ being an orthonormal basis in $L_2(T)$ and $\sum_{k=1}^{\infty} \lambda_k < \infty$, $\lambda_k \geq 0$, $k = 1, 2, \ldots$. Since functions of the type (3.16) are positive definite functions but not necessarily the converse, the set of positive definite functions is larger than the set of positive definite kernels. However, for continuous positive definite functions a one-to-one relationship can be obtained.

Theorem 3.4.3 ([28])**.** *Let $T = [a, b]$ be a compact interval and let $R : T \times T \to \mathbb{C}$ be continuous. Then R is a positive definite function if and only if*

$$\int_T \int_T f(t) R(t,s) f(s) dt ds \geq 0 \tag{3.19}$$

for all complex-valued continuous functions f with domain of definition including T.

Now the set of continuous functions is dense in $L_2(T)$ and therefore (3.19) implies (3.18) and hence the theorem shows that the set of continuous positive definite functions $T \times T \to \mathbb{C}$, with T being a compact interval, is equal to the set of positive definite kernels[29].

Due to the strong link to the theory of positive definite kernels, positive definite functions, in the sense (3.15), are often called positive definite kernels.

3.4.6 Gaussian Processes

Above we have discussed how to model the akf of stochastic process. To simplify the modeling task we also need to restrict the class of probability measures. A Gaussian random vector $\mathbf{X}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} > 0$ has pdf

$$\mathcal{N}(\mathbf{x}; \mathbf{m}, \boldsymbol{\Sigma}) := \frac{1}{\sqrt{\det 2\pi\boldsymbol{\Sigma}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

and to indicate this we write $\mathbf{X} \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma})$. This distribution is fully characterized by its mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. A Gaussian Process (GP) $\{\mathbf{X}(t)\}$ is a stochastic process for which all finite dimensional distributions are of the form above, i.e.

$$\begin{bmatrix} \mathbf{X}(t_1) \\ \vdots \\ \mathbf{X}(t_n) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m}(t_1) \\ \vdots \\ \mathbf{m}(t_n) \end{bmatrix}, \begin{bmatrix} C(t_1, t_1) & \ldots & C(t_1, t_n) \\ \vdots & \ldots & \vdots \\ C(t_n, t_1) & \ldots & C(t_n, t_n) \end{bmatrix} \right), \quad \forall t_i$$

[29] This should generalize to matrix valued kernels.

Modeling considerations. A GP is fully characterized by its mean and covariance functions, or equivalently by its mean and auto-correlation functions. The covariance function $C(t,s)$ is known as the kernel of the process. While Mercer's theorem provides a general parametrization, these can also often quite easily be tailored to the specific behaviour of the quantity that is to be modelled.

A significant simplification is obtained by parametrizing the akf (and the covariance function) as $R(t,s) = R(t-s)$. This means that the akf does not depend on absolute time, implying that the second order properties of the process do not change over time. This brings us to the next class of stochastic processes.

3.4.7 *Stationary stochastic processes*

A *stationary process* is characterized by that its probability measures do not change with time shifts. This means that these processes are characterized by that their finite dimensional distribution functions satisfy

$$F_{t_1+\Delta,\dots,t_n+\Delta}(x_1,\dots,x_n) = F_{t_1,\dots,t_n}(x_1,\dots,x_n),$$

$\forall t_i \in T,\ t_i \neq t_j,\ \Delta + t_i \in T,\ 1 \leq i,j \leq n,\ n \in \mathbb{N}$. For a process for which the finite dimensional pdfs $p_{t_1,\dots,t_n}(x_{t_1},\dots,x_{t_n})$ exist, these conditions can be expressed as

$$p_{t_1+\Delta,\dots,t_n+\Delta}(x_{t_1},\dots,x_{t_n}) = p_{t_1,\dots,t_n}(x_{t_1},\dots,x_{t_n})$$

Stationarity is sometimes referred to as *strict stationarity.* If $\begin{bmatrix}\mathbf{X}^T(t) & \mathbf{Y}^T(t)\end{bmatrix}^T$ is stationary we say that $\mathbf{X}$ and $\mathbf{Y}$ are *jointly stationary.*

Stationarity implies that the mean function is independent of time so we use the notation $m_{\mathbf{X}}$. Furthermore, the cross-correlation function and corss-covariance function do not depend on time-shifts for jointly stationary processes X and Y and therefore we can introduce the following notation

$$\begin{aligned} R_{\mathbf{X},\mathbf{Y}}(\tau) &:= R_{\mathbf{X},\mathbf{Y}}(\tau,0) = R_{\mathbf{X},\mathbf{Y}}(t,t-\tau) \\ C_{\mathbf{X},\mathbf{Y}}(\tau) &:= C_{\mathbf{X},\mathbf{Y}}(\tau,0) = C_{\mathbf{X},\mathbf{Y}}(t,t-\tau) \end{aligned}$$

The positivity condition (3.15) becomes

$$\sum_{i=1}^{m}\sum_{j=1}^{m} a_i^* R(t_i - t_j) a_j \geq 0, \quad \forall a_i \in \mathbb{C}^n,\ t_i \in T,\ m \in \mathbb{N} \qquad (3.20)$$

which, using the symmetry requirement $R(\tau) = R^T(-\tau)$ (see Exercise 3.9), can be expressed as that

$$\mathbf{T} = \begin{bmatrix} R(t_1-t_1) & R(t_1-t_2) & \dots & R(t_1-t_m) \\ R^T(t_1-t_2) & R(t_2-t_2) & \dots & R(t_2-t_m) \\ \vdots & \vdots & \ddots & \vdots \\ R^T(t_1-t_m) & R^T(t_2-t_m) & \dots & R(t_m-t_m) \end{bmatrix} \geq 0 \qquad (3.21)$$

for all matrices for the above type. Notice that $\mathbf{T}$ has the same block along its block diagonals. Such a matrix is called a block Toeplitz matrix, and Toeplitz matrix when R is scalar. A function $R : T \to \mathbb{R}^n$ satisfying (3.20) is said to be positive definite.

Wide-sense stationarity. A stochastic process is said to be *wide-sense stationary* (weakly stationary) if the mean function does not depend on time and if the auto-correlation function does not depend on time shifts.

For Gaussian processes wide-sense and strict stationarity are equivalent as the distribution functions only depend on the mean and covariance functions.

Quasi-stationarity. Certain non-stationary processes may behave more and more like a stationary process as time increases, for example it may hold that $m_X(t) \to m_X$ for some finite number m_X. A special class of such processes are quasi-stationary signals. For this we need the following definition

$$\overline{\mathbb{E}}\left\{f(t)\right\} = \lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} \mathbb{E}\left[f(t)\right]$$

which is defined whenever the limit on the right exists.

Definition 3.4.2. $\mathbf{X}(t)$ *is said to be a quasi-stationary signal if*

$$\begin{aligned} |m_{\mathbf{X}}(t)| &\leq C \quad \forall t \\ |R_{\mathbf{X},\mathbf{X}}(t,s)| &\leq C \quad \forall t,s \\ R_{\mathbf{X},\mathbf{X}}(\tau) &:= \overline{\mathbb{E}}\left\{\mathbf{X}(t)\mathbf{X}^T(t-\tau)\right\}, \quad \textit{exists}\ \forall \tau \end{aligned}$$

Two signals $\mathbf{X}(t)$ *and* $\mathbf{Y}(t)$ *are said to be jointly quasi-stationary if* $\begin{bmatrix}\mathbf{X}^T(t) & \mathbf{Y}^T(t)\end{bmatrix}^T$ *is quasi-stationary.*

Notice that deterministic signals may be quasi-stationary. For such signals quasi-stationarity means that the auto-correlation function is formed by taking the average over time, whereas for wide-sense stationary processes the auto-correlation function is formed by taking the average over the different outcomes ω for fixed time points t and s. Under weak conditions it holds for a stationary stochastic process $\mathbf{X}(t)$ that $\overline{\mathbb{E}}\left\{\mathbf{X}(t,\omega)\mathbf{X}(t-\tau,\omega)\right\} = R_{\mathbf{X},\mathbf{X}}(\tau)$ for all $\omega \in \Omega$ except for a set of probability measure zero - this is due to the law of large numbers that we will return to.

Frequency domain characterization. Recall the discussion that led to (3.16), namely that a product $f(t)f^*(s)$ is a positive definite function. Applying this to $f(t) = e^{i\omega t}$ gives that $e^{i\omega(t-s)}$ is a positive definite function. We also notice that this function is of the type $R(t-s)$, meaning that $R(\tau) = e^{i\omega\tau}$ is a positive definite function, i.e. it satisfies (3.20). Thus, summing such functions (having different ω) with positive weights $Q(\omega) \geq 0$ will also give a positive definite function. Extending this reasoning to integrals leads to a precise frequency domain representation of positive definite functions due to Herglotz and Bochner.

As preparation for this we need to extend the concept of a distribution function to the matrix-valued case.

Definition 3.4.3. *F is a matrix valued distribution function on $[a,b]$ (or $\mathbb{R}$), if $F(a) = 0$ (or $\lim_{\omega\to-\infty} F(\omega) = 0$), F is right-continuous, $F(\omega) - F(\mu)$ is non-negative definite for all $\omega \geq \mu$.*

Theorem 3.4.4. *i) Herglotz theorem. $R : T \to \mathbb{R}^{m\times m}$, with $T = \mathbb{Z}$, is a positive definite function, i.e. it satisfies* (3.20) *if and only if*

$$R(\tau) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\omega\tau} dF(\omega)$$

where F is an $m \times m$ matrix valued distribution function on $[-\pi,\pi]$.

ii) Bochner's theorem. $R : T \to \mathbb{R}^{m\times m}$, with $T = \mathbb{R}$ is a continuous and positive definite function, i.e. it satisfies (3.20)*, if and only if*

$$R(\tau) = \int_{-\infty}^{\infty} e^{i\omega\tau} dF(\omega)$$

where F is an $m \times m$ matrix valued distribution function on $\mathbb{R}$.

Proof. See [30]. □

The matrix valued distribution function F in Theorem 3.4.4 is called the *spectral distribution function*. Thus an autocorrelation function can be parametrized by its spectral distribution function. Under restrictions on R there exists a simpler characterization.

Corollary 3.4.1. *i) Suppose that $R : T \to \mathbb{R}^{m\times m}$, with $T = \mathbb{Z}$, is absolutely summable*

$$\sum_{\tau=-\infty}^{\infty} \|R(\tau)\|_F < \infty \tag{3.22}$$

Then R is a positive definite function, i.e. it satisfies (3.20) *if and only if*

$$R(\tau) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\omega\tau} Q(\omega) d\omega \tag{3.23}$$

for some continuous function $Q \in L_1(\mathbb{R})$, satisfying $Q(\omega) \geq 0$, $\omega \in [-\pi,\pi]$.

ii) Suppose that $R : T \to \mathbb{R}^{m\times m}$, with $T = \mathbb{R}$ belongs to $L_1(\mathbb{R})$. Then R is a continuous and positive definite function, i.e. it satisfies (3.20)*, if and only if*

$$R(\tau) = \int_{-\infty}^{\infty} e^{i\omega\tau} Q(\omega) d\omega \tag{3.24}$$

for some continuous function Q, satisfying $Q(\omega) \geq 0$, $\forall\omega$.

Proof. An absolutely summable sequence $\{R(\tau)\}$ can be represented as a Fourier integral (3.23) where

$$Q(\omega) = \sum_{\tau=-\infty}^{\infty} R(\tau) e^{i\omega\tau}$$

(e.g. Theorem 4.3.2 in [31]). Positivity of Q follows from Corollary 4.3.2 in [32]. Finally, dominated convergence now gives that Q must be continuous. The converse follows directly from Theorem 3.4.4 as $F(\omega) = \int_{-\pi}^{\omega} Q(\mu) d\mu$ is a matrix valued distribution function.

[30] L.L. Gihman and A.V. Skorohod. *The Theory of Stochastic Processes I.* Springer-Verlag, Berlin, 1974

[31] P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods.* Springer, 1991

[32] P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods.* Springer, 1991

For $T = \mathbb{R}$, $R \in L_1(\mathbb{R})$ implies that its Fourier transform Q is continuous and vanishes at infinity, see Theorem 9.6 in [33]. Positivity of Q follows by a limit procedure similar to the proof of Corollary 4.3.2 in [34].

Conversely, if (3.24) holds, then $\Phi \in L_1$ and hence R is continuous and vanishes at infinity by Theorem 9.6 in [35]. Furthermore, Theorem 3.4.4 gives that R is a positive definite function. □

Remark 3.4.1. *The set of positive definite kernels form a closed convex cone*[36] *under the point-wise convergence topology. Herglotz/Bochner's theorem shows that the set of complex exponentials is the generator of the cone, i.e. any element of the cone can be obtained as a linear combination from this set.*

Remark 3.4.2. *The theorem can be extended to multivariable functions, e.g.* $T = \mathbb{R}^m$. *Then* $\{e^{i\omega^T t} : \omega \in \mathbb{R}^m\}$ *takes the role as generator.*

When R is the akf of a (wide-sense) stationary process, the function Q is called the *spectrum* (or spectral density) of the process. When the process is multivariate, an off-diagonal element of the spectrum is called the *cross-spectrum* between the corresponding elements of the process. The same terminology applies for two disjoint sub-vectors of the process. In view of that R is the inverse Fourier transform of Q, the spectrum is the discrete time Fourier transform of R. We can thus see the spectrum as a function of $e^{i\omega}$ when $T = \mathbb{Z}$ and of $i\omega$ when $T = \mathbb{R}$. We will use the notations $\Phi(e^{i\omega}) = Q(\omega)$ and $\Phi(i\omega) = Q(\omega)$, respectively, for spectra of akfs in these two cases. With

$$\Phi(z) = \sum_{\tau=-\infty}^{\infty} R(\tau) z^{-\tau}, \text{ and } \Phi(s) = \int_{-\infty}^{\infty} R(\tau) e^{-s\tau} d\tau, \tag{3.25}$$

we thus have that the spectrum is given by $\Phi(e^{i\omega})$ and $\Phi(i\omega)$, respectively. For $T = \mathbb{Z}$, with Φ being rational, $\Phi(z)$ is a Laurent-series expansion convergent in an annulus including the unit circle $|z| = 1$. When we do not want to distinguish between the cases $T = \mathbb{Z}$ and $T = \mathbb{R}$, we will use the generic notation $\Phi(\omega)$.

Notice that the spectrum is Hermitian $\Phi^*(e^{i\omega}) = \Phi(e^{i\omega})$, $\Phi^*(i\omega) = \Phi(i\omega)$, respectively, see Exercise 3.10, which in the scalar case means that $\Phi(e^{-i\omega}) = \Phi(e^{i\omega})$ and $\Phi(-i\omega) = \Phi(i\omega)$, respectively.

Since

$$\mathbb{E}\left[\mathbf{X}(t)\mathbf{X}^T(t)\right] = R_{\mathbf{X}}(0) = \begin{cases} \frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi_{\mathbf{X}}(e^{i\omega}) d\omega & T = \mathbb{Z} \\ \int_{-\infty}^{\infty} \Phi_{\mathbf{X}}(i\omega) d\omega & T = \mathbb{R} \end{cases}$$

we can interpret $\Phi(\omega)$ as providing the distribution of the signal power over different frequencies.

Modeling considerations. The spectrum characterization provides a very convenient way to parametrize the akf of a (wide-sense) stationary process, or a stationary Gaussian process. A straightforward

[33]

[34] P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods.* Springer, 1991

[35]

[36] A cone is a set C for which $x \in C \Rightarrow cx \in C$ if $c > 0$.

parametrization is to use $\mathcal{B}_k(\omega) \geq 0$ and then take

$$\Phi(\omega) = \sum_{k=1}^{\infty} \alpha_k \mathcal{B}_k(\omega), \ \alpha_k \geq 0, \ k = 1, 2, \ldots$$

Another straightforward approach is to form

$$\Phi(\omega) = \tilde{H}(\omega)\tilde{H}^*(\omega)$$

for some continuous function $\tilde{H}(\omega)$.

Probability density functions are positive so *characteristic functions*[37] of pdfs that are symmetric about the origin can also be used as autocorrelation functions.

Example 3.10. *The characteristic function for a $\mathcal{N}(0, \Sigma)$ distributed random variable is given by*

$$R(\tau) = e^{-\frac{1}{2}\tau^T \Sigma \tau}$$

Now, conversely, if the characteristic function of a pdf is a positive function, then the pdf is a positive definite function. Clearly, the characteristic function for a zero mean Gaussian is a positive function and hence the pdf (omitting constant factors)

$$e^{-\frac{1}{2}x^T \Sigma^{-1} x}$$

is a positive definite function. As it has the same form as the characteristic function, we did not obtain a new class of akf's in this case. This function is known as the *Gaussian kernel.*

Example 3.11. *The Laplace distribution with mean 0 and variance λ has characteristic function*

$$R(\tau) = \frac{1}{1 + \frac{1}{2}\lambda\tau^2}$$

which then is a positive definite function. As R also is a positive function, the corresponding pdf

$$p(x) = \frac{1}{\sqrt{2\lambda}} e^{-\sqrt{\frac{2}{\lambda}}|x|}$$

is a positive definite function. This is known as the Laplace kernel.

The preceding examples generalize to that

$$e^{-\alpha|\tau|^p}, \ \frac{1}{1 + \alpha|\tau|^p}, \quad \alpha > 0, \ 0 < p \leq 2$$

are positive definite functions.

For the discrete time case, $T = \mathbb{Z}$, one possibility is to use functions $H(z)$ that are holomorphic in an open set containing the unit circle and take $\tilde{H}(\omega) = H(e^{i\omega})$. Such functions can be expanded in Laurent series

$$H(z) = \sum_{k=-\infty}^{\infty} H_k z^{-k}$$

[37] The characteristic function of a pdf is its inverse Fourier transform.

The equivalent in the continuous time case, $T = \mathbb{R}$, is to use $\tilde{H}(\omega) = H(i\omega)$, where

$$H(s) = \sum_{k=-\infty}^{\infty} H_k s^{-k}$$

This approach can be given a filtering interpretation which we turn to next.

3.4.8 Filtered white noise

The scalar case.

Definition 3.4.4. *A sequence $\{e(t)\}_{t=-\infty}^{\infty}$ of uncorrelated random variables with zero mean and variance λ is called white noise.*

When the sequence is independent it is called strict white noise.

For a white noise process it holds that

$$R(\tau) = \lambda\delta(\tau)$$

Hence the spectrum is $\Phi(\omega) = \lambda$, i.e. white noise has flat spectrum, meaning that it contains all frequencies with an equal amount.

Now let

$$s(t) = H(q)e(t) = \sum_{k=-\infty}^{\infty} h(k)e(t-k),$$

where $H(q) = \sum_{k=-\infty}^{\infty} h(k)q^{-k}$ is stable in the sense that[38]

$$\sum_{k=-\infty}^{\infty} |h(k)|^2 < \infty$$

Then $s(t)$ is stationary with spectrum[39]

$$\Phi_s(\omega) = \lambda|H(e^{i\omega})|^2$$

Thus the frequency domain characteristics of the signal can be modelled by the filter characteristics. We now further restrict $H(q)$.

Rational filters. A very common parametrization is to let H be a rational function

$$H(z) = \frac{C(z)}{D(z)},$$
$$C(z) = c_0 + c_1 z^{-1} + \dots c_{n_c} z^{-n_c}, \; D(z) = d_0 + d_1 z^{-1} + \dots d_{n_d} z^{-n_d}$$

where the coefficients in C and D are real-valued. Using the fundamental theorem of algebra, we can then factorize H, giving

$$H(z) = \frac{c_0}{d_0} z^{n_d - n_c} \frac{\prod_{i=1}^{n_c}(z - z_i)}{\prod_{i=1}^{n_d}(z - p_i)}$$

where z_i and p_i, $i = 1, \dots, n$ are the zeros and poles of H. The spectrum of $s(t)$ is given by

$$\Phi_s(e^{i\omega}) = \lambda|H(e^{i\omega})|^2 = \frac{\lambda c_0^2}{d_0^2} \frac{\prod_{i=1}^{n_c}(e^{i\omega} - z_i)\overline{(e^{i\omega} - z_i)}}{\prod_{i=1}^{n_d}(e^{i\omega} - p_i)\overline{(e^{i\omega} - p_i)}}$$

[38] Notice that this condition is implied by the bounded-input bounded-output (BIBO) stability condition, see Definition 1.2.1.

[39]

$$\begin{aligned}
&\mathbb{E}\left[s(t)s(t-\tau)\right] \\
&= \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} h(k)h(l)\mathbb{E}\left[s(t-k)s(t-\tau-l)\right] \\
&= \lambda \sum_{k=-\infty}^{\infty} h(k)h(k-\tau)
\end{aligned}$$

which only depends on τ and not t. Furthermore, the limit is well defined since

$$\left|\sum_{k=-\infty}^{\infty} h(k)h(k-\tau)\right| \le \sqrt{\sum_{k=-\infty}^{\infty} h^2(k) \sum_{k=-\infty}^{\infty} h^2(k-\tau)} < \infty$$

where the first inequality is Cauchy-Schwarz inequality and the second inequality follows from that the space of square summable sequences (called ℓ_2) is contained in the space of absolute summable sequences (called ℓ_1). $s(t)$ is thus wide-sense stationary and it spectrum is

$$\begin{aligned}
\Phi_s(\omega) &= \lambda \sum_{\tau=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} h(k)h(k-\tau)e^{-i\omega\tau} \\
&= \lambda \sum_{\tau=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} h(k)e^{-i\omega k}h(k-\tau)e^{-i\omega(\tau-k)} \\
&= \lambda \sum_{k=-\infty}^{\infty} h(k)e^{-i\omega k} \sum_{l=-\infty}^{\infty} h(l)e^{i\omega l} \\
&= \lambda|H(e^{i\omega})|^2
\end{aligned}$$

from which we see that we can take $c_0 = d_0 = 1$ without loss of generality since we can obtain any scaling through the choice of the variance $\lambda > 0$. We also see that there is no need to have poles or zeros at the origin since the corresponding product, e.g. $(e^{i\omega} - z_i)\overline{(e^{i\omega} - z_i)}$, is the constant 1. Next, we notice that since the coefficients in C and D are real

$$|H(e^{i\omega})|^2 = H(z)H(z^{-1})|_{z=e^{i\omega}}$$

so let us study the factor $(z - z_i)(z^{-1} - z_i)$ in this expression

$$\begin{aligned}(z - z_i)(z^{-1} - z_i) &= (1 - z_i z^{-1})(1 - z_i z) = z_i^2(z_i^{-1} - z^{-1})(z_i^{-1} - z)\\ &= z_i^2(z - z_i^{-1})(z^{-1} - z_i^{-1})\end{aligned}$$

but this means that by replacing the zero z_i by z_i^{-1} as zero in H gives the same spectrum, save for a constant. Notice, however, that the constant may be adjusted by the noise variance λ, meaning that we can obtain the same spectrum using z_i or z_i^{-1} as zero in H. Now if $|z_i| \geq 1$, $|z_i^{-1}| \leq 1$. This means that for a certain rational spectrum we can choose to have a zero in the unit disc or in the complement of the interior of the unit disc. The reader may be worried that the factor z_i^2 is complex if z_i is a complex zero. However, complex zeros always appear in pairs, z_i and $\bar{z}_i$, if the coefficients of the polynomial are real, so if perform the same operation on both zeros we obtain the real-valued factor $|z_i|^4$. As we will see later that there are good reasons for assigning all zeros to be in the unit disc.

The same considerations apply to the poles. However, as the poles represent singularities of the function H, which is required to be well defined on the unit circle, they cannot have magnitude one. Thus, a pole p outside the unit disc gives rise to exactly the same spectrum as a pole located inside the unit circle at $1/p$. However, the choice of pole locations determine if the filter will be causal ($h_k = 0$, $k < 0$), anti-causal ($h_k = 0$, $k > 0$) or non-causal. For H to be causal the poles have to be strictly inside the unit circle so that causal expansions such as

$$\frac{1}{z - p_i} = \frac{1}{z}\frac{1}{1 - p_i/z} = \frac{1}{z}\sum_{k=0}^{\infty} p_i^k z^{-k}$$

are well defined on the unit circle $|z| = 1$. All poles strictly outside the unit disc gives an anti-causal filter, and poles both inside and outside the unit circle gives a non-causal filter.

