Multidimensional inverse problems in imaging and identification using low-complexity models, optimal mass transport, and machine learning

AXEL RINGH

Doctoral Thesis
KTH Royal Institute of Technology
School of Engineering Sciences
Department of Mathematics
Divisions of Optimization and Systems Theory
Stockholm, Sweden 2018

ISBN 978-91-7873-050-6
TRITA-SCI-FOU 2018:53

Department of Mathematics
KTH Royal Institute of Technology
100 44 Stockholm, Sweden

Akademisk avhandling som med tillstånd av Kungliga Tekniska Högskolan framläggs till offentlig granskning för avläggande av teknologie doktorsexamen, torsdagen den 17 januari 2019 klockan 13:00 i sal F3, Lindstedtsvägen 26, Kungliga Tekniska Högskolan, Stockholm.

© Axel Ringh, 2018

Print: Universitetsservice US-AB, Stockholm, 2018

To Camille
My better half

To Emil
My other better half

Abstract

This thesis, which mainly consists of six appended papers, primarily considers a number of inverse problems in imaging and system identification.

In particular, the first two papers generalize results for the rational covariance extension problem from one to higher dimensions. The rational covariance extension problem stems from system identification and can be formulated as a trigonometric moment problem, but with a complexity constraint on the sought measure. The papers investigate a solution method based on variational regularization and convex optimization. We prove the existence and uniqueness of a solution to the variational problem, both when enforcing exact moment matching and when considering two different versions of approximate moment matching. A number of related questions are also considered, such as well-posedness, and the theory is illustrated with a number of examples.

The third paper considers the maximum delay margin problem in robust control: To find the largest time delay in a feedback loop for a linear dynamical system so that there still exists a single controller that stabilizes the system for all delays smaller than or equal to this time delay. A sufficient condition for robust stabilization is recast as an analytic interpolation problem, which leads to an algorithm for computing a lower bound on the maximum delay margin. The algorithm is based on bisection, where positive semi-definiteness of a Pick matrix is used as selection criteria.

Paper four investigate the use of optimal transport as a regularizing functional to incorporate prior information in variational formulations for image reconstruction. This is done by observing that the so-called Sinkhorn iterations, which are used to solve large scale optimal transport problems, can be seen as coordinate ascent in a dual optimization problem. Using this, we extend the idea of Sinkhorn iterations and derive a iterative algorithm for computing the proximal operator. This allows us to solve large-scale convex optimization problems that include an optimal transport term.

In paper five, optimal transport is used as a loss function in machine learning for inverse problems in imaging. This is motivated by noise in the training data which has a geometrical characteristic. We derive theoretical results that indicate that optimal transport is better at compensating for this type of noise, compared to the standard 2-norm, and the effect is demonstrated in a numerical experiment.

The sixth paper considers using machine learning techniques for solving large-scale convex optimization problems. We first parametrizes a family of algorithms, from which a new optimization algorithm is derived. Then we apply machine learning techniques to learn optimal parameters for given families of optimization problems, while imposing a fixed number of iterations in the scheme. By constraining the parameters appropriately, this gives learned optimization algorithms with provable convergence.

Keywords: inverse problems, convex optimization, variational regularization, trigonometric moment problems, optimal mass transport, computed tomography, machine learning, analytic interpolation, delay systems

Sammanfattning

Denna avhandling, som huvudsakligen består av de sex bifogade artiklarna, berör ett antal olika inversa problem med tillämpning inom bildrekonstruktion och systemidentifiering.

The två första artiklarna generaliserar resultat från litteraturen gällande det rationella kovariansutvidgningsproblemet, från det en-dimensionella fallet till det fler-dimensionella fallet. Det rationella kovariansutvidgningsproblemet har sitt ursprung inom systemidentifiering och kan formuleras som ett trigonometriska momentproblem. Momentproblemet är dock av icke-klassisk karaktär, eftersom det sökta måttet har ett bivillkor som begränsar dess komplexitet. Papperna undersöker olika metoder för att lösa problemet, metoder som alla bygger på variationell regularisering och konvex optimering. Vi undersöker både exakt och approximativ kovariansmatchning, och huvudresultaten är bevis av existens och unikhet vad gäller lösning till dessa olika problem. Artiklarna undersöker även ett antal relaterade frågor, så som välställdhet av problemen, och teorin är också illustrerad med ett antal olika exempel och tillämpningar.

Det tredje pappret behandlar ett problem inom robust reglering för linjära system: ett systems tidsfördröjningsmarginal. Tidsfördröjningsmarginalen är den längsta tidsfördröjning ett återkopplat linjärt dynamiskt system kan ha så att det fortfarande finns en enda regulator som stabiliserar systemet för alla tidsfördröjningar som är kortare. Artikeln undersöker ett tillräckligt villkor, och formulerar om detta som ett analytiskt interpolationsproblem. Detta leder till en algoritm för att beräkna en undre gräns för tidsfördröjningsmarginalen. Algoritmen bygger på intervallhalveringsmetoden, och använder Pick-matrisens teckenkaraktär som urvalskriterium.

Artikel fyra undersöker användandet av optimal masstransport som regulariseringsfunktion vid bildrekonstruktion. Idén är att använda optimal masstransport som ett avstånd mellan bilder, och på så vis kunna inkorporera förhandsinformation i rekonstruktionen. Mer specifikt görs detta genom att utvidga de så kallade Sinkhorn-iterationerna, som används för att beräkna lösningen till optimal masstransportsproblemet. Vi åstadkommer denna utvidgning genom att observera att Sinkhorn-iterationerna är ekvivalent med koordinatvis optimering i ett dualt problem. Med hjälp av detta tar vi fram en algoritm för att beräkna proximal-operatorn till optimal masstransportproblemet, vilket gör att vi kan lösa storskaliga optimeringsproblem som innehåller en sådan term.

I femte artikeln använder vi istället optimal masstransport som kostnadsfunktion vid träning av neurala nätverk för att lösa inversa problem inom bildrekonstruktion. Detta motiveras genom tillämpningar där bruset i data är av geometrisk karaktär. Vi presenterar teoretiska resultat som indikerar att optimal masstransport är bättre på att kompensera för denna typ av brus än till exempel 2-normen. Denna effekt demonstreras också i ett numerisk experiment.

Det sjätte pappret undersöker användandet av maskininlärning för att lösa storskaliga optimeringsproblem. Detta görs genom att först parametrisera en

familj av algoritmer, ur vilken vi också härleder en ny optimeringsmetod. Vi använder sedan maskininlärning för att ta fram optimala parametrar i denna familj av algoritmer, givet en viss familj av optimeringsproblem samt givet att bara ett fixt antal iterationer får göras i lösningsmetoden. Genom att begränsa sökrymden för algoritmparametrarna kan vi också garantera att den inlärda metoden är en konvergent optimeringsalgoritm.

Nyckelord: inversa problem, konvex optimering, variationell regularisering, trigonometriska momentproblem, optimal masstransport, datortomografi, maskininlärning, analytisk interpolation, system med tidsfördröjning

Acknowledgments

This work would not have been possible without the support from many people around me. To all of these persons I would like to extend a thank you.

First and foremost, I would like to thank my supervisor Johan Karlsson, who guided me through this journey. Thank you for your unconditional support, your genuine interest for the subject(s), your desire to learn new things, your unselfish attitude when it comes to sharing your knowledge with others, and your endless patience during our many and long discussions. It has been a pleasure to work with you, and to learn from you.

Second, I would like to direct a special thank you to Anders Lindquist, my co-supervisor. Thank you for introducing me to many of the topics in this thesis, and thank you for sharing both your great mathematical knowledge and your experience with the academic environment. I have also greatly enjoyed the stays in Shanghai, together with Johan, including the many evenings with discussions about various subjects. Thank you for your support, and for what you have taught me.

To all my other collaborators: Ozan Öktem, Jonas Adler, Sebastian Banert, Silun Zhang, and Xiaoming Hu. Thank you for great cooperation, and for teaching me so many things in a variety of different areas.

I would also like to thank the faculty at the division: Per Enqvist, Anders Forsgren, Xiaoming Hu, and Krister Svanberg. And all the PhD students and Postdocs that have been here during this time. For everything I have learned from you and with you, and for the open environment with "högt i tak" that allows for discussions on both professional and private matters.

Thank you Ozan Öktem, Jonas Adler, and Holger Kohr, for teaching me about inverse problems, programming, and many other things. And thank you also to the rest of the "inverse problems/imaging" group.

I would also like to thank many other persons at the department, including my fellow students who have created a nice atmosphere at the office. And thank you to the administration, especially to Ann-Britt, Diana, and Irene, for help with countless questions and issues.

Thank you also to the APICS/FACTAS team at INRIA Sophia Antipolis, for hosting me: For everything I learned during my stay, and for nice moments both during and outside of working hours. Thank you Fabien, Juliette, Laurent, Martine, and Syvain, and thank you Adam, Gibin, Konstantinos, and Sebastien.

There are also many persons outside of the professional sphere that deserve a thank you: To my parents, my brothers, and the rest of my family, for unconditional support on this journey. To Martin, Emil, and Björn, my three musketeers and Dalton brothers, for keeping me occupied outside of work with dinners, discussions, games, and other things. To Nabila, Xin, Evelina, Filippa, Jennifer, Agnes, and Emil. And to Camille, for standing by my side in all the ups and downs along this road.

Stockholm, November 2018
Axel Ringh

Table of Contents

Abstract **v**

Acknowledgments **ix**

Table of Contents **xi**

Part I: Introduction

1 Introductory overview **3**

2 Background **5**
2.1 Linear dynamical systems, analytic functions, and feedback control . 5
2.2 Stochastic processes and spectral estimation 15
2.3 Convex optimization and duality 22
2.4 Inverse problems, ill-posedness, and variational regularization 33
2.5 A note on machine learning and neural networks 40

3 Summary of papers **47**

Part II: Research Papers

A Multidimensional rational covariance extension **65**
A.1 Introduction 67
A.2 Main results 70
A.3 The multidimensional RCEP 75
A.4 Well-posedness and counter examples 80
A.5 Logarithmic moments and cepstral matching 83

A.6 The circulant problem . . . 87
A.7 Application to system identification . . . 90
A.8 Application to texture generation . . . 93
A.9 Application to image compression . . . 98
A.10 Appendix . . . 103

B Multidimensional rational covariance extension with approximate covariance matching 115
B.1 Introduction . . . 117
B.2 Rational covariance extension with exact matching . . . 121
B.3 Approximate covariance extension with soft constraints . . . 123
B.4 On the well-posedness of the soft-constrained problem . . . 127
B.5 Tuning to avoid a singular part . . . 129
B.6 Covariance extension with hard constraints . . . 132
B.7 On the equivalence between the two problems . . . 138
B.8 Estimating covariances from data . . . 141
B.9 Application to spectral estimation . . . 142
B.10 Application to system identification and texture reconstruction . . . 144
B.11 Conclusions . . . 147
B.12 Appendix . . . 147

C Lower bounds on the maximum delay margin by analytic interpolation 157
C.1 Introduction . . . 159
C.2 The delay margin problem . . . 160
C.3 Formulating and solving (C.2.6) using analytic interpolation . . . 162
C.4 Improving the lower bound using a constant shift . . . 164
C.5 Numerical example . . . 166
C.6 On the control implementation . . . 168
C.7 Conclusions and future directions . . . 169
C.8 Appendix . . . 169

D Generalized Sinkhorn iterations for regularizing inverse problems using optimal mass transport 173
D.1 Introduction . . . 175
D.2 Background . . . 177
D.3 The dual problem and generalized Sinkhorn iterations . . . 181
D.4 Inverse problems with optimal mass transport priors . . . 186
D.5 Application in computerized tomography . . . 190
D.6 Concluding remarks and further directions . . . 197
D.7 Appendix 1: Proof of Proposition D.3.4 . . . 198
D.8 Appendix 2: Proof of Theorem D.3.10 . . . 199
D.9 Appendix 3: Connection with method based on Dykstra's algorithm 200
D.10 Appendix 4: Parameters in the numerical examples . . . 202

E Learning to solve inverse problems using Wasserstein loss **209**
E.1 Introduction 211
E.2 Background 212
E.3 Learning a reconstruction operator using Wasserstein loss 216
E.4 Implementation and evaluation 218
E.5 Conclusions and future work 221
E.6 Appendix 1: Deferred proofs 221
E.7 Appendix 2: OMT for unbalanced marginals via Sinkhorn iterations 223
E.8 Appendix 3: Metric property of the cost function 225

F Data-driven nonsmooth optimization **231**
F.1 Introduction 233
F.2 Background 235
F.3 A new family of optimization solvers 239
F.4 Learning an optimization solver 251
F.5 Application to inverse problems and numerical experiments 256
F.6 Conclusions and future work 264

"We live on an island surrounded by a sea of ignorance. As our island of knowledge grows, so does the shore of our ignorance."

— John Horgan, attributed to John Archibald Wheeler

"We are just an advanced breed of monkeys on a minor planet of a very average star. But we can understand the Universe. That makes us something very special."

— Stephen Hawking

Part I: Introduction

1. Introductory overview

The human desire to understand the unknown has undoubtedly been an enabler for her development, and is also key to the creation and progression of science. An increased understanding of our surrounding has lead to descriptions of our environment, and these descriptions have been used to make predictions of the future. In order to make these descriptions and predictions more precise, mathematics have been used more and more extensively. This thesis deals with subjects related to mathematical modeling, model selection, and computational methods for these tasks. In particular, it deals with a number of questions and issues related to *control theory*, *system identification*, and *inverse problems*.

System identification and inverse problems are two categories of problems that deals with the extraction of information from an object which is not directly observable. Both of them can be described using the schematic representation in Figure 1.1. Here, F is simply a mapping that maps the input x to the output y. In the setting of system identification we can measure both x and y, however we do not know the mapping F. The goal is thus to recover a mathematical model of the *system* F, or an appropriate approximation thereof, from observations of the input x and the output y. System identification problems occur in many different areas, in particular in the field of control theory [73, 104, 72]. Expressed loosely, control theory deals with analysis and design of systems (feedback) that work in an autonomous fashion [32, 28, 60]. It is also in this context that this thesis deals with system identification.

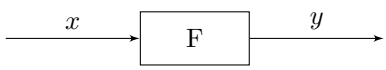


Figure 1.1: Schematic figure of a system F, mapping from input x to output y.

In the setting of inverse problems, on the other hand, we have a mathematical model for F and we also have access to measurements of the output y. However, in this setting we cannot directly measure the input x. The goal is thus to recover information about or completely reconstruct this unknown quantity from the information contained in the indirect observations $y = F(x)$ [37, 61]. Such problems

arise in several areas of science and engineering, and in particular in imaging such as *computed tomography* (CT) [81] and *magnetic resonance imaging* (MRI) [16]. These are noninvasive imaging modalities, and in the former the interior structure of an object under investigation is reconstructed from measurements of the average decay in intensity of X-ray beams sent through the object.

In many areas, including control theory, system identification, and inverse problem, the concept of something being *optimal* often occurs. This means that it is as good as it can be, given a certain criteria to measure "goodness". *Optimization* can then be said to be the theory to ensure that something is optimal and to derive algorithms for how to find the optimal points. The theory is well-established and provides many useful tools for other disciplines, especially when the problems are so-called convex optimization problems [75, 96, 8]. Many of the techniques and tools used in this thesis come from convex optimization.

As an example of this we can consider both system identification and inverse problems. The traditional way of tackling both of these problems have been using a so-called *model driven approach*. For system identification this has often meant using expert knowledge to specify a certain class of mappings F and then finding the mapping in this class which is optimal, given some suitable definition of "goodness", cf. [104, Sec. 4.1]. For inverse problems, in the subfield of variational regularization we look for a solution that minimizes data misfit, while also introducing a penalization for undesirable reconstructions [37, Chp. 5]. In both cases this imposes a prior on the unknown entity, a prior which is either explicitly or implicitly hand-crafted by the expert user and which hopefully reflects the real distribution of the unknown entity. However, the true priors are normally complex and not possible to state explicitly, e.g., there is no explicit expression for the distribution of images of cross-sections of human abdomen. Thus it is often hard to hand-craft model classes and regularizers. Therefore, so-called *data-drive approaches* has been suggested, especially under the name *machine learning* [10] and *deep learning* [48]. In this setting, the idea is to not hand-craft the prior, but instead try to "learn" it directly from data.

I have had the fortune to work on many different problems related all of these topics throughout my time as a PhD student, and this is what is presented in the this thesis. The thesis consists of two parts. The main scientific contribution is found in the second part, which contains a collection of appended papers. However, before that, the first part of the thesis contains another two chapters. The first of these chapters contains background material, in which the notions introduced above are made more precise. The second of these chapters contains a summary of the appended papers, and also clarifies the author's contribution in each case.

2. Background

Here we introduce some preliminary material. This is done to facilitate for the reader by introducing concepts needed in the appended papers in Part II. However, the material presented here is also intended to put the appended papers in a somewhat bigger context.

Notation The notation used is mostly standard, however to reduce the risk for confusion we will briefly introduce some of it here. To this end, for a set A we will denote the closure by $\bar{A}$ and the complement by A^C. Moreover, we introduce the two symbols $-\infty$ and $+\infty$ (the latter often just denoted ∞) which are such that $-\infty < x < +\infty$ for any $x \in \mathbb{R}$ and define the extended real numbers as $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$.[1] Finally, for $z \in \mathbb{C}$ we let z^* denote the complex conjugate, $\mathbb{C}_+ := \{z \in \mathbb{C} \mid z = a + ib,\ a, b \in \mathbb{R},\ a > 0\}$ denote the open complex right half plane, $\mathbb{D} := \{z \in \mathbb{C} \mid |z| < 1\}$ denote the open unit disc, and $\mathbb{T} := \{z \in \mathbb{C} \mid |z| = 1\}$ denote the unit circle.

2.1 Linear dynamical systems, analytic functions, and feedback control

Dynamical systems are used to model many phenomena in the world. Examples are models for different mechanical and electrical systems [60, Sec. 1.2], but also models for population dynamics and epidemics [76, Sec. 10.3 and 10.4]. The simplest type of dynamical systems are linear dynamical systems, yet these are powerful enough to mathematically model the behavior of many real-world systems. Examples of such systems are basic electrical circuits [28, Sec. 2.5.1], simple mechanical systems [28, Ex. 2.6 and 2.7], and mixing problems [28, Ex. 2.10]. Moreover, nonlinear dynamical systems are often analyzed by linearizing them [93, Sec. 5.4] [60, Sec. 4.3]. This section contains an overview of the theory for signals, systems, and feedback control, for linear dynamical systems. The exposition takes an input-output viewpoint, and the goal is also to highlight the connection to functional analysis and analytic

[1] For arithmetic rules including these two symbols, see, e.g., [42, p. 4] or [96, p. 24].

function theory, which provide powerful tools to analyze linear dynamical systems. The material presented is mainly from [55, 31, 32, 41, 86, 28, 83, 93]. A nice overview of the topic can also be found in [85].

Linear input-output mappings and the Laplace transform

A linear dynamical system can be described by a linear mapping $\mathcal{A}$ from some input space $\mathcal{U}$ to some output space $\mathcal{Y}$, as illustrated schematically in Figure 2.1. Here, u and y are normally called the input and output signal, respectively. In the case that the input and output spaces are function spaces on $\mathbb{R}$ or $\mathbb{R}_+$, we normally call the system a *continuous-time system*. The choice of domain for the function space in which the signals live normally depend on whether the system is assumed to be at rest at time $t = 0$, in which case one takes the domain of the function space to be $\mathbb{R}_+$. This is common in cases where the dynamics of the system can be described as a deterministic initial-value problem. Similarly, if the input and output spaces are sequence spaces on $\mathbb{Z}$ or $\mathbb{N}$, we normally call the system a *discrete-time system*. In this case, equivalent statements about the domain of the corresponding sequence spaces hold.

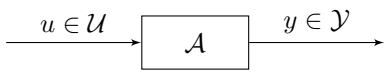


Figure 2.1: A linear dynamical system.

