SCUOLA NORMALE SUPERIORE DI PISA

CLASSE DI SCIENZE

CORSO DI PERFEZIONAMENTO IN
MATEMATICA PER LE TECNOLOGIE INDUSTRIALI

ANNO ACCADEMICO 2010-2011

TESI DI PERFEZIONAMENTO

Ph.D. Thesis

Deconvolution of Quantized-Input Linear Systems: An Information-Theoretic Approach

Sophie Marie Fosson

Advisor
Prof. Fabio Fagnani

September 19, 2011

Acknowledgements

I wish to thank my advisor Prof. Fabio Fagnani for his advisement and support during my Ph.D. research.

Many thanks also to Paolo Tilli for his inspiring suggestions and to all the people in Pisa and Torino who contributed to my work.

Finally, I wish to express my gratitude to my family.

I dedicate this dissertation to my mother.

Abstract

The deconvolution problem has been drawing the attention of mathematicians, physicists and engineers since the early sixties.

Ubiquitous in the applications, it consists in recovering the unknown input of a convolution system from noisy measurements of the output. It is a typical instance of inverse, ill-posed problem: the existence and uniqueness of the solution are not assured and even small perturbations in the data may cause large deviations in the solution. In the last fifty years, a large amount of estimation techniques have been proposed by different research communities to tackle deconvolution, each technique being related to a peculiar engineering application or mathematical set. In many occurrences, the unknown input presents some known features, which can be exploited to develop ad hoc algorithms. For example, prior information about regularity and smoothness of the input function are often considered, as well as the knowledge of a probabilistic distribution on the input source: the estimation techniques arising in different scenarios are strongly diverse.

Less effort has been dedicated to the case where the input is known to be affected by discontinuities and switches, which is becoming an important issue in modern technologies. In fact, *quantized* signals, that is, piecewise constant functions that can assume only a finite number of values, are nowadays widespread in the applications, given the ongoing process of digitization concerning most of information and communication systems. Moreover, *hybrid* systems are often encountered, which are characterized by the introduction of quantized signals into physical, analog communication channels.

Motivated by such consideration, this dissertation is devoted to the study of the deconvolution of continuous systems with quantized input; in particular, our attention will be focused on linear systems. Given the discrete nature of the input, we will show that the whole problem can be interpreted as a paradigmatic digital transmission problem and we will undertake an Information-theoretic approach to tackle it.

The aim of this dissertation is to develop suitable deconvolution algorithms for quantized-input linear systems, which will be derived from known decoding procedures, and to test them in different scenarios. Much consideration will be given to the theoretical analysis of these algorithms, whose performance will be rigorously described in mathematical terms.

Contents

Nomenclature

Acronyms and Abbreviations

APP	A Posteriori Probability
AWGN	Additive White Gaussian Noise
BCJR	Optimal Decoding Algorithm by Bahl, Cocke, Jelinek and Raviv
BER	Bit Error Rate
CBCJR	Causal BCJR
CBER	Conditional Bit Error Rate
EM	Expectation Maximization
EMP	Extended Markov Process
FTC	Fault Tolerant Control
i.p.m.	invariant probability measure
IRF	Iterated Random Functions
KF	Kalman Filter
LLMSE	Linear Least Mean Squares Estimate
LMSE	Least Mean Squares Estimate
MAP	Maximum A Posteriori
ML	Maximum Likelihood
MSE	Mean Square Error
OSA	One State Algorithm

p.m.	probability measure
r.v.	random variable
SNR	Signal-to-Noise Ratio
TSA	Two States Algorithm

Symbols

$\mathbb{1}_A$	indicator function: given a set A, $\mathbb{1}_A(x) = 1$ if $x \in A$, $\mathbb{1}_A(x) = 0$ otherwise
$\mathcal{B}(\Omega)$	Borel σ-algebra of Ω
$C_b(\Omega)$	space of the bounded continuous functions on Ω
$\mathbb{E}$	mean operator
erfc	complementary error function: $\text{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^{+\infty} e^{-s} \text{d}s$ for any $x \in \mathbb{R}$
$f_{(\cdot)}$	probability density function of continuous or hybrid (that is, involving both continuous and discrete events) random variables
I	identity operator
m-Lip(Ω)	set of functions $f : \Omega \to \mathbb{R}$ which are Lipschitz of constant m
P	probability on discrete random variables
$\mathbf{P}$	transition probability matrix of a Markov Chain
$P(\cdot, \cdot)$	transition probability kernel of a Markov Process

Other Notation

Capital letters will be used to name random variables

Boldface capital letters will be used to name vectors of random variables

$\mathbf{v}_a^b = (v_a, v_{a+1}, \dots, v_b)$ for any $\mathbf{v} \in \mathbb{R}^N$, $a, b, N \in \mathbb{N}$, $a \leq b \leq N$

Chapter 1

Introduction

The deconvolution problem is concerned with recovering the unknown input of a convolution system given measurements of the output.

Such an issue is ubiquitous in many scientific and technological domains, ranging from image restoration, astrophysics and geophysics to biological and biomedical systems and control of industrial processes. In these fields, the reconstruction of an object given an *image* of it, namely an indirect or inaccurate observation, is often undertaken and due to the physical impossibility of directly accessing the desired information. If the relation between object and image, i.e., between input and output, can be modeled as a convolution integral, then the problem is called *deconvolution.* This is the case, for instance, of linear differential systems.

Motivated by the many applications, the analysis of the deconvolution problem presents also interesting mathematical challenges for its nature of inverse, ill-posed problem, which have been attracting the attention of mathematicians since the early sixties. This dual theoretical and practical rationale has stimulated interdisciplinary research on this topic; significant, wide-ranging contributions have been made in the least thirty years and various resolutive techniques have been proposed. However, no universal solution is possible for deconvolution in its most general formulation: the practical aim and the mathematical structure (namely, the kind of system, its dimensions, its parameters) for each particular instance may lead to very different resolutive scenarios.

In that sense, deconvolution is destined to remain an open issue.

The mathematical formulation of the deconvolution problem (which will be mostly discussed in Chapter 2) is generally the following: given a convolution system

$$x(t) = \int_0^t \mathcal{K}(t-s)u(s)\mathrm{d}s \quad t \in [0,T] \tag{1.1}$$

where T is a (possibly infinite) time horizon, the aim is to recover the unknown input function $u(t)$ given measurements of the output $x(t)$ and the knowledge of the

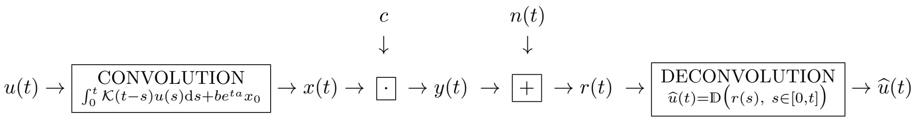
c
$n(t)$
$u(t)$
CONVOLUTION
$\int_0^t \mathcal{K}(t-s)u(s)\mathrm{d}s+be^{ta}x_0$
$x(t)$
$\cdot$
$y(t)$
$+$
$r(t)$
DECONVOLUTION
$\widehat{u}(t)=\mathbb{D}\big(r(s),\ s\in[0,t]\big)$
$\widehat{u}(t)$

Figure 1.1: Deconvolution aims to recover the unknown input $u(t)$ of a system given the noisy output $r(t)$. If this operation is not required to performed on-line, the deconvolution algorithm $\mathbb{D}$ may process the whole function $r(t)$, $t \in [0, T]$, to provide its estimation $\widehat{u}(t)$.

convolution kernel $\mathcal{K}(t)$ [1]. Typically, measurements are impaired by inaccuracies or incompleteness, which makes the issue more complicated.

1.1 Thesis contributions

An important example of convolution is provided by time-invariant, input/output linear systems:

$$\begin{cases} x'(t) = \mathrm{A}x(t) + \mathrm{B}u(t) \qquad t \in [0, T] \\ y(t) = \mathrm{C}x(t) \\ x(0) = x_0 \end{cases} \tag{1.2}$$

where $u : \mathbb{R} \to \mathbb{R}^p$, $x : \mathbb{R} \to \mathbb{R}^q$, $y : \mathbb{R} \to \mathbb{R}^r$ and A, B and C are consistent matrices, vectors or scalar constant values. The input/output relation is given by:

$$y(t) = \mathrm{C}\int_0^t e^{(t-s)\mathrm{A}}\mathrm{B}u(s)\mathrm{d}s + \mathrm{C}e^{t\mathrm{A}}\mathrm{B}x_0 \tag{1.3}$$

which is a convolution integral with kernel $\mathcal{K}(t) = e^{t\mathrm{A}}$.

Let us suppose the output $y(t)$ to be affected by an additive noise $n(t)$ so that

$$r(t) = y(t) + n(t)$$

is the measured function; thus, deconvolution aims at evaluating the unknown input $u(t)$ by processing the available data $r(t)$. The final result is an estimation $\widehat{u}(t)$ (see Figure 1.1).

The deconvolution of time-invariant, input/output linear systems is the main focus of this dissertation. A large amount of research has been devoted to this subject since the 1960s and classical results are generally concerned with systems whose input functions satisfy some conditions of regularity or smoothness. Moreover, most of deconvolution algorithms work *off-line*, that is, in the hypothesis that t represents the time, deconvolution is performed at the end of the transmission, say after the final time T (clearly, here T is supposed to be finite); in this case, a deconvolution algorithm $\mathbb{D}$ can

[1] Notice that this is a causal version of convolution, that is, the output x at time t depends only on the present and on the past, which will be the case considered in this work. More in general, the integration extremes might vary all over the real line

use the whole function $r(t)$, $t \in [0,T]$ to recover $u(t)$, but this cannot be implemented when a response about $u(t)$ is needed in real-time or when T is not finite.

Our contribution, instead, is to analyze

- quantized-input systems: the unknown object is a stepwise constant function assuming values in a finite set range;
- *on-line* deconvolution: $\widehat{u}(t)$ is provided exactly at time t (or after a fixed delay); this forces the algorithm to be *causal*, i.e., it processes only past and present information: $\widehat{u}(t) = \mathbb{D}(r(s),\ s \in [0,t])$.

The first point is motivated by modern applications, in which signals digitization is increasingly being adopted, and by the need of processing signals with abrupt changes. For example, the input of our system may be the output of a digital device or a signal carrying information about the status of an industrial process. In the simplest case, such status may be only “on” or “off”, so that the corresponding signal is *binary*, that is, may assume only two different values.

More in general, the binary case is relevant since it models a lot of real-world applications, from digital communications, where information is often encoded into sequences of bits, to the fault control of industrial, mechanical and transport processes, which basically aims to state whether a device works or not. On the other hand, it is the reference paradigm of all the quantized signals under the theoretical point of view. For that motivation, the analysis proposed in this dissertation is completed addressed to the binary case.

Concerning the second point, as already said, the dynamical model (1.1) is causal, that is the convolution on u at time t is performed on its present and past values. This suggests the idea that deconvolution can be done on-line: the input is estimated time after time, without waiting for final time T. More precisely, at any time t, a causal algorithm processes $r(s)$, $s \in [0.t]$, and provides $\widehat{u}(t)$ (with a possible delay).

Causality is not taken into account by classical deconvolution techniques, which generally work off-line and do not envisage a dynamical state representation of the problem. Nevertheless, in the applications, convolution systems are often associated with processes evolving in time and, in this framework, on-line deconvolution turns out to be necessary whenever a quick response is required or when the time horizon T is large. For instance, let us consider those industrial processes or communication systems that in principle could never stop working: for them, deconvolution (which could implemented, for example, for control purpose) cannot be performed “at the end”.

Classical deconvolution algorithms, which are efficient for smooth input functions and generally work off-line, are not fit for the causal, quantized framework. New methodologies have then to be developed, taking into account the prior information about the discrete nature of the unknown quantity and the call of on-line resolution.

The aim of this thesis is to tackle this challenge. Our basic intuition is that Information and Coding/Decoding Theory can provide the suitable tools to achieve the goal, since it naturally deals with digital transmission and recovery of discrete input

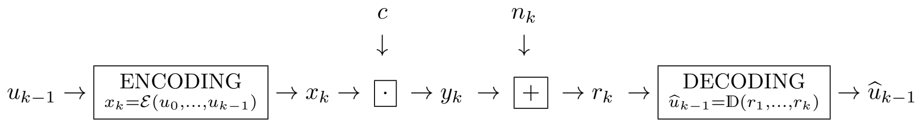
c
n_k
u_{k-1}
ENCODING
$x_k=\mathcal{E}(u_0,\dots,u_{k-1})$
x_k
$\cdot$
y_k
$+$
r_k
DECODING
$\widehat{u}_{k-1}=\mathbb{D}(r_1,\dots,r_k)$
$\widehat{u}_{k-1}$

Figure 1.2: Causal digital transmission paradigm at generic time $k \in \mathbb{N}$

messages, arising from a known input's alphabet, say a (finite) set of symbols. Moreover, even if classical decoding algorithms, such as the BCJR [10], work off-line, many on-line alternatives can be easily developed.

It is not our purpose to review here the fundamentals of Information Theory and Coding/Decoding techniques: the reader which is not familiar with basic digital communication can retrieve them, e.g., in [17],[126]. Here, we just recall the pattern of a typical digital transmission: a discrete sequence of symbols, said *input message*, has to be transmitted to a receiver, throughout a *channel*, say the physical medium that carries the message and which is generally affected by some noise. In order to make the communication reliable, before the transmission the input message is *encoded*, that is, some redundancy is added on it in order to preserve the information it carries even in case of undesired alterations or partial erasure in the channel passage. Finally, the receiver has to *decode* the message, i.e., he/it attempts to reconstruct the transmitted information on the basis of the received noisy message and given the knowledge of the used encoding rule. The Figure 1.1 represents a causal version of this model, in which the encoding function $\mathcal{E}$ at time k acts on the past and present input symbols $u_0, \dots . u_{k-1}$ to produce the symbol x_k; moreover, the estimation of u_{k-1} is performed on the basis of the actual lecture r_k and possibly using also the previous lectures $r_1, \dots, r_{k-1}$ when information storage is possible.

Coming back to deconvolution, our key idea is to treat the deconvolution of quantized-input linear systems as a decoding problem.

More in detail, the (causal) model represented in Figure 1.1 can be converted in the discrete one shown in Figure 1.1 whenever $u(t)$ is determined by a sequence of symbols $(u_0, u_1, u_2, \dots)$. In this case, our deconvolution system corresponds to a generic digital communication model in the sense that the following aspects are equivalent:

- quantized input $\Leftrightarrow$ digital message input;
- convolution of the input $\Leftrightarrow$ encoding of the input;
- noisy measurements $\Leftrightarrow$ transmission over a noisy channel;
- deconvolution $\Leftrightarrow$ decoding.

Just one difference has to be remarked in this scheme: while encoding is generally introduced to improve the transmission reliability, convolution is just an operation on the input imposed by the physics of the system and it may even worsen the communication. In other terms, we state that convolution and encoding are equivalent as they both perform some operation on the input, with no further considerations about their motivations and consequences.

Since quantized deconvolution techniques are not available, while many efficient decoding algorithms have been developed in the last fifty years, our suggestion is to use the latter in order to perform deconvolution. This will be the main task in this dissertation, which involves the adaptation of the classical decoding methodologies to our purpose.

We will derive and test different algorithms; in order to evaluate their performance, we will provide both simulations and rigorous theoretical results.

1.2 Summary of the thesis

1.2.1 Chapter 2

Chapter 2 is devoted to a general introduction on the deconvolution problem.

First, we illustrate some applicative examples arising from both classical and hybrid linear systems' literature, the second framework being closer to our purpose.

Afterwards, we propose an overview on the family of inverse problems, of which deconvolution constitutes a paradigmatic branch. We introduce the notion of inversion and the problematics related to it (such as ill-posedness and ill-conditioning) and we briefly describe the main resolutive methodologies, with particular attention to the distinction between deterministic and probabilistic approaches.

1.2.2 Chapter 3

In Chapter 3, we introduce the subject of the dissertation, namely, the deconvolution of linear systems with quantized input, in a general framework.

We describe the reference model and we recast it in an Information-theoretic perspective. The underlying idea is that decoding techniques, which are naturally developed for digital transmissions, are more suitable than classical deconvolution algorithms to work with quantized signals.

Afterwards, we give a suitable probabilistic setting and we define the performance goal in term of a minimization of a mean square cost. Finally, we introduce the deconvolution-decoding algorithms we intend to use, discussing their origins and main features.

1.2.3 Chapter 4

Chapter 4 is devoted to the differentiation problem, i.e., to the deconvolution of the system (1.2) in one dimension and with parameters $a = 0$, $b = 1$ and $c = 1$. This is the basic instance, however it presents the basic difficulties of the problem.

We test the previously introduced decoding algorithms to this case and we analyze their performance using the theories of Markov Processes and Markov Process in Random Environments. Particular attention is drawn to the asymptotic case, i.e., when time tends to infinity. The main result is an ergodic theorem that theoretically assesses the performance of the algorithms in terms of the mean square cost.

This chapter is partially based on the papers:

- F. Fagnani, S. M. Fosson, "An information theoretic approach to hybrid deconvolution problems", in Proceedings of 17th IFAC World Congress (Seoul, Korea), pp. 10112–10117, July 6-11, 2008.
- F. Fagnani, S. M. Fosson, "Deconvolution of linear systems with quantized input: a coding theoretic viewpoint", submitted to *Mathematics of Control, Signals, and Systems*, 2009, available at http://arxiv.org/abs/1001.3550.

1.2.4 Chapter 5

In Chapter 5 we present a generalization of the differentiation problem to one-dimensional, input/output linear systems. The approach is the same presented in Chapter 4, but the mathematical background is different since the state function assumes values

- in $\mathbb{N}$ in the differentiation case;
- in a not denumerable set (which can be in turn a compact interval or a Cantor set) in the generic case.

This implies, for example, that not all the algorithms presented in Chapter 4 can be efficiently implemented for generic linear systems among the algorithms, as complexity problems arise in the not denumerable framework. Moreover, extending the theoretical analysis proposed in Chapter 4 implies the study of Markov Processes in continuous state spaces; in particular, when such space is of Cantor space no standard argumentation can be used to study the asymptotic behavior of the related process. *Iterated Random Functions* will be then introduced for this purpose.

Finally, in this chapter we propose an analytical comparison between our algorithm and a Kalman Filter based technique. In particular, we exploit again the Iterated Random Functions to analyze the Kalman Filter and we evaluate performance and complexity of both methods.

This chapter is partially based on the papers:

- S. M. Fosson, P. Tilli, "Deconvolution of quantized-input linear systems: Analysis via Markov Processes of a low-complexity algorithm", in Proceedings of International Symposium MTNS 2010 (Budapest, Hungary), pp. 59–66, July 5-9, 2010.
- S. M. Fosson, "Analysis of a Deconvolution Algorithm for Quantized-Input Linear Systems through Iterated Random Functions", in Proceedings of 18th IFAC World Congress (Milano, Italy), pp. 11302–11307, Aug. 29 - Sept. 2, 2011.

1.2.5 Chapter 6

In Chapter 6, a Fault Tolerant Control application is studied. The system is multidimensional and the goal is to detect faults or failures in a process. In this model, we introduce also an active feedback with the aim of adjusting the system whenever a fault is detected: this takes a different viewpoint on the problem. Since the system is linear, the detection task actually is a deconvolution problem. We consider a particular example arising from flight control literature and we propose some optimal criteria for the design of a Fault Tolerant Control system.

This chapter is partially based on the paper:

- S. M. Fosson, “A Decoding Approach to Fault Tolerant Control of Linear Systems with Quantized Disturbance Input”, submitted to *International Journal of Control*, 2010, available at http://arxiv.org/abs/1011.2989.

Chapter 2

Overview on Deconvolution and Inverse Problems

The deconvolution problem consists in recovering the unknown input of a convolution system from the observation of the output.

Let us consider a (possibly infinite) time horizon T and a signal $u(t)$, $t \in [0, T]$, convolved with a convolution kernel $\mathcal{K}(t)$, producing $x(t)$:

$$x(t) = \int_0^t \mathcal{K}(t-s)u(s)\mathrm{d}s \quad t \in [0, T]. \tag{2.1}$$

The functions $\mathcal{K}$ and u are assumed to be so that the above integral makes sense. Deconvolution aims to determine the input $u(t)$ under the hypothesis that $\mathcal{K}(t)$ is known and that $x(t)$ can be observed (but with possible inaccuracies and/or sampling).

Such an issue occurs in various scientific, technological and industrial applications since convolution systems arise, e.g., in signal and image processing, astronomy, geophysics, seismology, and biomedical engineering. The pervasiveness of the problem has been strongly motivating the research since the early sixties and a large amount of literature has been produced by diverse scientific communities.

Since the early studies, deconvolution has been attracting the interest of pure and applied mathematicians given its nature of *inverse problem*. In general, an inverse problem consists in determining an unknown, not directly measurable quantity by observing its response to a probing signal; in other terms, one aims to recover an unknown quantity from the analysis of data connected with it by some physical laws. If the laws are known, one could reasonably expect to find the solution just by inverting them, but this is not the case whenever inaccuracies (even if small) affect the data. In fact, inverse problems typically are *ill-posed*, i.e., small errors in the data may cause large errors in the solution (see Section 2.3.1 for the exact definition given by Hadamard). Moreover, if the problem is discretized, namely it consists in a matrix linear system or is reduced to it for computational purpose, *ill-conditioning* may occur, which is again an error amplification phenomenon, but due in this case to numerical issues (see Section 2.3.3).

In particular, a problem that originally is well-posed may become ill-conditioned after discretization.

Ill-posedness and ill-conditioning underlie the main mathematical difficulty in inverse problems (and in particular in deconvolution), because they generally prevent the exact reconstruction of the solution. This is why such problems are typically faced with *estimation* techniques: one tries to approximate the solution on the basis of given measurements and minimizing a suitable distance functional between estimated and right solutions. Naturally, there is no universally convenient estimation technique, given the extent of the inverse problems' family. For instance, let us consider deconvolution: in order to establish a *good* estimation of the input we have to take account of the dimensions of the problem, the kind of convolution kernel and also the applicative goal of the system, that is, which is the fundamental information we need to know about the input. For example, in a control system the goal might be to determine if the input function overpasses a given threshold and the error functional should reflect this fact. Moreover, in many situations, the input function represents a physical signal and is known to have some properties, e.g., of boundedness or smoothness, which narrows the range of likely solutions: this must be considered in the development of an ad hoc estimation algorithm.

The techniques used for different models are very peculiar and this is why the literature about deconvolution is so wide and includes many disparate approaches.

This dissertation deals with the deconvolution of input/output linear systems with quantized input: the goal is to estimate the input by observing a noisy, sampled version of the output, under the hypothesis that the input is a step function that assumes only a finite number of constant values.

Before introducing our case study, we provide a general introduction to the deconvolution problem and we collect some bibliographical references. Given the extent of the problem, a comprehensive survey is not feasible. Here, we first present a few classical case studies and applications from inverse and deconvolution literature (Section 2.1) and then some specific examples, belonging to the family of the quantized input linear systems, that motivate our work (Section 2.2).

Afterwards, Section 2.3 will be devoted to a general discussion about inverse problems, aimed to illustrate the mathematical difficulties that deconvolution presents. In Sections 2.4 and 2.5 the most important approaches to deconvolution and inverse problems will be shown and finally some references are collected by topic in Section 2.6.

2.1 A few classical applications

In this section, a few instances of inverse and deconvolution problems are surveyed, coming from *classical* literature, which was oriented to continuous systems with inputs characterized by some regularity condition.

2.1.1 Image Processing

An image can be defined as a signal “carrying information about a physical object which is not directly observable”[18, Chapter 3]. In other terms, an image is a degraded representation of an object, where degradation is basically due to *blurring* in the image formation, that is, to disturbances such as relative motion between the camera and the object being captured, diffraction, aberration, atmospheric perturbations, and to *noise*, e.g., measurement inaccuracy, which is intrinsic to the image detection process. In this context, we call *image deconvolution* a post-processing of the detection of an image, aimed to reduce its degradation.

From a mathematical viewpoint, an image can be described by a function with domain in $\mathbb{R}^n$ where $n = 2$ or $n = 3$ respectively for the two-dimensional and three-dimensional cases. Let $f(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n$, be the intensity at $\mathbf{x}$ of an object and $g(\mathbf{x})$ the image produced by an optical instrument. In most situations, the imaging system can be approximated by a linear operator A, that is $g = Af$: moreover, in the case of space invariance, A is a convolution operator, say

$$(Af)(\mathbf{x}) = \int_{\mathbb{R}^n} h(\mathbf{x} - \mathbf{y}) f(\mathbf{y}) \mathrm{d}\mathbf{y} \tag{2.2}$$

where $h(\mathbf{x})$ is the so-called *Point Spread Function* (PSF) and models the blurring. Finally the detected image is the function

$$j(\mathbf{x}) = g(\mathbf{x}) + n(\mathbf{x}) \tag{2.3}$$

where $n(\mathbf{x})$ is an additive noise. The image deconvolution consists of estimating $f(\mathbf{x})$ given the noisy, blurred image $j(\mathbf{x})$.

This generic model can be retrieved in different fields, such as astronomy, microscopy and medical imaging.

2.1.2 Astronomy

Image deconvolution is fundamental in astronomy, where the reconstruction of scientific content from observations is widely faced. Typically, astronomical photograph is strongly affected by the large distance between camera and object and to the conditions in which images are recorded, such as turbulence and exposure time.

A leading example is the image deconvolution for telescopes. The mathematical model is basically the one given by (2.2, 2.3), the parameters of which depend on the characteristics of the telescope and the atmospheric conditions. It is important to notice that in the deconvolution problem, the PSF, say, the convolution kernel, is supposed to be known, but in the case of telescopes this is not true. The PSF can however be deduced from experimental data: for example, the telescope may be pointed to a reference object to reconstruct it; in many cases, this task is achieved with good precision, but inaccuracies in the PSF cannot be totally avoided.

Applications of deconvolution to telescopes have been studied for the Hubble Space Telescope (HST, see [140], http://hubblesite.org/) and more recently for the Large Binocular Telescope (LBT, see [19], http://medusa.as.arizona.edu/lbto/).

The deconvolution problem in case of unknown (or partially known) kernel, not discussed in this dissertation, has been widely studied in the last years and is referred to as *blind* (or *myopic*) deconvolution (see, e.g., [141]).

2.1.3 Seismology

Reflection seismology concerns the exploration of the Earth's internal structure by analyzing the reflectivity of subsurface layers and has been widely used in the last fifty years for mining and petroleum industrial purposes. Generally, this kind of exploration is performed recording the effects of an impulsive source, such as an explosion. The recorded data, say the seismic trace, is the result of the superposition of seismic wavelet replica (produced by the source) reflected from the interfaces of subsurface layers. More precisely, the model, early proposed in [128] is the following

$$z(k) = \sum_{i=1}^{k} \mu(i)w(k-i) + n(k) \tag{2.4}$$

where μ is the reflectivity function, w is the wavelet and n is an observational noise: this is an example of discrete convolution system. The goal is to estimate the reflectivity from the seismic trace z.

We remark that seismology is an early application of deconvolution, the first works being dated in the last fifties [128]. Later significant works are [7, 83, 117]. More recently, in [30] the problems of microseismic deconvolution have been discussed.

2.1.4 Tomographic Reconstruction

Deconvolution is involved in many biological and medical systems. In particular in the field of *medical imaging*, which studies how to create images of parts of the human body, the problem of recovering information from an image (for instance a radiography, a tomography, an electrocardiography) often arises. Actually, medical imaging is oriented to the *direct problem*, that is, to the procedure and the instrumentation to create image, while deconvolution is connected to the *inverse problem* of determining the biological quantity of interest (e.g., the activity of the heart, of the brain) just by observing the image (see Section 2.3 for a definition of direct and inverse problems).

In the last decades, *tomography*, that is, the study of three-dimensional objects through two-dimensional cross-sections or slices, has been widely studied and formalized from a mathematical viewpoint. The mathematical description of tomography can be formulated in terms of deconvolution, even if the relation between input and output is actually given by a line integral. Moreover, tomography is connected with projections and the model that one obtains is a Radon transform.

Different tomographic techniques have been developed. The first one dates 1971; introduced by Hounsfield, it is referred to as X-ray or computed tomography; it is based on the penetration of a certain part of the human body with X-rays from different directions, which can provide information about anatomical aspects.

A different tomographic technique is instead based on the injection or inhalation of radionuclide-labelled agents said radiopharmaceuticals; their distribution in the human body depends on physiological activities, such as blood flow and metabolism and can be detected by the γ-rays produced by the decay of radionuclides. Such technique, named emission computed tomography (ECT), is useful to infer the functions of the biological tissues of the organs and may be applied with different modalities (fro example, PET and SPECT, see [18, Section 8.3]).

Notice that tomography is used not only in the medical diagnosis, but also in many other fields, e.g., manufacturing industry, production quality control and security (for example, in the luggage check).

Computed tomography can be described by the following mathematical model. The object of interest is the linear attenuation function $f(\mathbf{x})$ which assesses the density of the body at point $\mathbf{x}$ and also represents the X-rays absorption by some tissue. X-rays are then used to determine f: a source generates a pencil beam of X-rays which propagates through the human body along a straight line L up to a detector. Let us call $I(\mathbf{x})$ the intensity of the beam at point $\mathbf{x}$ of L: then, the loss of intensity of the corresponding element dl of L is given by $\mathrm{d}I() = -f(\mathbf{x})\mathrm{d}l$ which is equivalent to

$$\log\left(\frac{I}{I_0}\right) = -\int_L f(\mathbf{x})\mathrm{d}l \tag{2.5}$$

where I_0 and I respectively are the intensity of the beam outgoing from the source and the intensity measured by the detector. Emissions and measurements are repeated by moving the source and the detector simultaneously and in the same direction, say θ_1; afterwards, the projection angle is changed and again source and detector are moved in the same, new direction θ_2. By repeating the procedure for a sufficient number of different angles, one obtains all the sufficient projections of $f(\mathbf{x})$. In particular, the projection of $f(\mathbf{x})$ along a direction θ is given by

$$Pf(\theta, s) = \int_L f(\mathbf{x})\mathrm{d}l = \int_{\mathbb{R}} f(s\theta + t\theta^{\perp})\mathrm{d}t. \tag{2.6}$$

X-ray tomographic reconstruction consists in estimating $f(\mathbf{x})$ from the projections $Pf(\theta, s)$ (which are usually noisy), where θ may assume different values according to the projections that are required to recover the image.

Notice that this is not properly a deconvolution problem, since the integral in (2.6) is not a convolution operator, but a projection operator known as *Radon transform*: however, it has been presented here since the similarity to deconvolution is evident.

For more details on tomography, see [52], [18, Section 8.2] and [36, Chapter 12].

2.1.5 Biomedical Engineering

Deconvolution is important not only in medical imaging, but also in other biomedical analysis. For example, it is involved in the study of the insulin secretion rate (ISR, [137, 116, 115]), which is a quantity used to assess the glucose regulation in humans. A concrete application of its study is then the diagnosis and therapy of diabetes.

ISR is not directly measurable *in vivo*. Insulin is secreted by the pancreas, usually in response to raised plasma glucose concentration, and then reaches the plasma; in the passage, a certain, not known percentage of it is extracted by the liver and this makes impossible to assess it. However, an equivalent amount of C-peptide (CP) is co-secreted with insulin and is not extracted by liver before reaching plasma. The ISR can then be retrieved by CP and in fact injection of glucose and measurement of CP is a common procedure (named intravenous glucose tolerance test, IVGTT) used to estimate ISR.

In particular, the mathematical law that connects ISR and CP is a convolution integral of kind:

$$\mathrm{CP}(t) = \int_{-\infty}^{t} \mathcal{K}(t-s)\mathrm{ISR}(s)\mathrm{d}s \tag{2.7}$$

where kernel $\mathcal{K}$ is the impulse response to the system [116] and CP and ISR are expressed as functions of time,

Hence the reconstruction of ISR is a deconvolution problem.

2.2 Applications of linear hybrid systems

Input/output linear systems have been representing a conspicuous class of deconvolution systems since the early studies, as many natural and technological processes are linear or can be linearized, that is, suitably approximated by linear equations. Consider

$$\begin{cases} x'(t) = \mathrm{A}x(t) + \mathrm{B}u(t) \qquad t \in [0,T] \\ y(t) = \mathrm{C}x(t) \\ x(0) = 0 \end{cases} \tag{2.8}$$

where $u(t) \in \mathbb{R}^m$, $x(t) \in \mathbb{R}^n$ and $y(t) \in \mathbb{R}^p$ respectively are the input, the state function and the output; $\mathrm{A} \in \mathbb{R}^{n\times n}$, $\mathrm{B} \in \mathbb{R}^{n\times m}$ and $\mathrm{C} \in \mathbb{R}^{p\times n}$ are constant matrices; T is a possibly infinite time horizon. Recovering the input $u(t)$ from $y(t)$ is a deconvolution operation, in fact

$$y(t) = \mathrm{C}\int_0^t e^{(t-s)\mathrm{A}}\mathrm{B}u(s)\mathrm{d}s. \tag{2.9}$$

Moreover, in many situations the time is discrete or discretized, for example for numerical implementation purposes. Fixed a suitable discretization step $\tau > 0$, let

$$u(t) = \sum_{k\in\mathbb{N}} u_k \mathbb{1}_{[k\tau,(k+1)\tau[}(t), \quad u_k \in \mathbb{R}^m \tag{2.10}$$

that is, the input is a piecewise constant function determined by a sequence of vectors of $\mathbb{R}^m$. In such case, the dynamical equation (2.9) becomes

$$y(k\tau) = \mathrm{A}^{-1}(e^{\tau\mathrm{A}} - \mathbb{I})e^{(k-1)\tau\mathrm{A}}\sum_{h=0}^{k-1}\mathrm{B}u_h e^{-h\tau\mathrm{A}}$$

such that we can write the following recursion (see Chapter 5 for details):

$$x_k = \mathrm{Q}x_{k-1} + \mathrm{W}u_{k-1}$$

where $x_k = x(k\tau)$, $\mathrm{Q} = e^{\tau\mathrm{A}}$ and $\mathrm{W} = \mathrm{A}^{-1}(e^{\tau\mathrm{A}} - \mathbb{I})$. A subcase of (2.10) is the quantized input case, namely the input can assume values in a finite set $\mathcal{U}$:

$$u(t) = \sum_{k\in\mathbb{N}} u_k \mathbb{1}_{[k\tau,(k+1)\tau[}(t), \quad u_k \in \mathcal{U}. \tag{2.11}$$

This will be the main subject of this dissertation, the interest of which arises from the spreading of digital technologies in many scientific and engineering fields. A digital device typically produces or processes a quantized signal; in our context, we may assume the input to be generated by a digital device and then to be naturally quantized or to be a continuous function quantized for successive processing purpose.

Notice that we consider the input to be quantized, while the linear system (2.9) is analogical, that is, continuous. This is a common occurrence in many modern technologies, where the integration of digital and analogical components is required. Such systems are generally known as *hybrid* systems are have been largely studied in the last decades.

In concrete applications, noise typically affects the measurements. The model should then envisage an observational noise n_k, so that the available output actually is the sequence

$$r_k = \mathrm{C}x_k + n_k, \; k \in \mathbb{N}.$$

In some cases (but not in this dissertation) also a *process noise* m_k is considered, so that $x_k = \mathrm{Q}x_{k-1} + \mathrm{W}u_{k-1} + m_k$.

The deconvolution (2.9) with (2.11) in general cannot be accomplished by classical algorithms, which typically require some regularity on the solution. Here, in fact, the input is not continuous and present abrupt changes.

It is well-known that in the presence of white noises n_k and m_k (by white noise we mean a zero-mean random process whose autocorrelation matrix is a multiple of the identity matrix), the states x_k of (2.9) with (2.10) can be efficiently estimated by the Kalman Filter, which provides the best estimate among the linear estimates in the sense of minimization of the mean square error (see Section 2.5.8). In particular, if the noises are Gaussian, the Kalman Filter is the best estimate, not only among the linear ones.

The Kalman Filter is largely adopted to study systems excited by discrete, abruptly changing signals. In particular, the optimal estimates of the states provided by Kalman Filter can then be used to estimate the u_k's (we will discuss this point in Chapter 5), but optimality on the input estimation is not given, which motivates us to study other approaches.

In order to motivate our work, we now present some important applications where the deconvolution of linear systems with quantized input is required.

2.2.1 Fault Detection

Almost all technological, industrial and transport processes are affected by faults. Let us consider an airplane: its flight is a process that involves the functioning of a large number of physical devices, which are used to accomplish the various tasks required to safe travel. However, such devices may undergo breakdowns and this may cause serious problems in the flight management. A trivial example of that should be the malfunctioning of the sensor that measures the distance between aircraft and earth: if its measurements are wrong and if the pilot does not detect the presence of an error, very dangerous consequences may be provoked.

It is now clear the importance of *fault detection*, that is, the capability of revealing abnormal behaviors of a device. In many situations, process models are dynamic linear systems of kind (2.8) in which the input envisages a fault component; more precisely, in the input are present the contributions of a known control function $f(t)$ and an unknown function representing the status (faulty or not faulty) of the process $z(t)$. The combination of $f(t)$ and $z(t)$ in the input depends on the context (see for instance [71, Chapter 5] for a discussion on this issue); for example we could have a product:

$$\begin{cases} x'(t) = \mathrm{A}x(t) + \mathrm{B}z(t)f(t) & t \in [0,T] \\ y(t) = \mathrm{C}x(t) \\ x(0) = x_0 \end{cases} \tag{2.12}$$

Assuming $z(t) \in \mathbb{R}$, $z(t) = 1$ may represent the fault-free status, the system being completely driven by $f(t)$, while $z(t) \in (0,1)$ may reflect a diminution of the effectiveness of the system, the effect of $f(t)$ being attenuated (see, e.g., [47]).

Fault detection in systems of kind (2.12) corresponds to the estimation of $z(t)$ given a noisy version of $y(t)$: in other terms, it is a deconvolution problem.

In many cases, it is not a classical deconvolution problem in the sense that $z(t)$ is affected by abrupt changes whenever the fault manifests suddenly. Moreover, quantization of $z(t)$ is often assumed, for instance if the fault occurs on a digital device. In the simplest case, a switch between two levels can be considered: let us imagine, for instance, that the device of interest normally assumes a constant value, while in case of fault, it switches off. Furthermore, in other cases, quantization can be assumed for numerical purposes. In both frameworks, we reduce to the hybrid setting described at the beginning of this section.

This fault detection issue will be studied in Chapter 6, with the purpose of designing a fault tolerant control system for a flight instance: we will discuss how to integrate a suitable fault detection with the introduction of a compensation input, with the aim of attenuating the bad effect of the fault.

2.2.2 Maneuvering Targets Tracking

The tracking problem is a military application of deconvolution of linear systems with quantized input. Suppose that one aims to track a manned maneuverable vehicle, for example an aircraft: the dynamics is given by a linear system whose state may represent the position and the velocity of the target vehicle, while the input may represent the value of a pilot-induced maneuver, e.g., an acceleration. The input estimation is then required for tracking.

Typically, the unknown maneuver is modeled as a stepwise constant function assuming values in a finite set. In particular, early results on tracking maneuvering targets consider just two levels: 0 and $u \neq 0$, respectively representing the no-maneuver and maneuver states [134]. The goal is then to detect if a maneuver has been done and in such case to estimate its unknown constant value u. More complex instances are studied by the so-called *generalized input estimation* [88] which assumes the maneuver to be a linear combination of known basic time functions.

When the basic time functions are stepwise constant, we obtain tracking models of kind (see [25]):

$$\begin{cases} x_{k+1} = \mathrm{Q}x_k + \mathrm{W}u_k + m_k \quad u_k \in \mathcal{U} \\ y_k = \mathrm{C}x_k + n_k \end{cases}$$

with where $x_k \in \mathbb{R}^n$ and $y_k \in \mathbb{R}^p$, $\mathrm{A} \in \mathbb{R}^{n\times n}$, $\mathrm{B} \in \mathbb{R}^{n\times m}$ and $\mathrm{C} \in \mathbb{R}^{p\times n}$ and $\mathcal{U}$ is a known finite set. Given such formulation, the problem can be faced, e.g., using a bank of $|\mathcal{U}|$ Kalman Filters, see [90, 25].

For more details, a long survey about maneuvering targets tracking has been recently written and divided into different papers (see http://ece.engr.uno.edu/isl/MTTSurveys.htm); [89] is the first part of this work.

2.2.3 Quantized Control Systems

The problem of stabilizing a time-invariant linear system with a finite number of control values is very actual. In many contexts, in fact, quantization is necessary for example to send information on communication channels with particular physical or technological constraints. Quantized Control Systems (QCS for short) are then all those linear systems on which control is performed through a quantized control input.

The basic model considered in Quantized Control is the following:

$$\begin{cases} x_{k+1} = \mathrm{Q}x_k + \mathrm{W}u_k \quad u_k \in \mathcal{U} \\ y_k = \mathrm{C}x_k \end{cases}$$

where $\mathcal{U}$ is a finite set and typically u is a feedback control obtained by quantization of some function of the output.

The recent literature on QCS faces problems such as the derivation of a quantization protocol that allows to maintain stability [44] and stabilization and reachability properties of quantized systems [22, 114, 50]; in these frameworks, the control input is known. Suppose instead, that one cannot access the control process, but needs to recover the control input sequence, just looking at the output of the system and knowing which quantization has been performed on the input: this is an example of deconvolution problem of linear systems with quantized input.

2.2.4 Mode estimation for switching systems

Nowadays, many industrial and control procedures can be described in terms of linear systems switching among different modes, that is, given a finite set of different time-invariant dynamical linear systems, in turn the process follows one of them, the switch among systems being governed by a logical device. In other terms, at any switch corresponds a change in the parameters of the system.

This is a typical example of hybrid problem, where the basic dynamics is continuous, but the decision on which is the “current dynamics ”is due to some digital element. The whole process is then affected by discontinuities and abrupt changes. A certain amount of work has been produced on this subject in the last years, the interest on it being motivated by the increasing introduction of digital checks and commands in continuous processes, in particular in control applications; see, e.g, [26, 39, 43, 23, 9], and the references therein for an overview on the problem.

Switching systems can be written as linear systems depending on a switching parameter. Slightly different models have been proposed in literature. For example, in [9], the model is the following:

$$\begin{cases} x_{k+1} = \mathrm{Q}(\lambda_{\mathrm{k}})x_k + \mathrm{W}(\lambda_{\mathrm{k}})u_k + m_k \\ y_k = \mathrm{C}(\lambda_k)x_k + n_k \end{cases}$$

where u_k, x_k, y_k, m_k and n_k are real vectors (m_k and n_k respectively represent a process noise and a measurement noise), while $\lambda_k \in \{1, 2, \dots, L\}$ is the *discrete state* or *mode of the system*, that is the parameter that decides the parameters of the system.

Consider the following simple one-dimensional case

$$\begin{cases} x_{k+1} = \mathrm{q}x_k + \mathrm{w}(\lambda_k)u_k \\ y_k = cx_k + n_k \end{cases}$$

with $\lambda_k \in \{0, 1\}$ and $\mathrm{w}(\lambda_k) = \lambda_k$ and suppose that the control input u_k is constant and equal to 1: the mode estimation in such case can be interpreted as a deconvolution problem; in particular, this is the model we will study in Chapter 5.

An important class of switching systems is given by *jump linear systems* [23] in which the switch is ruled by some Markov process. In such cases, we may face a deconvolution problem with stochastic input, which will be a main focus in this dissertation. Another possible instance is a linear system in which the noise is random, for example Gaussian, but its distribution parameters change in time according to Markov law [4].

Methods for mode estimation in switching systems typically arise from control and fault detection areas and in many cases an active approach in exploited, that is, the system is excited by a suitable control input; for what concern jump systems, probabilistic methods are used, such as the Bayes Methods, Maximum a Posteriori estimation and Kalman Filters.

2.3 Inverse problems

In 1976, Joseph B. Keller [81] called "two problems *inverses* of one another if the formulation of each involves all or part of the solution of the other. Often, for historical reasons, one of the two problems has been studied extensively for some time, while the other is new and not so well understood. In such cases, the former is called the *direct problem*, while the latter is called the *inverse problem*". This is probably the most general and frequently quoted definition, however slightly different interpretations are common. For example, from a physics-oriented viewpoint, one may define the direct problem as the one that follows the "natural direction" [36, Introduction] of cause-effect; in this setting, a direct mechanical problem is to compute the trajectory of an object given the forces acting on it, while the inverse problem is to retrieve the forces by observing the trajectory. Similarly, let us consider the heat equation with boundary and initial conditions: deducing the temperature at a given final time is a direct problem, while measuring the final temperature distribution and trying to determine it at earlier times is an inverse problem. An other example arises from scattering theory: the direct problem is the computation of the scattered waves from the knowledge of the source and obstacles, while the determination of the obstacles from source and waves is the inverse perspective. Many other physical examples can be found in the literature, see, e.g., [82, Section 1.1] and the survey paper [78].

By Keller's definition, some information about the solution of the direct problem is necessary to face an inverse problem. Clearly, a perfect knowledge of that solution would be the ideal condition, but in the practice this is never achieved because of measurement. The starting point for an inverse study is indeed the observation of the experimental data obtained by the direct process, on the basis of which, along with prior information on the process itself, one tries to recover the input. Observations are never completely reliable: usually, systematic errors occur due to the measurement instrumentation; moreover, instrumentation may capture only some data (for instance, samples at fixed time instants). In other terms, two aspects must be taken into consideration: first, the quantity that one aims to study is not directly observable, but can be inferred only through the observation of an *image* produced by the direct process; second, the measurements do not provide a perfect knowledge of the image, since some inaccuracies cannot be avoided. Inverse problems usually are affected by both aspects, but even issues characterized by of just one of them (either direct inaccurate measurement or indirect accurate measurement) may be incorporated in the family of inverse problems.

In mathematical terms, the most general formulation of an inverse problem is the following: given two spaces $\mathcal{X}$ and $\mathcal{Y}$ (the nature of those spaces will be discussed later) and an operator $\Phi : \mathcal{X} \to \mathcal{Y}$, let

$$\Phi(x) = y \quad x \in \mathcal{X}, y \in \mathcal{Y}. \tag{2.13}$$

The direct problem is to compute y given x and the *direct* operator Φ; the inverse problem, instead, consists in recovering x given a measurement $r = r(y) \in \mathcal{Y}$ of the direct solution y and by *inverting* the operator Φ. More precisely, the aim is to construct an operator $\Psi : \mathcal{Y} \to \mathcal{X}$ such that

$$\Psi(r) = x. \tag{2.14}$$

The problem is that, in general, the operator Ψ cannot be constructed for many motivations, which we depict through a few toy examples.

Example 1 (Sum) *Let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \mathbb{R}$ and let Φ be the linear transformation that computes the sum: $\Phi(x) = x_1 + x_2$ where $x = (x_1, x_2)$. We can imagine the operator Φ as a system which takes two numbers as input and computes their sum as output. In this case, it is clear that even if we exactly measure the output y we cannot retrieve the input $x = (x_1, x_2)$, in the sense that infinite values of x are possible. In other terms, the operator Φ is not invertible.*

Example 2 (Product-Division) *Let $\mathcal{X} = \mathcal{Y} = C_b(\mathbb{R})$ be the space of all the continuous and bounded functions on $\mathbb{R}$ and $\phi(x) = \varepsilon x$ where $\varepsilon > 0$ is very small. The inversion is clearly possible: if we know $y = \Phi(x)$, $x = \Psi(y) = \frac{y}{\varepsilon}$. Nevertheless, let us suppose that a small constant error affects the measurements, say $r = y + \sqrt{\varepsilon}$: then, $\Psi(r) = x + \frac{1}{\sqrt{\varepsilon}} >> x$. We conclude that inversion is possible, but* unstable *with respect to measurement errors, say small errors in the data may cause large errors in the inverse problem solution.*

Example 3 (Integration-Differentiation) *Another unstable example is provided by the couple integration-differentiation. Consider the integration as direct operation: let $\mathcal{X} = \mathcal{Y} = C^1[0,1]$ and $\Phi(x) = \int_0^t x(s)\mathrm{d}s$. Integration is known to be a* smoothing *operation, which inevitably causes a "loss of information" about the object to which it is applied. In fact, if $\int_0^t x(s)\mathrm{d}s = y(t)$, then $x(t) = y'(t)$, but if the measurement is affected by an oscillatory noise of small amplitude and high frequency, say $r(t) = y(t) + \varepsilon \sin(\frac{1}{\varepsilon^2}t)$, the inversion produces the solution $\Psi(r) = r'(t) = x(t) + \frac{1}{\varepsilon}\cos(\frac{1}{\varepsilon^2}t)$ which largely oscillates around the correct solution.*

Example 4 (Laplace equation) *Another classical example of instability is provided by Laplace equation ([65]):*

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{2.15}$$

with the following Cauchy conditions:

$$u(x,0) = \frac{1}{n}\cos(nx), \qquad \frac{\partial u}{\partial y}(x,0) = 0. \tag{2.16}$$

Given those data, our aim is to compute $u(x,y)$. The (unique) solution of (2.15-2.16) is

$$u(x,y) = \frac{1}{n}\cos(nx)\cosh(ny). \tag{2.17}$$

Now, if $n \to \infty$, $u(x,0) \to 0$ while for $y \neq 0$, $u(x,y) \to \infty$. In other terms, a large n produces a tiny oscillation in the data at $y = 0$, but causes a huge oscillation at any finite distance from $y = 0$.

Example 5 (Fredholm integral equation of the first kind) *Let us consider the Fredholm integral equation of the first kind [21, 78]*

$$y(s) = (Lx)(s) = \int_a^b \mathcal{K}(s,r)x(r)\mathrm{d}r \qquad x(r) \in C[a,b], y(s) \in C[c,d] \tag{2.18}$$

(notice that if $\mathcal{K}(s,r) = \mathcal{K}(s-r)$, (2.18) *is a convolution integral) with kernel K continuous and differentiable with respect to both s and r. Such problem is ill-posed because solutions may not exist for some functions: for example, if $y(s)$ is continuous but not differentiable on $[c,d]$, then the equation cannot have a continuous solution $x(r)$.*

A complete list of well-known ill-posed inverse differential problems can be retrieved in [78] and [82].

These examples have highlighted some typical inversion and stability issues, which are common among the inverse problems. A formalization of these concepts is provided by the Hadamard definition of ill-posedness, which is now recalled.

2.3.1 Ill-posedness

In 1902, Hadamard stated the conditions for a mathematical problem to be *well-posed* [64]:

1. *Existence* of a solution: for each item of data y in a given topological space $\mathcal{Y}$, there exists a solution x in a given topological space $\mathcal{X}$.

2. *Uniqueness* of the solution: the solution is unique in $\mathcal{X}$.

3. *Continuity* with respect to the data: x depends on y with continuity.

A problem is *ill-posed* when at least one of these three conditions is not fulfilled.

For what concerns inverse problems, it is easy to see that the Existence and Uniqueness conditions, respectively, are equivalent to surjectivity and injectivity of the direct operator. Continuity, instead, guarantees in particular that is y is affected by a small error δy, then also the error δx induced on the solution x is small. Notice also that the Continuity condition depends on the topology of the considered spaces; in the next, we will mainly consider the cases of Hilbert spaces.

Ill-posed problems are very frequent in mathematics and engineering. Moreover, the most of inverse problems are ill-posed, even if the corresponding direct problems are well-posed. Let us see some examples.

Example 6 (Continuation of Example 1) *The direct operator Φ is not injective. Therefore, the Uniqueness condition is not fulfilled by the inverse problem, which turns out to be ill-posed.*

Example 7 (Continuation of Examples 2-3) *For both examples, we have that $||r-y||_\infty \leq \varepsilon$, while the distance between $\Psi(x(t))$ and $\Psi(r(t))$ may be of order $\frac{1}{\varepsilon}$, at least for some $t \in \mathbb{R}$. In conclusion, continuity does not hold.*

For simplicity of exposition, from now onwards we will assume that

$$\begin{aligned} &\mathcal{X},\ \mathcal{Y} \text{ are Hilbert spaces} \\ &L : \mathcal{X} \to \mathcal{Y} \text{ is a linear, continuous operator.} \end{aligned} \tag{2.19}$$

In the setting (2.19), let us try to invert L, namely to solve

$$Lx = y. \tag{2.20}$$

Let us study its well-posedness, that is, the existence of a continuous operator

$$M : \mathcal{Y} \to \mathcal{X} \ \text{ s.t } \ M \circ L = I \tag{2.21}$$

where I is the identity operator. The continuity of M should guarantee that small perturbations on y produce only small perturbations on the solution $My = x$. First, the existence of the solution x is assured only if L is surjective; second, L must be injective, otherwise $Lx = y$ could have more than one solution. Moreover, continuity is assured only if $\mathrm{Im}(L)$ is closed, as a consequence of the Open Mapping Theorem (see, e.g., [130]).

This shows that many difficulties may arise in direct inversion due to ill-posedness. In order to tackle this issue, the most classical methods are *generalized inversion* and *regularization* methods, which are exposed in the next.

2.3.2 Least Squares and Generalized Inverses

In the previous section, we have highlighted the difficulties arising from ill-posed problems, which in general cannot be solved by a direct mathematical operation; even when it is possible to find out a solution, this one may be completely affected by noise, so not physically reliable.

The most classical method to tackle ill-posedness is provided by Least Squares and Generalized Inversion. We briefly expose them in the setting (2.19).

Our aim is to solve the equation $Lx = y$. The *Least Squares method* (or *pseudosolution*) consists in computing the solution of the following variational problem:

$$\underset{x\in\mathcal{X}}{\operatorname{argmin}} \, ||Lx - y||_{\mathcal{Y}}. \tag{2.22}$$

In Hilbert spaces, this problem is equivalent to solve the Euler equation [21]:

$$L^*Lx = L^*y \tag{2.23}$$

where L^* is the adjoint operator of L.

Notice that uniqueness is not assured if L is not injective. However, uniqueness can be re-established considering the solution of the equation (2.23) that has minimal norm: this is called a *Generalized solution* of the ill-posed problem and is indicated by x^G:

$$x^G = \underset{x\in\mathrm{S}}{\operatorname{argmin}} \, ||x||_{\mathcal{X}} \quad \text{where } \mathrm{S} = \left\{ \widehat{x} \in \mathcal{X} : \widehat{x} = \underset{x\in\mathcal{X}}{\operatorname{argmin}} \, ||Lx - y||_{\mathcal{Y}} \right\}. \tag{2.24}$$

In other terms, the Generalized solution is the Least Squares solution with minimal norm, which is unique since the set of solutions of (2.23) is convex and closed (see [36, Section 1.4]).

The existence of x^G is guaranteed only under some conditions: (2.22) corresponds to compute the projection of y on the closure of $\mathrm{Im}(L)$, hence the generalized solution exists only if such projection lies in $\mathrm{Im}(L)$, which holds if and only if

$$y = \mathrm{Im}(L) + \mathrm{Im}(L)^{\perp}.$$

Thus, no problems arise if $\mathrm{Im}(L)$ is closed: in this case the generalized solution exists and also the linear operator

$$L^G : \mathcal{Y} \to \mathcal{X}$$
$$L^G y = x^G$$

is continuous. Nevertheless, if $\mathrm{Im}(L)$ is not closed, then the Generalized solution exists if $y = \mathrm{Im}(L) + \mathrm{Im}(L)^{\perp}$ (which is a dense subspace of $\mathcal{Y}$), but the continuity of L^G is not provided.

To sum up the above, the uniqueness issue can be worked out using Generalized inverses solution, whose existence is guaranteed if the range of L is closed along with the continuity of L^G. On the other hand, if the range of L is not closed (for instance, when L is compact on a space of infinite dimension), L^G, if it exists, in general is not continuous. The closure of $\mathrm{Im}(L)$ is then a crucial point up to now.

Let us see some examples.

Example 8 (Continuation of Example 1) *In the Sum example, the Least Squares solution is given by* $\mathrm{argmin}_{x\in\mathbb{R}^2}(x_1+x_2-y)^2$*, which is not unique, since* $min_{x\in\mathbb{R}^2}(x_1+x_2-y)^2 = 0$ *for every* $x \in \mathbb{R}^2$ *such that* $x_1 + x_2 = y$*. Among them, the Generalized solution is the one with minimal norm, say* $x^G = \mathrm{argmin}_{x:x_1+x_2=y}\, x_1^2 + x_2^2$ *which can be easily solved by Lagrange multipliers method, the solution being* $x^G = \frac{y}{2}(1,1)$*. Notice that* x^G *is perpendicular to the kernel of the Sum operator generated by* $(1,-1)$*.*

Example 9 (Continuation of Example 3) *Let us consider the integration/differentiation problem supposing that* $\mathcal{X} = C[0,1]$*, with norm* $||x|| = \max_{t\in[0,1]}|x(t)|$*. Then Im(L)* $= \{f(t) \in C^1[0,1] : f(0) = 0\}$*. It is easy to check that Im(L) is not closed. For instance let us consider the sequence of functions* $f_n(t) = |x-1/2|^{1+1/n} - 1/2^{1+1/n}$*. For any* $n \in \mathbb{N}$*,* $f_n(t) \in$ *Im(L), but the sequence uniformly converges to* $|x-1/2| - 1/2$ *which has no derivative in* $t = 1/2$

The method of Generalized Inverses seems to be first introduced by Fredholm in 1903 [55], which gave a Generalized solution of an integral problem. The reader is referred to [99],[16] and to the references therein for a more general and comprehensive treatment on the subject; furthermore, an extensive on-line bibliography can be consulted at the web page of the International Linear Algebra Society: http://www.math.technion.ac.il/iic/

2.3.3 Ill-conditioning

Well-posedness is not sufficient to guarantee the *stability* of a system, since also numerical instability may affect the problem. Numerical instability is connected with discretization: in most situations, even if the problem is formulated in a continuous space, the experimental data are given by measurements performed *at a finite number of points* in the domain of definition and in time. In fact, the observable quantities may not always be accessible or the instrumentation to perform measurements may be digital, hence the data are gathered in the form of vectors and matrices.

Let us consider the discretized version of $Lx = y$, that is, suppose the operator L to be suitably approximated by a matrix $\mathbf{L}$. The system to invert is now: $\mathbf{y} = \mathbf{L}\mathbf{x}$ where $\mathbf{x}$ and $\mathbf{y}$ are consistent vectors. By Generalized Inversion, we can define $\mathbf{L}^G$ as $\mathbf{x}^G = \mathbf{L}^G\mathbf{y}$. Suppose now that an error $\delta\mathbf{y}$ occurs on the data and let $\delta\mathbf{x}^G$ be the induced error on the Generalized solution: then, $\delta\mathbf{x}^G = \mathbf{L}^G\delta\mathbf{y}$ and also $||\delta\mathbf{x}^G|| \leq ||\mathbf{L}^G||\ ||\delta\mathbf{y}||$ which combined with $||\mathbf{y}|| \leq ||\mathbf{L}||\ ||\mathbf{x}^G||$ produce the following inequality:

$$\frac{||\delta\mathbf{x}^G||}{||\mathbf{x}^G||} \leq ||\mathbf{L}||||\mathbf{L}^G||\ \frac{||\delta\mathbf{y}||}{||\mathbf{y}||} \tag{2.25}$$

which represents a bound for the induced error rate in function to the data error rate. The quantity $||\mathbf{L}||||\mathbf{L}^G||$ is named *condition number* and a problem with condition number considerably larger than one is said to be *ill-conditioned.*

Notice that ill-conditioning generally does not depend on the accuracy of the discretization (see [18, Section 4.4]): in fact, if the problem is ill-posed it may happen that the finer the discretization, the larger the condition number.