Spectral factorization - Rational scalar spectra. Above we have seen that a rational filter gives a rational spectrum. So is the converse true, i.e. given a rational positive function on $\mathbb{T}$ can we realize that as a stable causal filter with all its zeros in the unit disc? The answer is affirmative and follows from the spectral factorization theorem, which is a simple application of the Fejér-Riesz theorem (see Section B.4.1).

Theorem 3.4.5 (Spectral factorization theorem - Scalar case). *A positive function $\Phi(e^{i\omega}) \geq 0$, defined for all $\omega \in [-\pi, \pi]$, that is rational in*

$z = e^{i\omega}$ can be factorized as

$$\Phi(e^{i\omega}) = \lambda |H(e^{i\omega})|^2, \; H(z) = \frac{C(z)}{D(z)} = \frac{1 + c_1 z^{-1} + \ldots c_{n_c} z^{-n_c}}{1 + d_1 z^{-1} + \ldots + d_{n_d} z^{-n_d}} \tag{3.26}$$

where $z^{n_c} C(z)$ has all its zeros in the unit disc and $z^{n_d} D(z)$ all its zeros inside the unit circle.

Furthermore, $\Phi(e^{i\omega}) > 0$, $\forall \omega$ is equivalent to that all zeros of $z^{n_c} C(z)$ can be taken inside the unit circle.

Proof. See Appendix 3.A. □

Autocorrelation functions with rational spectra. A rational spectrum represents a structure of the spectrum so one may wonder which structure this translates into when it comes to the akf. The answer to this is obtained by way of the spectral factorization theorem which implies that a rational spectrum can be seen as if the stationary process is obtained by filtering white noise through a rational filter. Now a rational filter has a finite-dimensional state-space representation and from this representation we can obtain the structure of the akf. We begin with a general result for the auto-correlation functions for signals generated by a state-space model.

Lemma 3.4.1. *Let*

$$\begin{aligned} x(t+1) &= Ax(t+1) + w(t), \quad \mathbb{E}\left[x(0)\right] = m_x(0), \; \mathbb{E}\left[x(0)x^T(0)\right] = P(0) \\ y(t) &= Cx(t) + v(t), \quad \mathbb{E}\left[\begin{bmatrix} w(t+\tau) \\ v(t+\tau) \end{bmatrix} \begin{bmatrix} w(t) \\ v(t) \end{bmatrix}^T\right] = R\delta(\tau) \end{aligned}$$

where $w(t)$ and $v(t)$ have zero mean, and where

$$R = \begin{bmatrix} R_{wv} & R_{wv} \\ R_{vw} & R_{vv} \end{bmatrix}$$

Let $P(t) := R_x(t,t)$. Then

$$\begin{aligned} m_x(t+1) &= Am_x(t) \\ P(t+1) &= AP(t)A^T + R_{ww} \\ R_x(t+\tau, t) &= A^\tau P(t) \\ R_y(t,t) &= CP(t)C^T + R_{vv} \\ R_y(t+\tau, t) &= CA^{\tau-1}(AP(t)C^T + R_{wv}), \quad \tau > 0 \end{aligned}$$

For the stationary case we have the following result.

Corollary 3.4.2. *Adding that $\begin{bmatrix} x^T(t) & y^T(t) \end{bmatrix}^T$ is wide-sense stationary to the assumptions in Lemma 3.4.1, it holds that*

$$\begin{aligned} m_x &= 0 \\ P(t) &= P := APA^T + R_{ww} \qquad (3.27) \\ R_x(\tau) &= A^\tau P \\ R_y(0) &= CPC^T + R_{vv} \\ R_y(\tau) &= CA^{\tau-1}(APC^T + R_{wv}), \; \tau > 0 \qquad (3.28) \end{aligned}$$

The equation (3.27) is called a Lyapunov equation and has a unique solution when F has all its eigenvalues inside the unit circle[40].

Notice that a state-space representation of a transfer function is not unique, but that the akf of the output is unique. Thus, regardless of the realization (A, C, R), R_y defined by the above equations will be the same, and the same holds for the spectrum. Let us define $D = (CPC^T + R_{vv})/2$, and define $B := APC^T + R_{wv}$. Then

$$G(z) := \sum_{k=1}^{\infty} R_Y(\tau)z^{-\tau} + D = \sum_{k=1}^{\infty} CA^{\tau-1}Bz^{-\tau} + D = C(zI - A)^{-1}B + D \tag{3.29}$$

and hence the spectrum for $y(t)$ is given by

$$\Phi_y(z) = G(z) + G^T(z^{-1}) \tag{3.30}$$

The positive real part of a spectrum. The split (3.30) can be made for any spectrum (3.25) by taking

$$G(z) = \frac{1}{2}R(0) + \sum_{k=1}^{\infty} R(k)z^{-k}$$

so that

$$0 \le \Phi(e^{i\omega}) = G(e^{i\omega}) + G(e^{-i\omega}) = G(e^{i\omega}) + G^*(e^{i\omega}) = 2\mathrm{Re}\left\{G(e^{i\omega})\right\}$$

For a rational spectrum, we observe that G must have the same poles as H and therefore the order of G is the same as the order of H[41]

A function $G(z)$ satisfying $\mathrm{Re}\left\{G(e^{i\omega})\right\} \ge 0$ is said to be *positive real* (PR). Design of spectra can thus be done indirectly by designing positive real functions. A positive real function can also be taken as starting point for the spectral factorization theorem.

Corollary 3.4.3. *Suppose that Φ is a positive function that can be written as*

$$\Phi(e^{i\omega}) = G(e^{i\omega}) + G^*(e^{i\omega}) \tag{3.31}$$

where $G(z)$ is a rational function with poles strictly inside the unit circle. Then Φ can be factored as (3.26) *where H has all zeros inside the unit circle and the same poles as G.*

Proof. We refer to the reference given in the proof of Lemma 3.4.2 □

We will continue this discussion in the context of multivariable spectra.

The multivariable case. With $G \in L_2$, by $G(q)$ we mean the non-causal transfer function defined by the Fourier coefficients of G: $G(q) := \sum_{k=-\infty}^{\infty} g(k)q^{-k}$, and we do the same for $G \in H_2$ giving causal (or one-sided) transfer functions $G(q) = \sum_{k=0}^{\infty} g(k)q^{-k}$.

The idea of filtering white noise is straightforward to extend from the scalar case to the multivariable case, but also to filtering of a

[40] This means that the system is stable which is required for wide-sense stationarity.

[41] G is characterized in Lemma 4.3 in .
T. Söderström. *Discrete-Time Stochastic Systems. Estimation and control.* Prentice-Hall International, New York, 1994

wide-sense stationary process or a quasi-stationary process. However, so far we have ignored to exactly define what we mean with an expression like

$$y(t) = \sum_{k=-\infty}^{\infty} g(k)u(t-k)$$

when $\{u(t)\}$ is a wide-sense stationary process. When only a finite number of the impulse response coefficients are non-zero, the right-hand side is simply a sum of random vectors and hence well defined, but in the general case we need to define what we mean by the limit

$$\lim_{N\to\infty} \sum_{k=-N}^{N} g(k)u(t-k)$$

It would seem natural to define this as the limit for every outcome ω in the probability space on which $\{u(t)\}$ is defined, i.e.

$$y(t,\omega) = \sum_{k=-\infty}^{\infty} g(k)u(t-k,\omega) \tag{3.32}$$

For this expression to be meaningful we need to introduce some condition on the impulse response.

Returning to the question when the filtering expression (3.32) is meaningful when the input is a sequence of random variables, this would work well when the system is BIBO-stable and the realization $\{u(t,\omega)\}$ is bounded. However, we will often consider probabilistic models of signals consisting of random variables with infinite support, e.g. Gaussian random variables, where realizations are unbounded with probability 1, see Exercise 3.11. So can we still define the limit in a meaningful way in such a case? To answer this we first need the following definition.

Definition 3.4.5. *Let $\{X_N\}$ be a sequence of random variables. If there exists a set A with $\boldsymbol{P}(A) = 1$ and*

$$\lim_{N\to\infty} X_N(\omega) = 0$$

holds for all $\omega \in A$ then we say that X_N converges to 0 with probability one (w.p.1).

Alternative terminology is that X_N converges to 0 *almost everywhere* (a.e.) or *almost surely* (a.s.). If $X_N - X$ tends to zero w.p.1 we say that X_N tends to X w.p.1.

Theorem 3.4.6. *Let*

$$z_N(t) = \sum_{\tau=0}^{N} g(\tau)v(t-\tau)$$

where $\{g(\tau)\}$ is strictly stable and where

$$\mathbb{E}\left[|v(k)|\right] \le C \qquad \forall k. \tag{3.33}$$

Then $Z_N(t)$ converges to some random variable as $N \to \infty$ w.p.1.

Proof. See Appendix 3.B. □

Theorem 3.4.6 can be phrased in many ways. One can trade conditions on the input $\{v(t)\}$ for conditions on the system $\{g(k,)\}$.

Corollary 3.4.4. *The conclusion of Theorem 3.4.6 remain true if* $\{v(k)\}$*,* $k \leq t$*, is white noise and* $\{g_k\}$ *is strictly stable.*

There are several types of convergence besides convergence w.p.1 that can be defined in probability theory. One that will be particularly useful for us concerns random variables in the following class.

Definition 3.4.6. $\mathcal{L}_2^m = \mathcal{L}_2^m(\Omega, \mathcal{F}, \boldsymbol{P})$ *consists of the random vectors* $\mathbf{X}$ *defined on* $(\Omega, \mathcal{F}, \boldsymbol{P})$*,* $X : \Omega \to \mathbb{C}^m$*, for which* $\mathbb{E}\left[\|\mathbf{X}\|_F^2\right] < \infty$*.*

We can make $\mathcal{L}_2^m$ into a Hilbert space by introducing the inner product

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \mathbb{E}\left[\mathbf{Y}^* \mathbf{X}\right] = \int Y(\omega)^* X(\omega) \mathbf{P}(d\omega) \qquad (3.34)$$

so that the norm is $\|\mathbf{X}\|_2 = \sqrt{\mathbb{E}\left[\mathbf{X}^* \mathbf{X}\right]}$. It can be shown that

$$\lim_{n,m\to\infty} \|\mathbf{X}_n - \mathbf{X}_m\| = 0 \quad \Leftrightarrow \quad \exists \mathbf{X} : \|\mathbf{X}\|_2 < \infty, \lim_{n\to\infty} \|\mathbf{X}_n - \mathbf{X}\|_2 = 0$$

i.e. the criterion on the left, known as the Cauchy criterion, is equivalent to convergence. This means that $\mathcal{L}_2^n$ is complete. However, we cannot distinguish two random variables in $\mathcal{L}_2^m$ from each other if the difference has square expectation zero and hence the limit X above is only unique a.e. Uniqueness is obtained by dividing $\mathcal{L}_2^n$ into equivalence classes of random variables that are equal a.e. With this construct $\mathcal{L}_2^n$ is a Hilbert space. We thus need to keep in mind that random variables $\mathbf{X}$ and $\mathbf{Y}$ for which $\mathbb{E}\left[|\mathbf{X} - \mathbf{Y}|^2\right] = 0$ are considered to be the same. Another way of phrasing this is that when we say that a random variable is 0, it need not be identically zero for all outcomes but it only has to have zero variance, cf. the discussion in Section 1.1.2.

$\mathcal{L}_2^m$ can be seen as a space of functions $X : \Omega \to \mathbb{R}^m$ (that we happen to call random variables), just like L_2^m defined in Definition 3.4.1. Comparing (3.34) with (3.17) we see that the essential difference between $L_2^m(T)$ and $\mathcal{L}_2^m$ is the weighting with $\mathbf{P}(d\omega)$ in the inner-product. However, this does not change the principal properties of the spaces.

We can now talk about convergence in $\mathcal{L}_2^m$.

Definition 3.4.7. *We say that a sequence* $\{X_N\}$ *of random vectors converges to 0 in* $\mathcal{L}_2^m$ *if* $X_N \in \mathcal{L}_2^m$*,* $N = 1, 2, \ldots$ *and*

$$\|X_N\|_2 \to 0 \quad as \quad N \to \infty$$

As for convergence w.p.1, if $\{\mathbf{X}_N - \mathbf{X}\}$ converges to 0 in $\mathcal{L}_2^m$ we say that $\{\mathbf{X}_N\}$ converges to $\mathbf{X}$ in $\mathcal{L}_2^m$.

A useful convergence condition is that a sequence in $\mathcal{L}_2^m$ converges if and only if it is a Cauchy sequence in $\mathcal{L}_2^m$ since one does then not have to specify what the limit is. With this criterion it is easy to show that the limit $y(t) = G(q)u(t)$, with $\{u(t)\}$ being white noise and $G(e^{i\omega}) \in L_2(\mathbb{T})$ exists as element in $\mathcal{L}_2$, see Exercise 3.12.

Theorem 3.4.7. *Let*

$$y(t) = G(q)u(t)$$

where $G(q) = \sum_{k=-\infty}^{\infty} g(k)q^{-k}$.

i) Suppose that $\{u(t)\}$ is quasi-stationary process with spectrum Φ_u and let $G(q)$ have real valued impulse response $\{g(\tau)\}$ and be BIBO stable. Then $\{y(t)\}$ is quasi-stationary with spectrum

$$\Phi_{yy}(e^{i\omega}) = G(e^{i\omega})\Phi_{uu}(e^{i\omega})G^*(e^{i\omega}) \tag{3.35}$$

where X^* denotes the complex conjugate transpose of X.

ii) Suppose that $\{u(t)\}$ is a wide-sense stationary process with spectral distribution function F_u. Then $\{y(t)\}$ is wide-sense stationary with spectral distribution function F_y satisfying

$$dF_y(e^{i\omega}) = G(e^{i\omega})dF_u(e^{i\omega})G^*(e^{i\omega})$$

if and only if

$$\int_{-\pi}^{\pi} G(e^{i\omega})dF_y(e^{i\omega})d\omega < \infty$$

In particular, if $\{u(t)\}$ has spectrum Φ_{uu}, the spectrum of $\{y(t)\}$ is given by (3.35).

Proof. For i), see Appendix 2A in [42] for the scalar quasi-stationary case and for ii) see Theorem 1.1, Chapter 2 in [43] for the multivariable stationary case. □

Corollary 3.4.5. *Suppose that $\{u(t)\}$ is a wide-sense stationary process with spectrum $\Phi_u \in L_\infty(\mathbb{T})$ and that $G \in L_\infty(\mathbb{T})$. Then $y(t) = G(q)u(t)$ is a wide-sense stationary process with spectrum* (3.35).

Proof. Theorem 1.2, Chapter 2 in [44]. □

We have seen that scalar rational spectra correspond to akfs given by (3.28). A natural extension of rational spectra to the multivariable case is thus spectra which correspond to akfs with this structure. However, this structure translates into (3.30) with $G(z) = C(zI - A)^{-1}B + D$ for certain matrices (A, B, C, D). We can thus see (3.30) as a parametrization of a rational multivariable spectrum. The Positive Real lemma establishes exact conditions on A, B, C and D for $G(z)$ to be positive real, and hence for $\Phi(e^{i\omega})$ to be a spectrum.

Lemma 3.4.2 (Positive Real Lemma)**.** *Given $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $M = M^T \in \mathbb{R}^{(n+m)\times(n+m)}$, with* $\det(e^{i\omega}I - A) \neq 0$ *for $\omega \in \mathbb{R}$ and (A, B) controllable, the following two statements are equivalent:*

i)

$$\begin{bmatrix} (e^{i\omega}I - A)^{-1}B \\ I \end{bmatrix}^* M \begin{bmatrix} (e^{i\omega}I - A)^{-1}B \\ I \end{bmatrix} \geq 0, \quad \forall \omega \in \mathbb{R} \cup \{\infty\}$$

[42] L. Ljung. *System identification, Theory for the user*. System sciences series. Prentice Hall, Upper Saddle River, NJ, USA, second edition, 1999

[43] P.E. Caines. *Linear stochastic systems*. SIAM, 2018

[44] P.E. Caines. *Linear stochastic systems*. SIAM, 2018

ii) There exists a symmetric matrix $P \in \mathbb{R}^n \times n$ *such that*

$$M + \begin{bmatrix} P - A^T PA & -A^T PB \\ -B^T PA & -B^T PB \end{bmatrix} \geq 0$$

The corresponding equivalence for strict inequalities holds even if (A, B) *is not controllable.*

Proof. See [45]. □

Corollary 3.4.6. *Let* $G(z) = C(zI - A)^{-1}B + D$ *where* (A, B) *is controllable. Then G is positive real if and only if there exists a real symmetric matrix P such that*

$$\begin{bmatrix} P - A^T PA & C^T - A^T PB \\ C - B^T PA & D + D^T - B^T PB \end{bmatrix} \geq 0$$

Proof. Take

$$M = \begin{bmatrix} 0 & C \\ C^T & D + D^T \end{bmatrix}$$

in Lemma 3.4.2. □

In the continuous time case, G is positive real if $G(i\omega) + G^*(i\omega) \geq 0$ $\forall \omega$.

Lemma 3.4.3. *Given* $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $M = M^T \in \mathbb{R}^{(n+m)\times(n+m)}$, *with* $\det(i\omega I - A) \neq 0$ *for* $\omega \in \mathbb{R}$ *and* (A, B) *controllable, the following two statements are equivalent:*

i)

$$\begin{bmatrix} (i\omega I - A)^{-1}B \\ I \end{bmatrix}^* M \begin{bmatrix} (i\omega I - A)^{-1}B \\ I \end{bmatrix} \geq 0, \quad \forall \omega \in \mathbb{R} \cup \{\infty\}$$

ii) There exists a symmetric matrix $P \in \mathbb{R}^n \times n$ *such that*

$$M - \begin{bmatrix} A^T P + PA & PB \\ B^T P & 0 \end{bmatrix} \geq 0$$

The corresponding equivalence for strict inequalities holds even if (A, B) *is not controllable.*

Proof. See [46]. □

Corollary 3.4.7. *Let* $G(s) = C(sI - A)^{-1}B + D$ *where* (A, B) *is controllable. Then G is positive real if and only if there exists a real symmetric matrix P such that*

$$\begin{bmatrix} -PA - A^T P & C^T - PB \\ C - B^T P & D + D^T \end{bmatrix} \geq 0$$

Proof. Take

$$M = \begin{bmatrix} 0 & C \\ C^T & D + D^T \end{bmatrix}$$

in Lemma 3.4.3. □

[45] A. Rantzer. On the Kalman-Yakubovich-Popov lemma. *Systems & Control Letters*, 28:7–10, 1996

[46] J.C. Willems. Least squares stationary optimal control and the algebraic Riccati equation. *IEEE Trans. Aut. Control*, 16:621–634, 1971; and A. Rantzer. On the Kalman-Yakubovich-Popov lemma. *Systems & Control Letters*, 28:7–10, 1996

Theorem 3.4.8 (Spectral factorization - Multivariable finite dimensional case)**.** *Let $\Phi(z)$ be the spectrum of a stationary process given by* (3.30)*, with G having the structure* (3.29)*, and assume that $\Phi(z)$ is full rank for almost all z. Then there is a factorization*

$$\Phi(z) = H(z)\Sigma H^T(z^{-1})$$

where $H(z)$ is a square real rational transfer function matrix with all poles inside the unit circle, $\lim_{z\to\infty} H(z) = I$*, $H^{-1}(z)$ has all its poles in the unit disc, and $\Sigma > 0$. The factorization in unique.*

Proof. Theorem 4.1 in Chapter 9 in [47]. □

Remark 3.4.3. *That $\Phi(z)$ is full rank for almost all z is equivalent to that a Gaussian process with spectrum Φ cannot be generated as filtered white noise where the covariance matrix of the driving noise has rank lower than the dimension of the process.*

The way to construct H from the positive real part of the spectrum is given in the following corollary to Theorem 3.4.8.

Corollary 3.4.8. *Let $G(z) = C(zI - A)^{-1}B + D$ be the positive real part of the full rank spectrum Φ. Then $H(z)$ and Σ in Theorem 3.4.8 are given by*

$$\begin{aligned} H(z) &= I + C(zI - A)^{-1}K,\ K = -(ATC^T - B)\Sigma^{-1} \\ \Sigma &= 2D - CTC^T \end{aligned}$$

where T is the solution to

$$T = ATA^T + (ATC^T - B)(2D - CTC^T)^{-1}(ATC^T - B)^T$$

The general case. We will continue to elaborate on the structure of (wide-sense) stationary stochastic processes and their spectral factorization at the end of Chapter 4, when we have acquired the necessary tools from estimation theory.

3.4.9 Markov Processes

In its simplest form a *Markov Process* (MP) is described by a finite set of states $S = \{s_1, \ldots, s_{n_s}\}$ and a matrix of transition probabilities $\mathbf{P} \in \mathbb{R}^{n_s \times n_s}$, where the $i : j$th element is the probability of moving from state i to state j. To specify such a model thus corresponds to defining an $n \times n$ matrix with elements in the interval $[0, 1]$ such that the row sums are 1.

The model can be generalized to allow for that **P** varies with time. A further generalization is to allow for that the transition probabilities at a given time depends on some external action taken at that time. This leads to the richer class of *Markov Decision Processes* (MDP).

Yet, another generalization is to model the observations of the states. A simple approach is to consider a finite set of possible observations $\{o_1, \ldots, o_{n_o}\}$. The model for the observations then consists of

[47] B. D. O. Anderson and J.B. Moore. *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, New Jersey, 1979

a matrix $\mathbf{P}_o \in \mathbb{R}^{n_s \times n_o}$ where the ijth element is the probability of observing o_j given that the state is s_i. This is known as a *Hidden Markov Model* (HMM).

Combining an observation model and a model for the transition probabilities including actions leads to what is known as a *Partially Observed Markov Decision Process* (POMDP).

These models can be generalized to include a continuum of states and observations. A stochastic state-space model

$$\begin{aligned} x(t+1) &= f(x(t), u(t), w(t)) \\ y(t) &= h(x(t), u(t), v(t)) \end{aligned}$$

where w and v are stochastic processes and u a user controlled variable, is of this type.

3.4.10 A swatch of building blocks

Above, we have provided building blocks for constructing probabilistic models. Now, a disturbance may behave significantly different from an impulse response of a linear system. Different quantities may also have different domains, while disturbances and impulse responses are defined in the time domain, which is discrete for sampled data systems, the static non-linearity in a mechanical system has $\mathbb{R}$ as domain. Recalling that the better one can rank the actual behaviour, the better will the model be, it is important to use different types of models for quantities with different characteristics. The purpose of this section is to provide a brief swatch of commonly used models.

Let us see what we can achieve with a Gaussian Process with zero mean and kernel K

$$f(\cdot) \sim \mathcal{N}(0, K(\cdot, \cdot))$$

To emphasize the type of quantity we are working with we will use the kernel notation

$$K(f(x), f(y)) = \mathbb{E}\left[f(x)f(y)\right]$$

Disturbances and noise. Disturbances and noise often exhibit stationary behaviour over time. However, they may still have different characteristics in terms of how quickly they vary over time.

This type of behaviour can conveniently be modelled as a stationary stochastic process where a filter H is used to model the frequency behaviour. For a disturbance (or noise) sequence $\{v(t)\}$ this gives

$$K(v(t), v(s)) = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{i\omega}|^2 e^{i\omega(t-s)} d\omega$$

Taking a rational filter, the zeros and poles can be tuned so that some frequency domain characteristics are obtained. The higher the filter order, the more degrees of freedom to tweak the behaviour there will be. The two filters with Bode diagrams in Figure 3.3 generate the realizations in Figure 3.4.

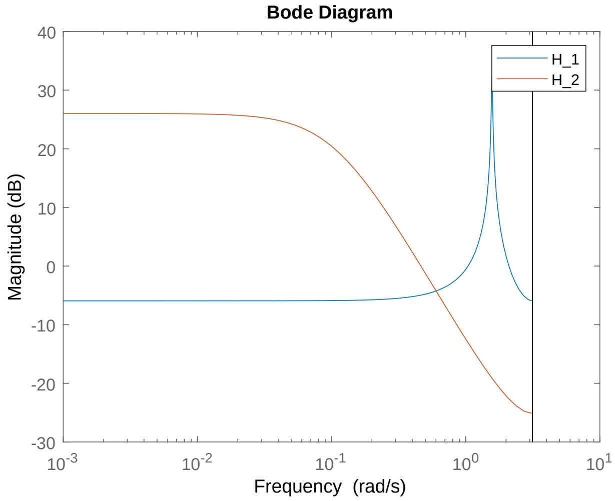

Figure 3.3: Bode diagrams of two filters.