In what follows we will consider the two case when the function space is defined on $\mathbb{R}$ and the sequence space on $\mathbb{Z}$. Moreover, we are particularly interested in *linear time-invariant* (LTI) *single-input-single-output* (SISO) systems. Time-invariance means that a time shift in the input results in the same time shift in the output, i.e., that the operator $\mathcal{A}$ commutes with the family of time-shift operators $\mathcal{T}_\mathfrak{t} : u(\cdot) \to u(\cdot - \mathfrak{t})$ for all $\mathfrak{t} \in \mathbb{R}$ in the continuous-time case, and with $\mathcal{T}_\mathfrak{t} : u_\cdot \to u_{\cdot - \mathfrak{t}}$ for all $\mathfrak{t} \in \mathbb{Z}$ in the discrete-time case. That $\mathcal{A}$ is SISO means that $\mathcal{U}$ and $\mathcal{Y}$ are real-valued function or sequence spaces. For LTI SISO systems, the operator $\mathcal{A}$ is a convolution operator [32, p. 15] [83, Chp. 7] [85, p. 195] [86, Sec. 2.3], i.e., for the continuous-time case we have that

$$y(t) = \mathcal{A}(u)(t) = [G * u](t) = \int_{-\infty}^{\infty} G(t - \tau)u(\tau)d\tau,$$

where G belongs to some suitable function/distribution space on $\mathbb{R}$,[2] and for discrete-time systems we have that

$$y_t = \mathcal{A}(u)_t = [g * u]_t = \sum_{j=-\infty}^{\infty} g_{t-j}u_j,$$

[2]Note that the convolution operator needs to be interpreted with some care. In fact, it is defined indirectly using the Fourier transform, see, e.g., [83, Sec. 7.5.3]. See also Remark 2.1.1.

where g belongs to some suitable sequence space on $\mathbb{Z}$. In mathematics, the function G and sequence g are often called the *convolution kernels*, however in control and signal processing they are usually called the *impulse response* of the respective system.[3] Finally, such systems are called *causal* if the output at the current time point do not depend on future input values. Formulated in terms of the impulse response, and system is causal if $G(t) = 0$ for $t < 0$ in the continuous-time case, and if $g_t = 0$ for $t < 0$ in the discrete-time case. In this case the convolutions take the form $y(t) = \int_{-\infty}^{t} G(t-\tau)u(\tau)d\tau$ and $y_t = \sum_{j=-\infty}^{t} g_{t-j}u_j$, respectively. Since the theory for continuous- and discrete-time systems are analogous in many case, from now on we will focus on the continuous-time case. In the end of the section, corresponding results and considerations will be summarize for the discrete-time case.

To this end, for a function $f : \mathbb{R} \to \mathbb{R}$ we define the (bilateral) *Laplace transform* $\mathcal{L}$ as

$$\mathcal{L}(f)(s) := \int_{-\infty}^{\infty} f(\tau)e^{-\tau s}d\tau,$$

where $s \in \mathbb{C}$. We say that the transform converges for an $s \in \mathbb{C}$ if $\int_{-\infty}^{\infty} |f(\tau)e^{-\tau s}|d\tau$ converges [85, p. 195] [86, Sec. 9.2], i.e., if the integral is absolutely convergent (cf. [31, p. 11]). If f is such that the transform has a nonempty region of convergence Ω, i.e., there exists an $\Omega \subset \mathbb{C}$, $\Omega \neq \emptyset$, so that for all $s \in \Omega$ the transform converges, we get that the Laplace transform of f is a function $\mathcal{L}(f) : \Omega \to \mathbb{C}$. We will denote this function $\hat{f}(s) := \mathcal{L}(f)(s)$. Using the Laplace transform we can now define the *transfer function* corresponding to the LTI SISO system $\mathcal{A}$. This is the Laplace transform of the operator $\mathcal{A}$, i.e., the Laplace transform of the convolution kernel/impulse response G.[4] However, for causal systems we have that $G(t) = 0$ for $t < 0$, and for causal systems we thus have that the transfer function is given by

$$\hat{G}(s) := \mathcal{L}(G)(s) = \int_{0}^{\infty} G(\tau)e^{-\tau s}d\tau.$$

Moreover, it can be shown that for this kind of functions, sometimes called right-sided functions, a nonempty region of convergence Ω is always a half-plane in $\mathbb{C}$ [31, Chp. 3] [85, pp. 195-196] [86, Sec. 9.2]. To see this, assume that the transform converges for some $s_0 \in \mathbb{C}$ such that $\Re(s_0) = a$. Then for any s with $\Re(s) \geq a$ we

[3] The name comes from the fact that they are obtained as output of the system when the input signal is taken to be an impulse at time $t = 0$. In the continuous-time case, this means the Dirac impulse $\delta(t)$ (the one-function in Fourier domain, cf. footnote 2), while in the discrete-time case the impulse is the sequence $\delta_0 = 1$ and $\delta_t = 0$ for $t \neq 0$.

[4] Formally, the Laplace transform of the operator is defined as the operator denoted by $\mathcal{L}(\mathcal{A})$, so that $\mathcal{L}(\mathcal{A}f) = \mathcal{L}(\mathcal{A})\big(\mathcal{L}(f)\big)$ for all f in the function space under consideration. It is well-known that when it converges, the Laplace transform of a convolution becomes a multiplication with the Laplace transform of the kernel [31, Chp. 10] [108, Thm. 3.6]. Strictly speaking, the Laplace transform of the operator is thus the multiplication operator that multiplies with the Laplace transform of the kernel, cf. Remark 2.1.1.

have that

$$\int_0^\infty |G(\tau)e^{-\tau s}|d\tau = \int_0^\infty |G(\tau)|e^{-\tau \Re(s)}d\tau \leq \int_0^\infty |G(\tau)|e^{-\tau a}d\tau,$$

and the last integral is convergent by assumption. Moreover, it can be shown that the corresponding function $\hat{G}(s)$ is in fact an *analytic function*[5] in the open half-space where it converges [31, Chp. 6], and that it is also bounded [31, Thm. 3.2].

Analytic functions, Hardy spaces, and connection to stability

So far we have avoided exact details of the function spaces $\mathcal{U}$ and $\mathcal{Y}$. However, to develop the theory further we need to specify these. In what follows, $\mathcal{U}$ and $\mathcal{Y}$ will in general be real function spaces. However, since the transfer function $\hat{G}(s)$ is an analytic function we will also need complex function spaces. To this end, let $(X, \mathfrak{A})$ and $(Y, \mathfrak{B})$ be two measure spaces. Often we will denote these spaces only X and Y, dropping the explicit notation for the σ-algebras $\mathfrak{A}$ and $\mathfrak{B}$. A function $f : X \to Y$ is said to be a *measurable function* if for all $B \in \mathfrak{B}$, $f^{-1}(B) \in \mathfrak{A}$.[6] Now, let $L_p(X)$ be the function space of all (potentially complex valued) measurable functions on X such that $\|f\|_{L_p}^p := \int_X |f(t)|^p dt < \infty$, for $p = 1, 2, \ldots$, and such that $\|f\|_{L_\infty} := \operatorname{ess\,sup}_{t\in X} |f(t)| < \infty$ for $p = \infty$, see, e.g. [42, Sec. 3.2], [55, Chp. 1], or [98, Chp. 3]. Moreover, let $\mathcal{H}_p(\mathbb{C}_+)$ denote the Hardy space of functions f that are analytic in $\mathbb{C}_+$ and such that

$$\|f\|_{\mathcal{H}_p(\mathbb{C}_+)}^p := \sup_{x>0} \int_{-\infty}^\infty |f(x+iy)|^p dy < \infty$$

for $p = 1, 2, \ldots$, and such that

$$\|f\|_{\mathcal{H}_\infty(\mathbb{C}_+)} := \sup_{x>0, y\in\mathbb{R}} |f(x+iy)| < \infty$$

for $p = \infty$, see, e.g., [55, Chp. 8]. In fact, for $f \in \mathcal{H}_p(\mathbb{C}_+)$ we have that $f(a+ib)$ belongs to $L_p(\mathbb{R})$ for all $a \geq 0$, when seen as a function of b. Moreover, for such f, any sequence of the form $\|f(a+i\cdot)\|_{L_p(\mathbb{R})}$ is a nonincreasing sequence in a, and thus $\|f\|_{\mathcal{H}_p(\mathbb{C}_+)}^p = \int_{-\infty}^\infty |f(iy)|^p dy = \|f(i\cdot)\|_{L_p(\mathbb{R})}^p =: \|f\|_{L_p(i\mathbb{R})}^p$, for $p = 1, 2, \ldots$, and similarly for $f \in \mathcal{H}_\infty(\mathbb{C}_+)$ [55, Chp. 8].

Now, *input-output stability* of a linear system can be defined in terms of that the operator $\mathcal{A} : \mathcal{U} \to \mathcal{Y}$ is a bounded linear operator between the two spaces. Therefore,

[5] Note that the term analytic normally refers to complex functions that locally have a power-series representation, while the term *holomorphic* refers to infinitely differentiable complex functions. However, the two notions are equivalent [98, Thm. 10.6 and 10.16], and thus we will make no difference between the two (cf. [98, Def. 10.2]).

[6] This is the definition used by Kallenberg [57, p. 3]. However, Kallenberg also states that a topological space is always equipped with the Borel σ-algebra, which is the σ-algebra generated by the topology [57, p. 2]. In contrast, Rudin defines measurable functions directly between a measure space and a topological space [98, p. 8], which is also what Friedman does [42, Sec. 2.1].

depending on the function spaces $\mathcal{U}$ and $\mathcal{Y}$ we choose, we get different notions of stability. One commonly considered notion of stability is when $\mathcal{A}$ is a bounded linear operator from $\mathcal{U} = L_2(\mathbb{R})$ to $\mathcal{Y} = L_2(\mathbb{R})$. This is of interest since the L_2-norm is often used as a measure of the energy in the signal. The idea of this type of stability is that an input signal with bounded energy should give an output signal with bounded energy, meaning that the energy amplification in the system is limited. We will thus refer to this kind of stability as L_2-stability or "stability in the energy sense". In this case, it can be shown that the operator norm is finite, i.e.,

$$\sup_{\substack{u \in L_2 \\ u \neq 0}} \frac{\|\mathcal{A}(u)\|_{L_2}}{\|u\|_{L_2}} < \infty,$$

if and only if $\hat{G}(s) \in \mathcal{H}_\infty(\mathbb{C}_+)$, see, e.g., [32, Sec. 2.3 and 2.5] [41, p. 54] [83, Sec. 7.5.3] [85, p. 199].

Another type of stability of interest is so-called *bounded-input-bounded-output* (BIBO) stability, which means that $\mathcal{A}$ is a bounded linear operator from $\mathcal{U} = L_\infty(\mathbb{R})$ to $\mathcal{Y} = L_\infty(\mathbb{R})$. This type of stability is arguably as natural as the energy stability, since the former can handle things like pure sinusoidal inputs which the latter cannot. BIBO stability can be shown to be equivalent to $G(t) \in L_1(\mathbb{R})$, see, e.g., [32, Sec. 2.3 and 2.5] [28, Thm. 5.1] [83, p. 161] or [85, p. 199] and references therein.[7]

When the region of convergence for the Laplace transform includes the imaginary axis we note that for $s = i\omega$ we get that

$$\hat{G}(i\omega) = \int_0^\infty G(\tau) e^{-i\omega\tau} d\tau = \int_{-\infty}^\infty G(\tau) e^{-i\omega\tau} d\tau,$$

which is the *Fourier transform* of the function G. The function $\hat{G}(i\omega)$ is called the *frequency response* of the system. The name comes from the fact that a sinusoidal input signal $u(t) = \sin(\omega_0 t)$ gives the sinusoidal output signal $y(t) = |\hat{G}(j\omega_0)| \sin(\omega_0 t + \angle \hat{G}(j\omega_0))$ [93, Sec. 8.4]. The frequency response has traditionally been a useful tool in control, e.g., properties of a system can be observed in the so-called Bode plot [93, Sec. 8.4]. Moreover, the frequency response function can also be used to determine stability of a closed-loop system via the Nyquist criteria [93, Sec. 9.2] [32, p. 37].

Remark 2.1.1. Although the presentation above serves as an introduction to the subject, it is not mathematically stringent. More precisely, the operator $\mathcal{A}$ is not defined via the convolution but via the Fourier transform. This means that the convolution above should be interpreted as $(G * u) := [\mathcal{F}^{-1} \hat{G} \mathcal{F}](u)$, where $\mathcal{F}$ is the Fourier transform [83, Sec. 7.5.3]. From this perspective, the transfer function (frequency response) is in fact more fundamental than the impulse response, since

[7] Note that the condition $G \in L_1(\mathbb{R})$ does not have a nice description in terms of the transfer function $\hat{G}$ [36, p. 102] [83, p. 161], the latter in fact being the more "fundamental" of the two, cf. Remark 2.1.1. This is most likely why L_2-stability is more commonly used in the literature.

the former gives rise to a notion of the latter and not the reverse. More over, in this case time invariance is defined as the commutativity of the Fourier transformed operator $\mathcal{F}\mathcal{A}\mathcal{F}^{-1}$ with the exponential group $e^{i\mathfrak{t}\omega}$, for all $\mathfrak{t} \in \mathbb{R}$ [83, Sec. 7.5.3], and causality is defined as that $f(t) = 0$ for $t \leq s$ implies that $\mathcal{A}(f)(t) = 0$ for $t \leq s$ [83, Sec. 7.1]. Finally, $\mathcal{F}^{-1}([\hat{G}\mathcal{F}](u))$ many not converge for all input signals in the intended input space $\mathcal{U}$. To remedy this one instead considers input signals from the signal space $\mathcal{U}_\kappa := \{e^{-\kappa t}u \mid u \in \mathcal{U}\}$ [54, Chp. 1]. If there is a finite κ_0 so that the inverse transform converges for all $u \in \mathcal{U}_{\kappa_0}$, then clearly it converges for all $\kappa \geq \kappa_0$. The set of input signals can thus be taken as $\bigcup_{\kappa \geq \kappa_0} \mathcal{U}_\kappa$. Note that in the above discussion, this corresponds to the Laplace transform converging in the half-plane $\Re(s) \geq \kappa_0$, and we call the system stable if $\kappa_0 \leq 0$.

Rational transfer functions, state-space representation, and finite-dimensional linear systems

A transfer function is called *rational* if it is the quotient of two polynomials, i.e., if $\hat{G}(s) = b(s)/a(s)$ where

$$a(s) = a_0 + a_1 s + \ldots + a_n s^k \tag{2.1.1a}$$

$$b(s) = b_0 + b_1 s + \ldots + b_m s^m. \tag{2.1.1b}$$

Here we will only consider the case of so-called *proper* systems, i.e., systems such that $k \geq m$. Moreover, we will always assume that a and b are coprime. Using the Laplace transform, it is easily seen that such a transfer function is equivalent to the operator $\mathcal{A}$ being the solution operator to a dynamical system that has a description in terms of a constant-coefficient ordinary differential equation (ODE) of the form

$$a_k y^{(k)}(t) + \ldots + a_1 \dot{y} + a_0 y = b_m v^{(m)}(t) + \ldots + b_1 \dot{v} + b_0 v. \tag{2.1.2}$$

For a rational transfer function we define the poles and zeros of the transfer function to be the zeros of $a(s)$ and $b(s)$, respectively. It is easily seen that it is analytic in the half-plane to the right of the right-most pole [86, p. 669]. Since the system is proper, if all poles are in the left half-plane $\mathbb{C}_-$ the function $\hat{G}(s)$ is in fact in $\mathcal{H}_\infty(\mathbb{C}_+)$ and thus stable. These arguments lead to the standard result from a basic course in control theory that "a system is stable if and only if all poles of the transfer function are in the left half-plan" [86, p. 697] [93, p. 240] (cf. cite[Sec. 3.1]foias1996robust). Moreover, if we define $u = v^{(m)}$ as the input signal, by the change of variable $x_1 = y$, …, $x_k = y^{(k-1)}$, $x_{k+1} = v$, …, $x_{k+m} = v^{m-1}$, we can write (2.1.2) as a system of first-order ODEs.

In the input-output approach so far described, one needs to keep track of the entire input signal $u(t)$ in order to compute $y(t)$. However, for many LTI SISO systems this is not necessary. The concept needed to get around this is that of a *state* of a dynamical system. A state $x(t)$ is defined as a representation of the system that contains enough information to describe how previous inputs affect future outputs: Formally, a state is a function $x(\cdot)$ so that for all t_0, the output

$y(t)$ for $t \geq t_0$ is uniquely determined by $x(t_0)$ and $u(t)$ for $t \geq t_0$ [28, Def. 2.1]. Moreover, an LTI SISO system is called *finite-dimensional* (or lumped) if it can be described in a finite-dimensional *state space form*, i.e., as

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{2.1.3a}$$
$$y(t) = Cx(t) + Du(t), \tag{2.1.3b}$$

where $x(t) \in \mathbb{R}^n$ is a vector-valued function, and A, B, C, and D are matrices of appropriate dimensions [28, Sec. 2.1.1]. This means that by definition, all finite-dimensional LTI SISO systems can be described by a system of first-order ODEs as in (2.1.3). Now, note that the change of variables defined above in order to write (2.1.2) as a system of first-order ODEs in fact defines a state space form for this higher-order ODE. Vice versa, any system of equations of the form (2.1.3) defines a proper rational transfer function via $\hat{G}(s) = D + C(sI - A)^{-1}B$ [28, pp. 15-16] [85, pp. 200-201] [93, Sec. 8.1].[8] This means that any LTI SISO system is finite-dimensional if and only if it has a rational transfer function [28, p. 14].

Remark 2.1.2. Note that for finite-dimensional LTI SISO systems, i.e., LTI SISO systems with a proper rational transfer function, BIBO stability is in fact also equivalent with all poles being in the left half-plane [28, Thm. 5.3].

If it is not possible to represent the system in the form (2.1.3) with a finite-dimensional state vector, then the system is called infinite-dimensional (or distributed) [28, Sec. 2.1.1]. One common example of such a system is the *delay system*, which can be described by the linear ODE

$$y(t) = u(t - \mathfrak{t}),$$

for some $\mathfrak{t} > 0$. In this case, to describe the evolution of $y(t)$ for $t \geq 0$ we need to know $u(t)$ for $t \geq 0$ and $u(t)$ for $t \in [-\mathfrak{t}, 0]$. The latter cannot be summarized in a finite-dimensional vector. In fact, in this case the operator $\mathcal{A}$ is a convolution with a shifted Dirac impulse, $G(t) = \delta_{\mathfrak{t}}(t)$, and the transfer function is $\hat{G}(s) = e^{-\mathfrak{t}s}$, cf. [28, Ex. 2.4], which as expected is not a rational function.

Remark 2.1.3. The idea of state is in fact what is used to define dynamical systems formally in mathematics. In this case, a dynamical system is a triplet $\{T, X, \phi_t\}$, where T is a time set, X is a state space, and $\{\phi_t\}_{t \in T}$ is a family of evolution operators $\phi_t : X \to X$ such that i) $\phi_0 =$ the identity operator, and ii) $\phi_{t_1+t_2} = \phi_{t_1} \circ \phi_{t_2}$. See, e.g., [63, Sec. 1.1] for details.

[8] Given a proper rational transfer function $\hat{G}(s) = b(s)/a(s)$, where a and b are coprime, any state space form (2.1.3), abbreviated (A, B, C, D), so that $\hat{G}(s) = D + C(sI - A)^{-1}B$ is called a *realization*. However, note that when forming $D+C(sI-A)^{-1}B$ we might get pole-zero cancellation between numerator polynomial and denominator polynomial. This happens if and only if the realization (A, B, C, D) is not *minimal*, which is equivalent with that is not both controllable and observable [28, Thm. 7.2].

Feedback control

The concept of *control* can be defined as designing the input signal $u(t)$ in order to steer the output signal $y(t)$, and in *feedback control* the value of $u(t)$ is often based on the value of $y(t)$ or its mismatch with a desired reference signal. This is often done by letting $u(t)$ be the output of another causal LTI SISO dynamical system with input depending on $y(t)$. To this end, let $P(s)$ denote the transfer function for a dynamical system and let $K(s)$ represented the controller, i.e., a transfer function where $u(t)$ is the output. In feedback control $K(s)$ is connected to $P(s)$, and these interconnections are often represented using block-diagrams. An example is shown in Figure 2.2, where $r(t)$ is a desired reference signal that we would like $y(t)$ to mimic, and $e(t) := r(t) - y(t)$ is the instantaneous error.

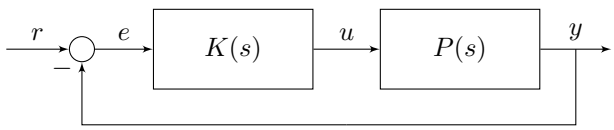


Figure 2.2: A feedback control system.