Example 10 (Continuation of Example 3) . *Let us set in $C^1([0,1])$ and discretize the integral problem by associating to each function in $C^1([0,1])$ a vector in R^N constructed by sampling the function at $\frac{k}{N}$, $k = 1, \ldots, N$, that is, $\mathbf{x} = (x_1, \ldots, x_N)$, $x_i = x(\frac{i}{N})$, $i = 1, \ldots N$ represents $x(t)$. Moreover, the discretized version of $y(t) = \int_0^t x(s)\mathrm{d}s$ is given by $y_i = \frac{1}{N}\sum_{j=1}^{i} x_i$ or, in matrices terms, $\mathbf{y} = \mathbf{L}\mathbf{x}$ where*

$$\mathbf{L} = \frac{1}{N}\begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 1 & 0 & \cdots & 0 \\ \vdots & & & \ddots & & \vdots \\ \vdots & & & & \ddots & \vdots \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Since

$$\mathbf{L}^G = \mathbf{L}^{-1} = N \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

and considering the matrix norm $|| \, ||_1$ *(*$||A||_1 = \max_{i=1,\dots,N} \sum_{j=1}^{N} |A_{i,j}|$*) we have* $||L||_1 = 1$ *and* $||L^{-1}||_1 = 2N$*. Hence the condition number is* $2N$*: the more the discretization is refined, the more the inverse problem is ill-conditioned.*

Notice that it may occur that a well-posed problem is ill-conditioned after discretization.

Generalized Inversion presents the same instability features for ill-posed and ill-conditioned problems. Regularization methods, introduced in the next section, can be considered as an improvement of Least Squares and Generalized Inversion in the sens of the robustness with respect to inaccuracies.

2.4 Regularization Methods

This section and the following one are aimed to review the most important methodologies introduced in the last fifty years to resolve inverse problems. The main techniques can be mainly divided into two branches: the deterministic and the probabilistic methods.

The deterministic methods are generally based on *regularization*, which represents an improvement with respect to Least Squares and Generalized Inverses solutions, that have been shown to be inadequate in many common occurrences (see Section 2.3.2).

Afterwards, probabilistic methods are described. Based on the consideration that a certain amount of uncertainty affects inverse systems, primarily due to the observational noise which usually can be modeled as a random variable, the stochastic methods are a more recent and intrinsically alternative branch with respect to regularization (even if in some cases analogies between regularization and probabilistic methods will be highlighted, in particular in Gaussian frameworks).

These are the two main branches that have been developed until nowadays: almost all inverse problems' methods refer to either of them. Naturally, some important subbranches can be individuated and are illustrated in the next.

2.4.1 Tikhonov regularization

Regularization methods are now introduced by illustrating the original idea of Tikhonov. Let us consider again the setting introduced in Section 2.3.2 and the system

$$Lx = y \quad x \in \mathcal{X}, y \in \mathcal{Y}.$$

It has been shown that Generalized Inversion consists in solving two minimization problems, say

$$\mathrm{S} = \{\widehat{x} \in \mathcal{X} : \widehat{x} = \arg\min_{x\in\mathcal{X}} ||Lx - y||_{\mathcal{Y}}\}$$

and

$$x^G = \arg\min_{x\in\mathrm{S}} ||x||_{\mathcal{X}}.$$

The seminal idea of Tikhonov [148] [149] can be viewed as the merging of these two problems into the following one:

$$x_\alpha^T = \arg\min_{x\in\mathcal{X}} \left(||Lx - y||_{\mathcal{Y}}^2 + \alpha||x||_{\mathcal{X}}^2\right) \tag{2.26}$$

The solution depends on the parameter α: it can be proved that x_α^T is unique and corresponds to

$$x_\alpha^T(r) = (L^*L + \alpha^2 I)^{-1}L^*y. \tag{2.27}$$

It is easy to prove that in Hilbert spaces

$$\lim_{\alpha\to 0}(L^*L + \alpha I)^{-1}L^* = L^G \tag{2.28}$$

under the hypothesis of existence of L^G, say $y = \mathrm{Im}(L) + \mathrm{Im}(L)^\perp$

The role played by the parameter α can be qualitatively explained as follows. Solving the least squares problem (2.22) corresponds to minimizing the residual between the measured data y and Lx, which implies some *fidelity* to the observed data y; in other terms, the observational noise is known (or is believed) to cause small perturbations. Nevertheless, this is not always so and if the noise has large amplitude such procedure may yield to mathematically correct, but physically non-admissible solutions. This consideration suggests to collect some information about the physics of the problem in order to individuate a set of admissible solutions and try to solve the problem therein.

The Tikhonov method is a first example of *regularization method.* More in general,

Definition 1 *A family of operators*

$$\{R_\alpha : \mathcal{Y} \to \mathcal{X},\ \ \alpha \in A\} \tag{2.29}$$

is said to be a regularizer *of the equation $Lx = y$ if for any $\alpha \in A$, R_α is continuous and for any $y \in \mathcal{Y}$,*

$$\lim_{\alpha\to 0} R_\alpha(y) = L^G y. \tag{2.30}$$

Notice that such definition requires the existence of the Generalized solution. The Tikhonov operator $(L^*L + \alpha I)^{-1}L^*$ clearly is a regularizer. Moreover, the Tikhonov method can be extended to many other regularization methods by substituting the quadratic conditions with other functionals, that is:

$$x_\alpha^R(y) = \arg\min_{x\in\mathcal{X}} \mathcal{F}(Lx - y) + \alpha\mathcal{G}(x), \quad \alpha \in (0, +\infty) \tag{2.31}$$

where $\mathcal{F}$ and $\mathcal{G}$ respectively represent a condition on the residual $Lx - y$ given by the measured data and some property imposed on the possible solutions x. x_α^R defines a regularization solution if $\lim_{\alpha\to 0} x_\alpha^R(y) = L^G y$.

The formulation (2.31) being definitely general, the problem has to be suitably defined according to the context. The choice on $\mathcal{F}$, $\mathcal{G}$ and α is arbitrary, but quantified on the basis of some crucial considerations. For example, as far as α is concerned, it can be intuitively understood that the more α is close to zero, the more reliability is assigned to the observations and vice verse a large α is equivalent to maximal adherence to the prior information.

Clearly, this is not enough to decide the *best* α, whose assessment is object of many discussions (see Section 2.4.2. On the other hand, the question of the choice on $\mathcal{F}$ and $\mathcal{G}s$ is more qualitative: for instance, in the example (2.26), $\mathcal{F}$ is the discrepancy between the measured output y and the output obtained by applying L on x, while $\mathcal{G}$ is the energy of the signal x.

Example 11 (Continuation of Example 3) *Let us consider again the discretized equation* $\mathbf{Lx} = \mathbf{y}$. *A Tikhonov solution of this equation is*

$$\mathbf{x^T} = (\mathbf{L}^T\mathbf{L} + \alpha\mathbb{I})^{-1}\mathbf{L}^T\mathbf{y}. \tag{2.32}$$

Let us study the conditioning:

$$\frac{||\delta\mathbf{x}^T||}{||\mathbf{x}^T||} \leq ||\mathbf{L}||||(\mathbf{L}^T\mathbf{L} + \alpha\mathbb{I})^{-1}\mathbf{L}^T||\ \frac{||\delta\mathbf{y}||}{||\mathbf{y}||}. \tag{2.33}$$

Considering the matrix norm $||\ ||_1$, *we have* $||L||_1 = ||L^T|| = 1$. *Moreover,*

$$\mathbf{L}^T\mathbf{L} = \frac{1}{N^2}\begin{pmatrix} N-1 & N-2 & N-3 & \cdots & \cdots & 1 \\ N-2 & N-3 & \cdots & \cdots & 1 & 1 \\ \vdots & & & & & \vdots \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Notice that if N is very large, $\mathbf{L}^T\mathbf{L}$ *tends to the null matrix and* $||(\mathbf{L}^T\mathbf{L}+\alpha\mathbb{I})^{-1}\mathbf{L}^T||_1 \to |\alpha^{-1}|$. *In such case, the condition number* $||\mathbf{L}||\ ||(\mathbf{L}^T\mathbf{L} + \alpha\mathbb{I})^{-1}\mathbf{L}^T||$ *tends to* $|\alpha^{-1}|$. *If* α *is a fixed positive constant, this represents a significant improvement with respect to Generalized Inversion, where the condition number was* $2N$, *hence increasing as far as the discretization was refined.*

Now, let us suppose to know that $||\mathbf{Lx} - \mathbf{y}||^2 \leq \sigma^2$, *namely the maximal* energy *of the noise is* σ^2. *Then, as a consequence of [45, Theorem 5.2],* $\alpha = \sigma$ *is a good choice for the Tikhonov parameter. In such case, the condition number is* σ^{-1}.

This example shows that Tikhonov regularization is more suitable than Generalized Inversion to treat ill-conditioning and when a considerable disturbance occurs.

Example 12 *Another observation can be done in the simple case $N = 1$ ($\mathbf{L} = 1$) which is not dramatically ill-posed. Let us suppose the maximal energy error to be σ^2, i.e. $|x - y| \leq \sigma$ where x and y are scalar real values. By direct inversion, we have $x^G = \mathbf{L}^{-1}y = y$, hence the distance between correct solution and inverse solution is at most σ. If we apply a Tikhonov regularization with $\alpha = \sigma$, $x^T_\alpha = \frac{1}{1+\sigma}y$ bounds the error in case of large noise. In fact, if x is the correct solution and assuming $y = x + \sigma$, $|\frac{1}{1+\sigma}y - x| = |\frac{1}{1+\sigma}(x + \sigma) - x| = \frac{\sigma}{1+\sigma}|1 - x|$, which is bounded with respect to the noise energy.*

In conclusion, regularization methods are based on the idea that some prior information about the regularity of the solution is known and can be efficiently used to reduce the noise effects. On the other hand, regularization does not solve the problems of continuity of the Generalized Inverse.

2.4.2 Choice of the regularization parameter α

Several methods have been proposed in the literature to choose an opportune value for the regularization parameter α. This is a critical point which has no universal solution: the best α for an inverse problem may be determined by the physical meaning of the model and by its applicative scope. Some of the most used methods are the Discrepancy Principle [96], L-curve method [67], Cross Validation [157].

2.4.3 Early regularization: a few historical notes

The term *regularization* was introduced by Tikhonov, which may be considered the original author of the regularization methods; its seminal work [148] was published in 1963 in the USSR. In the same period, another Russian mathematician, Ivanov, and Phillips in the USA worked on the similar arguments.

Ivanov was studying ill-posed problems since 1961 and in 1962 [73, 72], he introduced the notion of *quasi-solution* which can be briefly presented as follows. Given the usual system $Lx = y$, let the operator L to be invertible, linear and continuous. Fixed a certain y, let suppose that the solution $x = L^{-1}y$ belongs to a compact set $\mathcal{C} \subset \mathcal{X}$.

In this setting, quasi-solutions can be used to approximate solutions of $Lx = y$.

Given compact $\mathcal{C} \subset \mathcal{X}$, x_0 is said to be a quasi-solution of $Lx = y$ if x_0 minimizes the residual norm $||Lx - y||$ on $\mathcal{C}$:

$$\min\{||Lx - y|| : \ x \in \mathcal{C}\} = ||Lx_0 - y||. \tag{2.34}$$

The existence of a quasi-solution is not ensured, but Ivanov proved that if the solution of $Lx = y$ exists in $\mathcal{C}$, then it corresponds to the quasi-solution. Moreover, he proved stability with respect to bounded perturbations on the data and on the operator L.

In the same year, Phillips [113] proposed an approach for nonsingular linear integral equations of the first kind, with noisy data. First, he fixed two hypotheses: the noise is upper-bounded by a given constant k, say $||Lx - y||_\mathcal{Y} \leq k$, and the correct solution is known to be a “reasonably smooth function”. Now, given a family of estimated solutions, “with this [smoothness] assumption the best approximation we can choose is probably the function which is the smoothest in some sense”. Of the various smoothness conditions, in [113] Phillips chose to minimize the norm of the second derivative of the solution.

In general, the solution of this method (known also as *method of residual*) can be expressed as:

$$x^P = \arg\min\{\mathcal{G}(x) : ||Lx - y||_{\mathcal{Y}} \leq k\}.$$

To conclude this brief overview about early regularization algorithms, it should be noticed that the three methods discussed above (Tikhonov, Ivanov and Phillips) are strongly interrelated. In fact, even if arising from different contexts and requiring different conditions, their general structure as variational problems is very similar, as it can be appreciated in the following comparing scheme:

$$\begin{array}{ll} \text{Tikhonov} & x^T = \arg\min\{\mathcal{F}(Lx-y) + \alpha\mathcal{G}(x)\} \\ \text{Ivanov} & x^I = \arg\min\{\mathcal{F}(Lx-y) : \mathcal{G}(x) \leq h\} \\ \text{Phillips} & x^P = \arg\min\{\mathcal{G}(x) : \mathcal{F}(Lx-y) \leq k\} \end{array} \tag{2.35}$$

where the operators $\mathcal{F}$ and $\mathcal{G}$ usually are the the norms or square norms in the corresponding spaces.

2.4.4 Iterative methods

Regularizers can be constructed also by means of iterative schemes, typically arising from the solution of linear algebraic problems. Classical iterative methods are the fixed point iteration scheme, the Landweber’s method [45, Section 6.1], the Steepest Descent and the Conjugate Gradient method [18, Section 6.4]. A useful review of the main iterated techniques for linear algebraic systems in [28, Appendix C].

In the last decade, a novel iterative approach for deconvolution has been introduced by Fagnani and Pandolfi [48, 49, 47], which works under less restrictive conditions with respect to classical recursive methods.

The considered problem is the estimation of the input u of a linear, finite dimensional system

$$\begin{cases} x'(t) = Ax(t) + Bu(t) & t \in [0, T] \\ y(t) = Cx(t) & u(t) \in \mathbb{R}^m,\ x(t) \in \mathbb{R}^q,\ y(t) \in \mathbb{R}^p, \end{cases} \tag{2.36}$$

so that the input-output relation can be expressed as

$$y(t) = \int_0^t H(t-s)u(s)\mathrm{d}s \qquad H(t-s) = Ce^{A(t-s)}B. \tag{2.37}$$

The aim is to recover u from the observation of a noisy version of the output y.

The technique proposed for this causal problem originates from the algorithm proposed in [84, Section 1.3] and is based on two main features: first, a *model system* is associated with the real system (2.36) with the goal of testing the candidate approximations of u; second, a Tikhonov penalization is introduced to adjust the algorithm.

More in detail, let

$$\begin{cases} w'(t) = Aw(t) + Bv(t) \qquad t \in [0, T] \\ \\ z(t) = Cw(t) \end{cases} \tag{2.38}$$

be the model system, which depicts (2.36) and let us suppose that the measurements of y are noisy and taken only at time instants $\tau_k = \frac{kT}{N}$, $N \in \mathbb{N}$, $k = 1, \dots, N$, so that the available samples are $r_k = r(\tau_k) = y(\tau_k) + n_k$, $k = 1, \dots, N$ where n_k is a bounded disturbance, say $||n_k|| < h$. Now, the function v is recursively built up according to the following rule:

$$v_k = v|_{[\tau_{k-1}, \tau_k)} = \arg\min \left\{ ||z(\tau_k) - r_k||^2 + \alpha \int_{\tau_{k-1}}^{\tau_k} ||v(s)||^2 \mathrm{d}s \right\} \tag{2.39}$$

where α is the regularization parameter. Such v is then stepwise constant and is used to approximate the object of interest u. The analogy with Tikhonov regularization is evident, the only main difference being the recursivity.

Consistency, robustness and convergence of this method along with the admissible kernels and the case $T \to \infty$ have been discussed in [48, 49]. In [49] it is also shown that for small α, one can approximate v_k with $\frac{z(\tau_{k-1}) - r_{k-1}}{\alpha}$, which makes the algorithm very simple and low-complexity.

An interesting application of this iterative method to problems of disturbance reduction can be retrieved in [47].

2.5 Probabilistic Methods

In the previous section, the main regularization methods have been illustrated for inverse, ill-posed problems. A possible probabilistic nature of such problems has been not yet mentioned, but it is natural if we model the observational noise as a random variable. This is actually the starting point of the probabilistic analysis of the inverse problems: the noise function, in general, is not known and does not have a deterministic behavior. In some instances, it may be sufficient to know an upper bound of its amplitude, but when more information about its behavior is required to find the solution, one could try to model it according to an appropriate probabilistic distribution. In many cases, this turns out to be a very efficient approach.

Notice that the introduction of a random noise in the model is sufficient to give a probabilistic structure to the whole problem (the available data and the estimate of the input are now random variables, too). In many applications, also the unknown object is known to be ruled by some stochastic law, so that another source of uncertainty is present; this is a further prior information (if the law is given) that can be used to obtain the solution.

In some sense, the knowledge of the noise's distribution replaces the requirement of its boundedness and the knowledge if its maximal amplitude, while a given probabilistic law on the input may replace the smoothness conditions that very often are needed by regularization methods to give likely solutions.

In this section, the main probabilistic methodologies for inverse problems are illustrated. Notice that in this framework, inversion can be considered as a problem of *inference*, that is, an estimate of the solution is provided by gathering sufficient statistics from the available information (in spite of incompleteness and inaccuracies). Before introducing the methods, a short discussion about the noise law is provided.

2.5.1 How to model the noise

In most cases, and also in this dissertation, the noise on the data is supposed to be additive and modeled as a Gaussian (normal) random variable, with zero mean and covariance $\sigma^2\mathbb{I}$, independent from the other quantities participating in the system. This is what we call *white noise* and we indicate by $\mathcal{N}(0, \sigma^2\mathbb{I})$.

The additivity may be easily justified: usually, measurements' instrumentation do not amplify or compress the observed signal, their effect being more similar to a small shift. The zero-mean, Gaussian distribution, instead, may be motivated by three main issues. First, it is well known that the Central Limit Theorem states that the distribution of the sum of a sufficiently large number of mutually independent random variables (with finite sum and variance) is well-approximated by a Gaussian distribution. Since the noise is usually the accumulation of many small random contributions, the Gaussian model is the most likely or, at least, a good trade-off between reality and modeling requirements. Second, the noise typically is not caused by systematic error and this is well represented by a random variable with zero mean. Third, large noise oscillations are less probable than small ones (again in the case of no systematic error) and this is provided by the normal distribution with fixed variance.

2.5.2 Maximum Likelihood methods

The Maximum Likelihood (ML) principle, introduced by Fisher in the 1920s, can be intuitively explained as follows. Let us consider an unknown quantity x and some noisy measurement $y = Lx + n$. If the noise n obeys to a probabilistic distribution, the data y can be considered as realization of a random variable, that will be indicated by Y. Now, a ML estimate $\widehat{x}_{\mathrm{ML}}$ of x given y can be obtained by answering the following question: which is the x that maximizes the probability of having obtained the observed Y? In other terms, ML consists of solving the following maximization problem:

$$\widehat{x}_{\mathrm{ML}} = \arg\max_{x \in \mathcal{X}} f(Y = y|x) \tag{2.40}$$

where $f(Y = y|x)$ is the probability density function of Y given the parameter x. Notice that, up to now, no probabilistic distribution is assigned to x, which can be a deterministic object or a random function with unknown distribution.

This method is more reliable if y collects a sufficiently large number of measurements.

Let us consider the case in which uncertainty in the model is introduced by a white noise n, which is realization of a Gaussian random variable N. In this case, $f(Y = y|x)$ just depends on the distribution of N since $f(Y = y|x) = f(N = n = y - Lx)y|x)$. If N is Gaussian, the probability law is given by

$$f(Y = y|x) = f(N = n = y - Lx|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{||n||^2}{2\sigma^2}\right) \tag{2.41}$$

hence

$$\widehat{x} = \arg\max_{x\in\mathcal{X}} f(Y = y|x) = \arg\min_{x\in\mathcal{X}} ||n||^2 = \arg\min_{x\in\mathcal{X}} ||y - Lx||^2 \tag{2.42}$$

which actually is a Least Squares solution. In conclusion, ML corresponds to Least Squares method in case of white noise and as a consequence it is affected by the same negative effects: non-uniqueness of the solution and numerical instability.

Maximum likelihood takes its name by the likelihood function, which is defined as $L(\theta|y) = f_Y(y|\theta)$ where θ collects the parameters of the distribution. In other terms, likelihood *is* the p.d.f., but the difference lies in the role of the independent variables: likelihood is a function of θ (and y is a parameter), while f is a function of y.

Hence the maximization problem (2.42) corresponds to the maximization of the likelihood.

The interested reader could find a discussion about ML method in the case of Poisson noise in [18, Section 7.3]. This is important in image restoration, where often the data is the number of detected photons, the counting of which is rule by Poisson statistics.

2.5.3 Expectation-Maximization Algorithm

The Expectation-Maximization (EM) Algorithm is an iterative procedure commonly used in statistics to compute the ML estimate in presence of missing or hidden data; it was introduced by Dempster in 1977 [37].

Let us consider two random variables: Y, whose realizations can be observed, and Z, which represents the hidden data; the aim is to compute the maximum likelihood estimate of a set of unknown parameters θ.

Each iteration of the EM Algorithm is composed by two parts: the Expectation (E) step, which consists in computing the expectation of a likelihood function of the *complete* data (Y, Z) given an estimate θ_n of θ, and the Maximization (M) step in which such expectation is maximized and provides a new estimate θ_{n+1}.

More precisely, let

$$L(\theta_n|Y) = \log f_{Y|\theta_n}(y|\theta_n) \tag{2.43}$$

be the likelihood function of Y given θ, f being the probability density function (or a probability in the discrete case). By simple computations,

$$\begin{aligned} L(\theta_n|Y) &= \int \log f_{Y|\theta_n}(y|\theta_n)\ f_{Z|Y,\theta_n}(z|y,\theta_n)\mathrm{d}z \\ &= \int \log \frac{f_{Z,Y|\theta_n}(z,y|\theta_n)}{f_{Z|\theta_n}(z|\theta_n)}\ f_{Z|Y,\theta_n}(z|y,\theta_n)\mathrm{d}z \\ &= Q(\theta_n|\theta_n) + R(\theta_n|\theta_n) \end{aligned} \tag{2.44}$$

where

$$\begin{aligned} &Q(\theta|\theta_n) = \mathbb{E}_Z[L(\theta|Z,Y)|Y,\theta_n], \quad L(\theta|Z,Y) = \log f_{Z,Y|\theta}(z,y|\theta) \\ &R(\theta|\theta_n) = -\mathbb{E}_Z[\log f_{Z|Y,\theta}(z|y,\theta)|Y,\theta_n]. \end{aligned} \tag{2.45}$$

Using Jensen inequality, it can be proved that

$$L(\theta|Y) \geq Q(\theta|\theta_n) + R(\theta_n|\theta_n)$$

therefore a better estimate of L can be obtained by increasing Q at each iteration step. The EM procedure is then as follows: given an initial estimation θ_0, at each step $n = 0, 1, 2, \ldots,$

- E-step: compute $q_n(\xi) = Q(\xi|\theta_n) = \mathbb{E}_Z[L(\xi|Z,Y)|Y,\theta_n]$;
- M-step: compute $\theta_{n+1} = \mathrm{argmax}_\xi\, Q(\xi|\theta_n)$.

It follows that $Q(\theta_{n+1}|\theta_n) \geq Q(\theta_n|\theta_n)$, hence

$$\begin{aligned} L(\theta_{n+1}|Y) &\geq Q(\theta_{n+1}|\theta_n) + R(\theta_n|\theta_n) \\ &\geq Q(\theta_n|\theta_n) + R(\theta_n|\theta_n) = L(\theta_n|Y). \end{aligned}$$

The convergence of the EM Algorithm is studied, e.g,, in [92, Chapters 3-4].

The EM algorithm is commonly used in many applications in particular for tomography [132, 153] and image restoration [154], and many applicative variants have been recently studied, see, e.g., [123, 122, 53, 168]

2.5.4 Bayesian Methods and Maximum A Posteriori estimation

The Bayesian approach takes its name from the well-known Bayes' rule. Given two continuous random variables X taking values in $\mathcal{X}$ and Y taking values in $\mathcal{Y}$, the probability density function of X condition to Y, $f_{X|Y}(x|y)$ can be computed as

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)} = \frac{f_{Y|X}(y|x)f_X(x)}{\int_{x'\in\mathcal{X}} f_{Y|X}(y|x')f_X(x')} \tag{2.46}$$

Let us consider again the system $Lx = y$ and let us suppose that also the input x is generated according to some probabilistic distribution, i.e., it is a realization of a random variable X; then, we can rewrite the system in terms of a stochastic equation $LX = Y$. If the probability density function $f_X(x)$ of X is known, it constitutes an important prior information about the object of inference; the aim of Bayesian approach is exactly to take it in consideration and exploit it to reduce the uncertainty on x.

In this context, a Maximum A Posteriori (MAP) estimation can be performed. The MAP method answers the question: which is the value of x that maximizes the *a posteriori* p.d.f. of X, say the probability of X given the (noisy) measurements y?

In mathematical terms,

$$\widehat{x}_{\text{MAP}} = \arg\max_{x\in\mathcal{X}} f_{X|Y}(x|y). \tag{2.47}$$

Using the Bayes rule, this corresponds to

$$\widehat{x}_{\text{MAP}} = \arg\max_{x\in\mathcal{X}} f_{Y|X}(y|x) f_X(x) \tag{2.48}$$

which suggests that MAP is equivalent to ML when X has a uniform prior distribution; in other terms, MAP merges ML method and prior probabilistic information on the unknown quantity.

We finally notice that as well as ML is equivalent to Least Squares method in the presence of zero-mean, Gaussian noise, MAP corresponds to a regularization method to solve the problem $LX + N = Y$ where both noise N and unknown X are independent, zero-mean, Gaussian random variables, namely $N \sim \mathcal{N}(0, \sigma_N^2)$ and $X \sim \mathcal{N}(0, \sigma_X^2)$ (or $N \sim \mathcal{N}(0, \sigma_N^2\mathbb{I})$ and $X \sim \mathcal{N}(0, \sigma_X^2\mathbb{I})$ if we consider vector quantities). In fact,

$$\begin{aligned}\max_x f_{X|Y}(x|y) &= \max_x f_{Y|X}(y|x) f_X(x)\\ &= \max_x \frac{1}{\sigma_N\sqrt{2\pi}} \exp\left(-\frac{||Y-LX||^2}{2\sigma_N^2}\right) \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left(-\frac{||X||^2}{2\sigma_X^2}\right)\\ &= \min_x ||Y-LX||^2 + ||X||^2\end{aligned} \tag{2.49}$$

which is exactly a formulation of the Tikhonov regularization method (2.26).

2.5.5 Wiener Filter

It has already been noticed that ML with Gaussian noise is equivalent to Least Squares approach. Similarly, if both noise $N = Y - LX$ and X Gaussian and mutually independent, say $N \sim \mathcal{N}(0, \sigma_N^2)$ and $X \sim \mathcal{N}(0, \sigma_X^2)$

$$\begin{aligned}\widehat{x}_{\text{MAP}} = \arg\max_{x\in\mathcal{X}} f_{Y|X}(y|x) f_X(x) &= \arg\max_{x\in\mathcal{X}} \exp\left(-\frac{||n||^2}{2\sigma_N^2}\right) \exp\left(-\frac{||x||^2}{2\sigma_X^2}\right)\\ &= \arg\min_{x\in\mathcal{X}} \frac{||n||^2}{2\sigma_N^2} + \frac{||x||^2}{2\sigma_X^2}\\ &= \arg\min_{x\in\mathcal{X}} ||n||^2 + \frac{\sigma_N^2}{\sigma_X^2}||x||^2\end{aligned} \tag{2.50}$$

which corresponds to Tikhonov's regularization with quadratic terms and $\alpha = \frac{\sigma_N^2}{\sigma_X^2}$ In particular, applying (2.27),

$$\widehat{x}_{\text{MAP}} = \left(L^*L + \frac{\sigma_N^2}{\sigma_X^2} I \right)^{-1} L^*y = (\sigma_X^2 L^*L + \sigma_N^2 I)^{-1} \sigma_X^2 L^*y. \tag{2.51}$$

Notice that the quantity $\frac{\sigma_N^2}{\sigma_X^2}$ is exactly the inverse of the *signal-to-noise ratio* (SNR), say the ratio between the power of the signal and the power of the noise that corrupts it. The operator

$$G = \left(L^*L + \frac{1}{\text{SNR}} I \right)^{-1} L^* = (\sigma_X^2 L^*L + \sigma_N^2 I)^{-1} \sigma_X^2 L^* \tag{2.52}$$

is commonly known as Wiener filter and has been studied in the 1940s [161].

The Wiener filter is optimal, in the sense that it minimizes the mean square error $\mathbb{E}[|\widehat{X}_{MAP} - X|^2] = \mathbb{E}[|G(LX + N) - X|^2]$, in the case of Gaussian X; this can be easily proved by computing the derivative.

A slightly modification of the Wiener filter may be obtained introducing a regularization parameter α:

$$G_\alpha = (\sigma_X^2 L^*L + \alpha \sigma_N^2 I)^{-1} \sigma_X^2 L^*. \tag{2.53}$$

Adjusting α, different effects can be obtained: for instance, if $\alpha = 0$, then the estimate is unbiased but noisy (actually the noise power is not taken in account), while if $\alpha \to \infty$, the noise is somehow suppressed, but the estimate is definitely unlikely.

A generalization of (2.51) to the case of not independent X , say X is random vector with non-diagonal covariance matrix as assumed here, can be retrieved in [36, Section 3.8.3] or [18, Section 7.5]

2.5.6 Fourier Transform and Wavelet Deconvolution

In some situations, the deconvolution problem can be formulated in terms of Fourier Transforms. Given the convolution integral (on an infinite range)

$$x(t) = (\mathcal{K} * u)(t) = \int_{-\infty}^{+\infty} \mathcal{K}(t-s)u(s)\mathrm{d}s \tag{2.54}$$

and adding a noise s.t. $y(t) = x(t) + n(t)$, the problem can be treated in the frequency domain, where convolution is transformed into a product:

$$\mathcal{F}(y)(\omega) = \mathcal{F}(\mathcal{K})(\omega)\mathcal{F}(u)(\omega) + \mathcal{F}(n)(\omega) \tag{2.55}$$

where $\mathcal{F}$ indicates the Fourier transform. Since

$$\mathcal{F}(u)(\omega) = \mathcal{F}(\mathcal{K})(\omega)^{-1}\mathcal{F}(y)(\omega) + \mathcal{F}(\mathcal{K})(\omega)^{-1}\mathcal{F}(n)(\omega) \tag{2.56}$$

the Fourier transform of u could be naively estimated by $\widehat{\mathcal{F}(\S)}(\omega) = \mathcal{F}(\mathcal{K})(\omega)^{-1}\mathcal{F}(y)(\omega)$. This would be a correct solution if no noise occurred (and naturally under the hypothesis that the involved functions admit Fourier transform and that $\mathcal{F}(\mathcal{K})(\omega)$ is invertible) while if a Gaussian noise was introduced, this would be an unbiased estimator.

This solution however is not precise whenever $\mathcal{F}(\mathcal{K})(\omega)$ is very small, since this amplifies the noise in (2.56). So, even if the noise was small, its effect might be considerable.

The Fourier approach can be improved by introducing some regularization, for instance through a Wiener filter. Supposing that $X \sim \mathcal{N}(0, \sigma_X^2)$ and $N \sim \mathcal{N}(0, \sigma_N^2)$, the following estimator can be derived:

$$\widehat{\mathcal{F}(x)}(\omega) = G_\alpha(\omega)\mathcal{F}(y)(\omega) \qquad G_\alpha(\omega) = \frac{|\mathcal{F}(\mathcal{K})|\sigma_X^2}{|\mathcal{F}(\mathcal{K})|^2\sigma_X^2 + \alpha\sigma_N^2}. \tag{2.57}$$

For $\alpha = 1$, this is exactly the Wiener filter, which is optimal since it minimizes the mean square error. However, this method is not efficient whenever the signal X is not stationary: for instance, if X was representing an image with edges and ridges to be detected, G_α might smear them and make them non detectable.

Non-stationarities can be instead captured using *wavelets' bases*, which allow to analyze data at different scales. The wavelet approach to deconvolution has been widely studied in the last decades; we refer the interested reader to, e.g, [40, 1, 102, 103].

2.5.7 Bayesian Least Mean Squares estimation

In this section, a probabilistic version of the Least Squares method, introduced in Section 2.3.2, is illustrated.

In Section 2.3.2 the Least Squares method has been introduced. as a deterministic method to approximate the solution of an inverse problem $Lx = z$ by computing the x that minimizes the discrepancy term $||Lx - y||^2$, y being the available, noisy version of the output z. In a probabilistic framework, say when x and y are realizations of random variables X, Y (here supposed to be random vectors) it is natural to redefine such principle in terms of average: given measurements y of the noisy output Y, the minimization problem to solve is

$$\min_{x \in \mathcal{X}} \mathbb{E}[||X - x||^2 | Y = y]. \tag{2.58}$$

This solution, known as *Least Mean Squares Estimate* (or also *Bayesian* or *minimum mean-square-error* or *Minimum Variance* estimate; LMSE for short), is the basis of the Bayesian approach to Estimation Theory. Estimation theory aims to study how to extract information about a random experiment from the (noisy) observation of the experiment outcomes. Classical Estimation Theory concerns problems in which the unknown quantity is a deterministic parameter (for example, the parameter of a distribution, the mean of a random variable), while in the Bayesian approach a distribution is assigned to it.

From a mathematical viewpoint, given two mutually dependent random variables, the aim is to guess some features of the first random variable (its realizations or its distribution parameters) given some realizations of the second one.

The LMSE has been easily derived, see, e.g., [124, Proposition 1] and is given by:

$$\widehat{X} = \mathbb{E}[X|y] \tag{2.59}$$

and the corresponding estimate, given an observed realization y of Y, is

$$\widehat{x} = \widehat{X}(y) = \mathbb{E}[X|Y = y]. \tag{2.60}$$

The LMSE is unbiased. that is, $\mathbb{E}[X - \widehat{X}|Y] = \mathbb{E}[X|Y] - \mathbb{E}[\mathbb{E}[X|Y]|Y] = 0$ hence

$$\mathbb{E}[X] = \mathbb{E}[\widehat{X}]. \tag{2.61}$$

It is interesting to notice that the LMSE has the following particular straightforward formulation when the random vectors X and Y are Gaussian:

$$\widehat{X} = m_X + \Sigma_{XY}\Sigma_{YY}^{-1}(Y - m_Y) \tag{2.62}$$

where m_X and m_Y are the respectively means and Σ_{XY} and Σ_{YY} the covariance matrices of (X, Y) and of Y.

In many cases, the computation of the LMSE is not so straightforward as in the Gaussian case. It can then be useful to restrict the family of possible estimators for the sake of the calculus and despite of optimality. For example, we can focus on linear estimators namely of kind $AY + b$, A and b being constant values of consistent dimensions. When X and Y are zero-mean random variables, it can be proved that [124, Proposition 3]:

$$\min_{x=Ay+b} \mathbb{E}[||X - x||^2|Y = y] = \Sigma_{XY}\Sigma_{YY}^{-1}y. \tag{2.63}$$

We will indicate the Linear Least Mean Squares estimator (LLMSE) by $\mathbb{E}^*(X|Y) = \Sigma_{XY}\Sigma_{YY}^{-1}Y$. using the notion of [124] (notice that the symbol $\mathbb{E}^*$ is not a mean). It turns out that $\mathbb{E}^*(X|Y)$ corresponds to (2.62) when X and Y are zero-mean, Gaussian variables.

2.5.8 The Kalman Filter

The Kalman Filter is a recursive algorithm that calculates the Linear Least Mean Squares estimate in problems that admit a linear, *dynamical* state-representation.

Let us imagine that X is a vector whose components $X_1, X_2, \ldots, X_n$ are (or can be considered as) the values of a random process at different time instants and suppose to gather synchronized measurements $y_1, y_2 \ldots, y_n$. In this framework, one could aim to perform an *on-line* estimation, namely to give an estimation of X_t at time $t \leq n$

based on the present available data $y_1, \dots, y_t$. This is a common problem in many applications, for example when sudden estimation is required (e.g., in fault detection) and more in general when X is a high-dimensional vector.

The Kalman Filter, which dates back to 1960, answers to this question. Its was introduced in [79] for discrete problems, while the continuous version was presented a year later [80]. Here, we review the discrete version, following the exposition of [29, Section 7.2].

Let us consider the following dynamical system given: $k \in \mathbb{N}$,

$$\begin{cases} X_{k+1} = \mathrm{A}_k X_k + \mathrm{B}_k U_k & \text{stochastic difference equation} \\ Y_k = \mathrm{C}_k X_k + N_k & \text{measurement equation} \end{cases} \tag{2.64}$$

where U_k, X_k and Y_K are n-dimensional random vectors and A_k, B_k and C_k are deterministic matrices with consistent dimensions. The aim is to compute the best linear estimate of X_k given $Y_0, \dots, Y_k$ by an iterative method, computing at each step k the mean and the covariance matrix of the a posteriori distribution of X_k.

First, let us recall that given any X and a sequence of measures $Y_0, \dots, Y_k$, the LLMSE of X conditioned to $Y_0, \dots, Y_k$ can be recursively computed as follows (see [124, Proposition 4]): letting $\mathbb{E}^*[X|Y_0, \dots, Y_k]$ be such LLMSE [1], then

$$\begin{aligned} &\mathbb{E}^*[X|Y_0, \dots, Y_k] = \\ &\quad = \mathbb{E}^*[X|Y_0, \dots, Y_{k-1}] + \mathbb{E}^*\big[X - \mathbb{E}^*[X|Y_0, \dots, Y_{k-1}]|Y_k \mathbb{E}^*[Y_k|Y_0, \dots, Y_{k-1}]\big]. \end{aligned} \tag{2.65}$$

Now, let

$$\begin{aligned} &\mathbb{E}[U_k] = \mathbb{E}[N_k] = 0 \text{ for any } k = 0, 1, 2, \dots; \\ &\mathrm{Cov}[U_k, U_h] = \mathbb{E}[U_k U_h^T] = \Sigma_{U,k}\delta_{k,h}; \\ &\mathrm{Cov}[N_k, N_h] = \mathbb{E}[U_k U_h^T] = \Sigma_{N,k}\delta_{k,h}; \\ &\mathrm{Cov}[N_k, U_h] = \mathrm{Cov}[N_k, X_h] = 0 \text{ for all } k, h \in \mathbb{N}; \\ &\widehat{X}_{k|j} = \mathbb{E}^*[X_k|Y_0, Y_1, \dots, Y_k]; \\ &P_{k|j} = \mathrm{Cov}[\widehat{X}_{k|j} - X_k, \widehat{X}_{k|j} - X_k]; \\ &K_k = P_{k|k-1}\mathrm{C}_k^T\big[\mathrm{C}_k P_{k|k-1}\mathrm{C}_k^T + \sigma_{N,k}^2\big]^{-1} \quad \text{(Kalman gain matrix);} \end{aligned}$$

$\widehat{X}_{k|j}$ is called filtered, predicted or smoothed estimate of X_k respectively if $k = h$, $k > h$ or $k < h$. Now, the following result can be proved:

Theorem 1 *([79],[29, Theorem 7.2.2])*
For any $k = 0, 1, 2 \dots$, $\widehat{X}_{k|k}$ can be recursively computed as follows:

$$\begin{aligned} &\widehat{X}_{k+1|k+1} = \mathrm{A}_k \widehat{X}_{k|k} + K_{k+1}\big[Y_{k+1} - \mathrm{C}_{k+1} A_k \widehat{X}_{k|k}\big] \\ &P_{k|k} = [I - K_k \mathrm{C}_k] P_{k|k-1} \textit{ (covariance update)}; \\ &P_{k+1|k} = \mathrm{A}_k P_{k|k} \mathrm{A}_k^T + \mathrm{B}_k \Sigma_{U,k} \mathrm{B}_k^T \textit{ (covariance extrapolation)}. \end{aligned} \tag{2.66}$$

[1] $\mathbb{E}^*[X|Y_0, \dots, Y_k]$ is not a mean operator, it is just an easy to remember notation for LLMSE, in antithesis to the LMSE $\mathbb{E}[X|Y_0, \dots, Y_k]$

The initialization of the recursion depends on the framework. In many cases (and this will be the assumption in the next of this work, see Chapter 4), the initial value of X is fixed, say $X(0) = x_0$, then $\widehat{X}_{0|0} = x_0$ and $P_{0|0} = 0$.

The Kalman Filter is low-complexity and straightforward to implement, which makes its use widespread for state estimation in linear systems. Its first applications being concerned with spacecraft navigation (it was used to estimate the trajectories of the spacecrafts in the Apollo Program, [61, 91]), has been applied to all forms (aerospace, land, marine) of navigation problems, target tracking, industrial control, economics and many other areas. Sometimes, it is also implemented as part of more structured algorithms, e.g., the Mendel-Kormylo Minimum Variance Deconvolution procedure [93, 83].

In the literature, many works can be retrieved that compare the scheme and the performance of Kalman Filter to other probabilistic methods, such as Wiener Filter [135] or early Gauss least squares [136], which highlight that if on one hand the achievement of the Kalman Filter is unquestionable, on the other one some of its good features, in terms of reliability and structure, can be retrieved in the past estimation methodologies.

In Chapter 4, the Least Mean Squares estimation and the Kalman Filter will be retrieved and compared with the novel methods introduced in this dissertation.

2.5.9 Maximum Entropy Principle

The concept of entropy was introduced in Information Theory in 1948 by Shannon [131], but was in use earlier in statistical mechanics; it attempts to measure the *uncertainty*, or disorder, associated with a certain probabilistic system. The definition in the discrete case is the following: given a random vector $X = (X_1, \ldots, X_n)$ with probability distribution $P_i = P(X_i)$, $i = 1, 2, \ldots, n$, its entropy is the function

$$H(X) = -\sum_{i=1}^{n} P_i \log P_i. \tag{2.67}$$

Analogous definition can be given for continuous random variables.

Let us consider a random variable X whose distribution is not known, but possibly subject to some given constraints; moreover, some realizations of it may be observable. According to the Maximum Entropy Principle, stated by Jaynes in 1957 [75], the *best* way to estimate the unknown distribution is to choose the distribution that satisfies the constraints and maximizes the entropy. This guarantees to retain all the uncertainty not removed from the observations' measurements and then to avoid artefacts. In other terms, the chosen distribution is “maximally non-committal with regard to the missing information” [42].

Both entropy and Maximum Entropy Principle can be considered more philosophical than mathematical notions. Their assessment has been subjected to many debates and criticisms in the last sixty years, even if their efficacy is generally recognized in Information and Probabilistic Theory. Shannon's entropy has been constructed fixing some properties and then finding *the* function, up to a multiplicative factor [131, Theorem 2 -Section 1.6]; changing the basic assumptions, different definitions can be given (see, e.g., [107]).

On the other hand, the Maximum Entropy Principle, which is widely used in many scientific areas and in particular in image processing [160, 63, 62], actually has not been proved as a theorem and discussions on its validity are still open (see, e.g., [29, Chapter 3]).

The concept of entropy is also involved in the definition of the so-called Kullback-Leibler (KL) information (or divergence or pseudo-distance, [85]), which is a *relative* entropy, say a "distance" between probabilities P_i and Q_i:

$$\mathrm{KL}(P, Q) = \sum_{i=1}^{N} P_i \log \frac{P_i}{Q_i}. \tag{2.68}$$

The definition can clearly be extended to measures in the continuous case. The KL information is often used in image reconstruction, where P and Q may be not probabilities, but non-negative quantities [150, 27]. An example is given by emission tomography: the KL pseudo-distance is computed between the counts of emissions performed by a detector and the expected number of counts and since the emissions are ruled by a Poisson distribution, it has been shown that maximum likelihood solution corresponds to the minimization of the KL [27].

Furthermore, KL information (or equivalent entropy) is often used as penalty term $\mathcal{G}$ in regularization methods, based on the minimization of functionals of kind $\mathcal{F}(Lx - y) + \mathcal{G}(x)$ (see (2.35). In some cases, deterministic regularization is approached from a Bayesian viewpoint using Gaussian distribution as prior probabilistic information on the solution (see, e.g., [57, 87] and the references therein).

2.6 Topical Bibliography

We conclude this introductory chapter by providing some references about deconvolution and inverse problems for the interested reader. The references are collected in Table 2.6 according to the applications they address and the solution approaches they use.

	Image Restoration	Tomography	Biological and Biomedical systems
LS			[35, 137, 20]
REG	[151, 97, 11]		[137, 116, 115, 119, 77]
ITER	[97]		
ML	[11]		[116, 115]
B-MAP	[127, 11]		[35, 137, 116, 115, 20]
EM	[127, 138, 154, 53, 11]	[132, 153, 60, 123, 168]	
WF			[35]
KF	[11]		[115]
ME	[56, 160, 63, 106, 87]	[62]	
WV	[11, 138, 53, 139]	[168]	
LMS		[150]	[137]
	Geophysics and Seismology	Astronomy and Aerospace	
LS	[128, 112]	[141]	
REG	[164]	[95, 140, 141]	
ITER		[141]	
ML	[83]	[140, 141]	
B-MAP	[133]	[95, 140, 141]	
EM		[138, 95, 140, 141, 19]	
WF	[125, 112, 15, 6, 7]		
KF	[15, 7, 133, 118, 83, 120]	[61], [91]	
ME		[108, 95, 140, 141]	
WV		[138, 139, 141]	
LMS	[93, 120]		
	Quantized Input: Fault Detection, Targets Tracking, Switching Systems	Algorithms and Theory	
LS		[136]	
REG	[47]	[113, 73, 148, 157]	
ITER	[47]	[48, 49, 47]	
ML	[162, 163, 13, 12, 104]	[5]	
B-MAP	[31, 12]	[34]	
EM		[37, 127]	
WF		[161, 135]	
KF	[162, 163, 13, 46, 12, 66, 142] [90, 134, 25], [23]	[79, 80, 135, 136, 124]	
ME		[85, 75]	
WV		[40, 59, 1, 158, 102, 58, 51, 103]	
LMS		[124]	

Table 2.1: Some references on different instances of deconvolution. LS = Least Squares, REG = Regularization Methods; ITER = Iterative Methods; ML = Maximum Likelihood; B-MAP = Bayes Methods, Maximum a Posteriori; EM = Expectation Maximization; WF = Wiener Filter; KF = Kalman Filter; ME = Maximum Entropy; WV = Wavelet Methods; LMS = Least Mean Squares.

Chapter 3

Problem statement and Algorithms

After having introduced the deconvolution problem (Chapter 2), let us now describe the instance we will study in this dissertation. The aim of this chapter is to

- present the mathematical model;
- explain the approach we will undertake;
- expose the algorithms we will use.

After that, in Chapters 4 and 5 we will discuss the implementation of the aforementioned algorithms, both with simulations and theoretical analysis.

3.1 The model

Let us consider the following input/output linear system

$$\begin{cases} x'(t) = \mathrm{A}x(t) + \mathrm{B}u(t) \qquad t \in [0, T] \\ y(t) = \mathrm{C}x(t) \\ x(0) = 0 \end{cases} \tag{3.1}$$

where $u(t) \in \mathbb{R}^q$, $x(t) \in \mathbb{R}^n$, $y(t) \in \mathbb{R}^m$, A, B and C are constant matrices with consistent dimensions, and $[0, T]$ is a (possibly infinite) time horizon. $u(t)$ is the input and is supposed to be unknown, $x(t)$ is the state function, $y(t)$ is the output.

Our aim is to reconstruct $u(t)$ from $y(t)$, that is, to reverse the input-output convolution integral:

$$y(t) = \mathrm{C}x(t) = \mathrm{C} \int_0^t e^{(t-s)\mathrm{A}} \mathrm{B}u(s)\mathrm{d}s. \tag{3.2}$$

This is what we name *deconvolution of linear systems.* In the next, we will stick to this problem under the following assumptions.

Assumption 1 *The available output signal is a sampled, noisy version of $y(t)$:*

$$y_k = \mathrm{C}x_k + n_k$$

where $x_k = x(\tau k)$ and $y_k = y(\tau k)$, $\tau > 0$ being the constant sampling time, and n_k is an additive observational noise.

Assumption 2 $K = T/\tau \in \mathbb{N}$.

Assumption 3 *The n_k's are realizations of independent, identically distributed Gaussian random variables N_k's of 0 mean and covariance matrix $\sigma^2\mathbb{I}$.*

Assumption 4 *The input $u(t)$ is piecewise constant and quantized, that is, we fix a finite alphabet $\mathcal{U} \subset \mathbb{R}^q$ such that*

$$u(t) = \sum_{k=0}^{K-1} u_k \mathbb{1}_{[k\tau,(k+1)\tau[}(t) \qquad u_k \in \mathcal{U}. \tag{3.3}$$

Notice that for simplicity we are assuming a perfect synchronization between input and output: the sampling time τ is the same.

Assumption 5 *The u_k's are realizations of i.i.d. random variables.*

3.2 Information-theoretic approach

Under Assumption 4, $u(t)$, $t \in [0, T[$, is completely determined by the sequence of samples $u_k \in \mathcal{U}$, $k = 0, \ldots, K-1$. As a consequence, the state function $x(t)$ is identified by $x_k = x(k\tau) \in \mathcal{X}$, $k = 0, \ldots, K$:

$$\begin{aligned} x_k &= \int_0^{k\tau} e^{(k\tau - s)\mathrm{A}}\mathrm{B} \sum_{h=0}^{K-1} u_h \mathbb{1}_{h\tau,(h+1)\tau[}(s)\mathrm{d}s \\ &= e^{k\tau\mathrm{A}} \sum_{h=0}^{k-1} \int_{h\tau}^{(h+1)\tau} e^{-s\mathrm{A}}\mathrm{B}u_h \mathrm{d}s \\ &= e^{k\tau\mathrm{A}} \sum_{h=0}^{k-1} \left(e^{-h\tau\mathrm{A}} - e^{-(h+1)\tau\mathrm{A}}\right) \mathrm{A}^{-1}\mathrm{B}u_h. \end{aligned} \tag{3.4}$$

Notice that

$$\begin{aligned} x_{k+1} &= e^{(k+1)\tau\mathrm{A}} \sum_{h=0}^{k} \left(e^{-h\mathrm{A}} - e^{-(h+1)\mathrm{A}}\right) \mathrm{A}^{-1}\mathrm{B}u_h \\ &= e^{(k+1)\tau\mathrm{A}} \left(\sum_{h=0}^{k-1} \left(e^{-h\mathrm{A}} - e^{-(h+1)\mathrm{A}}\right) \mathrm{A}^{-1}\mathrm{B}u_h + \left(e^{-k\tau\mathrm{A}} - e^{-(k+1)\tau\mathrm{A}}\right) \mathrm{A}^{-1}\mathrm{B}u_k\right) \\ &= e^{\tau\mathrm{A}}x_k + \left(e^{\tau\mathrm{A}} - \mathbb{I}\right) \mathrm{A}^{-1}\mathrm{B}u_k. \end{aligned} \tag{3.5}$$

Defining

$$\mathrm{Q} := e^{\tau A}, \qquad \mathrm{W} := \left(e^{\tau \mathrm{A}} - \mathbb{I}\right) \mathrm{A}^{-1}\mathrm{B} \tag{3.6}$$

we can write the following recursive formula:

$$x_{k+1} = \mathrm{Q}x_k + \mathrm{W}u_k. \tag{3.7}$$

Since $x_0 = 0$,

$$x_k = \sum_{h=0}^{k-1} \mathrm{Q}^h \mathrm{W} u_{k-h-1}. \tag{3.8}$$

Hence, for any $k \in \mathbb{N}$

$$x_k \in \mathcal{X} = \left\{ \sum_{h=0}^{\infty} \mathrm{Q}^h \mathrm{W} \mu_h. \;\; \mu_h \in \mathcal{U} \cup \{0\} \subseteq \mathbb{R}^q \right\}. \tag{3.9}$$

Given this discrete setting, it is natural to interpret our deconvolution problem as a digital transmission paradigm. Let us recall that a standard digital transmission is composed by the following elements: an information message, i.e., a sequence of symbols that one aims to transmit to a receiver; a suitable *encoded* version of the input message; a noisy transmission channel over which the encoded message is sent; a decoder, i.e., the device that reads the output of the channel and recovers the information message. Typically, the symbols composing the input arise from a finite alphabet, said *source alphabet.* The decoder knows the source alphabet and recovers the information message within it.

Now, our deconvolution problem can be thought in these terms:

1. $(u_0, \dots, u_{K-1}) \in \mathcal{U}^K$ is an information message;

2. Convolution corresponds to encoding: $(x_1, \dots, x_K) \in \mathcal{X}^K$ is an encoded version of $(u_0, \dots, u_{K-1})$, the encoding procedure being defined by (3.7);

3. $y_k = \mathrm{C}x_k + n_k$, $k = 1, \dots, K$ can be interpreted as the passage of $(\mathrm{C}x_1 \dots, \mathrm{C}x_K)$ through an additive-noise channel (C is an amplification/compression factor);

4. Deconvolution corresponds to decoding.

Under the Assumption 3, the considered channel actually is a so-called Additive White Gaussian Noise (AWGN for short) channel, which is commonly used in communication models.

The only one difference between a typical digital transmission and our problem lies in the encoding philosophy. In digital transmissions, in fact, encoding is expressly conceived to improve the communication reliability; in our case, instead, encoding is imposed by the system itself. Hence, while in digital transmissions one designs both the encoding and the decoding schemes in order to recover the information message as well as possible, in our framework we can design only the decoding scheme.

The fact that deconvolution can be thought of as a decoding in the framework of quantized-input linear systems suggests us to *perform deconvolution using a decoding algorithm*: this will be our focus in the next. Before introducing the decoding procedures envisaged in this work, let us explain more in detail what a deconvolution algorithm is expected to do in our framework and how we can evaluate its performance.

3.2.1 Deconvolution and Decoding

In the next, we will denote by $\mathbf{y} = (y_1, \dots, y_K) \in \mathbb{R}^{m \times K}$ the vector of all available measures and by $\mathbf{y}_a^b = (y_a, y_{a+1}, \dots, y_b)$ the available measures from time a to time b, with $a, b \in \{1, \dots, K\}$, $a < b$.

In our framework, a *decoding algorithm* consists in a function $\mathbb{D}$, named *decoder*, such that

$$\begin{aligned} \mathbb{D} : \mathbb{R}^{m \times K} &\to \mathcal{U}^K \\ \widehat{\mathbf{u}} = \mathbb{D}(\mathbf{y}) &\in \mathcal{U}^K . \end{aligned} \tag{3.10}$$

$\widehat{\mathbf{u}}$ is the estimated input which in general will not coincide with the true input u, but is expected to fulfill some consistency property: when the variance of the noise and the sampling time go to 0, the error should converge (in some suitable sense) to 0.

Notice that classical deconvolution algorithms [148, 149] applied to our problem would produce estimates lying in $\mathbb{R}^{q \times K}$. This is the conceptual difference between decoding and classical deconvolution. In other terms, decoders are a particular class of deconvolution algorithms which exploit the prior information about the source alphabet and force the estimation of the input to be composed by symbols arising from that alphabet.

We say that a decoding algorithm is *causal* (with bounded delay $k_0\tau$. $k_0 \in \mathbb{N}$) if there exists a sequence of functions $\mathbb{D}_k : \mathbb{R}^{m \times k + k_0} \to \mathcal{U}$, $k = 1, 2, \dots, K$, such that $\widehat{u}_{k-1} = \mathbb{D}_k(\mathbf{y}_1^{k+k_0})$, $\widehat{u}_{k-1}$ being an estimate of u_{k-1}. Such a decoder estimates the unknown input signal in the current time interval $[(k-1)\tau, k\tau[$ exploiting the past and present information $\mathbf{y}_1^k$ and, in case, the future information $y_{k+1}^{k+k_0}$.

3.2.2 Probabilistic setting

Given the Assumptions 3 and 5, all the system assumes a probabilistic nature. In the sequel, we will use capital letters to indicate random variables: U_k will identify the input r.v. at time k, X_k the corresponding system function and $Y_k = \mathrm{C}X_k + N_k$, N_k being the Gaussian noise. Furthermore, $\widehat{U}_k = \mathbb{D}(\mathbf{Y})_k$ and $\widehat{X}_k = \mathrm{Q}\widehat{X}_{k-1} + \mathrm{W}\widehat{U}_{k-1}$ ($\widehat{X}_0 = 0$) will be respectively the estimated input and the estimated state. Finally, $\mathbf{U} = (U_0, \dots, U_{K-1})$, $\widehat{\mathbf{U}} = (\widehat{U}_0, \dots, \widehat{U}_{K-1})$, $\mathbf{Y} = (Y_1, \dots, Y_K)$, $\mathbf{Y} = (Y_1, \dots, Y_K)$, $\mathbf{Y}_a^b = (Y_a, \dots, Y_b)$, $a, b \in \{1, \dots, K\}$, $a < b$.

3.2.3 Performance Evaluation: The Mean Square Error

A fundamental issue in any deconvolution or decoding problem is the choice of the norm with respect to which errors are evaluated. In our context, we consider the Mean

Square Error (MSE):

$$\mathrm{MSE}(\mathbb{D}) = \sum_{k=0}^{K-1} \mathbb{E}\left[||U_k - \widehat{U}_k||^2_{\mathbb{R}^q}\right]$$

where $\widehat{U}_k = \mathbb{D}(\mathbf{Y})_k \in \mathcal{U}$. We now define $\mathbb{D}^*$ as the decoder minimizing $\mathrm{MSE}(\mathbb{D})$ among all the possible decoders. It can be constructed as follows: given the density $f_{\mathbf{Y}}(\mathbf{y})$ of $\mathbf{Y}$, notice that

$$\mathrm{MSE}(\mathbb{D}) = \sum_{k=0}^{K-1} \int_{\mathbb{R}^{m\times K}} \mathbb{E}\big[||U_k - \mathbb{D}(\mathbf{y})_k||^2_{\mathbb{R}^q}|\mathbf{Y} = \mathbf{y}\big] f_{\mathbf{Y}}(\mathbf{y})\mathrm{d}\mathbf{y}.$$

Hence, for any $\mathbf{y} \in \mathbb{R}^{m\times K}$,

$$\mathbb{D}^*(\mathbf{y})_k = \operatorname*{argmin}_{v\in\mathcal{U}} \mathbb{E}\big(||U_k - v||^2_{\mathbb{R}^q}|\mathbf{Y} = \mathbf{y}\big) = \operatorname*{argmin}_{v\in\mathcal{U}} \sum_{u\in\mathcal{U}} ||u - v||^2_{\mathbb{R}^q} \mathrm{P}(U_k = u|\mathbf{Y} = \mathbf{y}). \tag{3.11}$$

This turns out to be a finite optimization problem which can be solved by means of a marginalization procedure and a Bayesian inversion:

$$\mathrm{P}(U_k = u|\mathbf{Y} = \mathbf{y}) = \sum_{\mathbf{u}\in\mathcal{U}^K:\ u_k=u} \frac{f_{(\mathbf{Y}|\mathbf{X})}(\mathbf{y}|\mathcal{E}(\mathbf{u}))\mathrm{P}(\mathbf{U} = \mathbf{u})}{f_{\mathbf{Y}}(\mathbf{y})} \tag{3.12}$$

where $\mathcal{E}$ indicates the encoding function. Analogously, we can define $\mathbb{D}^{*k_0}$ as the decoder minimizing $\mathrm{MSE}(\mathbb{D})$ among all the possible causal decoders with delay k_0:

$$\mathbb{D}^{*k_0}(\mathbf{y})_k = \mathbb{D}^{*k_0}_k(\mathbf{y}_1^{k+1+k_0}) = \operatorname*{argmin}_{v\in\mathcal{U}} \sum_{u\in\mathcal{U}} ||u - v||^2_{\mathbb{R}^q} \mathrm{P}(U_{k-1} = u|\mathbf{Y}_1^{k+1+k_0} = \mathbf{y}_1^{k+1+k_0}) \tag{3.13}$$

where

$$\begin{aligned}&\mathrm{P}(U_k = u|\mathbf{Y}_1^{k+1+k_0} = \mathbf{y}_1^{k+1+k_0}) = \\ &= \sum_{\mathbf{u}\in\mathcal{U}^K:\ u_k=u} \frac{f_{(\mathbf{Y}_1^{k+1+k_0}|\mathbf{X}_1^{k+1+k_0})}(\mathbf{y}_1^{k+1+k_0}|\mathcal{E}(\mathbf{u}_0^{k+k_0}))\mathrm{P}(\mathbf{U}_0^{k+k_0} = \mathbf{u})}{f_{\mathbf{Y}_1^{k+1+k_0}}(\mathbf{y}_1^{k+1+k_0})}.\end{aligned} \tag{3.14}$$

3.3 Decoding Algorithms

The Bayesian inversions in (3.12) and (3.14) are numerically complex for large K. However, we can compute them respectively through the well-known BCJR algorithm [10] and a causal version of the BCJR.