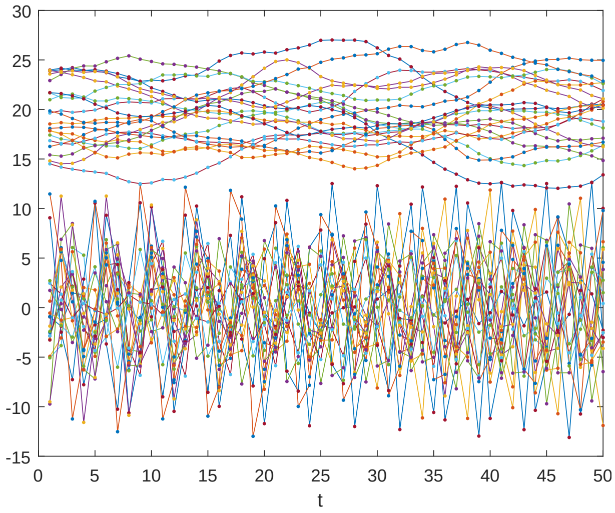

Figure 3.4: Example of realizations from two stochastic processes generated by two different filters H_1 and H_2. The mean has been shifted to 20 for the process corresponding to H_2 to improve visibility.

Impulse responses of stable linear systems. An impulse response $\{g(t)\}$ for a stable finite-dimensional linear system decays exponentially, with the rate of decay determined by the pole closest to the unit circle. The stable spline kernel

$$K(g(t), g(s)) = \eta_1\, \eta_2^{\min(t,s)}, \quad |\eta_2| < 1$$

models this type of non-stationary behaviour. Two features are modelled: i) the impulse response decays over time since $|\eta_2| < 1$, and ii) the response at one point in time is correlated with the response at other time instances in the same way as the response decays, through the use of $\min(t, s$. Figure 3.5 shows realizations for two different values of η_2.

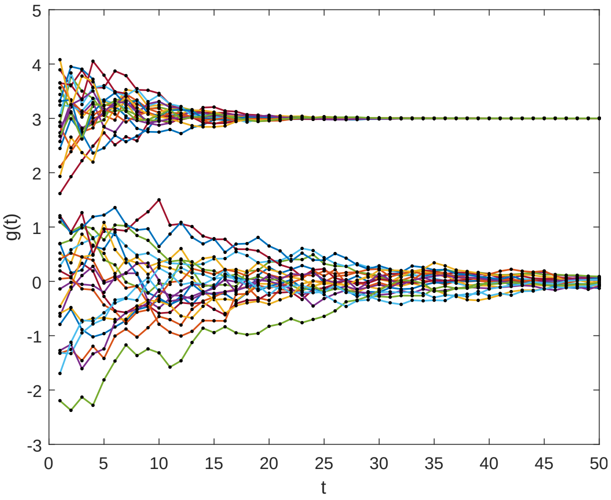

Figure 3.5: Realizations from the GP with the stable spline kernel with $\eta_2 = 0.7$ and $\eta_2 = 0.9$, where the latter leads to slower decay. An offset of 3 has been provided to the former to help visibility.

Gaussian kernel. A Gaussian kernel is often used when modeling a non-linear function $f : \mathbb{R} \to \mathbb{R}$

$$K(f(x), f(y)) = \eta_1\, e^{-\frac{|x-y|^2}{2\eta_2}},\ \eta_1 > 0,\ \eta_2 > 0$$

This means that the correlations between function values decreases exponentially with the squared distance between the points. Figure 3.6 illustrates this.

3.5 Summary

3.6 Exercises

3.1. Let $\hat{\boldsymbol{\xi}}_{MAP}(\mathbf{z})$ and $\hat{\boldsymbol{\xi}}_{PM}(\mathbf{z})$ be the MAP and PM estimators of $\boldsymbol{\xi}$. Let $\gamma = \gamma(\boldsymbol{\xi})$ be a function of the model parameters.

a. Define the maximum a posteriori estimate of γ as

$$\hat{\gamma}_{MAP}(\mathbf{z}) = \arg\max_{\gamma} p(\gamma, \mathbf{z})$$

Show by counterexample that $\hat{\gamma}_{MAP}(\mathbf{z}) \neq \gamma(\hat{\boldsymbol{\xi}}_{MAP}(\mathbf{z}))$ may hold. Show also that when γ is injective (one-to-one) equality holds.

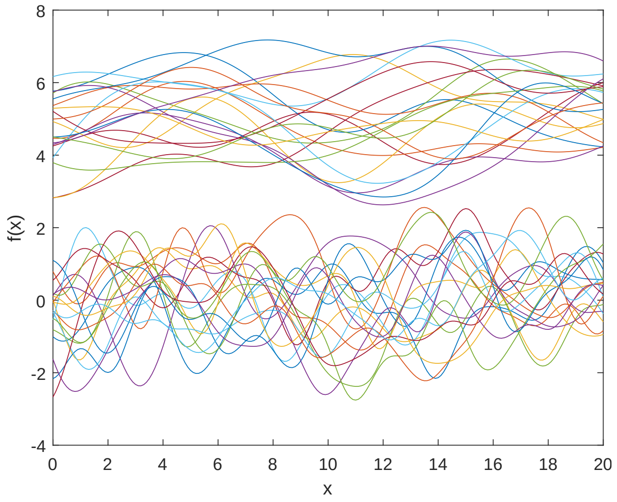

Figure 3.6: Realizations from the GP with a Gaussian kernel with $\eta_2 = 1$ and $\eta_2 = 10$. An offset of 5 has been added to the latter to help visibility.

b. Define the posterior mean of $\gamma(\xi)$ as

$$\bar{\gamma} = \int \gamma p(\gamma|\mathbf{z}) d\gamma$$

Show that this is the same estimator of γ as (3.3).

3.2. Let $p(\boldsymbol{\xi}, \mathbf{z})$ be the joint pdf for the model parameters $\boldsymbol{\xi}$ and the observations $\mathbf{z}$, and let $\hat{\boldsymbol{\xi}}_{MAP}$ and $\hat{\boldsymbol{\xi}}_{PM}$ be the MAP and PM estimators of $\boldsymbol{\xi}$, respectively. Suppose now that $\mathbf{q} = f(\mathbf{z})$, where f is one-to-one, is used as observations instead.

a. Determine the MAP estimator of $\boldsymbol{\xi}$.

b. Determine the PM estimator of $\boldsymbol{\xi}$.

3.3. Consider the state-space model

$$\begin{aligned} \mathbf{x}(t+1) &= \mathbf{A}(\boldsymbol{\theta})\mathbf{x}(t) + \mathbf{w}(t), \ \mathbf{x}(0) = 0 \\ \mathbf{y} &= \mathbf{C}(\boldsymbol{\theta})\mathbf{x}(t) + \mathbf{v}(t) \end{aligned} \tag{3.36}$$

where $\{\mathbf{w}(t)\}$ and $\{\mathbf{v}(t)\}$ are i.i.d. sequences of normal distributed random variables with zero mean and covariance $\mathbf{I}$.

a. Show that the relation between $\mathbf{y} := \begin{bmatrix}\mathbf{y}^T(1) & \dots & \mathbf{y}^T(N)\end{bmatrix}^T$, $\mathbf{x} := \begin{bmatrix}\mathbf{x}^T(1) & \dots & \mathbf{x}^T(N)\end{bmatrix}^T$ and $\mathbf{v} := \begin{bmatrix}\mathbf{v}^T(1) & \dots & \mathbf{v}^T(N)\end{bmatrix}^T$ can be expressed as

$$\mathbf{y} = \mathbf{H}(\boldsymbol{\theta})\mathbf{x} + \mathbf{v}, \quad \mathbf{x} \sim \mathcal{N}(0, \mathbf{R}_x(\boldsymbol{\theta})),\ \mathbf{v} \in \mathcal{N}(0, \mathbf{I})$$

for suitable choice of matrices $\mathbf{H}(\boldsymbol{\theta})$ and $\mathbf{R}_x(\boldsymbol{\theta})$.

b. Determine the covariance matrix $\mathbf{R}_y(\boldsymbol{\theta})$ for $\mathbf{y}$.

c. Determine the negative log-likelihood function when $\mathbf{y}$ is the observed variable.

d. Let $\mathbf{R}$ be a square non-singular matrix with matrix square root $\mathbf{R}^{1/2}$. Determine the negative log-likelihood function when $\mathbf{R}^{-1/2}\mathbf{y}$ is considered as the observation. Is it the same as in 3.3.c?

e. Determine the negative log-likelihood function when $\mathbf{R}_y^{-1/2}(\boldsymbol{\theta})\mathbf{y}$ is considered as the observation. Is it the same as in 3.3.c?

f. Compare the results in 3.3.d and 3.3.e.

g. Let $\mathbf{R}_y(\boldsymbol{\theta}) = \mathbf{L}(\boldsymbol{\theta})\mathbf{D}(\boldsymbol{\theta})\mathbf{L}^T(\boldsymbol{\theta})$ be the unique LDL-decomposition of $\mathbf{R}_y(\boldsymbol{\theta})$, i.e. $\mathbf{L}$ is lower unit[48] triangular and $\mathbf{D}$ positive definite diagonal. What is the negative log-likelihood function when the observation is considered to be $\mathbf{L}^{-1}(\boldsymbol{\theta})\mathbf{y}$? Is it the same as in 3.3.c? Compare this result with that in 3.3.e. What is the negative log-likelihood when the observation is taken as $L^{-1}(\boldsymbol{\theta})\mathbf{D}^{-1/2}(\boldsymbol{\theta})\mathbf{y}$?

h. What can you conclude from the above in regards to how parameter dependent transformations change the likelihood function?

i. Determine the smoothed estimate, $\mathbf{x}_s$ say, of $\mathbf{x}$ given $\mathbf{y}$.

j. In view of (3.36), it may seem reasonable to construct an estimator such that the difference between $y(t)$ and $\mathbf{C}(\boldsymbol{\theta})\mathbf{x}_s(t)$, where $\mathbf{x}_s(t)$ is the smoothed estimate of $\mathbf{x}(t)$, is small for $t = 1,\dots,N$. Argue that this corresponds to basing the estimator on the difference $\mathbf{y} - \mathbf{H}(\boldsymbol{\theta})\mathbf{x}_s$. What is the negative log-likelihood for $\mathbf{y} - \mathbf{H}(\boldsymbol{\theta})\mathbf{x}_s$? Is it the same as in 3.3.c?

   Study also the special case where $\mathbf{R}_x$ does not depend on $\boldsymbol{\theta}$ and where $\mathbf{H}(\boldsymbol{\theta}) = \boldsymbol{\theta} \in \mathbb{R}$. Which $\boldsymbol{\theta}$ maximizes the likelihood of $\mathbf{y} - \mathbf{H}(\boldsymbol{\theta})\mathbf{x}_s$? Does it seem to be a useful estimator?

k. Also, in view of (3.36), it may seem reasonable to construct an estimator such that not only the difference between $y(t)$ and $\mathbf{C}(\boldsymbol{\theta})\mathbf{x}_s(t)$ is small but also the one between $\mathbf{x}_s(t+1)$ and $\mathbf{F}(\boldsymbol{\theta})\mathbf{x}_s(t)$. Argue that this corresponds to basing the estimator on

$$\begin{bmatrix} \mathbf{y} - \mathbf{H}(\boldsymbol{\theta})\mathbf{x}_s \\ \mathbf{R}_x^{-1/2}(\boldsymbol{\theta})\mathbf{x}_s \end{bmatrix}$$

What is the negative log-likelihood for this quantity. Consider the same special case as in 3.3.j. Does the estimator maximizing the likelihood function appear reasonable?

3.4. Suppose that the model of interest is the nonlinear in the parameters model

$$\mathbf{z}^N = g(\boldsymbol{\theta}) + \mathbf{v}^N$$

where $g : \mathbb{R}^n \to \mathbb{R}^N$ is a known nonlinear function, where $\boldsymbol{\theta} \in \mathbb{R}^n$ and where $\mathbf{v}^N$ is assumed to be a random vector with i.i.d. distributed elements being $\mathcal{N}(0,1)$.

Let a simplified (linear in the parameters) model be

$$\mathbf{z}^N = \mathbf{T}\tilde{\boldsymbol{\theta}} + \mathbf{v}^N$$

where $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^m$.

a. Determine the ML-estimator of $\tilde{\boldsymbol{\theta}}$.

b. Formulate an indirect inference estimator of $\boldsymbol{\theta}$ based on the estimator in a. and the cost function (3.7).

[48] meaning that it has all ones on the diagonal.

c. Suppose that the weighting matrix is taken as $W = \mathbf{T}^T\mathbf{T}$. Interpret the effect this choice has on the estimator.

3.5. Suppose that (3.9) holds. Show that then $\hat{\mathbf{y}}(t|t-1;\boldsymbol{\theta})$ minimizes the MSE

$$\mathbb{E}\left[|\mathbf{y}(t) - f(\mathbf{y}^{t-1})|^2\right]$$

among all functions of $\mathbf{y}^{t-1}$.

3.6. Suppose that we have two samples of a Moving Average (MA) process:

$$\begin{aligned} y(1) &= w(1) + w(0) \\ y(2) &= w(2) + w(0) \end{aligned}$$

a. Suppose that $\{w(t)\}$ are i.i.d. $\mathcal{N}(0,1)$. Determine the posterior mean of $y(2)$ given $y(1)$, $\mathbb{E}\left[y(2)|y(1)\right]$. Show that the error $y(2) - \mathbb{E}\left[y(2)|y(1)\right]$ is independent of $y(1)$.

b. Suppose that $\{w(t)\}$ are i.i.d. random variables taking the values ± with equal probability. Compute the posterior mean $\mathbb{E}\left[y(2)|y(1)\right]$. Is the error $y(2) - \mathbb{E}\left[y(2)|y(1)\right]$ independent of $y(1)$?

3.7. Consider the model

$$M(\boldsymbol{\xi}) = f(\boldsymbol{\theta}) + \mathbf{v}, \quad \boldsymbol{\xi} = \begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{v} \end{bmatrix}$$

and suppose that the data has been generated by

$$\mathbf{z} = f(\boldsymbol{\theta}_o) + \mathbf{v}_o,$$

a. Suppose that $\hat{\boldsymbol{\theta}}$ is obtained by solving the minimization problem

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} J(\mathbf{z} - f(\boldsymbol{\theta}))$$

Suppose that the criterion J is such that the solution $\hat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}_o$. Show that then

$$\boldsymbol{\theta}_o - \hat{\boldsymbol{\theta}} \approx -\left(f'(\boldsymbol{\theta}_o)^T J''(\mathbf{v}_o) f'(\boldsymbol{\theta}_o)\right)^{-1} f'(\boldsymbol{\theta}_o)^T J'(\mathbf{v}_o)$$

b. Use 3.7.a to show that

$$\begin{aligned} &J''(\mathbf{v}_o)^{1/2}\left(f(\boldsymbol{\theta}_o) - f(\hat{\boldsymbol{\theta}})\right) \\ &\approx -J''(\mathbf{v}_o)^{1/2} f'(\boldsymbol{\theta}_o)\left(f'(\boldsymbol{\theta}_o)^T J''(\mathbf{v}_o) f'(\boldsymbol{\theta}_o)\right)^{-1}\left(J''(\mathbf{v}_o)^{1/2} f'(\boldsymbol{\theta}_o)\right)^T J''(\mathbf{v}_o)^{-1/2} J'(\mathbf{v}_o) \end{aligned}$$

Interpret this result in terms of an orthogonal projection.

c. Use 3.7.b to show that[49]

$$|f(\boldsymbol{\theta}_o) - f(\hat{\boldsymbol{\theta}})|_{J''(\mathbf{v}_o)} \approx |\mathcal{S} J''(\mathbf{v}_o)^{-1/2} J'(\mathbf{v}_o)|^2$$

where $\mathcal{S}$ is the n_θ-dimensional subspace spanned by the columns of $J''(\mathbf{v}_o)^{1/2} f'(\hat{\boldsymbol{\theta}})$, and where $\mathcal{S}\mathbf{v}_o$ denotes the orthogonal projection of $\mathbf{v}_o$ on the subspace $\mathcal{S}$.

[49] $|\mathbf{x}|^2_{\mathbf{Q}} = \mathbf{x}^T\mathbf{Q}\mathbf{x}$.

d. Consider now that $\boldsymbol{\theta} = \begin{bmatrix} \tilde{\boldsymbol{\theta}}^T & 0 \end{bmatrix}^T$. Compare the model fit, using

$$|f(\boldsymbol{\theta}_o) - f(\hat{\boldsymbol{\theta}})|_{J''(\mathbf{v}_o)}$$

as criterion, for the two cases when i) the entire vector $\boldsymbol{\theta}$ is estimated and, ii) only estimating $\tilde{\boldsymbol{\theta}}$ (assuming it known that the last element is zero). Conclude that overfitting occurs under general conditions when the model parameters are estimated by solving an optimization problem.

3.8. Let $f : \mathbb{R}^m \to \mathbb{R}^n$ and $A : \mathbb{R}^m \to \mathbb{R}^{n\times n}$. Show that

$$\left| \int A(x) f(x) dx \right|^2 \leq \int \|A(x)\|_F^2 dx \int |f(x)|^2 dx$$

3.9. Show that a positive definite function $K(s,t)$, i.e. a function satisfying (3.15), has to be symmetric: $K(t,s) = K^T(s,t)$.

3.10. Show that a spectrum Φ is symmetric, i.e. $\Phi(-\omega) = \Phi^T(\omega)$. Conclude that $\Phi^*(\omega) = \Phi(\omega)$ i.e. that $\Phi(\omega)$ is Hermitian.

3.11. Let $\{X(n)\}_{n=1}^{\infty}$ be a sequence of independent random variables which have distributions with unbounded support so that

$$\mathbf{P}(|X(n)| \geq c) = p(c) > 0, \quad \forall c > 0$$

Use Corollary D.2.1 to show that $\{X(n)\}_{n=1}^{\infty}$ is unbounded with probability 1.

3.12. Use the Cauchy criterion to show that

$$y(t) = G(q)u(t) = \sum_{k=-\infty}^{\infty} g(k) u(t-k)$$

exists as a limit in $\mathcal{L}_2$ when $\{e(t)\}$ is white noise.

3.A Proof of the scalar spectral factorization theorem - Theorem 3.4.5

The function Φ being rational in $e^{i\omega}$ means that

$$\Phi(\omega) = c\, e^{i\omega\tau}\, \frac{\prod_{k=1}^{n_z}(e^{i\omega} - z_k)}{\prod_{k=1}^{n_p}(e^{i\omega} - p_k)}$$

where there are no cancellations, where $z_k \neq 0$, $k = 1, \dots, n_z$ and $p_k \neq 0$, $k = 1, \dots, n_p$, and where the factor $e^{\omega\tau}$ corresponds to zeros of the numerator or denominator at the origin[50] but Φ being positive means that for all ω it has to hold

$$\begin{aligned} 1 = \frac{\Phi(\omega)}{\overline{\Phi(\omega)}} &= e^{i\omega 2\tau}\, \frac{\prod_{k=1}^{n_z}(e^{i\omega} - z_k)}{\prod_{k=1}^{n_z}(e^{-i\omega} - z_k)}\, \frac{\prod_{k=1}^{n_p}(e^{-i\omega} - p_k)}{\prod_{k=1}^{n_p}(e^{i\omega} - p_k)} \\ &= (-1)^{n_p - n_z} e^{i\omega(2\tau + n_z - n_p)}\, \frac{\prod_{k=1}^{n_p} p_k}{\prod_{k=1}^{n_z} z_k}\, \frac{\prod_{k=1}^{n_c}(e^{i\omega} - z_k)}{\prod_{k=1}^{n_c}(e^{i\omega} - z_k^{-1})}\, \frac{\prod_{k=1}^{n_p}(e^{i\omega} - p_k^{-1})}{\prod_{k=1}^{n_p}(e^{i\omega} - p_k)} \end{aligned}$$

For this function to be constant all factors need to cancel out. Firstly, $2\tau = n_p - n_z$. Secondly, since the numerator zeros are distinct from the denominator zeros the canellations have to take place within the rational functions

$$\frac{\prod_{k=1}^{n_z}(e^{i\omega} - z_k)}{\prod_{k=1}^{n_z}(e^{i\omega} - z_k^{-1})} \quad \text{and} \quad \frac{\prod_{k=1}^{n_p}(e^{i\omega} - p_k^{-1})}{\prod_{k=1}^{n_p}(e^{i\omega} - p_k)}$$

This means that for each numerator zero z_i there has to be another numerator zero z_k such that $z_k^{-1} = z_i$, and similarly for the denominator zeros. Thus, the denominator degree is even as the poles have to appear pairwise. Assuming the zeros are sorted so those inside the unit circle are p_k, $k = 1, \dots, n_p/2$, gives the denominator

$$\prod_{k=1}^{n_p/2}(e^{i\omega} - p_k)(e^{i\omega} - p_k^{-1}) = (-1)^{n_p/2} e^{i\omega n_p/2} \prod_{k=1}^{n_p/2} p_k^{-1} D(e^{i\omega})\overline{D(e^{i\omega})}$$

where $D(z) = \prod_{k=1}^{n_p/2}(z - p_k)$. We could do the same for the numerator, was it not for that the argument above is not valid for zeros at 1 as such factors cancel without a sibling. However, if we exclude all zeros at 1, say that the last m zeros $z_{n_z-m+1}, \dots, z_{n_z}$ are located at 1, the same argument as for the denominator gives that the numerator can be written

$$(-1)^{(n_z-m)/2} e^{i\omega(n_z-m)/2} \prod_{k=1}^{(n_z-m)/2} z_k^{-1}\;\; F(e^{i\omega})\overline{F(e^{i\omega})}\; (e^{i\omega} - 1)^m$$

where $F(z) = \prod_{k=1}^{(n_z-m)/2}(z - z_k)$, contains the zeros inside the unit circle. This gives that

$$\begin{aligned} \Phi(\omega) &= e^{i\omega(\tau - n_z + n_p - m)/2} \frac{|F(e^{i\omega})|^2}{|D(e^{i\omega})|^2}\, (e^{i\omega} - 1)^m \\ &= \tilde{c} \frac{|F(e^{i\omega})|^2}{|D(e^{i\omega})|^2}\, e^{-i\omega m/2}(e^{i\omega} - 1)^m \end{aligned}$$

[50] $\tau > 0$ corresponds to τ zeros of the numerator at the origin, whereas $\tau < 0$, corresponds to $-\tau$ zeros of the denominator at the origin.

where $\tilde{c} = c(-1)^{(n_z-m-n_p)/2}\prod_{k=1}^{(n_z-m)/2} z_k^{-1}/\prod_{k=1}^{n_p/2} p_k^{-1}$. For this to be real-valued we need $e^{-i\omega m/2}(e^{i\omega}-1)^m$ to be real valued, but

$$e^{-i\omega m/2}(e^{i\omega}-1)^m = (e^{i\omega/2}-e^{-i\omega/2})^m = i^m \sin^m\left(\frac{\omega m}{2}\right)$$

and hence m needs to be even. We can thus define $C(z) = F(z)(z-1)^{m/2}$ and conclude that Φ must be of the form

$$\Phi(\omega) = \tilde{c}\,\frac{|C(e^{i\omega})|^2}{|D(e^{i\omega})|^2}$$

where $C(z)$ and $D(z)$ are polynomials with their zeros in the unit disc and inside the unit circle, respectively.