A symbolic calculation based on the block-diagram in Figure 2.2 gives that

$$\left.\begin{array}{l} y = P(s)u \\ u = K(s)(r-y) \end{array}\right\} \quad \Longrightarrow \quad (1 + P(s)K(s))y = P(s)K(s)r,$$

and the system is called *well-posed* or *well defined* if

$$1 + P(s)K(s) \neq 0 \quad \text{for all } s \in \bar{\mathbb{C}}_+, \tag{2.1.4}$$

cf. [32, Sec. 3.1] [41, p. 53]. If (2.1.4) holds, then the *closed-loop transfer function* $PK/(1+PK)$ from r to y is analytic in $\bar{\mathbb{C}}_+$. Moreover, it is also bounded which means that the system is stable in energy sense. However, in this system u is not an *external signal* but an *internal signal*. This means that the above stability results do not tell us what happens with u. For this reason, the concept of *internal stability* has been introduced. A system is called *internally stable* if all bounded external signal gives rise to bounded internal signals [32, p. 35]. It can be shown that a transfer function for a feedback loop as show in Figure 2.2 is internally stable if and only if it fulfills (2.1.4) and there are no pole-zero cancellations of poles and zeros in $\mathbb{C}_+$ when forming the product $P(s)K(s)$ [32, p. 36].

A note on the factorization of $\mathcal{H}_p(\mathbb{C}_+)$-functions

As we saw earlier, the function spaces $\mathcal{H}_p(\mathbb{C}_+)$, and especially $\mathcal{H}_\infty(\mathbb{C}_+)$, turns out to play an important roll for the analysis and design of linear dynamical systems. It turns out that all functions $f \in \mathcal{H}_p(\mathbb{C}_+)$ have a particular structure, meaning that they can be uniquely factorized as a product of three types of analytic functions.

This kind of factorization is useful in many cases, and in this subsection we will thus summarize these results. Moreover, in the end of the subsection we will comment on the explicit relation to finite-dimensional linear dynamical systems. For an in-depth treatment of the subject, see, e.g., [55, 35] in which all results below can be found.

To this end, any $f \in \mathcal{H}_p(\mathbb{C}_+)$ can be uniquely factorized as

$$f(s) = \lambda B(s)S(s)F(s),$$

where $\lambda \geq 0$ is a constant, and $B, F, S \in \mathcal{H}_p(\mathbb{C}_+)$ are particular types of $\mathcal{H}_p(\mathbb{C}_+)$-functions. These different parts are presented presented below. First, $B(s)$ is called a *Blaschke product* if

$$B(s) = \left(\frac{s-1}{s+1}\right)^k \prod_n \frac{|1-\beta_n^2|}{1-\beta_n^2} \frac{s-\beta_n}{s+\bar{\beta}_n},$$

where

$$\sum_n \frac{\Re(\beta_n)}{1+|\beta_n|} < \infty. \tag{2.1.5}$$

Here, $\beta_1, \beta_2, \ldots$ are the zeros of f in $\mathbb{C}_+$ that are not in the point $s = 1$, and k is the (finite) order of a (potential) zero of f in $s = 1$. The condition (2.1.5) guarantees that the expression for the Blaschke product converge to an analytic function [55, p. 132]. Moreover, $[f/B](s)$ defines a $\mathcal{H}_p(\mathbb{C}_+)$-function without zeros in $\mathbb{C}_+$. As a final note on Blaschke products, note that $|B(s)| \leq 1$ for all $s \in \mathbb{C}_+$ and $|B(i\omega)| = 1$.

Furthermore, $S(s)$ is called a *singular function* and has the form

$$S(s) = e^{-\rho s} \exp\left[-\int_{-\infty}^{\infty} \frac{\omega s + i}{\omega + is} d\mu(\omega)\right]$$

where ρ is a nonnegative real number, and $d\mu$ is a finite, singular, nonnegative measure on $\mathbb{R}$ (including the possibility that $d\mu \equiv 0$). Moreover, $|S(s)| \leq 1$ for all $s \in \mathbb{C}_+$, $|S(iy)| = 1$ a.e., and $S(s) \neq 0$ for $s \in \mathbb{C}_+$, cf. [55, p. 133].

Any $\mathcal{H}_p(\mathbb{C}_+)$-function of the form $B(s)S(s)$ is called an *inner function*. The function $F(s)$ is called an *outer function* and has the form

$$F(s) = \exp\left[\frac{1}{\pi}\int_{-\infty}^{\infty} \log\left(|f(i\omega)|\right) \frac{\omega s + i}{\omega + is} \frac{1}{1+\omega^2}\, d\omega\right].$$

Some important properties of outer functions are that i) $F(x+i\omega)$ has nontangential limits $x \to 0$ which converge to in $L_p(\mathbb{R})$-norm to $f(i\omega)$ [55, p. 128-133], ii) $|F| = |f|$ a.e. on the imaginary axis (both B and S have absolute value 1 there), and iii) that $|F(s)| \geq |f(s)|$ for $s \in \mathbb{C}_+$. Moreover, since F does not have any zeros in $\mathbb{C}_+$ it is invertible on there, and the inverse is also an analytic function on $\mathbb{C}_+$.

Conversely, assume that f is a function of the form $f = BSF$ where i) B is a Blaschke product, ii) S is a singular function, and iii) F is constructed as an outer

function from a nonnegative measurable function $g \in L_p(\mathbb{R})$ such that

$$\int_{-\infty}^{\infty} \frac{\log\big(g(\omega)\big)}{1+\omega^2}\, d\omega > -\infty. \tag{2.1.6}$$

Then $f \in \mathcal{H}_p(\mathbb{C}_+)$ [35, Thm. 11.7]. Since an outer function F is uniquely defined from the nonnegative function $g \in L_p(\mathbb{R})$ fulfilling (2.1.6), such a function g is also sometimes referred to as an outer function.[9]

Finally, we will connect these different factors to concepts from control, and in particular to causal proper LTI SISO systems that are stable in energy sense. To this end, we first note the all L_2-stable finite-dimensional transfer functions $\hat{G}$ can be factorized into an L_2-stable *minimum phase* transfer function and an L_2-stable *all-pass* transfer function, $\hat{G} = \hat{G}_{\text{min-phase}}\hat{G}_{\text{all-pass}}$ [32, p. 91]. A minimum phase transfer function is a transfer function that has no zeros in $\mathbb{C}_+$ [32, p. 90][93, p. 283]. Moreover, an all-pass transfer function is a transfer function so that $|\hat{G}_{\text{all-pass}}(j\omega)| = 1$ for all ω. The name all-pass comes from the fact that, as we noted before, for a sinusoidal input signal $u(t) = \sin(\omega_0 t)$ we get a sinusoidal output signal $y(t) = |\hat{G}_{\text{all-pass}}(j\omega_0)|\sin(\omega_0 t + \angle\hat{G}_{\text{all-pass}}(j\omega_0)) = \sin(\omega_0 t + \angle\hat{G}_{\text{all-pass}}(j\omega_0))$, and thus the magnitude of the output signal is unchanged for all frequencies [32, p. 90]. Now, the factors introduced above for $\mathcal{H}_\infty(\mathbb{C}_+)$-functions generalize these concepts to infinite-dimensional systems that are energy stable. In fact, as outer functions have no zeros (and no poles) in $\mathbb{C}_+$, they can be interpreted as an extension of L_2-stable finite-dimensional minimum phase transfer functions. Moreover, inner functions have unit magnitude on the imaginary axis which means that they are the generalization of L_2-stable finite-dimensional all-pass functions. In fact, all L_2-stable finite-dimensional all-pass functions are Blaschke products with a finite number of zeros, cf. [32, p. 91]. As another example, note that the time-delay system introduced above, which was an infinite-dimensional system with transfer function $\hat{G}(s) = e^{-\mathfrak{t}s}$, corresponds to a singular function with $\rho = \mathfrak{t}$ and $d\mu \equiv 0$.

Summary of discrete-time systems

Here we will briefly summarize some of the corresponding results for discrete-time systems. In this case, the *Z-transform* takes the role of the Laplace transform. For a sequence $g := \{g_\ell\}_{\ell\in\mathbb{Z}}$ we define the Z-transform as [86, Chp. 10][10]

$$\hat{g} := \mathcal{Z}(g) := \sum_{\ell=-\infty}^{\infty} g_\ell z^{-\ell}.$$

[9] Note that in [35, Chp. 11], the notation $\mathfrak{H}^p$ is used to denote the Hardy spaces on $\mathbb{C}_+$ as we have defined them here, while H^p is used to denote a larger class of functions. Using the notation in [35] and comparing [35, Thm. 11.6] and [35, Thm. 11.7] we see, for example, that $g(\omega) = \log(1+|\omega|)$ defines an outer function in H^1 but not in $\mathfrak{H}^1$, since $\log(1+|\omega|)/(1+\omega^2) \in L_1(\mathbb{R})$ but $\log(1+|\omega|) \notin L_1(\mathbb{R})$.

[10] Note that one sometimes defines the Z-transform using positive powers, i.e., as $\sum_{\ell=-\infty}^{\infty} g_\ell z^{\ell}$. However, the two definitions are equivalent and all results are easily translated using, e.g., the Möbius transformation $M(z) = \frac{1}{z}$.

Similarly as before, the Z-transform of a causal LTI SISO system thus takes the form $\mathcal{Z}(g) = \sum_{\ell=0}^{\infty} g_\ell z^{-\ell}$, and any nonempty region of convergence thus has the form of the complement of a disc centered at the origin. Moreover, being a Laurent series of a complex variable it is clearly an analytic function in the region of convergence [98, Thm. 10.6].

We can now introduce the Hardy spaces on the complement of the unit disc: $\mathcal{H}_p(\bar{\mathbb{D}}^C)$ is the space of functions f that are analytic in $\bar{\mathbb{D}}^C$ such that $f \in L^p(\mathbb{T})$ [89, Def. 1.3.3], [55, p. 39]. However, since $\mathbb{T}$ has bounded total mass we have that $L_p(\mathbb{T}) \subset L_q(\mathbb{T})$ for all $1 \leq q \leq p \leq \infty$,[11] which thus mean that $\mathcal{H}_p(\bar{\mathbb{D}}^C) \subset \mathcal{H}_q(\bar{\mathbb{D}}^C)$ for all $1 \leq q \leq p \leq \infty$. Now we can use the same kind of arguments about stability as was done for the continuous-time case: A system is BIBO stable, i.e., a bounded linear map between $\mathcal{U} = \mathcal{Y} = \ell_\infty(\mathbb{Z})$, if and only if the impulse response g belongs to $\ell_1(\mathbb{Z})$ [86, Sec. 2.3.7] [83, Sec. 7.5.2]. Similarly, for $\mathcal{Y} = \mathcal{U} = \ell_2(\mathbb{Z})$ the system is stable if and only if $\hat{g} \in \mathcal{H}_\infty(\bar{\mathbb{D}}^C)$ [83, Lem. 7.2.3]. Moreover, equivalent derivations and observations on finite-dimensional systems, rational transfer functions, and finite difference equations can be made also in for discrete-time systems (cf. Section 2.2). This also leads to that for finite-dimensional proper LTI SISO systems, BIBO stability is equivalent with that all poles have magnitude less than one [28, Thm. 5.D3].

Remark 2.1.4. As a final note we observe that $\mathcal{H}_\infty(\mathbb{C}_+)$ and $\mathcal{H}_\infty(\bar{\mathbb{D}}^C)$ are in fact in a bijective correspondence, since the Möbius transformation $M(s) = \frac{s-1}{s+1}$ is a bijective conformal map from $\bar{\mathbb{D}}^C$ to $\bar{\mathbb{C}}_+$ (cf. [89, Sec. 1.4]) and since a function that is bounded remains bounded also after composition with the transform.

2.2 Stochastic processes and spectral estimation

This section formally introduces the concept of a stochastic process, second-order stationary processes, and the spectrum a of discrete-time second-order stationary process. It also introduces and motives the rational covariance extension problem, which is a spectral estimation problem that was posed by R.E. Kalman in 1981 [58]. The material presented below is a collection of some of the material presented in the text books [92, 42, 98, 57, 105, 72].

Random variables and stochastic processes

A *probability space* is a triplet $(\Omega, \mathfrak{A}, P)$, where $(\Omega, \mathfrak{A})$ is a measure space and P is a measure defined on $\mathfrak{A}$, such that $P(\Omega) = 1$. The intuitive interpretation of this definition is that the elements $A \in \mathfrak{A}$ are the random events that can occur and the value $P(A)$ is probability of the event A. One of the most fundamental notions in probability theory is a *random variable*, the generalization of which is a so-called *random element* [57, p. 24]. To define the latter, let $(\Omega, \mathfrak{A}, P)$ be a probability

[11] A more general statement is that for any measure space $(X, \mathfrak{A}, \mu)$, the two conditions i) $\sup_{A \in \mathfrak{A}} \mu(A) < \infty$, and ii) $L_p(\mathbb{T}) \subset L_q(\mathbb{T})$ for all $0 < q \leq p \leq \infty$, are equivalent [106, Thm. 2].

space and $(\Xi, \mathfrak{B})$ a measure space. A random element X is a measurable function $X : \Omega \to \Xi$, and the probability of the outcome $X(\omega) \in B$ is formally defined for any set $B \in \mathfrak{B}$ as

$$P(X(\omega) \in B) := P(X^{-1}(B)) = (P \circ X^{-1})(B).$$

In this case we can interpret $X(\omega) \in B \in \mathfrak{B}$ as an (indirect) observation of the random event $\omega \in \Omega$. Moreover, given a random element X, the *expected value* (*mean value*) of X is defined as $\mathbb{E}[X] := \int_\Omega X(\omega) dP(\omega)$.[12]

In this setting, a (complex) *random variable* is simply the name for a random element in the case when Ξ is equal to $\mathbb{R}$ ($\mathbb{C}$), and a (complex) *random vector* is a random element with Ξ equal to $\mathbb{R}^d$ ($\mathbb{C}^d$) for some $d = 2, 3, \ldots$. Moreover, a *stochastic process* can be defined as a random element that maps into a sequence space or function space [57, p. 24]. However, a stochastic process can equivalently be seen as a collection of random variables $\{X(\omega, t)\}_{t \in T}$, where T is some index set [57, p. 24] [92, Chp. 2]. Analogously to dynamical systems, if T is $\mathbb{N}$ or $\mathbb{Z}$ we say that it is a *discrete-time process* and if T is $\mathbb{R}_+$ or $\mathbb{R}$ we say that it is a *continuous-time process*. Moreover, for a fixed $\omega \in \Omega$, $X(\omega, t)$ is a function of the second argument t and this is called a *realization* or *sample path* of the process.

The mean value of a stochastic process is, by definition, a function of t since

$$m(t) := \mathbb{E}\big[X(\cdot, t)\big] = \int_\Omega X(\omega, t) dP(\omega) = \int_{\mathbb{R}} x[P \circ X(\cdot, t)^{-1}](x) dm,$$

where the last equality holds if we assume that the process is real-valued, and where dm is the standard Borel measure on $\mathbb{R}$. In a similar fashion we can define higher-order moments of the stochastic process, the most commonly used being the second-order moments. These are called the *covariances* and are defined as

$$c(t, s) = \mathbb{E}\big[(X(\cdot, t) - m(t))(X(\cdot, s) - m(s))^*\big] = \mathbb{E}\big[X(\cdot, t)X(\cdot, s)^*\big] - m(t)m(s)^*.$$

A process is said to be *second-order stationary* (or *weakly stationary*) if for some constant m, $m(t) = m$ a.e., and if the covariance function $c(t, s)$ is a function only of the argument $t - s$, i.e., with a slight abuse of notation if $c(t, s) = c(t - s)$ [72, p. 42]. Finally, a second-order stationary stochastic process is called *ergodic* if the mean and covariances of a realization are the same as the ensemble mean and covariances, i.e., if for the sample average

$$\hat{m}_T := \frac{1}{2T + 1} \sum_{t=-T}^{T} X(\omega, t)$$

we have that $\lim_{T \to \infty} \hat{m}_T = m = \mathbb{E}\big[X(\cdot, t)\big]$ for ω a.e. dP, and if for the sample covariances

$$\hat{c}_T(\tau) := \frac{1}{2T + 1} \sum_{t=-T}^{T} (X(\omega, t) - \hat{m}_T)(X(\omega, t - \tau) - \hat{m}_T)^*$$

[12] For definitions on how abstract integration is defined, see, e.g, [98, Chp. 1] [57, Chp. 1].

we have that $\lim_{T\to\infty} \hat{c}_T(\tau) = c(\tau)$ for all $\tau \in \mathbb{Z}$ and ω a.e. dP [72, Def. 13.1.3]. The concept is defined similarly for a continuous-time stochastic process [92, p. 17].

Spectra for discrete-time second-order stationary processes

For the remaining of Section 2.2 we will, unless otherwise explicitly stated, focus on discrete-time, second-order stationary, and ergodic processes. We therefore introduce a somewhat simplified notation. To this end, we will start by a slight abuse of notation and considering $\{y_t \in \mathbb{C}\}_{t\in\mathbb{Z}}$ both as a stochastic process, and as *time series*, the latter being nothing but one realization of the stochastic process. In the same spirit, we will still write things like $\mathbb{E}[y_t]$ for the expectation of the stochastic process, which due to the ergodicity assumption is the same as the average over the specific realization. Moreover, in addition to the above assumptions, by the second-order stationarity and ergodicity assumptions we will without loss of generality also assume that the time series is zero-mean. This means that we get a simplified expression for the covariances, namely

$$c_k = \mathbb{E}[y_t y_{t-k}^*].$$

The *power spectrum* of a stochastic process describes the average power distribution across frequencies in the process. This can be seen as a generalization of the energy spectrum of a deterministic signal, which describes the distribution of the signal energy across the frequency components of the signal [105, Chp. 1]. Formally, the power spectrum is defined as the nonnegative measure μ on the complex unit circle $\mathbb{T} := (-\pi, \pi]$ such that

$$c_k = \frac{1}{2\pi} \int_{\mathbb{T}} e^{ik\theta} d\mu(\theta), \quad k \in \mathbb{Z}$$

i.e., the nonnegative measure such that the covariances are Fourier coefficients of μ, see, e.g., [72, Chp. 3] or [105, Chp. 1]. In the case of a power spectrum with only absolutely continuous part, i.e., $d\mu(\theta) = \Phi(e^{i\theta})d\theta$, if Φ is also, e.g., continuously differentiable on $\mathbb{T}$ then we have that

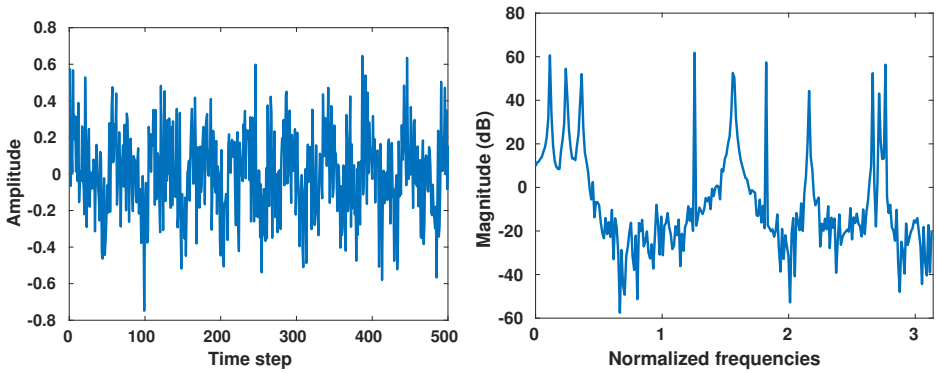
$$\Phi(e^{i\theta}) = \sum_{k=-\infty}^{\infty} c_k e^{-ik\theta}, \tag{2.2.1}$$

i.e., the series converges pointwise [108, Thm. 4.5].[13]

The spectral estimation problem can now be formulated as follows: From a finite realization of $\{y_t\}_{t\in\mathbb{Z}}$ estimate the power spectrum of the stochastic process. This is of interest since we in many cases can extract useful information from the power spectrum, which is hard to directly observe in the time series. As an illustration,

[13] Note that the assumption on continuous differentiability can be relaxed if one instead of pointwise convergence considers other types of convergence, cf. [98, pp. 88-92 and pp. 100-104] [108, Chp. 5].

consider Figure 2.3. Thus, spectral estimation has a lot of applications, for example in speech analysis [30, 38], medical diagnostics [2], system identification [72] [104, Sec. 3.5], and many other areas [105]. Because of its usefulness several different methods have also been developed for spectral estimation, such as the periodogram and the correlogram [105, Chp. 2], Burg's method (maximum entropy) [18, 19] [105, Sec. 3.9.3], the Capon method [105, Sec. 5.4], rational covariance extension [58, 44, 45, 27], and others [105].