The first aim of this section is then to introduce the BCJR and the causal BCJR schemes; afterwards, we will derive from them two low-complexity causal algorithms, named One State and Two States. These four decoding algorithms will be the ones implemented in the sequel of the dissertation to perform deconvolution.

3.3.1 The BCJR algorithm

Based on a forward-backward recursive procedure, the BCJR computes the a posteriori probabilities (APP) on states and state transitions of a Markov source, given the observed channel outputs.

In our framework, given the encoding rule $x_{k+1} = \mathrm{Q}x_k + \mathrm{W}u_k$, the states are $x_1, \dots, x_K$, while the state transitions are given by $u_0, \dots, u_{K-1}$; the observed channel outputs are $y_1, \dots, y_K$. Our interest is then to evaluate the APP, on the state transitions given $y_1, \dots, y_K$: these APP actually provide a soft decision on the input sequence.

Let us briefly remind the BCJR procedure.

BCJR Algorithm - Decoder $\mathbb{D}^*$

For $i, j \in \mathcal{X}$, we define the following probability density functions:

$$\begin{aligned} \alpha_k(i) &= f_{(X_k, \mathbf{Y}_1^k)}(i, \mathbf{y}_1^k) && k = 1, \dots, K \\ \beta_k(i) &= f_{(\mathbf{Y}_{k+1}^K | X_k)}(\mathbf{y}_{k+1}^K | i) && k = 0, \dots, K-1 \\ \Gamma_k(i,j) &= f_{(X_k, Y_k | X_{k-1})}(j, y_k | i) && k = 1, \dots, K. \end{aligned} \tag{3.15}$$

For any $k = 1, \dots, K$, the APP state transitions are obtained dividing:

$$\sigma_k(i,j) = f_{(X_k, X_{k-1}, \mathbf{Y})}(j, i, \mathbf{y}).$$

by $f_{\mathbf{Y}}(\mathbf{y})$. Given the following initial and final conditions:

$$\alpha_0(i) = \mathrm{P}(X_0 = i) = \begin{cases} 1 & \text{if } i = 0 \\ 0 & \text{otherwise.} \end{cases}$$

$$\beta_K(i) = 1 \text{ for any } i \in \mathcal{X}$$

for $k = 1, \dots, K$ we have

$$\sigma_k(i,j) = \alpha_{k-1}(i)\Gamma_k(i,j)\beta_k(j) \tag{3.16}$$

where $\alpha_k(i)$ and $\beta_k(i)$, $i \in \mathcal{X}$, can be respectively computed with a forward and a backward recursions:

$$\alpha_k(i) = \sum_{h \in \mathcal{X}} \alpha_{k-1}(h)\Gamma_k(h,i) \qquad \beta_k(i) = \sum_{h \in \mathcal{X}} \Gamma_{k+1}(i,h)\beta_{k+1}(h). \tag{3.17}$$

At this point,

$$\mathrm{P}(U_k = u | \mathbf{Y} = \mathbf{y}) = \frac{1}{f_{\mathbf{Y}}(\mathbf{y})} \sum_{i \in \mathcal{X}} \sigma_k(i, \mathrm{Q}i + \mathrm{W}u). \tag{3.18}$$

Substituting (3.18) in (3.11), we obtain $\mathbb{D}^*$.

Though not used in this work, we remind that the APP on the states can be computed as follows: given $\lambda_k(i) := f_{(X_k,\mathbf{Y})}(i,\mathbf{y})$, we have $\lambda_k(i) = \alpha_k(i)\beta_k(i)$ and $\mathrm{P}(X_k = i|\mathbf{Y} = \mathbf{y}) = \frac{\lambda_k(i)}{f_{\mathbf{Y}}(\mathbf{y})}$.

3.3.2 The Causal BCJR

Causal versions of the BCJR algorithm can be used to implement the decoder (3.13) with a bounded delay k_0. For $k = 1,\dots,K-k_0$,

$$\widetilde{\sigma}_k(i,j) = f_{(X_k,X_{k-1},\mathbf{Y}_1^{k+k_0})}(j,i,\mathbf{y}_1^{k+k_0}) = \alpha_{k-1}(i)\Gamma_k(i,j)\widetilde{\beta}_k(j) \tag{3.19}$$

where α_k and Γ_k are defined as above, while $\widetilde{\beta}_k(j) = f_{(\mathbf{Y}_{k+1}^{k+k_0}|X_k)}(\mathbf{y}_{k+1}^{k+k_0}|j)$. For $k > K-k_0$, we reduce to the classical formulation (3.16). We name CBCJR the purely causal BCJR algorithm, say with $k_0 = 0$.

CBCJR Algorithm - Decoder $\mathbb{D}^{*_0}$

Given $\alpha_k(i) = f_{(X_k,\mathbf{Y}_1^k)}(i,\mathbf{y}_1^k)$, $\Gamma_k(i,j) = f_{(X_k,Y_k|X_{k-1})}(j,y_k|i)$, $k = 1,\dots,K$ and the update rule $\alpha_k(i) = \sum_{h\in\mathcal{X}}\alpha_{k-1}(h)\Gamma_k(h,i)$, we compute

$$\widetilde{\sigma}_k(i,j) = f_{(X_k,x_{k-1},\mathbf{Y}_1^k)}(j,i,\mathbf{y}_1^k) = \alpha_{k-1}(i)\Gamma_k(i,j) \tag{3.20}$$

Thus

$$\mathrm{P}(U_k = u|\mathbf{Y}_1^k) = \frac{1}{f_{\mathbf{Y}}(\mathbf{y})}\sum_{(i,j)\in\mathcal{S}_u}\widetilde{\sigma}_k(i,j). \tag{3.21}$$

The CBCJR is optimal with respect to causal algorithms. However, causality has a price and the CBCJR algorithm clearly has worse performance than BCJR.

On the other hand, by comparing the efficiency of the two procedures, we gather that for both BCJR and CBCJR the required computations and storage locations increase with the number of transmitted bits, which is a drawback in case of long transmission.

This fact motivates the development of new suboptimal causal algorithms that improve the efficiency without substantial loss of reliability. To achieve that, we implement the CBCJR fixing the number of states, that is, at each step we save the n states with largest probability (where n is arbitrarily chosen) and we discard the others.

The algorithms in the cases $n = 1$ and $n = 2$, which are of great interest for their low complexity, are now introduced.

3.3.3 One State Algorithm

A suboptimal causal decoder

$$\mathbb{D}^{(1)} : \mathbb{R}^{m\times K} \to \mathcal{U}^K$$

can be derived from the CBCJR by assuming the most probable state to be the correct one. At any step $k = 0, 1, \dots,$ $\mathbb{D}^{(1)}$ decides on the current symbol by a single MAP procedure and upgrades the estimated state, which is the only one value that requires to be stored.

Consider (3.15), (3.19) and (4.12). Given the estimated state $\widehat{x}_{k-1}$, the decoding rule of $\mathbb{D}^{(1)}$ at time step k is given by (4.12) with no backward recursion $\widetilde{\beta}_k(j)$ and $\alpha_{k-1}(\widehat{x}_{k-1}) = 1$, $\alpha_{k-1}(j) = 0$ for any $j \neq \widehat{x}_{k-1}$. This reduces the decoding task to the comparison between two distances; in fact, the One State Algorithm (OSA for short) that implements $\mathbb{D}^{(1)}$ is as follows:

OSA - Decoder $\mathbb{D}^{(1)}$

Initialization: $\widehat{x}_0 = 0$.

For $k = 1, \dots, K$, given the received symbol $r_k \in \mathbb{R}^m$,

$$\begin{aligned} \widehat{u}_{k-1} &= \mathbb{D}^{(1)}(\mathbf{y})_{k-1} = \operatorname*{argmax}_{u \in \mathcal{U}} \mathrm{P}(U_{k-1} = u | Y_k = y_k, X_{k-1} = \widehat{x}_{k-1}) \\ &= \operatorname*{argmax}_{u \in \mathcal{U}} \Gamma_k(\widehat{x}_{k-1}, \mathrm{Q}\widehat{x}_{k-1} + \mathrm{W}u) \qquad (3.22) \\ \widehat{x}_k &= \mathrm{Q}\widehat{x}_{k-1} + \mathrm{W}\widehat{u}_{k-1}. \end{aligned}$$

Notice that we have used the fact that

$$\mathrm{P}(U_{k-1} = u | Y_k = y_k, X_{k-1} = \widehat{x}_{k-1}) = \mathrm{P}(X_k = \mathrm{Q}\widehat{x}_{k-1} + \mathrm{W}u | Y_k = y_k, X_{k-1} = \widehat{x}_{k-1}) \\ \frac{1}{f_{Y_k}(y_k)} \Gamma_k(\widehat{x}_{k-1}, \mathrm{Q}\widehat{x}_{k-1} + \mathrm{W}u). \qquad (3.23)$$

3.3.4 Two States Algorithm

By fixing $n = 2$, we derive a decoder

$$\mathbb{D}^{(2)} : \mathbb{R}^{m \times K} \to \mathcal{U}^K$$

that, at each step, estimates the current input bit and computes and stores the two most likely states along with the corresponding probabilities $\alpha_k(i)$ (defined by (3.15)). As for the One State Algorithm, the estimation of the input bit is performed by a MAP decoding rule (4.12) with no backward recursion and summing over the two “surviving” states. In detail, the recursive Two States Algorithm (TSA for short) that implements $\mathbb{D}^{(2)}$ is the following:

TSA - Decoder $\mathbb{D}^{(2)}$

For $k=1$, given the unique starting state $\widehat{x}_0 = 0 \in \mathbb{R}^n$, we estimate the first bit by a One State procedure:

$$\widehat{u}_0 = \mathbb{D}^{(2)}(\mathbf{y})_0 = \operatorname*{argmax}_{u \in \mathcal{U}} \mathrm{P}(U_0 = u | Y_1 = y_1, X_0 = 0). \tag{3.24}$$

Now, we have $|\mathcal{U}|$ possible states $\mathcal{X}_1 = \{\mathrm{W}u, \quad u \in \mathcal{U}\}$ and

$$\alpha_1(j) = f_{(X_1,Y_1)}(j, y_1) = f_{(Y_1|X_1)}(y_1|j)\mathrm{P}(X_1 = j) \quad j \in \{\mathrm{W}v_n, \ v_n \in \mathcal{U}\}.$$

We define

$$\begin{aligned} \widehat{x}_1(1) &= \operatorname*{argmax}_{j \in \mathcal{X}_1} \alpha_1(j) \\ \widehat{x}_1(2) &= \operatorname*{argmax}_{j \in \mathcal{X}_1 \setminus \{\widehat{x}_1(1)\}} \alpha_1(j). \end{aligned} \tag{3.25}$$

We normalize

$$\begin{aligned} \alpha_1^*(1) &= \frac{\alpha_1(\widehat{x}_1(1))}{\alpha_1(\widehat{x}_1(1)) + \alpha_1(\widehat{x}_1(2))} \\ \alpha_1^*(2) &= 1 - \alpha *_1 (2). \end{aligned} \tag{3.26}$$

We store $\widehat{x}_1(1), \widehat{x}_1(2), \alpha_1^*(1), \alpha_1^*(2)$ and we discard any other information.

For $k = 2, 3, \dots, K$, given $\widehat{x}_{k-1}(1), \widehat{x}_{k-1}(2), \alpha_{K-1}^*(1), \alpha_{K-1}^*(2)$:

$$\begin{aligned} &\widehat{u}_{k-1} = \mathbb{D}^{(2)}(\mathbf{y})_{k-1} = \\ &= \operatorname*{argmax}_{u \in \mathcal{U}} \mathrm{P}\big(U_{k-1} = u | Y_k = y_k, \widehat{X}_{k-1}(1) = \widehat{x}_{k-1}(1), \widehat{X}_{k-1}(2) = \widehat{x}_{k-1}(2)\big) = \\ &= \operatorname*{argmax}_{u \in \mathcal{U}} \{\alpha_{k-1}^*(1)\Gamma_k(\widehat{x}_{k-1}(1), \mathrm{Q}\widehat{x}_{k-1}(1) + \mathrm{W}u), \alpha_{k-1}^*(2)\Gamma_k(\widehat{x}_{k-1}(2), \mathrm{Q}\widehat{x}_{k-1}(2) + \mathrm{W}u)\}. \end{aligned}$$

Now, we have $2|\mathcal{U}|$ possible states $\mathcal{X}_k = \{\mathrm{Q}\widehat{x}_1(1) + \mathrm{W}u, \ \mathrm{Q}\widehat{x}_1(2) + \mathrm{W}u, \quad u \in \mathcal{U}\}$ and

$$\alpha_k(j) = f_{(X_k,Y_k)}(j, y_k) = f_{(Y_k|X_k)}(y_k|j)\mathrm{P}(X_k = j) \quad j \in \mathcal{X}_k.$$

We define

$$\begin{aligned} \widehat{x}_k(1) &= \operatorname*{argmax}_{j \in \mathcal{X}_k} \alpha_k(j) \\ \widehat{x}_k(2) &= \operatorname*{argmax}_{j \in \mathcal{X}_k \setminus \{\widehat{x}_k(1)\}} \alpha_k(j). \end{aligned} \tag{3.27}$$

We normalize

$$\begin{aligned} \alpha_k^*(1) &= \frac{\alpha_k(\widehat{x}_k(1))}{\alpha_k(\widehat{x}_k(1)) + \alpha_k(\widehat{x}_k(2))} \\ \alpha_k^*(2) &= 1 - \alpha_k(2). \end{aligned} \tag{3.28}$$

We store $\widehat{x}_k(1), \widehat{x}_k(2), \alpha_k(1), \alpha_k(2)$ and discard any other information.

Remark 1 *When the extreme case $\alpha_k(1) = 1$ occurs and $\widehat{x}_k(2)$ has null probability, the Two States Algorithm actually behaves as the One State Algorithm.*

Remark 2 *OSA and TSA are conceivable also if the state space $\mathcal{X}$ is infinite, as we force the number of surviving states to be finite. This makes them implementable also in the case of $K \to \infty$.*

3.4 Outline

After having introduced the model, the Information-theoretic approach and the decoding algorithms, we are ready to study in detail some instances of quantized-input linear systems. In particular, in the next we will always consider a *binary* quantization, which is the simplest case, but envisaging the main difficulties arising from quantization.

In Chapter 4, we will focus on the one-dimensional differentiation problem $x'(t) = u(t)$. We will implement all the decoding algorithms presented above and provide a complete theoretical description (in terms of Markov Processes) of the OSA and TSA procedures, which allows to study analytically their performance. We will pay particular attention to the asymptotic case $K \to \infty$.

In Chapter 5, we will extend the study to the one-dimensional linear problem $x'(t) = ax(t)+bu(t)$, $a, b \in \mathbb{R}$, under the stability condition $a < 0$. As we will see, the extension is not straightforward, due to some basic differences in the mathematical structure of the state space.

Finally, Chapter 6 will be devoted to a three-dimensional Fault Tolerant Control problem.

Chapter 4

The Differentiation Problem

In this chapter, we consider the system (3.1) with Assumptions 1-5, in one dimension ($q = n = m = 1$) and with A = 0, B = 1, C = 1. In other terms, we study the case when deconvolution is a differentiation problem. As already noticed in Chapter 2, differentiation is a typical inverse problem which is strongly affected by disturbances and measurements' inaccuracies.

The chapter is organized as follows. In Section 4.1, we describe the model, then in Section 4.2 we adapt to it the algorithms proposed in Chapter 3. Afterwards, we theoretically analyze the algorithms in the framework of Markov Processes (Sections 4.3 and 4.4), and finally we compare simulations and theoretical results.

4.1 Problem Statement

Let us consider the system

$$x(t) = \int_0^t u(s)\mathrm{d}s \qquad t \in [0, T] \tag{4.1}$$

where $u(t)$ and $x(t)$ are real functions ($u(t)$ is assumed to be integrable). The inverse system can be written as follows:

$$\begin{cases} x'(t) = u(t) \\ x(0) = 0 \end{cases} \tag{4.2}$$

Let us suppose that some additive noise affects the output, that is, the observable function is

$$y(t) = x(t) + n(t). \tag{4.3}$$

It is well known that the operation of differentiation is not robust with respect to noise perturbation, then the reconstruction of u from y cannot be simply done by differentiation. The goal is then to estimate u, using the available information on x and any a priori information on u. Several procedures can be exploited to accomplish this task and the choice is in general motivated by a suitable trade-off between precision of the solution and complexity of the algorithm.

4.1.1 Further Assumptions and Mean Square Error

In the next, we will stick to the problem (4.2) with Assumptions 1-5 and with the further specifications:

Assumption 6 *The input alphabet is binary:* $\mathcal{U} = \{0, 1\}$.

Assumption 7 *For* $k = 0 \dots K-1$*, the* U_k*'s are independent and uniformly distributed:* $\mathrm{P}(U_k = 0) = \mathrm{P}(U_k = 1) = \frac{1}{2}$*. In particular, the* U_k*'s are independent from the Gaussian noises* N_k*'s.*

Now the probabilistic setting is complete and we can resume the system as follows: given $X_0 = \widehat{X}_0 = 0$, for $k = 1, \dots, K$,

$$\begin{cases} U_{k-1} \sim \text{ Bernoulli } (1/2)\,; \\ X_k = X_{k-1} + \tau U_{k-1}; \\ N_k \sim \mathcal{N}(0, \sigma^2); \\ Y_k = X_k + N_k; \\ \widehat{U}_{k-1} = \mathbb{D}(\mathbf{Y})_{k-1}; \\ \widehat{X}_k = \widehat{X}_{k-1} + \tau \widehat{U}_{k-1}. \end{cases} \tag{4.4}$$

Notice that also X_k's are independent from N_k's.

In this one-dimensional setting, the Mean Square Error is given by

$$\mathrm{MSE}(\mathbb{D}) = \mathbb{E}\left(||\mathbf{U} - \widehat{\mathbf{U}}||^2\right) = \sum_{k=0}^{K-1} \mathbb{E}\left(|U_k - \widehat{U}_k|^2\right)$$

where $\widehat{\mathbf{U}} = \mathbb{D}(\mathbf{U})$. Under Assumption 6,

$$\mathrm{MSE}(\mathbb{D}) = \sum_{k=0}^{K-1} \mathbb{E}\left(|U_k - \widehat{U}_k|\right) = K\mathrm{BER}(\mathbb{D})$$

where

$$\mathrm{BER}(\mathbb{D}) = \frac{1}{K} \sum_{k=0}^{K-1} \mathrm{P}(\widehat{U}_k \neq U_k) = \frac{1}{K} \mathbb{E}(|\mathbf{U} - \widehat{\mathbf{U}}|) \tag{4.5}$$

is the Bit Error Rate (BER for short), a very common performance measure in digital transmissions that expresses the average number of bits in error. In our context, minimizing MSE($\mathbb{D}$) is equivalent to minimizing the BER($\mathbb{D}$) and, therefore, the optimal decoder $\mathbb{D}^*$ that performs this minimization coincides with the well-known Bit-MAP (Maximum a posteriori) decoder (see [126, 10]):

$$\mathbb{D}^*(\mathbf{y})_k = \underset{u \in \{0,1\}}{\operatorname{argmax}}\ \mathrm{P}(U_k = u | \mathbf{Y} = \mathbf{y}). \tag{4.6}$$

Its causal version is given by

$$\mathbb{D}^{*k_0}(\mathbf{y})_k = \underset{u\in\{0,1\}}{\operatorname{argmax}}\ \mathrm{P}(U_k = u|\mathbf{Y}_1^{k+1+k_0} = \mathbf{y}_1^{k+1+k_0}). \tag{4.7}$$

We introduce here also the Conditional Bit Error Rate, CBER for short:

$$\mathrm{CBER}(\mathbb{D}|\mathbf{U}) = \frac{1}{K}\sum_{k=0}^{K-1} \mathrm{P}(\widehat{U}_k \neq U_k|\mathbf{U}) = \frac{1}{K}\mathbb{E}(|\mathbf{U}-\widehat{\mathbf{U}}|\ |\mathbf{U}). \tag{4.8}$$

While the BER is a parameter that evaluates the *mean* performance of the transmission model, the CBER describes its behavior for *each* possible sent sequence. The CBER is then a relevant parameter for our system, whose decoding performance changes in function of the transmitted input.

For computational simplicity, from now onwards let

$$\tau = 1 \tag{4.9}$$

so that $\mathcal{X} = \{0,\ldots,K\}$ and in particular, if $X_0 = 0$, $X_k \in \{0,\ldots,k\}$. In the BCJR implementation of decoders (4.6) and (4.7), we obtain that $\alpha_k(i)$, $i = 0,1,\ldots,K$, is null for any $i > k$, while matrices Γ_k and σ_k are non-null only on diagonal and superdiagonal. By Assumption 7, $\mathrm{P}(X_k = j|X_{k-1} = i) = 1/2$ if $j = i, i+1$ and 0 otherwise. Recalling that the transition between X_k and Y_k is modeled by an AWGN channel, $f_{(Y_k|X_k)}(y_k|j) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(y_k-j)^2}{2\sigma^2}\right)$, we obtain

$$\begin{aligned}\Gamma_k(i,j) &= f_{(Y_k|X_k)}(y_k|j)\mathrm{P}(X_k = j|X_{k-1} = i)\\ &= \frac{1}{2\sigma\sqrt{2\pi}}\exp\left(-\frac{(y_k-j)^2}{2\sigma^2}\right) \quad \text{for } j = i, i+1.\end{aligned} \tag{4.10}$$

Given Γ_k, σ_k or its causal version $\widetilde{\sigma}_k$ can be recursively computed and the corresponding decoding rules are:

$$\text{BCJR} \qquad \mathbb{D}^*(\mathbf{y})_{k-1} = \begin{cases} 0 & \text{if } \sum_{i=0}^{k-1}\sigma_k(i,i+1) \leq \sum_{i=0}^{k-1}\sigma_k(i,i)\\ 1 & \text{otherwise.}\end{cases} \tag{4.11}$$

$$\text{CBCJR} \qquad \mathbb{D}^{*k_0}(\mathbf{y})_{k-1} = \begin{cases} 0 & \text{if } \sum_{i=0}^{k-1}\widetilde{\sigma}_k(i,i+1) \leq \sum_{i=0}^{k-1}\widetilde{\sigma}_k(i,i)\\ 1 & \text{otherwise.}\end{cases} \tag{4.12}$$

In the Appendix 4.7.9 we show that the CBCJR procedure actually is a causal LMSE (see Section 2.5.7).

4.2 Suboptimal Causal Decoding Algorithms

From simulations, we evaluate the performance gap between BCJR and CBCJR ($k_0 = 0$) as depicted in Figure 4.1: the two curves represent the corresponding BER's in

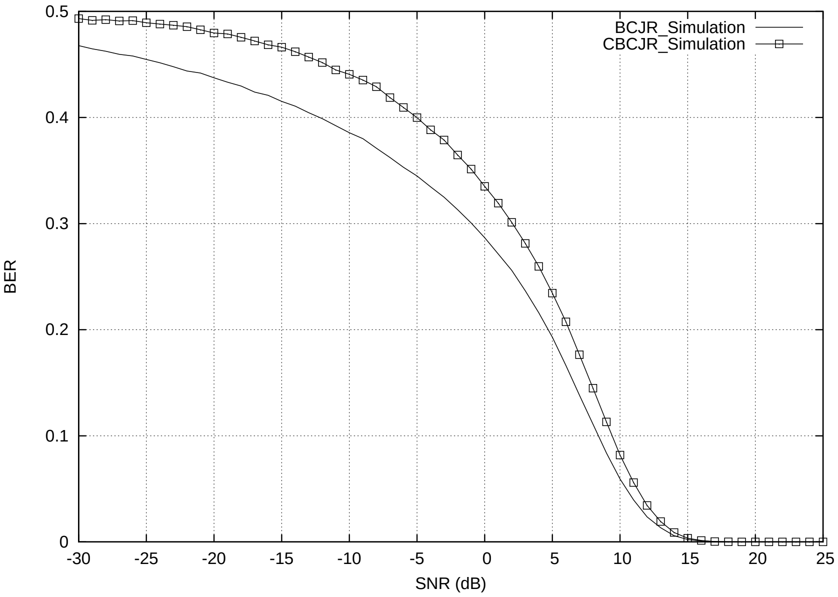

Figure 4.1: BCJR vs CBCJR.

	Computations	Storage Locations	Decoding Delay
BCJR	$O(K^2)$	$O(K^2)$	K
CBCJR	$O(K^2)$	$O(K^2)$	$k_0 = 0$

Table 4.1:

function of the Signal-to-Noise Ratio (SNR), here defined as $\tau^2/\sigma^2 = 1/\sigma^2$ and show the loss of reliability due to causality. These outcomes are the averages over 5000 transmissions of 100 bit messages. On the other hand, as said in Chapter 3, the loss of reliability is less dramatic than the high computational complexity that affects both BCJR and CBCJR in case of long time transmissions. This is why, in the next, we implement also the suboptimal causal algorithms OSA and TSA, introduced in Chapter 3 and now revised in the actual framework.

4.2.1 One State Algorithm

In the one-dimensional, binary input, differentiation case, the OSA has a very straightforward pattern:

OSA - Decoder $\mathbb{D}^{(1)}$

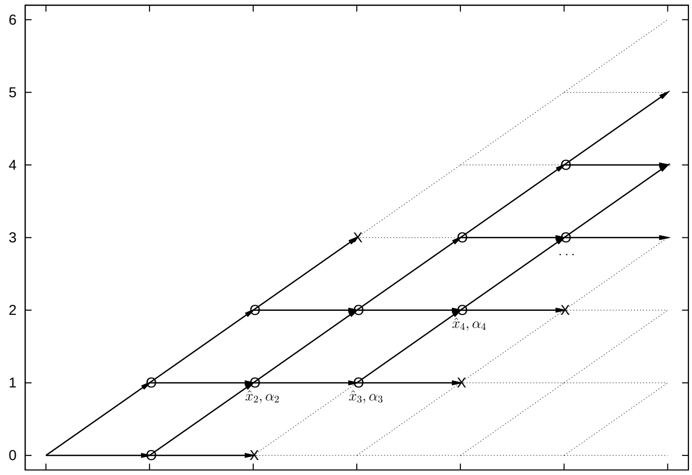
6
5
4
3
2
1
0
$\widehat{x}_2, \alpha_2$
$\widehat{x}_3, \alpha_3$
$\widehat{x}_4, \alpha_4$
...

Figure 4.2: Trellis representation of a possible evolution of the Two States Algorithm: circled nodes are the surviving states.

Initialization: $\widehat{x}_0 = 0$;

For $k = 1, \dots, K$, given the received symbol $y_k \in \mathbb{R}$,

$$\begin{aligned} \widehat{u}_{k-1} &= \mathbb{D}^{(1)}(\mathbf{y})_{k-1} = \operatorname*{argmax}_{u \in \{0,1\}} \mathrm{P}(U_{k-1} = u | Y_k = y_k, X_{k-1} = \widehat{x}_{k-1}) \\ &= \begin{cases} 0 & \text{if } \Gamma_k(\widehat{x}_{k-1}, \widehat{x}_{k-1}) \geq \Gamma_k(\widehat{x}_{k-1}, \widehat{x}_{k-1}+1) \\ 1 & \text{otherwise} \end{cases} \\ \widehat{x}_k &= \widehat{x}_{k-1} + \widehat{u}_{k-1} \end{aligned} \tag{4.13}$$

and given the equality (4.10) in the AWGN case,

$$\Gamma_k(\widehat{x}_{k-1}, \widehat{x}_{k-1}) \geq \Gamma_k(\widehat{x}_{k-1}, \widehat{x}_{k-1}+1) \quad \Leftrightarrow \quad |y_k - \widehat{x}_{k-1}| \leq |y_k - (\widehat{x}_{k-1}+1)|. \tag{4.14}$$

4.2.2 Two States Algorithm

The TSA scheme for the one-dimensional, binary input, differentiation problem is as follows:

TSA - Decoder $\mathbb{D}^{(2)}$

Initialization: $\widehat{x}_0 = 0$.

For $k = 1$: given the unique starting state $\widehat{x}_0 = 0$, we estimate the first bit by a One State procedure:

$$\begin{aligned}\widehat{u}_0 = \mathbb{D}^{(2)}(\mathbf{y})_0 &= \underset{u \in \{0,1\}}{\operatorname{argmax}} \mathrm{P}(U_0 = u | Y_1 = y_1, X_0 = 0) \\ &= \begin{cases} 0 & \text{if } |y_1| \leq |y_k - 1| \\ 1 & \text{otherwise.} \end{cases}\end{aligned} \tag{4.15}$$

Afterwards, the possible states are two: $\widehat{x}_1(0) = 0$ and $\widehat{x}_1(1) = 1$ and the corresponding probabilities $\alpha_1(0)$ and $\alpha_1(1)$ in our framework are given by

$$\begin{aligned}\alpha_1(j) &= f_{(X_1,Y_1)}(j, y_1) = f_{(Y_1|X_1)}(y_1|j)\mathrm{P}(X_1 = j) \\ &= f_{(Y_1|X_1)}(y_1|j)\mathrm{P}(U_0 = j) = \frac{1}{2} f_{(Y_1|X_1)}(y_1|j), \qquad j \in \{0, 1\}.\end{aligned}$$

We then normalize these probabilities so that $\alpha_1(0) + \alpha_1(1) = 1$ and we just store the couple of values $(\alpha_1(0), \widehat{x}_1(0))$, as this is sufficient to retrieve also $(\alpha_1(1), \widehat{x}_1(1)) = (1 - \alpha_1(0), \widehat{x}_1(0) + 1)$. For notational simplicity we rename the stored vector $(\alpha_1(0), \widehat{x}_1(0))$ as $(\alpha_1, \widehat{x}_1)$.

For $k = 2, 3, \ldots, K$: let us define $(A_{k-1}, \widehat{X}_{k-1}) \in [0, 1] \times \mathbb{N}$ as the random vector that represents the stored state along with the corresponding (normalized) probability. Given $(\alpha_{k-1}, \widehat{x}_{k-1})$,

$$\begin{aligned}&\widehat{u}_{k-1} = \mathbb{D}^{(2)}(\mathbf{y})_{k-1} = \\ &= \underset{u \in \{0,1\}}{\operatorname{argmax}} \mathrm{P}\big(U_{k-1} = u | Y_k = y_k, X_{k-1} = \widehat{x}_{k-1}, A_{k-1} = \alpha_{k-1}, \widehat{X}_{k-1} = \widehat{x}_{k-1}\big) = \\ &= \begin{cases} 0 & \text{if } \alpha_{k-1}\Gamma_k(\widehat{x}_{k-1}, \widehat{x}_{k-1}) + (1 - \alpha_{k-1})\Gamma_k(\widehat{x}_{k-1} + 1, \widehat{x}_{k-1} + 1) \geq \\ & \geq \alpha_{k-1}\Gamma_k(\widehat{x}_{k-1}, \widehat{x}_{k-1} + 1) + (1 - \alpha_{k-1})\Gamma_k(\widehat{x}_{k-1} + 1, \widehat{x}_{k-1} + 2) \\ 1 & \text{otherwise.} \end{cases}\end{aligned}$$

From step $k - 1$, three possible states arise: $\widehat{x}_{k-1}$, $\widehat{x}_{k-1} + 1$ and $\widehat{x}_{k-1} + 2$, whose probabilities are given by the forward recursion in (3.17):

$$\begin{aligned}&\alpha_k(\widehat{x}_{k-1}) = \alpha_{k-1}\Gamma_k(\widehat{x}_{k-1}, \widehat{x}_{k-1}) \\ &\alpha_k(\widehat{x}_{k-1} + 1) = \alpha_{k-1}\Gamma_k(\widehat{x}_{k-1}, \widehat{x}_{k-1} + 1) + (1 - \alpha_{k-1})\Gamma_k(\widehat{x}_{k-1} + 1, \widehat{x}_{k-1} + 1) \\ &\alpha_k(\widehat{x}_{k-1} + 2) = (1 - \alpha_{k-1})\Gamma_k(\widehat{x}_{k-1} + 1, \widehat{x}_{k-1} + 2).\end{aligned} \tag{4.16}$$

which can be reduced as follows in the case (4.10):

$$
\begin{aligned}
\alpha_k(\widehat{x}_{k-1}) &= \alpha_{k-1}\frac{1}{2\sigma\sqrt{2\pi}}\exp\left(-\frac{(y_k-\widehat{x}_{k-1})^2}{2\sigma^2}\right)\\
\alpha_k(\widehat{x}_{k-1}+1) &= \frac{1}{2\sigma\sqrt{2\pi}}\exp\left(-\frac{(y_k-(\widehat{x}_{k-1}+1))^2}{2\sigma^2}\right)\\
\alpha_k(\widehat{x}_{k-1}+2) &= (1-\alpha_{k-1})\frac{1}{2\sigma\sqrt{2\pi}}\exp\left(-\frac{(y_k-(\widehat{x}_{k-1}+2))^2}{2\sigma^2}\right).
\end{aligned}
$$

Since $|y_k-(\widehat{x}_{k-1}+1)| \neq \max\{|y_k-(\widehat{x}_{k-1}+j)|, j=0,1,2\}$, in the AWGN case $\alpha_k(\widehat{x}_{k-1}+1) \neq \min\{\alpha_k(\widehat{x}_{k-1}+j), j=0,1,2\}$. Hence, the state $\widehat{x}_{k-1}+1$ is never discarded and also the two "surviving" states are always adjacent. Therefore,

- we calculate $\alpha_{\min} = \min\{\alpha_k(\widehat{x}_{k-1}), \alpha_k(\widehat{x}_{k-1}+2)\}$.
- If $\alpha_{\min} = \alpha_k(\widehat{x}_{k-1})$, the surviving states are $(\widehat{x}_{k-1}+1, \widehat{x}_{k-1}+2)$ with probabilities $(\alpha_k(\widehat{x}_{k-1}+1), \alpha_k(\widehat{x}_{k-1}+2))$. We then store the lowest state along with the corresponding normalized probability: $(\alpha_k, \widehat{x}_k) = (\frac{\alpha_k(\widehat{x}_{k-1}+1)}{\alpha_k(\widehat{x}_{k-1}+1)+\alpha_k(\widehat{x}_{k-1}+2)}, \widehat{x}_{k-1}+1)$.
- Similarly, if $\alpha_{\min} = \alpha_k(\widehat{x}_{k-1}+2)$, $(\alpha_k, \widehat{x}_k) = (\frac{\alpha_k(\widehat{x}_{k-1})}{\alpha_k(\widehat{x}_{k-1})+\alpha_k(\widehat{x}_{k-1}+1)}, \widehat{x}_{k-1})$.

Remark 3 *Notice that if $\alpha_k = 1$, $\widehat{x}_k + 1$, then $\widehat{x}_{k+1} = \widehat{x}_k$; analogously, when $\alpha_k = 0$, $\widehat{x}_{k+1} = \widehat{x}_k + 1$. As a consequence, the unique initial state $\widehat{x}_0 = 0$ can be interpreted as a double state with all the probability in $\widehat{x}_0 = 0$, that is, $(\alpha_0, \widehat{x}_0) = (1, 0)$.*

The TSA procedure can be visualized on a trellis diagram in Figure 4.2. Both OSA and TSA schemes, that for simplicity have been here written for $\mathcal{U} = \{0, 1\}$, can be easily extended to any finite source alphabet.

4.2.3 Simulations and comparisons

In this section, we report the simulations' outcomes concerning the decoders $\mathbb{D}^{*_0}$, $\mathbb{D}^{(1)}$ and $\mathbb{D}^{(2)}$, respectively implemented with CBCJR, OSA and TSA. The results are the averages overall 5000 transmissions of 100 bit messages.

In Figure 4.3 we compare the efficiency of the three decoding schemes, in terms of BER: we evidence that two states are sufficient to achieve performance very close to the causal optimal CBCJR: the gain between $\mathbb{D}^{(2)}$ and $\mathbb{D}^{*_0}$ never exceeds 0.15 dB, while it achieves 0.8 dB between $\mathbb{D}^{(1)}$ and $\mathbb{D}^{*_0}$ for BER's values between 0.2 and 0.3. Furthermore, as reported in Table 4.2, OSA and TSA respectively require one and two storage locations, which makes them efficient even for long time transmissions and for a large number of states. Moreover, they have no delay.

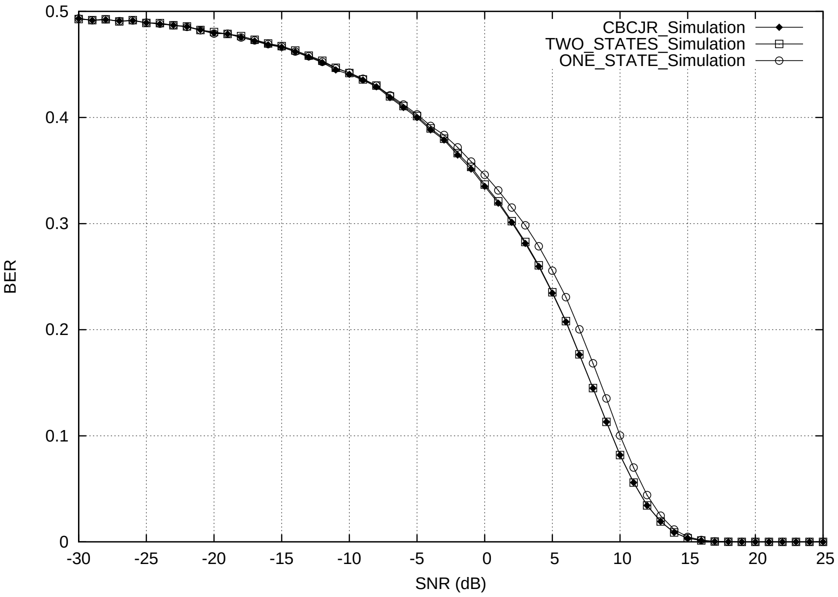

Figure 4.3: Performance comparison of different causal decoders.

	Computations	Storage Locations	Decoding Delay
BCJR	$O(K^2)$	$O(K^2)$	K
CBCJR	$O(K^2)$	$O(K)$	k_0
OSA	$O(K)$	1	0
TSA	$O(K)$	2	0

Table 4.2: Complexities comparison

4.3 Theoretic Analysis of the One State Algorithm

In this and in the following section, we propose an exhaustive theoretic analysis of the OSA and the TSA and provide a formal setting for the analytical evaluation of their performance. According to Definitions 4.5 and 4.8 in Section 1, the performance will be evaluated in terms of BER and CBER, which respectively describe the decoding for the "mean input" and for each possible input. The main results, for both OSA and TSA, consist in the assessment of the BER and the CBER in case of long time transmission; in particular, we will show that the performance in the "mean" case is equal to the performance obtained for "almost all" inputs.

Let us start with the analysis of the OSA. In order to state the main results, we have to introduce the following setting. Suppose to transmit a very large number of bits (say, $K \to \infty$) into our system and to decode the received message with the OSA; we define the stochastic process

$$D_k = \widehat{X}_k - X_k \in \mathbb{Z}, \quad k \in \mathbb{Z} \tag{4.17}$$

where $\widehat{X}_0 = 0$, $\widehat{X}_{k+1} = \widehat{X}_k + \widehat{U}_k$ and $\widehat{U}_k = \mathbb{D}^{(1)}(\mathbf{y})_k$ (see the algorithm (4.13). D_k represents the difference between the real and the estimated state values. Since $D_0 = 0$, the following recursive relationship holds:

$$D_{k+1} = D_k + \widehat{U}_k - U_k. \tag{4.18}$$

While the U_k's are mutually independent, each $\widehat{U}_k$ is function of U_k and D_k. Thus, the stochastic process $(D_k)_{k\in\mathbb{N}}$ is a Markov Chain.

By Markov Chain [129, Chapter 4] we intend any sequence of random variables $(X_n)_{n\in\mathbb{N}}$ assuming values in a denumerable set $\mathbf{X}$ (unless otherwise indicated, $\mathbf{X} = \mathbb{Z}$) and satisfying the Markov property: $\mathrm{P}(X_{n+1} = y|X_n = x, X_{n-1}, \ldots, X_0) = \mathrm{P}(X_{n+1} = y|X_n = x)$. If the chain is time-homogeneous, that is $\mathrm{P}(X_{n+1} = y|X_n = x) = \mathrm{P}(X_{n+m+1} = y|X_{n+m} = x)$, the *transition probabilities* $\mathbf{P}_{x,y} = \mathrm{P}(X_{n+1} = y|X_n = x)$ are the entries of a stochastic *transition probability matrix* $\mathbf{P} \in [0,1]^{\mathbf{X}\times\mathbf{X}}$.

A probability vector (p.v. for short) Φ ($\Phi \in [0,1]^{\mathbf{X}}$, $\sum_{x\in X} \Phi_x = 1$) such that $\Phi^T\mathbf{P} = \Phi^T$ is said to be *invariant* (or *stationary*) for the Markov Chain with transition probability matrix $\mathbf{P}$ (see, e.g., [129, Section 4.4]).

4.3.1 OSA Performance Theorems

At this point, we have the elements to state the main performance results concerning the OSA.

Theorem 2 (OSA - BER) *Let $\overline{\mathrm{q}}_d = \mathrm{P}[\widehat{U}_k \neq U_k|D_k = d]$, then*

$$\lim_{K\to\infty} \mathrm{BER}(\mathbb{D}^{(1)}) = \sum_{d\in\mathbb{Z}} \overline{\mathrm{q}}_d \Phi_d$$

where Φ is the unique invariant p.v. of $(D_k)_{k\in\mathbb{N}}$.

Theorem 3 (OSA - CBER) *Let π be the uniform Bernoulli probability measure over $\{0,1\}^{\mathbb{N}}$. Then, for the One State Algorithm,*

$$\lim_{K\to\infty} \mathrm{CBER}(\mathbb{D}^{(1)}|\mathbf{U}) = \lim_{K\to\infty} \mathrm{BER}(\mathbb{D}^{(1)}) \quad \textit{for } \pi\textit{-a.e. } \mathbf{U}.$$

Our next goal is to prove Theorem 2, on the basis of a classical ergodic result for Markov Chains, which is now reviewed. Instead, the proof of Theorem 3 is a bit more technical and requires some elements from the theory of Markov Chains in Random Environments (see Section 4.7.1). We then postpone it to the Appendix 4.7.2.

4.3.2 Review of the Ergodic Theorem for Markov Chains

Let $(X_n)_{n\in\mathbb{N}}$ be a Markov Chain on the state space $\mathbf{X} = \mathbb{Z}$.

Definition 2 *[129, Section 4.3] Two states $x, y \in \mathbb{Z}$* communicate *if there exist $n, m \in \mathbb{N}$ s.t. $(\mathbf{P}^n)_{x,y} > 0$ and $(\mathbf{P}^m)_{y,x} > 0$. Two states that communicate are said to belong to same* class. *If all the states communicate, say there is only one class, the Markov Chain is said to be* irreducible.

Definition 3 *[68, Section 3.2] A state j is said to be* positive recurrent *if* $\mathbb{E}(\tau_j|X_0 = j) < \infty$ *where* $\tau_j = \min\{n > 0 : X_n = j\}$.

Positive recurrence is a class property (see [68, Section 3.2]); thus, if a Markov Chain is irreducible and has one positive recurrent state, then all the states are positive recurrent. In such case, the Markov Chain is said to be *ergodic.*

Theorem 4 (Ergodic Theorem for Markov Chains) *[68, Propositions 3.3.1 - 3.3.2] An ergodic Markov Chain admits a unique invariant p.v. Φ. Furthermore, for every function f such that $\sum_{d\in\mathbb{Z}} |f(d)|\Phi_d < \infty$ and for every initial state $i \in \mathbb{Z}$,*

$$\lim_{K\to\infty} \sum_{k=0}^{K-1} \sum_{d\in\mathbb{Z}} f(d)(\mathbf{P}^k)_{i,d} = \sum_{d\in\mathbb{Z}} f(d)\Phi_d.$$

4.3.3 Proof of Theorem 2

Let us go back to the One State Algorithm. According to (4.18), $(D_k)_{k\in\mathbb{N}}$ is a denumerable homogeneous Markov Chain on $\mathbb{Z}$, with transition probabilities

$$\overline{\mathbf{P}}_{x,y} = \mathrm{P}(D_{k+1} = y|D_k = x) = \frac{1}{2}[\mathbf{P}_{x,y}(0) + \mathbf{P}_{x,y}(1)]$$

where $\mathbf{P}_{x,y}(u) = \mathrm{P}(D_{k+1} = y|D_k = x, U_k = u)$, $u \in \{0, 1\}$. Notice that the only non-null entries of $\mathbf{P}(u)$ are the following:

$$\mathbf{P}_{d,d+1}(0) = \frac{1}{2}\,\mathrm{erfc}\left(\frac{d+\frac{1}{2}}{\sqrt{2}\sigma}\right) \qquad \mathbf{P}_{d,d}(0) = 1 - \mathbf{P}_{d,d+1}(0)$$

$$\mathbf{P}_{d,d}(1) = \frac{1}{2}\,\mathrm{erfc}\left(\frac{d-\frac{1}{2}}{\sqrt{2}\sigma}\right) \qquad \mathbf{P}_{d,d-1}(1) = 1 - \mathbf{P}_{d,d}(1)$$

Now, we prove a few lemmas that will lead to the proof of Theorem 2.

Lemma 1 *$(D_k)_{k\in\mathbb{N}}$ is ergodic.*

Proof $\overline{\mathbf{P}}$ is tridiagonal and, for any $x, y \in \mathbb{Z}$, $\overline{\mathbf{P}}_{x,y} = \overline{\mathbf{P}}_{-x,-y}$ and $\overline{\mathbf{P}}_{x,y} > 0$ if and only if $|x - y| \leq 1$; by iteration, for any $n \in \mathbb{N}$, $(\overline{\mathbf{P}}^n)_{x,y} > 0$ if and only if $|x - y| \leq n$. Hence, given any couple of states $x, y \in \mathbb{Z}$ with distance $|x - y| = m$, $(\overline{\mathbf{P}}^m)_{x,y} > 0$ and $(\overline{\mathbf{P}}^m)_{y,x} > 0$, that is, $(D_k)_{k\in\mathbb{N}}$ is irreducible.

To prove that $(D_k)_{k\in\mathbb{N}}$ is positive recurrent, it suffices to apply the criterion proposed in [146, Exercise 3.3.3]: if there exists a function $\mathbf{g} \in \mathbb{R}^{+\mathbb{Z}}$ so that $\mathbf{g}_x \geq (\overline{\mathbf{P}}\mathbf{g})_x + \varepsilon$ for any $x \in \mathbb{Z} \setminus \{y\}$ and for some $\varepsilon > 0$, then y is a positive recurrent state. In our case, it is easy to prove that $y = 0$ is a positive recurrent state considering $\mathbf{g}_x = |x|$. Moreover, given that the chain is irreducible, if one state is positive recurrent, all states are so. ■

Lemma 2 $(D_k)_{k\in\mathbb{N}}$ *admits a unique invariant p.v.* Φ*, defined by*

$$\Phi_d = \Phi_0 \prod_{i=1}^{|d|} \frac{\overline{\mathbf{P}}_{i-1,i}}{\overline{\mathbf{P}}_{i,i-1}} \tag{4.19}$$

where $\Phi_0 = \left[1 + 2\sum_{d=1}^{\infty} \prod_{i=1}^{|d|} \overline{\mathbf{P}}_{i-1,i}/\overline{\mathbf{P}}_{i,i-1}\right]^{-1}$.

Proof $(D_k)_{k\in\mathbb{N}}$ is ergodic by Lemma 1, hence it admits a unique invariant p.v. Φ by Theorem 4. Let us then prove (4.19).

By $(\Phi^T\overline{\mathbf{P}})_d = \Phi_d^T$, for any $d \in \mathbb{Z}$, it follows that

$$\Phi_{d-1}\overline{\mathbf{P}}_{d-1,d} - \Phi_d\overline{\mathbf{P}}_{d,d-1} = c \quad (c \text{ constant}). \tag{4.20}$$

In particular, as $\Phi_d = \Phi_{-d}$ for any $d \in \mathbb{Z}$ (this is due to the uniqueness of Φ and to the symmetry of $\overline{\mathbf{P}}$), it suffices to substitute values $d = 0$ and $d = 1$ in (4.20) to conclude that $c = 0$; hence, relation (4.19) holds. ■

Notice that $c = 0$ corresponds to the property of time-reversibility of a Markov Chain (see Section 4.8 of [129]), hence one could even prove it by Theorem 4.2 in [129, Section 4.8], after having verified the *aperiodicity* [129, Section 4.4]

From Lemma 2 we deduce in particular that for any $d \in \mathbb{Z}$, $\Phi_d > 0$. Moreover, since $\overline{\mathbf{P}}_{i-1,i}/\overline{\mathbf{P}}_{i,i-1} < 1$ for $i \geq 1$, Φ_d has a maximum at $d = 0$ and it is monotone decreasing for $d > 0$.

Let us know conclude the proof of Theorem 2. Since

$$\mathrm{BER}(\mathbb{D}^{(1)}) = \frac{1}{K}\sum_{k=0}^{K-1} \mathrm{P}(\widehat{U}_k \neq U_k) = \frac{1}{K}\sum_{k=0}^{K-1}\sum_{d\in\mathbb{Z}} \overline{\mathrm{q}}_d(\overline{\mathbf{P}}^k)_{0,d}$$

the result follows from Theorem 4 and Lemma 1. In fact, given the ergodicity proved in Lemma 1, we can apply Theorem 4 with $f(d) = \overline{\mathrm{q}}_d$ (as $\overline{\mathrm{q}}_d$ is a probability, $\sum_{d\in\mathbb{Z}} \overline{\mathrm{q}}_d\Phi_d < \infty$ holds).

4.4 Theoretic Analysis of the Two States Algorithm

Similar to the OSA, the TSA procedure can be studied through the Markov Theory, which provides the instruments to compute both BER and CBER. The main results are collected in two Performance Theorems that we will state after having introduced the necessary setting.

As shown in Section 4.2.2, the Two States procedure stores, at each step, a state and its normalized probability, this information being sufficient to individuate also the second state and probability. Let $\widehat{X}_k$ be the r.v. representing the stored state, X_k the current correct state, $D_k = \widehat{X}_k - X_k$ and A_k the r.v corresponding to the probability of $\widehat{X}_k$: now, the stochastic process $(A_k, D_k)_{k\in\mathbb{N}}$ in $[0,1] \times \mathbb{Z}$ is a Markov Process, whose definition (which actually extends the definition of Markov Chain from a denumerable to a continuous set) is now given.

Consider a set $\mathbf{X}$ endowed with a countably generated σ-field $\mathcal{F}$. A *transition probability kernel* (or *Markov probability kernel*, see, e.g., [94, Section 3.4.1]) on $(\mathbf{X}, \mathcal{F})$ is an application $P : \mathbf{X} \times \mathcal{F} \to [0, 1]$ such that
(i) for each $F \in \mathcal{F}$, $P(\cdot, F)$ is a non-negative measurable function;
(ii) for each $x \in \mathbf{X}$, $P(x, \cdot)$ is a probability measure (p.m. for short) on $(\mathbf{X}, \mathcal{F})$.

Given a bounded measurable function v on $(\mathbf{X}, \mathcal{F})$, we denote by Pv the bounded measurable function on $(\mathbf{X}, \mathcal{F})$ defined as

$$(Pv)(x) = \int_{\mathbf{X}} v(y)P(x, \mathrm{d}y). \tag{4.21}$$

Further, let μ be a measure on $(\mathbf{X}, \mathcal{F})$: we define the measure μP

$$(\mu P)(F) = \int_{\mathbf{X}} P(x, F)\mu(\mathrm{d}x) \qquad F \in \mathcal{F}. \tag{4.22}$$

We define the n-th power of the transition kernel P simply putting $P^1(x, F) = P(x, F)$ and $P^n(x, F) = \int_{\mathbf{X}} P(x, \mathrm{d}y)P^{n-1}(y, F)$. It is easy to see that $P^n(x, F)$ are transition kernels, too. Corresponding actions on bounded functions and on measures will be respectively denoted by $P^n v$ and μP^n.

A measure ψ on $(\mathbf{X}, \mathcal{F})$ is said to be *invariant* for the transition kernel P if $\psi P = \psi$ (see, e.g., [94, (10.1)]).

We define a *homogeneous Markov Process on space* $(\mathbf{X}, \mathcal{F})$ *with transition kernel* P as a sequence of $\mathbf{X}$-valued random variables $(X_n)_{n\in\mathbb{N}}$ such that, for any $x \in \mathbf{X}$ and $F \in \mathcal{F}$,

$$\mathrm{Prob}(X_{n+1} \in F | X_n = x, X_{n-1}, \dots, X_0) = \mathrm{Prob}(X_{n+1} \in F | X_n = x) = P(x, F)$$

for any $n \in \mathbb{N}$. The evolution of $(X_n)_{n\in\mathbb{N}}$ is completely described once we fix a probability law μ of X_0 on $(\mathbf{X}, \mathcal{F})$; if μ is invariant, then the Markov Process is said to be stationary: all the r.v.'s X_n are distributed according to μ. Notice also that for any $x \in \mathbf{X}$ and $F \in \mathcal{F}$, $\mathrm{Prob}(X_{m+n} \in F | X_m = x) = P^n(x, F)$ for any $m, n \in \mathbb{N}$.

4.4.1 TSA Performance Theorems

$(A_k, D_k)_{k\in\mathbb{N}}$ is a Markov Process in $([0, 1] \times \mathbb{Z}, \mathcal{B}([0, 1]) \times \mathcal{P}(\mathbb{Z}))$ where $\mathcal{B}([0, 1])$ is the Borel σ-field on $[0, 1]$ and $\mathcal{P}(\mathbb{Z})$ is the discrete σ-field of $\mathbb{Z}$. In order to completely define the process, we provide also an initial distribution $\mathcal{L} \times \kappa$, $\mathcal{L}$ and κ respectively being the usual Lebesgue measure on $[0, 1]$ and the counting measure on $\mathbb{Z}$.

The transition probability kernel will be explicitly computed in the Appendix 4.7.3.

Theorem 5 (TSA - BER) *Let* $\overline{q}(\alpha, d) := P(\widehat{U}_k \neq U_k | \alpha_k = \alpha, D_k = d)$*, then*

$$\lim_{K\to\infty} \mathrm{BER}(\mathbb{D}^{(2)}) = \int_{[0,1]\times\mathbb{Z}} \overline{q}\ \mathrm{d}\widetilde{\phi}$$

where $\widetilde{\phi}$ is the unique the invariant p.m. of the kernel of $(A_k, D_k)_{k\in\mathbb{N}}$.

Theorem 6 (TSA - CBER) *Let π be the uniform Bernoulli probability measure over $\{0,1\}^{\mathbb{N}}$. Then, for the TSA,*

$$\lim_{K\to\infty} \mathrm{CBER}(\mathbb{D}^{(2)}|\mathbf{U}) = \lim_{K\to\infty} \mathrm{BER}(\mathbb{D}^{(2)}) \quad \textit{for } \pi\textit{-a.e. } \mathbf{U}.$$

Our next goal is to prove Theorem 2 using the Ergodic Theorem for Markov Processes, which is now reviewed. Instead, we refer the reader to the Appendix 4.7.8 for the proof of Theorem 6.

4.4.2 Review of the Ergodic Theorem for Markov Processes

From now onwards, we will assume $\mathbf{X}$ to be a *locally compact separable metric space*: under this topological condition we can easily prove the existence of an invariant measure (see [94, Section 12.3]). Let $\mathcal{B}(\mathbf{X})$ be the Borel σ-algebra of $\mathbf{X}$.

Definition 4 *[94, Sections 6.1.1, 11.3.1] Let P be a transition kernel on $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$. If $P(\cdot, O)$ is a lower semicontinuous function for any open set $O \in \mathcal{B}(\mathbf{X})$, then P is said to be* weak Feller*. Moreover, we say that P verifies the* Drift Condition *if there exist a compact set $\mathrm{C} \subset \mathbf{X}$, a constant $b < \infty$ and a function $V : \mathbf{X} \to [0, \infty]$ not always infinite such that*

$$\Delta V(x) := \int_{\mathbf{X}} P(x, \mathrm{d}y)V(y) - V(x) \le -1 + b\mathbb{1}_{\mathrm{C}}(x) \tag{4.23}$$

for every $x \in \mathbf{X}$.

Proposition 1 [94, Theorem 12.3.4] *If a transition kernel P is weak Feller and verifies the Drift Condition, then it admits an invariant p.m..*

Under some further conditions, also the uniqueness of the invariant measure can be proved.

Definition 5 *[94, Section 4.2.1] For any $B \in \mathcal{B}(\mathbf{X})$, let $\tau_B = \min\{n > 0 : X_n \in B\}$. $(X_n)_{n\in\mathbb{N}}$ is said to be μ-irreducible if there exists a measure μ on $\mathcal{B}(\mathbf{X})$ such that for every $x \in \mathbf{X}$, $\mu(B) > 0$ implies $\mathrm{P}(\tau_B < +\infty | X_0 = x) > 0$.*

A μ-irreducible Markov Process whose kernel admits an invariant p.m. is said to be *positive recurrent*[94] and

Proposition 2 [94, Theorem 10.0.1, Proposition 10.1.1] *The kernel of a positive recurrent Markov Process admits a unique invariant p.m..*

Furthermore,

Definition 6 *[68, Definitions 2.2.2, 2.4.1] A set $B \in \mathcal{B}(\mathbf{X})$ is said to be* invariant *if $P(x, B) \geq \mathbb{1}_B(x)$ for every $x \in \mathbf{X}$.*
A p.m. μ on $\mathcal{B}(\mathbf{X})$ is said to be ergodic *if $\mu(B) = 0$ or $\mu(B) = 1$ for every invariant set $B \in \mathcal{B}(\mathbf{X})$.*

Proposition 3 [68, Proposition 2.4.3] *If a Markov Process admits a unique invariant p.m. μ, then μ is ergodic.*

A fundamental issue for our analysis is the Ergodic Theorem of Markov Processes, which is the transposition into stochastic terms of the Birkhoff's Individual Ergodic Theorem ([159, Theorem 1.14]). Here we report its version under the ergodicity condition for an invariant p.m.; for a more general treatise, see [54, 68].

Theorem 7 (Ergodic Theorem for Markov Processes) *[68, Theorem 2.3.4 - Proposition 2.4.2] Assume that a kernel P on $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$ admits an ergodic invariant p.m. μ. Then, for any non-negative function $v \in L_1(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$,*

$$\lim_{K \to \infty} \frac{1}{K} \sum_{k=0}^{K-1} (P^k v)(x) = \int_{\mathbf{X}} v \, \mathrm{d}\mu \quad \textit{for} \;\; \mu\textit{-a.e.} \;\; x \in \mathbf{X}.$$

Finally, we report a result of direct convergence for the iterates of the kernel, in the case of aperiodic behavior.

Definition 7 *[8, Section 2] A Markov Process is said to be* strongly aperiodic *it there exist a set $A \subseteq \mathbf{X}$, a probability measure ν on A and a finite number $c > 0$ such that $P(x, B) \geq c\nu(B)$ for any $x \in A, B \in \mathcal{B}(\mathbf{X})$.*

Now, let $||P^n(x, \cdot) - \mu|| = 2 \sup_{B \in \mathcal{B}(X)} |P^n(x, B) - \mu(B)|$ be the total variation norm between the measures $P^n(x, \cdot)$ and μ.

Proposition 4 *[152, Theorem 4.1 (i)] For a positive recurrent, strongly aperiodic Markov Process with invariant p.m. μ, $||P^n(x, \cdot) - \mu(\cdot)|| \to 0$ as $n \to \infty$ for μ-a.e. $x \in \mathbf{X}$.*

4.4.3 Proof of Theorem 5

The proof of Theorem 5 consists of three steps. First, we prove that the kernel of $(A_k, D_k)_{k \in \mathbb{N}}$ admits an invariant p.m. $\widetilde{\phi}$,; second, we prove the uniqueness of such invariant p.m.; third, we show how to apply the Ergodic Theorem 7 to achieve the thesis.

Lemma 3 *The kernel of $(A_k, D_k)_{k \in \mathbb{N}}$ admits an invariant p.m. $\widetilde{\phi}$.*

The proof requires some technical computation and is postponed to Appendix 4.7.5.