3.B Proof of Theorem 3.4.6

We have $|z_N(t)| \leq \eta(N)$ where

$$\eta(N) = \sum_{\tau=0}^{N} |g(\tau)||v(t-\tau)|$$

Hence, if we can prove that $\{\eta(N)\}$ converges with probability one, the theorem follows from Stone-Weierstrass' theorem, see for example Theorem 7.10 p. 148[51]. Let $\varepsilon > 0$ be arbitrary and $M > N$. Then

$$\mathbf{P}(|\eta(M)-\eta(N)| \geq \varepsilon) \leq \frac{1}{\varepsilon}\sum_{\tau=N+1}^{M} |g(\tau)|\mathrm{E}\left[|v(t-\tau)|\right] \leq \frac{C}{\varepsilon}\sum_{\tau=N+1}^{\infty} |\mathbf{g}(\tau)| \tag{3.37}$$

where we first have used Chebyshev's inequality and in the second inequality (3.33). Hence

$$\sum_{N=1}^{\infty}\sup_{M>N}\mathbf{P}(|\eta(M)-\eta(N)| \geq \varepsilon) < \frac{C}{\varepsilon}\sum_{N=1}^{\infty}\sum_{\tau=N+1}^{\infty}|g(\tau)| = \frac{C}{\varepsilon}\sum_{k=2}^{\infty}(k-1)|g(k)| < \infty$$

and the Cauchy version of Borel-Cantelli's lemma, Corollary D.2.1, implies that $\{\eta(N)\}$ converges almost surely.

For Corollary 3.4.4, Chebyshev's inequality, Lemma D.2.2, with $\varphi(x) = x^2$ and using that the input is white gives

$$\mathbf{P}(|\eta(M)-\eta(N)| \geq \varepsilon) \leq \frac{1}{\varepsilon^2}C\sum_{\tau=N+1}^{\infty}|g(\tau)|^2 \qquad \forall M > N$$

and the proposition in Theorem 3.4.6 is true if

$$\sum_{N=1}^{\infty}\sum_{\tau=N+1}^{\infty}|g(\tau)|^2 = \sum_{k=2}^{\infty}(k-1)|g(k)|^2 < \infty.$$

[51] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, London, 1976

4
Estimation Theory

This chapter treats more formally the problem of estimating model-parameters from observations. Recall that in the formalism of Chapter 3 model parameters ξ^N are random variables, and hence, under our measurement model (3.1), the observation $\mathbf{z}^N$ is the realization of a random variable, which we denote $\mathbf{Z}$. We will thus discuss the inference problem of estimating one random vector given an observation of another random vector.

We can quantify the properties of an estimator in terms of its pdf. Generally, a good estimator has a distribution peaked around the quantity of interest. A quality measure for an estimator is called a risk function, and a commonly used risk function is the *Mean-Square Error* (MSE). With ξ denoting the model parameter and $\hat{\xi}(\mathbf{Z})$ its estimator, the MSE is defined as

$$\mathrm{MSE}\left[\hat{\xi}(\mathbf{Z})\right] := \mathbb{E}\left[(\hat{\xi}(\mathbf{Z}) - \xi)(\hat{\xi}(\mathbf{Z}) - \xi)^T\right]$$

which is small for a good estimator. Notice that the MSE is a matrix when $n > 1$ in which case the trace of the quantity above is what one would normally call the MSE.

4.1 Information contents in random signals

We start this chapter with some general considerations regarding information in observations of random variables.

4.1.1 Information contents in events

Consider two events A and B. When the sample space is restricted to A, i.e. whenever an outome occurs that is not in A it is discarded, the probability of the event B changes. The probability that both A and B occurs is $\mathbf{P}(A \cap B)$ and the probability that we will have A is $\mathbf{P}(A)$ which gives then that the probability of B occuring when A has occured is

$$\mathbf{P}(B|A) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)} \tag{4.1}$$

provided $\mathbf{P}(A) > 0$.

When the probability of B does not change if we are given that A has occured, i.e.

$$\mathbf{P}(B|A) = \mathbf{P}(B),$$

we say that the events are *independent*.

We can express the conditional $\mathbf{P}(B|A)$ using $\mathbf{P}(A|B)$

$$\mathbf{P}(B|A) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)} = \frac{\mathbf{P}(A|B)\mathbf{P}(B)}{\mathbf{P}(A)}$$

This is Bayes rule from which we see that if A and B are independent (according to the definition above)

$$\mathbf{P}(A|B) = \mathbf{P}(A)$$

From (4.1) we have see that see that when $\mathbf{P}(A) > 0$, independence is equivalent to

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B|A) = \mathbf{P}(A)\mathbf{P}(B)$$

This relation relation is typically taken as the definition of independence of events.

4.1.2 Information Contents in Random Variables

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be the probability space in which our probabilistic model is defined. We will be interested in the information contained in an observation of a random variable Y, i.e. what information does the outcome $Y(\omega)$ carry? It is obvious that we can determine if the event (3.11) has occured or not and in fact we can determine if events of the type $\{\omega : Y(\omega) \in B\}$, where B is a Borel set, has occured or not. It turns out that the events we can determine if they have occured or not are exactly those belonging to the σ-algebra generated by Y,

$$\sigma(Y) := \sigma\left(\{\omega : Y(\omega) \in B,\ B \in \mathcal{B}\}\right),$$

The σ-algebra $\sigma(Y)$ thus represents the events whose occurence we can determine by observing Y. $\sigma(Y)$ is a sub σ-algebra to $\mathcal{F}$, i.e. if $A \in \sigma(Y)$, then $A \in \mathcal{F}$. This means that there are typically events in $\mathcal{F}$ which we cannot determine if they have occured or not by observing Y. For example, suppose that there is another real valued random variable X. From the knowledge of Y we then only get partial information about X, i.e. we typically cannot determine the occurence of all events in $\sigma(X)$. So what can be said in such a case? Consider that we only know that $Y \in B_Y$ and that $\mathbf{P}(Y \in_Y) > 0$. What can we then say about the probability that $X \in B_X$? Well, from Section 4.1.1 we have that the probability changes to

$$\mathbf{P}_{X|Y}(B_X|B_Y) := \frac{\mathbf{P}(X \in B_X \cap Y \in B_Y)}{\mathbf{P}(Y \in B_Y)} \tag{4.2}$$

For each fix B_Y such that the denominator is positive, this is a probability measure over $(\Omega, \sigma(X))$. In the context of estimation we will call this the conditional distribution function of X.

For example

$$\mathbf{P}_{X|Y}(B_X|y) = \frac{\mathbf{P}(X \in B_X \cap Y = y)}{\mathbf{P}(Y = y)} \tag{4.3}$$

would give the probability of different events related for X when $Y = y$ is observed.

Example 4.1. *Suppose that X is Bernouilli distributed with equal probabilities for taking the values 0 and 1 and suppose that Y takes the values $-1, 0, 2$, and that the joint distribution of X, Y is given by*

x	y	$\boldsymbol{P}(X = x, Y = y)$
-1	-1	1/8
1	-1	2/8
-1	0	3/8
1	2	2/8

This gives

x	y	$\boldsymbol{P}_{X\|Y}(X = x\|Y = y)$
-1	-1	$\frac{\boldsymbol{P}(X=-1,Y=-1)}{\boldsymbol{P}(Y=-1)} = \frac{\frac{1}{8}}{\frac{1}{8}+\frac{2}{8}} = \frac{1}{3}$
1	-1	$\frac{\boldsymbol{P}(X=1,Y=-1)}{\boldsymbol{P}(Y=-1)} = \frac{\frac{2}{8}}{\frac{1}{8}+\frac{2}{8}} = \frac{2}{3}$
-1	0	$\frac{\boldsymbol{P}(X=-1,Y=0)}{\boldsymbol{P}(Y=-1)} = \frac{\frac{3}{8}}{\frac{3}{8}} = 1$
1	2	$\frac{\boldsymbol{P}(X=1,Y=2)}{\boldsymbol{P}(Y=-1)} = \frac{\frac{2}{8}}{\frac{2}{8}} = 1$

■

However, there is a technical problem with (4.3), namely that it is not well defined when $\mathbf{P}(Y = y) = 0$. For full generality we therefore need a more abstract definition of the conditional probability distribution function. What we would like is that $\mathbf{P}_{X|Y}(B_X|y)$ is a function such that when we integrate it with respect to y over a region B_Y, taking the probability measure for Y into account, we obtain $\mathbf{P}(X \in B_X \cap Y \in B_Y)$

$$\int_{Y \in B_Y} \mathbf{P}_{X|Y}(B_X|Y(\omega))\mathbf{P}(d\omega) = \mathbf{P}(X \in B_X \cap Y \in B_Y) \tag{4.4}$$

As noted above, in addition to (4.4) we also need that $\mathbf{P}_{X|Y}(B_X|B_Y)$ is a probability measure on $\sigma(X)$ for any fix B_Y. Fortunately it can be shown that there exists a function satisfying these requirements, however it is beyond this exposition to show this.

The expectation of the conditional distribution for X given $Y = y$ is given by

$$\int_\Omega X(\omega)\mathbf{P}_{X|Y}(X(d\omega)|y)$$

cf. (3.13). We can also define the conditional expectation of X given Y as

$$\mathbb{E}\left[X|Y\right](\omega) = \int_\Omega X(\bar{\omega})\mathbf{P}_{X|Y}(X(d\bar{\omega})|Y(\omega)) \tag{4.5}$$

This is a random variable which on the event $Y = y$ has as outcome the mean of the conditional distribution for this observation of Y.

This random variable has the same expectation as X

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[X|Y\right]\right] &= \int_\Omega \mathbb{E}\left[X|Y\right](\omega)\mathbf{P}(d\omega) = \int_\Omega\int_\Omega X(\bar{\omega})\mathbf{P}_{X|Y}(X(d\bar{\omega})|Y(\omega))\mathbf{P}(d\omega)\\ &= \int_\Omega X(\bar{\omega})\int_\Omega \mathbf{P}_{X|Y}(X(d\bar{\omega})|Y(\omega))\mathbf{P}(d\omega)\\ &= \int_\Omega X(\bar{\omega})\mathbf{P}\left(\{\omega:\ X(\omega)\in X(d\bar{\omega})\}\cap(Y\in\Omega)\right)\\ &= \int_\Omega X(\bar{\omega})\mathbf{P}(d\bar{\omega}) = \mathbb{E}\left[X\right]\end{aligned}$$

Example 4.2 (Example 4.1 continued). *The conditional mean is given by*

$$\mathbb{E}\left[X|Y=-1\right] = -1\times \boldsymbol{P}_{X|Y}(X=-1|Y=1) + 1\times \boldsymbol{P}_{X|Y}(X=1|Y=1) = -1\times\frac{1}{3} + 1\times\frac{2}{3} = \frac{1}{3}$$

$$\mathbb{E}\left[X|Y=0\right] = -1\times \boldsymbol{P}_{X|Y}(X=-1|Y=0) + 1\times \boldsymbol{P}_{X|Y}(X=1|Y=0) = -1\times 1 + 0 = -1$$

$$\mathbb{E}\left[X|Y=2\right] = -1\times \boldsymbol{P}_{X|Y}(X=-1|Y=2) + 1\times \boldsymbol{P}_{X|Y}(X=1|Y=2) = 0 + 1\times 1 = 1$$

Example 4.3. *In the case where the probability distributions of X and Y can be represented by the joint probability distribution function (pdf) $p(x,y)$, the conditional distribution function of X given Y can be expressed in terms of the* conditional pdf *(cpdf)*

$$p_{X|Y}(x|y) = \begin{cases}\frac{p_{X,Y}(x,y)}{p_Y(y)} & p(y) > 0\\ 0 & p_Y(y) = 0\end{cases}$$

where $p_Y(y)$ is the marginal pdf

$$p_Y(y) = \int p_Y(x,y)dx$$

The conditional expectation can the be expressed as

$$\mathbb{E}\left[X|Y\right] = \int_{-\infty}^{\infty} x p_{X|Y}(x|Y)dx$$

■

4.1.3 Independent Random Variables

So when does one random variable Y not contain any information about another random variable X? It is when the probability of any event in $\sigma(X)$ does not change when we observe an event in $\sigma(Y)$, i.e. when

$$\mathbf{P}_{X|Y}(B_X|B_Y) = \frac{\mathbf{P}(X\in B_X \cap Y\in B_Y)}{\mathbf{P}(Y\in B_Y)} = \mathbf{P}(X\in B_X) \tag{4.6}$$

i.e. when

$$\mathbf{P}(X\in B_X\cap Y\in B_Y) = \mathbf{P}(X\in B_X)\mathbf{P}(Y\in B_Y) \quad \forall B_X\in\sigma(X),\ B_Y\in\sigma(Y) \tag{4.7}$$

Since the sub-σ-algebras $\sigma(X)$ and $\sigma(Y)$ are generated by sets of the type $\{\omega:\ X(\omega) < c\}$ it is sufficient to prove that (4.6) holds for all sets of this type. In this case the conditional probability is given by

$$\mathbf{P}_{X|Y}(B_X|y) = \mathbf{P}(X\in B_X)$$

We conclude that when (4.7) holds, Y does not carry any information about X. We then say that X and Y are independent. When X, Y have a joint pdf $pX, Y(x,y)$, X and Y are independent if and only if

$$p_{X,Y}(x,y) = p_X(x)p_Y(y), \quad \forall x,y$$

A consequence of independence is that

$$\mathbb{E}\left[X|Y\right](\omega) = \int_\Omega X(\bar{\omega})d\mathbf{P}_{X|Y}(\bar{\omega}|Y(\omega)) = \int_\Omega X(\bar{\omega})d\mathbf{P}(\bar{\omega}) = \mathbb{E}\left[X\right] \quad \text{a.e.}$$

4.2 *Estimation of random variables*

4.2.1 *Minimizing the Mean-Square Error*

Let X be a random vector defined on a probability space $(\Omega,,\mathbf{P})$. Suppose that we would like to estimate X, i.e. provide a guess $\hat{x}$. One possible quality measure is the mean-square error

$$\text{MSE}\left[\hat{x}\right] = \mathbb{E}\left[|X-\hat{x}|^2\right]$$

The MSE can be split in two terms

$$\text{MSE}\left[\hat{x}\right] = \mathbb{E}\left[|X-\mathbb{E}\left[X\right]|^2\right] + |\mathbb{E}\left[X\right]-\hat{x}|^2 \tag{4.8}$$

The first term is called the variance error (it is simply the variance of X) and the second term the bias (systematic) error. We see that here the variance error is something we cannot influence while we can eliminate the bias error by taking $\hat{x} = \hat{x}^* := \mathbb{E}\left[X\right]$. This gives the minimimum mean-square error (MMSE)

$$\text{MMSE} := \text{MSE}\left[\hat{x}^*\right] = \mathbb{E}\left[|X-\mathbb{E}\left[X\right]|^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}^2[X] \le \text{MSE}\left[\hat{x}\right], \quad \forall \hat{x}$$

Suppose now that we are given the information that another random vector Y, defined on the same probability space, has taken the value y. Can we then improve our estimate of X? Well, now the probability distribution for X has changed to the conditional distribution $P_{X|Y}(X \in B_X|Y=y)$ so the MSE is now given by

$$\text{MSE}\left[\hat{x}|Y=y\right] = \mathbb{E}\left[|X-\hat{x}|^2|Y=y\right]$$

However, we can expand this expression into a variance term and a bias term just as in (4.8)

$$\text{MSE}\left[\hat{x}|Y=y\right] = \mathbb{E}\left[|X-\mathbb{E}\left[X|Y=y\right]|^2|Y=y\right] + |\mathbb{E}\left[X|Y=y\right]-\hat{x}|^2 \tag{4.9}$$

which again is minimized by taking $\hat{x}$ to be the mean of X, but now the conditional mean $\mathbb{E}\left[X|Y=y\right]$. This means that the optimal $\hat{x}$ now will depend on the observation y: $\hat{x} = \hat{x}^*(y) = \mathbb{E}\left[X|Y=y\right]$. The minimum MSE is now a function of y and given by

$$\begin{aligned}\text{MMSE}(y) &= \text{MSE}\left[\hat{x}^*(y)|Y=y\right] = \mathbb{E}\left[|X-\mathbb{E}\left[X|Y=y\right]|^2|Y=y\right] \\ &= \mathbb{E}\left[X^2|Y=y\right] - \mathbb{E}^2[X|Y=y]\end{aligned} \tag{4.10}$$

We would expect that our observation should improve the quality of the estimate, i.e. that

$$\text{MMSE}(y) < \text{MMSE}(\mathbb{E}[X])$$

but is this true, and, if so, how much better is $\hat{x}(y)$ than $\hat{x}$? To examine this we re-write the MSE as

$$\begin{aligned}\text{MSE}[\hat{x}] = \mathbb{E}\left[|X-\hat{x}|^2\right] = \mathbb{E}_Y\left[\mathbb{E}\left[|X-\hat{x}|^2|Y\right]\right] = \int \mathbb{E}\left[|X-\hat{x}|^2|Y=y\right]p_Y(y)dy \\ = \int \text{MSE}[\hat{x}|Y=y]\,p_Y(y)dy \qquad (4.11)\end{aligned}$$

Thus, while $\hat{x}^* := \mathbb{E}[X]$ minimizes this integral expression, $\hat{x}^*(y) := \mathbb{E}[X|Y=y]$ minimizes the integrand at the point y so clearly

$$\text{MSE}[\mathbb{E}[X|Y=y]\,|Y=y] \leq \text{MSE}[\mathbb{E}[X]\,|Y=y]$$

Furthermore, we observe that when X and Y are independent we have equality since conditioning on Y then does not change the distribution of X.

We now pose the question how to estimate X from an arbitrary observation of Y. We will then allow $\hat{x}$ to be a function of Y, $\hat{x}=\hat{x}(Y)$. We notice that the decomposition (4.11) is still valid in this case and that taking $\hat{x}(Y)=\hat{x}^*(Y)=\mathbb{E}[X|Y]$ will for each $Y=y$ minimize the integrand, and hence minimize the MSE. Taking the expectation of (4.10) gives

$$\text{MMSE} = \text{MSE}[\mathbb{E}[X|Y]] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[\mathbb{E}^2[X|Y]\right] \qquad (4.12)$$

which is the minimum MSE that can be achieved if $\hat{x}$ is allowed to be a function of Y, and which is achieved by taking $\hat{x}(Y)=\mathbb{E}[X|Y]$.

4.2.2 *The Internal Structure of the Conditional Expectation**

Above, we have seen that the conditional expection of X given the observed variable Y gives the MMSE estimator. In this section we will try to understand a little bit better why this is the case.

In general terms estimating a random variable X from another random variable Y means how well we can mimick the behaviour of X using Y. Now X is defined by its probability distribution over arbitrary sets in $\mathcal{F}$. In fact suppose that there is another random variable Z such that

$$\int_A Z(\omega)\mathbf{P}(d\omega) = \int_A X(\omega)\mathbf{P}(d\omega) \quad \forall A \in \sigma(X) \qquad (4.13)$$

Then it follows that $Z=X$ except possibly on a set of measure zero. To see this suppose that $Z \neq X$ on a set $A \in \sigma(X)$ for which $P(A) > 0$. Then we can take the subset on A either for which $Z > X$ or $Z < X$, whichever has non-zero measure. This subset also belong to $\sigma(X)$ and clearly the two integrals above cannot be the same then. Hence we can take the set of pairs $\{\{A \in \sigma(A), \int_A X(\omega)\mathbf{P}(d\omega)\}\}$ as definition of a random variable.

So a natural question is if we can use our observed variable Y to construct a new random variable $Z = Z(Y)$ which has a probability distribution as close as possible to the above. Since Z is a function of Y we can only try to match (4.13) for events $A \in \sigma(Y)$, i.e. we can try to find Z such that

$$\int_A Z((Y(\omega))\mathbf{P}(d\omega) = \int_A X(\omega)\mathbf{P}(d\omega) \quad \forall A \in \sigma(Y) \tag{4.14}$$

It is far from obvious that there is such a function and that, if so, it is a measurable function so that Z is a random variable so let us study a couple of examples.

Example 4.4 (Example 4.2 continued). *In this example, the sets of $\sigma(Y)$ are $Y = -1$, $Y = 0$, $Y = 2$, and unions of these. For example*

$$\int_{Y=-1} X(\omega)\boldsymbol{P}(d\omega) = -1 \times \boldsymbol{P}(X = -1, Y = -1) + 1 \times \boldsymbol{P}(X = 1, Y = -1) = -1 \times \frac{1}{8} + 1 \times \frac{2}{8} = \frac{1}{8}$$

whereas

$$\int_{Y=-1} \mathbb{E}\left[X|Y\right](\omega)\boldsymbol{P}(d\omega) = \mathbb{E}\left[X|Y = -1\right]\boldsymbol{P}(Y = -1) = \frac{1}{3} \times \frac{3}{8} = \frac{1}{8}$$

showing that $\mathbb{E}[X|Y](\omega)$ has the same mean as X over the event $Y = -1$. The same can be shown for the other sets of $\sigma(Y)$.

Example 4.5. *Suppose that $X(\omega) = 1$ for $\omega \in A_X$ and zero otherwise. Similarly $Y(\omega) = 1$ for $\omega \in A_Y$ and zero otherwise. Let $B_X = \{X(\omega) : \omega \in A_X\}$ and $B_Y = \{Y(\omega) : \omega \in A_Y\}$.*

Then $\sigma(Y) = \{0, A_Y, A_Y^c, \Omega\}$ and Z has to be constant on A_Y (and A_Y^c). Let us call this value z. In order to satisfy (4.14) we need

$$\int_{A_Y} Zd\boldsymbol{P} = \int_{A_Y} Xd\boldsymbol{P} = \boldsymbol{P}(A_X \cap A_Y)$$

but since Z is constant, $Z = z$, on A_Y

$$\int_{A_Y} Zd\boldsymbol{P} = z\boldsymbol{P}(A_Y)$$

we have

$$z = \frac{\boldsymbol{P}(A_X \cap A_Y)}{\boldsymbol{P}(A_Y)} = \frac{\boldsymbol{P}(X \in B_X \cap Y \in B_Y)}{\boldsymbol{P}Y \in B_Y)} = \boldsymbol{P}_{X|Y}(B_X|B_Y)$$

Repeating these calculations for the event B_Y^c gives that Z should take the value $\boldsymbol{P}(B_X|B_Y^c)$ on this set. We thus have

$$Z = \begin{cases} \boldsymbol{P}_{X|Y}(B_X|B_Y) & Y \in B_Y \\ \boldsymbol{P}_{X|Y}(B_X|B_Y^c) & Y \in B_Y^c \end{cases}$$

Consider now the conditional expectation (4.5)

$$\mathbb{E}\left[X|Y\right](\omega) = \int_\Omega X(\bar{\omega})\boldsymbol{P}_{X|Y}(X(d\bar{\omega})|Y(\omega))$$

For $Y(\omega) \in B_Y$ this evaluates to $\boldsymbol{P}_{X|Y}(B_X|B_Y)$ and for $Y(\omega) \in B_Y^c$ to $\boldsymbol{P}_{X|Y}(B_X|B_Y^c)$. Thus Z coincides with the conditional expectation $\mathbb{E}[X|Y]$.

■

The preceeding two examples suggests that the conditional expectation $\mathbb{E}[X|Y]$ satisfies (4.14) and this is indeed true. We will not prove this in full generality but restrain ourselves to the simple case that $X = \mathbf{1}_{B_X}$, generalizing Example 4.5. Then

$$\int_{Y\in B_Y} X(\omega)\mathbf{P}(d\omega) = \mathbf{P}(X \in B_X \cap Y \in B_Y)$$

while

$$\int_{Y\in B_Y} \mathbb{E}[X|Y](\omega)\mathbf{P}(d\omega) = \int_{Y\in B_Y}\int_{\Omega} X(\bar{\omega})\mathbf{P}_{X|Y}(d\bar{\omega}|Y(\omega))\mathbf{P}(d\omega)$$
$$= \int_{Y\in B_Y} \mathbf{P}_{X|Y}(B_X|Y(\omega))\mathbf{P}(d\omega) = \mathbf{P}(X \in B_X \cap Y \in B_Y)$$

The general result is obtained by a constructive procedure where X is built up as a limit of indicator functions.

The condition (4.14) can be taken as a definition of the conditional expectation. From this we realize that the conditional expectation is not unique: If Z satisfies (4.14) we can change it on a subset of Ω of measure zero, i.e. for which the probability measure $\mathbf{P}$ is zero, and still maintain (4.14). We say that the conditional expectation is uniquely defined a.e. (almost everywhere).