(a) Time series data.

(b) Periodogram estimate.

Figure 2.3: An example of a time series is shown in Figure 2.3a and the corresponding periodigram estimate is shown in Figure 2.3b. From the time series itself it is hard to tell, but as can be seen in the periodigram there are 10 prominent frequencies present in the data.

The rational covariance extension problem

Rational covariance extension is a spectral estimation procedure, and to understand it better we will first introduce the correlogram. The latter is a spectral estimation method where the estimated spectrum takes the form

$$\Phi_{\text{corr}}(e^{i\theta}) = \sum_{k=-n}^{n} c_k e^{-ik\theta},$$

and where $\{c_k\}_{k=-n}^{n}$ is an estimate of the true covarainces obtained from the finite realization of $\{y_t\}_{t\in\mathbb{Z}}$.[14,15] Comparing this to (2.2.1), we see that the correlogram

[14] Note that the correlogram coincides with the periodogram when the biased covariance estimates are used [105, p. 24].

[15] Here we have on purpose used a somewhat sloppy notation: We do not make an explicit difference between estimated and true covariances. The reason for this is that in what follows we will make the assumption that we have sufficiently good estimates of the covariances, and we will not explicitly consider estimation procedures for covariances nor errors in the obtained estimates.

in fact sets the higher-order covariances to zero. This has several drawbacks, for example that the function obtained might not be nonnegative. In the rational covariance extension problem, one instead seeks an extension of the covariance sequence $\{c_k\}_{k=-n}^{n}$ so that the obtained spectrum (2.2.1) is a nonnegative rational function. In particular, in [58], R.E. Kalman posed the problem of finding all extensions of the covariance sequence such that the spectrum is nonnegative and rational, and where the degree of both the numerator and the denominator is bounded by n, i.e., given a sequence of covariances $c := \{c_k\}_{k=-n}^{n}$ find all nonnegative functions $\Phi(e^{i\theta})$ so that

$$\text{(RCEP)} \qquad \begin{cases} c_k = \dfrac{1}{2\pi}\displaystyle\int_{\mathbb{T}} e^{ik\theta}\Phi(e^{i\theta})d\theta, & k = -n, \dots, -1, 0, 1, \dots, n \\ \Phi(e^{i\theta}) = \dfrac{P(e^{i\theta})}{Q(e^{i\theta})}, & P \text{ and } Q \in \bar{\mathfrak{P}}_+. \end{cases}$$

Here, $\bar{\mathfrak{P}}_+$ denotes the set of real-valued nonnegative trigonometric polynomials, i.e.,

$$\bar{\mathfrak{P}}_+ := \left\{ P \in C(\mathbb{T}) \;\middle|\; P(e^{i\theta}) := \sum_{k=-n}^{n} p_k e^{-ik\theta},\ p_{-k} = p_k^*,\ P(e^{i\theta}) \geq 0 \text{ for all } \theta \in \mathbb{T} \right\},$$

where $C(\mathbb{T})$ denotes the set of continuous functions on $\mathbb{T}$. Moreover, $\bar{\mathfrak{P}}_+$ is in fact a closed convex cone (cf. Section 2.3), and with $\mathfrak{P}_+$ we denote the interior which corresponds to all strictly positive trigonometric polynomials.

The interest in (RCEP) comes from stochastic realization theory, and in what follows we will clarify this connection. For an in-depth treatment of stochastic realization theory see [72], and for the rational covariance extension problem see in particular [72, Sec. 12.5 and 12.6] and references therein. To this end, in the rational covariance extension problem we are not only interested in obtaining an estimate of the spectrum, but we also want a model for how the stochastic process is generated. Here, the stochastic process is modeled as coloring of white noise by filtering it through a BIBO-stable causal finite-dimensional LTI SISO system, and we want to identify the latter. This theory is thus tightly linked with the theory presented in Section 2.1, except that the input and output spaces $\mathcal{U}$ and $\mathcal{Y}$ are now spaces of discrete-time processes instead of sequences.

Let $\{u_t\}_{t\in\mathbb{Z}}$ be a Gaussian white noise process, meaning that i) $\{u_t\}_{t\in\mathbb{Z}}$ is a Gaussian process, i.e., that any finite collection $(u_{t_1}, \dots, u_{t_\ell})$ is a Gaussian random vector, and ii) that the power spectrum of the process is constant: $\Phi_u(e^{i\theta}) \equiv 1$.[16] Moreover, let $\{y_t\}$ be the output of a BIBO-stable causal finite-dimensional LTI SISO system $\mathcal{A}$ with impulse response $g := \{g_k\}_{k\in\mathbb{Z}}$, where $g_k = 0$ for $k < 0$ since the system is causal. In particular, this means that $\mathcal{A}$ can be formulated as a so-called autoregressive-moving-average (ARMA) filter [105, Chp. 3], i.e., the

[16] An equivalent definition for a Gaussian white noise process is that all random variables u_t are Gaussian and independent.

input-output relation between the two time series can be expressed by the finite difference equation

$$y_t + \sum_{k=1}^{n} a_k\, y_{t-k} \;= \sum_{k=0}^{n} b_k\, u_{t-k}.$$

Using the machinery from Section 2.1, we apply the Z-transform[17] and get that the transfer function $\hat{g}(z)$ can be expressed as

$$\hat{g}(z) = \sum_{k\in\mathbb{Z}} g_k z^{-k} = \frac{\sum_{k=0}^{m} b_k\, z^{-k}}{\sum_{k=0}^{n} a_k\, z^{-k}} = \frac{b(z)}{a(z)}, \tag{2.2.2}$$

where $a(z^{-1})$ and $b(z^{-1})$ are complex polynomials with coefficients $\{a_k\}_{k=0}^{n}$, where $a_0 = 1$, and $\{b_k\}_{k=0}^{n}$, respectively.

Since $\mathcal{A}$ is linear, it can be shown that the power spectrum of the output process is given by the power spectrum of the input process and the transfer function $\hat{g}$. This relation is derived in most standard books, see, e.g., [105, Sec. 1.4] [104, Sec. 3.5] [92, Sec. 4.2], however, we will do this here as well since it nicely illustrates the connection between the rational covariance extension problem and identification of a finite-dimensional linear stochastic system (cf. [105, Sec. 3.2]). To this end, let $\tilde{c}_k$ be the covariances of the process u_t. Using the relation $y_t = \mathcal{A}(u)_t = \sum_{k\in\mathbb{Z}} g_k u_{t-k}$ we get that

$$\begin{aligned} c_k = \mathbb{E}[y_t y_{t-k}^*] &= \mathbb{E}\left[\left(\sum_{k_1\in\mathbb{Z}} g_{k_1} u_{t-k_1}\right)\left(\sum_{k_2\in\mathbb{Z}} g_{k_2} u_{t-k-k_2}\right)^*\right] \\ &= \sum_{k_1\in\mathbb{Z}}\sum_{k_2\in\mathbb{Z}} g_{k_1} g_{k_2}^* \mathbb{E}[u_{t-k_1} u_{t-k-k_2}^*] = \sum_{k_1\in\mathbb{Z}}\sum_{k_2\in\mathbb{Z}} g_{k_1} g_{k_2}^* \tilde{c}_{k+k_2-k_1}. \end{aligned}$$

Now, assuming that $\{y_t\}$ has an absolutely continuous spectrum that is, e.g., continuously differentiable, from (2.2.1) we obtain

$$\begin{aligned} \Phi(e^{i\theta}) &= \sum_{k\in\mathbb{Z}} c_k e^{-ik\theta} = \sum_{k\in\mathbb{Z}}\sum_{k_1\in\mathbb{Z}}\sum_{k_2\in\mathbb{Z}} g_{k_1} g_{k_2}^* \tilde{c}_{k+k_2-k_1} e^{-ik\theta} \\ &= \sum_{k_1\in\mathbb{Z}} g_{k_1} e^{-ik_1\theta} \sum_{k_2\in\mathbb{Z}} g_{k_2}^* e^{ik_2\theta} \sum_{k_3\in\mathbb{Z}} \tilde{c}_{k_3} e^{-ik_3\theta} = |\hat{g}(e^{i\theta})|^2 \Phi_u(e^{i\theta}). \end{aligned}$$

Finally, using i) that $\Phi_u(e^{i\theta}) \equiv 1$ since $\{u_t\}$ is a white noise process, and ii) the relation in (2.2.2), we get that

$$\Phi(e^{i\theta}) = |\hat{g}(e^{i\theta})|^2 \Phi_u(e^{i\theta}) = |\hat{g}(e^{i\theta})|^2 = \frac{|b(e^{i\theta})|^2}{|a(e^{i\theta})|^2} = \frac{\sum_{k=-m}^{m} p_k e^{-ik\theta}}{\sum_{k=-n}^{n} q_k e^{-ik\theta}} = \frac{P(e^{i\theta})}{Q(e^{i\theta})}$$

where P and $Q \in \bar{\mathfrak{P}}_+$. To summarize, this shows that if the stochastic process is generated by filtering white noise through a BIBO-stable causal finite-dimensional

[17] Note that z^{-1} corresponds to the (unit) delay operator.

LTI SISO system $\mathcal{A}$, then the power spectrum Φ is the quotient of two trigonometric polynomials. In fact, it can be shown that the converse is also true [92, pp. 98-99] (cf. Section 2.1). The final link that turns solving (RCEP) into a procedure for system identification is to obtain the filter coefficients $\{a_k\}_{k=0}^n$ and $\{b_k\}_{k=0}^n$ from the trigonometric polynomials P and Q. This is possible using so-called *spectral factorization*. In short, the spectral factorization theorem states that for any $P \in \bar{\mathfrak{P}}_+$ we have that $P(e^{i\theta}) = |b(e^{i\theta})|^2$ for some polynomial $b(z^{-1})$ of degree bounded by n, and where we can take all zeros of $b(z)$ to be in the closed unit disc $\bar{\mathbb{D}}$ [105, Sec. 3.2][92, p. 99][33, Thm. 1.1].

To summarize some of the early results on the rational covariance extension problem, in [44, 45] it was shown that (RCEP) has a solution if and only if the Toeplitz matrix of covariances is positive definite, i.e., if and only if

$$T(c) = \begin{bmatrix} c_0 & c_{-1} & \dots & c_{-n} \\ c_1 & c_0 & \dots & c_{-n+1} \\ \vdots & & \ddots & \vdots \\ c_n & c_{n-1} & \dots & c_0 \end{bmatrix} \succ 0.$$

It was also shown that for each c so that $T(c) \succ 0$ and for each numerator polynomial $P \in \mathfrak{P}_+$, there is a denominator polynomial $Q \in \mathfrak{P}_+$ so that P/Q is a solution to (RCEP). In [44, 45] it was also conjectured that for each pair (c, P) there is a unique Q so that (RCEP) holds, which was shown to be true in [27]. A constructive method for computing the unique Q was developed in [23, 24]. This method uses convex optimization, and also gives an alternative proof for the existence and uniqueness of a denominator polynomial Q. The results can be summarized as in the following theorem.

Theorem 2.2.1 ([23, 24])**.** *The rational covariance extension problem* (RCEP) *has a solution if only if $T(c) \succ 0$. For such c and any $P \in \mathfrak{P}_+$, there is a unique $\hat{Q} \in \mathfrak{P}_+$ such that $\Phi = P/\hat{Q}$ is a solution to* (RCEP)*. Moreover, $\Phi = P/\hat{Q}$ is the unique optimal solution to the convex optimization problem*

$$\begin{aligned} \min_{\substack{\Phi \in L_1(\mathbb{T}) \\ \Phi \geq 0}} \quad & \int_{\mathbb{T}} P(e^{i\theta}) \log \frac{P(e^{i\theta})}{\Phi(e^{i\theta})} \frac{d\theta}{2\pi} \\ \text{subject to} \quad & c_k = \int_{\mathbb{T}} e^{ik\theta} \Phi(e^{i\theta}) \frac{d\theta}{2\pi}, \quad k = -n, \dots, 0, 1, \dots, n. \end{aligned} \tag{2.2.3}$$

Furthermore, the unique $\hat{Q}$ is the solution to the dual problem

$$\min_{Q \in \mathfrak{P}_+} \quad \langle c, q \rangle - \int_{\mathbb{T}} P \log(Q) \frac{d\theta}{2\pi},$$

where $q = [q_n^*, \dots, q_0, \dots, q_n]^T$ and where $Q(e^{i\theta}) = \sum_{k=-n}^n q_k e^{-ik\theta}$.

Connection to analytic interpolation As a final note, we observe that the rational covariance extension problem can also be formulated as an analytic interpoilation problem with a rationality constraint on the interpolant, see, e.g., [27, p. 1843]. To this end, consider a sequence of covariances c. We call a function $f : \mathbb{C} \to \mathbb{C}$ positive-real if i) f is analytic in $\bar{\mathbb{D}}^C$, i.e., $f(z)$ can be written as $f(z) = \sum_{k=0}^{\infty} f_k z^{-k}$ for $|z| > 1$, and ii) $\Re\big(f(z)\big) \geq 0$ in $\bar{\mathbb{D}}^C$ [22, p. 822]. Moreover, we define $f^*(z) := \overline{f(\bar{z}^{-1})} = \sum_{k=0}^{\infty} \bar{f}_k z^k$, which is analytic in $\mathbb{D}$, and note that this means that $\Phi(z) := f(z) + f^*(z) \geq 0$ on $\mathbb{T}$. The rational covariance extension problem is then equivalent to looking for a positive real f of the form

$$f(z) = \gamma \frac{\tilde{b}(z)}{\tilde{a}(z)}, \tag{2.2.4}$$

where $\gamma > 0$ is a constant and $\tilde{a}$ and $\tilde{b}$ are Schur polynomials[18] of degree n, such that $f_0 = \frac{1}{2}c_0$ and $f_k = c_k$ for $k = 1, \ldots, n$. This last set of conditions can equivalently be interpreted as interpolation conditions for f and its first n complex derivatives as $z \to \infty$, or expressed in terms of $f^*(z)$ as interpolation conditions on the function and its first n complex derivatives in the origin $z = 0$. It can thus be seen as a Nevanlinna-Pick interpolation problem [32, Sec. 9.2] [43, Sec. I.2], but where the condition on f being positive-real of the form (2.2.4) makes it nonstandard. This problem has been investigated in great depth, leading to generalizations of the results to other Nevanlinna-Pick-type interpolation problems [22]. However, there is a subtle difference between the two formulations, and the problem (RCEP) needs to be slightly rephrased in order to make them equivalent, cf. [95, Sec. 4] (Section A.4 in this thesis).

2.3 Convex optimization and duality

As we saw an example of in Theorem 2.2.1, optimization is a useful tool in many areas. This section will define and explain some of the concepts brought up in Theorem 2.2.1, such as a *dual problem*. In particular, the form of optimization we will consider here are problems of the form: Given a Banach or Hilbert space X and a function $f : X \to \overline{\mathbb{R}}$, called the *objective function*, find $\hat{x}$ that minimizes f. We often write this problem as $\min_{x \in X} f(x)$. However, in some cases it is not possible to find a minimizer. For example, there might only be a sequence of points x_k such that $f(x_k)$ converges to a greatest lower bound of f but where x_k does not converge to a feasible point that attains this greatest lower bound. Therefore, in cases where we do not know if a minimizer exists we use the notation

$$\inf_{x \in X} f(x). \tag{2.3.1}$$

In many cases we want to find the global minimum $\hat{x}$, i.e., $\hat{x}$ such that $f(\hat{x}) \leq f(x)$ for all $x \in X$. However, very often we can only guarantee that a minimizer is a

[18] Schur polynomials are polynomials that are monic and with roots inside the unit disc.

local minimizer, i.e., that there exists an $\varepsilon > 0$ such that $f(\hat{x}) \leq f(x)$ for all $x \in B_\varepsilon(\hat{x}) := \{x \in X \mid \|\hat{x} - x\| \leq \epsilon\}$ [75, p. 177]. That is, unless the problem is a so-called *convex optimization* problem since in this case any local minimizer is a global minimizer, cf. [75, p. 191] [96, p. 264] [8, Prop. 11.4]. In this section we will present part of the rich theory, especially the one related to convex optimization. Standard references in this area are [75, 96, 8]. In particular, the emphasis in this section is on *duality* in convex optimization, which has been of great use in many of the appended articles. However, before we can introduce this properly we need to introduce some other concepts.

Epigraphs and proper, convex, lower semi-continuous functions

A problem of the form (2.3.1) is only of interest if $f(x) > -\infty$ for all $x \in X$, and if there is at least one point $x_0 \in X$ so that $f(x_0) < \infty$. Such functions are called *proper* [96, p. 24]. One might ask why functions that can take the value ∞ is of interest in optimization. One answer is that this allows us to also write constrained optimization problems in the form (2.3.1) by using indicator functions, i.e.,

$$\mathrm{I}_{\mathcal{C}}(x) = \begin{cases} 0, & \text{if } x \in \mathcal{C} \\ \infty, & \text{else,} \end{cases}$$

where $\mathcal{C} \subset X$. In fact, in this setting $\min_{x \in \mathcal{C}} f(x)$ is equivalent to $\min_{x \in X} f(x) + \mathrm{I}_{\mathcal{C}}(x)$, since any minimizer of the latter must clearly be such that $x \in \mathcal{C}$. With this in mind, one normally defines the *effective domain* of a function as the set

$$\mathrm{dom}_f := \{x \in X \mid f(x) < \infty\}.$$

The concept of a proper function can also be defined using the *epigraph*. The epigraph of a function $f : X \to \overline{\mathbb{R}}$ is defined as the set [75, p. 192] [96, p. 23] [8, Def. 8.1]

$$\mathrm{epi}_f := \{(r, x) \in \mathbb{R} \times X \mid f(x) \leq r\},$$

i.e., it is the set of all points above the graph of the function. Using this definition, and if we introduce the convention that the "vertical direction" in epi_f is the $\mathbb{R}$-direction, then a function f is proper if $\mathrm{epi}_f \neq \emptyset$ and if it contains no vertical lines. Moreover, f is called a *convex* function if epi_f is a convex set. This definition is equivalent with another definition often used, namely that f is called convex if for all $x_0, x_1 \in X$ and for all $\alpha \in [0, 1]$, $f(\alpha x_0 + (1 - \alpha)x_1) \leq \alpha f(x_0) + (1 - \alpha)f(x_1)$ [75, p. 192] [96, Thm. 4.1] [8, Prop. 8.4]. This latter definition means that a linear approximation of the function between two points is always an over-estimate of the function along the corresponding line. Some of the concepts introduced above are summarized graphically in Figure 2.4.

The last property of interest that we will introduce here is *lower semi-continuity*. A function is continuous if for any sequence $\{x_k\}$ such that $x_k \to \tilde{x}$ as $k \to \infty$,

we have that $\lim_{k\to\infty} f(x_k) = f(\tilde{x})$ [42, p. 110]. A function is called lower semi-continuous if instead $\lim_{k\to\infty} f(x_k) \leq f(\tilde{x})$ [96, p. 51]. We can equivalently define upper semi-continuity, and a function is thus continuous if and only if it is both lower and upper semi-continuous. Now, in terms of the epigraph, one can show that a function is lower semi-continuous if and only if the epigraph is a closed set [96, Thm. 7.1] (cf. [8, Thm. 9.9]).

To conclude this subsection, we note that proper, convex and lower semi-continuous functions are ideal when working with optimization problems. This is because any local minimizer is a global minimizer [75, p. 191] [96, p. 264] [8, Prop. 11.4], and given some extra conditions one can also assert the existence of a minimizer, cf. [96, Thm. 27.1] [8, Prop. 11.14]. For example, the minimum is attained if the sublevel sets, which is the family of sets defined by $\{x \in X \mid f(x) \leq \alpha\}$ for $\alpha \in \mathbb{R}$, are compact, cf. [8, Prop. 11.11 and 11.14].