Lemma 4 *$\widetilde{\phi}$ is unique (and ergodic).*

Proof $(A_k, D_k)_{k\in\mathbb{N}}$ is $(\mathcal{L}\times\kappa)$-irreducible (the proof of this fact is in the Appendix 4.7.6), then positive recurrent by Lemma 3. Thus, $\widetilde{\phi}$ is unique by Proposition 2 and ergodic by Proposition 3. ■

Given these two Lemmas, we now evaluate the BER by means of the Ergodic Theorem 7. The BER is given by

$$\begin{aligned}
&\mathrm{BER}(\mathbb{D}^{(2)}) = \\
&\quad = \frac{1}{K}\sum_{k=0}^{K-1} P(\widehat{U}_k \neq U_k) = \frac{1}{K}\sum_{k=0}^{K-1}\int_0^1 \sum_{d\in\mathbb{Z}} P(\widehat{U}_k \neq U_k, A_k = \alpha, D_k = d)\mathrm{d}\alpha \\
&\quad = \frac{1}{K}\sum_{k=0}^{K-1}\int_0^1 \sum_{d\in\mathbb{Z}} P(\widehat{U}_k \neq U_k | A_k = \alpha, D_k = d) P^k\big((1,0);(\mathrm{d}\alpha, d)\big).
\end{aligned}$$

the initial state $(1,0)$ being discussed in the Remark 3. As $\overline{q}(\alpha,d) = P(\widehat{U}_k \neq U_k|\alpha_k = \alpha, D_k = d)$, (notice that $\overline{q}(\alpha,d)$ actually does not depend on k) we have $\mathrm{BER}(\mathbb{D}^{(2)}) = \frac{1}{K}\sum_{k=0}^{K-1}(P^k\overline{q})(1,0)$.

Given Lemma 4, by the Ergodic Theorem 7,

$$\lim_{K\to\infty}\frac{1}{K}\sum_{k=0}^{K-1}(P^k\overline{q})(\alpha,d) = \int_{[0,1]\times\mathbb{Z}} \overline{q}\ \mathrm{d}\widetilde{\phi} \quad \widetilde{\phi}\text{-a.e. } (\alpha,d).$$

This result cannot be immediately applied to evaluate the BER since the convergence is not assured for *all* the initial states. In particular, let $N\subset[0,1]\times\mathbb{Z}$ be the negligible set for which there is no convergence and let $N_0 = \{\alpha\in[0,1] : (\alpha,0)\in N\}$. Now, recalling Remark 3,

$$\begin{aligned}
&\mathrm{BER}(\mathbb{D}^{(2)}) = \\
&\quad = \frac{1}{K}\overline{q}(1,0) + \frac{1}{K}\sum_{k=1}^{K-1}\int_{\alpha_1\in[0,1]}\sum_{d_1\in\mathbb{Z}} P((1,0),(\mathrm{d}\alpha_1,d_1))(P^{k-1}\overline{q})(\alpha_1,d_1) \\
&\quad = \frac{1}{K}\overline{q}(1,0) + \frac{1}{K}\sum_{k=1}^{K-1}\int_{\alpha_1\in[0,1]} P((1,0),(\mathrm{d}\alpha_1,0))(P^{k-1}\overline{q})(\alpha_1,0).
\end{aligned}$$

By the Lebesgue's Dominated Convergence Theorem,

$$\lim_{K\to\infty}\mathrm{BER}(\mathbb{D}^{(2)}) = \int_{\alpha_1\in[0,1]} P((1,0),(\mathrm{d}\alpha_1,0))\lim_{K\to\infty}\frac{1}{K}\sum_{k=1}^{K-1}(P^{k-1}\overline{q})(\alpha_1,0).$$

Notice that $\mathcal{L}(N_0) = 0$, otherwise $\widetilde{\phi}(N_0\times\{0\}) = \int_{[0,1]\times\mathbb{Z}} P(\omega, N_0\times\{0\})\widetilde{\phi}(\mathrm{d}\omega) > C_{\varepsilon,0}\mathcal{L}(N_0) > 0$ by Proposition 7 in the Appendix 4.7.6. By Lemma 6 in the Appendix

4.7.7, this implies that $P((1,0), N_0 \times \{0\}) = 0$. Finally,

$$\begin{aligned}\lim_{K\to\infty} \mathrm{BER}(\mathbb{D}^{(2)}) &= \int_{\alpha_1 \in [0,1]\setminus N_0} P((1,0),(\mathrm{d}\alpha_1, 0)) \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K-1} (P^{k-1}\overline{q})(\alpha_1, 0) \\ &= \int_{\alpha_1 \in [0,1]\setminus N_0} P((1,0),(\mathrm{d}\alpha_1, 0)) \int_{[0,1]\times\mathbb{Z}} \overline{q}\ \mathrm{d}\widetilde{\phi} = \int_{[0,1]\times\mathbb{Z}} \overline{q}\ \mathrm{d}\widetilde{\phi}\end{aligned}$$

as $(\alpha_1, 0) \notin N$. The function $\overline{q}(\alpha, d)$ is explicitly computed in the Appendix 4.7.4.

4.4.4 Direct Convergence to $\widetilde{\phi}$

The explicit construction of an invariant p.m. is an intricate issue in the not countable framework. When ergodic results are available, one can approximate it by several procedures (see, e.g, [68, Chapter 12]). In our framework, we can obtain an approximation by Proposition 4, which states the direct convergence of the iterates $P^n(\cdot,\cdot)$ to the invariant p.m.. Before illustrating that, let us prove that the hypotheses of Proposition 4 hold.

Proposition 5 *The Markov Process $(A_k, D_k)_{k\in\mathbb{N}}$ is strongly aperiodic.*

Proof Let us consider the probability measure $\mathcal{L} \times \delta_{\bar{d}}$ on $([0,1]\times\mathbb{Z}, \mathcal{B}([0,1])\times\mathcal{P}(\mathbb{Z}))$, where $\mathcal{L}$ is the Lebesgue measure and $\delta_{\bar{d}}(d) = 1$ if $d = \bar{d}$, 0 otherwise. By Proposition 7, $P((\alpha, d), M \times \{d\}) > \frac{1}{2}C_{\varepsilon,d}\mathcal{L}(M)$, $C_{\varepsilon,d} > 0$. Then, considering the Definition 7 with $\nu = \mathcal{L}\times\delta_{\bar{d}}$, $c = \frac{1}{2}C_{\varepsilon,d}$ and $A = [0,1]\times\{\bar{d}\}$, the proposition is proved. ■

This result along with Proposition 4 yields:

Corollary 1 (Direct Convergence) $||P^n((\alpha, d),\cdot) - \widetilde{\phi}|| \to 0$ *as* $n \to \infty$ *for* ϕ*-a.e.* $(\alpha, d) \in [0,1]\times\mathbb{Z}$.

4.5 Analytic vs Simulations' outcomes

To conclude our analysis of the OSA and TSA, we compare the simulations' outcomes with the theoretic results: we expect the BER's obtained by the simulations of sufficiently long transmissions to be consistent to the analytic computations.

By Theorems 2 and 5, the BER's can be computed once we know the corresponding invariant distributions. While for the OSA the invariant p.v. is explicitly given by (4.19), for the TSA we have approximated the invariant p.m. using the Corollary 1. In particular, we have discretized the kernel P into a matrix, afterwards we have computed the iterates P^n for a sufficiently large n, so that to obtain an equilibrium condition, that is, a matrix whose rows are all equal up to numerical roundoff . At this point, any row of the matrix is a discretized, approximated version of the invariant p.m.

In Figures 4.4 and 4.5, we compare analytic and simulations' outcomes respectively for OSA and TSA: as expected, they do not present substantial differences.

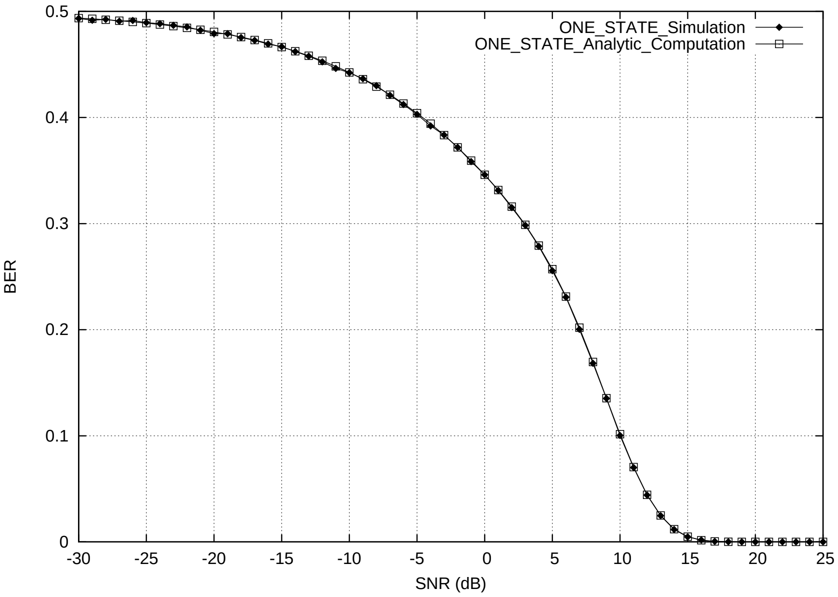

Figure 4.4: OSA: analytic computation vs simulation.

4.6 Conclusions

In this chapter, we have studied the differentiation problem in one dimension and in case of binary input generated by a Bernoulli source. The algorithms BCJR, CBCJR, OSA and TSA have been implemented to estimate the unknown input and simulations have been proposed. Furthermore, we have provided an exhaustive theoretical analysis of the performance of the OSA and the TSA, in terms of Markov Chains and Markov Processes. In particular, we have described the behavior of these algorithms in the asymptotic case - which is the most interesting for all those applications that envisage long time transmissions - exploiting the ergodic properties of the associated Markov Chains or Processes. The proposed algorithms are known to be sensible to the input sequences, say they have different performance for different inputs; however, we have proved that for *almost all* possible input sequences the algorithms behave as in the mean case (i.e., the case obtained by averaging the possible input sequences).

4.7 Appendix

4.7.1 Markov Chains in Random Environments

Consider a countable set Θ and a family of transition probability kernels $\{P_\theta, \theta \in \Theta\}$ on a space $(\mathbf{X}, \mathcal{F})$. Given a σ-field $\mathcal{B}$ of Θ, let $(\theta_n)_{n\in\mathbb{N}}$ and $(X_k)_{k\in\mathbb{N}}$ respectively be sequences of Θ-valued and $\mathbf{X}$-valued r.v's. $P_{\theta_k}(X_k, F)$ can now be interpreted as the transition probability of X_k to set F depending on the r.v θ_k, which represents to so-

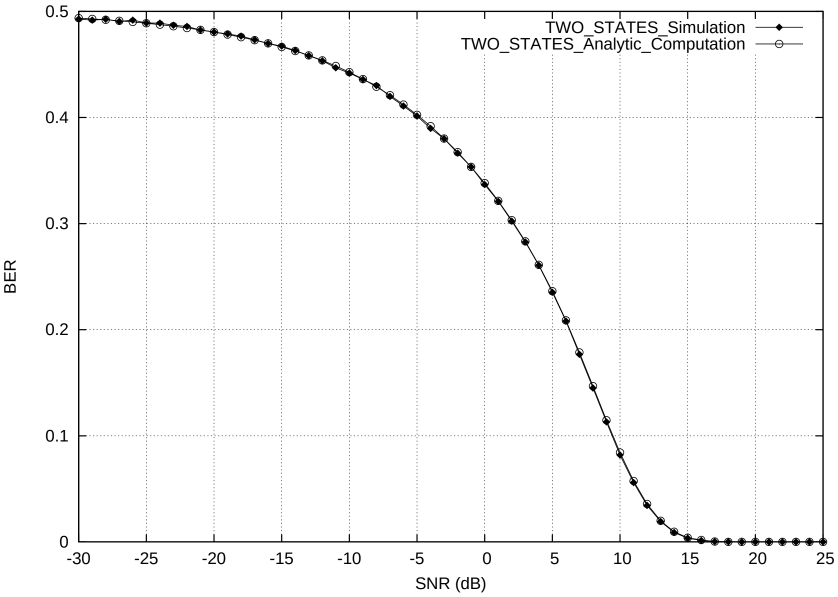

Figure 4.5: TSA: analytic computation vs simulation.

called *random environment.*
We say that $(X_k)_{k\in\mathbb{N}}$ with $(\theta_n)_{n\in\mathbb{Z}}$ is a Markov Chain in Random Environment (or MCRE) if

$$\begin{aligned} &P(X_{k+1}\in F|X_k,\dots,X_0,(\theta_n)_{n\in\mathbb{Z}}) = P_{\theta_k}(X_k,F) \quad \text{a.s.} \\ &\quad \text{for all } F\in\mathcal{F} \text{ and } k=0,1,\dots \end{aligned} \tag{4.24}$$

Let us define $\Theta^{\mathbb{N}} = \prod_0^{+\infty}\Theta$ and $\mathcal{B}^{\mathbb{N}} = \prod_0^{+\infty}\mathcal{B}$. An important feature of a MCRE is that we can always associate to it a classical Markov Process. In fact, given any $x\in\mathbf{X}$ and $\underline{\theta} = (\theta_0,\theta_1,\dots)\in\Theta^{\mathbb{N}}$ and denoting by T the left sequence shift on $\Theta^{\mathbb{N}}$ (that is, $T\underline{\theta} = \underline{\widetilde{\theta}}$ with $\underline{\widetilde{\theta}}_n = \underline{\theta}_{n+1}$ for any $n\in\mathbb{N}$), we can introduce the following transition probability kernel on $(\mathbf{X}\times\Theta^{\mathbb{N}},\mathcal{F}\times\mathcal{B}^{\mathbb{N}})$:

$$P\big((x,\underline{\theta}),F\times B\big) = P_{\theta_0}(x,F)\mathbb{1}_B(T\underline{\theta}) \tag{4.25}$$

which determines a Markov Process $\big(X_k,T^k(\theta_n)_{n\in\mathbb{N}}\big)_{k\in\mathbb{N}}$ on $(\mathbf{X}\times\Theta^{\mathbb{N}},\mathcal{F}\times\mathcal{B}^{\mathbb{N}})$. From now onwards, we will refer to it as to the *Extended Markov Process*, EMP for short.

Remark 4 *If the random environments θ_n's are independent and identically distributed then $(X_k)_{k\in\mathbb{N}}$ is a Markov Process with transition probability kernel $P(x,F) = \mathbb{E}\left[P_{\theta_0}(x,F)\right]$. In other terms, $(X_k)_{k\in\mathbb{N}}$ is the Markov Process moving in the* average environment.

In this framework, we prove the following

Proposition 6 *Let $(X_k)_{k\in\mathbb{N}}$ with $(\theta_n)_{n\in\mathbb{N}}$ be a MCRE on $\mathbf{X}\times\Theta^{\mathbb{N}}$. Suppose the random environments θ_n's to be independent, identically distributed with distribution π_0 on $(\Theta,\mathcal{B})$ and define the distribution $\pi = \times_{n=0}^{\infty}\pi_0$ over $(\Theta^{\mathbb{N}},\mathcal{B}^{\mathbb{N}})$. Moreover, suppose the kernel $P(\cdot) = \mathbb{E}[P_{\theta_0}(\cdot,\cdots)]$ of the Markov Process $(X_k)_{k\in\mathbb{N}}$ (see Remark 4) to admit an invariant p.m. ϕ. Then,*

$$\psi = \phi \times \pi \tag{4.26}$$

is an invariant p.m. for the EMP $\big(X_k, T^k(\theta_n)_{n\in\mathbb{N}}\big)_{k\in\mathbb{N}}$ over $(\mathbf{X}\times\Theta^{\mathbb{N}},\mathcal{F}\times\mathcal{B}^{\mathbb{N}})$.

Proof Let $\omega = (x,\underline{\theta}) \in \mathbf{X}\times\Theta^{\mathbb{N}}$. ψ is an invariant p.m. for $(X_k, T^k(\theta_n)_{n\in\mathbb{N}})_{k\in\mathbb{N}}$ if

$$\int_{\mathbf{X}\times\Theta^{\mathbb{N}}} P(\omega, F\times B)\psi(\mathrm{d}\omega) = \psi(F\times B)$$

for any $F\times B$ such that $F\in\mathcal{F}$, $B\in\mathcal{B}^{\mathbb{N}}$. Now,

$$\begin{aligned}\int_{\mathbf{X}\times\Theta^{\mathbb{N}}} P(\omega, F\times B)\psi(\mathrm{d}\omega) &= \int_{\mathbf{X}}\int_{\Theta^{\mathbb{N}}} P_{\theta_0}(x,F)\mathbb{1}_B(\theta_1,\theta_2,\dots)\pi(\mathrm{d}\underline{\theta})\phi(\mathrm{d}x)\\ &= \pi(B)\int_{\mathbf{X}}\sum_{\theta_0\in\Theta} P_{\theta_0}(x,F)\pi_0(\theta_0)\phi(\mathrm{d}x)\\ &= \pi(B)\int_{\mathbf{X}} P(x,F)\phi(\mathrm{d}x) = \pi(B)\phi(F) = \psi(F\times B)\end{aligned}$$

where we have exploited the fact that ϕ is invariant. ■

This Proposition is a partial extension of the Theorem 5 in [100], which states the same result in the case of denumerable state space $\mathbf{X}$ and attests also the inverse implication (that is, all the invariant p.m.'s are product measures of kind (4.26) still in the denumerable framework.

For a more detailed treatise on MCRE's, we refer the reader to [32, 33, 100, 101].

4.7.2 OSA: Proof of Theorem 3

From equation (4.18), $(D_k)_{k\in\mathbb{N}}$ with $(U_k)_{k\in\mathbb{N}}$ turns out to be a countable MCRE. This is the right way to look at $(D_k)_{k\in\mathbb{N}}$ if we want to understand its behavior with respect to typical instances of the input $\mathbf{U} = (U_0, U_1, \dots)$. For any $x, y\in\mathbb{Z}$, we have

$$P(D_{k+1} = y | D_k = x, D_{k-1},\dots,D_0;\mathbf{U}) = \mathbf{P}_{x,y}(U_k).$$

Consider the space $(\mathbb{Z}\times\{0,1\}^{\mathbb{N}}, \mathcal{P}(\mathbb{Z})\times\prod_0^\infty\mathcal{P}(\{0,1\}))$ endowed with the initial distribution $\kappa\times\pi$, where κ is the counting measure on $\mathbb{Z}$ and π is the usual uniform Bernoulli measure on $\{0,1\}^{\mathbb{N}}$. Given $x,y\in\mathbb{Z}$, $\mathbf{u} = (u_0,u_1,\dots)\in\{0,1\}^{\mathbb{N}}$ and $B\in\prod_0^\infty\mathcal{P}(\{0,1\})$, the EMP is defined by the transition probability kernel

$$P\big((x,\mathbf{u});\{y\}\times B\big) = \mathbf{P}_{x,y}(u_0)\mathbb{1}_B(T\mathbf{u}). \tag{4.27}$$

By Proposition 6, an invariant probability measure exists for our EMP and we explicitly compute it: in fact, let ϕ be a p.m. on $(\mathbb{Z},\mathcal{P}(\mathbb{Z}))$ given by $\phi(\{d\}) = \Phi_d$, Φ_d being the

invariant p.v. defined in the Lemma 2, for any integer d. Then, $\psi = \phi \times \pi$ is an invariant p.m. for the EMP.
We can verify that ψ is ergodic by the following criterion (see Chapter 3 of [33]). Let $\mathbf{P}^{(n)}(U_0, \ldots U_{n-1})$ the n-step transition matrix whose entries are

$$\mathbf{P}^{(n)}_{x,y}(U_0, \ldots U_{n-1}) = \mathrm{P}(D_n = y | D_0 = x, U_0, \ldots U_{n-1}). \tag{4.28}$$

If for each $x, y \in \mathbb{Z}$ and π-a.e. $\mathbf{U}$ there exist $n = n(x, y, \mathbf{U}) \in \mathbb{N}$ and $z = z(x, y, \mathbf{U}, n) \in \mathbb{Z}$ such that $\mathbf{P}^{(n)}_{x,z}(U_0, \ldots, U_{n-1})\mathbf{P}^{(n)}_{y,z}(U_0, \ldots U_{n-1}) > 0$, then ψ is ergodic. In our context it is easy to check that given any couple of starting states x and y, after $n > |x - y|$ steps we have a non-null probability of having joined a common state z.
Define $\mathrm{q}_d(U_k) = P[\widehat{U}_k \neq U_k | D_k = d, U_k] = \mathbf{P}_{d,d+1}(U_k) + \mathbf{P}_{d,d-1}(U_k)$ ($\overline{\mathrm{q}}_d$ is actually the mean of q_d). For any $K \in \mathbb{N}$ and given $D_0 = 0$, the CBER can be expressed as follows:

$$\mathrm{CBER}(\mathbb{D}^{(1)}) = \frac{1}{K} \sum_{k=0}^{K-1} \sum_{d \in \mathbb{Z}} \mathrm{q}_d(U_k) \mathbf{P}^{(n)}_{0,d}(U_0, U_1, \ldots U_{k-1}) \tag{4.29}$$

Notice that, since the U_k's $k \in \mathbb{N}$ are independent, $\mathbf{P}^{(n)}(U_0, U_1, \ldots, U_{k-1}) = \mathbf{P}(U_0)\mathbf{P}(U_1) \cdots \mathbf{P}(U_{k-1})$.

Consider $\omega = (x, \mathbf{U})$ and the function $g(\omega) = \mathrm{q}_x(U_0)$: we have that

$$\sum_{d \in \mathbb{Z}} \mathrm{q}_d(U_k) \mathbf{P}_{x,d}(U_0, \ldots U_{k-1}) = (P^k g)(x, \mathbf{U})$$

and notice that

$$\mathrm{CBER}(\mathbb{D}^{(1)}) = \frac{1}{K} \sum_{k=0}^{K-1} (P^k g)(0, \mathbf{U}). \tag{4.30}$$

Now, by the Ergodic Theorem 7:

$$\lim_{K \to \infty} \frac{1}{K} \sum_{k=0}^{K-1} P^k g(\omega) = \int_{\mathbb{Z} \times \{0,1\}^{\mathbb{N}}} g(\omega) \psi(\mathrm{d}\omega) \quad \text{for } \psi\text{-a.e. } \omega. \tag{4.31}$$

Notice that, as pointed out after Lemma 2, $\phi(\{d\}) > 0$ for any $d \in \mathbb{Z}$; then, a set $\{d\} \times B$, $d \in \mathbb{Z}$, $B \subset \{0,1\}^{\mathbb{N}}$, is ψ-negligible if and only if $\pi(B) = 0$. Hence, in (4.31), "ψ-a.e. ω" is equivalent to "for any $d \in \mathbb{Z}$ and π-a.e. $\mathbf{U}$".

This, along with the equality (4.30), implies that

$$\lim_{K \to \infty} \mathrm{CBER}(\mathbb{D}^{(1)}) = \int_{\mathbb{Z} \times \{0,1\}^{\mathbb{N}}} g(\omega) \psi(\mathrm{d}\omega) \quad \text{for } \pi\text{-a.e. } \mathbf{U}. \tag{4.32}$$

Finally, recalling that $\psi = \phi \times \pi$,

$$\int_{\mathbb{Z} \times \{0,1\}^{\mathbb{N}}} g(\omega) \psi(\mathrm{d}\omega) = \sum_{d \in \mathbb{Z}} \sum_{U_0 = 0,1} \mathrm{q}_d(U_0) \pi(U_0) \Phi_d = \sum_{d \in \mathbb{Z}} \overline{\mathrm{q}}_d \Phi_d.$$

4.7.3 TSA: Computation of the Transition Probabilities

In the next pages, we compute the probability of moving from a state $(\alpha, d) \in [0,1] \times \mathbb{Z}$ to a set of type $(0,\beta) \times \{d'\}$, $\beta \in (0,1], d' \in \mathbb{Z}$, for the Markov Process $(A_k, D_k)_{k \in \mathbb{N}}$ defined in Section 3.2. Let $P_u\big((\alpha,d),(0,\beta)\times\{d'\}\big)$ be the transition probability given the transmitted bit u: $P\big((\alpha,d),(0,\beta)\times\{d'\}\big) = \frac{1}{2}P_0\big((\alpha,d),(0,\beta)\times\{d'\}\big)+\frac{1}{2}P_1\big((\alpha,d),(0,\beta)\times\{d'\}\big)$ are null if $d' \notin \{d-1,d,d+1\}$, if $d' = d+1$ and $u=1$ or if $d'=d-1$ and $u=0$; we now compute the non-null instances. Given $(\alpha,d) \in (0,1)\times\mathbb{Z}$ and $x \in \{\alpha,(1-\alpha)^{-1},1\}$, $y \in \{d-1,d,d+1\}$, $z \in (0,1)$, we define:

$$\begin{aligned} c_\alpha &= \sqrt{\frac{\exp(1/\sigma^2)}{\alpha(1-\alpha)}} \\ h_{x,y}(z) &= \frac{\sigma^2 \log\left(x\frac{1-z}{z}\right) + y + \frac{1}{2}}{\sigma\sqrt{2}} \\ H_{x,y}(z) &= \frac{1}{2}\operatorname{erfc}\left(h_{x,y}(z)\right). \end{aligned} \tag{4.33}$$

Notice that these quantities depend on the noise variance σ^2, even if the notation does not emphasize that.

Case 1: $d' = d, u = 0$.

$$\begin{aligned} &P_0\big((\alpha,d),(0,\beta)\times\{d\}\big) = \mathrm{Prob}(\zeta_3 \le \zeta_1 \le \beta(\zeta_1+\zeta_2)|A_k=\alpha, D_k=d, U_k=0) \\ &= \begin{cases} 0 & \text{if } \alpha=0 \text{ or if } \alpha\in(0,1) \text{ and } \beta \le \frac{1}{1+c_\alpha} \\ H_{\alpha,d}(\beta) - H_{\alpha,d}\left(\frac{1}{1+c_\alpha}\right) & \text{if } \alpha \in (0,1) \text{ and } \beta > \frac{1}{1+c_\alpha} \\ H_{1,d}(\beta) & \text{if } \alpha = 1. \end{cases} \end{aligned} \tag{4.34}$$

Case 2: $d' = d, u = 1$.

$$\begin{aligned} &P_1\big((\alpha,d),(0,\beta)\times\{d\}\big) = \\ &\qquad = \mathrm{Prob}\big((\zeta_3 \ge \zeta_1)\cap(\beta\zeta_3 \ge (1-\beta)\zeta_2)|A_k=\alpha, D_k=d, U_k=1\big) \\ &= \begin{cases} H_{\frac{1}{1-\alpha},d}(\beta) & \text{if } \alpha=0 \text{ or if } \alpha\in(0,1) \text{ and } \beta \le \frac{c_\alpha}{1+c_\alpha} \\ H_{\frac{1}{1-\alpha},d}\left(\frac{c_\alpha}{1+c_\alpha}\right) & \text{if } \alpha \in (0,1) \text{ and } \beta > \frac{c_\alpha}{1+c_\alpha} \\ 0 & \text{if } \alpha = 1. \end{cases} \end{aligned} \tag{4.35}$$

Case 3: $d' = d+1, u = 0$.

$$\begin{aligned} &P_0\big((\alpha,d),(0,\beta)\times\{d+1\}\big) = \\ &\qquad = \mathrm{Prob}\big((\zeta_3 \ge \zeta_1)\cap(\beta\zeta_3 \ge (1-\beta)\zeta_2)|A_k=\alpha, D_k=d, U_k=0\big) \\ &= \begin{cases} H_{\frac{1}{1-\alpha},d+1}(\beta) & \text{if } \alpha=0 \text{ or if } \alpha\in(0,1) \text{ and } \beta \le \frac{c_\alpha}{1+c_\alpha} \\ H_{\frac{1}{1-\alpha},d+1}\left(\frac{c_\alpha}{1+c_\alpha}\right) & \text{if } \alpha \in (0,1) \text{ and } \beta > \frac{c_\alpha}{1+c_\alpha} \\ 0 & \text{if } \alpha = 1. \end{cases} \end{aligned} \tag{4.36}$$

Case 4: $d' = d-1, u = 1$.

$$P_1\big((\alpha,d),(0,\beta)\times\{d-1\}\big) =$$
$$= \mathrm{Prob}(\zeta_3 \le \zeta_1 \le \beta(\zeta_1+\zeta_2)|A_k=\alpha, D_k=d, U_k=1)$$
$$= \begin{cases} 0 & \text{if } \alpha=0 \text{ or if } \alpha\in(0,1) \text{ and } \beta\le\frac{1}{1+c_\alpha} \\ H_{\alpha,d-1}(\beta) - H_{\alpha,d-1}\left(\frac{1}{1+c_\alpha}\right) & \text{if } \alpha\in(0,1) \text{ and } \beta>\frac{1}{1+c_\alpha} \\ H_{1,d-1}(\beta) & \text{if } \alpha=1. \end{cases} \tag{4.37}$$

Remark 5 *: Since $c_\alpha > 2$, then $\frac{1}{1+c_\alpha} < \frac{1}{3} < \frac{2}{3} < \frac{c_\alpha}{1+c_\alpha}$.*

In summary:

$$P\big((\alpha,d),(0,\beta)\times\{d\}\big) =$$
$$\frac{1}{2}\begin{cases} H_{1,d}(\beta) & \text{if } \alpha=0 \text{ or if } \alpha=1 \\ H_{\frac{1}{1-\alpha},d}(\beta) & \text{if } \alpha\in(0,1) \text{ and } \beta\le\frac{1}{1+c_\alpha} \\ H_{\alpha,d}(\beta) - H_{\alpha,d}\left(\frac{1}{1+c_\alpha}\right) + H_{\frac{1}{1-\alpha},d}(\beta) & \\ \qquad \text{if } \alpha\in(0,1) \text{ and } \frac{1}{1+c_\alpha}<\beta\le\frac{c_\alpha}{1+c_\alpha} & \\ H_{\alpha,d}(\beta) - H_{\alpha,d}\left(\frac{1}{1+c_\alpha}\right) + H_{\frac{1}{1-\alpha},d}\left(\frac{c_\alpha}{1+c_\alpha}\right) & \\ \qquad \text{if } \alpha\in(0,1) \text{ and } \beta>\frac{c_\alpha}{1+c_\alpha} & \end{cases} \tag{4.38}$$

$$P\big((\alpha,d),(0,\beta)\times\{d+1\}\big) =$$
$$\frac{1}{2}\begin{cases} H_{\frac{1}{1-\alpha},d+1}(\beta) & \text{if } \alpha=0 \text{ or if } \alpha\in(0,1) \text{ and } \beta\le\frac{c_\alpha}{1+c_\alpha} \\ H_{\frac{1}{1-\alpha},d+1}\left(\frac{c_\alpha}{1+c_\alpha}\right) & \text{if } \alpha\in(0,1) \text{ and } \beta>\frac{c_\alpha}{1+c_\alpha} \\ 0 & \text{if } \alpha=1 \end{cases} \tag{4.39}$$

$$P\big((\alpha,d),(0,\beta)\times\{d-1\}\big) =$$
$$\frac{1}{2}\begin{cases} 0 & \text{if } \alpha=0 \text{ or if } \alpha\in(0,1) \text{ and } \beta\le\frac{1}{1+c_\alpha} \\ H_{\alpha,d-1}(\beta) - H_{\alpha,d-1}\left(\frac{1}{1+c_\alpha}\right) & \text{if } \alpha\in(0,1) \text{ and } \beta>\frac{1}{1+c_\alpha} \\ H_{1,d-1}(\beta) & \text{if } \alpha=1. \end{cases} \tag{4.40}$$

4.7.4 TSA: Computation of $\overline{q}(\alpha,d)$

The function $\overline{q}$ on $[0,1]\times\mathbb{Z}$ defined in the Corollary 5 is given by $\overline{q}(\alpha,d) = \frac{1}{2}\mathrm{P}(\widehat{U}_k = 1|U_k=0, A_k=\alpha, D_k=d) + \frac{1}{2}\mathrm{P}(\widehat{U}_k=0|U_k=1,A_k=\alpha,D_k=d)$. Note that

$$\mathrm{P}(\widehat{U}_k=1|U_k=0,A_k=\alpha,D_k=d) =$$
$$= \mathrm{Prob}\big(\alpha f_{(Y_{k+1}|X_{k+1})}(y_{k+1}|\widehat{x}_k+1) + (1-\alpha)f_{(Y_{k+1}|X_{k+1})}(y_{k+1}|\widehat{x}_k+2)$$
$$> \alpha f_{(Y_{k+1}|X_{k+1})}(y_{k+1}|\widehat{x}_k) + (1-\alpha)f_{(Y_{k+1}|X_{k+1})}(y_{k+1}|\widehat{x}_k+1)\big)$$
$$= \frac{1}{2}\mathrm{erfc}\left(\frac{\sigma^2\log z_1 + d + \frac{1}{2}}{\sqrt{2}\sigma}\right)$$

where z_1 is the positive solution of the equation $(1-\alpha)e^{-\frac{1}{\sigma^2}}z^2 + (2\alpha-1)z - \alpha = 0$. Similarly,

$$\mathrm{P}(\widehat{U}_k = 0 | U_k = 1, A_k = \alpha, D_k = d) = 1 - \frac{1}{2}\mathrm{erfc}\left(\frac{\sigma^2 \log z_1 + d - \frac{1}{2}}{\sqrt{2}\sigma}\right)$$

hence

$$\overline{q}(\alpha, d) = \frac{1}{2}\left[\frac{1}{2}\mathrm{erfc}\left(\frac{\sigma^2 \log z_1 + d + \frac{1}{2}}{\sqrt{2}\sigma}\right) + 1 - \frac{1}{2}\mathrm{erfc}\left(\frac{\sigma^2 \log z_1 + d - \frac{1}{2}}{\sqrt{2}\sigma}\right)\right].$$

Naturally, if $\alpha = 1$, then $\overline{q}(\alpha, d) = \frac{1}{2}\left[\frac{1}{2}\mathrm{erfc}\left(\frac{d+\frac{1}{2}}{\sqrt{2}\sigma}\right) + 1 - \frac{1}{2}\mathrm{erfc}\left(\frac{d-\frac{1}{2}}{\sqrt{2}\sigma}\right)\right] = \overline{\mathrm{q}}_d$ and we reduce to the One State case.

4.7.5 TSA: Proof of the Lemma 3

We prove that the kernel of (A_k, D_k) satisfies both the Weak Feller Property and the Drift Condition; the result will then follow from Proposition 1. First, we check the Drift Condition. By equations (4.38)- (4.40) in the Appendix,

$$\begin{aligned} P\big((\alpha, d), [0,1] \times \{d+1\}\big) &= \frac{1}{4}\mathrm{erfc}\left(\frac{\sigma^2 \log\sqrt{\frac{\alpha}{1-\alpha}} + d + 1}{\sigma\sqrt{2}}\right) \\ P\big((\alpha, d), [0,1] \times \{d-1\}\big) &= \frac{1}{2} - \frac{1}{4}\mathrm{erfc}\left(\frac{\sigma^2 \log\sqrt{\frac{\alpha}{1-\alpha}} + d}{\sigma\sqrt{2}}\right). \end{aligned} \tag{4.41}$$

In particular, $P\big((\alpha, d), [0,1] \times \{d+1\}\big)$ and $P\big((\alpha, d), [0,1] \times \{d-1\}\big)$ have values in $[0, 1/2]$ and are monotone respectively decreasing and increasing with respect to α. Now, let us define

$$\delta_d = \frac{1}{2(|d| + 10)} \tag{4.42}$$

and

$$V(\alpha, d) = \begin{cases} d^2 & \text{if } d \geq 0, \alpha \geq \delta_d \text{ or if } d < 0, \alpha \leq 1 - \delta_d; \\ d^2 + 2|d| & \text{otherwise.} \end{cases} \tag{4.43}$$

We are going to prove that V fulfills the Drift inequality for some compact C:

$$\Delta V(\alpha, d) = \int_{[0,1]\times\mathbb{Z}} P\big((\alpha, d), \mathrm{d}(\alpha', d')\big)V(\alpha', d') - V(\alpha, d) \leq -1 + b\mathbb{1}_{\mathrm{C}}(\alpha, d) \tag{4.44}$$

for every $(\alpha, d) \in [0,1] \times \mathbb{Z}$. In order to individuate C, let us find out the values of (α, d) such that (4.44) holds with $\mathbb{1}_{\mathrm{C}}(\alpha, d) = 0$. Recall that $P\big((\alpha, d),\ A \times \{d'\}\big) > 0 \Rightarrow$

$d' \in \{d-1, d, d+1\}$ for any $\alpha \in [0,1]$, $A \in \mathcal{B}([0,1])$.
In the next, let us use the notation $\omega = (\alpha, d)$, $\omega' = (\alpha', d')$.

If $d \geq 0$,

$$\begin{aligned}
\Delta V(\omega) &= \int_0^1 \sum_{d'=d-1}^{d+1} P(\omega, (\mathrm{d}\alpha', d'))V(\omega') - V(\omega) \\
&= \sum_{d'=d-1}^{d+1} \left[\int_0^{\delta_d} P(\omega, (\mathrm{d}\alpha', d'))(2d' + d'^2) + \int_{\delta_d}^1 P(\omega, (\mathrm{d}\alpha', d'))d'^2\right] - V(\omega) \\
&= \sum_{d'=d-1}^{d+1} \left[\int_0^1 P(\omega, (\mathrm{d}\alpha', d'))d'^2 + \int_0^{\delta_d} P(\omega, (\mathrm{d}\alpha', d'))2d'\right] - V(\omega) \\
&= \sum_{d'=d-1}^{d+1} \left[P(\omega, [0,1] \times \{d'\})d'^2 + P(\omega, [0, \delta_d] \times \{d'\})2d'\right] - V(\omega) \\
&= d^2 + 2d[P(\omega, [0,1] \times \{d+1\}) - P(\omega, ([0,1] \times \{d-1\})] \\
&\quad + P(\omega, [0,1] \times \{d+1\}) + P(\omega, [0,1] \times \{d-1\}) + 2dP(\omega, [0, \delta_d] \times \mathbb{Z}) \\
&\quad + 2[P(\omega, [0, \delta_d] \times \{d+1\}) - P(\omega, [0, \delta_d] \times \{d-1\})] - V(\omega).
\end{aligned}$$

As $P(\omega, [0,1] \times \{d+1\}) + P(\omega, [0,1] \times \{d-1\}) \leq \frac{1}{2}$ (see equations (4.41)) and $P(\omega, [\beta_1, \beta_2] \times \mathbb{Z}) \leq G(\beta_2 - \beta_1)$ (see Lemma 6 in the Appendix 4.7.7).

$$\begin{aligned}
\Delta V(\omega) &\leq d^2 + 2d[P(\omega, [0,1] \times \{d+1\}) - P(\omega, [0,1] \times \{d-1\})] + \frac{1}{2} \\
&\quad + 2(d+1)G\delta_d - V(\omega) \qquad (4.45) \\
&\leq d^2 + 2d[P(\omega, [0,1] \times \{d+1\}) - P(\omega, [0,1] \times \{d-1\})] + \frac{1}{2} + G - V(\omega)
\end{aligned}$$

where we exploited that $2(d+1)G\delta_d < G$ by the definition (4.42) of δ_d.

If $d < 0$, by analogous computation we obtain again the inequality (4.45). Let us study the behavior of this bound for every $\omega \in [0,1] \times \mathbb{Z}$, according to the partition of $[0,1] \times \mathbb{Z}$ into four subsets given by the definition of V.

Subset 1: If $d \geq 0$ and $\alpha \geq \delta_d$, $V(\omega) = d^2$ and

$$P(\omega, [0,1] \times \{d+1\}) \leq \frac{1}{4}\mathrm{erfc}\left(\frac{\sigma^2 \log\sqrt{\frac{\delta_d}{1-\delta_d}} + d}{\sigma\sqrt{2}}\right)$$

$$P(\omega, [0,1] \times \{d-1\}) \geq \frac{1}{2} - \frac{1}{4}\mathrm{erfc}\left(\frac{\sigma^2 \log\sqrt{\frac{\delta_d}{1-\delta_d}} + d}{\sigma\sqrt{2}}\right)$$

hence inequality (4.45) becomes

$$\Delta V(\omega) \leq G + d\left[\operatorname{erfc}\left(\frac{\sigma^2 \log\sqrt{\frac{\delta_d}{1-\delta_d}} + d}{\sigma\sqrt{2}}\right) - 1\right] + \frac{1}{2}$$
$$= G + d\left[\operatorname{erfc}\left(\frac{-\frac{\sigma^2}{2}\log(2d+19) + d}{\sigma\sqrt{2}}\right) - 1\right] + \frac{1}{2}.$$

As erfc$(x) \in (0,1)$ whenever $x > 0$, then for d is sufficiently large the quantity in the square bracket is negative. Moreover, this quantity is multiplied by d; hence, there necessarily exists an integer $d_0^+ > 0$, depending on the noise σ, such that for any $d > d_0^+$, $\Delta V(\omega) \leq -1$.

Subset 2: If $d < 0$ and $\alpha \leq 1 - \delta_d$,

$$P(\omega, [0,1] \times \{d+1\}) \geq \frac{1}{4}\operatorname{erfc}\left(\frac{-\sigma^2 \log\sqrt{\frac{\delta_d}{1-\delta_d}} + d + 1}{\sigma\sqrt{2}}\right)$$
$$P(\omega, [0,1] \times \{d-1\}) \leq \frac{1}{2} - \frac{1}{4}\operatorname{erfc}\left(\frac{-\sigma^2 \log\sqrt{\frac{\delta_d}{1-\delta_d}} + d + 1}{\sigma\sqrt{2}}\right)$$

hence inequality (4.45) becomes

$$\Delta V(\omega) \leq G + d\left[\operatorname{erfc}\left(\frac{-\sigma^2 \log\sqrt{\frac{\delta_d}{1-\delta_d}} + d + 1}{\sigma\sqrt{2}}\right) - 1\right] + \frac{1}{2}$$
$$= G + d\left[\operatorname{erfc}\left(\frac{\frac{\sigma^2}{2}\log(-2d+19) + d + 1}{\sigma\sqrt{2}}\right) - 1\right] + \frac{1}{2}.$$

The computation is now analogous to the previous case and we conclude that there necessarily exists an integer $d_0^- < 0$, depending on the noise, such that for any $d < d_0^-$, $\Delta V(\omega) \leq -1$.

Subset 3: If $d \geq 0$ and $\alpha < \delta_d$, $V(\omega) = d^2 + 2d$; moreover, we have no tight bounds for $P(\omega, [0,1] \times \{d+1\})$ and $P(\omega, [0,1] \times \{d-1\})$: we can just notice that their difference is smaller than $\frac{1}{2}$. Substituting it in (4.45) we obtain

$$\Delta V(\omega) \leq d^2 + G + \frac{1}{2} + d - d^2 - 2d = G + \frac{1}{2} - d$$

hence $\Delta V(\omega) \leq -1$ if $d > d_1 = G + \frac{3}{2}$.

Subset 4: If $d < 0$ and $\alpha > 1 - \delta_d$, $V(\omega) = d^2 - 2d$; as $P(\omega, [0,1] \times \{d+1\}) - P(\omega, [0,1] \times \{d-1\}) \geq -\frac{1}{2}$,

$$\Delta V(\omega) \leq G + \frac{1}{2} + d$$

and $\Delta V \leq -1$ if $d < -d_1$.

Now, it is easy to verify that the subsets of $[0,1] \times \mathbb{Z}$ not yet considered form the compact set $\big([0, \delta_d] \times \{0, \dots, d_1\}\big) \cup \big([\delta_d, 1] \times \{0, \dots, d_0^+\}\big) \cup \big([0, 1-\delta_d] \times \{d_0^-, \dots, -1, 0\}\big) \cup \big([1-\delta_d, 1] \times \{-d^1, \dots, -1, 0\}\big)$. For simplicity, we can consider the bigger compact set $\mathrm{C} = [0,1] \times \{-d_{\mathrm{C}}, \dots, d_{\mathrm{C}}\}$, where $d_{\mathrm{C}} = \max\{d_0^+, -d_0^-, d_1\}$: now, it is easy to check that for any $\omega \in \mathrm{C}$ the Drift Condition is satisfied whenever $b \geq G + d_{\mathrm{C}} + \frac{3}{2}$.

We now check the Weak Feller Property. Given any open interval $I \subset [0,1]$ and $d' \in \mathbb{Z}$, the continuity of $P(\cdot, I \times \{d'\})$ can be easily verified by the equations (4.38)-(4.40) (Section 4.7.3): $P((\alpha, d), I \times \{d'\})$ is piecewise defined as combination of H, which is a continuous function; moreover, it is straightforward to check that the continuity holds also at the connection points. Furthermore,
(a) any open set on the real line (hence on $[0,1]$) is a countable union of disjoint intervals;
(b) if f_N is a monotone increasing sequence of lower semicontinuous functions such that $f_N \uparrow f$ pointwise, then f is lower semicontinuous.

By (a), any open set O in $[0,1]$ can be expressed as $O = \cup_{n=1}^{\infty} I_n$, with I_n mutually disjoint open intervals in $[0,1]$. Moreover, $f_N(\omega) = P(\omega, (\cup_{n=1}^{N} I_n) \times \{d'\}) \leq 1$ fulfills the hypotheses of statement (b), hence its pointwise limit $f(\omega) = P(\omega, (\cup_{i=1}^{\infty} I_i) \times \{d'\}) = P(\omega, O \times \{d'\})$ is lower semicontinuous. As any open set of the product topology can be expressed as $\cup_{n \in \mathbb{Z}} (O_n \times \{n\})$, O_n open in $[0,1]$, the lower semicontinuity is extended to all the open sets.

4.7.6 TSA: Proof of the $(\mathcal{L} \times \kappa)$-irreducibility of $(A_k, D_k)_{k \in \mathbb{N}}$

In this paragraph, we complete the proof of the Lemma 4 and then of the Theorem 5 showing the $(\mathcal{L} \times \kappa)$-irreducibility of $(A_k, D_k)_{k \in \mathbb{N}}$ in the space $([0,1] \times \mathbb{Z}, \mathcal{B}([0,1]) \times \mathcal{P}(\mathbb{Z}))$. For this purpose, we first prove that any non-negligible Borel subset of kind $M \times \{d'\} \subset [0,1] \times \mathbb{Z}$ is achievable with positive probability from any (α, d), in one or two steps, if $d' \in \{d-1, d, d+1\}$ and M is sufficiently far from the extreme points of $[0,1]$:

Lemma 5 *For any $\varepsilon > 0$, $d \in \mathbb{Z}$, there exists a constant $C_{\varepsilon,d} > 0$ such that the following inequalities hold for every $(\alpha, d) \in [0,1] \times \mathbb{Z}$ and $M \in \mathcal{B}([\varepsilon, 1-\varepsilon])$:*

$$P\big((\alpha, d), M \times \{d\}\big) \geq C_{\varepsilon,d} \mathcal{L}(M)$$
$$P^2\big((\alpha, d), M \times \{d+1\}\big) \geq C_{\varepsilon,d} \mathcal{L}(M)$$
$$P^2\big((\alpha, d), M \times \{d-1\}\big) \geq C_{\varepsilon,d} \mathcal{L}(M)$$

where $\mathcal{L}$ is the Lebesgue measure.

Proof First, we prove the lemma on the open intervals $(\beta_1, \beta_2) \subset [\varepsilon, 1-\varepsilon]$. Let $\bar{\alpha} = \frac{1}{1-\alpha}$. Consider the first inequality. On the basis of the equations (4.38) and Remark 5, the following cases may occur:

Case 1: If $\alpha = 0$, $(\beta_1, \beta_2) \subset [\varepsilon, 1-\varepsilon]$ or if $\alpha \in (0,1)$, $(\beta_1, \beta_2) \subset [\varepsilon, \frac{1}{1+c_\alpha}] \subseteq [\varepsilon, \frac{1}{3}]$:

$$\begin{aligned} P\big((\alpha,d),(\beta_1,\beta_2)\times\{d\}\big) &= \frac{1}{2}H_{\bar{\alpha},d}(\beta_2) - \frac{1}{2}H_{\bar{\alpha},d}(\beta_1) \\ &= \frac{1}{2\sqrt{\pi}}\int_{h_{\bar{\alpha},d}(\beta_2)}^{h_{\bar{\alpha},d}(\beta_1)} e^{-t^2}\mathrm{d}t \\ &= -\frac{1}{2\sqrt{\pi}}\int_{\beta_1}^{\beta_2} e^{-h^2_{\bar{\alpha},d}(z)}\frac{\partial}{\partial z}h_{\bar{\alpha},d}(z)\mathrm{d}z \\ &\geq \frac{1}{2\sqrt{\pi}}(\beta_2-\beta_1)\min_{z\in(\beta_1,\beta_2)}\left(-e^{-h^2_{\bar{\alpha},d}(z)}\frac{\partial}{\partial z}h_{\bar{\alpha},d}(z)\right) \\ &\geq \frac{1}{2\sqrt{\pi}}(\beta_2-\beta_1)\min_{z\in(\beta_1,\beta_2)}\left(-\frac{\partial}{\partial z}h_{\bar{\alpha},d}(z)\right)\min\left\{e^{-h^2_{\bar{\alpha},d}(\beta_1)}, e^{-h^2_{\bar{\alpha},d}(\beta_2)}\right\}. \end{aligned}$$

By definition (4.33), for any x, y, $\frac{\partial}{\partial z}h_{x,y}(z) = \frac{\sigma}{z(z-1)\sqrt{2}} \leq -\sigma 2\sqrt{2}$; moreover,

$$\min\left\{e^{-h^2_{\bar{\alpha},d}(\beta_1)}, e^{-h^2_{\bar{\alpha},d}(\beta_2)}\right\} \geq \min\left\{e^{-h^2_{\bar{\alpha},d}(\varepsilon)}, e^{-h^2_{\bar{\alpha},d}(1-\varepsilon)}\right\} =: m_{\bar{\alpha},d}.$$

Notice now that for any $d \in \mathbb{Z}$, $m_{\bar{\alpha},d} \to 0$ if and only if $\alpha \to 1$; nevertheless, if $\alpha \to 1$, also $(1+c_\alpha)^{-1} \to 0$ and in particular there will be some α such that $(1+c_\alpha)^{-1} < \varepsilon$, which contradicts the hypothesis $\beta_1 \geq \varepsilon$. Hence, we can conclude that

$$P\big((\alpha,d),(\beta_1,\beta_2)\times\{d\}\big) \geq \sigma\sqrt{2/\pi}\ \min_{\alpha} m_{\alpha,d}(\beta_2-\beta_1) > 0$$

where the minimum has to be computed for α satisfying the initial hypotheses.

Case 2: If $\alpha = 1$,$(\beta_1, \beta_2) \in [\varepsilon, 1-\varepsilon]$ or if $\alpha \in (0,1)$, $(\beta_1, \beta_2) \subset [\frac{c_\alpha}{1+c_\alpha}, 1-\varepsilon] \subseteq [\frac{2}{3}, 1-\varepsilon]$: by analogous procedure, we obtain

$$P\big((\alpha,d),(\beta_1,\beta_2)\times\{d\}\big) \geq \sigma\sqrt{2/\pi}\ \min_{\alpha} m_{\alpha,d}(\beta_2-\beta_1) > 0$$

where $m_{\alpha,d} = \min\left\{e^{-h^2_{\alpha,d}(\varepsilon)}, e^{-h^2_{\alpha,d}(1-\varepsilon)}\right\} > 0$ and its minimum is computed for α satisfying the above hypotheses. The positiveness holds since for any $d \in \mathbb{Z}$, $m_{\alpha,d} \to 0$ if and only if $\alpha \to 0$, which implies $\frac{c_\alpha}{1+c_\alpha} \to 1$ and contradicts $\beta_2 \leq 1-\varepsilon$.

Case 3: Otherwise: it is straightforward to verify that

$$P\big((\alpha,d),(\beta_1,\beta_2)\times\{d\}\big) \geq \sigma\sqrt{2/\pi}\ (m_{\alpha,d} + m_{\bar{\alpha},d})(\beta_2-\beta_1).$$

Finally, if we consider

$$\bar{m}(\alpha,d,\beta_1,\beta_2) = \begin{cases} m_{\bar{\alpha},d} \text{ if } \alpha = 0 \text{ or if } \alpha \in (0,1) \text{ and } \varepsilon < \beta_1 < \beta_2 \leq \frac{1}{1+c_\alpha}; \\ m_{\alpha,d} \text{ if } \alpha = 1 \text{ or if } \alpha \in (0,1) \text{ and } \frac{c_\alpha}{1+c_\alpha} < \beta_1 < \beta_2 \leq 1-\varepsilon; \\ m_{\alpha,d} + m_{\bar{\alpha},d} \text{ otherwise.} \end{cases} \tag{4.46}$$

and

$$C^{(1)}_{\varepsilon,d} = \sigma\sqrt{\frac{2}{\pi}} \min_{\substack{\alpha\in[0,1]\\(\beta_1,\beta_2)\subset[\varepsilon,1-\varepsilon]}} \bar{m}(\alpha,d,\beta_1,\beta_2) \tag{4.47}$$

we conclude that for any $\varepsilon > 0$, $d \in \mathbb{Z}$,

$$P\big((\alpha,d),(\beta_1,\beta_2)\times\{d\}\big) \geq C^{(1)}_{\varepsilon,d}(\beta_2-\beta_1) \qquad C^{(1)}_{\varepsilon,d} > 0. \tag{4.48}$$

Let us prove the second inequality, on the basis of equations (4.39). In this case, the component d of the state moves to $d+1$, which is not always possible in one step. In particular, there are two situations in which the transition probability is null: $\alpha = 1$ and when $\beta_1 = \frac{c_\alpha}{1+c_\alpha}$ (and given the continuity of (4.39), problems occur whenever $\alpha \to 1$ or $\beta_1 \to \frac{c_\alpha}{1+c_\alpha}$).

Both issues can be solved considering two-step transition: roughly speaking, if α is close to 1, a first step is used to move α away from 1 (and d remains constant); at this point, the probability to move d to $d+1$ is positive. On the other hand, when β_1 is close to $\frac{c_\alpha}{1+c_\alpha}$ a first step is used to move d to $d+1$ and a second one to move the component α to the desired interval (and now this is possible since we reduce to the case in which d remains constant, previously studied).

Let us assess this qualitative argumentation.

Case 1: If $\alpha = 0$, $(\beta_1,\beta_2) \subset [\varepsilon, 1-\varepsilon]$ or if $\alpha \in (0, 1-\delta_1]$ for some small $\delta_1 > 0$, $(\beta_1,\beta_2) \subset [\varepsilon, \frac{c_\alpha}{1+c_\alpha}]$ $\left(\frac{c_\alpha}{1+c_\alpha} \leq 1-\varepsilon\right)$:

$$P\big((\alpha,d),(\beta_1,\beta_2)\times\{d+1\}\big) \geq \sigma\sqrt{2/\pi} \min_{\alpha\in[0,1-\delta_1]} m_{\bar{\alpha},d+1}(\beta_2-\beta_1) > 0 \tag{4.49}$$

where the positiveness of $\min_{\alpha\in[0,1-\delta_1]} m_{\bar{\alpha},d+1} > 0$ has been discussed above.

Case 2: If $\alpha \in (0,1-\delta_1]$, $\beta_1 \in [\varepsilon, \frac{c_\alpha}{1+c_\alpha} - \delta_2]$ for some small $\delta_1, \delta_2 > 0$ and $\beta_2 \in [\frac{c_\alpha}{1+c_\alpha}, 1-\varepsilon]$: the transition probability depends on β_1, not on β_2, and

$$P\big((\alpha,d),(\beta_1,\beta_2)\times\{d+1\}\big) \geq \sigma\sqrt{2/\pi} \min_{\alpha\in(0,1-\delta_1]} m_{\bar{\alpha},d+1}\left(\frac{c_\alpha}{1+c_\alpha} - \beta_1\right)$$

where $\frac{c_\alpha}{1+c_\alpha} - \beta_1 \geq \delta_2 \geq \delta_2(\beta_2-\beta_1)$.

Let us now consider the cases that require two steps to move with non-null probability into the desired set. For this purpose, notice that

$$\begin{aligned}&P^2\big((\alpha,d),(\beta_1,\beta_2)\times\{d+1\}\big) = \\ &\quad = \int_0^1 \sum_{d'=d,d+1} P\big((\alpha,d),(\mathrm{d}\alpha',d')\big)P\big((\alpha',d'),(\beta_1,\beta_2)\times\{d+1\}\big)\end{aligned}$$

Case 3: If $\alpha \in (0,1-\delta_1]$, $\beta_1 \in (\frac{c_\alpha}{1+c_\alpha} - \delta_2, \beta_2)$ and $\beta_2 \in [\frac{c_\alpha}{1+c_\alpha}, 1-\varepsilon]$, we exploit that

$$\begin{aligned}&P^2\big((\alpha,d),(\beta_1,\beta_2)\times\{d+1\}\big) \geq \\ &\quad \geq \int_0^1 P\big((\alpha,d),(\mathrm{d}\alpha',d+1)\big)P\big((\alpha',d+1),(\beta_1,\beta_2)\times\{d+1\}\big)\end{aligned} \tag{4.50}$$

As $P\big((\alpha', d+1), (\beta_1, \beta_2) \times \{d+1\}\big) \geq C^{(1)}_{\varepsilon,d+1}(\beta_2 - \beta_1)$ by (4.48),

$$\begin{aligned} P^2\big((\alpha, d), (\beta_1, \beta_2) \times \{d+1\}\big) &\geq C^{(1)}_{\varepsilon,d+1}(\beta_2 - \beta_1) P\big((\alpha, d), ([0,1], d+1)\big) \\ &\geq C^{(1)}_{\varepsilon,d+1}(\beta_2 - \beta_1) P\big((\alpha, d), ([\varepsilon, 1-\varepsilon], d+1)\big) \geq \\ &\geq C^{(1)}_{\varepsilon,d+1}(\beta_2 - \beta_1) \sigma\sqrt{2/\pi}(1-2\varepsilon) \min_{\alpha \in (0, 1-\delta_1]} m_{\bar{\alpha}, d+1}. \end{aligned} \tag{4.51}$$

Case 4: If $\alpha \in (1-\delta_1, 1]$, we exploit that

$$\begin{aligned} &P^2\big((\alpha, d), (\beta_1, \beta_2) \times \{d+1\}\big) \geq \\ &\geq \int_0^1 P\big((\alpha, d), (\mathrm{d}\alpha', d)\big) P\big((\alpha', d), (\beta_1, \beta_2) \times \{d+1\}\big). \end{aligned} \tag{4.52}$$

From (4.49), a sufficient condition to have $P\big((\alpha', d), (\beta_1, \beta_2) \times \{d+1\}\big) > \sigma\sqrt{2/\pi}\ \min_{\alpha \in [0,1-\delta_1]} m_{\bar{\alpha'}, d+1}(\beta$ $\beta_1)$ is $\alpha' \in (0, 1-\delta_1]$ for some small $\delta_1 > 0$ and $\beta_2 \leq \frac{c_{\alpha'}}{1+c_{\alpha'}}$. The latter corresponds to $c_{\alpha'} \geq \frac{\beta_2}{1-\beta_2}$, that is, $\alpha'^2 - \alpha' + \exp\left(\frac{1}{\sigma^2}\right)\left(\frac{1-\beta_2}{\beta_2}\right)^2 \geq 0$. This holds for any α' if $4\exp\left(\frac{1}{\sigma^2}\right)\left(\frac{1-\beta_2}{\beta_2}\right)^2 \geq 1$, otherwise for $\alpha' \in [0, \zeta(\beta_2)] \cup [1-\zeta(\beta_2), 1]$ where $\zeta(\beta_2) = \frac{1-\sqrt{1-4\exp\left(\frac{1}{\sigma^2}\right)\left(\frac{1-\beta_2}{\beta_2}\right)^2}}{2}$.