Suppose now that $X = f(Y)$ where f is a measurable function. Then we see immediately that $Z = X = f(Y)$ satisfies (4.14), i.e.

$$\mathbb{E}[f(Y)|Y] = f(Y) \quad a.e.$$

More generally we see that when $X = VW$ where $V \in \sigma(Y)$ and W is another random variable,

$$\mathbb{E}[X|Y] = \mathbb{E}[VW|Y] = V\mathbb{E}[W|Y] \quad a.e., \quad \forall V \in \sigma(Y) \tag{4.15}$$

4.2.3 A Hilbert space interpretation

By embedding the estimation problem in a Hilbert space setting we will be able to handle a wide range of estimation problems. Therefore, let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and let $\mathcal{L}_2$ be the Hilbert space defined in Definition 3.4.6 with $m = 1$ (for a start).

Consider now that we would like to use a random variable Y to estimate another random variable X in MMSE sense. We notice that

$$\text{MSE}[g(Y)] = \|X - g(Y)\|_2^2$$

To approach the problem of minimizing this quantity we form the subspace $\mathcal{S}(Y)$ consisting of all elements Z in $\mathcal{L}_2$ for which the σ-algebra generated by the corresponding random variable, let us call it $\sigma(Z)$, is a subset of $\sigma(Y)$. This essentially means the subspace of all measurable functions of Y which have bounded second moment.

From the Hilbert space theory (Appendix C) we then know that there is a unique element in $\mathcal{S}(Y)$ solving

$$\min_{Z\in\mathcal{S}(Y)} \|X - Z\|_2 \tag{4.16}$$

and that this element is uniquely determined by the orthogonality conditition

$$\langle X - Z, W \rangle = 0 \quad \forall W \in \mathcal{S}(Y) \tag{4.17}$$

in other words Z should be the orthogonal projection of X on $\mathcal{S}(Y)$. We observe that $\mathcal{S}(Y)$ is infinite dimensional so it does not seem trivial to find the projection. However, spurred by the results in Section 4.2.1 let us take $\mathbb{E}[X|Y]$ as candidate. We have for any $W \in \mathcal{S}(Y)$

$$\begin{aligned}\langle X - \mathbb{E}[X|Y], W \rangle &= \mathbb{E}[(X - \mathbb{E}[X|Y])W] \\ &= \mathbb{E}[XW] - \mathbb{E}[W\mathbb{E}[X|Y]] = \mathbb{E}[XW] - \mathbb{E}[\mathbb{E}[WX|Y]] \\ &= \mathbb{E}[XW] - \mathbb{E}[WX] = 0 \end{aligned} \tag{4.18}$$

where the third equality follows from (4.15). We have thus proved that the conditional expectation is the orthogonal projection of X on $\mathcal{S}(Y)$ which we write as

$$\mathbb{E}[X|Y] = X_{\|\mathcal{S}(Y)} \tag{4.19}$$

A rather remarkable result considering that $\mathcal{S}(Y)$ is infinite dimensional.

We recoqnize (4.12) as Pythagoras relation

$$\|X - X_{\|\mathcal{S}(Y)}\|^2 = \|X\|^2 - \|X_{\|\mathcal{S}(Y)}\|^2 \tag{4.20}$$

We can extend the setting to the case where both $\mathbf{X}$ and $\mathbf{Y}$ are m-dimensional random vectors by considering the space $\mathcal{L}_2^m$. Then $\mathbb{E}[\mathbf{X}|\mathbf{Y}]$ is a vector of the conditional expectations $\mathbb{E}[\mathbf{X}(i)|\mathbf{Y}]$ and each element of $\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Y}]$ is orthogonal to all measurable functions of $\mathbf{Y}$ that have bounded second moments. We can then extend Pythagoras relation (4.20) to a matrix equality[1]

$$\lfloor \mathbf{X} - \mathbf{X}_{\|\mathcal{S}(\mathbf{Y})}, (\mathbf{X} - \mathbf{X}_{\|\mathcal{S}(\mathbf{Y})})^T \rfloor = [\mathbf{X}, \mathbf{X}] - \left[\mathbf{X}_{\|\mathcal{S}(\mathbf{Y})}, \mathbf{X}_{\|\mathcal{S}(\mathbf{Y})}\right] \tag{4.21}$$

Using that $\mathbb{E}\left[\mathbf{X}_{\|\mathcal{S}(\mathbf{Y})}\right] = \mathbb{E}[\mathbf{X}]$ (recall that the conditional mean has the same mean as the random variable itself), we can re-write this as

$$\operatorname{Cov}\left\{\mathbf{X} - \mathbf{X}_{\|\mathcal{S}(\mathbf{Y})}\right\} = \operatorname{Cov}\{\mathbf{X}\} - \operatorname{Cov}\left\{\mathbf{X}_{\|\mathcal{S}(\mathbf{Y})}\right\} \tag{4.22}$$

It follows that

$$\operatorname{Cov}\left\{\mathbf{X} - \hat{\mathbf{X}}(\mathbf{Y})\right\} \geq \operatorname{Cov}\left\{\mathbf{X} - \mathbf{X}_{\|\mathcal{S}(\mathbf{Y})}\right\} \tag{4.23}$$

for any (measurable) estimator $\hat{\mathbf{X}}(\mathbf{Y})$ with equality iff $\hat{\mathbf{X}}(\mathbf{Y}) = \mathbf{X}_{\|\mathcal{S}(\mathbf{Y})} = \mathbb{E}[\mathbf{X}|\mathbf{Y}]$ a.e.

4.2.4 *Linear Estimators*

It is not always easy to compute the conditional mean, and also the entire joint distribution of X, Y may not be known. It may therefore be attractive to project on other subspaces than $\mathcal{S}(Y)$. The derivations in the preceeding section still hold. However, (4.22) holds only if

[1] See notations.

the subspace is such that the projection has the same mean as the estimated variable.

Consider estimating the random variable $X \in \mathbb{R}$ from the random vector $\mathbf{Y} \in \mathbb{R}^N$, Sacrificing accuracy, an estimator simpler to compute requiring only second order moments is the linear estimator

$$\hat{X} = \mathbf{LY}$$

To solve this case we let $\mathcal{S}_L(Y)$ be the subspace $\mathcal{S}_L(Y) = \text{Span}\{\mathbf{Y}(1), \dots, \mathbf{Y}(n)\}$. Since this is a finite dimensional subspace spanned by $\mathbf{Y}(1), \dots, \mathbf{Y}(n)$ it is sufficient that the orthogonality condition (4.17) holds for $W = \mathbf{Y}(1), \dots, \mathbf{Y}(n)$

$$\langle X - \mathbf{LY}, \mathbf{Y}(k) \rangle = 0 \; k = 1, \dots, n \tag{4.24}$$

which, using (C.9), can be written

$$\mathbf{L}\lfloor \mathbf{Y}, \mathbf{Y}^T \rfloor = \lfloor X, \mathbf{Y}^T \rfloor \tag{4.25}$$

i.e. the orthogonal projection is given by

$$X_{\|\mathcal{S}_L(\mathbf{Y})} = \lfloor X, \mathbf{Y}^T \rfloor \lfloor \mathbf{Y}, \mathbf{Y}^T \rfloor^{-1} \mathbf{Y} \tag{4.26}$$

We call $X_{\|\mathcal{S}_L(\mathbf{Y})}$ the optimal linear estimator (OLE) of X given $\mathbf{Y}$. The OLE only depends on the second order properties of $X, \mathbf{Y}$. Since $\mathcal{S}_L(\mathbf{Y}) \subset \mathcal{S}(\mathbf{Y})$ it follows that

$$\|X - X_{\|\mathcal{S}(\mathbf{Y})}\|^2 \leq \|X - X_{\|\mathcal{S}_L(\mathbf{Y})}\|^2$$

Unless the means of X and $\mathbf{Y}$ are zero, the OLE may be biased. A simple way to improve the estimator is then to extend $\mathbf{Y}$ with a constant element, we denote the resulting subspace $\mathcal{S}_{L_e}(\mathbf{Y})$. This is equivalent to adding a constant term to the OLE. Adjusting this term such that the mean of the OLE is the same as the mean of X gives the smallest MSE[2]. The simplest way to do this is to construct the OLE for $X - \mathbb{E}[X]$ given $\mathbf{Y} - \mathbb{E}[\mathbf{Y}]$ and then to add $\mathbb{E}[X]$ to the estimator:

$$X_{\|\mathcal{S}_{L_e}(\mathbf{Y})} =$$
$$\lfloor X - \mathbb{E}[X], (\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^T \rfloor \lfloor \mathbf{Y} - \mathbb{E}[\mathbf{Y}], (\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^T \rfloor^{-1} (\mathbf{Y} - \mathbb{E}[\mathbf{Y}]) + \mathbb{E}[X]$$

The OLE can be more compactly expressed as

$$X_{\|\mathcal{S}_{L_e}(\mathbf{Y})} = \text{Cov}\{X, \mathbf{Y}\} \, \text{Cov}^{-1}\{\mathbf{Y}\} \, (\mathbf{Y} - \mathbb{E}[\mathbf{Y}]) + \mathbb{E}[X] \tag{4.27}$$

which has covariance

$$\text{Cov}\left\{X_{\|\mathcal{S}_{L_e}(\mathbf{Y})}\right\} = \text{Cov}\{X, \mathbf{Y}\} \, \text{Cov}^{-1}\{\mathbf{Y}\} \, \text{Cov}\{\mathbf{Y}, X\}$$

and Pythagoras relation (4.22) becomes

$$\text{Cov}\left\{\mathbf{X} - \mathbf{X}_{\|\mathcal{S}(\mathbf{Y})}\right\} = \text{Cov}\{\mathbf{X}\} - \text{Cov}\{X, \mathbf{Y}\} \, \text{Cov}^{-1}\{\mathbf{Y}\} \, \text{Cov}\{\mathbf{Y}, X\} \tag{4.28}$$

We can also estimate a random vector $\mathbf{X} = \begin{bmatrix} X(1) & \dots X(m) \end{bmatrix}^T$ using linear combinations of $\mathbf{Y}$. With now $\mathbf{L}$ being a matrix, each element of $\mathbf{X}$ can be estimated as above rendering exactly the same

[2] Why?

expression as above for $\mathbf{X}_{\|\mathcal{S}_{L_e}(\mathbf{Y})}$ and its covariance, and the matrix expression (4.22) for Pythagoras relation remain intact.

For future use we file the information that whenever we encounter an expression that can be written on the form

$$\lfloor x, y \rfloor \lfloor y, y \rfloor^{-1} \lfloor y, x \rfloor \tag{4.29}$$

for vectors x and y whose elements are in a Hilbert space, then we can interpret this expression as the norm of the projection of x on the span of y.

4.2.5 *Maximum A Posteriori Estimation*

The Maximum A Posteriori (MAP) estimate given $Y = y$ is defined as the mode of the posterior density

$$\hat{X}_{MAP}(y) = \arg\max p_{X|Y}(x|y)$$

4.2.6 *Other Estimation Criteria**

A generalization of the MSE is to include a positive weighting function W which may depend on X (the variable we want to estimate), giving

$$\mathbb{E}\left[W(X)\|X - \hat{X}(Y)\|^2\right]$$

as criterion. The optimal estimator for this criterion is

$$\hat{X}(Y) = \frac{\mathbb{E}\left[W(X)X|Y)\right]}{\mathbb{E}\left[W(X)|Y)\right]}$$

while the criterion

$$\mathbb{E}\left[\|X - \hat{X}(Y)\|\right]$$

leads to the optimal estimator being the median of the conditional distribution.

4.3 *Wold decomposition*

Let us now return to the issue of how to model a stationary stochastic process $\{v(t)\}_{t=-\infty}^{\infty}$ that we discussed extensively in Section 3.4.7. It turns out that there is yet more to say on this subject and we will consider the more general class of wide-sense stationary processes.

We will consider the Wold decomposition which has as basis a linear projection of $v(t)$ on the span of its past $v(t-1), v(t-2), \ldots$. Denoting this projection $\hat{v}(t|t-1)$ and the prediction error $v(t) - \hat{v}(t|t-1)$ by $e(t)$, we can write

$$v(t) = e(t) + \hat{v}(t|t-1) = e(t) + \sum_{k=1}^{\infty} \alpha(k) v(t-k) \tag{4.30}$$

for some sequence $\{\alpha(k)\}_{k=1}^{\infty}$. However, we can now replace $v(t-1)$ by the same type of expansion, giving

$$\begin{aligned} v(t) &= e(t) + \alpha(1)v(t-1) + \sum_{k=2}^{\infty} \alpha(k)v(t-k) \\ &= e(t) + \alpha(1)e(t-1) + \sum_{k=2}^{\infty} \beta(k)v(t-k) \end{aligned}$$

for a new sequence of coefficients $\{\beta(k)\}$. Continuing like this suggests that we should obtain

$$v(t) = \sum_{k=0}^{\infty} h(k)e(t-k) \tag{4.31}$$

for some sequence $\{h(k)\}$, for which $h(0) = I$. Since $e(t)$ is orthogonal to the past and $e(s)$, $s < t$, is a function of the past, $\{e(t)\}$ form an uncorrelated sequence. In view of (4.30) we can see $e(t)$ as the linearly unpredictable part of $v(t)$ given the entire past of the process. The process $\{e(t)\}$ is therefore called the *innovation-process* and the quantity

$$\Sigma := \mathbb{E}\left[\left(v(t) - \hat{v}(t|t-1)\right)\left(v(t) - \hat{v}(t|t-1)\right)^T\right]$$

is called the *prediction error matrix*.

The representation (4.31) emphasizes that one could model a stationary process as filtered white noise. However, the representation does not cover all stationary processes.

Example 4.6. *Suppose that*

$$v(t) = x, \quad -\infty < t < \infty$$

where x is a random variable. Clearly $\{v(t)\}$ is stationary. Then $\hat{v}(t|t-1) = v(t-1) = x$ and $e(t) = 0$. We can thus not represent $v(t)$ on the form (4.31).

In the example, x is perfectly predictable given the past of $v(t)$. A formal definition of this concept is that a process $\{v(t)\}$ is said to be *linearly singular* (or linearly, purely deterministic) if $\hat{v}(t|t-k) = v(t)$ for a pair t, k, $k > 0$. It then follows from stationarity that this holds for any t and by iterating over k it follows that this also holds for any $k > 0$. A process for which $\hat{v}(t|t-k) \neq v(t)$, for some $k \in \mathbb{N}$ holds is said to be linearly non-deterministic. For a wide-sense stationary process it then holds that $\hat{v}(t|t-k) \neq v(t)$, $\forall t$ and $\forall k \in \mathbb{N}$.

It turns out that any stationary process can be represented if a linearly singular term is added to (4.31). We begin with an example.

Example 4.7. *Suppose that*

$$v(t) = e(t) + 0.5e(t-1) + x, \quad -\infty < t < \infty \tag{4.32}$$

where x is a random variable and $\{e(t)\}$ is zero mean white noise. From Example 4.6, $v_d(t) := x$ is linearly singular. We now show that $v_d(t)$ also can be perfectly predicted using the past of $v(t)$. For this, let

$$\hat{v}_d^N(t) := \frac{1}{N}\sum_{k=1}^{N} v(t-k) = x + \frac{1}{N}\sum_{k=1}^{N} e(t-k) + 0.5\frac{1}{N}\sum_{k=1}^{N} e(t-1-k)$$

which tends to $v_d(t) = x$ as $N \to \infty$ since the two last terms tend to zero by the law of large numbers as $\{e(t)\}$ is white noise. Thus $v_d(t)$ is perfectly predictable given the past of $v(t)$ and since $\{e(t)\}$ is an uncorrelated sequence, (4.32) *can be written as*

$$v(t) = \sum_{k=0}^{\infty} h(k)e(t-k) + v_d(t), \; -\infty < t < \infty, \tag{4.33}$$

with $h(0) = 1$, $h(1) = 0.5$, $h(k) = 0$, $k \geq 2$ and where $v_d(t) = x$ is a linearly singular process.

Next, notice that for a process that can be decomposed as (4.31) it must hold that

$$\hat{v}(t|t-s) = \sum_{k=s}^{\infty} h(k)e(t-k)$$

and that

$$\mathbb{E}\left[\hat{v}(t|t-s)\hat{v}^T(t|t-s)\right] = \sum_{k=s+1}^{\infty} h(k)\Sigma h^T(k) \to 0 \quad \text{as } s \to \infty$$

Hence[3]

$$\lim_{s\to\infty} \hat{v}(t|t-s) = 0$$

A process with this property is said to be *linearly regular* (linearly, purely non-deterministic). Such processes can be characterized in several different ways.

Theorem 4.3.1 (Theorem 6.13 in [4]). *The following statements are equivalent for a wide-sense stationary process $\{v(t)\}$:*

i) $\{v(t)\}$ is linearly regular

ii) $v(t) = A(q)w(t)$ for some $A(q) = \sum_{k=0}^{\infty} a(k)q^{-k}$ where $\{w(t)\}$ is white noise.

iii) $v(t)$ belongs to its remote past $\mathcal{S}_{-\infty}(v)$, $v(t) \in \mathcal{S}_{-\infty}(v)$, defined as $\mathcal{S}_{-\infty}(v) = \cap_{t=-\infty}^{\infty} \mathcal{S}_t(v)$ where

$$\mathcal{S}_t(v) = \text{Span}\{v(s) : -\infty < s \leq t\},$$

Notice that with necessity the limit $\sum_{k=0}^{\infty} a^T(k)Ka(k)$, where $K := \mathbb{E}\left[w(t)w^T(t)\right]$, exists in ii) as a wide-sense stationary process is assumed to have finite variance.

We are now ready the following decomposition result.

Theorem 4.3.2 (Wold Decomposition, Theorem 6.11 in [5]). *A wide-sense stationary stochastic process $\{v(t)\}_{t=-\infty}^{\infty}$ can be uniquely decomposed as*

$$v(t) = v_r(t) + v_d(t), \tag{4.34}$$

where $\{v_r(t)\}$ and $\{v_d(t)\}$ are uncorrelated and linearly regular and linearly singular, respectively.

[3] Here we see $\{v(t|t-s)\}_{s=1}^{\infty}$ as elements of the Hilbert space of random variables with finite variance, meaning that a random variable that is 0 has variance 0.

The linearly regular term $v_r(t)$ can be written as

$$v_r(t) = H(q)e(t) \tag{4.35}$$

where $H(q) = \sum_{k=0}^{\infty} h(k)q^{-k}$ and $\{e(t)\}$ is the innovation-process for $\{v(t)\}$, having the same dimension as $\{v(t)\}$, which has the following properties

i) $\mathbb{E}\left[e(t)\right] = 0$ *and* $\mathbb{E}\left[e(t)e^T(t-\tau)\right] = \Sigma\,\delta(\tau)$ *where* $\Sigma \geq 0$ *is the prediction error matrix.*

ii) $\mathbb{E}\left[v(0)e^T(t-k)\right] = h(k)\Sigma$, *where* $h(0)\Sigma = \Sigma = \Sigma h^T(0)$

iii)

$$\sum_{k=0}^{\infty} h(k)\Sigma h^T(k) < \infty \tag{4.36}$$

iv) $\mathbb{E}\left[e(t)v_d^T(t)\right] = 0$ *for all* $s, t \in \mathbb{Z}$.

v) $e(t) \in \mathcal{S}_t(v)$

and where the linearly singular term $v_d(t) \in S_{-\infty}(v)$ for all $t \in \mathbb{Z}$, implying that

$$\mathbb{E}\left[\left(v_d(t) - v_d(t)_{\|\mathcal{S}_{t-1}(v)}\right)\left(v_d(t) - v_d(t)_{\|\mathcal{S}_{t-1}(v)}\right)^T\right] = 0$$

For multivariable processes, the linearly singular term of a process may not be the only quantity that is perfectly predictable.

Example 4.8. *Suppose that*

$$v(t) = \begin{bmatrix} w(t) \\ w(t-1) \end{bmatrix}$$

where $\{w(t)\}$ is white noise with variance λ. Here we see that while the best estimate of $v_1(t)$ is zero, we can predict $v_2(t)$ perfectly by $v_1(t-1)$, giving

$$\hat{v}(t|t-1) = \begin{bmatrix} 0 \\ v_1(t-1) \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} v(t-1)$$

and hence the prediction error matrix. given by

$$\Sigma = \mathbb{E}\left[\begin{bmatrix} w(t) \\ 0 \end{bmatrix}\begin{bmatrix} w(t) \\ 0 \end{bmatrix}^T\right] = \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix}$$

is singular since $v_2(t)$ can be perfectly predicted.

The *rank* of a stationary process $\{v(t)\}$ is the rank of its prediction error matrix. Thus when this matrix is not full rank, there are linear combinations of $v(t)$ that can be perfectly predicted. However, there is also another issue, namely that the representation (4.39) is not unique, see Exercise 4.6. This is the reason for that $h(0)$ is only determined as in ii) in the theorem.

Corollary 4.3.1. *When the process $\{v(t)\}$ is full rank the decomposition in Theorem 4.3.2 is unique and $h(0) = I$ and $H(z) := \sum_{k=0}^{\infty} = \sum_{k=0}^{\infty} h(k)z^{-k} \in H_2$ so that the expression* (4.39) *for the linearly regular term can be written*

$$v_r(t) = H(q)e(t)$$

Proof. Being wide-sense stationary, the variance of $v(t)$ is finite and hence

$$\infty > \mathbb{E}\left[v(t)v^T(t)\right] \geq \mathbb{E}\left[v_r(t)v_r^T(t)\right] = \sum_{k=0}^{\infty} h(k)\Sigma h^T(k)$$

where the last equality is due to Theorem 3.4.7. Now if Σ is full rank this implies that $H(z) \in H_2$. □

In Theorem 3.4.4 we characterized a wide-sense stationary process in terms of its spectral distribution function. A matrix valued distribution function F inherits the properties of a one dimensional distribution function.

Theorem 4.3.3. *A matrix valued distribution function F is differentiable almost everywhere, has at most a countable number of discontinuities, and can be decomposed into three terms*

$$F = F_a + F_d + F_s$$

where F_u is absolutely continuous, F_d is piecewise constant and F_s, called the singular part, is continuous with zero derivative almost everywhere. Furthermore, the derivative $F' \in L_1(\mathbb{T})$.

Proof. We refer to Sections 4 and 7 of [6]. For showing that $F' \in L_1$, we note that

$$\infty > F(\pi) \geq \int_{-\pi}^{\pi} F'(e^{i\omega})d\omega = \int_{-\pi}^{\pi} |F'(e^{i\omega})|d\omega$$

where the equality follows from the non-increasing property of F. □

Let us now relate the Wold decomposition to this representation. First we notice that Theorem 3.4.7 implies that v_r has an absolutely continuous spectral distribution function with spectrum

$$\Phi_{v_r}(e^{i\omega}) = A(e^{i\omega})\Sigma A^*(e^{i\omega}), \quad A(e^{i\omega}) := \sum_{k=0}^{\infty} a(k)e^{-i\omega k}$$

This means that v_r contributes to the absolute continuous part of the spectral distribution function. The question is now if the linearly singular part v_d can contribute to this part as well.

Example 4.9. *Let $\{e(t)\}$ and $\{w(t)\}$ be mutually independent with $\{e(t)\}$ being white noise with variance λ and with $\{w(t)\}$ having the spectrum*

$$\Phi_w(e^{i\omega}) = \mathbb{1}_{[-1,1]}(\omega)$$

Then $\{w(t)\}$ is linearly singular[7] *whereas $\{e(t)\}$ is linearly regular. Now*

[6] N. Wiener and P. Masani. The prediction theory of multivariate stochastic processes: I. The regularity condition. *Acta Math*, 98:111–150, 1957

[7] See Exercise **??**.

set

$$v(t) = \begin{bmatrix} e(t) \\ w(t) \end{bmatrix}$$

Then the spectral distribution function of $\{v(t)\}$ *is absolutely continuous and the spectrum is*

$$\Phi_v(e^{i\omega}) = \begin{bmatrix} \lambda & 0 \\ 0 & \Phi_w(e^{i\omega}) \end{bmatrix}$$

From the preceding example we see that indeed the linearly singular part of a wide-sense process can contribute to the absolute continuous part of the spectral distribution function. Notice that in the example, the prediction error matrix is singular since $w(t)$ can be perfectly predicted from its past which is available as part of the past of $v(t)$, i.e. we have

$$\mathbf{\Sigma} = \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix}$$

It turns out that the situation changes radically if we exclude this case. To proceed with this, notice that $F' = F_a$ almost everywhere since F_d is piecewise constant, except at a countable number of points, and F_s has zero derivative almost everywhere. Hence, if v_r is not full rank then F' will be singular almost everywhere, which in turn means that $\det F' = 0$ almost everywhere so that $\log\det F'$ is not in $L_1(\mathbb{T})$. It turns out that we can characterize a full rank process by this condition.