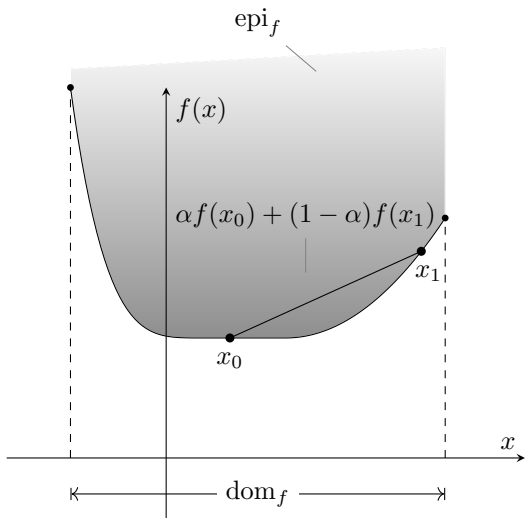


Figure 2.4: Figure illustrating the notions epi$_f$, dom$_f$, and convexity. Note also that the function is lower semi-continuous, and thus the epigraph is closed.

Dual spaces, differentials, and adjoints

To define duality in convex optimization, we first need to define duality in the sense of functional analysis. To this end, consider two normed linear spaces X and Y. First we note that a linear operator $\mathcal{A} : X \to Y$ is continuous if and only if it is bounded [75, p. 144] [42, Thm. 4.4.2], where the operator norm is defined as

$$\|\mathcal{A}\|_{\text{op}} := \sup_{\substack{x\in X \\ x\neq 0}} \frac{\|\mathcal{A}(x)\|_Y}{\|x\|_X} = \sup_{\substack{x\in X \\ \|x\|_X=1}} \|\mathcal{A}(x)\|_Y.$$

Therefore the term bounded and continuous will be used interchangeably. Next, for a normed linear space X we define the (normed) *dual space* of X, denoted X^*, as the space of all bounded linear functionals on X. A fundamental result for what follows is that this dual space X^* is in fact a Banach space, i.e., a complete normed linear space [75, Sec. 5.2][42, p. 150]. However, if this is a Banach space we can also define the dual space of X^*, which consists of all bounded linear functionals on X^*. This space is called the second dual of X and is denote X^{**} [75, Sec. 5.6][42, Sec. 4.10]. Now, for any two elements $x_1^*, x_2^* \in X^*$ and any two scalars $a_1, a_2 \in \mathbb{R}$, note that for any $x \in X$ we have by the standard definition of operations on functionals that $(a_1x_1^* + a_2x_2^*)(x) = a_1x_1^*(x) + a_2x_2^*(x)$. This shows that, in fact, all $x \in X$ can be seen as linear functionals on X^*. Moreover, it can be easily shown that they all correspond to bounded linear functionals, since by the definition of the operator norm we have that $|x^*(x)| \leq \|x^*\|_{\text{op}}\|x\|_X$ and thus the operator norm of $x \in X$ as a functional on X^* is simply $\|x\|_X$ (cf. [75, pp. 115-116] [42, p. 159]). We can thus identify X with a subset of X^{**}. However, in general $X \neq X^{**}$; spaces for which $X = X^{**}$ are called *reflexive* [75, p. 116] [42, p. 160].

In passing, we also note that Hilbert spaces are self-dual, i.e., $H^* = H$ [75, p. 109] [98, Thm. 4.12]. This means that all bounded linear functionals h^* on a Hilbert space can be written as an inner product $h^*(\cdot) = \langle h_1, \cdot \rangle_H$, for some $h_1 \in H$. This motivates the notation for the so-called *dual paring* [75, Sec. 5.2] in Banach spaces, i.e., a mapping $\langle \cdot, \cdot \rangle_X : X^* \times X \to \mathbb{R}$ where for any fixed elements $x \in X$ and $x^* \in X^*$ we define $\langle x^*, x \rangle_X := x^*(x)$. Using this dual paring we can define *weak convergence* and *weak* convergence* [75, Sec. 5.10][42, p. 161 and 170]. A sequence $\{x_k\}_k \subset X$ is said to converge weakly to x if for every $x^* \in X^*$ we have that $\langle x^*, x_k \rangle_X \to \langle x^*, x \rangle_X$. Similarly, a sequence $\{x_k^*\}_k \subset X^*$ is said to converge in weak* to x^* if for every $x \in X$ we have that $\langle x_k^*, x \rangle_X \to \langle x^*, x \rangle_X$. Using these types of convergence, a weaker version of the Bolzano-Weierstrass theorem about bounded sequences on $\mathbb{R}$ can be generalized to Banach spaces, cf. [97, Thm. 3.6] and [98, Thm. 11.29].

Remark 2.3.1. Note the difference between weak convergence and weak* convergence on X^*. In the former we take "test points" from the (normally larger) space X^{**}, while in the latter we take "test points" from the (normally smaller) space X. Only for reflexive spaces are the two notions equal.

Using the above notions, we can now define well-known concepts such as *directional derivatives*, *gradients*, and *subgradients*. To this end, a function $f : X \to \overline{\mathbb{R}}$ is said to have a *Gâteaux differential* (directional derivative) at a point x in the direction h if

$$\delta f(x; h) := \lim_{\alpha \to 0} \frac{1}{\alpha} \Big(f(x + \alpha h) - f(x) \Big) = \frac{\partial}{\partial \alpha} f(x + \alpha h) \Big|_{\alpha = 0}$$

exists [75, p. 171]. If it exists for all $h \in X$, the function is called *Gâteaux differentiable* at x. Moreover, the function is said to be *Fréchet differentiable* at x if

there exists a bounded linear operator $\delta f(x;\cdot) : X \to \mathbb{R}$ such that [75, p. 172]

$$\lim_{\|h\|_X \to 0} \frac{|f(x+h) - f(x) - \delta f(x;h)|}{\|h\|_X} = 0.$$

It is customary to have the same notation for the two expressions since if the function is Fréchet differentiable in a point x, then it is also Gâteaux differentiable in this point and the two differentials are equal [75, p. 173]. The notion of Fréchet differentiable is thus stricter than that of Gâteaux differentiable. Moreover, note that $\delta f(x;\cdot)$ is a bounded linear functional on X and thus by definition $\delta f(x;\cdot) \in X^*$. This observation can be used to generalize the concept of a gradient. This can be done by noting that i) $\mathbb{R}^n$ is a Hilbert space and thus $(\mathbb{R}^n)^* = \mathbb{R}^n$, and ii) that a directional derivative in a direction h of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is given by the inner product with the gradient, i.e., $\delta f(x;h) = \langle \nabla f(x), h \rangle_{\mathbb{R}^n}$ [97, pp. 218-219]. With this in mind we define the gradient of a function $f : X \to \mathbb{R}$, if it exists, as the element $\nabla f \in X^*$ such that the Fréchet derivative in a direction h can be written as [75, p. 175]

$$\delta f(x;h) = \langle \nabla f(x), h \rangle_X.$$

The interest in the differentials from an optimization perspective comes from the fact that the differentials gives information about candidate optimal solutions. More precisely, for everywhere Gâteaux differentiable functions we must, in the unconstrained case, have that $\delta f(x_0;h) = 0$ for all $h \in X$ in order for x_0 to be an extreme point and hence a potential minima [75, p. 178]. Equivalent statements can also be made for the constrained case, generalizing the first-order optimality conditions (known as the KKT-conditions, cf. [96, pp. 280-281]) to Banach spaces, cf. [75, Sec. 7.7]

Even if a function does not poses a Gâteaux of Fréchet differential we can define another notion of differential with is very useful, especially in the case of convex functions. This is the so-called *subdifferential*. The subdifferential of a function $f : X \to \overline{\mathbb{R}}$ at a point x_0, denoted $\partial f(x_0)$, is defined as the set

$$\partial f(x_0) := \{x^* \in X^* \mid f(\tilde{x}) \geq f(x_0) + \langle x^*, \tilde{x} - x_0 \rangle_X \text{ for all } \tilde{x} \in X\},$$

cf. [75, p. 237] [96, p. 214] [8, Def. 16.1]. This means that the mapping $x \mapsto \partial f(x)$ is a *set-valued map*, which we denote by $\partial f : X \rightrightarrows X^*$. Moreover, if no such x^* exists then we say that $\partial f(x) = \emptyset$. Also, by definition if $f(x) = \pm\infty$ we say that $\partial f(x) = \emptyset$. While $\partial f(x)$ is called the subdifferential, an element $x^* \in \partial f(x)$ is called a *subgradient*. One of the nice properties of the subdifferential is that $\hat{x}$ is a global minimizer of a proper function f if and only if 0 is a subgradient in x, i.e., if and only if $0 \in \partial f(\hat{x})$. This is easily seen from the definition, and sometimes referred to as Fermat's rule [8, Thm. 16.2].

Finally, we shortly turn our attention to operators between Banach spaces. For a continuous linear operator $\mathcal{A} : X \to Y$, the *adjoint operator* is defined as the unique

continuous linear operator $\mathcal{A}^* : Y^* \to X^*$ such that for all $x \in X$ and $y \in Y^*$ [75, Sec. 6.5][42, Def. 4.13.1]

$$\langle y^*, \mathcal{A}x \rangle_Y = \langle \mathcal{A}^* y^*, x \rangle_X.$$

Dual optimization problems

Having introduced the above notions from functional analysis, we can now introduce the notion of duality in optimization. Loosely expressed, the idea behind duality is to analyze the problem (2.3.1) via another problem. For this to make sense the two problems need to be linked in some way. Moreover, the related problem needs to be "easier" and more "well-behaved" than the original problem, for example it could be convex and the minimum could be attained. Duality theory can also be used to derive so-called optimality conditions, which are necessary and/or sufficient conditions that characterize which $x \in X$ that could be optimal solutions to (2.3.1). An example of this mentioned above are the so-called KKT-conditions, cf. [96, pp. 280-281].

We will here precede as is often done classically, and present *Lagrangian* and *Fenchel* duality separately. However, the two methods are equivalent [77], and the last paragraph will present a somewhat more unified viewpoint.

Lagrangian duality Given the problem to minimize $f : X \to \overline{\mathbb{R}}$ subject to $x \in \mathcal{C} \subset X$, a *relaxation* of the problem is any optimization problem

$$\begin{aligned} \inf_{x \in X} \quad & f_{\text{relax}}(x) \\ \text{subject to} \quad & x \in \mathcal{C}_{\text{relax}} \end{aligned}$$

such that $\mathcal{C} \subset \mathcal{C}_{\text{relax}}$ and such that $f_{\text{relax}}(x) \leq f(x)$ for all $x \in \mathcal{C}$. From this it is clear that if $\hat{x}$ is optimal to the former and $\hat{x}_{\text{relax}}$ is optimal to the latter, then $f_{\text{relax}}(\hat{x}_{\text{relax}}) \leq f(\hat{x})$. This kind of formulation is especially interesting if the *relaxation is tight*, i.e., if $\hat{x}_{\text{relax}} \in \mathcal{C}$ while also fulfilling $f_{\text{relax}}(\hat{x}_{\text{relax}}) = f(\hat{x}_{\text{relax}})$. In this case, $\hat{x}_{\text{relax}}$ is clearly a minimum also to the original problem.

This is a rather general example of the principle explained above, in which the original problem is analyzed via a related problem. On particular type of relaxation is so-called Lagrangian relaxation. To explain this theory in a general setting, we need the concept of a *convex cone* and its *dual cone* (*positive conjugate cone*). For a normed linear space X, a subset $P \subset X$ is called a *cone* if for all $x \in P$ we have that $\alpha x \in P$ for all scalars $\alpha \geq 0$ [75, p. 18] [96, p. 13] [8, p. 1]. Moreover, the cone is called a convex cone if it is also a convex set. Now, give a cone $P \subset X$, we say that $x_1 \geq_P x_2$ (sometimes just abbreviated $x_1 \geq x_2$) if $x_1 - x_2 \in P$. Moreover, we say that $x_1 >_P x_2$ if $x_1 - x_2$ is in the interior of P [75, p. 214]. Finally, we call

$$P^* := \{x^* \in X^* \mid \langle x^*, x \rangle_X \geq 0 \text{ for all } x \in P\}$$

the dual cone of P [75, p. 157 and 215].

Remark 2.3.2. Note that the inequality with respect to a proper convex cone P, i.e., a cone such that $P \cap -P = \{0\}$, introduces a partial ordering on X. An illustrative example is $X = \mathbb{R}$ and the convex cone given by the nonnegative real numbers: $P = \mathbb{R}_+$. In this case we recover the "ordinary inequality" between real numbers, since for the latter we trivially have that $\alpha_1 \geq \alpha_2$ is equivalent with $\alpha_1 - \alpha_2 \geq 0$. In fact, in this case the cone inequality gives rise to a total ordering of $\mathbb{R}$. However, in general inequalities defined with respect to convex cones only give a partial ordering. As an example, consider the convex cone that is the positive orthant in $\mathbb{R}^2$. In this case, it induces no ordering between, e.g., $[1, 0]^T$ and $[0, 1]^T$. Nonetheless, inequalities with respect to convex cones can be seen as an attempt to generalize well-known ideas and concepts regarding inequalities from the real numbers.

Having introduced the above notions we now consider the optimization problem

$$\inf_{x \in X} \quad f(x) \tag{2.3.2a}$$

$$\text{subject to} \quad g(x) \leq_P 0. \tag{2.3.2b}$$

Here, X, Y are Banach spaces, $f : X \to \overline{\mathbb{R}}$ is a convex function, $P \subset Y$ is a convex cone, and $g : X \to Y$ is a convex function with respect to the convex cone $P \subset Y$. That g is convex means that $g(\alpha x_1 + (1 - \alpha)x_2) \leq_P \alpha g(x_1) + (1 - \alpha)g(x_2)$ for all $x_1, x_2 \in X$ and $\alpha \in [0, 1]$ [75, p. 215] [8, Def. 19.22] (cf. the definition of convexity for functions $f : X \to \overline{\mathbb{R}}$ given above). Note that the convexity of g is not intrinsic to the function, but depends on the cone P. Also note that the convexity of g with respect to the cone P ensures that the feasible region $\mathcal{C}_{g,P} := \{x \in X \mid g(x) \leq_P 0\}$ is convex. We now consider a special type of relaxation for the problem (2.3.2) called the Lagrangian relaxation. To this end, we introduce the *Lagrangian function*

$$L(x, y^*) := f(x) + \langle y^*, g(x) \rangle_Y, \tag{2.3.3}$$

where $y^* \in Y^*$ are known as the Lagrangian multipliers [75, Sec. 8.3] [96, p. 280] [8, Rem. 19.24]. Now, for a fixed $y^* \in Y^*$ consider the problem

$$\inf_{x \in X} L(x, y^*). \tag{2.3.4}$$

First, the feasible region, i.e., $\operatorname{dom}_{L(\cdot, y^*)}$, in (2.3.4) clearly contains that of (2.3.2), i.e., $\operatorname{dom}_{f + I_{\mathcal{C}_{g,P}}}$. Second, for each fixed $y^* \in P^*$ we have that $L(x, y^*) = f(x) + \langle y^*, g(x) \rangle_Y \leq f(x)$ for all $x \in \mathcal{C}_{g,P}$ and thus for all x fulfilling (2.3.2b). This shows that for each $y^* \in P^*$, (2.3.4) is a relaxation of (2.3.2).

Having obtained a relaxation of the original problem (2.3.2), we can now try to make the relaxation as tight as possible. This can be done by introducing the so-called *dual function* $\varphi : Y^* \to \overline{\mathbb{R}}$, which is simply defined as $\varphi(y^*) := \inf_{x \in X} L(x, y^*)$ [75, p. 223]. Making the relaxation as tight as possible is then to find the $y^* \in P^*$ that maximizes φ, i.e., we get the *Lagrangian dual problem*

$$\sup_{y^* \in P^*} \varphi(y^*) = \sup_{y^* \in P^*} \inf_{x \in X} L(x, y^*) = \sup_{y^* \in P^*} \inf_{x \in X} f(x) + \langle y^*, g(x) \rangle_Y. \tag{2.3.5}$$

Under appropriate conditions one can show that (2.3.2) and (2.3.5) have the same value, and that the optimal solution is actually attained in the primal and/or the dual problem. The most common such condition is the so-called Slater condition: If the primal problem (2.3.2) has a finite optimal value, and if there exists an $x_0 \in X$ such that $g(x_0) <_P 0$, then the dual problem (2.3.5) takes the same value and there exists a $\hat{y}^* \in P^*$ that achieves it [75, p. 224]. Moreover, in this case the conditions for optimality can be expressed as saddle-point conditions for the Lagrangian (2.3.3), cf. [75, Chp. 8], [96, Chp. 28], [8, Sec. 19.4]. The theory can also be extended to handle equality constraints, however these must be affine for the primal problem to be convex, cf. [8, Sec. 19.3].

Remark 2.3.3. As a final remark on Lagrangian relaxation we note that the Lagrangian (2.3.3) of a problem (2.3.2) can be constructed irrespective of if the problem is convex or not. Some of the theory presented above can also be extended to the nonconvex case, cf. [75, Sec. 8.4, 9.3, and 9.4].

Fenchel duality Given a function $f : X \to \overline{\mathbb{R}}$, the *convex conjugate* (*Fenchel conjugate* or *Legendre transform*) is defined as the function $f^* : X^* \to \overline{\mathbb{R}}$ [75, Sec. 7.10] [96, Chp. 12] [8, Def. 13.1]

$$f^*(x^*) := \sup_{x \in X} \langle x^*, x \rangle_X - f(x).$$

From this definition it follows that the objective function value of the globally optimal solution to (2.3.1) is given by $-f^*(0)$. But f^* also have many other interesting properties. For example, let us define the second conjugate of f as the convex conjugate of f^*, which we denote f^{**}. For reflexive spaces X we have that if f is proper, then $f = f^{**}$ if and only if f is lower semi-continuous and convex, cf. [8, Prop. 13.32][75, p. 198]. In fact, in this case of reflexive spaces we have that f^{**} is the closed convex hull of f, cf. [96, Thm. 12.2].

Now, note that irrespectively of if f is convex or not, f^* is always a convex function [75, p. 196] [8, Prop. 13.11]. Moreover, from the definition it follows that if f is proper then $f^*(x^*) > -\infty$ for all x^*. In this case we also have that for all $x^* \in X^*$, $f^*(x^*) = \sup_{x \in X} \langle x^*, x \rangle_X - f(x) \geq \langle x^*, x \rangle_X - f(x)$ for all $x \in X$, with the implicit understanding that the left side might be ∞ and the right side might be $-\infty$. For all $x \in \mathrm{dom}_f$, i.e., where $f(x)$ is finite, the right hand side is larger than $-\infty$ and we can rearrange the terms to read $f(x) + f^*(x^*) \geq \langle x^*, x \rangle_X$, where the left hand side might still be ∞. Moreover, it is trivially extended from $x \in \mathrm{dom}_f$ to $x \in X$ since f is proper and thus $f^*(x^*) > -\infty$ for all $x^* \in X^*$. This gives us the so-called Fenchel-Young inequality, namely that if f is a proper function then

$$\langle x^*, x \rangle_X \leq f(x) + f^*(x^*) \quad \text{for all } x \in X \text{ and } x^* \in X^*, \tag{2.3.6}$$

cf. [96, p. 105][8, Prop. 13.13].

Next, consider the problem $\inf_{x\in X} f(x)+g(x)$ where $f, g : X \to \overline{\mathbb{R}}$ are proper functions. The *Fenchel dual problem* of this optimization problem is defined to be

$$\sup_{x^*\in X^*} -f^*(x^*) - g^*(-x^*),$$

cf. [8, Def. 15.10]. If we let $\nu := \inf_{x\in X} f(x)+g(x)$ and $\nu^* := \sup_{x^*\in X^*} -f^*(x^*) - g^*(-x^*)$, then $\nu \geq \nu^*$, cf. [8, Prop. 15.12]. This is called the *duality gap*. In some cases it can be shown that the duality gap is zero and that the dual problem in fact attains its optimal solution. One such example is when X is a Hilbert space, f and g are proper, convex, and lower semi-continuous, and 0 belongs to the interior of the set $\mathrm{dom}_f - \mathrm{dom}_g$ [8, p. 91 and Prop. 15.13].

This type of duality can also be extended to incorporate other problems. The first type is of the form $\inf_{x\in X} f(x) - g(x)$, where $g(x)$ is a concave function. This can be done by handling $-g$ as a convex function, but can also be done by an appropriate definition of the *concave conjugate* of concave function, see [75, Sec. 7.11 and 7.12] or [96, p. 308 and Thm. 31.1]. The second type of extension is to so-called *Fenchel-Rockafellar duality*. In this case one considers problems of the form $\inf_{x\in X} f(x)+g(Lx)$, where $L : X \to Y$ is a bounded linear operator. As we will see in Section 2.4, this type of problems is common in variational regularization of inverse problems. The dual of this problem is $\sup_{y^*\in Y^*} -f^*(L^*y^*) - g^*(-y^*)$, cf. [8, Def. 15.19], and similarly the duality gap can sometimes be shown to be zero, cf. [8, Prop. 15.22].