Given the initial hypothesis $\beta_1 \leq 1-\varepsilon$, $\zeta(\beta_2) \geq \frac{1-\sqrt{1-4\exp\left(\frac{1}{\sigma^2}\right)\left(\frac{\varepsilon}{1-\varepsilon}\right)^2}}{2} =: \eta$. Since $\eta \leq \frac{1}{2}$, $\eta < 1-\delta_1$. Now, reducing the domain of integration to $[0, \eta]$, we obtain

$$\begin{aligned} &P^2\big((\alpha, d), (\beta_1, \beta_2) \times \{d+1\}\big) \geq \\ &\geq \int_0^\eta P\big((\alpha, d), (\mathrm{d}\alpha', d)\big) P\big((\alpha', d), (\beta_1, \beta_2) \times \{d+1\}\big) \\ &\geq \int_0^\eta P\big((\alpha, d), (\mathrm{d}\alpha', d)\big) \sigma\sqrt{2/\pi}\ m_{\bar{\alpha'}, d+1}(\beta_2 - \beta_1) \\ &\geq \sigma\sqrt{2/\pi} \min_{\alpha' \in [0,\eta]} m_{\bar{\alpha'}, d+1}(\beta_2 - \beta_1) P\big((\alpha, d), ([0, \eta], d)\big) \\ &\geq \sigma\sqrt{2/\pi} \min_{\alpha' \in [0,\eta]} m_{\bar{\alpha'}, d+1}(\beta_2 - \beta_1) C^{(1)}_{\varepsilon,d} \eta. \end{aligned} \tag{4.53}$$

Finally, gathering the bounds obtained in the previous four cases, we obtain

$$P^2\big((\alpha, d), (\beta_1, \beta_2) \times \{d+1\}\big) \geq C^{(2)}_{\varepsilon,d}(\beta_2 - \beta_1). \tag{4.54}$$

where $C^{(2)}_{\varepsilon,d} = \delta_2(1-2\varepsilon)\eta\sigma\sqrt{2/\pi} \min_{\alpha \in [0,1-\delta_1]} m_{\bar{\alpha}, d+1} \min\{C^{(1)}_{\varepsilon,d}, C^{(1)}_{\varepsilon,d+1}\} > 0$.

We omit the proof of the third inequality as it is analogous to the second one: by the same argumentation, we obtain a suitable constant $C^{(3)}_{\varepsilon,d}$. Finally, for any small $\varepsilon > 0$ and $d \in \mathbb{Z}$, $C_{\varepsilon,d} = \min\{C^{(1)}_{\varepsilon,d}, C^{(2)}_{\varepsilon,d}, C^{(3)}_{\varepsilon,d}\}$.

The thesis is now proved for any open interval in $[\varepsilon, 1-\varepsilon]$. The generalization to all the open sets in $[\varepsilon, 1-\varepsilon]$ is straightforward since any open set on the real line is countable union of disjoint open intervals. Finally, we can extend the result to all the Borelians in $[\varepsilon, 1-\varepsilon]$. Remind that for any Lebesgue measurable set M (in particular, for any Borelian) in $\mathbb{R}$ there exists a sequence of open sets O_n such that $M \subset \cap_{n=1}^{\infty} O_n$ and $\mathcal{L}(M) = \mathcal{L}(\cap_{n=1}^{\infty} O_n)$, see [130]. As any finite intersection of open sets is open, we have

$$P^r\big((\alpha, d), \cap_{n=1}^{N} O_n \times \{d'\}\big) \geq C_\varepsilon \mathcal{L}(\cap_{n=1}^{N} O_n) \geq C_\varepsilon \mathcal{L}(\cap_{n=1}^{\infty} O_n) = C_\varepsilon \mathcal{L}(M)$$

for any $d' \in \{d-1, d, d+1\}$ and $r = 1, 2$ according to the value of d'. This inequality holds for any $N \in \mathbb{N}$, hence

$$\lim_{N\to\infty} P^r\big((\alpha, d), \cap_{n=1}^{N} O_n \times \{d'\}\big) = P^r((\alpha, d), \cap_{n=1}^{\infty} O_n \times \{d'\}) \geq C_\varepsilon \mathcal{L}(M).$$

■

By this lemma, it follows in particular that for any $M \in \mathcal{B}([\varepsilon, 1-\varepsilon])$,

$$\begin{cases} P^{2|d-d'|}\big((\alpha, d), M \times \{d'\}\big) \geq C_{\varepsilon,d}^{|d-d'|} \mathcal{L}(M) & \text{if } d \neq d'; \\ P\big((\alpha, d), M \times \{d\}\big) \geq C_{\varepsilon,d} \mathcal{L}(M). \end{cases}$$

Moreover,

Proposition 7 *For any $M \in \mathcal{B}([0,1])$ with $\mathcal{L}(M) > 0$,*

$$\begin{cases} P^{2|d-d'|}\big((\alpha, d), M \times \{d'\}\big) > \frac{1}{2} C_{\varepsilon,d}^{|d-d'|} \mathcal{L}(M) & \textit{if } d \neq d'; \\ P\big((\alpha, d), M \times \{d\}\big) > \frac{1}{2} C_{\varepsilon,d} \mathcal{L}(M). \end{cases}$$

In particular, $(A_k, D_k)_{k\in\mathbb{N}}$ is $(\mathcal{L} \times \kappa)$-irreducible, κ being the counting measure.

Proof By the previous lemma, this result holds when $M \in \mathcal{B}([\varepsilon, 1-\varepsilon])$ given any $\varepsilon > 0$. Now, if we consider any $M \in \mathcal{B}([0,1])$ with $\mathcal{L}(M) = \lambda > 0$, we have $\mathcal{L}(M \cap [\varepsilon, 1-\varepsilon]) = \mathcal{L}(M) - \mathcal{L}(M \cap [\varepsilon, 1-\varepsilon]^{\mathrm{c}}) \geq \lambda - 2\varepsilon$ and we can always choose $\varepsilon = \varepsilon(\lambda)$ such that $\lambda > 2\varepsilon$. For instance, let us choose $\varepsilon = \frac{\lambda}{4}$, so that $\lambda - 2\varepsilon = \frac{\lambda}{2}$. Therefore,

$$\begin{aligned} P^{2|d-d'|}\big((\alpha, d), M \times \{d'\}\big) &\geq P^{2|d-d'|}\big((\alpha, d), (M \cap [\varepsilon, 1-\varepsilon]) \times \{d'\}\big) \\ &\geq C_{\varepsilon,d} \mathcal{L}(M \cap [\varepsilon, 1-\varepsilon]) > \frac{\lambda}{2} C_{\varepsilon,d}^{|d-d'|} \end{aligned}$$

when $d \neq d'$, and similarly when $d = d'$. ■

4.7.7 TSA: an upper bound for the transition probability kernel

Lemma 6 *There exists a real positive constant G such that*

$$P\big((\alpha, d), M \times \mathbb{Z}\big) \leq G\mathcal{L}(M)$$

for any $(\alpha, d) \in [0,1] \times \mathbb{Z}$ and $M \in \mathcal{B}([0,1])$.

Proof First, we prove the lemma when M is an open interval. Consider the equations (4.38) - (4.40): given (α, d), $P\big((\alpha, d), (\beta_1, \beta_2) \times \mathbb{Z}\big)$ is equal to a sum of integrals of type $\int_{\beta_1}^{\beta_2} e^{-h^2_{x,y}(z)}(-h'_{x,y}(z))\mathrm{d}z$ with $x \in \{\alpha, \frac{1}{1-\alpha}\}$ and $y \in \{d-1, d, d+1\}$ according to the instance. As we have shown in the Proof of Lemma 2, $h'_{x,y}(z) = \frac{\sigma}{z(z-1)\sqrt{2}}$, hence $g(z) = -e^{-h^2_{x,y}(z)}h'_{x,y}(z) > 0$ for every $z \in (0,1)$. Furthermore, as $h''_{x,y}(z) = (h'_{x,y}(z))^2\frac{\sqrt{2}}{\sigma}(1-2z)$, $g'(z) = 2e^{-h^2_{x,y}(z)}(h'_{x,y}(z))^2\left(h_{x,y}(z) - \frac{\sqrt{2}}{\sigma}\left(\frac{1}{2} - z\right)\right) = 0$ at $z_0 \in (0,1)$, z_0 being the unique solution of the equation $h_{x,y}(z) = \frac{\sqrt{2}}{\sigma}(\frac{1}{2} - z)$; hence $g(z)$ is increasing in $(0, z_0)$, decreasing in $(z_0, 1)$ and admits a maximum in $z_0 \in (0,1)$. In conclusion, $\int_{\beta_1}^{\beta_2} g(z)\mathrm{d}z \leq G(\beta_2 - \beta_1)$, $G = g(z_0)$.
The extension to all the open sets is trivial as any open set is countable union of disjoint intervals. Finally, as for any $M \in \mathcal{B}([0,1])$ there exists a sequence of open sets O_n such that $M \subset \cap_{n=1}^{\infty} O_n$ and $\mathcal{L}(M) = \mathcal{L}(\cap_{n=1}^{\infty} O_n)$ (see [130]), for any $n \in \mathbb{N}$ we can write

$$P\big((\alpha, d), \cap_{n=1}^{\infty} O_n \times \mathbb{Z}\big) \leq P\big((\alpha, d), \cap_{n=1}^{N} O_n \times \mathbb{Z}\big) \leq G\mathcal{L}(\cap_{n=1}^{N} O_n)$$

as any finite intersection of open sets is open. The result follows from the arbitrariness of N. ■

4.7.8 TSA: Proof of Theorem 6

The process $(A_k, D_k)_{k\in\mathbb{N}}$ with $(U_k)_{k\in\mathbb{N}}$ is an instance of MCRE. The corresponding EMP in $\Omega = [0,1] \times \mathbb{Z} \times \{0,1\}^{\mathbb{N}}$ is defined by the following transition probability kernel:

$$P\big((\alpha, d, \mathbf{u}), A \times \{d'\} \times B\big) = P_{u_0}\big((\alpha, d), A \times \{d'\}\big)\mathbb{1}_B(T\mathbf{u}) \tag{4.55}$$

where $\mathbf{u} = (u_0, u_1, \dots) \in \{0,1\}^{\mathbb{N}}$, $A \in \mathcal{B}([0,1])$, $d' \in \mathbb{Z}$, $B \in \mathcal{P}(\{0,1\}^{\mathbb{N}})$.

$P_{u_0}\big((\alpha, d), A \times \{d'\}\big)$ can be assessed by equations (4.34)-(4.37). Moreover, we denote by $P_{u_0,\dots u_{k-1}}\big((\alpha, d), A \times \{d'\}\big)$ the probability of moving from $(\alpha, d) \in [0,1] \times \mathbb{Z}$ to the set $A \times \{d'\}$, $A \in \mathcal{B}([0,1])$, in k-steps, given the input sequence $(u_0, \dots, u_{k-1}) \in \{0,1\}^k$. By Proposition 6, $\widetilde{\psi} = \widetilde{\phi} \times \pi$ ($\widetilde{\phi}$ being defined in Lemma 3), is an invariant p.m. for the EMP. Moreover,

Lemma 7 *$\widetilde{\psi}$ is ergodic.*

Proof Let $F \subset \Omega$ be an invariant set: by Definition 6, to prove the ergodicity of $\widetilde{\psi}$ is sufficient to show that $\widetilde{\psi}(F) > 0$ implies $\widetilde{\psi}(F) = 1$.

Then, let us suppose $\widetilde{\psi}(F) > 0$. We name

$$\begin{aligned}
&\mathcal{U}_F = \big\{\mathbf{u} \in \{0,1\}^{\mathbb{N}} : (\alpha, d, \mathbf{u}) \in F \text{ for some } (\alpha, d) \in [0,1] \times \mathbb{Z}\big\};\\
&\mathcal{U}_0 = \big\{\mathbf{u} \in \{0,1\}^{\mathbb{N}} : \mathbf{u} \text{ contains infinitely many 0's and 1's}\big\};\\
&\mathcal{U}_0^n = \big\{\mathbf{u} \in \mathcal{U}_0 : \mathbf{u} \text{ contains at least a 0 and a 1 in its first } n \text{ bits } \big\},\ n \geq 2.
\end{aligned}$$

Given the transition probability kernel (4.55), if $\mathbf{u} \in \mathcal{U}_F$ then also $T\mathbf{u} \in \mathcal{U}_F$ and since π is an ergodic measure with respect to the shift operator T (see [159, Section 1.5]) and $\pi(\mathcal{U}_F) > 0$ (otherwise $\widetilde{\psi}(F) = 0$), we have that $\pi(\mathcal{U}_F) = 1$ by the Birkhoff's Individual Ergodic Theorem ([159, Theorem 1.14]).

By analogous reasoning, $\pi(\mathcal{U}_0) = 1$. Furthermore, $\mathcal{U}_0^n \subset \mathcal{U}_0^{n+1}$, then $\mathcal{U}_0^n \uparrow \mathcal{U}_0$. This implies the existence of an $n_0 \geq 2$ such that $\pi(\mathcal{U}_0^{n_0}) > 0$.

At this point, let us consider the equations (4.34)-(4.37): by applying the procedure used to prove Lemma 5 and Proposition 7, it is easy to verify that for any $(\alpha, d) \in (0,1) \times \mathbb{Z}$,

$$\begin{aligned}
&P_0\big((\alpha,d), M \times \{d\}\big) > 0 \ \text{ for any } M \in \mathcal{B}(\ (1/3,1]\ ),\ \mathcal{L}(M) > 0;\\
&P_1\big((\alpha,d), M \times \{d\}\big) > 0 \ \text{ for any } M \in \mathcal{B}(\ [0,2/3)\ ),\ \mathcal{L}(M) > 0;\\
&P_0\big((\alpha,d), M \times \{d+1\}\big) > 0 \ \text{ for any } M \in \mathcal{B}(\ [0,2/3)\ ),\ \mathcal{L}(M) > 0;\\
&P_1\big((\alpha,d), M \times \{d-1\}\big) > 0 \ \text{ for any } M \in \mathcal{B}(\ (1/3,1]\ ),\ \mathcal{L}(M) > 0.
\end{aligned} \tag{4.56}$$

where $\frac{1}{3}$ and $\frac{2}{3}$ are sufficient, not necessary bounds derived from Remark 5. These inequalities yield to

$$\begin{aligned}
&P_{01}\big((\alpha,d), M \times \{d\}\big) > 0 \ \text{ for any } M \in \mathcal{B}(\ [0,1]\ ),\ \mathcal{L}(M) > 0;\\
&P_{10}\big((\alpha,d), M \times \{d\}\big) > 0 \ \text{ for any } M \in \mathcal{B}(\ [0,1]\ ),\ \mathcal{L}(M) > 0
\end{aligned} \tag{4.57}$$

which may interpreted as follows: whenever the input sequence contains a couple of bits 01 or 10, the component α can reach any non-negligible subset in $(0,1)$. Notice also that we are not considering the negligible cases $\alpha = 0$ and $\alpha = 1$, which may prevent the one-step transition (see (4.34)-(4.37)). Maintaining this hypothesis, let us consider $(\alpha, d, \mathbf{u}) \in F$ such that $\mathbf{u} \in \mathcal{U}_0^{n_0} \cap \mathcal{U}_F$ (remind that $\pi(\mathcal{U}_0^{n_0}) > 0$ and $\pi(\mathcal{U}_F) = 1$, then $\pi(\mathcal{U}_0^{n_0} \cap \mathcal{U}_F) > 0$).

Now, let us consider the evolution $(\alpha, d, \mathbf{u}) \in F$, reminding that it cannot get out of F. As $\mathbf{u} \in \mathcal{U}_0^{n_0}$, $\mathbf{u}$ contains at least one couple 01 or 10 in its first bits and given (4.57), after n_0 steps α could have been reached all the interval $(0,1)$. Moreover, for any bit sequence, d can maintain its position. Hence, we can conclude that

$$(0,1) \times \{d\} \times \{T^{n_0}\mathbf{u}\} \subset F. \tag{4.58}$$

Now, the fact that $\mathcal{U}_0^{n_0}$ is not negligible implies that we can always choose $\mathbf{u} \in \mathcal{U}_0^{n_0}$ such that $\mathcal{V}_{\mathbf{u}} = \{T^n\mathbf{u},\ n \in \mathbb{N}\}$ has measure $\pi(\mathcal{V}_{\mathbf{u}}) = 1$, as a consequence of [159, Theorem 1.14]. Hence,

$$[0,1] \times \{d\} \times \mathcal{V}_{\mathbf{u}} \subset F \tag{4.59}$$

Finally, let us consider the evolution of the component $d \in \mathbb{Z}$: from equations (4.56), there is a positive probability that, in n steps, d achieves any integer $d' \in D_n$ where $D_n = \{d - m_1, d - m_1 + 2, \dots, d + n - m_1\}$, m_1 being the number of 1's in the corresponding n-bit input sequence. Hence,

$$[0,1] \times D_n \times T^n \mathcal{V}_{\mathbf{u}} \subset F \tag{4.60}$$

where $T^n \mathcal{V}_{\mathbf{u}} = \mathcal{V}_{\mathbf{u}}$ except for at most a π negligible set. Given that for any n, $D_n \subset D_{n+1}$, in particular, D_{n+1} has one more element than D_n, then $D_n \uparrow \mathbb{Z}$. This finally proves that

$$[0,1] \times \mathbb{Z} \times \mathcal{V}_{\mathbf{u}} \subset F \tag{4.61}$$

except for at most a π negligible set. This implies

$$\widetilde{\psi}(F) = \widetilde{\phi}([0,1] \times \mathbb{Z})\pi(\mathcal{V}_{\mathbf{u}}) = 1. \tag{4.62}$$

■

Given $q(\alpha, d, U_k) = P(\widehat{U}_k \neq U_k | U_k, A_k = \alpha, D_k = d)$,

$$\begin{aligned} \mathrm{CBER}(\mathbb{D}^{(2)}|\mathbf{U}) &= \frac{1}{K}\sum_{k=0}^{K-1} P(\widehat{U}_k \neq U_k|\mathbf{U}) = \\ &= \int_0^1 \sum_{d\in\mathbb{Z}} \frac{1}{K}\sum_{k=0}^{K-1} q(\alpha, d, U_k) P_{(U_0,\dots U_{k-1})}\big((1,0),(\mathrm{d}\alpha, d)\big). \end{aligned} \tag{4.63}$$

Now, let $g(\alpha, d, \mathbf{U}) = q(\alpha, d, U_0)$: it is easy to verify that

$$(P^k g)(\alpha, d, \mathbf{U}) = \int_0^1 \sum_{d'\in\mathbb{Z}} q(\alpha', d', U_k) P_{(U_0,\dots U_{k-1})}\big((\alpha, d),(\mathrm{d}\alpha', d')\big)$$

then

$$\mathrm{CBER}(\mathbb{D}^{(2)}|\mathbf{U}) = \frac{1}{K}\sum_{k=0}^{K-1} (P^k g)(1, 0, \mathbf{U}). \tag{4.64}$$

By the Ergodic Theorem 7,

$$\lim_{K\to\infty} \frac{1}{K}\sum_{k=0}^{K-1} (P^k g)(\omega) = \int_\Omega g \, \mathrm{d}\widetilde{\psi} \quad \text{for } \widetilde{\psi}\text{-a.e. } \omega \in \Omega$$

Let $N \subset \Omega$ be the negligible set for which there is no convergence and let $N_{0,\mathbf{U}} = \{\alpha \in [0,1] : (\alpha, 0, \mathbf{U}) \in N\}$. By the same argumentation used in Corollary 5, $P_{U_0}((1,0), N_{u,\mathbf{U}} \times \{0\}) = 0$ and

$$\begin{aligned} &\mathrm{CBER}(\mathbb{D}^{(2)}|\mathbf{U})) = \\ &= \frac{1}{K} g(1, 0, \mathbf{U}) + \frac{1}{K}\sum_{k=1}^{K-1} \int_{\alpha_1\in[0,1]} P_{U_0}((1,0),(\mathrm{d}\alpha_1, 0))(P^{k-1}g)(\alpha_1, 0, T\mathbf{U}) \\ &\overset{K\to\infty}{\longrightarrow} \int_{\alpha_1\in[0,1]\setminus N_{0,\mathbf{U}}} P_{U_0}((1,0),(\mathrm{d}\alpha_1, 0)) \int_\Omega g \, \mathrm{d}\widetilde{\psi} = \int_\Omega g \, \mathrm{d}\widetilde{\psi} \quad \pi\text{-a.e.} \mathbf{U} \in \{0,1\}^{\mathbb{N}}. \end{aligned}$$

which proves the thesis, as

$$\int_\Omega g \, \mathrm{d}\widetilde{\psi} = \int_{[0,1]} \sum_{d\in\mathbb{Z}} \sum_{u\in\{0,1\}} q(\alpha, d, u)\widetilde{\phi}(\mathrm{d}\alpha, d)\pi_0(u) = \int_{[0,1]\times\mathbb{Z}} \overline{q} \, \mathrm{d}\widetilde{\phi}.$$

4.7.9 CBCJR vs LMSE

Let us consider the binary input case. While the CBCJR computes the estimate

$$\widehat{u}_{k-1}^{CBCJR}(\mathbf{y}_1^k) = \arg \min_{v\in\{0,1\}} \mathbb{E}[|U_{k-1} - v|^2|\mathbf{Y}_1^k = \mathbf{y}_1^k]$$

the (causal) LMSE (see Section 2.5.7) computes

$$\widehat{u}_{k-1}^{LMSE}(\mathbf{y}_1^k) = \arg \min_{v\in\mathbb{R}} \mathbb{E}[||U_{k-1} - v||^2|\mathbf{Y}_1^k = \mathbf{y}_1^k].$$

Thus, the only one difference lies in the space where the minimum is calculated. In particular,

Proposition 8

$$\widehat{u}_{k-1}^{CBCJR} = \begin{cases} 0 & \textit{if } \widehat{u}_{k-1}^{LMSE} \leq \frac{1}{2} \\ 1 & \textit{otherwise.} \end{cases} \tag{4.65}$$

Proof We know that (see 4.12)

$$\widehat{u}_{k-1}^{CBCJR} = \begin{cases} 0 & \text{if } \sum_{i=0}^{k-1} \widetilde{\sigma}_k(i, i+1) \leq \sum_{i=0}^{k-1} \widetilde{\sigma}_k(i, i) \\ 1 & \text{otherwise.} \end{cases}$$

where $\widetilde{\sigma}_k(i, j) = f_{(X_k, X_{k-1}, \mathbf{Y}_1^k)}(j, i, \mathbf{y}_1^k)$. Hence,

$$\begin{aligned} &\sum_{i=0}^{k-1} \sigma_k(i, i+1) - \sigma_k(i, i-1) = \\ &= \sum_{i=0}^{k-1} f_{(U_{k-1}, X_{k-1}, \mathbf{Y}_1^k)}(1, i, \mathbf{y}_1^k) - f_{(U_{k-1}, X_{k-1}, \mathbf{Y}_1^k)}(0, i, \mathbf{y}_1^k) \\ &= f_{(U_{k-1}, \mathbf{Y}_1^k)}(1, \mathbf{y}_1^k) - f_{(U_{k-1}, \mathbf{Y}_1^k)}(0, \mathbf{y}_1^k) \end{aligned}$$

Noting that

$$\begin{aligned} \widehat{u}_{k-1}^{LMSE} &= \mathbb{E}[U_{k-1}|\mathbf{Y}_1^k = \mathbf{y}_1^k] = P(U_{k-1} = 1|\mathbf{Y}_1^k = \mathbf{y}_1^k) - P(U_{k-1} = 0|\mathbf{Y}_1^k = \mathbf{y}_1^k) \\ &= \frac{f_{(U_{k-1}, \mathbf{Y}_1^k)}(1, \mathbf{y}_1^k) - f_{(U_{k-1}, \mathbf{Y}_1^k)}(0, \mathbf{y}_1^k)}{f_{\mathbf{Y}_1^k}(\mathbf{y}_1^k)} \end{aligned}$$

we obtain the thesis. ■

Chapter 5

One-dimensional Linear Systems

In this chapter, we study the deconvolution of a generic quantized-input, one-dimensional, linear input/output system.

The structure is analogous to the one of Chapter 4: we introduce the problem and the assumptions, state the algorithm we intend to use, and develop a theoretical analysis, which in this case is mainly based on results about Iterated Random Functions.

Furthermore, the last part (Section 5.6) is devoted to the comparison of our techniques with Kalman Filtering method.

5.1 Problem Statement

We consider the input/output linear system

$$\begin{cases} x'(t) = ax(t) + bu(t) \qquad t \in [0,T] \\ y(t) = cx(t) \\ x(0) = 0 \end{cases} \tag{5.1}$$

where $u(t)$, $x(t)$ and $y(t)$ are real functions and respectively represent the input, the state function and the output; a, b and c are non-null real constants, $[0,T]$ is a possibly infinite time horizon. $u(t)$ is supposed to be unknown, while $y(t)$ is accessible, but possibly affected by an observational noise.

Our aim is to reconstruct $u(t)$ from $y(t)$, that is, to reverse the input/output convolution integral:

$$y(t) = cb\int_0^t e^{a(t-s)}u(s)\mathrm{d}s. \tag{5.2}$$

This problem is the natural extension of the differentiation problem (4.2) studied in Chapter 4. Our aim is to extend the Information-Decoding setting and apply the algorithms introduced in Chapter 4 in this more general framework, under analogous assumptions on the quantization of the input and the sampling of the output. We will observe that the dynamics of the present system is different, mainly since in the differentiation case the sampled version of x was a vector of natural numbers, while now the state space is not countable. This is drawback from the complexity implementation viewpoint: as we will in the next, not all the algorithms previously introduced can be applied and in particular only the One State Algorithm turns out to be efficient. Its performance will be theoretically analyzed in the framework of Markov Process and using the Iterated Random Functions theory. Finally, a comparison with the Kalman Filter, which is the most used algorithm for linear systems, is proposed.

In the next, we will stick to the problem (5.1)) under Assumptions 1-5 of Section 3.1 and Assumptions 6-7 of Section 4.1.1, that is: the input is stepwise constant with constant sampling time step τ, it is quantized over two levels 0 and 1 and is generated by a Bernoulli source; the output is sampled and synchronized with the input; a Gaussian noise affects the measurements of the output.

Moreover, we state

Assumption 8 *The system is stable, that is $a < 0$.*

Notice that the case $a = 0$ actually corresponds to the differentiation problem, see Chapter 4).

Assumption 6 induces a discrete description for $x(t)$:

$$\begin{aligned} x_k := x(k\tau) &= be^{ak\tau}\int_0^{k\tau} e^{-as}\sum_{h=0}^{K-1} u_h \mathbb{1}_{[h\tau,(h+1)\tau[}(s)ds \\ &= be^{ak\tau}\sum_{h=0}^{k-1} u_h \int_{h\tau}^{(h+1)\tau} e^{-as}ds \\ &= \frac{b}{a}(e^{a\tau}-1)e^{a(k-1)\tau}\sum_{h=0}^{k-1} u_h e^{-ah\tau}. \end{aligned} \tag{5.3}$$

Defining

$$\mathrm{q} := e^{a\tau}, \qquad \mathrm{w} := \frac{b}{a}(e^{a\tau}-1) = \frac{b}{a}(\mathrm{q}-1) \tag{5.4}$$

we can write the following recursive formula:

$$x_k = \mathrm{q}x_{k-1} + \mathrm{w}u_{k-1}. \tag{5.5}$$

By trivial computations, we obtain also

$$x_k = \mathrm{w}\sum_{h=0}^{k-1} u_{k-h-1}\mathrm{q}^h$$

which shows that each x_k assumes values in the set

$$\mathcal{X} = \mathrm{w}\left\{\sum_{h=0}^{\infty} \mu_h \mathrm{q}^h,\ \mu_h \in \{0,1\}\right\}$$

that includes all the x_k's, $k \in \mathbb{N}$. The structure of $\mathcal{X}$ will play a fundamental role in the deconvolution's performance. For computational simplicity, from now onwards let

$$\tau = 1 \text{ and } b > 0.$$

Notice that $\mathcal{X} \subseteq \mathrm{w}[0, \frac{1}{1-\mathrm{q}}]$ and $\mathrm{q} \in (0,1)$ by Assumption 8. Now, two possible cases have to be distinguished.

5.1.1 Case $\mathrm{q} \in \left[\frac{1}{2}, 1\right)$.

If $\mathrm{q} \in \left[\frac{1}{2}, 1\right)$, then $\mathcal{X} \equiv \mathrm{w}\left[0, \frac{1}{1-\mathrm{q}}\right]$. This can be proved as follows. Given any $x \in \mathrm{w}\left[0, \frac{1}{1-\mathrm{q}}\right]$, we construct a series $\mathrm{w}\sum_{h=0}^{\infty} \mu_h \mathrm{q}^h = x$, $\mu_h \in \{0,1\}$ defining on by one the coefficients μ_h. The series being positive termed, the procedure is:
For $h = 0$, fix

$$\begin{cases} \mu_0 = 1 & \text{if } x \geq \mathrm{w} \\ \mu_0 = 0 & \text{otherwise} \end{cases}$$

For $h = 1, 2, 3, \ldots$, fix

$$\begin{cases} \mu_h = 1 & \text{if } x \geq \mathrm{w}\sum_{i=0}^{h-1} \mu_i + \mathrm{w}\mathrm{q}^h \\ \mu_h = 0 & \text{otherwise} \end{cases}$$

For any $h \in \mathbb{N}$, we then obtain a polynomial $\mathrm{w}\sum_{i=0}^{h} \mu_i \mathrm{q}^i \leq x$. In case of equality, the property is proved; otherwise, we have to show that

$$x \leq \mathrm{w}\sum_{i=0}^{h} \mu_i \mathrm{q}^i + \mathrm{w}\sum_{i=h+1}^{\infty} \mathrm{q}^i \tag{5.6}$$

This is obvious if $\mu_i = 1$ for any $i = 0, \ldots, h$. Otherwise, if there exists at least one null coefficient between 0 and h, let us consider the null coefficient with greater index, that is, pick j such that $\mu_j = 0$ and $\mu_i = 1$ for any $i = j+1, \ldots, h$. Then, $x < \mathrm{w}\sum_{i=0}^{j-1} \mu_i \mathrm{q}^i + \mathrm{w}\mathrm{q}^j$, otherwise it should have been $\mu_j = 1$. Now, in order to prove the bound (5.6) it is sufficient to show that

$$\sum_{i=0}^{j-1} \mu_i \mathrm{q}^i + \mathrm{q}^j \leq \sum_{i=0}^{h} \mu_i \mathrm{q}^i + \sum_{i=h+1}^{\infty} \mathrm{q}^i \tag{5.7}$$

This is obtained by easy computations:

$$\begin{aligned} \mathrm{q}^j &\leq \sum_{i=j}^{h} \mu_i \mathrm{q}^i + \sum_{i=h+1}^{\infty} \mathrm{q}^i \\ &= \sum_{i=j+1}^{\infty} \mathrm{q}^i = \frac{\mathrm{q}^{j+1}}{1-\mathrm{q}}. \end{aligned} \tag{5.8}$$

Finally,

$$\mathrm{q}^j \leq \frac{\mathrm{q}^{j+1}}{1-\mathrm{q}} \quad \Leftrightarrow \quad \mathrm{q} \geq \frac{1}{2}$$

and this proves (5.6). Now, we know that

$$\mathrm{w}\sum_{i=0}^{h} \mu_i \mathrm{q}^i \leq x \leq \mathrm{w}\sum_{i=0}^{h} \mu_i \mathrm{q}^i + \mathrm{w}\frac{\mathrm{q}^{h+1}}{1-\mathrm{q}} \tag{5.9}$$

and in the limit case $h \to \infty$, this becomes $x = \mathrm{w}\sum_{i=0}^{\infty} \mu_i \mathrm{q}^i$.
In conclusion, $\mathrm{w}[0, \frac{1}{1-\mathrm{q}}] \subseteq \mathcal{X}$ and given that the opposite inclusion holds by definition, we have the equivalence

$$\mathcal{X} \equiv \left[0, \frac{1}{1-\mathrm{q}}\right]. \tag{5.10}$$

5.1.2 Case $\mathrm{q} \in (0, \frac{1}{2})$.

If $\mathrm{q} < \frac{1}{2}$, $\mathcal{X}$ is a Cantor set. It can be constructed from the interval $\mathrm{w}\left[0, \frac{1}{1-\mathrm{q}}\right]$ by deleting the elements that cannot be represented by the series, that is, the subintervals $\mathrm{w}\left(\frac{\mathrm{q}}{1-\mathrm{q}}, 1\right)$, $\mathrm{w}\left(\frac{\mathrm{q}^2}{1-\mathrm{q}}, \mathrm{q}\right) \cup \mathrm{w}\left(1 + \frac{\mathrm{q}^2}{1-\mathrm{q}}, 1+\mathrm{q}\right)$, etc. More precisely,

$$\mathcal{X} = \mathrm{w}\left[0, \frac{1}{1-\mathrm{q}}\right] - \mathrm{w}\sum_{m=0}^{\infty}\bigcup_{n=1}^{2^m}\left(\mathrm{p}_{m,n} + \frac{\mathrm{q}^{m+1}}{1-\mathrm{q}},\ \mathrm{p}_{m,n} + \mathrm{q}^m\right) \tag{5.11}$$

where $\mathrm{p}_{m,1}, \ldots, \mathrm{p}_{m,2^m}$ are the binary polynomials in q of degree at most $m-1$ ($\mathrm{p}_{-1,1} = 0$ by convention).

Notice that $\mathrm{w}\sum_{i=0}^{\infty} \mu_i \mathrm{q}^i$ is a bijective map from $\{0,1\}^{\mathbb{N}}$ to $\mathcal{X}$ if $\mathrm{q} < \frac{1}{2}$. The surjectivity is obvious, while as far as the injectivity is concerned, suppose that $\sum \mu_n \mathrm{q}^n = \sum \nu_n \mathrm{q}^n$, $\mu_n, \nu_n \in \{0,1\}$ but with some different coefficients; for instance, let $\mu_n = \nu_n$ for $n = 0, \ldots, m-1$, $\mu_m = 0$ and $\nu_m = 1$ for some m. Since $\mathrm{q} < \frac{1}{2}$, $\frac{\mathrm{q}^{m+1}}{1-\mathrm{q}} < \mathrm{q}^m$, hence $\sum \mu_n \mathrm{q}^n < \sum \nu_n \mathrm{q}^n$: this proves that there cannot be two series with the same sum, but different coefficients.

The geometrical characterization of $\mathcal{X}$ strongly affects the performance of our deconvolution algorithm, that will be shortly introduced. Before that, we need to change our perspective on the problem, describing it in Information theoretic, probabilistic terms.

5.1.3 Probabilistic Setting and Performance Evaluation

Given the assumptions stated in the last section, let us now rewrite our dynamical system in probabilistic terms:

$$\begin{cases} U_{k-1} \sim \mathrm{Ber}\,(1/2) \\ N_k \sim \mathcal{N}(0, \sigma^2) \\ X_k = \mathrm{q}X_{k-1} + \mathrm{w}U_{k-1} \quad (X_0 = 0) \\ Y_k = cX_k + N_k \end{cases} \tag{5.12}$$

Moreover, we will denote by $\widehat{U}_k$ the estimate of the bit U_k obtained by a deconvolution-decoding algorithm $\mathbb{D}$, that is, $\widehat{U}_k = \mathbb{D}(\mathbf{U})_k$ (see Section 4.1).

As in Chapter 4 (see in particular Section 4.1), we measure the performance of a deconvolution-decoding algorithm $\mathbb{D}$ in terms of Mean Square Error:

$$\mathrm{MSE}(\mathbb{D}) = \sum_{k=0}^{K-1} \mathbb{E}(U_k - \hat{U}_k)^2$$

which, in case of binary input, corresponds to

$$\mathrm{MSE}(\mathbb{D}) = \sum_{k=0}^{K-1} \mathrm{P}(U_k \neq \hat{U}_k) = K\mathrm{BER}(\mathbb{D}).$$

5.2 One State Algorithm

Let us implement the OSA in order to estimate the input of the system (5.12): in detail, the pattern is as follows.

OSA - Decoder $\mathbb{D}^{(1)}$

Initialization: $\widehat{x}_0 = 0$.

For $k = 1, \dots, K$, given the received symbol $y_k \in \mathbb{R}$, estimate the current bit and the current state:

$$\widehat{u}_{k-1} = \mathbb{D}^{(1)}(\mathbf{y})_{k-1} \begin{cases} 0 \text{ if } |y_k - c\mathrm{q}\widehat{x}_{k-1}| \leq |y_k - (c\mathrm{q}\widehat{x}_{k-1} + c\mathrm{w})| \\ 1 \text{ otherwise.} \end{cases}$$
$$\widehat{x}_k = \mathrm{q}\widehat{x}_{k-1} + \mathrm{w}\widehat{u}_{k-1}. \tag{5.13}$$

We recall the causality of the OSA: $\widehat{u}_{k-1} = \mathbb{D}^{(1)}(\mathbf{y})_{k-1} = \mathbb{D}^{(1)}(y_1, y_2, \dots y_k)_{k-1}$.

Given the estimation of the current state x_k, the decoder estimates the possible transmitted signal according to the dynamics of the system. As the input is binary, at each step we have just two possible signals and we decide between them evaluating the distance between them and the acquired output sample y_k.

We recall that the OSA is suboptimal, but presents two main good properties: (a) it is low-complexity, both for number of computations and storage locations; (b) it is causal, that is, it uses only the past and the present information to decode the current bit. Therefore, (a) it can be applied to our case in which the number of states is (not countably) infinite and (b) it can be used on-line, making unnecessary the complete transmission before starting deconvolution, this feature being fundamental to study long time transmissions.

We observe that BCJR, CBCJR and TSA (introduced in Chapter 4) are not efficient in this framework. In fact, BCJR and CBCJR cannot be applied for complexity issues: even if we consider the number K of transmitted bits to be finite, the state x_K may assume 2^K values and the complexity of these algorithms grows exponentially. The TSA, instead, has no complexity problems but has performance too similar to the OSA (in spite of a slightly higher complexity) because of the structure of the state space $\mathcal{X}$: the two best states turn out to be very close to each other, which does not improve the information provided by the OSA. Its implementation is then not motivated.

5.3 Theoretic Analysis of the OSA through IRF

The theoretical analysis of the OSA performance is based on the definition of the random difference between the real and the estimated state, at each step $k = 1, \dots, K$:

$$\begin{cases} D_k = \widehat{X}_k - X_k = \mathrm{q}D_{k-1} + \mathrm{w}(\widehat{U}_{k-1} + U_{k-1}) \\ D_0 = 0 \end{cases} \tag{5.14}$$

Given $D_{k-1} = z$, $D_k \in \{\mathrm{q}z, \mathrm{q}z + \mathrm{w}, \mathrm{q}z - \mathrm{w}\}$. Considering $K \to \infty$, $(D_k)_{k\in\mathbb{N}}$ with $D_0 = 0$ is a Markov Process on the state space $\mathrm{w}\left\{\sum_{h=0}^{\infty} \mu_h \mathrm{q}^h,\ \mu_h \in \{-1, 0, 1\}\right\}$; the structure is analogous for $\mathcal{X}$: this state space is the interval $\mathcal{D} = \left[\frac{b}{a}, -\frac{b}{a}\right]$ if $\mathrm{q} \geq \frac{1}{3}$, otherwise it is a Cantor set included in $\mathcal{D}$. Neglecting the given initial state, let us generically say that $(D_k)_{k\in\mathbb{N}}$ is Markov Process on $\mathcal{D}$, starting from $D_0 = z$, $z \in \mathcal{D}$, and evolving according to a transition probability kernel P, which in turn is ruled by the independent stochastic process $(U_{k-1}, N_k)_{k=1,2,\dots}$. A suitable way to describe such model is given by Iterated Random Functions' theory; before doing that, let us state the main result of this work based on the evolution of D_k.

5.3.1 Performance Theorem

Through IRF theory, it is possible to prove that

Theorem 8 *Under the stability condition* $a < \log\left(\frac{1}{3+\sqrt{\frac{2}{e\pi}}}\right)$,

$$\lim_{K\to\infty} \mathrm{BER}(\mathbb{D}^{(1)}) = \int_{\mathcal{D}} g \mathrm{d}\mu \tag{5.15}$$

where $g(z) = \mathrm{P}(\widehat{U}_k \neq U_k | D_k = z)$ is the probability of bad decoding at step $k = 0, 1, \dots$ given the knowledge of D_k, and μ is the unique invariant probability measure for the kernel P of $(D_k)_{k \in \mathbb{N}}$.

Notice that $g(z)$ is time-invariant (i.e., does not depend on k) and can be analytically computed. In fact, given $D_{k-1} = z$, $D_k = \mathrm{q}z$ if and only if $\widehat{U}_k = U_k$, $D_k = \mathrm{q}z + \mathrm{w}$ if and only if $\widehat{U}_k = 1$ and $U_k = 0$, $D_k = \mathrm{q}z - \mathrm{w}$ if and only if $\widehat{U}_k = 0$ and $U_k = 1$ and

$$\begin{aligned} P(z, \mathrm{q}z + \mathrm{w}) &= \mathrm{P}(\widehat{U}_k = 1, U_k = 0 | D_k = z) \\ &= \frac{1}{4} \mathrm{erfc} \left(\frac{c\mathrm{q}z + c\mathrm{w}/2}{\sigma\sqrt{2}} \right) \\ P(z, \mathrm{q}z - \mathrm{w}) &= \mathrm{P}(\widehat{U}_k = 0, U_k = 1 | D_k = z) \\ &= \frac{1}{4} \mathrm{erfc} \left(\frac{-c\mathrm{q}z + c\mathrm{w}/2}{\sigma\sqrt{2}} \right) \\ P(z, \mathrm{q}z) &= 1 - P(z, \mathrm{q}z + \mathrm{w}) - P(z, \mathrm{q}z - \mathrm{w}) \end{aligned} \tag{5.16}$$

$$g(z) = P(z, \mathrm{q}z + \mathrm{w}) + P(z, \mathrm{q}z - \mathrm{w}). \tag{5.17}$$

Furthermore, the limit probability measure μ can be numerically evaluated.

5.3.2 Iterated Random Functions

Let $(\mathcal{D}, d)$ be a complete metric space and $\mathcal{S}$ be a measurable space. Consider a measurable function $w : \mathcal{D} \times \mathcal{S} \to \mathcal{D}$ and for each fixed $s \in \mathcal{S}$, $w_s(x) := w(x, s)$, $x \in \mathcal{D}$. Let $(I_k)_{k \in \mathbb{N}}$ be a stochastic sequence in $\mathbf{S}$ such that $I_0, I_1, \dots$ are independent, identically distributed. Then, the set $\{w_{I_k}(x),\ k \in \mathbb{N}\}$ is a family of random functions. The systems obtained by iterating such random functions, called IRF or Iterated Functions Systems (IFS), are studied for diverse purposes: for example, IRF with contractive properties are used to construct fractal sets, see [69, 38]. More interesting for our study is the exploitation of IRF to study Markov Processes. In particular, given an IRF and a starting state $x \in \mathcal{D}$, we can define the induced Markov Process $(Z_k(x))_{k \in \mathbb{N}}$ as

$$Z_k(x) := w_{I_{k-1}} \circ w_{I_{k-2}} \circ w_{I_{k-3}} \circ \cdots \circ w_{I_0}(x) \quad (k \geq 1) \tag{5.18}$$

and analyze its asymptotic behavior through the properties of $w_{I_k}(x), k \in \mathbb{N}$. It has been proved that if the $w_{I_k}(x)$ have *some* contractive properties, the transition probability kernel of $Z_n(x)$ converges to a limit probability measure, unique for all initial states $x \in \mathcal{D}$. The required contractive properties may be slightly different: [38] studied the case of Lipschitz functions $w_{I_k}(x)$ "contracting on average", while similar results have been obtained by [144] without the continuity requirement on $w_{I_k}(x)$, by [143] for "locally contractive" functions and by [74] for "non-separating on average" functions. A useful survey on the argument has been recently proposed by [70].

Let us show how to exploit the IRF theory in our framework.

The evolution of $(D_k)_{k \in \mathbb{N}}$ can be modeled by IRF. We consider the complete metric space $\mathcal{D}$ naturally endowed with the Euclidean metric d of $\mathbb{R}$, the measurable space

$\mathbf{S} = \{0,1\} \times \mathbb{R}$ and the stochastic process $I_k = (U_k, N_{k+1})$, $k = 0, 1, 2, \dots$, on $\mathcal{S}$, and we define the random function

$$w_{I_k}(x) = \mathrm{q}x + \mathrm{w}\mathbb{1}_{(\mathrm{cq}x+\mathrm{cw}\left(\frac{1}{2}-U_k\right),+\infty)}(N_{k+1}) - \mathrm{w}U_k, \quad x \in \mathcal{D} \tag{5.19}$$

that describes the dynamics of $(D_k)_{k\in\mathbb{N}}$. The key result for our purpose is the following theorem (here stated for compact spaces), which does not require continuity. Let d_W be the Wasserstein (or Kantorovich) distance between probability measures defined as

$$d_W(\mu, \nu) = \sup_{f\in 1\text{-Lip}(\mathcal{D})} \int f \mathrm{d}\mu - \int f \mathrm{d}\nu \tag{5.20}$$

where 1-Lip($\mathcal{D}$) indicates the set of all the Lipschitz functions with Lipschitz constant equal to 1 on $\mathcal{D}$ (see [144] and [121, Section 2.1, Example 3.2.2] for more details on this metric).

Theorem 9 *Stenflo Theorem [144, Theorem 1].*
Suppose that there exists a constant $l < 1$ such that

$$\mathbb{E}[d(w_{I_0}(x), w_{I_0}(y))] \leq l\ d(x, y) \tag{5.21}$$

for all $x, y \in \mathcal{D}$, $(\mathcal{D}, d)$ being a compact metric space. Then there exists a unique invariant probability measure μ for the Markov Process Z_n and there exists a positive constant $\gamma_{\mathcal{D}}$ such that

$$\sup_{x\in\mathcal{D}} d_W(P^n(x,\cdot), \mu(\cdot)) \leq \frac{\gamma_{\mathcal{D}}}{1-l} l^n \qquad n \geq 0 \tag{5.22}$$

where $P^n(x,\cdot)$ is the n-step transition probability kernel of the Markov Process $Z_n(x)$ and

Theorem 8 can be proved applying the Stenflo Theorem.
Proof of Theorem 8.
Let us analyze the condition (5.21). Consider $x, y \in \mathcal{D}$ with $x > y$ (recall that $\mathrm{q} > 0, \mathrm{w} > 0$). Let $H = H(x, y, I_0)$ and $\mathcal{I}_u$ be defined by

$$\begin{aligned} H &:= \mathbb{1}_{\left(\mathrm{cq}y+\mathrm{cw}\left(\frac{1}{2}-U_0\right),\mathrm{cq}x+\mathrm{cw}\left(\frac{1}{2}-U_0\right)\right)}(N_1) \\ \mathcal{I}_u &:= \frac{1}{\sqrt{2\pi}\sigma} \int_{\mathrm{cq}y+\mathrm{cw}\left(\frac{1}{2}-u\right)}^{\mathrm{cq}x+\mathrm{cw}\left(\frac{1}{2}-u\right)} e^{-\frac{n^2}{2\sigma^2}} \mathrm{d}n \\ &= \frac{1}{2}\mathrm{erfc}\left(c\frac{\mathrm{q}y + \mathrm{w}\left(\frac{1}{2}-u\right)}{\sigma\sqrt{2}}\right) - \frac{1}{2}\mathrm{erfc}\left(c\frac{\mathrm{q}x + \mathrm{w}\left(\frac{1}{2}-u\right)}{\sigma\sqrt{2}}\right). \end{aligned}$$

Hence,

$$\begin{aligned}
&\mathbb{E}\left[|w_{(U_0,N_1)}(x) - w_{(U_0,N_1)}(y)|\right] = \mathbb{E}\left[|\mathrm{q}(x-y) - \mathrm{w}H|\right] \\
&= \sum_{u\in\{0,1\}} \mathrm{P}(U_0 = u)\frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}} f_{N_1}(n)|\mathrm{q}(x-y) - \mathrm{w}H|\mathrm{d}n \\
&= \frac{1}{2}\sum_{u\in\{0,1\}}\int_{\mathbb{R}} e^{-\frac{n^2}{2\sigma^2}}|\mathrm{q}(x-y) - \mathrm{w}H|\mathrm{d}n \\
&= \frac{1}{2}\sum_{u\in\{0,1\}} |\mathrm{q}(x-y) - \mathrm{w}|\mathcal{I}_u + \mathrm{q}(x-y)(1-\mathcal{I}_u).
\end{aligned} \tag{5.23}$$

If $\mathrm{q}(x-y) > \mathrm{w}$, then $\mathbb{E}\left[|w_{(U_0,N_1)}(x) - w_{(U_0,N_1)}(y)|\right] < \mathrm{q}(x-y)$ and the contraction would be proved with $l = \mathrm{q}$. This is never the case when $\mathrm{q} < \frac{1}{3}$, $|x-y| \leq -2\frac{b}{a} = 2\frac{\mathrm{w}}{1-\mathrm{q}} < \frac{\mathrm{w}}{\mathrm{q}}$ for every $x, y \in \mathcal{D}$.

Let us then consider $\mathrm{q}(x-y) < \mathrm{w}$. We can write

$$\begin{aligned}
&\mathbb{E}\left[|w_{(U_0,N_1)}(x) - w_{(U_0,N_1)}(y)|\right] \\
&\quad= \frac{1}{2}\sum_{u\in\{0,1\}} (\mathrm{w} - \mathrm{q}(x-y))\mathcal{I}_u + \mathrm{q}(x-y)(1-\mathcal{I}_u) \\
&\quad\leq \frac{1}{2}\sum_{u\in\{0,1\}} \mathrm{w}\mathcal{I}_u + \mathrm{q}(x-y).
\end{aligned} \tag{5.24}$$

The last expression is obtained by neglecting $-\sum_{u\in\{0,1\}}$ q $(x-y)\mathcal{I}_u$, which is the sum of two second degree terms in $(x-y)$, since, by the integral mean value theorem,

$$\mathcal{I}_u = \frac{1}{\sqrt{2\pi}\sigma}c\mathrm{q}(x-y)e^{-\frac{n_0^2}{2\sigma^2}} \tag{5.25}$$

for some $n_0 \in \left[c\mathrm{q}y + c\mathrm{w}\left(\frac{1}{2} - u\right), c\mathrm{q}x + c\mathrm{w}\left(\frac{1}{2} - u\right)\right]$, $(n_0 \neq 0)$. The remaining terms are of order one, then $\frac{1}{2}\sum_{u\in\{0,1\}} \mathrm{w}\mathcal{I}_u + \mathrm{q}(x-y)$ is a suitable approximation of the mean when $x \to y$. Notice also that

$$\frac{1}{2}\sum_{u\in\{0,1\}} \mathrm{w}\mathcal{I}_u + \mathrm{q}(x-y) = F(x) - F(y) \tag{5.26}$$

where

$$F(x) = \mathrm{q}x - \frac{\mathrm{w}}{4}\mathrm{erfc}\left(\frac{c\mathrm{q}x + c\frac{\mathrm{w}}{2}}{\sigma\sqrt{2}}\right) - \frac{\mathrm{w}}{4}\mathrm{erfc}\left(\frac{c\mathrm{q}x - c\frac{\mathrm{w}}{2}}{\sigma\sqrt{2}}\right).$$

Therefore, the thesis is achieved if $F(x)$ is a contraction; since $F(x)$ is differentiable and monotone increasing, its Lipschitz constant is the maximum of its first derivative:

$$F'(x) = \mathrm{q} + \frac{c\mathrm{w}\mathrm{q}}{2\sigma\sqrt{2\pi}}\left[\exp\left(-\frac{\left(c\mathrm{q}x + c\frac{\mathrm{w}}{2}\right)^2}{2\sigma^2}\right) + \exp\left(-\frac{\left(c\mathrm{q}x - c\frac{\mathrm{w}}{2}\right)^2}{2\sigma^2}\right)\right] \tag{5.27}$$

In order to find the maximum of $F'(x)$, let us compute:

$$\begin{aligned} &F''(x) = \\ &= -\frac{c\mathrm{wq}}{2\sigma\sqrt{2\pi}} \frac{2\left(c\mathrm{q}x + c\frac{\mathrm{w}}{2}\right)c\mathrm{q}}{2\sigma^2} \exp\left(-\frac{\left(c\mathrm{q}x + c\frac{\mathrm{w}}{2}\right)^2}{2\sigma^2}\right) + \\ &\quad - \frac{c\mathrm{wq}}{2\sigma\sqrt{2\pi}} \frac{2\left(c\mathrm{q}x - c\frac{\mathrm{w}}{2}\right)c\mathrm{q}}{2\sigma^2} \exp\left(-\frac{\left(c\mathrm{q}x - c\frac{\mathrm{w}}{2}\right)^2}{2\sigma^2}\right) \end{aligned}$$

which is null for x satisfying:

$$\left(\mathrm{q}x + \frac{\mathrm{w}}{2}\right)\exp\left(-\frac{c^2\mathrm{qw}x}{\sigma^2}\right) + \left(\mathrm{q}x - \frac{\mathrm{w}}{2}\right) = 0 \tag{5.28}$$

a solution of which is $x = 0$. Now, considering that $F'(x)$ is a mixture of two Gaussians, two cases may occur: (a) $x = 0$ is the maximum of $F'(x)$; (b) $x = 0$ is a minimum for $F'(x)$ and there are two symmetric maxima ($F''(x)$ is an even function) at $x_0 \in (0, \frac{\mathrm{w}}{1-\mathrm{q}}]$ and $-x_0$, but x_0 cannot be analytically computed from the exponential equation (5.28). By studying the sign of $F''(x)$ for $x \to 0$, it is easy to see that $x = 0$ is a maximum point only for $\frac{c^2\mathrm{w}^2}{\sigma^2} < 4$, that is, only for large noise, which makes this case not really interesting.

On the other hand, when $x = 0$ is a minimum point, $F(x)$ is contractive only under some conditions. In particular, consider $x > 0$ and σ^2 close to zero: by (5.28), $|x - \frac{\mathrm{w}}{2\mathrm{q}}|$ tends to zero more quickly than σ^2, hence $\exp\left(-\frac{\left(c\mathrm{q}x - c\frac{\mathrm{w}}{2}\right)^2}{2\sigma^2}\right)$ tends to one and the maximum of $F'(x)$ (see (5.27) may assume very large values.

More in general, we observe that the points $x = \pm\frac{\mathrm{w}}{2\mathrm{q}}$ are undesired as they are the unique points where the OSA fails: for these values, the error probability given by (5.16) is at least $\frac{1}{4}$, no matter which is the noise variance. This "singular" phenomenon is more evident when the noise is small; in terms of $F(x)$, it causes large variations, hence the loss of the contractivity, in a neighborhood of the point $\pm\frac{\mathrm{w}}{2\mathrm{q}}$, the radius of the neighborhood being larger for smaller σ^2.

This problem is bypassed if we consider $\mathrm{q} < \frac{1}{3}$ (which corresponds to require a "stronger" stability for the system (5.1), which forces $\pm\frac{\mathrm{w}}{2\mathrm{q}}$ to be outside the state space $\mathcal{D}$. Under this assumption, for any $x \in \mathcal{D}$

$$\begin{aligned} \exp\left(-\frac{\left(c\mathrm{q}x + c\frac{\mathrm{w}}{2}\right)^2}{2\sigma^2}\right) &< \exp\left(-\frac{\left(-c\mathrm{q}\frac{\mathrm{w}}{1-\mathrm{q}} + c\frac{\mathrm{w}}{2}\right)^2}{2\sigma^2}\right) \\ \exp\left(-\frac{\left(c\mathrm{q}x - c\frac{\mathrm{w}}{2}\right)^2}{2\sigma^2}\right) &< \exp\left(-\frac{\left(c\mathrm{q}\frac{\mathrm{w}}{1-\mathrm{q}} - c\frac{\mathrm{w}}{2}\right)^2}{2\sigma^2}\right) \end{aligned} \tag{5.29}$$

hence

$$F'(x) \leq \mathrm{q} + \frac{c\mathrm{wq}}{\sigma\sqrt{2\pi}} \exp\left(-\frac{c^2\mathrm{w}^2 \left(\frac{1-3\mathrm{q}}{2(1-\mathrm{q})}\right)^2}{2\sigma^2} \right) \tag{5.30}$$

that is, a sufficient condition for $F'(x) < 1$ is

$$\mathrm{q}\frac{-\frac{b}{a}c}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(\frac{b}{a}c\right)^2 (1-3\mathrm{q})^2}{8\sigma^2} \right) < 1 \tag{5.31}$$

where we used the fact that $\mathrm{w} = -\frac{b(1-\mathrm{q})}{a}$. As $\max_{t\geq 0} te^{-t^2} = \frac{1}{\sqrt{2e}}$,

$$\mathrm{q}\frac{-\frac{b}{a}c}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(\frac{b}{a}c\right)^2 (1-3\mathrm{q})^2}{8\sigma^2} \right) \leq \frac{2\mathrm{q}}{\sqrt{\pi}(1-3\mathrm{q})} \frac{1}{\sqrt{2e}}$$

which is smaller than 1 whenever

$$\mathrm{q} < \frac{1}{3+\sqrt{\frac{2}{e\pi}}} \tag{5.32}$$

which approximately corresponds to $a < -5/4$. This is then the required condition to have average contraction; notice that it depends only on a and not on the other parameters or on the noise.

At this point, the hypotheses of Stenflo's Theorem are fulfilled, hence the existence and uniqueness of a limit probability measure μ is assured for $(D_k)_{k\in\mathbb{N}}$; in particular, it is assured for *any* initial state $D_0 \in \mathcal{D}$.

Now, let us prove the convergence of the BER. First, notice that

$$\begin{aligned} \mathrm{BER}(\mathbb{D}^{(1)}) &= \frac{1}{K}\sum_{k=0}^{K-1} \mathrm{P}(\widehat{U}_k \neq U_k) = \\ &= \frac{1}{K}\sum_{k=0}^{K-1} \int_{\mathcal{D}} \mathrm{P}(\widehat{U}_k \neq U_k | D_k = z) P^k(0, \mathrm{d}z) \\ &= \frac{1}{K}\sum_{k=0}^{K-1} (P^k g)(0) \end{aligned} \tag{5.33}$$

where $(P^k g)(0) = \int_{\mathcal{D}} P(0,\mathrm{d}z)g(z)$. Moreover, we have that $g \in L_g$-Lip$(\mathcal{D})$ where $L_g = \max_{z\in\mathcal{D}} |g'(z)|$ and

$$|g'(z)| = \frac{c\mathrm{q}}{2\sigma\sqrt{2\pi}} \left| -e^{-\frac{(c\mathrm{q}z+c\mathrm{w})^2}{2\sigma^2}} + e^{-\frac{(-c\mathrm{q}z+c\mathrm{w})^2}{2\sigma^2}} \right| \leq \frac{c\mathrm{q}}{\sigma\sqrt{2\pi}} \tag{5.34}$$

then $L_g \leq \frac{c\mathrm{q}}{\sigma\sqrt{2\pi}}$ is finite. Since for any $L > 0$

$$\sup_{f\in L\text{-Lip}(\mathcal{D})}\left|\int f\mathrm{d}(\mu-\nu)\right| = \sup_{f\in L\text{-Lip}(\mathcal{D})} L\left|\int \frac{1}{L}f\mathrm{d}(\mu-\nu)\right|$$
$$\leq \sup_{f\in 1\text{-Lip}(\mathcal{D})} L\left|\int f\mathrm{d}(\mu-\nu)\right| = L\ d_W(\mu,\nu)$$

we have

$$\begin{aligned}\sup_{x\in\mathcal{D}}\left|(P^k g)(x) - \int g\mathrm{d}\mu\right| &= \sup_{x\in\mathcal{D}}\left|\int g(z)P^k(x,\mathrm{d}z) - \int g\mathrm{d}\mu\right| \\ &\leq \sup_{x\in\mathcal{D}}\sup_{f\in L_g\text{-Lip}}\left|\int f(z)P^k(x,\mathrm{d}z) - \int f\mathrm{d}\mu\right| \\ &= \sup_{x\in\mathcal{D}} L_g d_W(P^k(x,\cdot),\mu(\cdot)) \overset{k\to\infty}{\longrightarrow} 0.\end{aligned} \tag{5.35}$$

The convergence is then assured also for the Cesàro sum, for any starting state $x \in \mathcal{D}$:

$$\frac{1}{K}\sum_{k=0}^{K-1}(P^k g)(x) \overset{K\to\infty}{\longrightarrow} \int g\mathrm{d}\mu \quad \forall x \in \mathcal{D}. \tag{5.36}$$

As convergence holds for any initial state, in particular it holds for $x = 0$, which is our case of interest. ■

5.3.3 Observations

The OSA performance improve if the values of a, b and c increase (under the constraints $a < 0$, $b > 0$ and $c > 0$). This occurs because the distance between the possible states is $c\mathrm{w} = cb\frac{1-e^a}{-a}$ and larger distances may counterbalance a larger noise in the decoding. $c^2\mathrm{w}^2$ can be also interpreted as the signal power per channel use, hence we can define the signal-to-noise ratio as SNR= $\frac{c^2\mathrm{w}^2}{\sigma^2}$ and evaluate the performance of the OSA with respect to it. In Figure 5.1 we depict the Bit Error Rate for $b = c = 1$ and $a = -5/4, -2, -10$ with respect to the SNR (expressed in dB): we notice that for the same SNR, smaller values of a are slightly preferable for mid SNR values: this is due to the fact that if $\mathrm{q} = e^a$ is very small, the system "loses its memory", i.e., in the iteration $x_{k+1} = \mathrm{q}x_k + \mathrm{w}u_k$ the value of x_k becomes less significant; analogously, an incorrect estimate of x_k by the OSA is less dramatic.