Theorem 4.3.4 (7.10 Main Theorem I in [8]). *A wide-sense stationary process is full rank if and only if its spectral distribution function F is such that* $\log\det F' \in L_1(\mathbb{T})$.

Corollary 4.3.2. *For a full rank process,*

$$\log\det\Sigma = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\det F'(e^{i\omega})d\omega$$

For full rank processes we now have the following addendum to the Wold decomposition

Theorem 4.3.5 (7.11 Main Theorem II in [9]). *For a full rank wide-sense stationary process,* v_r *corresponds to the absolutely continuous part and* v_d *to the piecewise constant and singular parts of the spectral distribution function, respectively.*

Corollary 4.3.3 (7.12 Main Theorem III in [10]). *A wide-sense stationary process is full rank and regular if and only if its spectral distribution function is absolutely continuous with spectrum satisfying* $\log\det\Phi \in L_1(\mathbb{T})$.

From Theorem 4.3.5 we see that for a full rank process we cannot have that the linearly singular part of the process has a spectrum as in Example 4.3.

There is still one more loose end to tie up. From Wolds decomposition we have that a full rank linearly regular wide-sense stationary process can be written as

$$v(t) = H(q)e(t) = \sum_{k=0}^{\infty} h(k)e_{t-k}$$

and according to Theorem 4.3.1 its spectrum is given by

$$\Phi_v(e^{i\omega} = H(e^{i\omega})\mathbf{\Sigma}H^*(e^{i\omega})$$

The spectral factorization theorem, Theorem 3.4.8, shows that for wide-sense stationary processes where the spectrum has a rational structure and is of full rank almost everywhere there is a particular spectral factorization such that H has all its poles strictly inside the unit circle and $H^{-1}(z)$ all its poles in the unit disc. We now proceed with a general spectral factorization theorem. For this, let L_1 and L_2 be the classes of absolutely integrable and square integrable functions on $[-\pi, \pi]$, respectively, see Definition B.2.2.

Theorem 4.3.6. *Let* $\Phi \in L_1(\mathbb{T})$ *be non-negative, and let* $\log\det\Phi \in L_1(\mathbb{T})$*. Then there exists* $\mathbf{\Sigma} > 0$ *and* $H \in L_2(\mathbb{T})$ *for which the Fourier coefficients satisfy*

$$h(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} H(e^{i\omega})e^{i\omega k}d\omega = 0, \quad k < 0$$

and $h(0) = I$*, such that*

$$\Phi(e^{i\omega}) = H(e^{i\omega})\mathbf{\Sigma}H^*(e^{i\omega}) \quad \textit{almost everywhere}$$

The function H *is the radial limit of some* $\tilde{H} = \sum_{k=0}^{\infty}\tilde{h}(k)z^{-k} \in H_2$ *having the properties that* $\tilde{H}^{-1}(z)$ *is holomorphic i* $|z| > 1$*, that* $\tilde{H}(\infty) \geq 0$ *and that*

$$\log\det\mathbf{\Sigma} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\log\det\Phi(e^{i\omega})d\omega \tag{4.37}$$

Proof. We start with not requiring $h(0) = I$. Then, with the left-hand side of (4.37) being

$$2\log|\det\tilde{H}(\infty)|$$

all results except that $\tilde{H}^{-1}$ is holomorphic in $|z| > 1$ follows from Theorem 7.13 in [11]. For the last result we use the approach in the proof of Theorem 1 in [12]. For this we will use the transformation $z \to 1/z$ so that $\tilde{H}$ is analytic in $|z| < 1$. Since $\tilde{H} \in H_2$ it is also bounded on $|z| < 1$ and we can apply Jensen's inequality (Theorem 15.19 in [13]) to $\det\tilde{H}(z)$ giving

$$\begin{aligned}\log|\det\tilde{H}(0)| &\leq \frac{1}{2\pi}\int_{-\pi}^{\pi}\log|\det\tilde{H}(re^{i\omega})|d\omega \\ &\leq \frac{1}{2\pi}\int_{-\pi}^{\pi}\log|\det\tilde{H}(e^{i\omega})|d\omega, \quad 0 < r < 1\end{aligned}$$

Now $|\det\tilde{H}^*(re^{i\omega})| = |\det\tilde{H}(re^{i\omega})|$, and hence we can write this as

$$\begin{aligned}\log|\det\tilde{H}(0)\tilde{H}^*(0)| &\leq \frac{1}{2\pi}\int_{-\pi}^{\pi}\log|\det\tilde{H}(re^{i\omega})\tilde{H}^*(re^{i\omega})|d\omega \\ &\leq \frac{1}{2\pi}\int_{-\pi}^{\pi}\log|\det\Phi(e^{i\omega})|d\omega, \quad 0 < r < 1\end{aligned}$$

[11] N. Wiener and P. Masani. The prediction theory of multivariate stochastic processes: I. The regularity condition. *Acta Math*, 98:111–150, 1957

[12] P.E. Caines and L. Gerencsér. A simple proof for a spectral factorization theorem. *IMA Journal of Mathematical Control & Information*, 8:39–44, 1991

[13] W. Rudin. *Real and Complex Analysis*. McGraw-Hill, London, 1986

However, by (4.37) the lower bound equals the upper bound[14]. We thus have

$$\log|\det \tilde{H}(0)\tilde{H}^*(0)| = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|\det \tilde{H}(re^{i\omega})\tilde{H}^*(re^{i\omega})|d\omega, \quad 0 < r < 1$$

but according to Jenssen's formula (Theorem 5.18 in [15]) for this to hold $\det \tilde{H}(z)$ can have no zeros in $|z| < r$. The result now follows since $0 < r < 1$. An alternative is to base the proof on Theorem 17.17 in [16]. Now $2\log|\det \tilde{H}(\infty)|$ being finite means that $\tilde{H}(\infty) = \tilde{h}(0) > 0$, and hence we can factorize $\tilde{H}(z)$ so that $h(0) = I$ and take $\mathbf{\Sigma} = h(0)h^T(0) > 0$. □

The formula (4.37) is known as Szegös formula. We now summarize the results for full rank processes.

Theorem 4.3.7 (Wold Decomposition full rank processes). *A wide-sense stationary stochastic process $\{v(t)\}_{t=-\infty}^{\infty}$ is full rank if and only if its spectral distribution function F satisfies $\log\det F' \in L_1(\mathbb{T})$.*

Such a process can be uniquely decomposed as

$$v(t) = v_r(t) + v_d(t), \tag{4.38}$$

where $\{v_r(t)\}$ and $\{v_d(t)\}$ are uncorrelated and linearly regular and linearly singular, respectively.

The linearly regular process $\{v_r(t)\}$ can be expressed as

$$v_r(t) = H(q)e(t) \tag{4.39}$$

where $\{e(t)\}$ is white noise with its covariance matrix $\mathbf{\Sigma} > 0$ being the prediction error matrix and $e(t) \in \mathcal{S}_t(v)$, $t \in \mathbb{Z}$, and where $H(z) \in H_2$, $H(\infty) = I$ and with $H^{-1}(z)$ holomorphic in $|z| > 1$. The spectrum of v_r is $\Phi_r(e^{i\omega}) = H(e^{i\omega})\mathbf{\Sigma}H^*(e^{i\omega})$.

The linearly singular process $\{v_d(t)\}$ corresponds to the piecewise constant and singular parts of the spectral distribution function. Furthermore $\mathbb{E}\left[e(t)v_d^T(t)\right] = 0$ for all $s, t \in \mathbb{Z}$ and $v_d(t) \in S_{-\infty}(v)$ for all $t \in \mathbb{Z}$, implying that

$$\mathbb{E}\left[\left(v_d(t) - v_d(t)_{\|\mathcal{S}_{t-1}(v)}\right)\left(v_d(t) - v_d(t)_{\|\mathcal{S}_{t-1}(v)}\right)^T\right] = 0$$

It is common to restrict the model class to some parametrized family, such as

$$H(q) = \frac{C(q)}{D(q)} = \frac{1 + c_1 q^{-1} + \ldots + c_{n_c} q^{-n_c}}{1 + d_1 q^{-1} + \ldots + d_{n_d} q^{-n_d}} \tag{4.40}$$

where $z^{n_c}C(z)$ has no zeros outside the unit disc and where $z^{n_d}D(z)$ has no zeros on or outside the unit disc and where c_i, $i = 1, \ldots, n_c$ and d_i, $i = 1, \ldots, n_d$ are parameters (possibly constrained to some set). We stress, however, that when disturbances and noise undergo non-linear transformations in the model, the above model class is not sufficient as only second order characteristics are modeled. Then the finite dimensional distributions of the process need further modeling.

[14] Recall that in the proof we have $z \to 1/z$, meaning that the point $z = 0$ in the proof corresponds to $z = \infty$ in the theorem.

[15] W. Rudin. *Real and Complex Analysis*. McGraw-Hill, London, 1986

[16] W. Rudin. *Real and Complex Analysis*. McGraw-Hill, London, 1986

An important exception is when $\{v(t)\}$ is a Gaussian process. Then $\{e(t)\}$ will be an iid Gaussian process since uncorrelated Gaussian random variables are independent. This is an important result so we state it separately.

Corollary 4.3.4. *Any regular full rank stationary Gaussian stochastic process $\{v(t)\}$ can be decomposed as*

$$v(t) = H(q)e(t) + m \tag{4.41}$$

where $\{e(t)\}$ is a sequence of zero mean iid Gaussian random variables with $\mathbb{E}\left[e(t)e^T(t)\right] = \Sigma$ for some $\Sigma > 0$, and where $H(z) := \sum_{k=0}^{\infty} h(k)z^{-k}$, with $h(0) = I$, has no zeros or poles on or outside of the unit circle. The constant m is the mean of the process.

When higher order moments of $v(t)$ come into play in the model, a simple modification of (4.39) and (4.40) is to add the assumption that $\{e(t)\}$ is a sequence of iid random variables with some (possibly parametrized) pdf p_e.

In Exercise 4.7 it is discussed what it means that a spectrum looses rank.

4.4 *Exercises*

4.1. Let $X \in \mathbb{R}^m$ and $Y \in \mathbb{R}^n$ be two random vectors that are jointly Gaussian:

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m_X \\ m_Y \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix} \right)$$

Derive the conditional probability density function for X given Y. What is the conditional mean and the conditional covariance matrix?

Compare with the OLE. What conclusions can you draw?

4.2. A typical situation is that the distribution of the observed variable Y is known when the variable X to be estimated is given, i.e. $p(y|x)$. Suppose that X is Bernouilli distributed with probability p that $X = 0$ and suppose that $Y|X = 0$ is $\mathbb{N}(2, 1)$ and $Y|X = 1$ is $\chi^2(3)$. What is the conditional distribution of X given Y?

4.3. *The Monty Hall problem.* Suppose that you are in a game show where a car is hidden behind one of three closed doors. Initially you choose one of the doors and then the game host, which knows where the car is hidden, opens one of the other doors which is empty. You are now given the option of keeping the door that you selected in the first place or to change to the other closed door. Compute the posterior probabilities for which door the car hides behind, given your initial choice and the game hosts choice. What is the posterior mean? Which door has the maximum a posteriori probability? What is the optimal strategy and what are the winning chances?

This problem caused a big media ruccus in 1990 with an ensuing torrent of mails from the public in regards to the correct solution. Interestingly, 62% of the answers coming from PhDs were incorrect. Google after you have solved the problem!

4.4. Prove that (4.28) is positive using Schur complement.

4.5. Consider the problem of estimating $\mathbf{X}$ given $\mathbf{Z}$ when

$$\text{Cov}\left\{\begin{bmatrix}\mathbf{X}\\ \mathbf{Z}\end{bmatrix}\right\} = \begin{bmatrix}\Sigma_{\mathbf{X},\mathbf{X}} & \Sigma_{\mathbf{X},\mathbf{Y}}\\ \Sigma_{\mathbf{Y},\mathbf{X}} & \Sigma_{\mathbf{Y},\mathbf{Y}}\end{bmatrix}$$

Show that the largest estimation error in the MSE-sense is obtained when the joint distribution of $\mathbf{X}, \mathbf{Z}$ is Gaussian.

4.6. Show by construction that the innovations-process is not unique in Example 4.8.

4.7. Let $\Phi(e^{i\omega}) = H(e^{i\omega})\mathbf{\Sigma}H^*(e^{i\omega})$ be full rank for almost all ω.

a) Show that a process with Φ as spectrum is full rank.

b) Show that $H(e^{i\omega})$ cannot be a tall matrix, i.e. have more rows than columns.

c) Show that if $H(e^{i\omega})$ is a 'fat' matrix, i.e it has more columns than rows, then the spectrum can be written as $\Phi(e^{i\omega}) = \tilde{H}(e^{i\omega})\tilde{\mathbf{\Sigma}}\tilde{H}^*(e^{i\omega})$ where $\tilde{H}$ is square, $\tilde{H}(\infty) = I$ and $\tilde{\mathbf{\Sigma}} > 0$.

d) Suppose that $H(z)$ looses rank at z_o. Show that then there is a non-zero input u such that

$$y(t) = H(q)u(t) = \sum_{k=0}^{\infty} h(k)u(t-k) = 0, \quad t \in \mathbb{Z}$$

Such a point z_o is called a zero of $H(q)$.

e) Let $H(z)$ have the form

$$H(z) = C(zI - A)^{-1}B + D \tag{4.42}$$

Reformulate the statement in the spectral factorization theorem that "$H(z)$ can be taken such that $H^{-1}(z)$ is holomorphic in $|z| > 1$" in terms of the zeros of $H(z)$.

f) Give a necessary condition on the rank of D for $H^{-1}(z)$ to be holomorphic in $|z| > 1$.

g) Under the condition in f) and the matrix inversion lemma, Lemma A.2.1, to derive an expression for $H^{-1}(z)$.

h) Use the condition in f) and the state space realization

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

of $H(q)$ to derive an expression for $H^{-1}(z)$.

A
Matrix Algebra

A.1 Matrix norms

$|x|$ denotes the Euclidean norm of a vector x

$$|x| = \sqrt{\sum_k |x_k|^2}$$

The Frobenius norm for a matrix $A \in \mathbb{C}^{n \times m}$ is defined as

$$\|A\|_F = \sqrt{\sum_{i,j} |A_{i,j}|^2}$$

and the operator (or 2-) norm as

$$\|A\|_2 = \sup_x \frac{|Ax|}{|x|} = \overline{\sigma}(A)$$

where $\overline{\sigma}(A)$ is the largest singular value of A.

A.2 Matrix Inversion Lemma

Next follows a useful result on matrix inversion.

Lemma A.2.1. *Suppose that $A \in \mathbb{R}^{n \times n}$ and $C \in \mathbb{R}^{m \times m}$ are invertible. Then for $B \in \mathbb{R}^{n \times m}$, $D \in \mathbb{R}^{m \times n}$*

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} \tag{A.1}$$

A.3 Block Matrix Inversion

Let $\Delta_A = D - CA^{-1}B$ and $\Delta_D = A - BD^{-1}C$. Then

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B\Delta_A^{-1}CA^{-1} & -A^{-1}B\Delta_A^{-1} \\ -\Delta_A^{-1}CA^{-1} & \Delta_A^{-1} \end{bmatrix} = \begin{bmatrix} \Delta_D^{-1} & -\Delta_D^{-1}D^{-1}B \\ -D^{-1}C\Delta_D^{-1} & D^{-1} + D^{-1}C\Delta_D^{-1}BD^{-1} \end{bmatrix}$$

whenever the inverses exist.

A.3.1 Inverse of an Inner Product

As an application of the formulae above, let

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$$

and let $\theta_{Y_i|Y_j} = \langle Y_i, Y_j\rangle\langle Y_j, Y_j\rangle^{-1}$ be the coordinates for the projection of the elements of Y_i on the linear span of Y_j. Then

$$\langle Y, Y\rangle^{-1} = \begin{bmatrix} \langle Y_1Y_2, Y_1Y_2\rangle^{-1} & -\theta_{Y_2|Y_1}\langle Y_2Y_1, Y_2Y_1\rangle^{-1} \\ -\theta^T_{Y_1|Y_2}\langle Y_1Y_2, Y_1Y_2\rangle^{-1} & \langle Y_2Y_1, Y_2Y_1\rangle^{-1} \end{bmatrix} =: \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$$

Thus $Y_{1\|Y_2} = -A^{-1}BY_2$ and the "norm" of the error $Y_1 - Y_{1\|Y_2}$ is A^{-1}. Similarly $Y_{2\|Y_1} = -BD^{-1}Y_1$ with "norm" of the error being D^{-1}. Notice that this holds regardless of how Y is divided into Y_1 and Y_2.

A.4 Schur Complement

Consider the symmetric block-matrix

$$Z = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$$

If $A > 0$ then

$$\begin{aligned} Z > 0 &\Leftrightarrow C - B^T A^{-1} B > 0 \\ Z \geq 0 &\Leftrightarrow C - B^T A^{-1} B \geq 0 \end{aligned}$$

If $C > 0$ then

$$\begin{aligned} Z > 0 &\Leftrightarrow A - BC^{-1}B^T > 0 \\ Z \geq 0 &\Leftrightarrow A - BC^{-1}B^T \geq 0 \end{aligned}$$

For

$$Z = \left\langle \begin{bmatrix} X \\ Y \end{bmatrix}, \begin{bmatrix} X \\ Y \end{bmatrix} \right\rangle$$

we get the following special case: When $\langle X, X\rangle > 0$, Z is positive definite iff no linear combination of X can be linearly predicted exactly by Y. Symmetrically, when $\langle Y, Y\rangle > 0$, Z is positive definite iff no linear combination of Y can be linearly predicted exactly by X.

A.5 Completing the square

Lemma A.5.1. *Suppose that* $\in \mathbb{R}^n$, $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{n\times m}$ *and* $\boldsymbol{P}_y \in \mathbb{R}^{n\times n}$, $\boldsymbol{P}_y > 0$. *Then*

$$(\mathbf{y} - \mathbf{A}\mathbf{x})^T\mathbf{P}_y^{-1}(\mathbf{y} - \mathbf{A}\mathbf{x}) = (\mathbf{x} - \hat{\mathbf{x}})\mathbf{A}^T\mathbf{P}_y^{-1}\mathbf{A}(\mathbf{x} - \hat{\mathbf{x}}) + \|\mathbf{y} - \hat{\mathbf{y}}\|^2_{\mathbf{P}_y^{-1}} \tag{A.2}$$

where

$$\begin{aligned} \hat{\mathbf{x}} &= (\mathbf{A}^T\mathbf{P}_y^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{P}_y^{-1}\mathbf{y} \\ \hat{\mathbf{y}} &= \mathbf{A}\hat{\mathbf{x}} \end{aligned}$$

Suppose in addition that $\mathbf{z} \in \mathbb{R}^m$ *and* $\mathbf{P}_x \in \mathbb{R}^{m\times m}$, $\mathbf{P}_x > 0$. *Then*

$$\begin{aligned} &(\mathbf{y} - \mathbf{A}\mathbf{x})^T\mathbf{P}_y^{-1}(\mathbf{y} - \mathbf{A}\mathbf{x}) + (\mathbf{x} - \mathbf{z})^T\mathbf{P}_x^{-1}(\mathbf{x} - \mathbf{z}) \\ &= (\mathbf{x} - \tilde{\mathbf{x}})^T\mathbf{S}^{-1}(\mathbf{x} - \tilde{\mathbf{x}}) + (\mathbf{y} - \mathbf{A}\mathbf{z})^T\mathbf{T}^{-1}(\mathbf{y} - \mathbf{A}\mathbf{z}) \end{aligned} \tag{A.3}$$

where

$$\begin{aligned}
\mathbf{T} &= (\mathbf{P}_y + \mathbf{A}\mathbf{P}_x\mathbf{A}^T) \\
\mathbf{S} &= (\mathbf{P}_x^{-1} + \mathbf{A}^T\mathbf{P}_y^{-1}\mathbf{A})^{-1} = \mathbf{P}_x - \mathbf{P}_x\mathbf{A}^T\mathbf{T}^{-1}\mathbf{A}\mathbf{P}_x \\
\tilde{\mathbf{x}} &= \mathbf{z} + \mathbf{L}(\mathbf{y} - \mathbf{A}\mathbf{z}) \\
\mathbf{L} &= \mathbf{P}_x\mathbf{A}^T\mathbf{T}^{-1}
\end{aligned}$$

Proof. Expanding (A.2)

$$\begin{aligned}
(\mathbf{y} - \mathbf{A}\mathbf{x})^T\mathbf{P}_y^{-1}(\mathbf{y} - \mathbf{A}\mathbf{x}) &= \mathbf{x}^T\mathbf{A}^T\mathbf{P}_y^{-1}\mathbf{A}\mathbf{x} - 2\mathbf{y}^T\mathbf{P}_y^{-1}\mathbf{A}\mathbf{x} + \mathbf{y}^T\mathbf{P}_y^{-1}\mathbf{y} \\
&= (\mathbf{x} - \hat{\mathbf{x}})^T\mathbf{A}^T\mathbf{P}_y^{-1}\mathbf{A}(\mathbf{x} - \hat{\mathbf{x}}) - \hat{\mathbf{x}}^T\mathbf{A}^T\mathbf{P}_y^{-1}\mathbf{A}\hat{\mathbf{x}} + \mathbf{y}^T\mathbf{P}_y^{-1}\mathbf{y}
\end{aligned}$$

We see that $\mathbf{x} = \hat{\mathbf{x}}$ is the solution to the minimization problem

$$\arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_{\mathbf{P}_y^{-1}}$$

meaning that $\hat{\mathbf{y}} = \mathbf{A}\hat{\mathbf{x}}$ is the orthogonal projection of $\mathbf{y}$ on the column span of $\mathbf{A}$ and that therefore $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to the column span of $\mathbf{A}$, in turn implying that

$$\mathbf{y}^T\mathbf{P}_y^{-1}\mathbf{y} - \hat{\mathbf{x}}^T\mathbf{A}^T\mathbf{P}_y^{-1}\mathbf{A}\hat{\mathbf{x}} = \|\mathbf{y}\|^2_{\mathbf{P}_y^{-1}} - \|\hat{\mathbf{y}}\|^2_{\mathbf{P}_y^{-1}} = \|\mathbf{y} - \hat{\mathbf{y}}\|^2_{\mathbf{P}_y^{-1}}$$

For (A.3) we have

$$\begin{aligned}
&(\mathbf{y} - \mathbf{A}\mathbf{x})^T\mathbf{P}_y^{-1}(\mathbf{y} - \mathbf{A}\mathbf{x}) + (\mathbf{x} - \mathbf{z})^T\mathbf{P}_x^{-1}(\mathbf{x} - \mathbf{z}) \\
&= \mathbf{x}^T(\mathbf{P}_x^{-1} + \mathbf{A}^T\mathbf{P}_y^{-1}\mathbf{A})\mathbf{x} - 2\mathbf{x}^T(\mathbf{A}^T\mathbf{P}_y^{-1}\mathbf{y} + \mathbf{P}_x^{-1}\mathbf{z}) + \mathbf{y}^T\mathbf{P}_y^{-1}\mathbf{y} + \mathbf{z}^T\mathbf{P}_x^{-1}\mathbf{z} \\
&= (\mathbf{x} - \mathbf{S}\mathbf{w})^T\mathbf{S}^{-1}(\mathbf{x} - \mathbf{S}\mathbf{w}) - \mathbf{w}^T\mathbf{S}\mathbf{w} + \mathbf{y}^T\mathbf{P}_y^{-1}\mathbf{y} + \mathbf{z}^T\mathbf{P}_x^{-1}\mathbf{z} \qquad \text{(A.4)} \\
& \qquad \text{(A.5)}
\end{aligned}$$

where

$$\mathbf{w} = \mathbf{A}^T\mathbf{P}_y^{-1}\mathbf{y} + \mathbf{P}_x^{-1}\mathbf{z}$$