As a final remark in this paragraph, we note that the convex conjugate of a function f can be interpreted as defining nonvertical supporting hyperplanes for epi_f. To see this, note that $\mathrm{epi}_f \subset \mathbb{R}\times X$, so a linear functional in this space is of the form $(r^*, x^*) \in \mathbb{R}\times X^*$. Since the hyperplane is nonvertical we have $r^* \neq 0$, and thus without loss of generality we can always take $r^* = -1$ by simply scaling appropriately. For a fixed $x_0^* \in X^*$, assume that the sup in the definition of f^* is attained in a point x_0. Now, consider the linear functional $(r, x) \mapsto \langle x_0^*, x\rangle_X - r$ and the corresponding hyperplane $\langle x_0^*, x\rangle_X - r = f^*(x_0^*)$. This gives that in the point $x = 0$ we have $r = -f^*(x_0^*)$, and in the point $x = x_0$ we have $r = f(x_0)$.[19] The latter point is thus a point on the boundary of epi_f, however for all $(r, x) \in \mathrm{epi}_f$ we have that $r \geq f(x)$ and thus by the definition of the Fenchel conjugate we have that

$$\langle x_0^*, x\rangle - r \leq \langle x_0^*, x\rangle - f(x) \leq f^*(x_0^*) \quad \forall\, x \in X. \tag{2.3.7}$$

Therefore, epi_f is contained on one side of the hyperplane $\langle x_0^*, x\rangle_X - r = f^*(x_0^*)$. This means that it is a supporting hyperplane of epi_f that is tangential in the point $(f(x_0), x_0)$. A graphic illustration of this is shown in Figure 2.5. Moreover, since we assumed that $f^*(x_0^*) = \langle x_0^*, x_0\rangle - f(x_0)$, using (2.3.7) we can see that this also means that x_0^* is a subgradient of f in the point x_0. In fact, the converse is also

[19] This is true since we assumed that the supremum that defines $f^*(x_0^*)$ is attained in x_0. Equivalently, in $(x_0^*, x_0) \in X^*\times X$ the Fenchel-Young inequity (2.3.6) is an equality.

true, i.e., if f is a proper function then we have the equivalence

$$x_0^* \in \partial f(x_0) \quad \Longleftrightarrow \quad \begin{array}{l}\text{the Fenchel-Young inequality (2.3.6)} \\ \text{is fulfilled with equality in } (x_0^*, x_0),\end{array}$$

cf. [96, Thm. 23.5][8, Prop. 16.9].

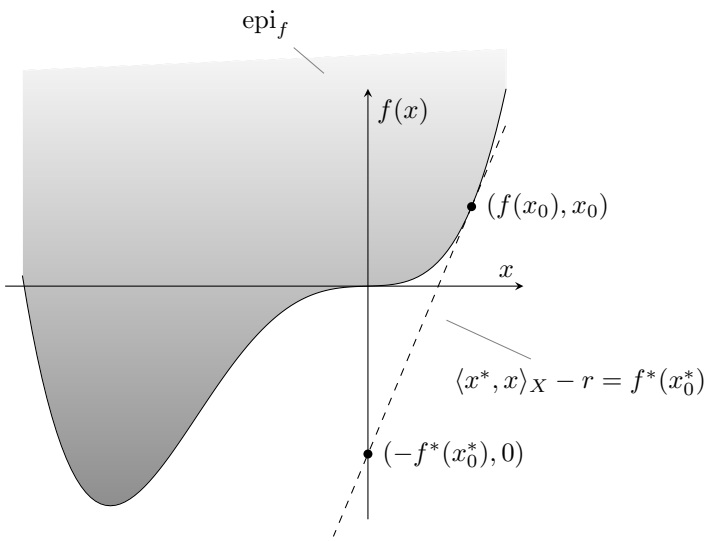


Figure 2.5: The convex conjugate defines supporting hyperplanes for the epigraph.

Lagrangian duality via convex conjugates In this paragraph we will see that the two dualities defined above are tightly linked. In fact, strong duality for Fenchel-type primal-dual problems can be obtained using Lagrangian-type arguments, and strong duality for Lagrangian-type primal-dual problems can be obtained using Fenchel-type arguments [77]. However, the material can also be unified using a perturbation-type theory, and this is the path taken here. The material presented is a summary of [96, Chp. 29-30], [8, Chp. 19].[20]

To this end, let X, Y be reflexive Banach spaces and consider the standard minimization problem (2.3.1), where f is assumed to be convex. As noted before, this formulation can include constraints by defining f to be ∞ on some regions. We now define a *perturbation function* to f as a function $\Gamma : X \times Y \to \overline{\mathbb{R}}$ such that

$$\Gamma(x, 0) = f(x).$$

[20] Although the material can be found in the above references, this presentation is inspired by presentations found on the two web pages `https://math.stackexchange.com/questions/948862/fenchel-dual-vs-lagrange-dual` and `https://mathematix.wordpress.com/2017/05/07/lagrange-vs-fenchel-duality/`, which both give a nice overview of it.

Here, we consider convex perturbation functions, i.e., functions Γ such that $\text{epi}_\Gamma \subset \mathbb{R} \times X \times Y$ is a convex set. The primal problem is given by $\inf_{x \in X} \Gamma(x, 0)$, but for each $y \in Y$ we may also consider $\inf_{x \in X} \Gamma(x, y)$. This is a function of y, and we define $\nu : Y \to \overline{\mathbb{R}}$ to be

$$\nu(y) := \inf_{x \in X} \Gamma(x, y).$$

Now, consider the convex conjugate of Γ, i.e., $\Gamma^* : X^* \times Y^* \to \overline{\mathbb{R}}$ defined by

$$\Gamma^*(x^*, y^*) := \sup_{\substack{x \in X \\ y \in Y}} \langle x^*, x \rangle_X + \langle y^*, y \rangle_Y - \Gamma(x, y).$$

From this definition we see that

$$\begin{aligned} \Gamma^*(0, y^*) &= \sup_{\substack{x \in X \\ y \in Y}} \langle y^*, y \rangle_Y - \Gamma(x, y) = \sup_{y \in Y} \langle y^*, y \rangle_Y - \inf_{x \in X} \Gamma(x, y) \\ &= \sup_{y \in Y} \langle y^*, y \rangle_Y - \nu(y) = \nu^*(y^*). \end{aligned}$$

Now note that the solution to (2.3.1) is by definition given by $\nu(0)$. However, since we assume that the spaces are reflexive we have that $Y^{**} = Y$. Therefore $\nu^{**} : Y \to \overline{\mathbb{R}}$. Moreover, $\nu^{**}(y) \leq \nu(y)$ for all $y \in Y$ (cf. the paragraph on Fenchel duality above) and if ν is proper, convex and lower semi-continuous then $\nu^{**} = \nu$, cf [8, Prop. 19.11 and 19.12]. Furthermore, by definition

$$\nu^{**}(y) = \sup_{y^* \in Y^*} \langle y, y^* \rangle_{Y^*} - \nu^*(y^*).$$

Taking all of this together we get a primal-dual pair of problems via

$$\inf_{x \in X} \Gamma(x, 0) = \inf_{x \in X} f(x) = \nu(0) \geq \nu^{**}(0) = \sup_{y^* \in Y^*} -\nu^*(y^*) = \sup_{y^* \in Y^*} -\Gamma^*(0, y^*).$$

Lagrangian duality can now be recovered from this frame-work by perturbing (2.3.2) and considering

$$\begin{aligned} \inf_{x \in X} \quad & f(x) \\ \text{subject to} \quad & g(x) \leq y, \end{aligned}$$

cf. [75, p. 216]. By introducing the set $\mathcal{C}_{g,P}(y) := \{x \in X \mid g(x) \geq_P y\}$, the indicator function $I_{\mathcal{C}_{g,P}(\cdot)}(x, y)$, and by defining $\Gamma(x, y) := f(x) + I_{\mathcal{C}_{g,P}(\cdot)}(x, y)$, we can recover similar results as described in the paragraph on Lagrangian duality. Also the saddle-point properties can be recovered by an appropriate definition of a Lagrangian for Γ, cf. [96, p. 296 and Cor. 30.5.1] [8, pp. 280-281].

Remark 2.3.4. As a final remark in this section we note that in the above presentation of optimization, convex problems have indirectly been portrayed as if they are "easy" to solve. Although convex problems are normally easier than nonconvex problems,

convex problems can also be "hard" to solve. For example, the problem of finding the minimizer of a multidimensional polynomial $p : \mathbb{R}^n \to \mathbb{R}$ is well-known to be a "hard" problem. Still this can be reformulated as a convex optimization problem, in fact even a linear one, namely (cf. [66, pp. 5-6])

$$\begin{aligned} \inf_{\mu \in \mathrm{BV}(\mathbb{R}^n)} \quad & \int_{\mathbb{R}^n} p(x) d\mu(x) \\ \text{subject to} \quad & \mu(\mathbb{R}^n) = 1 \\ & \mu \geq 0, \end{aligned}$$

where $\mathrm{BV}(\mathbb{R}^n)$ is the space of singed finite measures of bounded variation, i.e. the dual space of the space of continuous functions that vanish at infinity, cf. [98, Thm. 6.19]. In the above formulation, the inequality $\mu \geq 0$ simply means that μ is a finite measure, or expressed equivalently that any Hahn-decomposition $\mathbb{R}^n = A \cup B$ for μ, such that A carries the positive mass and B the negative mass, is such that $\mu(B) = 0$ [42, Sec. 1.10] [98, pp. 119-126]. If an optimal solution to the above problem exists and is finite, then it is given by an impulse in the global minimizer $\hat{x}$ of $p(x)$, and thus solves the original "hard" problem.

Another example is the standard quadratic programming problem $\min_{x \in \mathbb{R}^n} x^T Q x$ subject to $\sum_{i=1}^n x_i = 1$ and $x \geq 0$, but where the Hessian Q is not necessarily positive definite. This is known to be a "hard" problem, but it can be relaxed to a convex problem over the cone of so-called completely positive matrices, i.e., matrices in the convex cone $\{A \in \mathbb{R}^{n \times n} \mid A = \sum_{j=1}^m a_j a_j^T \text{ where } a_j \in \mathbb{R}^n_+ \text{ and } m \text{ finite}\}$. Moreover, this relaxation can then be shown to be tight, meaning that the original "hard" problem can be solved by solving this convex relaxation. For more details we refer the read to [34] and references therein.

2.4 Inverse problems, ill-posedness, and variational regularization

Mathematically, an inverse problem can be stated as the problem of reconstructing an entity $f_{\text{true}} \in X$ representing the object under investigation from data $g \in Y$, assuming that the two are related according to

$$g = \mathcal{A}(f_{\text{true}}) + \delta g. \tag{2.4.1}$$

Here, $\mathcal{A} : X \to Y$ is the so-called *forward operator*, which models how the data is formed in the absence of noise. Moreover, X and Y are suitable Hilbert or Banach spaces, often denote the *reconstruction space* and *data space*, respectively. Finally, δg is a Y-valued random element which correspond to the noise that will inevitably be present in data. Note that more advanced noise models than simple additive noise can of course also be considered; here we limit ourselves to this case just in order to simplify the exposition.

In contrast, the *direct problem* can be see as obtaining a sufficiently good model for the forward operator $\mathcal{A}$. If possible, this is often done by first-principle modeling, however, this might be hard or even impossible. One example is when $\mathcal{A}$ represents the solution operator of a PDE that might not have a unique closed-form-type expression for the solution. In some cases one can also apply system identification procedures to obtain a model for the forward operator. Either way, aspects that are important when deriving the model is that the operator $\mathcal{A}$ captures the relevant physics while still remaining mathematically tractable. Here, we will not dwell further on these aspects and instead we assume that we have access to an appropriate forward operator $\mathcal{A}$.

Solving inverse problems are of course (relatively) easy if the inverse operator $\mathcal{A}^{-1} : Y \to X$ exists, the inverse is "well-conditioned", and the noise level is low. However, many inverse problems of interest are so-called *ill-posed* inverse problems, which loosely speaking means that there exists no "well-condition" inverse. In this case, an arbitrarily small amount of noise could lead to an arbitrarily bad approximation of f_{true}. The notion ill-posed was introduce by Hadamard: An inverse problem is said to be *well-posed* if

i) for each data there exists at least one solution to the problem,
ii) for each data the solution is unique,
iii) the solution depends continuously on data,

and otherwise it is called an ill-posed problem [37, p. 31] [61, p. 9].

A special case of interest that gives rise to ill-posed problems is when $\mathcal{A} : X \to Y$ is a so-called *compact linear operator* and when X is infinite-dimensional. To be precise, a continuous linear operator $\mathcal{A} : X \to Y$, where X and Y are Banach spaces, is called compact if for all bounded $\Omega \subset X$ the image $\mathcal{A}(\Omega) \subset Y$ is compact [42, p. 186] [61, Def. A.31]. Now, any operator with a nonempty kernel clearly gives rise to an ill-posed inverse problem since a solution will not be unique. Therefore, only operators that are injective (one-to-one) can be give rise to well-posed problems. However, if $\mathcal{A}$ is compact and injective, although the inverse $\mathcal{A}^{-1}$ exists it is an unbounded operator if $\dim(X) = \infty$ [61, Thm. 1.17] and therefore not continuous [42, Thm. 4.4.2], thus violating point iii) above. This means that any inverse problem involving a compact linear operator $\mathcal{A} : X \to Y$ with $\dim(X) = \infty$ will be ill-posed.

In order to solve ill-posed inverse problems one uses *regularization*. A regularization is a parametrized family of operators $\{\mathcal{A}_\theta^\dagger\}_\theta$ that approximates the inverse mapping. In particular, the family should be such that when the noise δg in the data goes to zero, there is a (at least implicit) selection rule for the parameter θ so that $\mathcal{A}_\theta^\dagger(g) \to f$ when $\theta \to 0$ according to this selection rule [37, Def. 3.1] (cf. [61, Def. 2.1 and Def. 2.3]). One way which is often used to construct such an operator $\mathcal{A}_\theta^\dagger$ is by using so-called *variational regularization*.

Variational regularization

Variational regularization is a type of regularization in which the reconstruction problem is formulated as an optimization problem. This means that we define the reconstruction operator $\mathcal{A}_\theta^\dagger : Y \to X$ as

$$\mathcal{A}_\theta^\dagger : g \mapsto \arg\min_f \mathcal{D}(\mathcal{A}(f), g) + \theta \mathcal{S}(f). \tag{2.4.2}$$

In this formulation, $\mathcal{D} : Y \times Y \to \mathbb{R}$ is the *data discrepancy* function, $\mathcal{S} : X \to \mathbb{R}$ is the *regularization function*, and $\theta \in \mathbb{R}_+$ is the regularization parameter which controls the trade-off between $\mathcal{D}$ and $\mathcal{S}$. These functions need to be chosen appropriately: $\mathcal{D}$ should to be a relevant measure of the data-misfit, and $\mathcal{S}$ needs to encode relevant *a priori* information of the type of reconstructions sought. The latter is done implicitly by letting S penalize undesirable solutions. This means that the functions need to be designed for each application. However, luckily, in many cases there are "standard functions" to try like $\|\mathcal{A}(f) - g\|_Y$ for data discrepancy and 1-norm-type functions as regularization, the latter intended to promote a suitable notion of sparsity in the reconstruction [17].

The formulation in (2.4.2) can also be interpreted from a statistical perspective. To this end, let $p_{\mathtt{g}}(g \mid f)$ be the likelihood of data g, i.e., the probability distribution of observing data g given the "parameters" f. Now, taking $\theta = 0$ in (2.4.2) and taking the data discrepancy function to be the negative log-likelihood, i.e., $D(\mathcal{A}(f), g) = -\log(p_{\mathtt{g}}(g \mid f))$, the operator $\mathcal{A}_{\theta=0}^\dagger$ is the *maximum likelihood* estimator. However, for ill-posed inverse problems this approximate inverse operator is typically still unstable with respect to data, meaning that the solution might not be unique and that small changes in g gives rise to large changes in the reconstruction $\hat{f}_g := \mathcal{A}_{\theta=0}^\dagger(g)$, cf. [9, Sec. 2.1]. To stabilize the reconstruction operator, one can take a Bayesian perspective and introduce a Gibbs prior on f with density $p_{\mathtt{f}}(f) = \frac{1}{c} e^{-\theta \mathcal{S}(f)}$, where c is an appropriate scaling constant. Using Bayes rule, one finds that the *posterior* density takes the form $p_{\mathtt{f}}(f \mid g) \propto p_{\mathtt{g}}(g \mid f) p_{\mathtt{f}}(f)$, and by taking the negative logarithm of this we identify that $\mathcal{A}_\theta^\dagger$ as defined in (2.4.2) is the *maximum a posteriori* estimator [9, Sec. 2.2].

Examples of inverse problems

Finally, we will here outline a few examples of inverse problems. These problems are computed tomography (CT), an inverse problem in magentization, and the rational covariance extension problem.

Computed tomography CT is a noninvasive imaging modality for investigating internal two- or three-dimensional structures of objects by using penetrating waves or particles. It has a wide range of applications, e.g., X-ray CT [82, 81] in medical imaging and electron tomography (ET) [84, 80] in biology and material science. In this section we will mainly focus on the mathematics associated with X-ray CT.

In imaging, the domain of the forward operator $\mathcal{A}$ is a set of real-valued functions $f : \Omega \to \mathbb{R}$, where $\Omega \subset \mathbb{R}^n$ is normally compact and $n = 2$ or 3. In X-ray CT, $f(x)$ corresponds to the attenuation of the X-ray in the point x of the object. To describe a simple model for the physics, let $I(x)$ be the intensity of the X-ray in the point x. Given the attenuation $f(\cdot)$, an X-ray with intensity $I(\cdot)$ that travels a small distance Δx through a point x will have a change (decrease) in intensity described by

$$\Delta I(x) \approx -f(x)I(x)\Delta x.$$

More rigorously, this is described by $\frac{d}{dx}I(x) = -f(x)I(x)$, and integrating this equation along the entire line L that the X-ray travels trough the object gives a model for the data acquisition, namely

$$\log\left(\frac{I_{\text{in}}}{I_{\text{out}}}\right) = \int_L f(x)dx,$$

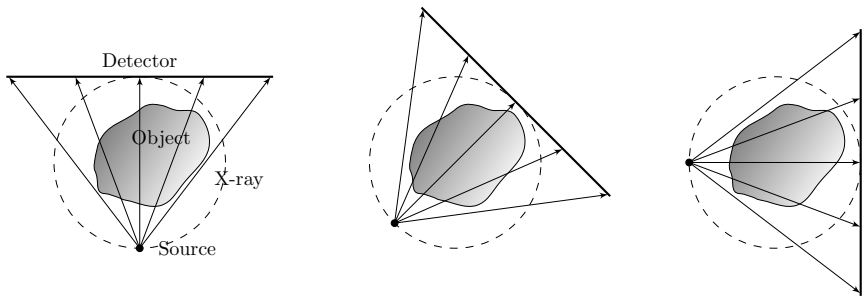
cf. [81, p. 1]. Here, I_{in} is the intensity of the incoming X-ray beam and I_{out} is the intnesity of the outgoing X-ray beam. Both of these are known: I_{in} by the design of the machine and I_{out} by measuring it using detectors. Thus, in tomographic imaging the data acquisition is normally modeled as line integrals of the function f. Moreover, this means that the range of the operator is a set of real-valued functions on the set of lines $\mathbb{M}$ in $\mathbb{R}^n$. This forward operator $\mathcal{A}$ is called the *ray transform*.