Furthermore, we have reported the (numerically approximated) densities of the limit probability measures in the cases $a = -5/4$ and $a = -2$ in Figures 5.2-5.3 for $\sigma^2 = 1$: as expected, the densities are symmetric and with global maximum in zero, which corresponds to null error in the state estimation; their support is a Cantor set. Since at each step a product by $\mathrm{q} = e^a$ is performed, for $a = -2$, the density is close to a sequence of spikes, while for $a = -5/4$ it is more distributed over the state space.

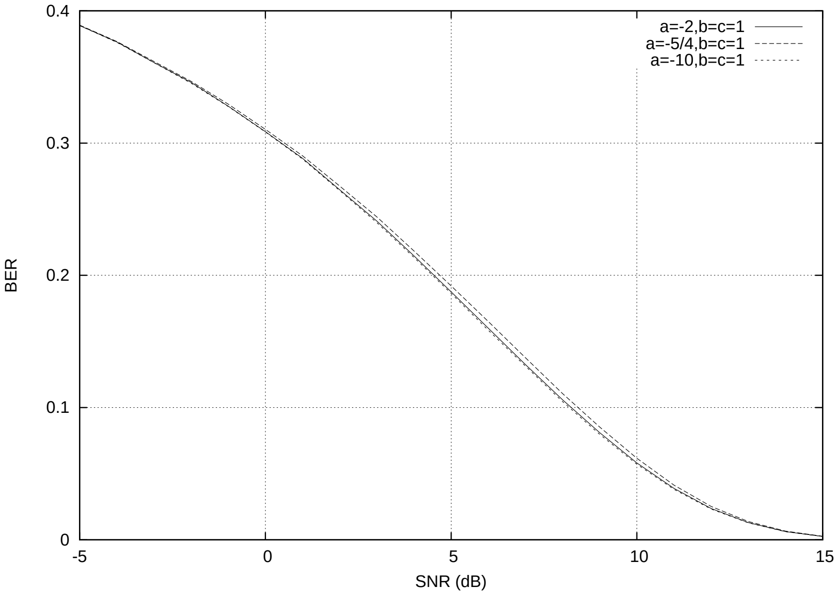

Figure 5.1: BER for $a = -2, -10, -5/4$.

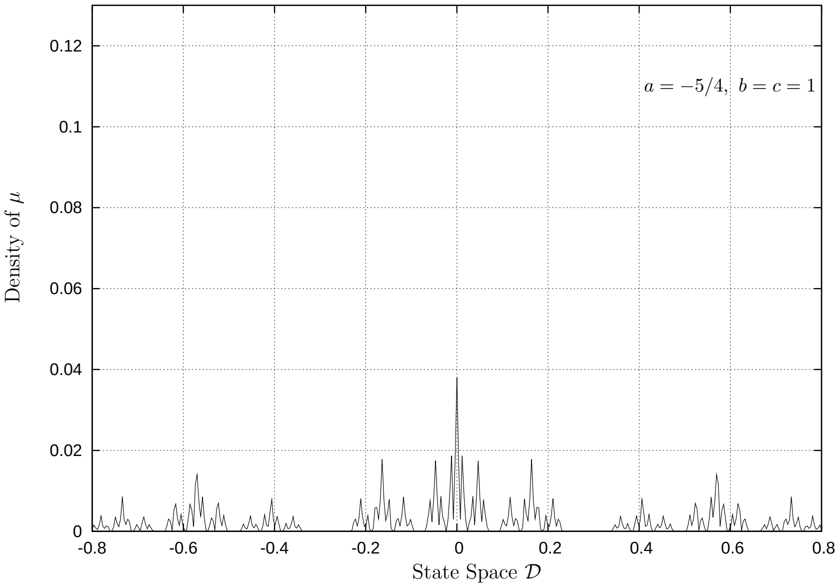

Figure 5.2: Density of the Limit Probability Measure μ for $a = -5/4$, $b = c = 1$, $\sigma^2 = 1$.

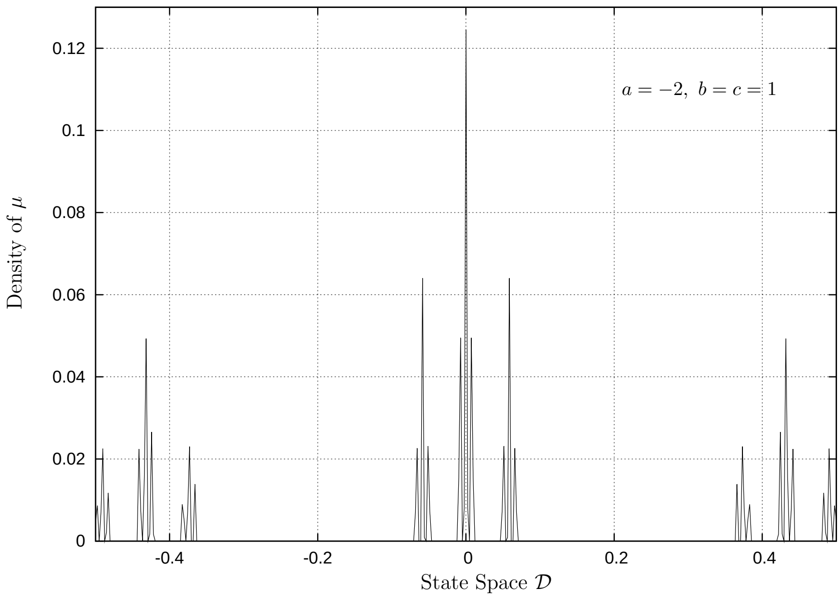

Figure 5.3: Density of the Limit Probability Measure μ for $a = -2$, $b = c = 1$, $\sigma^2 = 1$.

In the next (see Remark 6), we will show that the probability measure μ has support on the Cantor set $\mathcal{D}_{\mathcal{C}} = \mathrm{w}\big\{z \in \mathcal{D} : z = \mathrm{w}\sum_{n=0}^{\infty} \alpha_n \mathrm{q}^n$ for some $\alpha_n \in \{-1, 0, 1\},\ n = 0, 1, 2\ldots\big\}$.

5.4 Alternative Theoretic Analysis with Markov Processes

In the differentiation case, an analogous of the Performance Theorem was proved using the Ergodic Theorem of Markov Processes (see Theorem 7 and equation (4.31)); in that case, IRF could not be used since no contractive property occurred.

Though in the present case the use of IRF is very efficient, one might wonder if we could prove the Performance Theorem 5.15 using Markov Processes. The answer is yes, but with the introduction of some non-classical Markov tool as explained in the next.

As shown in Chapter 4, typically, the steps to describe the asymptotic behavior of the system using the Ergodic Theorem for Markov Processes 7 are the following: (1) prove the existence of an i.p.m.; (2) prove its uniqueness, classically obtained through the ϕ-irreducibility; (3) uniqueness implies ergodicity, which is the condition that makes the Ergodic Theorem true. Nevertheless, the Ergodic Theorem holds *almost everywhere*, while in our case, we fix the starting point. In other terms, In order to have the expected result, we have to prove that our starting point does not belong to the negligible set that does not lead to convergence.

We will see that, for the current problem, existence is obtained with no efforts. On the other hand, uniqueness is a tricky issue: ϕ-irreducibility is typically obtained with respect to the Lebesgue measure, but this is not the case, since the i.p.m.'s are singular with respect to it; moreover, finding an other measure ϕ is a very tricky issue. Sometimes, a ϕ is provided by the so-called topological irreducibility, but the SEEMS to be not the case.

Anyway, uniqueness can be proved in an original way, that is, exploiting the contractive properties of the operator $T : \mu \to \mu P$ acting on probability measures and the Banach Fixed Point Theorem. At this point, we will have all the conditions required for the Ergodic Theorem in a slightly different version, say convergence holds for *any* starting state, but it is *weak*, which is sufficient in our context.

We notice that this procedure exactly leads to the same results obtained with IRF; however, we report it because of its elements of originality.

Proposition 9 (Existence of an i.p.m.) *The transition probability kernel of the Markov Process* $(D_k)_{k\in\mathbb{N}}$ *on* $(\mathcal{D}, \mathcal{B}(\mathcal{D}))$, $\mathcal{D} = \left[-\frac{\mathrm{w}}{1-\mathrm{q}}, \frac{\mathrm{w}}{1-\mathrm{q}}\right]$ *admits at least one invariant probability measure.*

Proof $\mathcal{D}$ is compact. This property, along with the weak-Feller property, guarantees the existence of an i.p.m. (see for example [68, Theorem 7.2.3]). The weak-Feller property is easily proved: if $f \in C_b(\mathcal{D})$, then $Pf(\xi) = \sum_{i\in\{-\mathrm{w},0,\mathrm{w}\}} P(\xi, \mathrm{q}\xi+i)f(\mathrm{q}\xi+i) \in C_b(\mathcal{D})$ since $P(\xi, \mathrm{q}\xi + i)$, $i \in \{-\mathrm{w}, 0, \mathrm{w}\}$, are continuous and bounded as functions of ξ.

Proposition 10 (Uniqueness of the i.p.m.) *The condition* $\mathrm{q} < \frac{1}{3+\sqrt{\frac{2}{e\pi}}}$ *is sufficient so that the transition probability kernel of* $(D_k)_{k\in\mathbb{N}}$ *admits a unique i.p.m..*

Proof Let μ and ν be two probability measures on $\mathcal{D}$. We define the operator

$$T : \mu \to T\mu = \mu P \tag{5.37}$$

on the space of the probability measures on $\mathcal{D}$. Now, we prove the following lemmas.

Lemma 8 *If* $\mathrm{q} < \frac{1}{3+\sqrt{\frac{2}{e\pi}}}$, *for any* $f \in 1$ *-Lip*$(\mathcal{D})$, $Pf \in h$*-Lip*$(\mathcal{D})$ *with* $h < 1$.

Proof Given any $f \in$ 1-Lip$(\mathcal{D})$ and $\xi, \zeta \in \mathcal{D}$,

$$\begin{aligned}
&Pf(\xi) - Pf(\zeta)\\
&= \sum_{i\in\{-\mathrm{w},0,\mathrm{w}\}} f(\mathrm{q}\xi + i)P(\xi, \mathrm{q}\xi + i) - f(\mathrm{q}\zeta + i)P(\zeta, \mathrm{q}\zeta + i)\\
&= f(\mathrm{q}\xi) + \sum_{i\in\{-\mathrm{w},\mathrm{w}\}} [f(\mathrm{q}\xi + i) - f(\mathrm{q}\xi)]P(\xi, \mathrm{q}\xi + i)+\\
&- f(\mathrm{q}\zeta) - \sum_{i\in\{-\mathrm{w},\mathrm{w}\}} [f(\mathrm{q}\zeta + i) - f(\mathrm{q}\zeta)]P(\zeta, \mathrm{q}\zeta + i).
\end{aligned}$$

Adding and removing the quantity $[f(\mathrm{q}\xi + \mathrm{w}) - f(\mathrm{q}\xi)]P(\zeta, \mathrm{q}\zeta + \mathrm{w}) + [f(\mathrm{q}\xi - \mathrm{w}) - f(\mathrm{q}\xi)]P(\zeta, \mathrm{q}\zeta - \mathrm{w})$, using the Lipschitz property of f and recalling that $P(\zeta, \mathrm{q}\zeta + \mathrm{w}) +$

$P(\zeta, \mathrm{q}\zeta - \mathrm{w}) \leq \frac{1}{2}$ for any $\zeta \in \mathcal{D}$, we obtain

$$\begin{aligned}
&Pf(\xi) - Pf(\zeta) = \\
&[f(\mathrm{q}\xi + \mathrm{w}) - f(\mathrm{q}\xi)][P(\xi, \mathrm{q}\xi + \mathrm{w}) - P(\zeta, \mathrm{q}\zeta + \mathrm{w})] \\
&\ + [f(\mathrm{q}\xi - \mathrm{w}) - f(\mathrm{q}\xi)][P(\xi, \mathrm{q}\xi - \mathrm{w}) - P(\zeta, \mathrm{q}\zeta - \mathrm{w})] \\
&\ + [f(\mathrm{q}\xi + \mathrm{w}) - f(\mathrm{q}\xi) - f(\mathrm{q}\zeta + \mathrm{w}) + f(\mathrm{q}\zeta)]P(\zeta, \mathrm{q}\zeta + \mathrm{w}) \\
&\ + [f(\mathrm{q}\xi - \mathrm{w}) - f(\mathrm{q}\xi) - f(\mathrm{q}\zeta - \mathrm{w}) + f(\mathrm{q}\zeta)]P(\zeta, \mathrm{q}\zeta - \mathrm{w}) \\
&\ + f(\mathrm{q}\xi) - f(\mathrm{q}\zeta) \leq \\
&\leq \mathrm{w}\big|P(\xi, \mathrm{q}\xi + \mathrm{w}) - P(\zeta, \mathrm{q}\zeta + \mathrm{w})\big| \\
&+ \mathrm{w}\big|P(\xi, \mathrm{q}\xi - \mathrm{w}) - P(\zeta, \mathrm{q}\zeta - \mathrm{w})\big| + \mathrm{q}|\xi - \zeta|
\end{aligned}$$

which exactly corresponds to (5.24), hence the proof follows from (5.24)-(5.32).

Lemma 9 *In the hypotheses of Lemma 8, T is a contraction on the metric space of the probability measures on $\mathcal{D}$, endowed with the Wasserstein metric d_W, that is,*

$$d_W(\mu P, \nu P) \leq h \ d_W(\mu, \nu), \quad h < 1. \tag{5.38}$$

Proof Given a function $f \in L_1(\mathcal{D}, \mathcal{B}(\mathcal{D}), \mu P)$,

$$\begin{aligned}
\int f \ d(\mu P) &= \int_{\xi \in \mathcal{D}} \int_{\zeta \in \mathcal{D}} f(\xi) P(\zeta, d\xi) \mu(d\zeta) \\
&= \int Pf \ d\mu
\end{aligned}$$

Hence

$$\begin{aligned}
d(\mu P, \nu P) &= \sup_{f \in 1\text{-Lip}(\mathcal{D})} \left(\int f \ d(\mu P) - \int f \ d(\nu P) \right) \\
&= \sup_{f \in 1\text{-Lip}(\mathcal{D})} \left(\int Pf \ d\mu - \int Pf \ d\nu \right).
\end{aligned}$$

By Lemma 8, if $f \in 1\text{-Lip}(\mathcal{D})$, then $Pf \in h\text{-Lip}(\mathcal{D})$, $h < 1$. Then, for any probability measures μ, ν on $(\mathcal{D}, \mathcal{B}(\mathcal{D}))$ and $f \in 1\text{-Lip}(\mathcal{D})$:

$$\begin{aligned}
&\int Pf \ d\mu - \int Pf \ d\nu \leq \\
&\leq \sup_{g \in h\text{-Lip}(\mathcal{D})} \left(\int g d\mu - \int g d\nu \right) = h \ d(\mu, \nu)
\end{aligned}$$

In particular,

$$\begin{aligned}
&d(\mu P, \nu P) = \\
&= \sup_{f \in 1\text{-Lip}(\mathcal{D})} \left(\int Pf \ d\mu - \int Pf \ d\nu \right) \leq h \ d(\mu, \nu)
\end{aligned}$$

with $h < 1$. At this point, we can conclude the proof of Theorem 10. In fact, under the required conditions, the operator T is a contraction in the space of the probability measures on $(\mathcal{D}, \mathcal{B}(\mathcal{D}))$, then it admits a unique fixed point, i.e., there is a unique probability measure μ such that $T\mu = \mu$, or equivalently $\mu P = \mu$. In conclusion, μ is the unique i.p.m. for $(D_k)_{k \in \mathbb{N}}$.

Now the Performance Theorem is proved using [94, Proposition 12.1.4]. We recall that a sequence of probability measures $(\mu_k)_{k\in\mathbb{N}}$ on a space $\mathcal{D}$ is said to be *tight* if for any $\varepsilon > 0$ there exists a compact subset $\mathrm{C} \subseteq \mathcal{D}$ such that $\liminf_{k\to\infty} \mu_k(\mathrm{C}) \geq 1 - \varepsilon$; moreover, the transition probability kernel P of a Markov Process is said to be *bounded in probability on average* if for each initial condition x the sequence $\frac{1}{K}\sum_{k=0}^{K-1} P^k(x, \cdot)$, $K \in \mathbb{N}$ is tight (see [94, Chapter 12]). Then,

Proposition 11 *[94, Proposition 12.1.4]*
If the transition probability kernel P of a Markov Process is weak Feller, bounded in probability on average and admits a unique i.p.m. μ, then

$$\textit{for every } x \in \mathcal{D}, \quad \frac{1}{K}\sum_{k=0}^{K-1} P^k(x, \cdot) \Rightarrow \mu \tag{5.39}$$

where $\Rightarrow$ indicates the weak convergence*:*

$$\mu_n \Rightarrow \mu \textit{ if } \lim_{n\to\infty} \int f \mathrm{d}\mu_n = \int f \mathrm{d}\mu \textit{ for every } f \in \mathbb{C}_b(\mathcal{D}). \tag{5.40}$$

In our case, tightness is trivially assured by the compactness of $\mathcal{D}$. Hence, the previous propositions lead to the weak convergence, which is sufficient for our purpose (recall that the function g in Theorem 5.15 is bounded and continuous).

Remark 6 *Let $\mathcal{D}_\mathcal{C} = \mathrm{w}\big\{z \in \mathcal{D} : z = \mathrm{w}\sum_{n=0}^{\infty} \alpha_n \mathrm{q}^n$ for some $\alpha_n \in \{-1, 0, 1\}$, $n = 0, 1, 2 \ldots\big\}$. $\mathcal{D}_\mathcal{C}$ is a Cantor (hence closed) set and is an invariant set for our Markov process: $P^n(z, \mathcal{D}_\mathcal{C}) = 1$ if $z \in \mathcal{D}_\mathcal{C}$, for any $n \in \mathbb{N}$. It follows from* (5.39) *and the Portmanteau Theorem [68, Theorem 1.4.16] that $\mu(\mathcal{D}_\mathcal{C}) = 1$.*

5.5 A few simulations

In the next, we report the outcomes of some simulations of our transmission system. Recalling the pattern of the One State Algorithm, notice that $|c\mathrm{w}| = |c\frac{b}{a}(1 - \mathrm{q})|$, which represents the distance between the two possible transmitted signals at each step, plays a fundemental role. A larger value of $|c\mathrm{w}|$ is then desirable, since, as already mentioned, a larger distance improves the reliability of our estimation technique. On the other hand, $c^2\mathrm{w}^2$ can be interpreted as the energy per channel use of our transmission system, then for the applications its value cannot be increased too much.

In the next, we will represent the $\mathrm{BER}(\mathbb{D}^{(1)})$ in function of the SNR of our transmission, that is, $c^2\mathrm{w}^2/\sigma^2$. This quantity represents the proportion between signal and

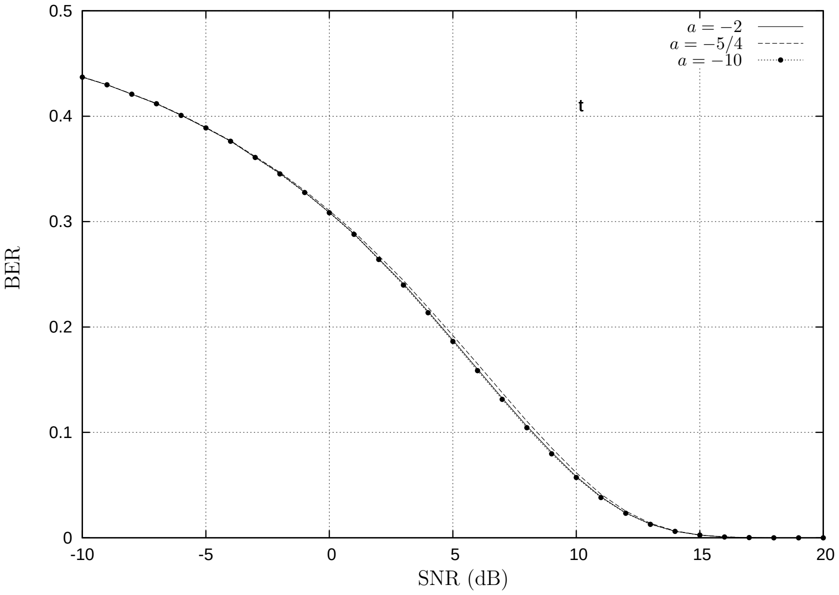

Figure 5.4: Simulations with $b = c = 1$, $a = -5/4, -2, -10$.

noise energies and is usually assumed as reference parameter to study the quality of a transmission.

Given that cw is the leading parameter, in the simulations we fix and $b = c = 1$, while a can vary. In Figure 5.4 the cases $a = -5/4$, $a = -2$ and $a = -10$ are represented: the graphs show the BER($\mathbb{D}^{(1)}$) in function of $SNR = c^2 \mathrm{w}^2/\sigma^2$ expressed in dB. We can notice that the performance improve (that is, the BER decreases) as a decreases; the performance of the cases $a = -2$ and $a = -10$ are very close.

In Figure 5.5 we show a comparison between the Bit Error Rate obtained by analytic computation and by the simulations in the case $b = c = 1$ and $a = -1$: the graphs are coincident. This is only an example, but a perfect consistency between simulated and analytic results has been observed to hold in every case.

5.6 One State Algorithm vs Kalman Filter

As in the Appendix 4.7.9 we have compared the causal BCJR and the causal Least Means Squares Estimation procedures, in the next we compare the One State Algorithm and the (causal) Linear Least Mean Square Estimation, computed through the Kalman Filter, and we propose a theoretical performance analysis.

We know that the Kalman Filter (see Section 2.5.8) is a widely used recursive algorithm that computes the causal linear least squares estimate of the states of a

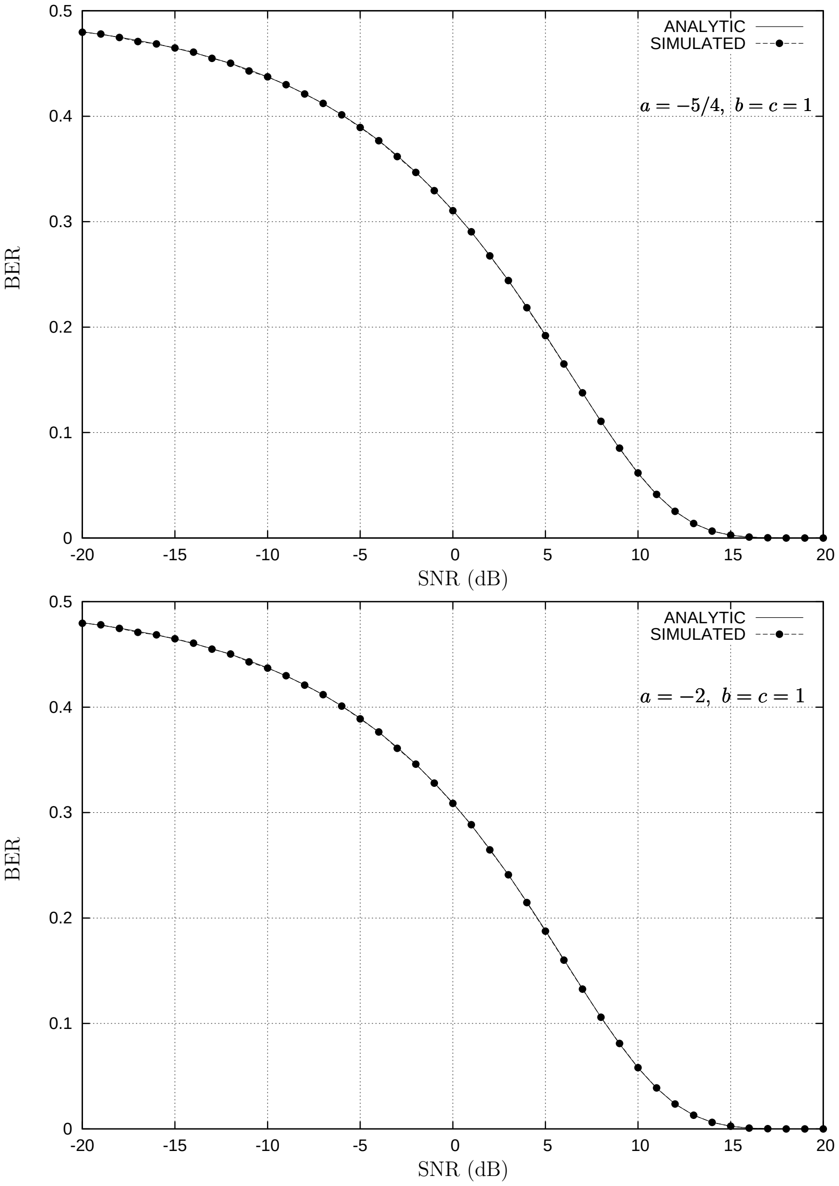

Figure 5.5: Analytic vs Simulated Bit Error Rate (system with $b = c = 1$, $a = -5/4$ and $a = -2$): the results are consistent.

dynamical linear system. Let us consider our dynamical model (5.12)

$$\begin{cases} U_{k-1} \sim \text{Ber}\,(1/2) \\ N_k \sim \mathcal{N}(0, \sigma^2) \\ X_k = \mathrm{q} X_{k-1} + \mathrm{w} U_{k-1} \quad (X_0 = 0) \\ Y_k = c X_k + N_k \end{cases}$$

with $U_{k-1} \in \{-1, 1\}$ (notice that in this Section we assume the binary inputs to be -1 and 1 instead of 0 and 1, the motivation of which will be clear in a while).

For any $k = 1, 2, \dots$, the Kalman Filter (KF for short) computes the Linear Least Mean Squares Estimate (LLMSE, see Section 2.5.7) of X_k given the measurements $Y_1, Y_2, \dots, Y_k$, which we will denote by

$$\text{LLMSE}[X_k | Y_1, \dots Y_k] = \widehat{X}_k^{\text{FK}}. \tag{5.41}$$

In general, the LLMSE of a random variable X given a random variable Z is known to be

$$\text{LLMSE}[X|Z] = \text{cov}[X, Z](\text{cov}[Z, Z])^{-1} Z \tag{5.42}$$

(see [124, Proposition 3]) and the KF allows to compute it in the case (5.41) through a low-complexity iterative algorithm.

Our purpose being to estimate U_k, $k = 0, 1, \dots$, given $\text{LLMSE}[X_k | Y_1, \dots Y_k]$ we will be able to compute

$$\text{LLMSE}[U_{k-1} | Y_1, \dots Y_k] \tag{5.43}$$

and finally we will define

$$\widehat{U}_{k-1}^{\text{KF}} = \text{sgn}\left(\text{LLMSE}[U_{k-1} | Y_1, \dots Y_k]\right) \tag{5.44}$$

where sgn is the sign function:

$$\text{sgn}(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{cases}.$$

The aim of this section is to compare the KF approach and the One State Algorithm. Interesting conclusions will be drawn in terms of complexity and performance: in particular, we will show that the two algorithms have very similar patterns and that the OSA performs better in case of low SNR (in case of long time transmissions), the latter result being analytically obtained using the IRF theory.

Let us first describe the KF approach to estimate the input of the system 5.12.

5.6.1 The Kalman Filter approach

KF is typically used to provide $\text{LLMSE}[X_k | Y_1, \dots, Y_k]$ for linear dynamical systems of kind

$$\begin{cases} X_k = a_1 X_{k-1} + a_2 u_{k-1} + a_3 M_k \\ Y_k = a_4 X_k + N_k \end{cases} \tag{5.45}$$

where $a_1, a_2, a_3, a_4 \in \mathbb{R}^1$, u_k is a known control input, M_k and N_k are random disturbances with $\mathbb{E}[M_k] = \mathbb{E}[N_k] = 0$, $\text{cov}[M_k, M_h] = \eta_k \delta_{kh}$, $\text{cov}[N_k, N_h] = \theta_k \delta_{kh}$ (where $\delta_{kh} = 1$ if $k = h$ and 0 otherwise), respectively called *process noise* and *measurement noise*. If the noises are Gaussian, the Kalman Filter is optimal.

In our system 5.12, no control input neither process noise are present; however, we can consider our input U_k as the process noise (since $\mathbb{E}[U_k] = 0$ and $\text{cov}[U_k, U_h] = \delta_{kh}$) and the control input to be null. Furthermore, in our case the measurement noise is Gaussian and $\theta_k = \sigma^2$ is constant.

In this setting, the KF procedure is as follows (see [124] or [29, Chapter 7] for more details). Let

$$\Sigma_k = \mathbb{E}\big[\,(X_{k+1} - \text{LLMSE}[X_{k+1}|Y_1, \ldots, Y_k])^2\,\big] \tag{5.46}$$

then,

1. Initialization: $\widehat{x}_0^{\text{KF}} = 0$ and $\Sigma_0 = \mathbb{E}[(X_1)^2] = \mathbb{E}[(\mathrm{w}U_0)^2] = \mathrm{w}^2$;

2. For any $k = 1, 2, \ldots$, given the measurements $y_1, \ldots, y_k$

$$\begin{aligned} K_k &= \frac{c\Sigma_{k-1}}{c^2\Sigma_{k-1} + \sigma^2} \\ \Sigma_k &= \mathrm{q}^2(1 - cK_k)\Sigma_{k-1} + \mathrm{w}^2 \\ \widehat{x}_k^{\text{KF}} &= \text{LLMSE}[X_k|y_1, \ldots, y_k] = \mathrm{q}\widehat{x}_{k-1}^{\text{KF}} + K_k(y_k - c\mathrm{q}\widehat{x}_{k-1}^{\text{KF}}) \end{aligned} \tag{5.47}$$

K_k is sometimes called the *Kalman gain.*

Given the causal LLMSE of the states, we can also compute the causal LLMSE of the inputs U_k, $k \in \mathbb{N}$. In fact,

Lemma 10

$$\text{LLMSE}[U_{k-1}|Y_1, \ldots, Y_k] = \text{LLMSE}[U_{k-1}|Y_k - c\mathrm{q}\widehat{X}_{k-1}^{KF}] \tag{5.48}$$

Proof By [124, Proposition 4b - Property 4],

$$\begin{aligned} &\text{LLMSE}[U_{k-1}|Y_1, \ldots, Y_k] = \text{LLMSE}[U_{k-1}|Y_1, \ldots, Y_{k-1}] + \\ &+ \text{LLMSE}\big[U_{k-1} - \text{LLMSE}[U_{k-1}|Y_k, \ldots, Y_{k-1}]\big|Y_k - \text{LLMSE}[Y_k|Y_1, \ldots, Y_{k-1}]\big] \end{aligned} \tag{5.49}$$

Applying the formula (5.42)

$$\text{LLMSE}[U_{k-1}|Y_1, \ldots, Y_{k-1}] = 0$$

since $\text{cov}[U_{k-1}, Y_j] = 0$ for any $j = 1, \ldots, k-1$. Furthermore, as $Y_k = c(\mathrm{q}X_{k-1} + \mathrm{w}U_{k-1}) + N_k$

$$\begin{aligned} \text{LLMSE}[Y_k|Y_1, \ldots, Y_{k-1}] &= \text{LLMSE}[c(\mathrm{q}X_{k-1} + \mathrm{w}U_{k-1}) + N_k|Y_1, \ldots, Y_{k-1}] \\ &= c\mathrm{q}\widehat{X}_{k-1}^{\text{KF}} \end{aligned}$$

[1] KF can be applied to n-dimensional systems with coefficients varying in time: for our purpose, here we outline only the one-dimensional case, with constant coefficients.

by the linearity of (5.42) (see [124, Propositon 4a - Property 5]), and this proves the lemma.

Lemma 11

$$\mathrm{LLMSE}[U_{k-1}|Y_1,\dots Y_k] = \mathrm{w}\frac{K_k}{\Sigma_{k-1}}(Y_k - c\mathrm{q}\widehat{X}^{KF}_{k-1}) \tag{5.50}$$

Proof By the previous lemma, $\widehat{U}^{\mathrm{KF}}_{k-1} = \mathrm{LLMSE}[U_{k-1}|Y_k - c\mathrm{q}\widehat{X}^{\mathrm{KF}}_{k-1}]$ By (5.42),

$$\mathrm{LLMSE}[U_{k-1}|Y_k - c\mathrm{q}\widehat{X}^{\mathrm{KF}}_{k-1}] = \mathbb{E}[U_{k-1}(Y_k - c\mathrm{q}\widehat{X}^{\mathrm{KF}}_{k-1})]\left(\mathbb{E}[(Y_k - c\mathrm{q}\widehat{X}^{\mathrm{KF}}_{k-1})^2]\right)^{-1}(Y_k - c\mathrm{q}\widehat{X}^{\mathrm{KF}}_{k-1}) \tag{5.51}$$

$$\mathbb{E}[U_{k-1}(Y_k - c\mathrm{q}\widehat{X}^{\mathrm{KF}}_{k-1})] = \mathbb{E}[U_{k-1}Y_k] = c\mathrm{w}. \tag{5.52}$$

Moreover,

$$\mathrm{LLMSE}[X_k|Y_1,\dots,Y_{k-1}] = \mathrm{LLMSE}[\mathrm{q}X_{k-1} + \mathrm{w}U_{k-1}|Y_1,\dots,Y_{k-1}] = \mathrm{q}\widehat{X}^{\mathrm{KF}}_{k-1}) \tag{5.53}$$

hence

$$\begin{aligned}\mathbb{E}[(Y_k - c\mathrm{q}\widehat{X}^{\mathrm{KF}}_{k-1})^2] &= \mathbb{E}[(Y_k - c\mathrm{LLMSE}[X_k|Y_1,\dots,Y_{k-1}])^2] \\ &= \mathbb{E}[(N_k + cX_k - c\mathrm{LLMSE}[X_k|Y_1,\dots,Y_{k-1}])^2] \\ &= \mathbb{E}[N_k^2] + c^2\mathbb{E}[X_k - \mathrm{LLMSE}[X_k|Y_1,\dots,Y_{k-1}])^2] \\ &= \sigma^2 + c^2\Sigma_{k-1}\end{aligned} \tag{5.54}$$

and finally,

$$\mathrm{LLMSE}[U_{k-1}|Y_1,\dots Y_k] = \frac{c\mathrm{w}}{\sigma^2 + c^2\Sigma_{k-1}}(Y_k - c\mathrm{q}\widehat{X}^{\mathrm{KF}}_{k-1}) = \mathrm{w}\frac{K_k}{\Sigma_{k-1}}(Y_k - c\mathrm{q}\widehat{X}^{\mathrm{KF}}_{k-1}) \tag{5.55}$$

Finally, given that $U_k \in \{-1, 1\}$, we force its estimate to lie in the same set by computing:

$$\begin{aligned}\widehat{U}^{\mathrm{KF}}_{k-1} &= \mathrm{sgn}\left(\mathrm{LLMSE}[U_{k-1}|Y_1,\dots Y_k]\right) \\ &= \mathrm{sgn}\left(Y_k - c\mathrm{q}\widehat{X}^{\mathrm{KF}}_{k-1}\right).\end{aligned} \tag{5.56}$$

as $\mathrm{w} > 0$, $\Sigma_k > 0$ and $K_k > 0$ for any $k \in \mathbb{N}$.

Let us remark also that

Lemma 12

$$\Sigma_k = \mathrm{w}^2 + \mathrm{w}^2\sum_{i=1}^{k}\mathrm{q}^{2i}\prod_{j=k-i+1}^{k}(1 - cK_j) = \mathrm{w}^2 + \mathrm{w}^2\sum_{i=1}^{k}\mathrm{q}^{2i}\prod_{j=k-i+1}^{k}\frac{1}{1 + \frac{c^2}{\sigma^2}\Sigma_{j-1}} \tag{5.57}$$

and in particular it is an increasing, convergent sequence. Furthermore, K_k is an increasing, convergent sequence in $\left(0, \frac{1}{c}\right)$.

One State Algorithm	Kalman Filter
$\widehat{x}_k^{\text{OSA}} = \text{q}\widehat{x}_{k-1}^{\text{OSA}} + \text{w sgn}(y_k - c\text{q}\widehat{x}_{k-1}^{\text{OSA}})$	$\widehat{x}_k^{\text{KF}} = \text{q}\widehat{x}_{k-1}^{\text{KF}} + K_k(y_k - c\text{q}\widehat{x}_{k-1}^{\text{KF}})$
$\widehat{u}_{k-1}^{\text{OSA}} = \text{sgn}(y_k - c\text{q}\widehat{x}_{k-1}^{\text{OSA}})$	$\widehat{u}_{k-1}^{\text{KF}} = \text{sgn}(y_k - c\text{q}\widehat{x}_{k-1}^{\text{KF}})$
$\widehat{x}_k^{\text{OSA}} \in \text{w}\{\sum_{i=0}^{\infty} \mu_i \text{q}^i,\ \mu_i \in \{-1,1\}\}$	$\widehat{x}_k^{\text{KF}} \in \mathbb{R}$

Table 5.1: One State Algorithm vs LLMSE estimation through KF.

Proof The expression (5.57) can be trivially proved for $k = 1$:

$$\Sigma_1 = \text{w}^2 + \text{q}^2(1 - cK_1)\Sigma_0 = \text{w}^2 + \text{q}^2(1 - cK_1)\text{w}^2 \tag{5.58}$$

and $K_1 = \frac{1}{c}\frac{1}{1+\frac{c^2\text{w}^2}{\sigma^2}}$. By induction, if the expression holds for a given k, it holds also for $k+1$:

$$\begin{aligned}
\Sigma_{k+1} &= \text{w}^2 + \text{q}^2(1 - cK_{k+1})\Sigma_k \\
&= \text{w}^2 + \text{q}^2(1 - cK_{k+1})\left(\text{w}^2 + \text{w}^2\sum_{i=1}^{k}\text{q}^{2i}\prod_{j=k-i+1}^{k}(1 - cK_j)\right) \\
&= \text{w}^2 + \text{q}^2(1 - cK_{k+1})\text{w}^2 + \text{w}^2\sum_{i=1}^{k}\text{q}^{2i+2}\prod_{j=k-i+1}^{k+1}(1 - cK_j) \\
&= \text{w}^2 + \text{w}^2\sum_{i=1}^{k+1}\text{q}^{2i+2}\prod_{j=k-i+1}^{k+1}(1 - cK_j).
\end{aligned} \tag{5.59}$$

Σ_k is then increasing, which implies that also K_k is so. On the other hand $0 < K_k = \frac{1}{c}\frac{1}{1+\frac{\sigma^2}{c^2\Sigma_{k-1}}} < \frac{1}{c}$ is bounded, hence convergent. Moreover, also Σ_k is bounded and convergent. In particular, since $1 - cK_j < 1 - cK_1 = \frac{1}{1+\text{SNR}}$, we have

$$\text{w}^2 \leq \Sigma_k < \text{w}^2 + \text{w}^2\sum_{i=0}^{k}\text{q}^{2i}\left(\frac{1}{1+\text{SNR}}\right)^i < \text{w}^2 + \text{w}^2\frac{1}{1 - \frac{\text{q}^2}{1+\text{SNR}}}. \tag{5.60}$$

This lemma shows that the Kalman gain K_k tends to stabilize to an equilibrium value. which will be exploited in the performance analysis.

At this point, we can compare the OSA procedure and the estimation of U_k based on LLMSE and KF now introduced. In the Table 5.1, we have rewritten the OSA in the case of input in $\{-1, 1\}$ and at the corresponding update using the Kalman Filter. It is interesting to notice that the decision on the transmitted U_k is performed in the same way, but the estimation of the state X_k is different: in the OSA, we compare the Euclidean distances and we force the estimate $\widehat{x}_k^{\text{OSA}}$ to lie in the set where X_k lies; using the KF method, the state is updated using the Kalman gain K_k and the state estimate can be any real number.

From the point of view of the complexity, both procedures perform a few operations at each step and require to store only a state value at each step.

The performance are now studied.

5.6.2 Theoretical Analysis of the Kalman Filter Method through IRF

The aim of this section is to evaluate the performance of the KF method in terms of BER:

$$\mathrm{BER(KF)} = \frac{1}{K}\sum_{k=0}^{K-1} \mathrm{P}(U_k \neq \widehat{U}_k)$$

In particular, in our next theorem we will analytically determine

$$\lim_{K\to\infty} \mathrm{BER}(KF)$$

using the Iterated Random Functions; afterwards, we will compare the KF and OSA performance in terms of BER.

Let us introduce the IRF setting for the KF method.

As in the OSA case, the evaluation of the BER can be done studying the Markov Process

$$D_k = X_k^{\mathrm{KF}} - X_k, \quad k \in \mathbb{N} \tag{5.61}$$

on $\mathbb{R}$ with initial state $D_0 = 0$. As it will be clear at the end of the section the asymptotic BER can be easily obtained if $(D_k)_{k\in\mathbb{N}}$ is *strongly ergodic*, that is, if asymptotically the states are distributed according to an invariant probability measure. Notice that

$$\begin{aligned} D_k &= \mathrm{q}X_{k-1}^{\mathrm{KF}} + K_k(Y_k - \mathrm{q}cX_{k-1}^{\mathrm{KF}}) - X_k \\ &= \mathrm{q}X_{k-1}^{\mathrm{KF}} + K_k(c - \mathrm{q}X_{k-1} + c\mathrm{w}U_{k-1} + N_k - \mathrm{q}cX_{k-1}^{\mathrm{KF}}) - \mathrm{q}X_{k-1} - \mathrm{w}U_{k-1} \\ &= \mathrm{q}(1-cK_k)D_{k-1} - \mathrm{w}(1-cK_k)U_{k-1} + K_kN_k. \end{aligned} \tag{5.62}$$

and in particular let us remark that $(D_k)_{k\in\mathbb{N}}$ is time-nonhomogeneous.

The evolution of the Markov Process $(D_k)_{k\in\mathbb{N}}$ and its ergodic properties can be obtained by the Iterated Random Functions. In fact, if we define the IRF

$$w_{I_k}(x) = \mathrm{q}(1-cK_k)x - \mathrm{w}(1-cK_k)U_{k-1} + K_kN_k, \quad x \in \mathbb{R} \tag{5.63}$$

where $I_k = (U_{k-1}, N_k)$ is a random process on $\{-1,1\}\times\mathbb{R}$ and the I_k are mutually independent, we can exploit Stenflo Theorem 9 in order to state that that if $k\to\infty$, then the transition probability $P^k(x,\cdot)$ of $(D_k)_{k\in\mathbb{N}}$ tends (with respect to the Wasserstein distance) to a probability measure $\mu(\cdot)$, independent on the initial state x, .

Before keeping on the analysis, we have to observe that (5.63) defines a time-nonhomogeneous IRF, with a deterministic term K_k varying in time: this prevents the immediate application of the Stenflo Theorem. However, we will apply it to the *limit* IRF and then deduce our main result as explained as follows.

Let us consider $K = \lim_{k\to\infty} K_k$ and define the *limit* IRF

$$\widetilde{w}_{I_k}(x) = \mathrm{q}(1-cK)x - \mathrm{w}(1-cK)U_{k-1} + KN_k, \quad x \in \mathbb{R}. \tag{5.64}$$

This time-homogeneous IRF fulfills the hypothesis of Stenflo Theorem 9, since it is always contractive:

$$|\widetilde{w}_{I_0}(x) - \widetilde{w}_{I_0}(y)| = \mathrm{q}(1-cK)|x-y| \tag{5.65}$$

where $\mathrm{q}(1-cK) < \mathrm{q}(1-cK_1) = \frac{\mathrm{q}}{1+\mathrm{SNR}} < 1$ and given any $x \in \ R$

$$\begin{aligned}\mathbb{E}[|\widetilde{w}_{I_0}(x) - x|] &\le (1-\mathrm{q}(1-cK))|x| + \mathrm{w}(1-cK)\mathbb{E}[|U_0|] + K\mathbb{E}[|N_1|] \\ &= (1-\mathrm{q}(1-cK))|x| + \mathrm{w}(1-cK) + K\sqrt{\frac{2\sigma^2}{\pi}} < +\infty.\end{aligned} \tag{5.66}$$

Then, if $P_l(\cdot,\cdot)$ is the transition probability of the limit process, there exists a probability measure $\widetilde{\mu}$ such that

$$P_l^n(x,\cdot) \xrightarrow{d_W} \widetilde{\mu} \text{ for } n \to \infty, \forall x \in \mathbb{R}$$

and $\widetilde{Z}_n(x) = \widetilde{w}_{I_{n-1}} \circ \widetilde{w}_{I_{n-2}} \circ \cdots \circ \widetilde{w}_{I_0}(x)$ tends a.s to a random variable $\widetilde{Z}$ distributed according to $\widetilde{\mu}$. More precisely,

$$d_W\left(P_l^n(x,\cdot),\widetilde{\mu}\right) \le \frac{[\mathrm{q}(1-cK)]^n}{1-\mathrm{q}(1-cK)}\mathbb{E}(d(\widetilde{w}_{I_0}(x),x)).$$

At this point, we can prove the following result about our original time-nonhomogeneous process

Theorem 10

$$\lim_{K\to\infty} \mathrm{BER}(KF) = \int g \ \mathrm{d}\widetilde{\mu} \tag{5.67}$$

where $\widetilde{\mu}$ *is the limit probability measure of the limit IRF* (5.64) *and* $g(z) = \mathrm{P}(\widehat{U}_k \neq U_k | D_k = z)$.

In order to prove this theorem, we need the following Lemma. If $Z_n(x) = w_{I_{n-1}} \circ w_{I_{n-2}} \circ \cdots \circ w_{I_0}(x)$, then

Lemma 13

$$d\left(Z_n(x), \widetilde{Z}\right) \to 0 \textit{ in probability for any } x \in \mathbb{R}. \tag{5.68}$$

Proof We have that

$$\begin{aligned} Z_n(x) &= \left(\mathrm{q}^n \prod_{i=0}^{n-1} l_i\right)x + \mathrm{w}\sum_{i=0}^{n-1} U_{n-1-i}\left(\mathrm{q}^i \prod_{j=n-1-i}^{n-1} l_j\right) + \sum_{i=0}^{n-1} K_{n-1}N_{n-1}\left(\mathrm{q}^i \prod_{j=n-i}^{n-1} l_j\right) \\ \widetilde{Z} &= \lim_{n\to\infty} \mathrm{w}l \sum_{i=0}^{n-1} U_{n-1-i}(\mathrm{q}l)^i + K\sum_{i=0}^{n-1} N_{n-1}(\mathrm{q}l)^i \end{aligned} \tag{5.69}$$

where $K_k \nearrow K$ and $l_k = 1 - cK_k \searrow l$. Now,

$$\begin{aligned}&\lim_{n\to\infty} \mathbb{E}[|Z_n(x) - \widetilde{Z}|]\\ &\leq \lim_{n\to\infty} \mathrm{w} \sum_{i=0}^{n-1} \mathbb{E}[|U_{n-1-i}|]\mathrm{q}^i \left(\prod_{j=n-1-i}^{n-1} l_j - l^{i+1} \right) + \sum_{i=0}^{n-1} \mathrm{q}^i \mathbb{E}[|N_{n-1}|] \left| K_{n-1} \prod_{j=n-i}^{n-1} l_j - Kl^i \right|\end{aligned} \tag{5.70}$$

Let us prove that the last expression tends to zero. First, notice that $\mathbb{E}[|U_n|]$ and $\mathbb{E}[|N_n|]$ are bounded (and constant for any $n \in \mathbb{N}$). Recalling that that $l_j \searrow l$ and $l_j \in (0,1)$ for any j, let us fix a small $\varepsilon > 0$ and let $i_\varepsilon \in \mathbb{N}$ be such that for any $i \geq i_\varepsilon$, $q^i < \varepsilon$. Then,

$$\sum_{i=0}^{n-1} \mathrm{q}^i \left(\prod_{j=n-1-i}^{n-1} l_j - l^{i+1} \right) = \sum_{i=0}^{i_\varepsilon - 1} \mathrm{q}^i \left(\prod_{j=n-1-i}^{n-1} l_j - l^{i+1} \right) + \sum_{i=i_\varepsilon}^{n-1} \mathrm{q}^i \left(\prod_{j=n-1-i}^{n-1} l_j - l^{i+1} \right)$$

and

$$\begin{aligned}\lim_{n\to\infty} \sum_{i=0}^{i_\varepsilon - 1} \mathrm{q}^i \left(\prod_{j=n-1-i}^{n-1} l_j - l^{i+1} \right) &\leq \lim_{n\to\infty} \sum_{i=0}^{i_\varepsilon - 1} \mathrm{q}^i \left(l_{n-1-i}^{i+1} - l^{i+1} \right)\\ &= \sum_{i=0}^{i_\varepsilon - 1} \mathrm{q}^i \lim_{n\to\infty} \left(l_{n-1-i}^{i+1} - l^{i+1} \right) = 0\end{aligned}$$

as i_ε is finite, while

$$\begin{aligned}\lim_{n\to\infty} \sum_{i=i_\varepsilon}^{n-1} \mathrm{q}^i \left(\prod_{j=n-1-i}^{n-1} l_j - l^{i+1} \right) &\leq \lim_{n\to\infty} \sum_{i=i_\varepsilon}^{n-1} \mathrm{q}^i\\ &= \lim_{n\to\infty} \frac{\mathrm{q}^{i_\varepsilon} - \mathrm{q}^n}{1 - \mathrm{q}} = \frac{\mathrm{q}^{i_\varepsilon}}{1 - \mathrm{q}} < \frac{\varepsilon}{1 - \mathrm{q}}.\end{aligned}$$

Since ε can be chosen arbitrarily small, this limit is zero and we can conclude that

$$\lim_{n\to\infty} \mathrm{w} \sum_{i=0}^{n-1} \mathbb{E}[|U_{n-1-i}|]\mathrm{q}^i \left(\prod_{j=n-1-i}^{n-1} l_j - l^{i+1} \right) = 0.$$

Analogously,

$$\sum_{i=0}^{n-1} \mathrm{q}^i \left| K_{n-1} \prod_{j=n-i}^{n-1} l_j - Kl^i \right| = \sum_{i=0}^{i_\varepsilon - 1} \mathrm{q}^i \left| K_{n-1} \prod_{j=n-i}^{n-1} l_j - Kl^i \right| + \sum_{i=i_\varepsilon}^{n-1} \mathrm{q}^i \left| K_{n-1} \prod_{j=n-i}^{n-1} l_j - Kl^i \right|$$

and

$$\lim_{n\to\infty} \sum_{i=0}^{i_\varepsilon - 1} \mathrm{q}^i \left| K_{n-1} \prod_{j=n-i}^{n-1} l_j - Kl^i \right| = \sum_{i=0}^{i_\varepsilon - 1} \mathrm{q}^i \lim_{n\to\infty} \left| K_{n-1} \prod_{j=n-i}^{n-1} l_j - Kl^i \right| = 0$$

as i_ε is finite, while

$$\begin{aligned}\lim_{n\to\infty}\sum_{i=i_\varepsilon}^{n-1}\mathrm{q}^i\left|K_{n-1}\prod_{j=n-i}^{n-1}l_j - Kl^i\right| &\leq \lim_{n\to\infty}\sum_{i=i_\varepsilon}^{n-1}\mathrm{q}^i\frac{1}{c}\\ &= \frac{1}{c}\lim_{n\to\infty}\frac{\mathrm{q}^{i_\varepsilon}-\mathrm{q}^n}{1-\mathrm{q}} = \frac{1}{c}\frac{\mathrm{q}^{i_\varepsilon}}{1-\mathrm{q}} < \frac{1}{c}\frac{\varepsilon}{1-\mathrm{q}}.\end{aligned}$$

Since ε can be chosen arbitrarily small,

$$\lim_{n\to\infty}\sum_{i=0}^{n-1}\mathbb{E}[|N_{n-1}|]\mathrm{q}^i\left|K_{n-1}\prod_{j=n-i}^{n-1}l_j - Kl^i\right| = 0.$$

In conclusion, $\mathbb{E}[|Z_n(x)-\widetilde{Z}|]\overset{n\to\infty}{\longrightarrow}0$, hence the convergence in probability is proved by the Chebichev inequality:

$$\mathbb{P}[|Z_n(x)-\widetilde{Z}|<\varepsilon] \leq \frac{\mathbb{E}[|Z_n(x)-\widetilde{Z}|]}{\varepsilon}\overset{n\to\infty}{\longrightarrow}0. \tag{5.71}$$

■

Proof of Theorem 10.
The convergence in probability stated by the previous Lemma implies the convergence in distribution, that is if μ_n^x is measure probability of $Z_n(x)$, then $\int f\mu_n^x \to \int f\widetilde{\mu}$ for any $x \in R$ where $\mu f = \int f \mathrm{d}\mu$ and f is any continuous and bounded function in $\mathbb{R}$.

This is sufficient to prove the convergence of the BER. In fact,

$$\begin{aligned}g(z) &= \mathrm{P}(\mathrm{sgn}(N_{k+1}+c\mathrm{w}U_k+c\mathrm{q}z)=1|U_k=-1)\mathrm{P}(U_k=-1)+\\ &\quad+\mathrm{P}(\mathrm{sgn}(N_{k+1}+c\mathrm{w}U_k+c\mathrm{q}z)=-1|U_k=1)\mathrm{P}(U_k=1)\\ &=\frac{1}{4}\mathrm{erfc}\left(\frac{c\mathrm{w}-c\mathrm{q}z}{\sigma\sqrt{2}}\right)+\frac{1}{4}\mathrm{erfc}\left(\frac{c\mathrm{w}+c\mathrm{q}z}{\sigma\sqrt{2}}\right)\end{aligned} \tag{5.72}$$

and $P^k g(x) = \int g\ \mathrm{d}\mu_k^x \to \int g\ \mathrm{d}\widetilde{\mu}$ for any x, then

$$\lim_{K\to\infty}\mathrm{BER}(KF) = \lim_{K\to\infty}\frac{1}{K}\sum_{k=0}^{K-1}P^k g(0) = \int g\ \mathrm{d}\widetilde{\mu}.$$

■

5.6.3 Design Criteria

In the previous sections, we have analytically evaluated the asymptotic MSE for both the OSA and the KF based method, using the IRF theory.

The conclusion we achieve is that neither algorithm is definitely better than the other one, the performance depending on the system parameters and noise conditions.

It is not intuitive to understand for which values of a, b, c and σ^2 the OSA is better than the KF and vice-versa, but given these parameters, one can compute the corresponding i.p.m. and BER and decide which method is preferable. In the next we illustrate these observations presenting some instances (see Figure 5.6), where we fix $b = c = 1$ and vary a and the noise variance. From the graphs we deduce that a stronger stability (i.e., smaller a) leads to more similar performance. This can be explained noting that a smaller a corresponds to a smaller q, then to the reduction of the weight assigned to the state estimate in both algorithms (see Table 5.1). In particular, in the limit case $a \to -\infty$, OSA and KF coincide.

On the other hand, when a is close to zero, the OSA performs better than the KF method for high values of SNR. This can be motivated by the OSA "hard" decision on the states (that is, the state estimates are forced to belong to the state space): in fact, for high SNR, the hard procedure gives the correct state with higher probability, while the KF will give a good (real) estimate, but in general not the true value.

A final consideration is about the complexity: both algorithms just store one state value at each step and perform very easy operations (linear operation for KF and a sign computation for OSA) to update the state. On the other hand, the KF method has an additional operation, that is the computation of the Kalman gain K_k; moreover, the KF is less straightforward to implement (and is not optimal) when the unknown input has not the characteristics of a white noise, while the OSA scheme is easily adapted to different source distributions.

5.7 Conclusions

In this chapter, we have studied the deconvolution of one-dimensional, linear systems in case of binary input generated by a Bernoullian source. We have exploited the approach used for the differentiation problem, even if the mathematical setting turns out to be different: in particular, the state space is no more $\mathbb{N}$, but a compact interval. This prevents the implementation of the proposed algorithm except for the OSA, for complexity issues (the state space is not denumerable). We have then implemented OSA, simulated the system and provided a theoretical analysis of the performance using the Iterated Random Functions, computing the asymptotic BER for *sufficiently* stable systems. We have also shown that analogous results can be achieved using the Markov Processes' theory and the Fixed Point Theorem.

Afterwards, we have developed a Kalman Filter based method to recover the input and analyzed it with the IRF. Finally, we have compared the KF with the OSA and concluded that neither is definitely better than the other: their performance depend on the particular instance, say on the parameters of the system and the measurement noise. Given the knowldege of parameters and noise, one can analytically choose the more suitable algorithm, which is very useful from the design viewpoint.

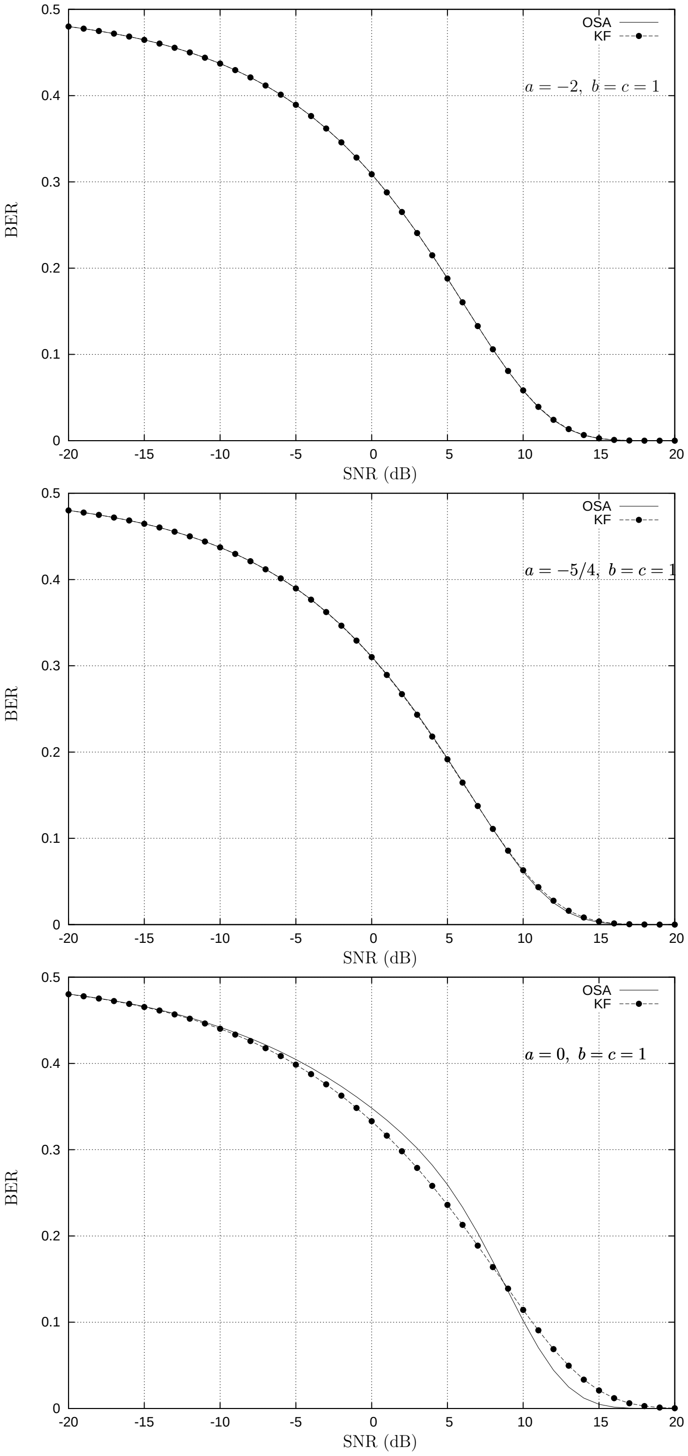

Figure 5.6: OSA vd KF: Bit Error Rate in function of the SNR for $b = c = 1$ and $a = -2, -5/4, 0$

Chapter 6

A Fault Tolerant Control Problem

Fault Tolerant Control (FTC for short, see [24], [71], [41]) aims to cancel or contain the consequences of faults in an automation system. Such an operation is fundamental in modern technological processes, which are required to assure robust performance, stability and safety even in case of partial malfunctions or degradations. Often, robustness is achieved by redundancy, say by the introduction of many control components like sensors; nevertheless, this sophistication naturally increases the probability of breakdown and then continues to motivate the research on reliable control systems.