Using the Matrix Inversion Lemma (A.1),

$$\mathbf{T}^{-1} = (\mathbf{P}_y + \mathbf{A}\mathbf{P}_x\mathbf{A}^T)^{-1} = \mathbf{P}_y^{-1} - \mathbf{P}_y^{-1}\mathbf{A}\mathbf{S}\mathbf{A}^T\mathbf{P}_y^{-1} \qquad \text{(A.6)}$$

and furthermore, using Exercise 6.1,

$$\begin{aligned}
\mathbf{P}_x^{-1} - \mathbf{P}_x^{-1}\mathbf{S}\mathbf{P}_x^{-1} &= \mathbf{P}_x^{-1}\mathbf{S}(\mathbf{S}^{-1} - \mathbf{P}_x^{-1}) = \mathbf{P}_x^{-1}(\mathbf{P}_x^{-1} + \mathbf{A}^T\mathbf{P}_y^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{P}_y^{-1}\mathbf{A} \\
&= (\mathbf{I} + \mathbf{A}^T\mathbf{P}_y^{-1}\mathbf{A}\mathbf{P}_x)^{-1}\mathbf{A}^T\mathbf{P}_y^{-1}\mathbf{A} \\
&= \mathbf{A}^T\mathbf{P}_y^{-1}(\mathbf{I} + \mathbf{A}\mathbf{P}_x\mathbf{A}^T\mathbf{P}_y^{-1})^{-1}\mathbf{A} \\
&= \mathbf{A}^T(\mathbf{P}_y + \mathbf{A}\mathbf{P}_x\mathbf{A}^T)^{-1}\mathbf{A} = \mathbf{A}^T\mathbf{T}^{-1}\mathbf{A} \qquad \text{(A.7)}
\end{aligned}$$

Also

$$\begin{aligned}
\mathbf{P}_x^{-1}\mathbf{S}\mathbf{A}^T\mathbf{P}_y^{-1} &= \mathbf{P}_x^{-1}(\mathbf{P}_x^{-1} + \mathbf{A}^T\mathbf{P}_y^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{P}_y^{-1} = (\mathbf{I} + \mathbf{A}^T\mathbf{P}_y^{-1}\mathbf{A}\mathbf{P}_x)^{-1}\mathbf{A}^T\mathbf{P}_y^{-1} \\
&= \mathbf{A}^T(\mathbf{I} + \mathbf{P}_y^{-1}\mathbf{A}\mathbf{P}_x\mathbf{A}^T)^{-1}\mathbf{P}_y^{-1} = \mathbf{A}^T(\mathbf{P}_y + \mathbf{A}\mathbf{P}_x\mathbf{A}^T)^{-1} = \mathbf{A}^T\mathbf{T}^{-1} \\
& \qquad \text{(A.8)}
\end{aligned}$$

Using (A.6)–(A.8) gives

$$\begin{aligned}
&-\mathbf{w}^{\mathbf{T}}\mathbf{S}\mathbf{w} + \mathbf{y}^T\mathbf{P}_y^{-1}\mathbf{y} + \mathbf{z}^T\mathbf{P}_x^{-1}\mathbf{z} = \\
&\mathbf{y}^T(\mathbf{P}_y^{-1} - \mathbf{P}_y^{-1}\mathbf{A}\mathbf{S}\mathbf{A}^T\mathbf{P}_y^{-1})\mathbf{y} + \mathbf{z}^T(\mathbf{P}_x^{-1} - \mathbf{P}_x^{-1}\mathbf{S}\mathbf{P}_x^{-1})\mathbf{z} - 2\mathbf{z}^T(\mathbf{P}_x^{-1}\mathbf{S}\mathbf{A}^T\mathbf{P}_y^{-1})\mathbf{y} \\
&\mathbf{y}^T\mathbf{T}^{-1}\mathbf{y} + \mathbf{z}^T\mathbf{A}^T\mathbf{T}^{-1}\mathbf{A}\mathbf{z} - 2\mathbf{z}^T\mathbf{A}^T\mathbf{T}^{-1}\mathbf{y} = (\mathbf{y} - \mathbf{A}\mathbf{z})^T\mathbf{T}^{-1}(\mathbf{y} - \mathbf{A}\mathbf{z}) \quad \text{(A.9)}
\end{aligned}$$

Next

$$\begin{aligned}
\mathbf{S}\mathbf{w} &= (\mathbf{P}_x - \mathbf{P}_x\mathbf{A}^T\mathbf{T}^{-1}\mathbf{A}\mathbf{P}_x)(\mathbf{A}^T\mathbf{P}_y^{-1}\mathbf{A}\mathbf{y} + \mathbf{P}_x^{-1}\mathbf{z}) \\
&= \mathbf{P}_x\mathbf{A}^T\mathbf{T}^{-1}(\mathbf{T} - \mathbf{A}\mathbf{P}_x\mathbf{A}^T)\mathbf{P}_y^{-1}\mathbf{y} + \mathbf{z} - \mathbf{P}_x\mathbf{A}^T\mathbf{T}^{-1}\mathbf{A}\mathbf{z} \\
&= \mathbf{P}_x\mathbf{A}^T\mathbf{T}^{-1}\mathbf{y} + \mathbf{z} - \mathbf{P}_x\mathbf{A}^T\mathbf{T}^{-1}\mathbf{A}\mathbf{z} \qquad \text{(A.10)}
\end{aligned}$$

Inserting (A.9) and (A.10) in (A.4) now gives the result. □

B
Bits and Pieces of Complex Analysis

We follow the notation in [1], from which also much of the results, proofs, and references, can be found. The symbols C, $D_+(r)$ and $D_-(r)$ will denote the sets $|z| = 1$, $|z| < r$ and $r < |z| < \infty$, respectively, of the extended complex plane. $D_+(1)$ and $D_(1)$ will be denoted by D_+ and $D-$, respectively.

B.1 Holomorphic functions

Definition B.1.1. *Suppose that the limit*

$$\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}$$

exists. Then this limit is called the derivative of f at z_0. A function f is said to be holomorphic (analytic) in an open set Ω if it is differentiable at every point in Ω.

Theorem B.1.1 (Cauchy integral formula)**.** *Let f be analytic in the open set Ω and let $Q \subset \Omega$ denote a closed contour. Then*

$$f^{(k)}(z) = \frac{n!}{2\pi i} \oint_Q \frac{f(s)ds}{(s-z)^{n+1}}$$

From the Cauchy integral formula the following important result follows.

Theorem B.1.2 (Theorem 10.16 in [2])**.** *Every holomorphic function $f(z)$ in an open set Ω, can be represented by a unique power series expansion*

$$f(z) = \sum_{k=0}^{\infty} c_k (z-a)^n \tag{B.1}$$

in an open disc belonging to Ω and centered at a.

Corollary B.1.1. *The coefficients in the power series expansion are given by*

$$c_k = \frac{f^{(k)}(a)}{k!}$$

In particular, if f is holomorphic in $D_+(r)$, $a = 0$ can be used in the theorem.

[1] N. Wiener and P. Masani. The prediction theory of multivariate stochastic processes: I. The regularity condition. *Acta Math*, 98:111–150, 1957

B.2 Hardy classes H_p and limits of functions in H_p

We now add some regularity to holomorphic functions.

Definition B.2.1. *The Hardy class H_p, $p > 0$, consists of all complex-valued holomorphic functions f on D_+ for which there exists a constant M such that*

$$\int_{-\pi}^{\pi} |f(re^{\omega})|^p d\omega \leq M, \quad 0 < r < 1$$

From Theorem B.1.2, a function $f \in H_p$, $p > 0$, on D_+ can be represented as

$$f(re^{i\omega}) = \sum_{k=0}^{\infty} c_k r^k e^{i\omega k}, \quad 0 \leq r < 1 \tag{B.2}$$

The class H_2 can be exactly characterized by this expansion.

Theorem B.2.1 (Theorem 17.12 [3]). *Suppose that f is analytic in D_+ and*

$$f(z) = \sum_{k=0}^{\infty} c_k z^k \tag{B.3}$$

Then $f \in H_2$ if and only if $\sum_{k=0}^{\infty} |c_k|^2 < \infty$.

Now if the stronger condition

$$\sum_{k=0}^{\infty} |c_k| < \infty \tag{B.4}$$

holds, (B.2) implies that the limit $\lim_{r \to 1-} f(re^{i\omega})$ is well defined and define the function

$$f(e^{i\omega}) = \sum_{k=0}^{\infty} c_k e^{i\omega k}$$

on C. This means that we can see the expansion (B.3) as valid in $|z| \leq 1$. Notice also that the function f is absolute integrable on C

$$\int_{-\pi}^{\pi} |f(e^{i\omega})| d\omega = \int_{-\pi}^{\pi} |\sum_{k=0}^{\infty} c_k e^{i\omega k}| d\omega \leq 2\pi \sum_{k=0}^{\infty} |c_k| < \infty$$

For a more precise formulation of this type of result we introduce a function class on C.

Definition B.2.2. *The class L_p consists of all complex-valued measurable functions f on C for which*

$$\int_{-\pi}^{\pi} |f(e^{i\omega})|^p d\omega < \infty$$

Theorem B.2.2 (Theorem 2.6 in [4]). *Suppose $f_+ \in H_p$, $p > 0$. Then*

(a) $f(e^{i\omega}) = \lim_{r \to 1-} f_+(re^{i\omega})$ exists a.e. on C and $f \in L_p$.

(b) The convergence is also in the L_p-topology.

It turns out that we can characterize the magnitude of f_+ at the origin through its limit function in L_p, somewhat reminiscent of the Cauchy integral formula. This is an important result from estimation point of view. Introducing $g(z) = \sum_{k=0}^{\infty} \bar{c}_k z^{-k}$, and making the change of variable $z = e^{i\omega}$ gives

$$\begin{aligned}
\frac{1}{2\pi}\int_{-\pi}^{\pi} \log|f(e^{i\omega})|\, d\omega &= \frac{1}{4\pi}\int_{-\pi}^{\pi} \log|f(e^{i\omega})|^2\, d\omega \\
&= \frac{1}{4\pi}\int_{-\pi}^{\pi} \log f(e^{i\omega})g(e^{i\omega})\, d\omega \\
&= \frac{1}{4\pi i}\oint_C \log f(z)g(z)\, \frac{dz}{z} \\
&= \frac{1}{4\pi i}\oint_C \log f(z)\, \frac{dz}{z} + \frac{1}{4\pi i}\oint_C \log g(z)\, \frac{dz}{z}
\end{aligned}$$

Now, assume f is analytic in the set $D_+(r)$ for some $r > 1$, e.g. $f(z)$ is a rational function with all poles having radius larger than r. Then the first integral is obtained from the Cauchy integral formula as $\frac{1}{2}\log f(0)$ if f does not have any zeros on $D_+(r)$ then $\log f(z)$ is holomorphic on $D_+(r)$. Using the variable transformation $s = z^{-1}$ on the second integral, one then similarly get that this integral is equal to $\frac{1}{2}\log \overline{f(0)}$. We thus have shown that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \log|f(e^{i\omega})|\, d\omega = \frac{1}{2}\log f(0) + \frac{1}{2}\log\overline{f(0)} = \frac{1}{2}\log|f(0)|^2 = \log|f(0)|$$

under the assumption that f is holomorphic and non-zero in a set including C. We can write this as

$$|f(0)| = e^{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log|f(e^{i\omega})|\, d\omega}$$

The general result is as follows.

Corollary B.2.1. *Under the same assumptions as in Theorem B.2.2 and with $f_+ \neq 0$, then*

(c) $\log|f| \in L_1$ *on C and*

$$|f_+(0)| \leq e^{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log|f(e^{i\omega})|\, d\omega} \tag{B.5}$$

(d) Equality in (B.5) *can only hold if f_+ has no zeros on D_+.*

Notice that (c) implies that $f(e^{i\omega}) \neq 0$ a.e.

B.3 Mapping L_p to H_p

Theorem B.2.2 implies that $f \in L_p$, $1 \leq p \leq \infty$, is a limit of a function in H_p only if the Fourier coefficients

$$c_k := \frac{1}{2\pi}\int_{-\pi} f(e^{i\omega})e^{-i\omega k} d\omega$$

are zero for negative k.

It can be shown that for $1 < p < \infty$, setting $f_+(z) = \sum_{k=0}^{\infty} c_k z^k$ defines a function in H_p and that all functions in H_p, $1 < p < \infty$ can be obtained in this way (Exercise 25, Chapter 17 [5]).

For $f \in L_1$, the Fourier coefficients tend to zero as $|k| \to \infty$ due to the Riemann-Lesbegue lemma (5.14 in [6]) and hence

$$f_+(z) = \sum_{k=0}^{\infty} c_k z^k, \quad f_-(z) = \sum_{k=1}^{\infty} c_k z^{-k}$$

are defined point-wise on D_+, and D_-. These are called the inner and outer functions determined by f.

Theorem B.3.1 (Theorem 2.4 in [7]). *Let $f \in L_1$ on C then*

$$\lim_{r \to 1-} f_+(re^{i\omega}) + f_-(r^{-1}e^{i\omega}) = f(e^{i\omega})$$

Corollary B.3.1 (Corollary 2.5 [8]). *Suppose that $f \in L_p$, $p \geq 1$, on C and $f_- = 0$, then $f_+ \in H_p$ on D_+.*

Under the assumptions of the corollary, we see that we can view an $f \in L_p$ with Fourier coefficients $c_k = 0$ for $k < 0$, as defined on $|z| \leq 1$ by the function

$$f(z) = \sum_{k=0}^{\infty} c_k z^k$$

holomorphic in D_+ and where on C the expression in general is to be interpreted as a limit, but when $\sum_k |c_k| < \infty$, the expression is valid in the entire of $|z| \leq 1$. In view of Theorem B.2.1, the latter applies when $p = 2$. Notice that f may not be analytic on C as it may not be defined for any z outside C.

B.4 Positive functions on L_1

In this section we consider an extension of the Fejér-Riesz theorem to L_p.

Theorem B.4.1 (Fejér-Riesz theorem). *A trigonometric polynomial $g(e^{i\omega}) = \sum_{k=-n}^{n} c_k e^{i\omega k}$ assumes non-negative values if and only if it can be expressed as $g(e^{i\omega}) = |p(e^{i\omega})|^2$ for some polynomial $p(z) = \sum_{k=0}^{n} a_k z^k$. The polynomial can be chosen to have no roots in D_+, and is then unique up to a multiplicative constant of modulus one.*

Theorem B.4.2 (Theorem 2.8 in [9]). *Let $g \in L_p$ on C, $p \geq 0$, $g \geq 0$ and suppose that* $\log g \in L_1$. *Then there exists $f_+ \in H_p$ on D_+ without zeros in D_+ such that its radial limit f satisfies $|f| = g$ a.e. on C and*

$$|f_+(0)| = e^{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log g(e^{i\omega})\, d\omega}$$

As a particular application of Theorem B.4.2, suppose that g satisfies the conditions of the theorem with $p = 1$. Then $\sqrt{g} \in L_2$ since $g \in L_1$ and $\log \sqrt{g} = \frac{1}{2} \log g \in L_1$ since $\log g \in L_1$. Thus we can apply Theorem B.4.2 to $\sqrt{g}$ giving that there exists an $f \in H_2$ without zeros in D_+ such that $|f(e^{i\omega})|^2 = g(e^{i\omega})$ and

$$|f(0)|^2 = e^{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log g(e^{i\omega})\, d\omega}$$

[5]

[6]

B.5 Bibliography

[10] Chapter 10 in [11]

[10] N. Wiener and P. Masani. The prediction theory of multivariate stochastic processes: I. The regularity condition. *Acta Math*, 98:111–150, 1957

[11]

C

Vector Spaces

There are other spaces of elements where the geometry of the Euclidean space holds. A vector space is a set $\mathcal{V}$ of elements v, called vectors, on which two operations $+$ and $\cdot$, called vector addition and scalar multiplication, are defined such that for all vectors $u, v, w \in \mathcal{V}$ and scalars c, d it holds

1. Closure: $u + v \in \mathcal{V}$
2. Commutativity: $u + v = v + u$
3. Associativity: $(u + v) + w = u + (v + w)$
4. Additive identity: $\mathcal{V}$ contains an element, denoted by 0, such that $0 + v = v,\ \forall v \in \mathcal{V}$
5. Additive inverse: There exists a unique $x(v) \in \mathcal{V}$ such that $v + x(v) = 0$. $x(v)$ is called $-v$.
6. For any scalar c, $c \cdot v \in \mathcal{V}$
7. Distributivity: $c \cdot (u + v) = c \cdot u + c \cdot v$
8. Distributivity: $(c + d) \cdot v = c \cdot v + d \cdot v$
9. Associativity: $c \cdot (d \cdot v) = (cd) \cdot v$
10. Multiplicative identity: $1 \cdot v = v$

If the above hold when the scalars belong to the field of reals, $\mathcal{V}$ is said to be a real vector space, and when $c \in \mathbb{C}$, $\mathcal{V}$ is a complex vector space. We will consider complex vector spaces.

C.1 Inner Product Spaces

What gives the Euclidean space its geometry is the scalar product $\circ$. For two vectors x and y we have that the angle between the two vectors can be determined from

$$\cos(\alpha) = \frac{x \circ y}{\|x\| \, \|y\|} \tag{C.1}$$

The corresponding operator in a complex vector space is the inner product $\langle \cdot, \cdot \rangle : \ \mathcal{V} \times \mathcal{V} \to \mathbb{C}$, which, mimicking the properties of the scalar product, has to satisfy the following axioms for all $u, v, w \in \mathcal{V}$ and $\lambda \in \mathbb{C}$:

1. $\langle u + v, w\rangle = \langle u, w\rangle + \langle v, w\rangle$
2. $\langle \lambda u, v\rangle = \lambda\langle u, v\rangle$
3. $\langle u, v\rangle = \langle v, u\rangle^*$
4. $\langle v, v\rangle \geq 0$ with equality iff $v = 0$

A vector space endowed with an inner product is called an inner product space. For such spaces we can introduce a topology via the norm

$$\|v\| := \sqrt{\langle v, v\rangle}$$

It is easy to verify that this is a norm, i.e. that $\|\cdot\|$ satisfies i) $\|v\| \geq 0$, $\forall v \in \mathcal{V}$ with equality iff $v = 0$, ii) $\|\lambda v\| = |\lambda|\|v\|$, and iii) $\|u + v\| \leq \|u\| + \|v\|$.

It can be noted, although we will not make use of this, that a normed vector space $\mathcal{V}$ is an inner product space iff the parallelogram law

$$2\|u\|^2 + 2\|v\|^2 = \|u + v\|^2 + \|u - v\|^2$$

holds for all $u, v \in \mathcal{V}$. If this law holds then the inner product is given by

$$\langle x, y\rangle = \frac{1}{4}\left(\|u + v\|^2 - \|u - v\|^2\right)$$

The geometry in an inner product space becomes clear if we for a vector $u \in \mathcal{V}$ define its orthogonal projection on another vector $v \in \mathcal{V}$ as $u_{\|v} := \alpha v$ where α satisfies the normal equation

$$\langle u - \alpha v, v\rangle = 0 \tag{C.2}$$

i.e.

$$u_{\|v} = \frac{\langle u, v\rangle}{\langle v, v\rangle} v$$

Then

$$0 \leq \|u - u_{\|v}\|^2 = \|u\|^2 - \frac{|\langle u, v\rangle|^2}{\|v\|^2}$$

with equality iff $u = \lambda v$ for some $\lambda \in \mathbb{C}$, so that

$$0 \leq \frac{|\langle u, v\rangle|}{\|u\|\|v\|} \leq 1 \tag{C.3}$$

with the upper inequality true iff $u = \lambda v$ for some $\lambda \in \mathbb{C}$. As in (C.1) we can interpret the number in the middle above as the cosine of the angle between u and v. Since

$$\langle u - u_{\|v}, v\rangle = 0 \tag{C.4}$$

we say that $u - u_{\|v}$ is orthogonal to v (written $u - u_{\|v} \perp v$) and from this and the decomposition $u = (u - u_{\|v}) + u_{\|v}$ we obtain Pythagoras theorem

$$\|u\|^2 = \|u_{\|v}\|^2 + \|u - u_{\|v}\|^2 \tag{C.5}$$

The upper inequality in (C.3) is known as the Cauchy-Schwarz inequality. Using this, the geometric interpretation is completed by considering

$$\|u - \lambda v\|^2 = \|u - u_{\|v} + u_{\|v} - \lambda v\|^2 = \|u - u_{\|v}\|^2 + \|u_{\|v} - \lambda v\|^2 \geq \|u - u_{\|v}\|^2 \tag{C.6}$$

with equality only if $\lambda v = u_{\|v}$, which shows that $u_{\|v}$ is the vector in the direction of v that is closest to u, i.e. the notion of an orthogonal projection in an inner product space is consistent with the same notion in the Euclidean space.

C.2 *Subspaces and Orthogonal Projections*

A subspace $\mathcal{S}$ to a vector space $\mathcal{V}$ is a subset of $\mathcal{V}$ that is closed under addition and scalar multiplication, i.e. if $u, v \in \mathcal{S}$ then $\lambda_1 u + \lambda_2 v \in \mathcal{S}$ for any scalars λ_1 and λ_2.

Starting from a set of vectors $\{v_\alpha\}_{\alpha \in \mathcal{A}}\}$, we can generate a subspace by all finite linear combinations of these vectors. We denote such a subspace $\text{Span}\{\{v_\alpha\}_{\alpha \in \mathcal{A}}\}$.

A finite set of vectors $\{v_k\}_{k=1}^n$, $n < \infty$, is said to be linearly independent if the only solution to $\sum_{k=1}^n \alpha_k v_k = 0$ is $\alpha_1 = \ldots = \alpha_n = 0$. More generally, a set is said to be linearly independent if every finite collection of vectors from is linearly independent.

A vector space $\mathcal{V}$ is said to be finite dimensional if there is an $n < \infty$, the dimension of $\mathcal{V}$, $\dim[\mathcal{V}]$, such that $\mathcal{V}$ contains a linearly independent set of n vectors, whereas all sets of $n+1$ vectors are linearly dependent.

A basis for a vector space is a linearly independent set such that all vectors in the space can be uniquely represented as a finite linear combination of elements in the set, the basis elements. A basis exists for every vector space but it is not unique.

For an inner product space $\mathcal{V}$ with an n-dimensional subspace $\mathcal{S}$ having basis $\{v_1, \ldots, v_n\}$, we can for a vector u define its orthogonal projection on $\mathcal{S}$ as $u_{\|\mathcal{S}} := \sum_{k=1}^n \alpha_k v_k$ where, similarly to (C.2), the $\{\alpha_k\}$ are defined by the normal equations

$$\langle u - \sum_{k=1}^{n} \alpha_k v_k, v_l \rangle = 0, \quad l = 1, \ldots, n \tag{C.7}$$

which in matrix form becomes

$$\begin{bmatrix} \langle v_1, v_1 \rangle & \ldots & \langle v_1, v_n \rangle \\ \vdots & \vdots & \vdots \\ \langle v_n, v_1 \rangle & \ldots & \langle v_n, v_n \rangle \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} \langle u, v_1 \rangle \\ \vdots \\ \langle u, v_n \rangle \end{bmatrix} \tag{C.8}$$

If we form $v = \begin{bmatrix} v_1, \ldots, v_n \end{bmatrix}^T$ and $x = \begin{bmatrix} x_1, \ldots, x_m \end{bmatrix}^T$ define

$$\lfloor v, x^T \rfloor = \begin{bmatrix} \langle v_1, x_1 \rangle & \ldots & \langle v_1, x_m \rangle \\ \vdots & \vdots & \vdots \\ \langle v_n, x_1 \rangle & \ldots & \langle v_n, x_m \rangle \end{bmatrix}$$

we can write (C.8) in compact form as

$$\lfloor v, v^T \rfloor \alpha = \lfloor v, x \rfloor \tag{C.9}$$

where $\alpha = \begin{bmatrix} \alpha_1 & \dots & \alpha_n \end{bmatrix}^T$.

Considering an arbitrary point w in $\mathcal{S}$, we obtain the same inequality as in (C.6), i.e. $u_{\|\mathcal{S}}$ is the unique point closest to u in $\mathcal{S}$. We call this the orthogonal projection of u on $\mathcal{S}$. We can also define the orthogonal complement $\mathcal{S}^\perp$ of $\mathcal{S}$ to be the set of all vectors in $\mathcal{V}$ orthogonal to all vectors in $\mathcal{S}$, i.e.

$$\mathcal{S}^\perp = \{u \in \mathcal{V} : \langle u, v_k \rangle = 0, \ k = 1, \dots, n\}$$

Clearly $\mathcal{S}^\perp$ is a subspace. Furthermore, we can uniquely decompose any vector $u \in \mathcal{V}$ into $u = u_{\|\mathcal{S}} + u\mathcal{S}$, where $u_{\|\mathcal{S}}$ is the orthogonal projection on $\mathcal{S}$ and where $u\mathcal{S} := u - u_{\|\mathcal{S}}$. We write $\mathcal{V} = \mathcal{S} \oplus \mathcal{S}^\perp$.