The set of lines $\mathbb{M} \subset \mathbb{R}^n$ is a manifold,[21] and to express the action of $\mathcal{A}$ we can introduce coordinates on this manifold. To this end, note that a line in $\mathbb{R}^n$ can be described by a directional vector ω on the unit sphere S^{n-1}, to which the line is parallel, and a point $x \in \mathbb{R}^n$ that it passes through. However, this description is redundant since any of the points along the line can be chosen. To reduce this redundancy, we also enforce that the point must be in the orthogonal complement of the space spanned by unit direction chosen. One set of coordinates on $\mathbb{M}$ is thus $(\omega, x) \in S^{n-1} \times \mathbb{R}^n$ with $x \in \omega^\perp$, where $\omega^\perp \subset \mathbb{R}^n$ is the unique plane through the origin with $\omega \in S^{n-1}$ as its normal vector. In the aforementioned coordinates, the ray transform is expressible as [81, Chp. 2]

$$\mathcal{A}(f)(\omega, x) := \int_{-\infty}^{\infty} f(x + t\omega)dt. \tag{2.4.3}$$

Using the ray transform, tomographic data is modeled as values of $\mathcal{A}(f)(\omega, x)$ for a sampling of $\omega \in S^{n-1}$ and $x \in \omega^\perp$. This is illustrated in Figure 2.6 for a so-called fan-beam geometry. With slight abuse of terminology, one refers to a data point as the "projection" of f along the line given by (ω, x). However, although the continuous transform is in principle invertible [81, Thm. II.2.1], when data is only available from a finite subsample $\{(\omega_k, x_k)\}_{k=1}^{\ell}$ this is no longer true. In fact, in this case the inverse problem is ill-posed [81, pp. 35-36].

[21] In fact, $\mathbb{M}$ is often called the real projective space [68, Ex. 1.5] or a Grassmanian manifold [68, Ex. 1.36].



(a) First position of the source-detector pair.

(b) Second position of the source-detector pair.

(c) Third position of the source-detector pair.

Figure 2.6: Figure illustrating the data collection in X-ray CT using a fan-beam geometry [82, Sec. 3.1.2] [81, Sec. III.3]. The source-detector pair is rotated around the object, and examples of three different rotations are shown in Figures 2.6a, 2.6b, and 2.6c. For each position of the source-detector pair a number of X-rays are emitted from the source and the intensity of these X-rays, after passing through the object, are registered at the detector. Each such detection corresponds to a line integral of the attenuation f at a certain sampling point (ω, x).

Magnetization of thin rock samples This inverse problem comes from geoscience. The goal is to recover the magnetization, $\mathbf{m}$, of a thin rock sample from measurements of the vertical component H_z of the field[22] $\mathbf{H}$ generated by the sample at a given hight h, cf. Figure 2.7. This is of interest for understanding the history and the earths magnetic field, and also the history of magnetic fields of other planets and astorids, see, e.g., [7, 6] and references therein.

Le Ω be a compact set in $\mathbb{R}^3$. The magnetization is then a vector field defined on Ω, i.e., $\mathbf{m} : \Omega \to \mathbb{R}^3$, and is denoted by

$$\mathbf{m} : (x, y, z) \mapsto \begin{bmatrix} m_x(x, y, z) \\ m_y(x, y, z) \\ m_z(x, y, z) \end{bmatrix}.$$

This magnetization $\mathbf{m}$ will produce a field in the ambient space, which we denote by $\mathbf{H}$ or $\mathbf{H}(\mathbf{m})$. For a fixed $\mathbf{m}$, $\mathbf{H}(\mathbf{m})$ is a vector field $\mathbf{H}(\mathbf{m}) : \mathbb{R}^3 \to \mathbb{R}^3$, $(x, y, z) \mapsto [H_x(x, y, z),\ H_y(x, y, z),\ H_z(x, y, z)]^T$. By Maxwell's equations for magnetostatics [51, Chp. 6], for points outside the support of $\mathbf{m}$, i.e., for $\mathbf{x} := (x, y, z) \in \mathbb{R}^3$ such

[22] The naming of this "magnetic field" is debated, see, e.g., [51, p. 271]. In what follows we will therefore only refer to it as $\mathbf{H}$ or as "the field".

that $\mathbf{x} \notin \Omega$, this field is given by

$$\mathbf{H}(\mathbf{m})(\mathbf{x}) = -\mu_0 \nabla \phi(\mathbf{m})(\mathbf{x}), \tag{2.4.4a}$$

$$\phi(\mathbf{m})(\mathbf{x}) = \frac{1}{4\pi} \int_\Omega \frac{\langle \mathbf{m}(\tilde{\mathbf{x}}), \mathbf{x} - \tilde{\mathbf{x}} \rangle_{\mathbb{R}^3}}{\|\mathbf{x} - \tilde{\mathbf{x}}\|^3_{\mathbb{R}^3}} d\tilde{\mathbf{x}}, \tag{2.4.4b}$$

cf. [7, Sec. 2.1]. Here, $\phi(\mathbf{m})$ is the scalar potential that exists by the fact that the filed $\mathbf{H}(\mathbf{m})$ is curl-free in the absence of free external currents [51, p. 269], and μ_0 is the magnetic permeability in vacuum.

In the case of thin samples, the sample is assumed to be a compact set Ω contained in a plane. For simplicity we assume that the plane is oriented in the x-y-plane, call it $\mathbb{R}^2_{z=0}$, and thus the z-unit vector $e_3 = (0, 0, 1)^T$ is normal to it. Moreover, it is assumed that data are measurements of H_z on a plane that is parallel to $\mathbb{R}^2_{z=0}$ and located at a height h (the plane $\mathbb{R}^2_{z=h}$). For an illustration of this set-up, see Figure 2.7. By using (2.4.4), an expression for the forward operator $\mathcal{A} : \mathbf{m} \mapsto H_z(\mathbf{m})$ can be derived. To this end, let P_z be the Poisson kernel for the upper half-space, given by $P_z(\mathbf{x}) := z/(2\pi \|x\|^{3/2}_{\mathbb{R}^3})$, and let P_z* denote the convolution with this kernel "at height" $z > 0$, i.e., the convolution is only performed in the first two coordinates while z is kept fix. Moreover, let R_x and R_y be the Riesz transforms defined for $f \in L_2(\mathbb{R}^2)$ by[23]

$$R_x(f)(x, y) := \lim_{\varepsilon \to 0} \frac{1}{2\pi} \int_{\mathbb{R}^2 \setminus B_\varepsilon(x,y)} f(\tilde{x}, \tilde{y}) \frac{x - \tilde{x}}{((x - \tilde{x})^2 + (y - \tilde{y})^2)^{3/2}} d\tilde{x} d\tilde{y},$$

where $B_\varepsilon(x, y) := \{(\tilde{x}, \tilde{y}) \in \mathbb{R}^2 \mid \sqrt{(\tilde{x} - x)^2 + (\tilde{y} - y)^2} < \varepsilon\}$ is the ball in $\mathbb{R}^2$ of radius ε centered around (x, y). R_y is defined in an equivalent fashion. The action of $\mathcal{A}$ can then be expressed as

$$\begin{aligned} \mathcal{A}(\mathbf{m})(\mathbf{x}) &= -\mu_0 \frac{\partial}{\partial z} \phi(\mathbf{m})(\mathbf{x}) \\ &= -\frac{\mu_0}{2} \frac{\partial}{\partial z} \left[P_z * \Big(R_1\big(m_x(\cdot, \cdot, z)\big) + R_2\big(m_y(\cdot, \cdot, z)\big) + m_z \Big) \right](\mathbf{x}), \end{aligned} \tag{2.4.5}$$

where $\mathbf{x} \in \mathbb{R}^3$ is a point with $z > 0$,[24] see [7, Thm. 2.1] and [6, Sec. 3.1].

Since all operators involved in (2.4.5) are linear, the forward operator $\mathcal{A}$ is a linear operator. Moreover, it is bounded and thus continuous [6, Sec. 3.2] and therefore is also has an adjoint operator [6, Sec. 3.3]. However, the kernel of the operator is nonempty [6, Prop. 2], which makes the recovery of $\mathbf{m}$ from measurements of $H_z(x, y, h)$ an ill-posed linear inverse problem.

The rational covariance extension problem as an inverse problem Although this thesis considered the rational covariance extension problem (RCEP) in

[23] Alternatively, the Riesz transform can be defined using the Fourier transform [6, Eq. (9)].

[24] When sampled on $\mathbb{R}^2_{z=h}$, as described above and indicated in Figure 2.7, clearly $z = h > 0$.

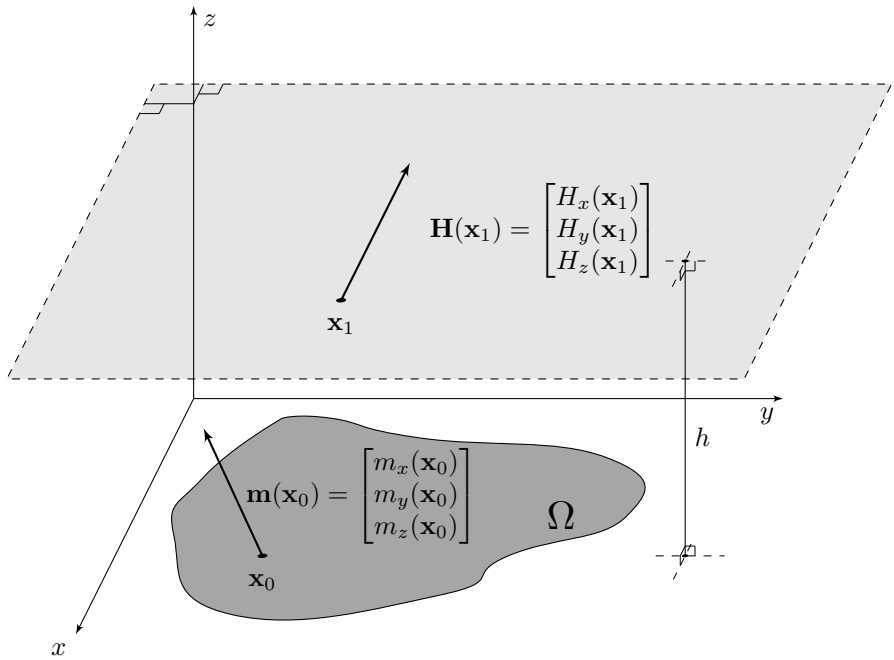


Figure 2.7: Figure illustrating the geometry of the inverse problem in magnetization. $\Omega \subset \mathbb{R}^2_{z=0} := \{(x, y, z) \in \mathbb{R}^3 \mid z = 0\}$ is the compact set on which the magnetization $\mathbf{m}$ has support. Measurements of H_z are taken at hight h above the thin specimen, on the plane parallel to $\mathbb{R}^2_{z=0}$.

the context of system identification, it can also be viewed as an inverse problem. To this end, let the forward operator be $\mathcal{A} : \mathrm{BV}(\mathbb{T}) \to \mathbb{C}^{2n+1}$, where, as before, BV is the space of signed finite measures of bounded variation on $\mathbb{T}$, i.e., the dual space of the space of the continuous functions on $\mathbb{T}$ (called $C(\mathbb{T})$) [75, Sec. 5.5] [98, Thm. 6.19] [42, Thm. 4.14.8]. The action of the operator is given by

$$\mathcal{A} : d\mu \mapsto \left[\int_{-\pi}^{\pi} e^{ik\theta} d\mu(\theta) \right]_{k=-n}^{n},$$

i.e., it maps a measure to the corresponding $2n + 1$ trigonometric moments. We can also derive the adjoint operator $\mathcal{A}^* : \mathbb{C}^{2n+1} \to \mathrm{BV}(\mathbb{T})^*$ by the following calculation:

$$\begin{aligned} \langle \mathcal{A}(d\mu), p \rangle_{\mathbb{C}^{2n+1}} &= \sum_{k=-n}^{n} \left(\int_{-\pi}^{\pi} e^{ik\theta} d\mu(\theta) \right) p_k^* = \int_{-\pi}^{\pi} \left(\sum_{k=-n}^{n} p_k^* e^{ik\theta} \right) d\mu(\theta) \\ &= \langle d\mu, \mathcal{A}^*(p) \rangle_{\mathrm{BV}(\mathbb{T})^*}, \end{aligned}$$

i.e., $\mathcal{A}^* : p \mapsto \sum_{k=-n}^{n} p_k^* e^{ik\theta}$. In fact, from this we see that the range of $\mathcal{A}^*$ is the set of all trigonometric polynomials (not necessarily real-valued), and thus $\mathcal{A}^* : \mathbb{C}^{2n+1} \to C(\mathbb{T}) \subset \mathrm{BV}(\mathbb{T})^*$. Therefore, although $\mathrm{BV}(\mathbb{T})$ is not reflexive, the adjoint of the adjoint operator can be identified with the operator itself, i.e., $\mathcal{A}^{**} = \mathcal{A}$.

The second constraint posed in (RCEP) can be seen as a constraint on the type of measure sought. In particular, the inverse problem is not to find any measure that matches the given data, but to parametrize all nonnegative measures with only absolutely continuous part $\Phi \in L_1(\mathbb{T})$ (that matches the data), and where the absolutely continuous part can be written as $\Phi = P/Q$ a.e. for P and $Q \in \bar{\mathfrak{P}}_+$. In fact, the results in Theorem 2.2.1 can now be interpreted from a perspective of variational regularization: The primal problem (2.2.3) is a variational regularization which finds the solution $d\mu = \Phi d\theta$ that matches the convariances and which has minimal distance to P in the *Kullback-Leibler* sense [47]. In this context, the result of the theorem states that this regularizing function "promotes rational solutions". Moreover, viewed from this perspective the formulation can be extended to other compactly supported moment problems, which has been done in [25, 26, 46]. This is also related to work on maximum entropy solutions to moment problems, see, e.g., [64, 65, 79, 12, 13, 14, 15].

Remark 2.4.1. As a final remark here, we note that this last example illustrates that the distinction between system identification and inverse problems is not as clear as it was indicated in Chapter 1. In Section 2.2, the rational covariance extension problem was derived from a system identification perspective, however in this preceding paragraph it was seen as an inverse problem. Similarly, estimating coefficients in differential equations can be called both a system identification problem, cf. [104, Chp. 6], and an inverse problem [37, Sec. 1.6] [61, Ex. 1.10], depending on the community to which one belongs.

2.5 A note on machine learning and neural networks

Although the current boom in machine learning research started relatively recently, machine learning has has been around for quite a while. Moreover, machine learning is a much larger area than just *neural networks*, cf. [10]. However, to limit the scope, this section will only present material on the latter since this is what is most relevant for the appended papers.

Machine learning with neural networks has been an active area of research since at least the 1940s [48, Sec. 1.2.1] [102, Sec. 5.1]. Moreover, it has several connections to other topics treated in this thesis: It has been linked to system identification at least since the early 1990s [74], and some of the methods used for training are closely related to calculus of variations and dynamical programming, see, e.g., [48, Sec. 6.6], [102, Sec. 5.5] and references therein. Some claim that the "new deep learning era" started with AlexNet [62], which is a deep neural network for image classification. Irrespectivly if this is completely true or not[25] it

[25] There were also other achievements around the same time, many of which were also in imaging

serve as a good example of an application where a machine learning (*data-driven*) method successfully outperformed "ordinary" (*model-driven*) methods. With [62], the authors competed in "ImageNet Large Scale Visual Recognition Challenge" in 2012 and won [100], reducing the classification errors in the image test set with around ten percentage point compared to the runner-up. This trend has continued, cf. [48, Fig. 1.12], and the machine learning methods reached "super-human" levels for image classification a few years later [53, 56]. This type of success stories has undoubtedly contributed to the surge in machine learning and deep learning research in the last couple of years. The following section is intended as a brief introduction to the area for a mathematically inclined audience.

In the following subsection we will first introduce *supervised* and *unsupervised* machine learning: What are the mathematical problems one tries to solve and how are they solved conceptually? This can be treated without explicitly introducing the concept of a *neural network*. In the subsequent subsection, we will introduce the concept of a neural network, and also shortly describe how the "learning" is normally done in practice, i.e., how the corresponding optimization problems are actually solved.

Supervised and unsupervised machine learning

In this section we will try to clarify the difference between *supervised* and *unsupervised* machine learning. These concepts are not always strictly defined in the literature, cf. [48, Sec. 5.1.3], why such an attempt will most likely fall short from some aspects. Nevertheless, in many cases machine learning using neural networks can be seen as "automatic parameter tuning" of certain parametrized operators. The automatic tuning is normally done by minimizing an appropriately chosen loss function, most of the time involving an expectation. To be more precise, let $\mathcal{B}_\gamma : A \to B$ be a parametrized operator with parameters $\gamma \in \Gamma \subset \mathbb{R}^m$. Today, the number of parameters m can easily be in the order of millions or more, cf. [48, Fig. 1.11]. Moreover, the sets of inputs A and outputs B of the operator can be very different for different problems: Sometimes they are finite sets, sometimes Hilbert or Banach spaces, and sometimes manifolds.

In *supervised learning* one considers $(A \times B)$-valued random elements $(\mathtt{a}, \mathtt{b})$ that follows some distribution $P_{(\mathtt{a},\mathtt{b})}$. The idea is to design the operator $\mathcal{B}_\gamma$ so that it can predict the outcome of $\mathtt{b}$ by only observing the outcome of $\mathtt{a}$. For a given parametrization of $\mathcal{B}_\gamma$ the goal is thus to find "optimal" parameters $\hat{\gamma}$ for the operator, where "optimal" is normally defined to be

$$\hat{\gamma} = \operatorname*{arg\,min}_{\gamma \in \Gamma} \ \mathbb{E}_{(\mathtt{a},\mathtt{b})}\big[\mathcal{D}(\mathcal{B}_\gamma(\mathtt{a}), \mathtt{b})\big].$$

Here, $\mathcal{D} : B \times B \to \mathbb{R}$ is a suitable distance measure (cf. the data discrepancy functional in variational regularization, Section 2.4), and as indicated the expectation

applications. One example is an architecture for unsupervised learning for feature detection, e.g., face detection in images [67]. These results also got public outreach in ordinary press [78].

is taken over the joint probability distribution $P_{(\mathtt{a},\mathtt{b})}$. However, this problem can normally not be solved directly. Instead, this optimization problem is approximated by considering a finite number of pairs (a_i, b_i) that are assumed to be independent and identically distributed (i.i.d.) realizations of $(\mathtt{a}, \mathtt{b})$. The expectation is then approximated with a finite average over these pairs, which gives the problem

$$\hat{\gamma} = \arg\min_{\gamma\in\Gamma} \ \frac{1}{N}\sum_{i=1}^{N} \mathcal{D}(\mathcal{B}_\gamma(a_i), b_i). \tag{2.5.1}$$

For suitable choices of $\mathcal{B}_\gamma$ and $\mathcal{D}$ this is a smooth optimization problem in γ, and in this case a stationary point to this cost function could be found by using gradient descent. However, for large N computing the gradient of the above expression is too time consuming. Instead, one typically uses a stochastic-type optimization algorithm, like stochastic gradient decent [48, Sec. 5.9], in which a small number $n \ll N$ of the samples $\{(a_i, b_i)\}_{i=1}^N$ are randomly selected (called a *minibatch*) and a gradient with respect to γ is computed based on these samples. How the gradient computation is done will be explained in the next subsection.

In contrast, *unsupervised* machine learning considers A-valued random elements that follows some distribution $P_{\mathtt{a}}$. Here the goal is to find the optimal parameters

$$\hat{\gamma} = \arg\min_{\gamma\in\Gamma} \ \mathbb{E}_{\mathtt{a}}\big[\mathcal{S}(\mathcal{B}_\gamma(\mathtt{a}))\big], \tag{2.5.2}$$

where $\mathcal{S} : B \to \mathbb{R}$ is a suitable loss function, cf. [48, Eq. (5.102)]. For an example of such a problem, see Remark 2.5.1 at the end of the subsection. Again, since this can normally not be solved directly[26] we consider a finite number of N i.i.d. samples a_i of $\mathtt{a}$ and approximate the optimization problem with

$$\arg\min_{\gamma\in\Gamma} \ \frac{1}{N}\sum_{i=1}^{N} \mathcal{S}(\mathcal{B}_\gamma(a_i)).$$

Similarly to the supervised case, a stationary point to this cost function is normally obtained using some type of stochastic gradient descent.

One risk with a cost function of the form (2.5.1) or (2.5.2) is that of overfitting the given data, especially when the parametrized operators have millions of parameters. Therefor it is common to use some type of regularization for the parameters γ in the training, see, e.g., [48, Chp. 7], [74] or [102, Sec. 4.4 and 5.6.3] and references therein. For the problem (2.5.1), and similarly for (2.5.2), this means that one instead considers

$$\arg\min_{\gamma\in\Gamma} \ \mathbb{E}_{(\mathtt{a},\mathtt{b})}\big[\mathcal{D}(\mathcal{B}_\gamma(\mathtt{a}), \mathtt{b})\big] + \theta\mathcal{G}(\gamma),$$

where $\mathcal{G} : \Gamma \to \mathbb{R}$ is a regularization functional that promotes certain structures in the set of parameters, and θ is the regularization parameter.