The problem of upholding the functionality of an apparatus affected by a disturbance is ubiquitous in the industrial and transport fields. In particular, FTC systems are widely applied in those contexts where human health and environment are concerned, for example, in the design of mechanical and chemical plants; nuclear power reactors; medical systems; aircrafts, helicopters and spacecrafts; automotive engines, railway and marine vehicles. Another application is in the communication networks (for instance, wireless sensor networks), where the aim of FTC is to avoid unexpected interruptions of data flow in case of troubled connectivity or impaired nodes. In general, a satisfying FTC design prevents total failures and stops, with the ultimate objective of reducing health, environmental and economic damages.

The literature about FTC is definitely widespread and contributions arise from diverse applied mathematical domains. Several survey works report the main theoretical concepts and provide classifications of the outstanding FTC approaches, with detailed references; for example, we refer the reader to the recent review by [167], which supplies a comprehensive bibliography, and to [76], [145], [110], [162].

As far as the applications are concerned, flight control has been motivating FTC research since 1970s, given the evident danger that aircraft faults may cause to human safety. A significant amount of papers have been produced on the argument, pertaining to the many different aspects that characterize a safe flight dynamics. For a general overview, see [142], [46], and the up-to-date book by [41] which in Chapter II provides the list of the most common flight control systems, with the relative references.

In this work, we consider a linear model with a multiplicative disturbance factor, which is common in flight control ([165]); in particular, we adopt the system presented by [3, 2] and studied also by [166] and [47] as an application test.

Even if FTC systems can be designed in many different ways according to the specific aim they are conceived for, in general they all have to perform the following main tasks:

1. the fault detection: the controller makes a binary decision on the presence of a malfunction;

2. the fault identification: the controller determines or estimates the size of the disturbance; when necessary, identification is preceded by the fault isolation, that is, the location of the impaired component;

3. an active compensation to the fault, i.e., the reconfiguration of the system inputs and/or parameters in order to maintain, as much as possible, the integrity of the process.

Fault Detection and Identification (FDI) can be undertaken in diverse ways. In the cited works, in particular [41] a comprehensive discussion about the most popular FDI schemes is presented: among them, we recall the unknown input observers (UIO, [111], [156]) and residual generation; the Kalman filtering; the statistical methods and the more recent techniques based on neural networks ([98]).

In this chapter, we present a novel approach to FDI. Our setting is a continuous, time-invariant, linear system in which a quantized disturbance input is introduced. Such a hybrid model, that combines discrete and continuous dynamics, is motivated by the upcoming digitalization of modern devices: a quantized disturbance may represent the switches of actuators and sensors or a malfunction in a digital component; moreover, it may describe the behavior of any mechanical device that is known to occupy only certain positions and also it may be the approximation of a continuous disturbance.

Results about FTC for hybrid systems are not very common; in part, they can be retrieved in the extensive discussion about the detection of abrupt changes in dynamical systems, whose leading work is by [14] (while some further contributions are given by [86] and [105]). The problem of estimating brusque alterations is always actual (see, e.g., [155] and [147], which respectively concern medical imaging and ground-penetrating radar issues) and in general is approached by classical estimation techniques, such as Kalman filtering.

Recently, input quantization in linear systems has been studied with the aim of reducing the effects of a coarse quantization ([109], [44]). In this work, instead, our purpose is to detect faults using the information that the disturbance input is quantized, assuming the quantization to be sufficiently accurate.

In order to detect and evaluate a quantized input disturbance, we propose an Information theoretic approach: given the discrete nature of the disturbance, we suggest to perform FDI using the One State Algorithm introduced in Chapter 3.

The structure of the chapter is the following: in Section 6.1, we describe the problem; in Section 6.2, we introduce the decoding algorithm we intend to use for the fault detection; in Section 6.3, we provide a theoretical analysis aimed at deriving optimal design criteria (Section 6.4). The analysis includes considerations about the sensitivity of system to the false alarm (*false positive*) and to miss detection (*false negative*), promptness of detection and reconfiguration. In Section 6.5 we show a few significant simulations about a specific numerical example, arisen from flight control literature; finally, in Sections 6.6 and 6.7 we propose some considerations about possible quantization errors and a few concluding observations.

6.1 Problem Statement

Let us consider processes that can be modeled by the following finite-dimensional, input/output linear system:

$$\begin{cases} \dot{x}(t) = \mathrm{A}x(t) + \mathrm{B}z(t)f(t) \quad t \in [0,T] \\ x(0) = 0 \\ y(t) = \mathrm{C}x(t) \end{cases} \tag{6.1}$$

where $x(t) \in \mathbb{R}^n$, $y(t) \in \mathbb{R}^m$, $f(t)$ and $z(t)$ are scalar functions and A, B and C are constant matrices with consistent dimensions. The output $y(t)$ is supposed to be noisy observable; $f(t)$ is a known input signal, while $z(t)$ is a disturbance modeling some fault in the system. Typically, $z(t) \in (0,1]$; if $z(t) = 1$, the system operates in its nominal (i.e., fault-free) regime and is totally driven by $f(t)$: this is the condition that one aims at reproducing when $z(t) \in (0,1)$, i.e., when some unexpected breakdown, interruption or loss of effectiveness affects the dynamics. In order to achieve that, a control input $u(t)$ is introduced, which adjusts the dynamics as follows:

$$\begin{cases} \dot{x}(t) = \mathrm{A}x(t) + \mathrm{B}z(t)\big(f(t) + u(t)\big) \quad t \in [0,T] \\ x(0) = 0 \\ y(t) = \mathrm{C}x(t) \end{cases} \tag{6.2}$$

Notice that to maintain the error-free behavior, say $\mathrm{B}z(t)\big(f(t) + u(t)\big) = \mathrm{B}f(t)$, in principle it is sufficient to fix $u(t) = f(t)\left(\frac{1}{z(t)} - 1\right)$, but, in the real applications, this is actually impossible for the following motivations. Generally, the disturbance z is not known and the controller can access it only through the (noisy) observation of the output y. In order to determine z one has to perform a deconvolution, that is, to invert the solution of system (6.2), given by:

$$y(t) = \mathrm{C}x(t) = \mathrm{C}\int_0^t e^{(t-s)\mathrm{A}}\mathrm{B}z(s)(f(s) + u(s))\mathrm{d}s. \tag{6.3}$$

Under this condition, the inversion of expression (6.3) becomes tricky: deconvolution is in fact known to be an ill-posed and ill-conditioned problem, that is, the uniqueness of solution is not guaranteed and also small errors in the data may raise large errors in the solution. In conclusion, the reconstruction of $z(t)$ by inversion may produce outcomes very far from the correct ones; for this reason, an estimation approach to the problem is the most suitable one.

In the next, Assumptions 1, 2 and 3 of Section 3.1 are supposed to hold, that is:

- the available output signal is a sampled, noisy version of $y(t)$:
  $$y_k = y(k\tau) + n_k = \mathrm{C}x(k\tau) + n_k \quad k = 0, 1, \ldots, K$$
  where $\tau > 0$ is a constant sampling time and n_k is an additive observational noise;
- $K = T/\tau \in \mathbb{N}$;
- the n_k's are realizations of independent, identically distributed Gaussian random variables N_k's of 0 mean and covariance matrix $\sigma^2\mathbb{I}$.

Moreover,

Assumption 9 *The disturbance function $z(t)$ is known to be quantized over two levels, in particular $z(t)$ can assume only the values $\zeta_0 = 1$ and $\zeta_1 \in (0, 1)$.*

In general, no prior stochastic information is available on the behavior of $z(t)$.

ζ_0 and ζ_1 may respectively represent the fault-free and the faulty conditions. Such a binary situation occurs in engineering issues such as the abrupt blocking of an actuator, the sudden disconnection of a component, the use of alarm sensors and more in general in the presence of any device that may switch on or off. In the next, we will refer to the jumps from ζ_0 and ζ_1 and vice-versa as *switch points*.

Under Assumption 9, Fault Detection and Identification are coincident: the decision on the fault presence automatically determines also its size. We have to remark that this is not the typical FTC paradigm, as in most applications the “faulty” value ζ_1 is not known and moreover in many cases it is continuous (for example, it belongs to an interval of real numbers). In this more common framework, the unknown value must be identified and quantization can be assumed only if it is sufficiently coarse to ensure a reliable Identification.

Here, we however consider a binary, perfect quantization for the following reasons. First, the use of digital devices, which work within finite sets of values, is nowadays widespread and increasing also in control systems. The state of a digital device typically can assume a finite number of possible levels and a fault of it may be represented by an undesired switch. In this context, the signals are not continuous and quantization is naturally implied by the problem itself. On the other hand, the choice of a *binary* quantization, which in many cases is too restrictive, has mainly an introductory purpose: since our approach to FTC is novel and arising from a different research field, it is preferable to consider the simplest scenario. More complicated (and more realistic) cases will be addressed in future work.

Coming back to our model, our aim is to estimate $z(t)$ as well as possible in order to provide the best feedback compensation to the system. Clearly, the estimation has to be performed on-line, that is, each time a sample is acquired (notice that the sampling inevitably undertakes some delay): each τ instants the controller tries to detect possible faults and consequently updates the system design.

For mathematical simplicity, the switch points of $z(t)$ are supposed to occur at the time instants $k\tau$, in order to have synchronization with the output sampling. Hence, we can write:

$$z(t) = \sum_{k=0}^{K-1} z_k \mathbb{1}_{[k\tau,(k+1)\tau)}(t) \quad z_k \in \{\zeta_0, \zeta_1\} \tag{6.4}$$

This forced synchronization is acceptable since it will not affect the performance of our algorithm in a sensible way. In particular, it will just cause some negligible difference in its delay (see Section 6.2).

Now, $z(t)$ is equivalent to the binary sequence $(z_0, \dots, z_{K-1}) \in \{\zeta_0, \zeta_1\}^K$: the estimation problem is actually discrete. Let $\hat{z}_k$ be an estimate of z_k: since the operation must be performed on-line, we expect $\hat{z}_{k-1} = \mathcal{D}(r_1, \dots, r_k)$, where $\mathcal{D}$ indicates a detection/estimation function.

Taking account of the conditions mentioned before, the natural definition of the control input is:

$$u(t) = f(t) \sum_{k=0}^{K-1} \left(\frac{1}{\hat{z}_{k-1}} - 1 \right) \mathbb{1}_{[k\tau,(k+1)\tau)}(t) \tag{6.5}$$

As the first measurement is performed at time τ, the initial value $\hat{z}_{-1}$ is arbitrarily fixed; for example it has sense to initialize it with $\hat{z}_{-1} = 1$, which means that no compensation is introduced into the system in $[0, \tau)$.

Let us now plug $z(t)$ and $u(t)$ given by (6.4) and (6.5) into (6.3): if $t \in [k\tau, (k+1)\tau)$

for some $k \in \{0, \dots, K-1\}$,

$$\begin{aligned}
x(t) &= \int_0^t e^{(t-s)\mathrm{A}}\mathrm{B}z(s)(f(s)+u(s))\mathrm{d}s \\
&= \int_0^t e^{(t-s)\mathrm{A}}\mathrm{B}\sum_{j=0}^{K-1} z_j \mathbb{1}_{[j\tau,(j+1)\tau)}(s)\left(f(s)+f(s)\sum_{j=0}^{K-1}\left(\frac{1}{\hat{z}_{j-1}}-1\right)\mathbb{1}_{[j\tau,(j+1)\tau)}(s)\right)\mathrm{d}s \\
&= \int_0^{k\tau} e^{(t-s)\mathrm{A}}\mathrm{B}\sum_{j=0}^{K-1} z_j \mathbb{1}_{[j\tau,(j+1)\tau)}(s)\left(f(s)+f(s)\sum_{j=0}^{K-1}\left(\frac{1}{\hat{z}_{j-1}}-1\right)\mathbb{1}_{[j\tau,(j+1)\tau)}(s)\right)\mathrm{d}s \\
&+ \int_{k\tau}^t e^{(t-s)\mathrm{A}}\mathrm{B}\sum_{j=0}^{K-1} z_j \mathbb{1}_{[j\tau,(j+1)\tau)}(s)\left(f(s)+f(s)\sum_{j=0}^{K-1}\left(\frac{1}{\hat{z}_{j-1}}-1\right)\mathbb{1}_{[j\tau,(j+1)\tau)}(s)\right)\mathrm{d}s \\
&= \sum_{j=0}^{k-1}\frac{z_j}{\hat{z}_{j-1}}\int_{j\tau}^{(j+1)\tau} e^{(t-s)\mathrm{A}}\mathrm{B}f(s)\mathrm{d}s + \frac{z_k}{\hat{z}_{k-1}}\int_{j\tau}^{t} e^{(t-s)\mathrm{A}}\mathrm{B}f(s)\mathrm{d}s \\
&= e^{t-k\tau}\sum_{j=0}^{k-1}\frac{z_j}{\hat{z}_{j-1}}\int_{j\tau}^{(j+1)\tau} e^{(k\tau-s)\mathrm{A}}\mathrm{B}f(s)\mathrm{d}s + \frac{z_k}{\hat{z}_{k-1}}\int_{j\tau}^{t} e^{(t-s)\mathrm{A}}\mathrm{B}f(s)\mathrm{d}s \\
&= e^{(t-k\tau)\mathrm{A}}x_k + \frac{z_k}{\hat{z}_{k-1}}\mathrm{M}_{t,k\tau}, \qquad t \in [k\tau,(k+1)\tau)
\end{aligned}$$

where

$$x_k = x(k\tau)$$

and

$$\mathrm{M}_{a,b} := \int_b^a e^{(a-s)\mathrm{A}}\mathrm{B}f(s)\mathrm{d}s \qquad a,b \in \mathbb{R}. \tag{6.6}$$

At this point, the evolution of system (6.2) with (6.4) and (6.5) can be written recursively as

$$\begin{cases}
x_0 = 0 \\
\hat{z}_{-1} = 1 \\
x(t) = e^{(t-k\tau)\mathrm{A}}x_k + \frac{z_k}{\hat{z}_{k-1}}\mathrm{M}_{t,k\tau} \qquad k = 0,\dots,K-1,\ t \in [k\tau,(k+1)\tau) \\
y(t) = \mathrm{C}x(t)
\end{cases} \tag{6.7}$$

In (6.7), we have not specified yet how we intend to determine the estimates $\hat{z}_k$: the detection algorithm will be introduced in Section 6.2.

Notice that $u(t)$ is computed and introduced in the system each τ time instants. Given a generic interval $[k\tau,(k+1)\tau)$, $u(t)$ is deceptive when a switch occurs at $k\tau$, as it is based on the estimate $\hat{z}_{k-1}$ relative to the previous interval; thus, the delay τ underlies a temporary, unavoidable deviation (even in case of correct detection) from the right trajectory. This issue will be widely discussed in the next; for the moment, let us just observe that switch points cause the most of the problems in our FTC model. For this reason, *permanent* interruptions, i.e., *failures* (which involve just one switch point) are definitely preferable than *transient* faults for our purpose, though this should appear as a paradox in the practice.

6.1.1 Illustrative Example: a Flight Control Problem

Before discussing our FDI algorithm, let us notice that systems of kind (6.1) are common in flight control literature to model different aspects of the aerospace dynamics. A typical example is the following: if we consider the matrices

$$\mathrm{A} = \begin{bmatrix} -0.5162 & 26.96 & 178.9 \\ -0.6896 & -1.225 & -30.38 \\ 0 & 0 & -14 \end{bmatrix} \qquad \mathrm{B} = \begin{bmatrix} -175.6 \\ 0 \\ 14 \end{bmatrix} \qquad \mathrm{C} = [1 \quad 12.43 \quad 0]$$

the system (6.1) represents the longitudinal short-period mode of an F4-E jet with additional horizontal canards, in supersonic conditions. The vector $x(t)$ determines the longitudinal trajectory: its three entries respectively represent the normal acceleration, the pitch rate and the deviation of elevator deflection from the trim position. The output $y(t)$ is the C^* response, a parameter that synthesizes the aircraft response to the pilot inputs; typically, the C^* response must lie in a given admissible flight envelope. This application example is illustrated in the Appendix D.1 of the book by [3] and studied also in [2], [166], [47].

Referring to this example, $f(t)$ can be interpreted as the elevator deflection command and $z(t)$ as the indicator of the status of the elevators: $z = \zeta_0$ may attest a good status, while the switch to $z = \zeta_1$ may denote an abrupt loss of effectiveness. In such a case, the controller is required to detect the accident and introduce a suitable control input $u(t)$ in order to recover the optimal trajectory, say the one imposed by the flight plan. In terms of the output $y(t)$, one aims at maintaining or at bringing it it back within the prescribed envelope.

In this context, it makes sense to suppose that the elevator cannot recover its efficiency during the flight: this is a case of failure, which will be our case study in the next.

This Flight Control Problem will be retrieved later and used as test application for the implementation of our detection algorithm, which is introduced in the next section.

6.2 Fault Detection: One State Algorithm

Given the quantization of $z_k \in \{\zeta_0, \zeta_1\}$, it makes sense to settle the same set for the estimation: $\hat{z}_k \in \{\zeta_0, \zeta_1\}$ and to use one of the decoding algorithms proposed in Chapter 3 to identify the disturbance. In particular, since the considered process, in principle, may never end, we have chosen to implement the causal, low-complexity One State Algorithm (OSA).

The key idea of the One State procedure is to recursively provide an estimate $\hat{x}_k$ of the state $x_k = x(k\tau)$ and of z_{k-1} given the current lecture y_k and the estimate $\hat{x}_{k-1}$ of the previous state x_{k-1}.

OSA - Decoder $\mathbb{D}^{(1)}$

Initialization: $\hat{x}_0 = 0$, $\hat{z}_{-1} = 1$.

For $k = 1, \ldots, K$:
System evolution: $x_k = e^{\tau \mathrm{A}} x_{k-1} + \frac{z_{k-1}}{\hat{z}_{k-2}} \mathrm{M}_{k\tau,(k-1)\tau}$.
Measurement: $y_k = \mathrm{C} x_k + n_k$.

$$\text{Disturbance Estimation: } \hat{z}_{k-1} = \begin{cases} \zeta_0 & \text{if } ||y_k, \mathrm{C}e^{\tau \mathrm{A}} \hat{x}_{k-1} + \frac{\zeta_0}{\hat{z}_{k-2}} \mathrm{CM}_{k\tau,(k-1)\tau}||_{\mathbb{R}^m} \\ & \leq ||y_k, \mathrm{C}e^{\tau \mathrm{A}} \hat{x}_{k-1} + \frac{\zeta_1}{\hat{z}_{k-2}} \mathrm{CM}_{k\tau,(k-1)\tau}||_{\mathbb{R}^m} \\ \zeta_1 & \text{otherwise} \end{cases}$$

State Estimation: $\hat{x}_k = e^{\tau \mathrm{A}} \hat{x}_{k-1} + \frac{\hat{z}_{k-1}}{\hat{z}_{k-2}} \mathrm{M}_{k\tau,(k-1)\tau}$.

We observe that the OSA neglects the evolution of the system except for the time instants $k\tau$, $k \in \mathbb{N}$, that is, it considers a discretized version.

As already noticed, the system does not have compensation in $[0, \tau)$. For the binary nature of each z_k, the process of estimation/detection reduces here to the comparison of two distances. Moreover, the storage required is of two locations (one float for the current state and one boolean for the current disturbance): the algorithm is definitely low-complexity.

Remark 7 *The One State algorithm can be easily extended to the case of quantization with $q > 2$ levels, by comparing the distances between the received symbol and the q possible signals; nevertheless, the theoretical analysis of the performance, as we propose it in the next Section, would be definitely more complicated.*

6.3 Theoretical Analysis of the One State Algorithm

In this section, we provide a theoretical analysis of the performance of the One State Algorithm with the final aim to determine optimal design criteria for our FTC system. We will focus on the system (6.7) with a failure, that is, in the presence of just one switch point $T_F = k_F \tau \in [0, T]$, $k_F \in \mathbb{N}$, such that

$$z(t) = \begin{cases} \zeta_0 = 1 & t \in [0, T_F) \\ \zeta_1 \in (0, 1) & t \in [T_F, T] \end{cases} \tag{6.8}$$

or equivalently, $z_k = \zeta_0$ for $k = 0, 1, \ldots, k_F - 1$ and $z_k = \zeta_1$ for $k = k_F, 1, \ldots, K - 1$.

Switch points are critical since they always cause deviations from the desired trajectory due to the detection delay. In fact, $u(t)$ is deceptive in $[k\tau, (k+1)\tau)$ when a switch occurs at $k\tau$, as it is function of $\hat{z}_{k-1}$, which in turn depends on the evolution of the system in $[0, k\tau)$.

Considering the case of one switch point is then oriented to isolate and understand this phenomenon, as well as motivated by the ubiquity of failure occurrences in the applications.

6.3.1 Probabilistic setting

Assuming the measurement noises n_k's to be realizations of independent, zero-mean, Gaussian random variables, a certain amount of uncertainty affects all the system (6.7); in particular also $\hat{z}_{k-1}$, $x(t)$ and $y(t)$ are random variables as they are directly or indirectly functions of the noise. The evolution of our FTC procedure in probabilistic terms is as follows (capital letters indicate random variables):

$$\begin{cases} X_0 = 0, \ \hat{X}_0 = 0, \ \hat{Z}_{-1} = \zeta_0 = 1 \\ X_k = e^{\tau \mathrm{A}} X_{k-1} + \frac{z_{k-1}}{\hat{Z}_{k-2}} \mathrm{M}_{k\tau} \qquad k = 1, \dots, K \\ Y_k = \mathrm{C} X_k + N_k \\ \hat{Z}_{k-1} = \mathbb{D}_1(Y_k, \hat{X}_{k-1}, \hat{Z}_{k-2}) \\ \hat{X}_k = e^{\tau \mathrm{A}} \hat{X}_{k-1} + \frac{\hat{Z}_{k-1}}{\hat{Z}_{k-2}} \mathrm{M}_{k\tau} \end{cases} \tag{6.9}$$

where $\mathbb{D}_1$ indicates the One State decoding/detection function. We recall that $z(t)$ is not supposed to be driven by any known stochastic rule.

6.3.2 Aim of the analysis

The performance of the One State Algorithm must be determined in terms of a suitable *distance* between the desired and the real output. Let us consider the Flight Control Problem (Section 6.1.1) as reference: the output $y(t)$, which summarizes the features of the aircraft's trajectory, must be maintained in a prescribed flight envelope. This can be interpreted in two ways: the distance between desired and real output must be (a) bounded in a certain range (b) minimized as much as possible, tolerating infrequent and temporary excursions outside the flight envelope. For our model, these conditions can be effectively formulated as follows:

1. choose τ so that the maximal amplitude $\mathbb{M}$ of the deviations is minimized;

2. choose τ so that the probability $\mathbb{P}$ that no deviations occur is maximized.

Since in a flight context both conditions may be important, we propose to merge the two criteria in this way: fixed a small tolerance ε, we define the optimal τ as

$$\tau_{\mathrm{opt}} = \operatorname*{argmin}_{\tau \in S_\varepsilon} \mathbb{M}, \qquad S_\varepsilon = \{\tau : \ \mathbb{P} > 1 - \varepsilon\}. \tag{6.10}$$

In other terms, we first individuate a possible set S_ε so that $\mathbb{P}$ is above a safety threshold $1 - \varepsilon$ and then we choose $\tau \in S_\varepsilon$ that minimizes the maximal amplitude. This trade-off strategy has been derived observing that, for our model, a larger τ corresponds to a larger $\mathbb{P}$, but also to larger deviations at switch points (even in case of correct detection), i.e., to a larger $\mathbb{M}$. In the next, we will specialize the criterion (6.10) to our setting, in terms of an error function (Sections 6.3.3 and 6.3.4); afterwards, we will discuss how to compute τ_{opt} in some specific cases (Section 6.4) and test it in the Flight Control Problem.

6.3.3 Error Function and Probability of n-step Error Decay

The Error Function we adopt to represent the distance between the desired and the real output is given by the discrete stochastic process $(E_k)_{k=0,1,\dots}$ that describes the difference between the state X_k and the nominal (i.e., fault-free) state $x^N(t) := \int_0^t e^{(t-s)\mathrm{A}}\mathrm{B}f(s)ds$, at time instants $k\tau$, $k = 0, 1, \dots, K$:

$$\begin{cases} E_0 = 0 \\ E_k = X_k - x^N(k\tau) \\ \quad\; = e^{\tau \mathrm{A}} E_{k-1} + \left(\frac{z_{k-1}}{\hat{Z}_{k-2}} - 1\right) \mathrm{M}_{k\tau} \quad k = 1, \dots, K. \end{cases} \tag{6.11}$$

Let us notice that

Lemma 14 *For any $k_0, n \in \mathbb{N}$, the events $\{E_{k_0+n} = e^{n\tau A}E_{k_0}\}$ and $\{\hat{Z}_{k-1} = z_k \text{ for all } k = k_0, k_0+1, \dots k_0+n\}$ coincide.*

Proof It immediately follows from the definition of E_k: for any $n \in \mathbb{N}$, the event $\{E_{k+1} = e^{\tau\mathrm{A}}E_k\}$ coincides to $\{\hat{Z}_{k-1} = z_k\}$ and then $\{E_{k_0+n} = e^{n\tau\mathrm{A}}E_{k_0}\}$ coincides to $\{\hat{Z}_{k_0-1} = z_{k_0}, \hat{Z}_{k_0} = z_{k_0+1}, \dots, \hat{Z}_{k_0+n-1} = z_{k_0+n}\}$. ■Notice that under the

hypothesis of the proposition and if A is asymptotically stable, E_k exponentially decays to zero, regardless of the initial value E_{k_0}. Moreover, observe that $\{\hat{Z}_{k-1} = z_k\}$ is not the event of correct detection $\{\hat{Z}_k = z_k\}$, since the feedback in the system implies a delay τ; however, if z_k is constant over the considered interval, the two events coincide.

Afterwards, let

$$\begin{cases} D_0 = 0 \\ D_k = \hat{X}_k - X_k = e^{\tau\mathrm{A}}D_{k-1} + \frac{\hat{Z}_{k-1} - z_{k-1}}{\hat{Z}_{k-2}}\,\mathrm{M}_{k\tau} \end{cases}$$

where the $\hat{X}_k$'s are the states estimated by the One State Algorithm. Given $k_0, n \in \mathbb{N}$, $k_0 \geq 1$, we define the probability of n-step error decay (EDP^n for short) as

$$\mathrm{EDP}^n(k_0, d, \zeta, \eta) =$$
$$\mathrm{P}\left(E_{k_0+n} = e^{n\tau\mathrm{A}}E_{k_0} \middle| D_{k_0-1} = d, \hat{Z}_{k_0-2} = \zeta, z_k = \eta \text{ for any } k = k_0 - 1, \dots, k_0 + n - 1)\right)$$

where $d \in \mathbb{R}^n$, $\zeta, \eta \in \{\zeta_0, \zeta_1\}$. Let us now reformulate the optimization problem (6.10) in a more precise way. For simplicity, from now onwards we will assume $y(t) \in \mathbb{R}$, as in the Flight Control Problem.

Before the failure, $\mathbb{P}$ corresponds to $\mathrm{EDP}^{T_F/\tau-1}(1,0,\zeta_0,\zeta_0)$, while after the failure it corresponds to $\mathrm{EDP}^{(T-T_F)/\tau-1}(k_F+1,D_{k_F},\zeta_0,\zeta_1)$. Furthermore, the maximal deviation $\mathbb{M}$ can be approximated by $||(\mathrm{C}E_0,\dots,\mathrm{C}E_K)||_\infty$ (this is an approximation since the peak may be placed at any time instant in $(0,K\tau]$, not only at instants $k\tau$). In conclusion, fixed a small tolerance $\varepsilon>0$,

$$\tau_{\text{opt}} = \underset{\tau\in S_\varepsilon}{\operatorname{argmin}}||(\mathrm{C}E_0,\dots,\mathrm{C}E_K)||_\infty$$
$$S_\varepsilon = \{\tau>0: \ \mathrm{EDP}^{T_F/\tau-1}(1{,}0,\zeta_0,\zeta_0)>1-\varepsilon, \ \mathrm{EDP}^{(T-T_F)/\tau-1}(k_F+1,D_{k_F},\zeta_0,\zeta_1)>1-\varepsilon\}. \tag{6.12}$$

Notice that if $\tau\in S_\varepsilon$, with probability $1-\varepsilon$ there are no detection errors and $||(\mathrm{C}E_0,\dots,\mathrm{C}E_K)||_\infty = \mathrm{C}E_{k_F+1}$. We will retrieve this issue in Section 6.4.

6.3.4 Evaluation of the Probability of n-step Error Decay

The evaluation of EDP^n for the One State Algorithm, which is necessary to assess the formula (6.12), is the central result of our theoretical analysis and will be used in the next section to compute τ_{opt} in some significant instances.

Proposition 12 *Let $y(t)\in\mathbb{R}$ and σ^2 be the variance of N_k; let us call*

$$S_k^w = \mathrm{C}e^{\tau\mathrm{A}}\hat{X}_{k-1} + \frac{w}{\hat{Z}_{k-2}}\mathrm{CM}_{k\tau}\in\mathbb{R}, \quad w\in\{\zeta_0,\zeta_1\}$$

the two possible received signals estimated by the One State Algorithm at time step k. Then,

$$\mathrm{EDP}^n(k_0,d,\zeta,\eta) =$$
$$= \frac{1}{2^n}\operatorname{erfc}\left(-\frac{\left|\frac{\zeta_0-\zeta_1}{2\zeta}\mathrm{CM}_{k_0\tau}\right| + \mathrm{C}e^{\tau\mathrm{A}}d\left[\left(1-2\mathbb{1}_{\{\zeta_0\}}(\eta)\right)\left(1-2\mathbb{1}_{(S_k^{\zeta_1},+\infty)}(S_k^{\zeta_0})\right)\right]}{\sigma\sqrt{2}}\right)\cdot$$
$$\cdot\prod_{m=1}^{n-1}\operatorname{erfc}\left(-\frac{\left|\frac{\zeta_0-\zeta_1}{2\eta}\mathrm{CM}_{(k_0+m)\tau}\right|}{\sigma\sqrt{2}} - \frac{\mathrm{C}e^{(m+1)\tau\mathrm{A}}d\left[\left(1-2\mathbb{1}_{\{\zeta_0\}}(\eta)\right)\left(1-2\mathbb{1}_{(S_{k+m}^{\zeta_1},+\infty)}(S_{k+m}^{\zeta_0})\right)\right]}{\sigma\sqrt{2}}\right). \tag{6.13}$$

In order to prove this proposition, we need a few technical lemmas. Let us define the following detection error probability P_{det}: given $k\in\mathbb{N}$, $d\in\mathbb{R}^n$ and $\zeta\in\{\zeta_0,\zeta_1\}$,

$$P_{\text{det}}(k,d,\zeta) = \mathrm{P}\left(\hat{Z}_k\neq z_k|D_k=d,\hat{Z}_{k-1}=\zeta\right).$$

EDP is connected to P_{det} by the following law:

Lemma 15

$$\mathrm{EDP}^n(k_0,\mathrm{d},\zeta,\eta) =$$
$$= \left(1-P_{\text{det}}(k_0-1,\mathrm{d},\zeta)\right)\big|_{z_{k_0-1}=\eta}\prod_{m=1}^{n-1}\left(1-P_{\text{det}}(k_0+m-1,e^{m\tau A}\mathrm{d},\eta)\right)\big|_{z_{k_0+m-1}=\eta}.$$

Proof By Lemma 14, we have

$$\begin{aligned}
&\mathrm{EDP}^1(k_0,\mathrm{d},\zeta,\eta) = \mathrm{P}\left(E_{k_0+1} = e^{\tau\mathrm{A}}E_{k_0}\middle|D_{k_0-1}=\mathrm{d},\hat{Z}_{k_0-2}=\zeta, z_{k_0-1}=z_{k_0}=\eta\right)\\
&= \mathrm{P}\left(\hat{Z}_{k_0-1}=z_{k_0}\middle|D_{k_0-1}=\mathrm{d},\hat{Z}_{k_0-2}=\zeta, z_{k_0-1}=z_{k_0}=\eta\right)\\
&= 1-P_{\mathrm{det}}(k_0-1,\mathrm{d},\zeta)\big|_{z_{k_0-1=\eta}}
\end{aligned}$$

that is, the error decays when the detection is correct. Notice that this relation between EDP and P_{det} subsists in virtue of the condition $z_{k_0-1}=z_{k_0}$: if k_0 were a switch point, the feedback delay would produce a deviation in the Error Function in case of correct detection.

Generalizing to n steps,

$$\begin{aligned}
&\mathrm{EDP}^n(k_0,\mathrm{d},\zeta,\eta) =\\
&= P(\hat{Z}_{k_0-1}=\hat{Z}_{k_0}=\cdots=\hat{Z}_{k_0+n-2}=\eta\big|D_{k_0-1}=d,\hat{Z}_{k_0-2}=\zeta)\\
&= \mathrm{P}\left((D_{k_0},\hat{Z}_{k_0-1})=(e^{\tau\mathrm{A}}d,\eta)|(D_{k_0-1},\hat{Z}_{k_0-2})=(d,\zeta)\right)\cdot\\
&\cdot\prod_{m=1}^{n-1}\mathrm{P}\left((D_{k_0+m},\hat{Z}_{k_0+m-1})=(e^{(m+1)\tau\mathrm{A}}d,\eta)\big|(D_{k_0+m-1},\hat{Z}_{k_0+m-2})=(e^{m\tau\mathrm{A}}d,\eta)\right)\\
&= \mathrm{EDP}^1(k_0,\mathrm{d},\zeta,\eta)\prod_{m=1}^{n-1}\mathrm{EDP}^1(k_0+m,e^{m\tau\mathrm{A}}\mathrm{d},\eta,\eta)\\
&= \left(1-P_{\mathrm{det}}(k_0-1,\mathrm{d},\zeta)\right)\big|_{z_{k_0-1}=\eta}\prod_{m=1}^{n-1}\left(1-P_{\mathrm{det}}(k_0+m-1,e^{m\tau A}\mathrm{d},\eta)\right)\big|_{z_{k_0+m-1}=\eta}.
\end{aligned}$$

■

At this point, let us compute P_{det}.

Lemma 16 *For any $k=1,2,\dots,K$,*

$$\begin{aligned}
&P_{\mathrm{det}}(k-1,d,\zeta)=\\
&=\frac{1}{2}\mathrm{erfc}\left(\frac{\left|\frac{\zeta_0-\zeta_1}{2\zeta}\mathrm{CM}_{k\tau}\right|+\mathrm{C}e^{\tau\mathrm{A}}d\left[\left(1-2\mathbb{1}_{\{\zeta_0\}}(z_{k-1})\right)\left(1-2\mathbb{1}_{(S_k^{\zeta_1},+\infty)}(S_k^{\zeta_0})\right)\right]}{\sigma\sqrt{2}}\right).
\end{aligned}\tag{6.14}$$

Proof Under the hypothesis that $z_{k-1}=\zeta_1$, P_{det} is given by:

$$\begin{aligned}
&P_{\mathrm{det}}(k-1,d,\zeta)|_{(z_{k-1}=\zeta_1)} = \mathrm{P}\left(\hat{Z}_{k-1}=\zeta_0\middle|D_{k-1}=d,\hat{Z}_{k-2}=\zeta,z_{k-1}=\zeta_1\right)\\
&= \mathrm{P}\left(|Y_k-S_k^{\zeta_0}|<|Y_k-S_k^{\zeta_1}|\ \middle|D_{k-1}=d,\hat{Z}_{k-2}=\zeta,z_{k-1}=\zeta_1\right)\\
&= \begin{cases}
\mathrm{P}\left(Y_k<\frac{S_k^{\zeta_1}+S_k^{\zeta_0}}{2}\ \middle|D_{k-1}=d,\hat{Z}_{k-2}=\zeta,z_{k-1}=\zeta_1\right) & \text{if } S_k^{\zeta_1}>S_k^{\zeta_0}\\
\mathrm{P}\left(Y_k\geq\frac{S_k^{\zeta_1}+S_k^{\zeta_0}}{2}\ \middle|D_{k-1}=d,\hat{Z}_{k-2}=\zeta,z_{k-1}=\zeta_1\right) & \text{otherwise.}
\end{cases}
\end{aligned}$$

If $S_k^{\zeta_1} > S_k^{\zeta_0}$:

$$\begin{aligned}
&\mathrm{P}\left(Y_k < \frac{S_k^{\zeta_1}+S_k^{\zeta_0}}{2}\ \middle|D_{k-1}=d, \hat{Z}_{k-2}=\zeta, z_{k-1}=\zeta_1\right) = \\
&= \mathrm{P}\left(Y_k < \mathrm{C}e^{\tau\mathrm{A}}\hat{X}_{k-1} + \frac{\zeta_0+\zeta_1}{2\zeta}\mathrm{CM}_{k\tau}\ \middle|D_{k-1}=d\right) \\
&= \mathrm{P}\left(\mathrm{C}X_k + N_k < \mathrm{C}e^{\tau\mathrm{A}}\hat{X}_{k-1} + \frac{\zeta_0+\zeta_1}{2\zeta}\mathrm{CM}_{k\tau}\ \middle|D_{k-1}=d\right) \\
&= \mathrm{P}\left(\mathrm{C}e^{\tau\mathrm{A}}X_{k-1} + \frac{\zeta_1}{\zeta}\mathrm{CM}_{k\tau} + N_k < \mathrm{C}e^{\tau\mathrm{A}}\hat{X}_{k-1} + \frac{\zeta_1+\zeta_0}{2\zeta}\mathrm{CM}_{k\tau}\ \middle|D_{k-1}=d\right) \\
&= \mathrm{P}\left(N_k < \mathrm{C}e^{\tau\mathrm{A}}d + \frac{\zeta_0-\zeta_1}{2\zeta}\mathrm{CM}_{k\tau}\right) \\
&= \frac{1}{2}\mathrm{erfc}\left(\frac{-\mathrm{C}e^{\tau\mathrm{A}}d + \frac{\zeta_1-\zeta_0}{2\zeta}\mathrm{CM}_{k\tau}}{\sigma\sqrt{2}}\right).
\end{aligned}$$

The last step depends on the Gaussian distribution of N_k; notice also that $\frac{\zeta_1-\zeta_0}{\zeta}\mathrm{CM}_{k\tau} = S_k^{\zeta_1} - S_k^{\zeta_0} > 0$. It follows also that for $S_k^{\zeta_1} \leq S_k^{\zeta_0}$:

$$\begin{aligned}
&\mathrm{P}\left(Y_k \geq \frac{S_k^{\zeta_1}+S_k^{\zeta_0}}{2}\ \middle|D_{k-1}=d, \hat{Z}_{k-2}=\zeta, z_{k-1}=\zeta_1\right) = \\
&\qquad = 1 - \frac{1}{2}\mathrm{erfc}\left(\frac{-\mathrm{C}e^{\tau\mathrm{A}}d + \frac{\zeta_1-\zeta_0}{2\zeta}\mathrm{CM}_{k\tau}}{\sigma\sqrt{2}}\right)
\end{aligned}$$

where $\frac{\zeta_1-\zeta_0}{\zeta}\mathrm{CM}_{k\tau} = S_k^{\zeta_1} - S_k^{\zeta_0} \leq 0$. Summing up,

$$\begin{aligned}
&P_{\det}(k-1,d,\zeta)|_{(z_{k-1}=\zeta_1)} = \\
&= \mathrm{P}\left(|Y_k - S_k^{\zeta_0}| < |Y_k - S_k^{\zeta_1}|\ \middle|D_{k-1}=d, \hat{Z}_{k-2}=\zeta, z_{k-1}=\zeta_1\right) \\
&= \begin{cases} \frac{1}{2}\mathrm{erfc}\left(\frac{-\mathrm{C}e^{\tau\mathrm{A}}d+\frac{\zeta_1-\zeta_0}{2\zeta}\mathrm{CM}_{k\tau}}{\sigma\sqrt{2}}\right) & \text{if } S_k^{\zeta_1} > S_k^{\zeta_0} \\ 1-\frac{1}{2}\mathrm{erfc}\left(\frac{-\mathrm{C}e^{\tau\mathrm{A}}d+\frac{\zeta_1-\zeta_0}{2\zeta}\mathrm{CM}_{k\tau}}{\sigma\sqrt{2}}\right) & \text{otherwise.} \end{cases}
\end{aligned}$$

This actually corresponds to the false negative probability. The false positive probability $P_{\det}(k-1,d,\zeta)|_{(z_{k-1}=\zeta_0)}$ can be computed in the same way and the result is:

$$\begin{aligned}
&P_{\det}(k-1,d,\zeta)|_{(z_{k-1}=\zeta_0)} = \mathrm{P}\left(\hat{Z}_{k-1}=\zeta_1\middle|D_{k-1}=d, \hat{Z}_{k-2}=\zeta, z_{k-1}=\zeta_0\right) \\
&= \mathrm{P}\left(|Y_k - S_k^{\zeta_1}| < |Y_k - S_k^{\zeta_0}|\ \middle|D_{k-1}=d, \hat{Z}_{k-2}=\zeta, z_{k-1}=\zeta_0\right) \\
&= \begin{cases} 1-\frac{1}{2}\mathrm{erfc}\left(\frac{-\mathrm{C}e^{\tau\mathrm{A}}d-\frac{\zeta_1-\zeta_0}{2\zeta}\mathrm{CM}_{k\tau}}{\sigma\sqrt{2}}\right) & \text{if } S_k^{\zeta_1} > S_k^{\zeta_0} \\ \frac{1}{2}\mathrm{erfc}\left(\frac{-\mathrm{C}e^{\tau\mathrm{A}}d-\frac{\zeta_1-\zeta_0}{2\zeta}\mathrm{CM}_{k\tau}}{\sigma\sqrt{2}}\right) & \text{otherwise.} \end{cases}
\end{aligned}$$

The thesis is then proved. ■

Remark 8 *By the definition of D_k, we have*

$$P_{\text{det}}(k,d,\zeta) = \mathrm{P}\left(\hat{Z}_k \neq z_k, D_{k+1} = e^{\tau \mathrm{A}} d + \frac{z_k^c - z_k}{z_{k-1}} \mathrm{M}_{(k+1)\tau} \middle| D_k = d, \hat{Z}_{k-1} = z_{k-1}\right) \tag{6.15}$$

where z_k^c indicates the complementary of z_k in $\{\zeta_0, \zeta_1\}$. This probability may be interpreted as the transition probability of the Markov Process

$$(D_k, \hat{Z}_{k-1})_{k=0,1,\ldots}$$

in the state space $\mathbf{D} \times \{\zeta_0, \zeta_1\}$, $\mathbf{D} \subset \mathbb{R}^n$, with starting state $(D_0, \hat{Z}_{-1}) = (0, \zeta_0)$. A thorough analysis of this process using Markov Theory should provide more general results than ours, but this approach is too complex when the problem is multidimensional.

Remark 9 *If $d = 0 \in \mathbb{R}^n$,*

$$\begin{aligned} P_{\text{det}}(k-1,0,\zeta) &= \frac{1}{2} erfc\left(\frac{\left|\frac{\zeta_0 - \zeta_1}{2\zeta} \mathrm{CM}_{k\tau}\right|}{\sigma\sqrt{2}}\right) \\ &= \frac{1}{2} erfc\left(\frac{|S_k^{\zeta_0} - S_k^{\zeta_1}|/2}{\sigma\sqrt{2}}\right). \end{aligned} \tag{6.16}$$

This expression suggests an Information-theoretic interpretation of our problem. In fact, the presence of the Gaussian noise in the data lecture can be thought as if signal $\mathrm{C}x_k$ were transmitted on an AWGN channel. If $D_{k-1} = 0$, $\mathrm{C}x_k$ can be $S_k^{\zeta_0}$ or $S_k^{\zeta_1}$. Moreover, if we shift the signals by their average, so that they become antipodal $\pm\frac{S_k^{\zeta_0} - S_k^{\zeta_1}}{2}$, the average energy per channel use at step k is $\mathcal{E}_k = \left(\frac{S_k^{\zeta_0} - S_k^{\zeta_1}}{2}\right)^2$. Given that the spectral density of the Gaussian noise is $N_0 = 2\sigma^2$, the argument of the erfc function in (6.16) turns out to be the square root of the so called Signal-to-Noise Ratio (SNR), defined as $SNR_{k,\tau} = \mathcal{E}_k/N_0$, of our ideal channel. The subscripts emphasize the dependence of the SNR on time and on parameter τ.

Generally, the SNR compares the magnitudes of the transmitted signal and of the channel noise and it is widely used in Information Theory to describe channel performance. In our framework, the SNR determines the reliability of the detection, say the reliability of the channel where $\mathrm{C}x_k$ is ideally transmitted. This remark emphasizes that our problem is analogous to a common digital-transmission paradigm and bears out the idea of using decoding techniques to the detection task.

In the next, we will use the common dB notation for the SNR, that is, we express it as $10\log_{10}$ of its value.

Remark 10 *Since typically $\zeta_1 < \zeta_0$, by expression (6.16) we have*

$$P_{\text{det}}(k-1,0,\zeta_1) < P_{\text{det}}(k-1,0,\zeta_0).$$

Given that $\hat{Z}_{k-2} = \zeta_1$ is generally more likely when $z_{k-2} = \zeta_1$ (otherwise our detection method would be improper), we can conclude that our detection algorithm is more reliable after the failure, or, in other terms, it is more sensitive to false positives.

Proof of Proposition 12.

Notice that z_k is assumed to be constant in $[k_0 - 1, k_0 + n - 1]$, that is, we consider the system before or after a failure event. By Lemmas 15 and 16, we have

$$\begin{aligned}
&\mathrm{EDP}^n(k_0, d, \zeta, \eta) = \\
&= \frac{1}{2}\mathrm{erfc}\left(- \frac{\left| \frac{\zeta_0 - \zeta_1}{2\zeta} \mathrm{CM}_{k_0\tau} \right| + \mathrm{C}e^{\tau\mathrm{A}} d \left[\left(1 - 2\mathbb{1}_{\{\zeta_0\}}(\eta)\right) \left(1 - 2\mathbb{1}_{(S_k^{\zeta_1}, +\infty)}(S_k^{\zeta_0})\right) \right]}{\sigma\sqrt{2}} \right) \cdot \\
&\quad \cdot \prod_{m=1}^{n-1} \frac{1}{2}\mathrm{erfc}\left(- \frac{\left| \frac{\zeta_0 - \zeta_1}{2\eta} \mathrm{CM}_{(k_0+m)\tau} \right|}{\sigma\sqrt{2}} + \right. \\
&\qquad\qquad \left. - \frac{\mathrm{C}e^{(m+1)\tau\mathrm{A}} d \left[\left(1 - 2\mathbb{1}_{\{\zeta_0\}}(\eta)\right) \left(1 - 2\mathbb{1}_{(S_{k+m}^{\zeta_1}, +\infty)}(S_{k+m}^{\zeta_0})\right) \right]}{\sigma\sqrt{2}} \right).
\end{aligned}$$

■

Let us now briefly distinguish the behavior of EDP^n before and after the failure.

6.3.5 False positive evaluation

Let us suppose the system to be affected by a failure according to the model (6.8) with $k_F \geq 1$, that is, the system is not faulty from the beginning. In particular, since there is no compensation at the first time step (or equivalently $\hat{Z}_{-1} = \zeta_0$), no false positive is produced at $k = 0$. Then, studying EDP in $[1, k_F)$ actually corresponds to evaluate the probability that no false positives occur during the whole pre-failure transient regime. Given that $D_0 = 0$, we have

$$\mathrm{EDP}^{k_F - 1}(1, 0, \zeta_0, \zeta_0) = \prod_{m=1}^{k_F - 1} \frac{1}{2}\mathrm{erfc}\left(- \frac{\left| \frac{\zeta_0 - \zeta_1}{2\zeta_0} \mathrm{CM}_{m\tau} \right|}{\sigma\sqrt{2}} \right). \tag{6.17}$$

Since $E_1 = 0$ and $D_0 = 0$, then $\mathrm{EDP}^{k_F - 1}(1, 0, \zeta_0, \zeta_0) = P(E_{k_F} = 0) = P(D_{k_F} = 0)$.

6.3.6 Switch Point

Suppose that $D_{k_F} = 0$, then in particular, $\hat{Z}_{k_F - 1} = z_{k_F - 1}$ and $\hat{Z}_{k_F - 1} \neq z_{k_F}$. In other terms, the detection is correct, but the compensation, based on the detection at the previous step, is not efficient in correspondence of a switch point. Our detection method cannot control what happens at step at step k_F, that is, in the time interval $[T_F, T_F + \tau)$.

6.3.7 False negative evaluation

Given that we cannot control the system immediately after the switch point, it is likely that $E_{k_F + 1} \neq 0$. We now want to study the probability of decay of the Error Function towards zero, which actually corresponds to the evaluation of the false negatives. In fact, under the hypothesis $D_{k_F} = 0$ (i.e., no false positives and in particular $\hat{Z}_{k_F-1} = \zeta_0$),

$$\mathrm{EDP}^{K-k_F-1}(k_F+1,0,\zeta_0,\zeta_1) = \mathrm{EDP}^1(k_F+1,0,\zeta_0,\zeta_1)\prod_{m=1}^{n-1}\mathrm{EDP}^1(k_F+1+m,0,\zeta_1,\zeta_1)$$
$$\frac{1}{2}\mathrm{erfc}\left(-\frac{\left|\frac{\zeta_0-\zeta_1}{2\zeta_0}\mathrm{CM}_{(k_F+1)\tau}\right|}{\sigma\sqrt{2}}\right)\prod_{m=1}^{n-1}\frac{1}{2}\mathrm{erfc}\left(-\frac{\left|\frac{\zeta_0-\zeta_1}{2\zeta_1}\mathrm{CM}_{(k_F+m+1)\tau}\right|}{\sigma\sqrt{2}}\right). \tag{6.18}$$

The considerations about EDP made in the last sections are now specialized to the case of constant input $f(t)$.

6.3.8 Constant input $f(t)$

If the input $f(t)$ is constant the last expression can be simplified and analytically evaluated, as the system evolution does not depend on time step k. Let us fix $f \equiv 1$: we have

$$\mathrm{M}_{k\tau} = \mathrm{M}_\tau := (e^{\tau\mathrm{A}} - \mathbb{I})\mathrm{A}^{-1}\mathrm{B}$$

for any $k = 1,\dots,K$. Hence,

$$\mathrm{EDP}^n(1,0,\zeta_0,\zeta_0) = \left[\frac{1}{2}\mathrm{erfc}\left(-\frac{\left|\frac{\zeta_0-\zeta_1}{2\zeta_0}\mathrm{CM}_\tau\right|}{\sigma\sqrt{2}}\right)\right]^n \tag{6.19}$$

for any $n \in \mathbb{N}$ such that $n+1 \leq k_F$ and

$$\mathrm{EDP}^n(k_F+1,0,\zeta_0,\zeta_1) = \frac{1}{2}\mathrm{erfc}\left(-\frac{\left|\frac{\zeta_0-\zeta_1}{2\zeta_0}\mathrm{CM}_\tau\right|}{\sigma\sqrt{2}}\right)\left[\frac{1}{2}\mathrm{erfc}\left(-\frac{\left|\frac{\zeta_0-\zeta_1}{2\zeta_1}\mathrm{CM}_\tau\right|}{\sigma\sqrt{2}}\right)\right]^{n-1}. \tag{6.20}$$

In terms of signal-to-noise ratio, we can write

$$\sqrt{\mathrm{SNR}_\tau(\eta)} = \frac{\left|\frac{\zeta_1-\zeta_0}{2\eta}\mathrm{CM}_\tau\right|}{\sigma\sqrt{2}}$$

so that

$$\mathrm{EDP}^n(1,0,\zeta_0,\zeta_0) = \left[\frac{1}{2}\mathrm{erfc}\left(-\sqrt{\mathrm{SNR}_\tau(\zeta_0)}\right)\right]^n$$
$$\mathrm{EDP}^n(k_F+1,0,\zeta_0,\zeta_1) = \frac{1}{2}\mathrm{erfc}\left(-\sqrt{\mathrm{SNR}_\tau(\zeta_0)}\right)\left[\frac{1}{2}\mathrm{erfc}\left(\sqrt{\mathrm{SNR}_\tau(\zeta_1)}\right)\right]^{n-1}.$$

Under the hypothesis $0 < \zeta_1 < \zeta_0 = 1$, $\mathrm{SNR}_\tau(\zeta_0) < \mathrm{SNR}_\tau(\zeta_1)$, that is $\mathrm{EDP}^m(k_0,0,\zeta_0,\zeta_0) < \mathrm{EDP}^m(k_1,0,\zeta_1,\zeta_1)$; in other terms, our detection algorithm is more sensitive to false positives, then our fault tolerant control method is more efficient *after* the failure. Thus, the suitable design criteria for the pre-failure state will automatically be appropriate also for the post-failure state, recalling that in general we ask EDP to be larger than a given threshold (see (6.22)).

For this motivation, in the next we will focus on the pre-failure framework and, for brevity, we will adopt this notation:

$$\mathrm{SNR}_\tau = \mathrm{SNR}_\tau(\zeta_0) \quad \text{and} \quad \mathrm{EDP}^n = \mathrm{EDP}^n(k_0, 0, \zeta_0, \zeta_0) = \left[\frac{1}{2}\mathrm{erfc}\left(-\sqrt{\mathrm{SNR}_\tau}\right)\right]^n. \qquad (6.21)$$

The next section is devoted to assess the optimal design criteria for our FTC system, in a few instances, on the basis of the theoretical analysis developed in this section.

6.4 Design Criteria

On the basis of the previous analysis, let us assess the optimal design criteria for our FTC system in two different input instances: $f(t)$ constant and $f(t)$ sinusoidal. As far as the first case in concerned, we will show that the theoretic analysis of Section 6.3 provides the instruments to determine the optimal sampling step in an analytic way. On the other hand, when the input is not constant some difficulties arise in the analytical computation.

6.4.1 Design Criteria in the case of constant input $f(t)$

Let us evaluate τ_{opt} when $f(t) \equiv 1$. If there are no detection errors, the maximal deviation in the output is in the interval $(k_F\tau, (k_F+1)\tau]$ and is equal to $\max_{t\in(0,\tau]} |\frac{\zeta_1-\zeta_0}{\zeta_0}\mathrm{CM}_t|$ where $\mathrm{M}_t = (e^{t\mathrm{A}} - \mathbb{I})\mathrm{A}^{-1}\mathrm{B}$. Let us approximate it by $\mathrm{C}E_{k_F+1} = |\frac{\zeta_1-\zeta_0}{\zeta_0}\mathrm{CM}_\tau|$. Given the definition of τ_{opt} in (6.12), our aim is then to provide

$$\tau_{\mathrm{opt}} = \underset{\tau\in S_\varepsilon}{\mathrm{argmin}} |\mathrm{CM}_\tau| \qquad (6.22)$$

where $S_\varepsilon = \{\tau > 0 :\ \mathrm{EDP}^{T_F/\tau-1}(1, 0, \zeta_0, \zeta_0) > 1-\varepsilon, \mathrm{EDP}^{(T-T_F)/\tau-1}(k_F+1, D_{k_F}, \zeta_0, \zeta_1) > 1-\varepsilon\}$.

6.4.1.1 Application to the Flight Control Problem

Let us now compute τ_{opt} for the Flight Control Problem introduced in Section 6.1.1, in the case of constant input $f(t)$. In the Figure 6.1, the graph of CM_τ in function of τ is shown. In particular, we notice that CM_τ is negative for any $\tau > 0$, achieves a global minimum at $\tau_0 = 0.55$ and converges to a constant value for a sufficiently large τ. Then, if $\tau > \tau_0$, $\max_{t\in(0,\tau]} |\mathrm{CM}_t| = |\mathrm{CM}_{\tau_0}|$, that is, the peak is fixed and we cannot control it. This undesired occurrence can be prevented by imposing

$$\tau \in (0, \tau_0].$$

In this interval, CM_τ is monotone decreasing and we exactly have $\max_{t\in(0,\tau]} |\mathrm{CM}_t| = |\mathrm{CM}_\tau|$. Then, fixed the tolerance ε, our aim (see (6.12)) is the computation of

$$\tau_{\mathrm{opt}} = \underset{\tau\in(0,\tau_0]:\mathrm{EDP}^{W/\tau}>1-\varepsilon}{\mathrm{argmin}} |\mathrm{CM}_\tau| \qquad (6.23)$$

where $W = n\tau$ indicates the length of the window we are considering.

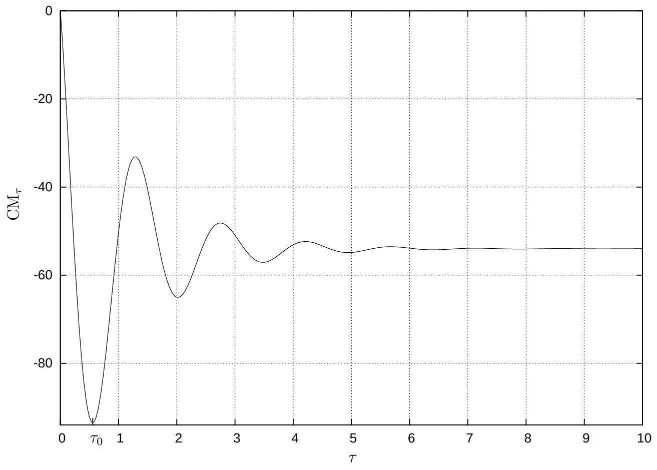

Figure 6.1: CM_τ.

Notice that

$$\mathrm{EDP}^{W/\tau} = \left[\frac{1}{2}\mathrm{erfc}\left(-\sqrt{\mathrm{SNR}_\tau}\right)\right]^{W/\tau} = \left[\frac{1}{2}\mathrm{erfc}\left(-\frac{|\frac{\zeta_1-\zeta_0}{2\zeta_0}\mathrm{CM}_\tau|}{\sigma\sqrt{2}}\right)\right]^{W/\tau}$$

is monotone increasing as a function of τ. Then, let $\tau_m = \tau_m(\varepsilon)$ be the minimum τ in $(0, \tau_0]$ such that $\mathrm{EDP}^{W/\tau} > 1 - \varepsilon$ (if it exists). Then

$$\tau_{\mathrm{opt}} = \underset{\tau \geq \tau_m}{\mathrm{argmin}} |\mathrm{CM}_\tau| = \tau_m. \tag{6.24}$$

Now, let assign numerical values to the parameters and solve the corresponding instance: if

$$\begin{aligned} &\zeta_0 = 1 \quad \zeta_1 = \frac{1}{2} \quad \sigma^2 = 2 \\ &\varepsilon = 10^-3 \quad W = 20 \end{aligned} \tag{6.25}$$

then $\tau_{\mathrm{opt}} = 0.12$ as shown in Figure 6.2. The value of τ_{opt} clearly depends on the noise and in particular there can exist noise values for which there is no τ making $\mathrm{EDP}^{W/\tau} > 1 - \varepsilon$: for instance, this occurs if we consider $\sigma^2 > 34.72$ in the example (6.25) (the range of admissible σ^2's with the corresponding τ_{opt}'s is shown in Figure 6.3). In such occurrences, one should allow a lower threshold $1 - \varepsilon$.