C.3 Hilbert Spaces

It is easy to see that the limit point of convergent sequences in finite dimensional subspaces to normed spaces also belong to the subspace in question. Topologically, finite dimensional subspaces are always closed. A set $\mathcal{S}$ in a normed space is said to be open if there to every point $v \in \mathcal{S}$ exists a neighbourhood $\{u : \|u - v\| < \varepsilon\} \subset \mathcal{S}$, $\varepsilon > 0$. A set is closed if its complement is open. Matters become somewhat more complicated when considering subspaces of infinite dimensions.

Example C.1. *Let $\mathcal{V} = C[0,1]$, the space of continuous function on the interval $[0,1]$. Clearly this is a vector space under standard definitions of addition and scalar multiplication. We take the norm to be $\|v\| = \max_{0 \le x \le 1} |v(x)|$ (this is not an inner product space).*

Now, let $\mathcal{S}$ be the subspace to $\mathcal{V}$ consisting of all polynomials. Then the sequence of monomials $v_k : v_k(x) := x^k$, $k = 1, 2, \dots$ converges to the discontinuous function

$$v^*(x) := \begin{cases} 0, & 0 \le x < 1 \\ 1, & x = 1 \end{cases}$$

in the used norm, i.e. to a function not even belonging to $\mathcal{V}$, and even less to $\mathcal{S}$.

Example C.2. *Let the vector space $\mathcal{V}$ consist of the real numbers over the field of rationals, i.e. the scalars we use are rational, equipped with the inner product $\langle u, v \rangle = uv$. Now consider the subset $\mathcal{S}$ consisting of rational numbers. Clearly, $\mathcal{S}$ is a subspace over the field of rationals. This subspace is not closed as there are rational sequences that converge to irrational numbers (this is in fact a way to extend rational numbers to reals).*

When working with infinite dimensional subspaces we must therefore typically require that the subspace is closed as otherwise orthogonal projections in the spirit above may not even belong to the subspace onto which we project.

Apart from the geometrical properties discussed in the preceeding section, the Euclidean space possess another desirable property namely that convergence of a sequence $\{x_k\}$ is equivalent to that the sequence is a Cauchy sequence, i.e. for every $\varepsilon > 0$ there exists an N such that $\|x_k - x_l\| < \varepsilon$ when $k, l > N$. This holds also for subspaces to an Euclidean space. A metric space with the property that every Cauchy sequence converges to an element in the space is said to be complete. A Hilbert space is a complete inner product space. Perhaps not surprising, there is a strong connection between closedness and completeness: In a Hilbert space a subspace is closed iff it is complete. Furthermore, as in the Euclidean space, a finite dimensional subspace is complete.

Above we have seen that in a finite dimensional subspace $\mathcal{S}$ to an inner product space $\mathcal{V}$ there is a vector in the subspace that is closest to a given vector $u \in \mathcal{V}$, c.f. with the Euclidean space. In a Hilbert space this generalizes to infinite dimensional subspaces:

Theorem C.3.1. *Let $\mathcal{S}$ be a closed subspace to a Hilbert space $\mathcal{H}$ and let $u \in \mathcal{H}$ be given. Then there is a unique vector $v \in \mathcal{S}$ such that $u - v \perp w$ for all $w \in \mathcal{S}$. The vector v solves*

$$\min_{v \in \mathcal{S}} \|u - v\|$$

For Hilbert spaces we can thus talk about the orthogonal projection on $\mathcal{S}$ even when $\mathcal{S}$ is infinite dimensional and any vector $u \in \mathcal{V}$ can uniquely be split into $u = u_{\|\mathcal{S}} + u\mathcal{S}$ where $u_{\|\mathcal{S}} \in \mathcal{S}$ and where $u\mathcal{S} \in \mathcal{S}^{\perp}$, the orthogonal complement to $\mathcal{S}$ defined as $\mathcal{S}^{\perp} = \{v : v \perp \mathcal{S}\}$. There is dual formulation to the problem in Theorem C.3.1 which has $u_{\perp\mathcal{S}}$ as solution.

Corollary C.3.1. *Let $\mathcal{S}$ be a closed subspace to a Hilbert space $\mathcal{H}$ and let $u \in \mathcal{H}$ be given. Consider the linear variety*

$$\mathcal{L}_u = \{x = u + v,\ v \in \mathcal{S}\}$$

Then the problem

$$\min_{x \in \mathcal{L}_u} \|x\|$$

has a unique solution $u_{\perp\mathcal{S}}$. The solution is the unique $x \in \mathcal{L}_u$ satisfying

$$\langle x, v \rangle = 0 \quad \forall v \in \mathcal{S}$$

C.4 Orthonormal bases

A subset in an inner product space is said to be an orthonormal set if all vectors in have norm 1 and for any pair $u, v \in$, $\langle u, v \rangle = 0$ when $u \neq v$.

Given an orthonormal sequence $\{e_k\}_{k=1}^{\infty}$ in a Hilbert space $\mathcal{H}$, we can take $\mathcal{S}_k$ to be the span of $\{e_k\}_{k=1}^{n}$. Then the orthonormality gives

$$\|u_{\|\mathcal{S}_k}\|^2 = \sum_{k=1}^{n} |\langle u, e_k \rangle|^2$$

and from Pythagoras theorem (C.5) it follows that

$$\|u_{\|\mathcal{S}_k}\|^2 = \sum_{k=1}^{n} |\langle u, e_k \rangle|^2 \leq \|u\|^2, \quad k = 1, 2, \ldots$$

Since this holds for every finite k, Bessel's inequality

$$\sum_{k=1}^{\infty} |\langle u, e_k \rangle|^2 \leq \|u\|^2 \tag{C.10}$$

follows. A remarkable consequence of this inequality is that if one has an uncountable orthonormal set in an inner product space $\mathcal{V}$, then for a given $u \in \mathcal{V}$, at most a countable set of Fourier coefficients $\langle u, e \rangle$, $e \in$ can be non-zero. In this case we can thus still associate u to a series

$$\sum_{k=1}^{\infty} \langle u, e_k \rangle e_k$$

where $\{e_k\}$ is an enumeration of the elements in that have non-zero inner products with u.

Now, suppose that we have a series $\sum_{k=1}^{\infty} \alpha_k e_k$ in a Hilbert space $\mathcal{H}$. Then this series is convergent, to v say, iff $\sum_{k=1}^{\infty} |\alpha_k|^2$ is convergent. This follows since $\sum_{k=1}^{\infty} \alpha_k e_k$ then is a Cauchy sequence.

Furthermore, in this case $\alpha_k = \langle v, e_k \rangle$.

In particular it holds that for any $u \in \mathcal{H}$, $\sum_{k=1}^{\infty} \langle u, e_k \rangle e_k$ is convergent. However, the series may not correspond to u despite that, as per the preceeding paragraph, the Fourier coefficients $\langle u, e_k \rangle$ are the same as for u. Denoting the series by v, it is easy to see that $u - v$ is orthogonal to every e_k. Thus for $u - v$ to be non-zero must mean that the orthonormal sequence does not span the whole of $\mathcal{H}$. This is guaranteed by requiring the orthonormal sequence $\{e_k\}_{k=1}^{\infty}$ to be what is called a complete (or total) orthonormal basis in $\mathcal{H}$. This means that the span of the sequence is dense in $\mathcal{H}$, i.e. the closure of the span is $\mathcal{H}$ itself.

For a complete orthonormal basis equality holds in (C.10) which is then known as Parseval's relation which can be seen as a generalization of Pythagoras theorem. Conversely, if Parseval's relation holds for every $u \in \mathcal{H}$, then the orthonormal set is complete.

So for which Hilbert spaces does there exist a countable complete orthogonal set? It turns out that the space must be separable which means that it has a countable subset which is dense and for such spaces every orthonormal set is countable.

D
Probability Theory

D.1 Transformation of random variables

Lemma D.1.1 (Transformation of random variables)**.** *Suppose that the random vector* $\mathbf{x} \in \mathbf{X} \subseteq \mathbb{R}^n$ *has pdf* $p_x(\mathbf{x})$*. Let* $f : \mathbf{X} \to \mathbf{Y} \in \mathbb{R}^n$ *be injective and continuously differentiable. Then* $\mathbf{y} = f(\mathbf{x}) \in \mathbf{Y}$ *has pdf* p_y *defined by*

$$p_y(f(\mathbf{x})) = \frac{p_x(\mathbf{x})}{|\det f'(\mathbf{x})|} \tag{D.1}$$

Proof: Let $\mathbf{A} \in \mathbf{X}$. Then using the change of variables formula

$$\begin{aligned}
&\int_{\mathbf{A}} p_x(\mathbf{x})d\mathbf{x} = \mathbf{P}(\mathbf{x} \in \mathbf{A}) = \mathbf{P}(\mathbf{y} \in f(\mathbf{A})) = \int_{f(\mathbf{A})} p_y(\mathbf{y})d\mathbf{y} \\
&= \left\{ \begin{array}{c} \mathbf{y} = f(\mathbf{x}) \\ d\mathbf{y} = |\det f'(\mathbf{x})|d\mathbf{x} \end{array} \right\} \\
&= \int_{\mathbf{A}} p_y(f(\mathbf{x}))|\det f'(\mathbf{x})|d\mathbf{x}
\end{aligned}$$

Since this holds for any measurable $\mathbf{A}$, comparing the first integral with the last one, (D.1) must hold. □

D.2 Limits of random variables

A fundamental result that is useful for establishing convergence w.p.1 is the Borel-Cantelli lemma. Let $\{A_n\}$ be events in a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and consider the set

$$\limsup_n A_n = \cap_{n=1}^{\infty} \cup_{m=n}^{\infty} A_m$$

This event contains the outcomes that belong to the events A_n an infinitely number of times and is also denoted A_n infinitely often, A_n i.o.

Lemma D.2.1 (Theorem 4.2.1 in [1]. The Borel-Cantelli lemma)**.** *Let* $\{A_n\}$ *be events in a probability space* $(\Omega, \mathcal{F}, \boldsymbol{P})$ *for which*

$$\sum_{n=1}^{\infty} \boldsymbol{P}(A_n) < \infty$$

Then

$$\boldsymbol{P}(A_n \ i.o.) = 0$$

Corollary D.2.1. *If the events $\{A_n\}$ are independent, then*

$$\sum_{n=1}^{\infty} \boldsymbol{P}(A_n) = \infty \quad \Rightarrow \quad \boldsymbol{P}(A_n \textit{ i.o. }) = 1$$

Corollary D.2.2 (Proposition 6.4 in [2]). *Let $\{X(n)\}$ be a sequence of random variables. Suppose that for every $\varepsilon > 0$*

$$\sum_{n=1}^{\infty} \sup_{m \geq n} \boldsymbol{P}(|X(n) - X(m)| \geq \varepsilon) < \infty$$

then $\{X(n)\}$ converges a.e.

The Chebyshev inequality is often used to show that the summability condition in the Borel-Cantelli lemma is satisfied.

Lemma D.2.2 (p.48 [3]. Chebyshev's inequality). *Let φ be a symmetric, strictly positive and increasing function on $(0, \infty)$ and let X be a random variable such that $\mathbb{E}\left[\varphi(X)\right)]$. Then*

$$\boldsymbol{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}\left[\varphi(X)\right]}{\varphi(u)}$$

D.3 The normal distribution

Definition D.3.1. *A random vector $\mathbf{X} \in \mathbb{R}^n$ with pdf*

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) := \frac{1}{\sqrt{\det(2\pi)\boldsymbol{\Sigma}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \ \boldsymbol{\Sigma} > 0$$

is said to be normal distributed and has mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

D.3.1 Singular normal distributions

When $\boldsymbol{\Sigma}$ is singular with rank $m < n$, $\mathbf{X} - \boldsymbol{\mu}$ is restricted to the range space of Σ. With the factorization $\boldsymbol{\Sigma} = \mathbf{K}\mathbf{K}^T$ where $\mathbf{K} \in \mathbb{R}^{n \times m}$, $\mathbf{X}$ can be given the following description

$$\mathbf{X} = \mathbf{K}\mathbf{Z} + \boldsymbol{\mu}, \quad \text{where } \mathbf{Z} \in \mathcal{N}(0, \mathbf{I})$$

Thus the outcomes of $\mathbf{X} - \boldsymbol{\mu}$ belong to the m-dimensional subspace $\mathbb{R}^n_{\mathbf{K}} := \{\mathbf{x} = \mathbf{K}\mathbf{z},\ \mathbf{z} \in \mathbb{R}^m\}$ of $\mathbb{R}^n$. Then $\mathbf{X}$ does not have pdf in an ordinary sense. In order to be able to write

$$\mathbf{P}(\mathbf{X} \in \mathbf{B}) = \int_{\mathbf{B}} p_x(\mathbf{x}) d\mathbf{x}$$

where $\mathbf{B} - \boldsymbol{\mu} \in \mathbb{R}^n_{\mathbf{K}}$, we need to interpret $d\mathbf{x}$ as a volume measure over $\mathbb{R}^n_{\mathbf{K}}$ so that

$$\int_{\mathbb{R}^n_{\mathbf{K}}} p_x(\mathbf{x}) d\mathbf{x} = 1$$

With this interpretation, the scaling of volume by the transformation $\mathbf{x} - \boldsymbol{\mu} = \mathbf{K}\mathbf{z}$ is given by $d\mathbf{x} = \sqrt{\det \mathbf{K}^T\mathbf{K}}\, d\mathbf{z}$ [4]. With $\mathbf{B} = \{\mathbf{x} = \mathbf{K}\mathbf{z} + \boldsymbol{\mu},\ \mathbf{z} \in$

[4] S. Krantz and H. Parks. *Geometric Integration Theory.* Birkhäuser Boston, Boston, 2008. Chapter 5

A}

$$\begin{aligned}
&\int_{\mathbf{A}} p_z(\mathbf{z})d\mathbf{z} = \mathsf{P}(\mathbf{z} \in \mathbf{A}) = \mathsf{P}(\mathbf{x} \in \mathbf{B}) = \int_{\mathbf{B}} p_x(\mathbf{x})d\mathbf{x} \\
&= \left\{ \begin{array}{c} \mathbf{x} = \mathbf{K}\mathbf{z} + \boldsymbol{\mu} \\ d\mathbf{x} = \sqrt{\det \mathbf{K}^T\mathbf{K}} d\mathbf{z} \end{array} \right\} \\
&= \int_{\mathbf{A}} p_x(\mathbf{K}\mathbf{z} + \boldsymbol{\mu})\sqrt{\det \mathbf{K}^T\mathbf{K}}\, d\mathbf{z}
\end{aligned}$$

Since this holds for any measurable **A**, we can formally write

$$p_x(\mathbf{K}\mathbf{z} + \boldsymbol{\mu}) = \frac{1}{\sqrt{\det \mathbf{K}^T\mathbf{K}}} p_z(\mathbf{z}) = \frac{1}{\sqrt{\det(2\pi)\mathbf{K}^T\mathbf{K}}} e^{-\frac{1}{2}\mathbf{z}^T\mathbf{z}}$$

which we can express in terms of **x** through the relation $\mathbf{z} = (\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T(\mathbf{x} - \boldsymbol{\mu})$

$$p_x(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi)\mathbf{K}^T\mathbf{K}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{K}(\mathbf{K}^T\mathbf{K})^{-2}\mathbf{K}^T(\mathbf{x}-\boldsymbol{\mu})} \tag{D.2}$$

We here recognize that $\mathbf{K}(\mathbf{K}^T\mathbf{K})^{-2}\mathbf{K}^T$ is the Moore-Penrose pseudo-inverse, denoted $\boldsymbol{\Sigma}^+$, of $\boldsymbol{\Sigma} = \mathbf{K}\mathbf{K}^T$, and that $\det \mathbf{K}^T\mathbf{K}$ is the pseudo-determinant of $\boldsymbol{\Sigma}$, denoted $\det^* \boldsymbol{\Sigma}$. We can thus write

$$p_x(\mathbf{x}) = \frac{1}{\sqrt{\det^*(2\pi)\boldsymbol{\Sigma}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^+(\mathbf{x}-\boldsymbol{\mu})}$$

D.3.2 The expectation of a normal pdf

Lemma D.3.1. *Let $\mathcal{N}(y; m, P)$ denote the pdf of a normal distribution with mean $m \in \mathbb{R}^n$ and covariance P. Let $X \sim \mathcal{N}(m, P_x)$, $X \in \mathbb{R}^m$. Then*

$$\mathbb{E}\left[\mathcal{N}(y; AX, P_y)\right] = \mathcal{N}(y; Am, P_y + AP_xA^T) \tag{D.3}$$

Proof. Completing the square using Lemma (A.5.1) gives

$$\mathbb{E}\left[N(y; AX, P_y)\right] \tag{D.4}$$

$$\begin{aligned}
&= \int \frac{e^{-\frac{1}{2}(y-Ax)^TP_y^{-1}(y-Ax)}}{(2\pi)^{n/2}\sqrt{\det P_y}} \frac{e^{-\frac{1}{2}(x-m)^TP_x^{-1}(x-m)}}{(2\pi)^{m/2}\sqrt{\det P_x}} dx \\
&= \int \frac{e^{-\frac{1}{2}\left((y-Ax)^TP_y^{-1}(y-Ax)+(x-m)^TP_x^{-1}(x-m)\right)}}{(2\pi)^{(n+m)/2}\sqrt{\det P_yP_x}} dx \\
&= \int \frac{e^{-\frac{1}{2}\left((y-Am)^TT^{-1}(y-Am)+(x-\hat{x})^TS^{-1}(x-\hat{x})\right)}}{(2\pi)^{(n+m)/2}\sqrt{\det P_yP_x}} dx
\end{aligned} \tag{D.5}$$

where T, S and $\hat{x}$ are defined in Lemma (A.5.1). We can further write this as

$$\begin{aligned}
&\mathbb{E}\left[N(y; AX, P_y)\right] \\
&= \frac{e^{-\frac{1}{2}(y-Am)^TT^{-1}(y-Am)}}{(2\pi)^{(n)/2}\sqrt{\det P_yP_xS}} \int N(x; \hat{x}, S)dx \\
&= \frac{e^{-\frac{1}{2}(y-Am)^TT^{-1}(y-Am)}}{(2\pi)^{(n)/2}\sqrt{\det P_yP_xS^{-1}}}
\end{aligned} \tag{D.6}$$

but

$$\begin{aligned}
&\det P_y P_x S^{-1} = \det P_y P_x (P_x^{-1} + A^T P_y^{-1} A)^{-1} = \det P_y (I + A^T P_y^{-1} A P_x) \\
&= \det P_y \det(I + A^T P_y^{-1} A P_x) = \det P_y \det(I + A P_x A^T P_y^{-1}) \\
&= \det(P_y(I + A P_x A^T P_y^{-1}) = \det(P_y + A P_x A^T) = \det T \qquad \text{(D.7)}
\end{aligned}$$

Inserting this in (D.6) gives the result. □

D.4 Stein's Identity

Lemma D.4.1. *Let* **Z** *be distributed according to the canonical exponential family* (5.11)

$$p(\mathbf{z};\boldsymbol{\theta}) = e^{\boldsymbol{\theta}^T T(\mathbf{z}) - A(\boldsymbol{\theta})} h(\mathbf{z}) \qquad \text{(D.8)}$$

and let g be any differentiable function such that $\mathbb{E}\left[|g'(\mathbf{Z})|\right] < \infty$. Then

$$\mathbb{E}\left[g'(\mathbf{Z})\right] = -\mathbb{E}\left[\left(\frac{d}{d}\log h(\mathbf{Z}) + T'(\mathbf{Z})^T\boldsymbol{\theta}\right) g^T(\mathbf{Z})\right] \qquad \text{(D.9)}$$

Bibliography

[1] B. D. O. Anderson and J.B. Moore. *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, New Jersey, 1979.

[2] K. J. Åström and B. Wittenmark. *Computer-Controlled Systems*. Prentice Hall, Englewood Cliffs, N.J., 3rd edition, 1997.

[3] C. Berg, J.P.R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer-Verlag, New York, 1984.

[4] A.N. Borodin. *Stochastic Processes*. Birkhäuser, 2017.

[5] P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods*. Springer, 1991.

[6] P.E. Caines. *Linear stochastic systems*. SIAM, 2018.

[7] P.E. Caines and L. Gerencsér. A simple proof for a spectral factorization theorem. *IMA Journal of Mathematical Control & Information*, 8:39–44, 1991.

[8] L. Carleson. On convergence and growth of partial sums of Fourier series. *Acta Math*, 116, 1966.

[9] K.L. Chung. *A Course in Probability Theory*. Academic Press, Orlando, Florida 32887, 1974.

[10] R.V. Churchill and J.W. Brown. *Complex Variables and Applications*. McGraw-Hill.

[11] H. Cramér. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, 1946.

[12] C.A. Desoer and M. Vidyasagar. *Feedback Systems: Input-Output Properties*. Academic Press, New York, 1975.

[13] M. Galrinho. *System Identification with Multi-Step Least-Squares Methods*. Doctoral thesis, KTH, Stockholm, Sweden, 2018.

[14] L.L. Gihman and A.V. Skorohod. *The Theory of Stochastic Processes I*. Springer-Verlag, Berlin, 1974.

[15] R.A. Hunt. On the convergence of Fourier series, orthogonal expansions and their continuous analogues. In *Proc. Conf., Edwardsville, Ill., 1967*, Southern Illinois Univ. Press, pages 235–255, Carbondale, Ill., 1968.

[16] W. James and C. Stein. Estimation of quadratic loss. In *Proc. 4th Berkeley Symp. Math. Statist. Prob.*, volume 1, pages 361–379, Berkely, CA, USA, 1961.

[K)] H. K˙ *Eigenvalue Distribution of Compact Operators*.

[18] S.M. Kay. *Fundamentals of Statistical Signal Processing. Estimation Theory*. Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

[19] A. N. Kolmogorov. Une série de Fourier-Lebesque divergente presque partout. *Fund. Math.*, 4, 1923.

[20] S. Krantz and H. Parks. *Geometric Integration Theory*. Birkhäuser Boston, Boston, 2008. Chapter 5.

[21] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. John Wiley & Sons, New York, second edition edition, 1998.

[22] L. Ljung. *System identification, Theory for the user*. System sciences series. Prentice Hall, Upper Saddle River, NJ, USA, second edition, 1999.

[23] J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society, London*, 209:415–446, 1909.

[24] W. Philipp and W. Stout. Almost sure invariance principles for partial sums of weakly dependent random variables. *Memoirs of the Am. Math. Soc.*, 2(161), 1975.

[25] A. Rantzer. On the Kalman-Yakubovich-Popov lemma. *Systems & Control Letters*, 28:7–10, 1996.

[26] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, London, 1976.

[27] W. Rudin. *Real and Complex Analysis*. McGraw-Hill, London, 1986.

[28] A.N. Shiryaev. *Probability*. Springer, 2nd edition, 1989.

[29] T. Söderström. *Discrete-Time Stochastic Systems. Estimation and control*. Prentice-Hall International, New York, 1994.

[30] T. Söderström and P. Stoica. *System identification*. Prentice Hall, 1989.

[31] C.M. Stein. Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9(6):1135–1151, 1981.

[32] P. Stoica and T. Söderström. On reparameterization of loss functions used in estimation and the invariance principle. *Signal Processing*, 17(4):383–387, 1989.

[33] N. Wiener and P. Masani. The prediction theory of multivariate stochastic processes: I. The regularity condition. *Acta Math*, 98: 111–150, 1957.

[34] J.C. Willems. Least squares stationary optimal control and the algebraic Riccati equation. *IEEE Trans. Aut. Control*, 16:621–634, 1971.

[35] E. Wong and B. Hajek. *Stochastic Processes in Engineering Systems*. Springer-Verlag, New York, 1985.

[36] K. Zhou, J. C. Doyle, and K. Glover. *Robust and Optimal Control*. Prentice-Hall, 1996.