6.4.2 Design Criteria in the case of input $f(t) = \sin t$

When $f(t)$ is not constant, it is more difficult to study analytical design criteria as the quality of the detection depends on time. In particular, at each time step $k\tau$ the detection is affected by the values of $f(t)$, $t \in ((k-1)\tau, k\tau)$, then any detection step is different from the others and an analogous of (6.22) cannot be provided: roughly speaking, the optimum would be to change τ according to the shape of $f(t)$ in each considered interval.

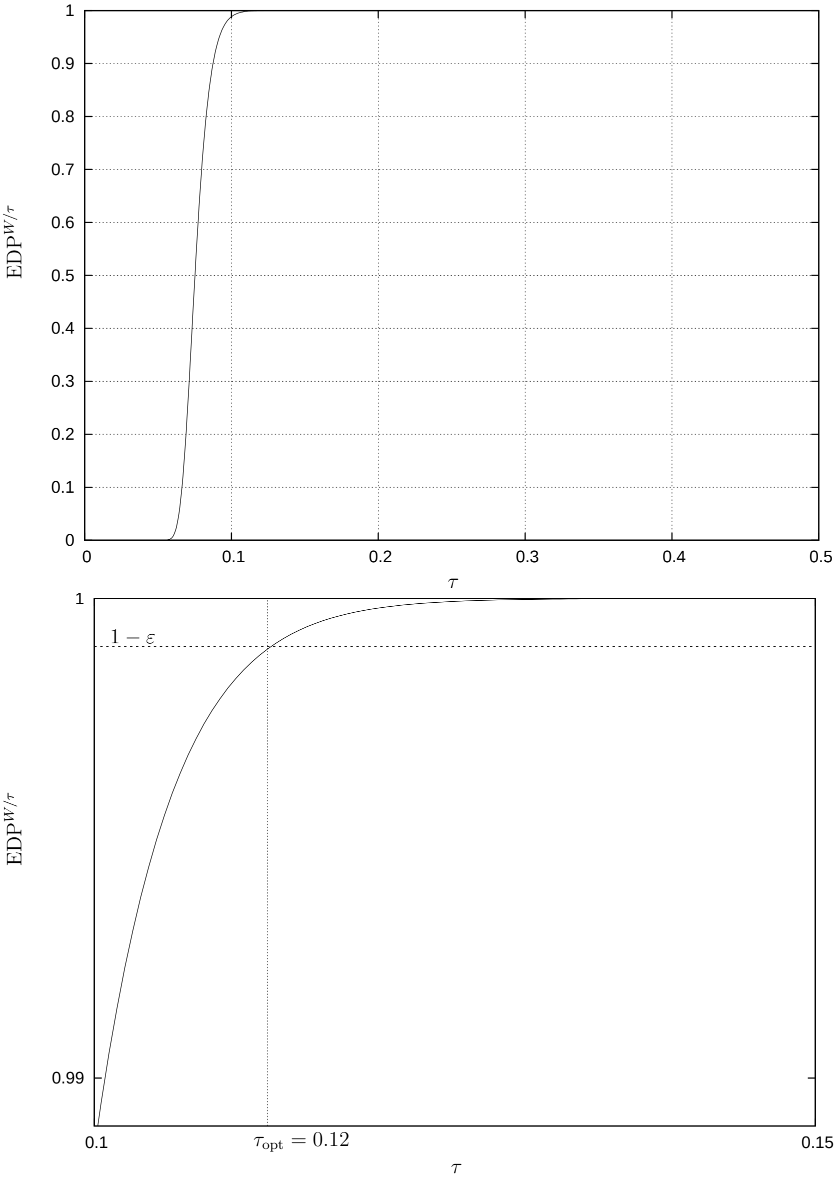

Figure 6.2: $\mathrm{EDP}^{W/\tau}$ in function of τ in the instance (6.25). The second graph is a zoom that allows to see that $\tau_{\text{opt}} = 0.12$.

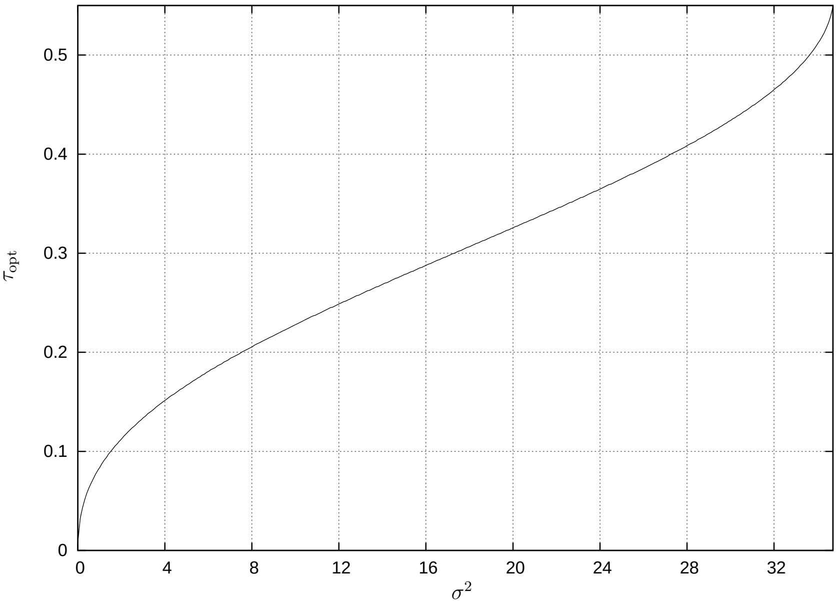

Figure 6.3: The optimal τ's as the noise variance σ^2 changes ($\zeta_0 = 1, \zeta_1 = \frac{1}{2}, \varepsilon = 10^-3, W = 20$).

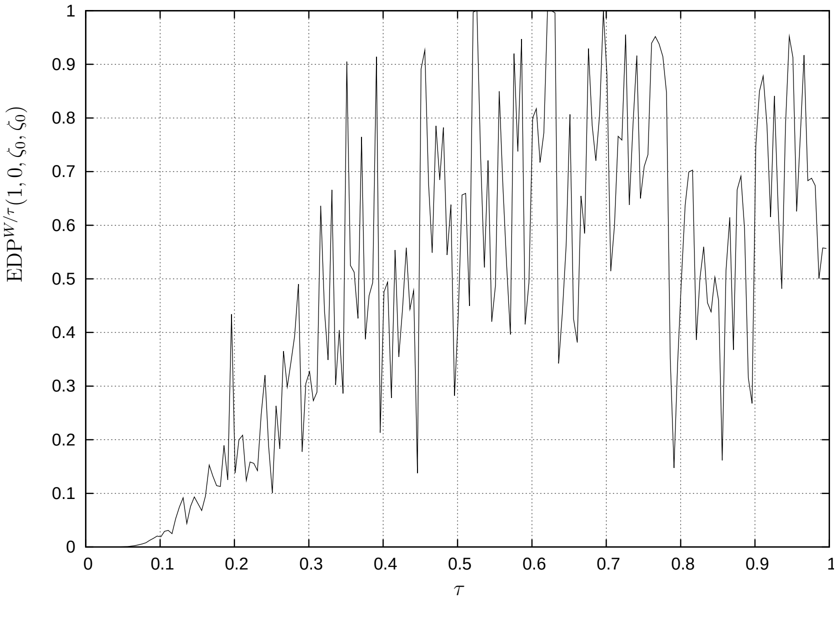

Figure 6.4: $\text{EDP}^{W/\tau}(1, 0, \zeta_0, \zeta_0)$ in function of τ in the instance (6.25) ($\zeta_0 = 1, \zeta_1 = \frac{1}{2}, \sigma^2 = 2, W = 20$).

When $f(t)$ is periodic, we can suggest some numerical computation in order to fix a suitable τ. In fact, if we compute $\mathrm{EDP}^{W/\tau}(1, 0, \zeta_0, \zeta_0)$ for a sufficiently large W, we get an idea about the sampling times that are more suitable. On the other hand, there is no way to control the amplitude of the deviation in case of failure, given its dependence on time. The idea is then to choose as sampling time that maximizes $\mathrm{EDP}^{W/\tau}(1, 0, \zeta_0, \zeta_0)$ or that makes it larger than a given threshold, understanding that this does not arrange the issue of the unavoidable deviation.

Let us illustrate these observations in the Flight Control Problem with $f(t) = \sin t$ and parameters given by (6.25). First, let us numerically compute $\mathrm{EDP}^{W/\tau}(1, 0, \zeta_0, \zeta_0)$ in function of τ, the result being presented in Figure 6.4: the graph shows a clear unsettled behavior which cannot be described analytically. However, it also suggests the values of τ that give an high $\mathrm{EDP}^{W/\tau}(1, 0, \zeta_0, \zeta_0)$ and which can then considered suitable.

More details about this instance can be retrieved in the simulations presented in the next section.

6.5 Flight Control Problem: a few simulations

In this section, we show some simulations concerning the application of the One State Algorithm to the Flight FTC example presented in the Paragraph 6.1.1 and studied in the previous paragraphs.

In a time interval $[0, T] = [0, 40]$, we suppose that a failure occurs at $T_F = 20$ and causes the switch of the disturbance function $z(t)$ from $\zeta_0 = 1$ to $\zeta_1 = 1/2$ ($\zeta_1 = 1/2$ might represent a loss of effectiveness of 50% of the elevator of the aircraft). The measurement noise is a Gaussian random variable $\mathcal{N}(0, 2)$. We consider both the cases of input $f \equiv 1$ and $f(t) = \sin t$ and we show the behavior of the One State procedure for different values of τ. The graphs represent the output $y(t)$ of the system.

Figure 6.5 reproduces the case $f \equiv 1$. The first graph compares the nominal system, that is, the desirable trajectory, to the faulty system with no compensation: after the failure, the trajectory of the latter is appreciably incorrect. In the other graphs, we introduce the compensation using the One State Algorithm: as proved in the Paragraph 6.4.1.1 , $\tau_{\mathrm{opt}} = 0.12$. In the second graph, we fix $\tau = 0.4$, which is larger than τ_{opt}: we obtain a correct detection at each step, but the unavoidable deviation is not optimized: in fact, considering τ_{opt} (third graph), we have a smaller peak after the failure. Furthermore, we see that also $\tau = 0.09$ is suitable, even if, the corresponding $\mathrm{EDP}^{W/\tau} > 1 - \varepsilon$. On the other hand, $\tau = 0.06$ assures a good detection only after the failure (this is consistent with our observation about the different sensitivity of false positives and false negatives), while a too small sampling time ($\tau = 0.001$) causes instability: the detection is not reliable and the Error is always non-null.

Figure 6.6 concerns the case $f(t) = \sin t$. Again, the output of the system with no compensation in the first graph undergoes an evident change after the failure at $T_F = 20$. Instead, applying the One State Algorithm with time step $\tau = 0.525$ (this value being suggested by the numerical computation of the EDP) allows to recover the nominal condition. The same occurs with $\tau = 0.35$, which is preferable for the smaller amplitude of the unavoidable deviation in correspondence to the switch point.

When $\tau = 0.3$, some detections fail (the error percentage is about 4%), but the output y is not dramatically affected by them. Furthermore, when $\tau = 0.01$ the error percentage is about 9%: many deviations occur, but they are not very large. In particular, they are quite null when the slope of $y(t)$ is steeper. In correspondence to the switch point a plain oscillation is present, but it is less remarkable than in the cases of larger τ.

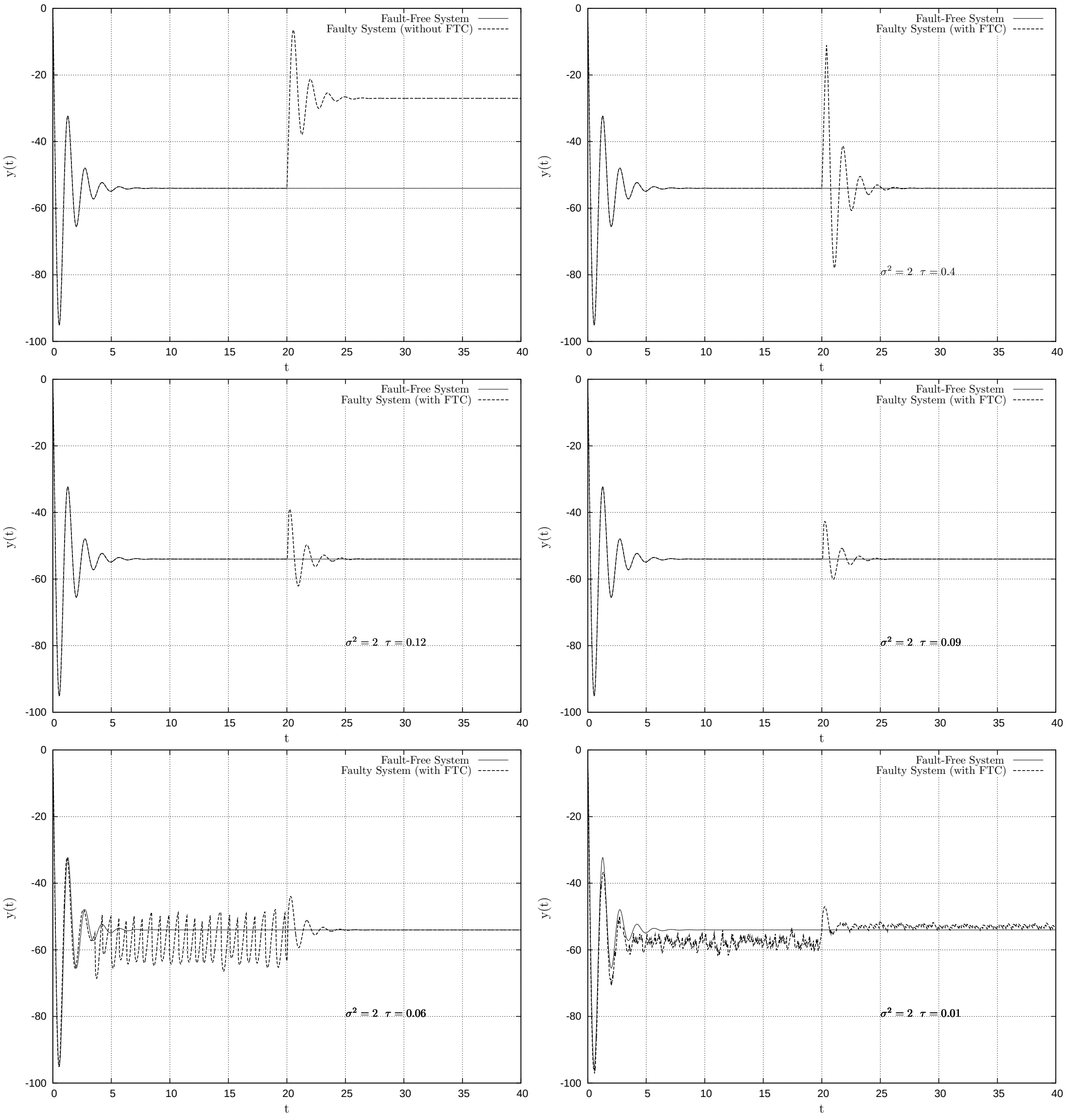

Figure 6.5: Fault-free System vs System with a failure at $T_F = 20$, with measurement noise of variance $\sigma^2 = 2$ and $f \equiv 1$. The x-axes represent the time, the y-axes the trajectories $y(t)$. Six different cases are shown: the first graph represents the system with no fault compensation (say, $u(t) \equiv 0$); the other ones are with compensation, respectively with time step τ equal to 0.4, 0.12, 0.09, 0.06, 0.01

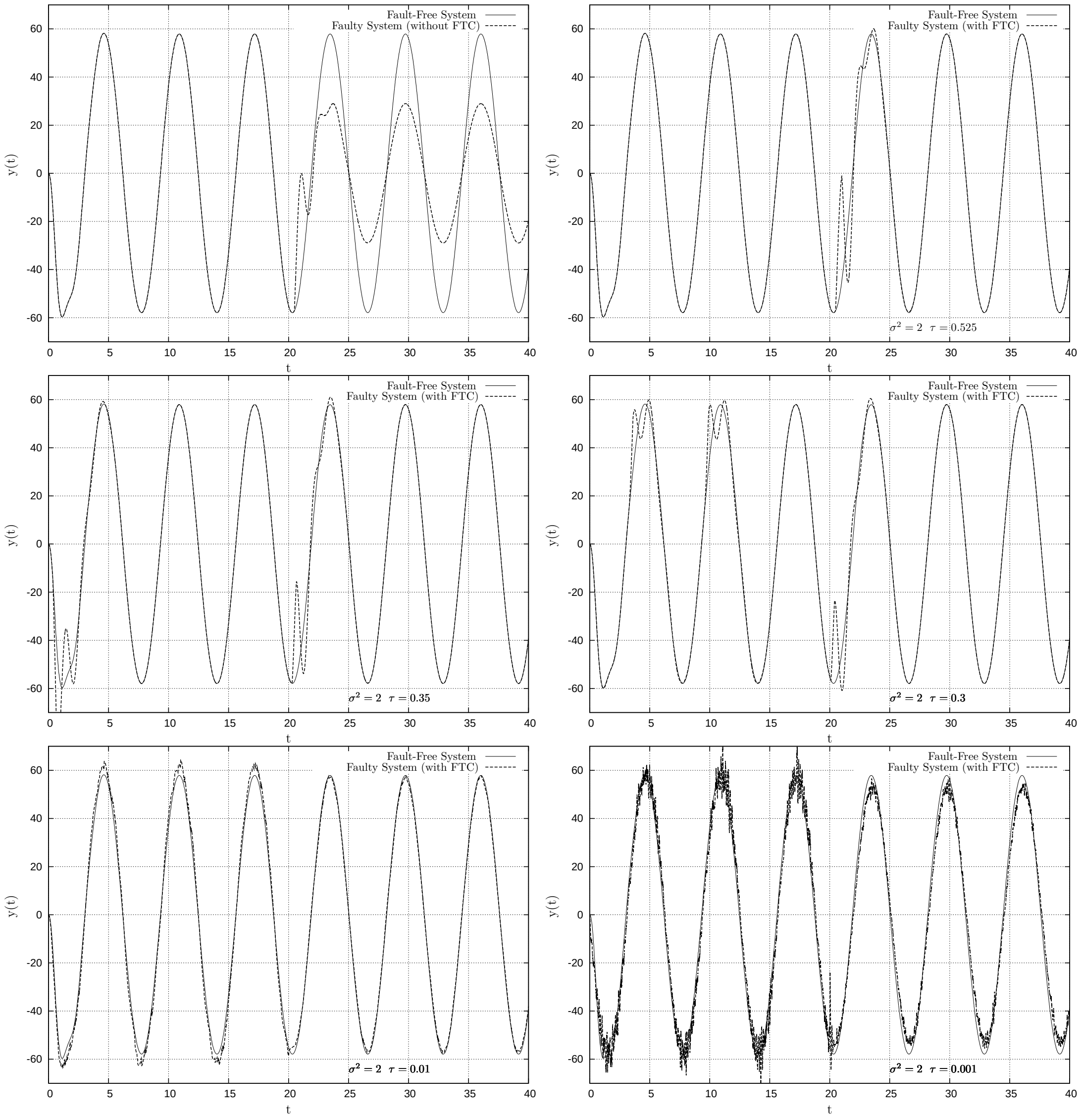

Figure 6.6: Nominal System vs System with a failure at $T_F = 20$, with measurement noise of variance $\sigma^2 = 2$ and $f(t) = \sin t$. The x-axes represent the time, the y-axes the trajectories $y(t)$. Six different cases are shown: the first graph represents the system with no fault compensation (say, $u(t) \equiv 0$); the other ones are with compensation, respectively with time step τ equal to 0.525, 0.35, 0.3, 0.01, 0.001

Decreasing τ again, the percentage of wrong detections does not overpass 10%, but for very small values of τ, the system is unstable (see for instance, the last graph corresponding to $\tau = 0.001$) and many oscillations occur.

6.6 Inaccurate quantization

The disturbance function considered in this chapter may assume only two known values. This is a simplified case that has been exploited to introduce our decoding approach to FTC in the easiest way, but real systems are in general more complicated. Two main realistic cases should be discussed:

1. the disturbance function is quantized, but assumes more than two values;
2. the disturbance function is not quantized.

Both problems require detection and also *identification* of the disturbance (see the introductory part of this chapter). Point (1) can be tackled with our approach (see Remark 7): if q is the number of quantization levels, it suffices to adapt the Disturbance Estimation task in the One State Algorithm performing a comparison among q Euclidean distances. Nevertheless, the corresponding theoretical analysis turns out to be more complicated. The second problem, instead, can be approached introducing a sufficiently fine quantization, taking into account that a larger number of quantization levels produces more precise results in spite of numerical complexity.

The examination of these issues is beyond our purpose, but some observations can be made about the following point: what happens if we apply the One State Algorithm when the value of the disturbance function $z(t)$ is constant, but unknown, after the failure?

More precisely, let us suppose that $z(t) \in \{\zeta_0, \alpha\}$ where $\zeta_0 = 1$ and $\alpha \in (0, 1)$ is not given, while $\hat{z}(t) \in \{\zeta_0, \zeta_1\}$, ζ_1 being fixed: the Algorithm works with two quantized states, but the failure state value that it considers may be incorrect. Then, let us wonder which error is produced by this inaccurate quantization; naturally, we expect that if α is sufficiently close to ζ_0, the outcomes will be sufficiently reliable. Notice also that the false positive discussion is not touched by this issue.

First of all, we have to distinguish the errors in the detection task and in the trajectories. The detection is correct if we become aware of the switch of $z(t)$ from ζ_0 to α; in this case, the One State Algorithm estimates $z(t)$ with ζ_1. It is easy to compute that the probability of incorrect detection is given by:

$$P(\hat{Z}_k = \zeta_0 | z_k = \alpha, D_k = \mathrm{d}, \hat{Z}_{k-1} = \zeta) =$$
$$= \frac{1}{2}\mathrm{erfc}\left(\frac{\left|\frac{\zeta_0+\zeta_1-2\alpha}{2\zeta}\mathrm{CM}_{k\tau,(k-1)\tau}\right| + \mathrm{C}e^{\tau\mathrm{A}}d\left[\left(1 - 2\mathbb{1}_{\{\zeta_0\}}(z_{k-1})\right)\left(1 - 2\mathbb{1}_{(S_k^{\zeta_1},+\infty)}(S_k^{\zeta_0})\right)\right]}{\sigma\sqrt{2}}\right). \tag{6.26}$$

The calculus is analogous to the one for the P_{det} and the result is similar, but with $\zeta_0 + \zeta_1 - 2\alpha$ instead of $\zeta_0 - \zeta_1$ Now, let us analyze it in the example proposed in Section 6.4.1.1, in the case of constant f: considering $\mathrm{d} = 0$, we have

$$P(\hat{Z}_k = \zeta_0 | z_k = \alpha, D_k = 0, \hat{Z}_{k-1} = \zeta) = \frac{1}{2}\mathrm{erfc}\left(\frac{\left|\frac{2\alpha-\zeta_0-\zeta_1}{2\zeta}\mathrm{CM}_\tau\right|}{\sigma\sqrt{2}}\right). \tag{6.27}$$

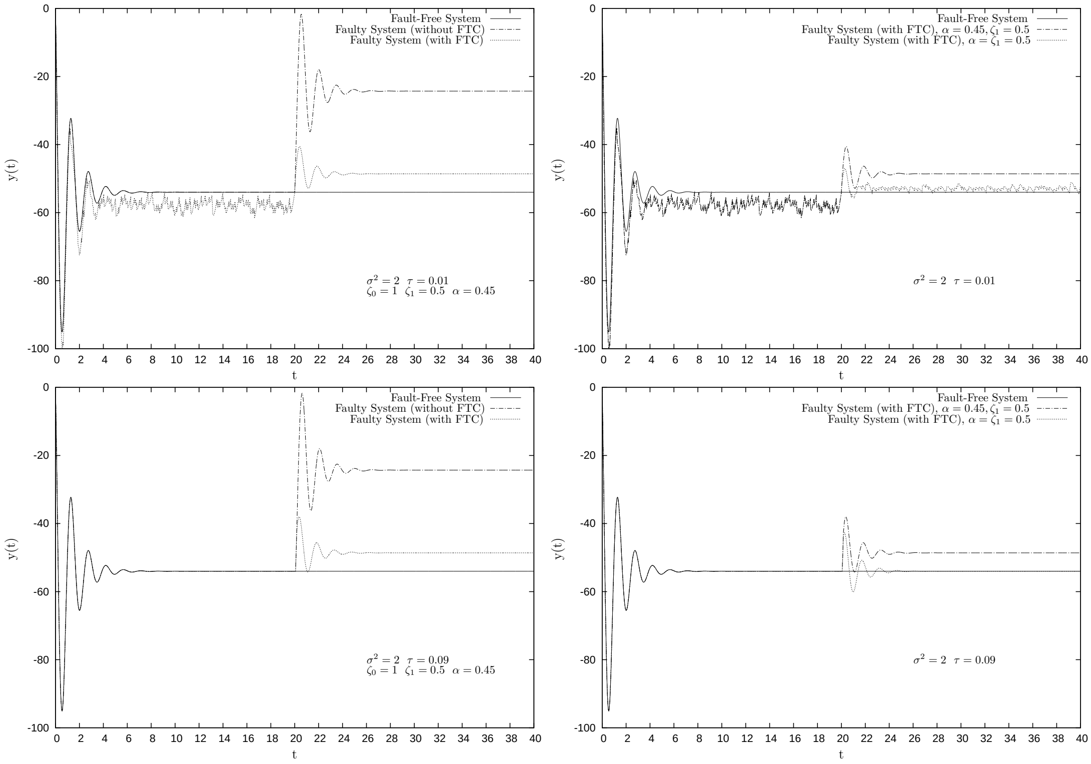

Figure 6.7: Error due to inaccurate quantization. The graphs represent the trajectories y in function of the time t. Above, the graphs for $\tau = 0.01$: after the failure, if $\alpha = \zeta_1 = 0.5$ the detection presents many errors, while if $\alpha = 0.45, \zeta_1 = 0.5$, it is exact (the corresponding trajectory is perfectly parallel to the desired one). Below, the graphs for $\tau = 0.09$: after the failure, in both cases $\alpha = \zeta_1 = 0.5$ and $\alpha = 0.45, \zeta_1 = 0.5$ the detection is correct, but in the second case the obtained trajectory is not the desired one. However, we have to notice the evident improvement with respect to the not controlled system: the trajectory error in case of control with inaccurate quantization is about 1/5 the error in the not controlled case.

This suggests that if $|\zeta_0+\zeta_1-2\alpha| > |\zeta_0-\zeta_1|$, the detection is better in the inaccurate quantization case. This is equivalent to require that $\alpha < \zeta_1$ in the case $\zeta_0 = 1$ and $\zeta_1, \alpha \in (0,1)$, which is a quite intuitive result.

Thus, in case of inaccurate quantization detection may even be more reliable that in the exact case. On the other hand, inaccurate quantization always produces an error in the trajectory: even in case of exact detection, the recovered trajectory will not be the fault-free one since the Error Function E_k does not decay to zero.

Let us present a few simulations: we exploit again our Flight Control numerical example and we apply our FTC design, in the hypotheses that $\zeta_1 = \frac{1}{2}$ is the quantization failure level considered by the algorithm, while $\alpha = 0.45$ is the real value assumed by $z(t)$ after the failure. Some outcomes are shown in Figure 6.7, where the graphs of the trajectories are reported. The first two graphs represent the case with $\tau = 0.01$. We have already said that in this case the detection (with perfect quantization) is not reliable either before and after the failure. Instead, if the quantization is inaccurate and $\alpha < \zeta_1$ as in our instance, the detection is correct after the failure. This can be appreciated in the second graph: the corresponding trajectory is parallel to the fault-free one, while the trajectory in the case $\alpha = \zeta_1$ is closer to the fault-free one, but very "noisy". The decision about which result is preferable depends on the applications: if we imagine $y(t)$ to be the trajectory of an aircraft, if $\alpha < \zeta_1$, $y(t)$ is in general more distant from the planned trajectory, but the flight for $\alpha = \zeta_1$ seems to be too disturbed.

The third and fourth graph show the instance $\tau = 0.09$, where perfect detection is achieved with $\alpha = \zeta_1 = 0.5$. Detection is correct in both cases, but only if $\alpha = \zeta_1$ we get back to the right trajectory after the failure. However, let us notice that an appreciable improvement is obtained also in the case $\alpha = 0.45, \zeta_1 = 1/2$ if compared to the faulty system without control: the distance between fault-free and "FTC with $\alpha = 0.45, \zeta_1 = 1/2$" trajectories is about 1/5 the distance between fault-free and "faulty, not controlled" trajectories.

6.7 Conclusions

In this chapter, an original Fault Tolerant Control method, based on Information and Decoding Theory, has been introduced. Given a linear system with a disturbance and supposing the disturbance function to be quantized over two levels, the detection task can be tackled by decoding techniques. In particular, we have used the low-complexity One State Algorithm. Its application to a Flight FTC problem has produced satisfactory outcomes even in case of relatively large noise in the data acquisition.

The low-complexity encourages the implementation of this method; moreover, adjusting the sampling time step τ, one can improve its performance, according to the different values of noise and of input f. In some cases, for instance when f is constant, an optimal value of τ can be analytically computed with sufficient precision, where the optimality is intended in terms of trade-off between convergence conditions and amplitude of the deviations. Other arrangements might be obtained changing the values and the number of levels of quantization.

Chapter 7

Conclusions

In this dissertation, we have studied the deconvolution of input/output linear systems with quantized input, where by “quantized” we indicate a digital signal that can assume only a finite number of values.

This is an example of *hybrid* system, the input and the output respectively being digital and analog, which is a fairly common scenario in modern engineering technologies.

Our aim has been the development of suitable deconvolution algorithms to recover the unknown input from noisy measurements of the output, taking account of the quantized nature of the input. This has been motivated by the fact that in the widespread literature on deconvolution the input functions are generally supposed to be regular, thus “classical” algorithms not suitable for hybrid scenarios.

Under the hypothesis that the available data is a sequence of samples picked from the output at regular time intervals, and supposing an exact synchronization between input and output, the problem turns out to be analogous to a typical digital transmission issue. In particular, the convolution can be interpreted as a sort of encoding of the input signal, while deconvolution becomes analogous to a classical *decoding* issue. This suggests to use *decoding* algorithms to perform deconvolution.

The development of such decoding algorithms has been the core of this thesis. Our starting point has been the well-known BCJR method, which is an iterative procedure that implements an optimal decoding. Nevertheless, BCJR has been shown to be too complex for our problem, in particular when long-time transmissions are considered. Thus, our efforts have been focused on the development of some low-complexity, iterative, causal algorithms (derived from the BCJR), which have been proved to be efficient in many situations. More precisely, we have introduced the Causal BCJR, say a version of BCJR processing only past and present information, which is optimal among the causal procedures, but whose complexity linearly increases in time. Afterwards, we have developed the One State Algorithm and the Two State Algorithm, which store and process finite information at any iterative step, which makes them very low-complexity.

These algorithms have been tested in three different scenarios, under the common hypothesis of binary input.

First, we have considered the one-dimensional differentiation problem. The direct system in this case has been represented by a discrete-time dynamical system evolving in $\mathbb{N}$ and the inversion takes account All the proposed algorithms have been implemented in this framework and simulations have been reported. Moreover, a complete theoretical analysis has been developed to evaluated their performance. Using the Ergodic Theory of Markov Processes and the theory of Markov Processes in Random Environments, the performance are computed in terms of a mean square error and for long-time transmissions.

Second, we have generalized our study to linear one-dimensional systems, under stability hypotheses. The mathematical set is very different: in particular, the system evolves in a compact Cantor space. The One State Algorithm has been implemented and shown to be efficient. Again, simulations' and theoretical results have been presented, the latter being based on Iterated Random Functions and Markov Processes Theory.

Afterwards, we have compared the One State Algorithm with the Kalman Filter, which is commonly used for optimal state estimation. The conclusion we have reached is that in our context there are some instances in which the One State Algorithm performs better than the Kalman Filter.

Third, we have studied an application to a multi-dimensional Fault Tolerant Control problem. The goal has been to design a control system that reveals the presence of faults in a process and introduces a suitable compensation in order to minimize the negative effects of the faults, then the system envisages also a feedback. We have implemented the One State Algorithm to detect the faults, such a task consisting in a deconvolution, and shown some simulations and theoretical considerations. In particular, we have observed that increasing the measurement delay we increase the quality of the detection, since more information is collected; on the other hand, a larger delay may cause serious damages, as the compensation is not promptly provided. In conclusion, a suitable trade-off must be achieved, which can be theoretically studied. We have to notice that in this setting, a complete theoretical analysis cannot be developed; nevertheless, fundamental design criteria can be theoretically provided at least for some instances, in terms of the trade-off above mentioned.

In conclusion, in this dissertation we have developed and analyzed deconvolution algorithms for quantized-input linear systems which have the following good features:

1. they are causal, hence they can be used to perform on-line deconvolution;
2. they are low-complexity and easy to implement, with no dramatic loss of efficiency with respect to the optimal BCJR algorithm;
3. they can be theoretically analyzed in the mathematical framework of Markov Process Theory.

Future work may be oriented to develop an exhaustive theoretical analysis in multi-dimensional frameworks and to extend our study to input signals with more than two quantization levels.

Bibliography

[1] F. Abramovich and B. W. Silverman. The vaguelette-wavelet decomposition approach to statistical inverse problems. *Biometrika*, 85:115–129, 1997.

[2] J. Ackermann. Robustness against sensor failures. *Automatica*, 20(2):211–215, 1984.

[3] J. Ackermann. *Sampled-data control systems: analysis and synthesis, robust system design*. Springer, Verlag New York, USA, 1985.

[4] G. Ackerson and K. Fu. On state estimation in switching environments. *IEEE Trans. Autom. Control*, 15(1):10 – 17, February 1970.

[5] J. Aldrich. R. A. Fisher and the making of maximum likelihood 1912–1922. *Statist. Sci.*, 12(3):162–176, 1997.

[6] J. R. Arora. Extraction of signals from noise through wiener filtering. *Pure and Applied Geophysics*, 92:5–18, 1971. 10.1007/BF00874987.

[7] V. K. Arya and H. D. Holden. Deconvolution of seismic data - an overview. *IEEE Trans. on Geosci. Electron.*, 16(2):95–98, apr. 1978.

[8] K. B. Athreya and P. Ney. A new approach to the limit theory of recurrent markov chains. *Trans. Amer. Math. Soc.*, 245:493–501, 1978.

[9] M. Baglietto, G. Battistelli, and L. Scardovi. Active mode observation of switching systems based on set-valued estimation of the continuous state. *Int. J. Robust Nonlinear Control*, 19(14):1521–1540, 2009.

[10] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv. Optimal decoding of linear codes for minimizing symbol error rate. *IEEE Trans. Inf. Theory*, IT-20:284–287, 1974.

[11] M. R. Banham and A. K. Katsaggelos. Digital image restoration. *IEEE Signal Processing Magazine*, 14(2):24–41, mar. 1997.

[12] M. Basseville. Detecting changes in signals and systems–a survey. *Automatica*, 24(3):309 – 326, 1988.

[13] M. Basseville and A. Benveniste. Desgin and comparative study of some sequential jump detection algorithms for digital signals. *IEEE Trans. Acoust. Speech Signal Process.*, 31(3):521 – 535, jun. 1983.

[14] M. Basseville and I. V. Nikiforov. *Detection of abrupt changes: theory and application.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[15] J. W. Bayless and E. O. Brigham. Application of the Kalman Filter to continuous signal restoration. *Geophysics*, 35(1):2–23, 1970.

[16] A. Ben-Israel and T. N. E. Greville. *Generalized Inverses: Theory and Applications.* Springer- Verlag New York Inc, 2 edition, 2003.

[17] S. Benedetto and E. Biglieri. *Principles of Digital Transmission: With Wireless Applications.* Kluwer Academic Publishers, Norwell, MA, USA, 1999.

[18] M. Bertero and P. Boccacci. *Introduction to inverse problems in imaging.* Institute of Physics Publishing, Bristol, 1998.

[19] M. Bertero, P. Boccacci, B. Anconelli, G. Desiderà, M. Carbillet, and H. Lanteri. High-resolution image reconstruction: the case of the large binocular telescope (lbt). *Astronomy with High Contrast Imaging III, eds. M. Carbillet, A. Ferrari, and C. Aime, EAS Publication Series*, (22):35–67, 2006.

[20] M. Bertero and M. Piana. *Complex Systems in Biomedicine*, chapter Inverse problems in biomedical imaging: modeling and methods of solution. Springer, Berlin, 2006.

[21] M. Bertero, T. A. Poggio, and V. Torre. Ill-posed problems in early vision. In *Proceedings of the IEEE*, volume 76, pages 869–889, 1988.

[22] A. Bicchi, A. Marigo, and B. Piccoli. On the reachability of quantized control systems. *IEEE Trans. on Autom. Control*, 4(47):546–563, 2002.

[23] L. Blackmore, S. Rajamanoharan, and B. C. Williams. Active estimation for jump markov linear systems. *IEEE Trans. Autom. Control*, 53(10):2223–2236, 2008.

[24] M. Blanke, M. Kinnaert, J. Lunze, M. Staroswiecki, and J. Schröder. *Diagnosis and Fault-Tolerant Control.* Springer, Verlag New York, USA, 2006.

[25] L. Bogler. Tracking a maneuvering target using input estimation. *IEEE Trans. Aerosp. Electron. Syst.*, 3:298–310, 1987.

[26] M. S. Branicky, V. S. Borkar, and S. K. Mitter. A unified framework for hybrid control: Model and optimal control theory. *IEEE Trans. Autom. Control*, 43:31–45, 1998.

[27] C. L. Byrne. Iterative image reconstruction algorithms based on cross-entropy minimization. *IEEE Trans. Image Process.*, 2(1):96–103, jan. 1993.

[28] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral methods: fundamentals in single domains.* Springer Verlag, Berlin Heidelberg, 2006.

[29] D. Catlin. *Estimation, Control and the Discrete Kalman Filter.* Springer Verlag, Berlin, 1989.

[30] J. Ching, A. C. To, and S. D. Glaser. Microseismic source deconvolution: Wiener filter versus minimax, fourier versus wavelets, and linear versus nonlinear. *The Journal of the Acoustical Society of America*, 115(6):3048–3058, 2004.

[31] E. Y. Chow and A. S. Willsky. Bayesian design of decision rules for failure detection. *IEEE Trans. Aerosp. Electron. Syst.*, AES-20(6):761 –774, nov. 1984.

[32] R. Cogburn. The ergodic theory of markov chains in random environments. *Zeitschrift Fur Wahrscheinl Und Verwandte Gebiete*, 66:109–128, 1984.

[33] R. Cogburn. On products of random stochastic matrices. *Contemporary Mathematics*, 50:199–213, 1986.

[34] D. Commenges. The deconvolution problem: Fast algorithms including the preconditioned conjugate-gradient to compute a map estimator. *IEEE Trans. Autom. Control*, 29(3):229 – 243, mar. 1984.

[35] G. De Nicolao and D. Liberati. Linear and nonlinear techniques for the deconvolution of hormone time-series. *IEEE Trans. Biomed. Eng.*, 40(5):440–455, may. 1993.

[36] G. Demoment and J. Idier. Inverse problems, ill-posed problems. In *Bayesian approach to inverse problems*, Digit. Signal Image Process. Ser., pages 25–40. ISTE, London, 2008.

[37] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977. With discussion.

[38] P. Diaconis and D. Freedman. Iterated random functions. *SIAM Review*, 41:45–76, 1999.

[39] E. A. Domlan, J. Ragot, and D. Maquin. Active mode estimation for switching systems. pages 1143–1148, July 2007.

[40] D. L. Donoho. Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Applied and Computational Harmonic Analysis*, 2:101–126, 1992.

[41] G. J. J. Ducard. *Fault-tolerant Flight Control and Guidance Systems: Practical Methods for Small Unmanned Aerial Vehicles.* Springer, 2009.

[42] M. Dudik. Maximum entropy density estimation and modeling geographic distributions of species, ph.d. thesis, 2007.

[43] J. Ragot E. A. Domlan and D. Maquin. Switching systems: Active mode recognition, identification of the switching law. *Journal of Control Science and Engineering*, 2007.

[44] N. Elia and S. K. Mitter. Stabilization of linear systems with limited information. *IEEE Trans. Autom. Control*, 46(9):1384–1400, sep 2001.

[45] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems.* Kluwer Academic Publishers, 2000.

[46] J. S. Eterno, J. L. Weiss, D. P. Looze, and A. S. Willsky. Design issues for fault tolerant-restructurable aircraft control. volume 24, pages 900–905, 1985.

[47] F. Fagnani, V. Maksimov, and L. Pandolfi. A recursive deconvolution approach to disturbance reduction. *IEEE Trans. Autom. Control*, 49(6):907–921, 2004.

[48] F. Fagnani and L. Pandolfi. A singular perturbation approach to a recursive deconvolution problem. *SIAM J. Control Optim.*, 40(5):1384–1405, 2002.

[49] F. Fagnani and L. Pandolfi. A recursive algorithm for the approximate solution of volterra integral equations of the first kind of convolution type. *Inverse Problems*, 19(1):23–47, 2003.

[50] F. Fagnani and S. Zampieri. Quantized stabilization of linear systems: complexity versus performance. *IEEE Trans. Autom. Control*, 49:1534–1548, 2004.

[51] J. Fan and J. Y. Koo. Wavelet deconvolution. *IEEE Trans. Inf. Theory*, 48(3):734 –747, mar. 2002.

[52] A. Faridani. Introduction to the mathematics of computed tomography. *Inside Out: Inverse Problems and Applications*, 47:1–46, 2003.

[53] M. A. T. Figueiredo and R. D. Nowak. An em algorithm for wavelet-based image restoration. *IEEE Trans. Image Process.*, 12(8):906–916, aug. 2003.

[54] S. R. Foguel. *The ergodic theory of Markov Processes.* Van Nostrand, Princeton, 1969.

[55] E. I. Fredholm. Sur une classe d'equations fonctionnelles. *Acta Mathematica*, 27:365–390, 1903.

[56] B. R. Frieden. Restoring with maximum likelihood and maximum entropy. *J. Opt. Soc. Am.*, 62(4):511–518, 1972.

[57] F. Gamboa and E. Gassiat. Bayesian methods and maximum entropy for ill-posed inverse problems. *Ann. Statist.*, 25(1):328–350, 1997.

[58] A. Gilbert and W. Keller. Deconvolution with wavelets and vaguelettes. *Journal of Geodesy*, 74:306–320, 2000. 10.1007/s001900050288.

[59] A. Graps. An introduction to wavelets. *IEEE Comput. Science Eng.*, 2(2):50–61, 1995.

[60] P. J. Green. Bayesian reconstructions from emission tomography data using a modified em algorithm. *IEEE Trans. Med. Imag.*, 9(1):84–93, mar. 1990.

[61] M. S. Grewal and A. P. Andrews. Applications of kalman filtering in aerospace 1960 to the present [historical perspectives]. *IEE Control Systems Magazine*, 30(3):69–78, jun. 2010.

[62] S. F. Gull and T. J. Newton. Maximum entropy tomography. *Appl. Opt.*, 25(1):156–160, 1986.

[63] S. F. Gull and J. Skilling. Maximum entropy method in image processing. *IEE Proceedings- Part F: Communications, Radar and Signal Processing,*, 131(6):646–659, oct. 1984.

[64] J. Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton Univiersity Bull.*, 13:49–52, 1902.

[65] J. Hadamard. *Lectures on Cauchy's problem in linear partial differential equations.* Yale University Press, 1923.

[66] C. Hajiyev and F. Caliskan. Fault detection in flight control systems via innovation sequence of kalman filter. In S. Tzafestas and G. Schmidt, editors, *Progress in system and robot analysis and control design*, volume 243 of *Lecture Notes in Control and Information Sciences*, pages 63–74. Springer Berlin / Heidelberg, 1999. 10.1007/BFb0110534.

[67] P. C. Hansen. Analysis of discrete ill-posed problems by means of the l-curve. *SIAM Rev.*, 34(4):561–580, 1992.

[68] O. Hernández-Lerma and J. B. Lasserre. *Markov Chains and Invariant Probabilities.* Birkhauser-Verlag, Basel, 2003.

[69] J. Hutchinson. Fractals and self-similarity. *Indiana Univ. Math. J.*, 30(5):713–747, 1981.

[70] M. Iosifescu. Iterated function systems. a critical survey. *Mathematical Reports*, 11(3):181—229, 2009.

[71] R. Isermann. *Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance.* Springer, 2006.

[72] V. K. Ivanov. Integral equations of the first kind and approximate solution of the inverse problem of potential theory. *Dokl. Akad. Nauk SSSR*, 142:998–1000, 1962.

[73] V. K. Ivanov. On linear problems which are not well-posed. *Dokl. Akad. Nauk SSSR*, 145:270–272, 1962.

[74] S. F. Jarner and R. L.Tweedie. Locally contracting iterated functions and stability of markov chains. *J. Appl. Probab.*, 38(2):494–507, 2001.

[75] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106(4):620–630, May 1957.

[76] J. Jiang. Fault-tolerant control systems - an introductory overview. *Automatica SINCA*, 31(1):161–174, 2005.

[77] M. Jiang, L. Xia, G. Shou, Q. Wei, F. Liu, and S. Crozier. Effect of cardiac motion on solution of the electrocardiography inverse problem. *IEEE Trans. Biomed. Eng.*, 56(4):923–931, apr. 2009.

[78] S. I. Kabanikhin. Definitions and examples of inverse and ill-posed problems. *J. Inv. Ill-Posed Problems*, 16(3):267–282, 2008.

[79] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[80] R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction theory. *Trans. ASME Ser. D. J. Basic Engrg.*, 83:95–108, 1961.

[81] J. B. Keller. Inverse problems. *Amer. Math. Monthly*, 83(2):107–118, 1976.

[82] A. Kirsch. *An introduction to the mathematical theory of inverse problems.* Springer-Verlag New York, Inc., 1996.

[83] J. J. Kormylo and J. M. Mendel. Maximum-likelihood seismic deconvolution. *IEEE Trans. Geosci. Remote Sens.*, GE-21(1):72–82, jan. 1983.

[84] A. V. Kryazhimskii and Y. S. Osipov. *Inverse Problems for Ordinary Differential Equations: Dynamical Solutions.* Gordon and Breach, London, 1995.

[85] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statistics*, 22:79–86, 1951.

[86] T. L. Lai and J. Z. Shan. Efficient recursive algorithms for detection of abrupt changes in signals and control systems. *IEEE Trans. Autom. Control*, 44(5):952–966, may 1999.

[87] G. Le Besnerais, J.-F. Bercher, and G. Demoment. A new look at entropy for solving linear inverse problems. *IEEE Trans. Inf. Theory*, 45(5):1565–1578, jul. 1999.

[88] H. Lee and M.-J. Tahk. Generalized input-estimation technique for tracking maneuvering targets. *IEEE Trans. Aerosp. Electron. Syst.*, 35:1388–1402, October 1999.

[89] X. R. Li and V. P. Jilkov. Survey of maneuvering target tracking-part i: dynamic models. *IEEE Transactions on Aerospace and Electronic Systems*, 39(4):1333–1364, 2003.

[90] D. Magill. Optimal adaptive estimation of sampled stochastic processes. *IEEE Trans. on Automat. Contr.*, 10:434–439, 1965.

[91] L. A. McGee and S. F. Schmidt. *Discovery of the Kalman filter as a practical tool for aerospace and industry.* National Aeronautics and Space Administration, Ames Research Center, Moffett Field, Calif., 1985.

[92] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* Wiley Series in Probability and Statistics. Wiley, 2nd edition, 2008.

[93] J. M. Mendel. Minimum-variance deconvolution. *IEEE Trans. Geosci. Remote Sens.*, GE-19(3):161–171, jul. 1981.

[94] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability.* Springer-Verlag, London, 1993.

[95] R. Molina, J. Nunez, F.J. Cortijo, and J. Mateos. Image restoration in astronomy: a bayesian perspective. *IEEE Signal Processing Magazine*, 18(2):11–29, mar. 2001.

[96] V. A. Morozov. *Methods for Solving Incorrectly Posed Problems.* Springer-Verlag, 1984.

[97] J. G. Nagy, R. J. Plemmons, and T. C. Torgersen. Iterative image restoration using approximate inverse preconditioning. *IEEE Trans. Image Process.*, 5(7):1151–1162, jul. 1996.

[98] M. R. Napolitano, C. D. Neppach, V. Casdorph, Naylor S., M. Innocenti, and G. Silvestri. A neural-network-based scheme for sensor failure detection, identification, and accommodation. *AIAA Journal of Guidance, Control, and Dynamics*, 18(6), 1995.

[99] M. Z. Nashed. *Generalized Inverses and Applications.* Academic Press, New York, 1976.

[100] K. Nawrotzki. Discrete open systems or markov chains in a random environment i. *Elektronische Informationsverarbeitung und Kybernetik*, 17(11/12):569–599, 1981.

[101] K. Nawrotzki. Discrete open systems or markov chains in a random environment ii. *Elektronische Informationsverarbeitung und Kybernetik*, 18(1/2):83–98, 1982.

[102] R. Neelamani, H. Choi, and R. Baraniuk. Wavelet-based deconvolution for ill-conditioned systems. volume 6, pages 3241–3244 vol.6, mar. 1999.

[103] R. Neelamani, Hyeokho Choi, and R. Baraniuk. Forward: Fourier-wavelet regularized deconvolution for ill-conditioned systems. *Signal Processing, IEEE Transactions on*, 52(2):418–433, feb. 2004.

[104] I. V. Nikiforov. A generalized change detection problem. *IEEE Trans. Inf. Theory*, 41(1):171–187, jan. 1995.

[105] I. V. Nikiforov. A simple recursive algorithm for diagnosis of abrupt changes in random signals. *IEEE Trans. Inf. Theory*, 46(7):2740–2746, nov 2000.

[106] D. Noll. Restoration of degraded images with maximum entropy. *J. of Global Optimization*, 10(1):91–103, 1997.

[107] N. R. Pal and S. K. Pal. Entropy: a new definition and its applications. *IEEE Trans. Syst. Man Cybern.*, 21(5):1260–1270, sep. 1991.

[108] E. Pantin and J.-L. Starck. Deconvolution of astronomical images using the multiscale maximum entropy method. *Astronomy and Astrophysics, Suppl. Ser*, 315:31–5, 1996.

[109] P. Park, Y. J. Choi, and S. W. Yun. Eliminating effect of input quantisation in linear systems. *Electronics Letters*, 44(7):456–457, 27 2008.

[110] R. Patton. Fault-tolerant control: the 1997 situation. In *Proc. of the 3rd IFAC Symp. on Fault Detection, Supervision and Safety for Technical Processes*, volume 2, pages 1033–1055, 1997.

[111] R. Patton and J. Chen. Observer-based fault detection and isolation: robustness and applications. *Control Eng. Pract.*, 5(5):671–682, 1997.

[112] K. L. Peacock and S. Treitel. Predictive deconvolution: Theory and practice. *Geophysics*, 34(2):155–169, 1969.

[113] D. L. Phillips. A technique for the numerical solution of certain integral equations of the first kind. *J. Assoc. Comput. Mach.*, 9:84–97, 1962.

[114] B. Picasso and A. Bicchi. On the stabilization of linear systems under assigned i/o quantization. *IEEE Trans. Automat. Control*, 52(10):1994–2000, 2007.

[115] G. Pillonetto and B. Bell. Deconvolution of nonstationary physical signals: a smooth variance model for insulin secretion rate. *Inverse Problems*, 20:367–383, 2004.

[116] G. Pillonetto, G. Sparacino, and C. Cobelli. Reconstructing insulin secretion rate after a glucose stimulus by an improved stochastic deconvolution method. *IEEE Trans. Biomed. Eng.*, 48(11):1352–1354, nov. 2001.

[117] K. N. Plataniotis, S. K. Katsikas, D. G. Lainiotis, and A. N. Venetsanopoulos. Optimal seismic deconvolution: distributed algorithms. *IEEE Trans. Geosci. Remote Sens.*, 36(3):779–792, may. 1998.

[118] S. Prasad and A. K. Mahalanabis. Adaptive filter structures for deconvolution of seismic signals. *IEEE Trans. Geosci. Remote Sens.*, 18(3):267–273, jul. 1980.

[119] A. J. Pullan, L. K. Cheng, M. P. Nash, C. P. Bradley, and D. J. Paterson. Noninvasive electrical imaging of the heart: Theory and model development. *Annals of Biomedical Engineering*, 29:817–836, 2001. 10.1114/1.1408921.

[120] J. Qu and T. L. Teng. Recursive stochastic deconvolution in the estimation of earthquake source parameters: synthetic waveforms. *Physics of The Earth and Planetary Interiors*, 86(4):301– 327, 1994.

[121] S. T. Rachev. *Probability Metrics and the Stability of Stochastic Models.* Wiley, New York, 1991.

[122] K. Rajan, L. M. Patnaik, and J. Ramakrishna. Linear array implementation of the em algorithm for pet image reconstruction. *IEEE Trans. Nuclear Science*, 42(4):1439–1444, aug. 1995.

[123] K. Rajan, L.M. Patnaik, and J. Ramakrishna. High-speed computation of the em algorithm for pet image reconstruction. *Nuclear Science, IEEE Transactions on*, 41(5):1721–1728, oct. 1994.

[124] I. Rhodes. A tutorial introduction to estimation and filtering. *IEEE Trans. Autom. Control*, 16(6):688–706, dec. 1971.

[125] R. B. Rice. Inverse convolution filters. *Geophysics*, 27:4–18, February 1962.

[126] T. Richardson and R. Urbanke. *Modern Coding Theory.* Cambridge University Press, March 2008.

[127] W. H. Richardson. Bayesian-based iterative method of image restoration. *J. Opt. Soc. Am.*, 62(1):55–59, 1972.

[128] E. A. Robinson. Predictive decomposition of seismic traces. *Geophysics*, 22(4):767–778, 1957.

[129] S. M. Ross. *Introduction to Probability Models, Ninth Edition.* Academic Press, Inc., Orlando, FL, USA, 2006.

[130] W. Rudin. *Real and complex analysis, 3rd ed.* McGraw-Hill, Inc., New York, NY, USA, 1987.

[131] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication.* The University of Illinois Press, Urbana, Ill., 1949.

[132] L. A. Shepp and Y. Vardi. Maximum likelihood reconstruction for emission tomography. *IEEE Trans. Med. Imag.*, 1(2):113–122, oct. 1982.

[133] C. S. Sims and M. R. D'Mello. Adaptive deconvolution of seismic signals. *IEEE Trans. Geosci. Electron.*, 16(2):99–103, apr. 1978.

[134] A. Singer. Estimating optimal tracking filter performance for manned maneuvering targets. *IEEE Trans. Aerosp. Electron. Syst.*, AES-6(4):473–384, 1970.

[135] R. Singer and P. Frost. On the relative performance of the kalman and wiener filters. *IEEE Trans. Autom. Control*, 14(4):390–394, aug. 1969.

[136] H. W. Sorenson. Least-squares estimation: from gauss to kalman. *IEEE Spectrum*, 7(7):63–68, jul. 1970.

[137] G. Sparacino and C. Cobelli. A stochastic deconvolution method to reconstruct insulin secretion rate after a glucose stimulus. *IEEE Trans. Biomed. Eng.*, 43(5):512–529, may. 1996.

[138] J.-L. Starck and F. Murtagh. Image restoration with noise suppression using the wavelet transform. *Astronomy and Astrophysics*, 288(1):342–348, 1994.

[139] J-.L. Starck, M. K. Nguyen, and F. Murtagh. Wavelets and curvelets for image deconvolution: a combined approach. *Signal Processing*, 83(10):2279 – 2283, 2003.

[140] J.-L. Starck, E. Pantin, and F. Murtagh. Deconvolution in astronomy: A review. *Publications of the Astronomical Society of the Pacific*, 114:1051–1069, 2002.

[141] J.-L. Starck, E. Pantin, and F. Murtagh. *Deconvolution and Blind Deconvolution in Astronomy*, pages 277–317. 2007.

[142] M. Steinberg. Historical overview of research in reconfigurable flight control. volume 219, pages 263–276, 2005.

[143] D. Steinsaltz. Locally contractive iterated function systems. *Ann. Probab.*, 27(4):1952–1979, 1999.

[144] O. Stenflo. Ergodic theorems for markov chains represented by iterated function systems. *Bull. Polish Acad. Sci. Math*, 49(1):27–43, 2001.

[145] R. F. Stengel. Intelligent failure-tolerant control. *IEEE Control Systems Magazine*, 11(4):14–23, June 1991.

[146] D. W. Stroock. *An Introduction to Markov Processes.* Springer-Verlag, Berlin, 2005.

[147] U. Sumbul, J.M. Santos, and J.M. Pauly. A practical acceleration algorithm for real-time imaging. *IEEE Trans. Med. Imag.*, 28(12):2042–2051, dec. 2009.

[148] A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.*, 4:1035–1038, 1963.

[149] A. N. Tikhonov and V. Y. Arsenin. *Solutions of ill-posed problems.* V. H. Winston and Sons, New York, 1977.

[150] D. M. Titterington. On the iterative image space reconstruction algorthm for ect. *IEEE Trans. Med. Imag.*, 6(1):52–56, mar. 1987.

[151] V. Torre and T. A. Poggio. On edge detection. *IEEE Trans. Pattern Anal. Mach. Intell*, PAMI-8(2):147–163, mar. 1986.

[152] R. L. Tweedie. Markov chains: Structure and applications. 19:817– 851, 2001.

[153] Y. Vardi, L. A. Shepp, and L. Kaufman. A statistical model for positron emission tomography. *J. Amer. Statist. Assoc.*, 80(389):8–37, 1985. With discussion.

[154] Y. Vardi and C-H. Zhang. *Reconstruction of binary images via the EM algorithm*, pages 297–316. Appl. Numer. Harmon. Anal. Birkhäuser Boston, Boston, MA, 1999.

[155] V. Venkatasubramanian, H. Leung, and B. Moorman. An interacting multiple-model-based abrupt change detector for ground-penetrating radar. *IEEE Geosci. Remote Sens. Letters*, 4(4):634–638, oct. 2007.

[156] N. Viswanadham and R. Srichander. Fault detection using unknown input observers. *Control Theory Adav. Technol.*, 3:91–101, 1987.

[157] G. Wahba. Practical approximate solutions to linear operator equations when the data are noisy. *SIAM J. Numer. Anal.*, 14(4):651–667, 1977.

[158] G. G. Walter and X. Shen. Deconvolution using meyer wavelets. *J. Integral Equations Appl.*, 11(4), 1999.

[159] P. Walters. *An Introduction to Ergodic Theory.* Springer-Verlag, New York, 2000.

[160] S. J. Wernecke and L. R. D'Addario. Maximum entropy image reconstruction. *IEEE Trans. Computer*, C-26(4):351–364, apr. 1977.

[161] N. Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series. With Engineering Applications.* The Technology Press of the Massachusetts Institute of Technology, Cambridge, Mass, 1949.

[162] A. S. Willsky. A survey of design methods for failure detection in dynamic systems. *Automatica–J. IFAC*, 12(6):601–611, 1976.

[163] A. S. Willsky and H. Jones. A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *IEEE Trans. Autom. Control*, 21(1):108–112, 1976.

[164] Z. S. Yao and R. G. Roberts. A practical regularization for seismic tomography. *Geophysical Journal International*, 138(2):293–299.

[165] D. Ye and G.-H. Yang. Adaptive fault-tolerant tracking control against actuator faults with application to flight control. *IEEE Trans. Control Syst. Techn.*, 14(6):1088–1096, 2006.

[166] J.-S. Yee, J. L. Wang, and B. Jiang. Actuator fault estimation scheme for flight applications. *Journal of Dynamic Systems, Measurement, and Control*, 124(4):701–704, 2002.

[167] Y. Zhang and J. Jiang. Bibliographical review on reconfigurable fault-tolerant control systems. *Annual Reviews in Control*, 32(2):229–252, December 2008.

[168] J. Zhou, J.-L. Coatrieux, A. Bousse, H. Shu, and L. Luo. A bayesian map-em algorithm for pet image reconstruction using wavelet transform. *IEEE Trans. Nuclear Science*, 54(5):1660–1669, oct. 2007.