[26] For an illustrative toy-example that does have a closed-form analytic solution, see [5, Ex. 4.1] which is Example F.4.1 in this thesis.

Remark 2.5.1. While unsupervised machine learning is sometimes used simply because data pairs (a_i, b_i) are to expensive to generate, e.g., by manual labeling of images, there are problems that are "intrinsically unsupervised". One example is the application in [5] (Paper F in this thesis). Another example is so-called *autoencoders* [48, Chp. 14]. This is a neural network that can be decomposed in two parts $\mathcal{B}_\gamma = \mathcal{B}^2_{\gamma_2} \circ \mathcal{B}^1_{\gamma_1}$, and that tries to learn the identity mapping. Normally this is done by minimizing a functional of the type $\mathcal{S}(\mathcal{B}_\gamma(\mathtt{a})) := \|\mathtt{a} - (\mathcal{B}^2_{\gamma_2} \circ \mathcal{B}^1_{\gamma_1})(\mathtt{a})\|$. While this sounds trivial, note that it is not by any means sure that the parametrization of $\mathcal{B}_\gamma$ is such that it can "easily" represent the identity operator. In fact, if the output of $\mathcal{B}^1_{\gamma_1}$ has a dimension which is much smaller than the input, while the output of $\mathcal{B}^2_{\gamma_2}$ has the same dimension as the input, then the optimization can be interpreted as learning a sparse representation of the possible outcomes under the distribution $P_{\mathtt{a}}$. This can be used, e.g., to learn a pair of compression-decompression algorithms for certain types of data, where $\mathcal{B}^1_{\gamma_1}$ would work as the compression algorithm while $\mathcal{B}^2_{\gamma_2}$ would work as the decompression algorithm. In many applications it is also assumed that the data distribution $P_{\mathtt{a}}$ has support on a lower-dimensional manifold. This method could then be interpreted as learning coordinate representations of this manifold: After training the low-dimensional input to $\mathcal{B}^2_{\gamma_2}$ are the local coordinates, and the network returns an appropriate point on the manifold in a higher-dimensional space, cf. [48, Chp. 14.6].

Neural networks and backpropagation

The specific paramtrization of the operator $\mathcal{B}_\gamma$ is normally called the *network architecture*. A commonly used family of architectures is so-called (*feed-forward*) *neural networks*, cf. [48, Sec. 6.0 and 6.4]. As indicated by the name, these were in the beginning intended as models of how the human brain might work [48, p. 13]. Specifically, in a neural network architecture the operator takes the form

$$\mathcal{B}_\gamma = f^n \circ \mathcal{A}^n_{\gamma_n} \circ \dots f^1 \circ \mathcal{A}^1_{\gamma_1}. \tag{2.5.3}$$

Here, $\mathcal{A}^i_{\gamma_i} : \mathbb{R}^{\ell_i} \to \mathbb{R}^{k_i}$ are affine operators, and $f^i : \mathbb{R}^{k_i} \to \mathbb{R}^{\ell_{i+1}}$ are nonlinear functions. Each component $f^i \circ \mathcal{A}^i_{\gamma_i}$ is called a *layer*. As indicated by the notation, $\gamma = [\gamma_1, \dots, \gamma_n]$ and we normally only optimize over the parameters in the affine operators. In some cases $\mathcal{A}^i_{\gamma_i}$ is allowed to be any affine operator, i.e., it is represented by a dense matrix and a vector, in which case the layer is called fully-connected. In other cases, certain structures are imposed, and, e.g., in imaging applications it has been seen that it is often useful to restrict the linear part of the operators $\mathcal{A}^i_{\gamma_i}$ to be convolutions [48, Chp. 9].[27] The motivation for this is to get networks that are translation invariant [48, p. 254] (cf. time-invariant systems in Section 2.1).

The functions f^i are normally so-called "pointwise" nonlinearities, or they have an action which is "local". To understand the terminology of "pointwise"

[27] To be more precise, the discrete operators are normally cross-correlation-type operators, since the kernel is not transposed before applying it, cf. [48, pp. 332-333].

and "local" action, note that the output of each operator $\mathcal{A}^i_{\gamma_i}$ is a vector $\mathbb{R}^{k_i}$. A "pointwise" nonlinearity simply means that f^i acts on each component of this vector independently, i.e., with a slight abuse of notation, that for $x \in \mathbb{R}^{k_i}$ the function is given by $f^i(x) = [f^i(x_1), \dots, f^i(x_{k_i})]^T$.[28] Commonly used examples of pointwise nonlinearities are the sigmoid function $f(x) = 1/(1+e^{-x})$ [48, Sec. 6.3.2], and the so-called rectified linear unite (ReLu) which is given by $f(x) = \max(0, x)$ [48, Sec. 6.3.1]. If f^i has a "local" action it means that it only acts on a relatively small number of samples in the vector, e.g., $f(x) = [f(x_1, x_2, x_3),\ f(x_2, x_3, x_4), \dots,\ f(x_{k_i}, x_1, x_2)]^T$. These are often called pooling layers and a common type in imaging applications is max pooling [48, Sec. 9.3], which is given by $f(x) = [\max_{i \in \mathcal{C}_j} x_i]_{j=1,\dots,\ell_{i+1}}^T$ where $\mathcal{C}_j$ are appropriate subsets of $\mathbb{Z}^{k_i}$ typically corresponding to a set of pixels that are neighboring to each other.

Having formally introduce the concept of a neural network, we will now briefly touch upon how machine learning problems of the form (2.5.1) and (2.5.2) are solved in practice. As briefly mentioned in the above subsection, this type of problems are normally solved using (stochastic) gradient-type methods. However, these gradients are normally computed using *automatic differentiation* [48, Sec. 6.5.9], cf. [50]. Due to the structure of (2.5.3) this can also be done in a computationally efficient way that also saves memory, which is called *backpropagation* [99] [48, Sec. 6.5]. To explain the idea, consider an operator $\mathcal{B}_\gamma$ that has the structure (2.5.3), and let $\mathcal{B}^j_\gamma$ be the first j layers, i.e., $\mathcal{B}^j_\gamma = f^j \circ \mathcal{A}^j_{\gamma_j} \circ \dots f^1 \circ \mathcal{A}^1_{\gamma_1}$ and $\mathcal{B}^n_\gamma = \mathcal{B}_\gamma$. For one of the elements in the sum (2.5.1), let us take the partial derivative with respect to γ_n and apply the chain rule (see, e.g., [97, Thm. 5.5]). This gives the expression

$$\frac{\partial}{\partial \gamma_n} \mathcal{D}(\mathcal{B}_\gamma(a_i), b_i) = \\ \left.\frac{\partial \mathcal{D}(w_{n+1}, b_i)}{\partial w_{n+1}}\right|_{w_{n+1}=\mathcal{B}_\gamma(a_i)} \left.\frac{\partial f^n(v_n)}{\partial v_n}\right|_{v_n=\mathcal{A}^n_{\gamma_n}(\mathcal{B}^{n-1}_\gamma(a_i))} \left.\frac{\partial \mathcal{A}^n_{\gamma_n}(w_n)}{\partial \gamma_n}\right|_{w_n=(\mathcal{B}^{n-1}_\gamma(a_i))}.$$

Here, the first term belongs to $\mathbb{R}^{1\times \ell_{n+1}}$, the second to $\mathbb{R}^{\ell_{n+1}\times k_n}$, and the third to $\mathbb{R}^{k_n \times \dim(\gamma_n)}$. Now, if we instead take the partial derivative with respect to γ_{n-1}, the first two terms will be identical. This means that these terms only need to be computed ones, while they can be used many times. Similarly, any partial derivative of γ_j, for $j \leq n-1$, can be written as

$$\frac{\partial}{\partial \gamma_j} \mathcal{D}(\mathcal{B}_\gamma(a_i), b_i) = \left.\frac{\partial \mathcal{D}(w_{n+1}, b_i)}{\partial w_{n+1}}\right|_{w_{n+1}=\mathcal{B}_\gamma(a_i)} \\ \prod_{k=j+1}^{n} \left(\left.\frac{\partial f^k(v_k)}{\partial v_k}\right|_{v_k=\mathcal{A}^k_{\gamma_k}(\mathcal{B}^{k-1}_\gamma(a_i))} \left.\frac{\partial \mathcal{A}^k_{\gamma_k}(w_k)}{\partial w_k}\right|_{w_k=(\mathcal{B}^{k-1}_\gamma(a_i))} \right)$$

[28] Note that in such a case we have the $\ell_{i+1} = k_i$.

$$\left.\frac{\partial f^j(v_j)}{\partial v_j}\right|_{v_j=\mathcal{A}^j_{\gamma_j}(\mathcal{B}^{j-1}_\gamma(a_i))} \left.\frac{\partial \mathcal{A}^j_{\gamma_j}(w_j)}{\partial \gamma_j}\right|_{w_j=(\mathcal{B}^{j-1}_\gamma(a_i))}.$$

By taking the derivatives in the order ∂_{γ_i} for $i = n, n-1, \ldots, 1$, and only storing the product of the terms that reappear, a lot of computational time and memory storage space is saved.

Remark 2.5.2. As a final note, observe that the theoretical understanding of neural networks is, to the best of my knowledge, still limited. It is therefore difficult to explain the successful application in certain areas, and also to predict to which extent they can be expected to generalize to other areas. However, recently there has been a lot of work on expanding the theoretically understanding, especially of (feed-forward) neural networks. Example of such works are [11] where they characterize the trade-off between complexity and approximation properties of deep neural networks, and [52, 101] where they interpreting them as discretizations of ODEs and PDEs, respectively. The learning problem itself has also been recasted as an optimal control problem, leading to new types of training algorithms based on the Pontryagin maximum principle [70]. A final example is [87, 88], where convolutional neural networks are investigated from the perspective of so-called convolutional sparse coding, which is related to the literature on sparse solutions to linear equations, cf. [17]. However, although the literature around this is growing very rapidly, I find that it is still safe to say that much remains to do when it comes to creating a fundamental theoretical understand of the area.

3. Summary of papers

The main part of this thesis is the six appended papers in Part II. Here follows a short summary of each paper, also clarifying the authors contributions to each of them.

Paper A: Multidimensional rational covariance extension

Paper A of this thesis contains material from the publications

- A. Ringh, J. Karlsson, and A. Lindquist. Multidimensional rational covariance extension with applications to spectral estimation and image compression. *SIAM Journal on Control and Optimization*, 54(4):1950–1982, 2016.
- A. Ringh, J. Karlsson, and A. Lindquist. Further results on multidimensional rational covariance extension with application to texture generation. In *IEEE Annual Conference on Decision and Control (CDC)*, pages 4038–4045. IEEE, 2017.

In particular, the main body of the paper is an edited version of the first paper. The paper in this thesis also contains an example in texture generation and Wiener system identification from the second paper.

Summary The paper investigates generalizations of the rational covariance extension problem [58, 44, 45, 27, 23] to higher dimensions. The approach taken in the paper is to extend the convex optimization problem proposed in [23], similar to what has already been done in [46]. We derive the form of the optimal solution, prove existence and uniqueness of it, and derive the dual problem. The papers also extend part of the work in [20, 21, 39], whereby both the denominator and the numerator polynomial are estimated by also using cepstral coefficients. Moreover, in the spirit of [71] the paper also shows that a discretized version of the problem can be used to find an approximate solution, cf. [94].

Contribution The main ideas of this paper emerged from discussions between all three authors. The author of this thesis has then been a main part in the

theoretical development around these ideas, has done the numerical implementations and simulations, and has been an active part in writing the manuscript.

Paper B: Multidimensional rational covariance extension with approximate covariance matching

Paper B of this thesis is an edited version of the paper

- A. Ringh, J. Karlsson, and A. Lindquist. Multidimensional rational covariance extension with approximate covariance matching. *SIAM Journal on Control and Optimization*, 56(2):913–944, 2018.

Summary This papers continues on the work in [95] (Paper A in this thesis), and considers approximate covariance matching formulations in the rational covariance extension framework. This is of interest since in applications the covaraince used will be estimated from a finite amount of data and thus contain errors, and since the condition that guarantees the existence of a solution is nontrivial to test. Such ideas have been considered previously in [103, 40] [4, Chp. B], and this work expand upon these. In fact, two different formulations for approximate matching are investigated, each one containing a tuning (regularization) parameter that indicates to which degree we want to enforce matching of the covariances. For both formulations we derive the form of the solution, and show existence and uniqueness. Moreover, we show that the two formulations are in fact equivalent, in the sense that there is an homotopy between the two sets of solutions obtained when varying the tuning parameters.

Contribution This paper is a continuation of the work in Paper A, and similarly it is also the result of a cooperation between all three authors. The author of this thesis has thus been an active part in formulating the research questions, developing the theoretical results, performing the numerical experiments, and writing of the manuscript.

Paper C: Lower bounds on the maximum delay margin by analytic interpolation

Paper C of this thesis is an edited version of

- A. Ringh, J. Karlsson, and A. Lindquist. Lower bounds on the maximum delay margin by analytic interpolation. Accepted to *IEEE Annual Conference on Decision and Control (CDC)*. IEEE, 2018.

A preprint of the above paper is available as arXiv preprint arXiv:1803.09487.

Summary The maximum delay margin of a plant is a robustness measure on how sensitive the plant is to delay in the feedback loop. However, how to compute the maximum delay margin of a general plant is still an unsolved problem. This work considers finding lower bounds on the maximum delay margin, and builds on [90, 91] where lower bounds are derived using a sufficient condition of "small gain"-type. This gives rise to an analytic interpolation problem, which in [90, 91] is simplified by introducing a rational approximation of the irrational transfer function coming from the delay-term. Instead, we omit this approximation and tackle the interpolation problem directly by using analytic function theory. This only leads to a marginal improvement of the lower bound, but the direct approach also gives an increased understanding of the sufficient condition. This in turn enables us to introduce of a tuning parameter that can be used to obtain better lower bounds. The improvement is finally demonstrated in numerical examples.

Contribution The idea to solve the problem directly using analytic interpolation theory is due to J. Karlsson after a discussion with J. Chen, one of the authors of [90, 91]. The rest of the work is, just as the previous two papers, the result of a joint effort by all three authors.

Paper D: Generalized Sinkhorn iterations for regularizing inverse problems using optimal mass transport

Paper D of this thesis is an edited version of the paper

- J. Karlsson, and A. Ringh. Generalized Sinkhorn iterations for regularizing inverse problems using optimal mass transport. *SIAM Journal on Imaging Sciences*, 10(4):1935–1962, 2017.

Summary The optimal mass transport problem is a geometric framework for how to transport masses in an optimal way [107]. Although it can be formulated as a linear programming problem, when the two marginals have large dimensions, the size of this linear program becomes prohibitively large. In particular, this is the case when using it to compute the distance between two images. A recent development to address this builds on using an entropic barrier term and solving the resulting optimization problem using so-called Sinkhorn iterations [29]. In this work we show how these results can be used and extended in order to solve other optimization problems involving an optimal transport term. In particular, we derive iterations similar to the Sinkhorn iterations for computing the proximal operator of the optimal transport distance. Moreover, in many cases of interest the matrix that defines the transportation cost gets a Toeplitz-block-Toeplitz structure. We utilize this to speed up the computations and reduce the memory requirements of the algorithm by doing matrix-vector multiplications using the fast Fourier Transform. Finally, this opens up for using it in variational regularization in inverse problems by using variable splitting techniques, and this is demonstrate by an example in CT.

Contribution The research was initiated by J. Karlsson, and the idea of using optimal mass transport as a regularizer in inverse problem is due to him. The idea of using variable splitting is due to myself. I have also done most of the coding and numerical examples. All parts of the paper is the result of a close collaboration between the authors.

Paper E: Learning to solve inverse problems using Wasserstein loss

Paper E of this thesis is an updated and edited version of the paper

- J. Adler, A. Ringh, O. Öktem, and J. Karlsson. Learning to solve inverse problems using Wasserstein loss. arXiv preprint arXiv:1710.10898, 2017.

The results of the paper were presented at the workshop Optimal Transport & Machine Learning at the conference Advances in Neural Information Processing Systems (NIPS) in 2017.

Summary In supervised machine learning, pairs (f_i, g_i) of ground-truths f_i and corresponding input g_i are used to optimize ("learn") the parameters in a parametrized operator ("neural network"). This paper investigates what happens if these pairs (f_i, g_i) are corrupted by noise. In particular, it considers the case of using machine learning for solving ill-posed inverse problems in imaging. In this case the input g_i is measurement data corresponding to f_i. However, in this work noise is introduced by letting the data g_i be generated from a "geometrically distorted" version of f_i. This will of course affect the quality of the learned reconstruction operator, but the degree of the degradation will depend on how the training is done, i.e., which cost function that is used in the optimization. We derive theoretical results that indicates that training with standard mean squared error loss could give a reconstruction operator which severely degrade the quality of the reconstructions, while training with optimal transport loss could give a reconstruction operator that better compensate for these distortions. We also perform a numerical experiment in CT by training a neural network on this kind of distorted data, and the results of this experiment are in line with the theoretical predictions.

Contribution The idea of using optimal mass transport as loss function in training emerged in discussions between all four authors. The implementation is based on the code associated with [1] and [59] (Paper D in this thesis), and has been done by J. Adler. The author of this thesis has contributed to the theoretical results in the paper, and also to the writing of the paper.

Paper F: Data-driven nonsmooth optimization

Paper F of this thesis is an edited version of the paper

- S. Banert, A. Ringh, J. Adler, J. Karlsson, and O. Öktem. Data-driven nonsmooth optimization. arXiv preprint arXiv:1808.00946, 2018.

The paper has been submitted for publication, and the results have been presented at the SIAM Conference on Imaging Science in 2018.

Summary The paper considers the use of machine learning to "learn" an optimization solver. This has been considered before, notably in the "LISTA paper" [49], but also recently in [69, 3]. The idea is that the objective function evaluated in the output of the neural network is a natural loss function for training, which means that training can be done in an unsupervised fashion. However, what differentiate this work from previous work is the parametrization of the neural network and the consequences of such a parametrization. Here, the key idea is to first specify a class of optimization algorithms using a generic iterative scheme involving only linear operations and applications of proximal operators. The architecture is inspired from unrolling iterative schemes for solving optimization problems, and truncating them after a finite number of iterations. This makes the network suitable for solving large-scale optimization problems with a possibly nonsmooth objective function, and also leads to provable convergence for some of the trained networks. To demonstrate the possibilities of the approach, we consider examples arising in tomographic reconstruction and image deconvolution, and in particular a family of total variation regularization problems.

Contribution Most authors have contributed to most parts of the work. However, a rough outline is as follows: The idea to use unsupervised learning for solving inverse problems via variational regularization is due to J. Adler and O. Öktem. The idea to consider the schemes in (F.3.1) and (F.4.3) is due to by myself and J. Adler. The fixed-point analysis in Section F.3 was done by myself and J. Karlsson, while the proofs of Theorems F.3.1 and F.3.10 are due to S. Banert. I have implemented the code for the numerical experiments, with help from J. Adler and the code corresponding to [1]. Most of the writing has been done by myself together with S. Banert.

Copyright notice

As indicated above, some of the material in Part II of this thesis has been published elsewhere. The following is a full disclosure of the copyright holders of the corresponding material.

- Paper A:
  - Sections A.1 through A.7, as well as sections A.9 and A.10: © 2016 Society for Industrial and Applied Mathematics.
  - Section A.8: © 2018 IEEE.

  This copyright includes all figures referenced in the corresponding sections. In particular, this means that IEEE holds the copyright for figures A.6, A.7, A.8, and A.9, while Society for Industrial and Applied Mathematics holds the copyright for the remaining figures in this paper.
- Paper B: © 2018 Society for Industrial and Applied Mathematics.
- Paper C: © 2018 IEEE.
- Paper D: © 2017 Society for Industrial and Applied Mathematics.

In accordance with guidelines from the IEEE, the following statement also needs to be included in this copyright notice.

> In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of KTH Royal Institute of Technology's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to `http://www.ieee.org/publications_standards/publications/rights/rights_link.html` to learn how to obtain a License from RightsLink. If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

References

[1] J. Adler and O. Öktem. Learned primal-dual reconstruction. *IEEE Transactions on medical imaging*, 37(6):1322–1332, 2018.

[2] A.N. Amini, E.S. Ebbini, and T.T. Georgiou. Noninvasive estimation of tissue temperature via high-resolution spectral analysis techniques. *IEEE Transactions on Biomedical Engineering*, 52(2):221–228, 2005.

[3] M. Andrychowicz, M. Denil, S. Gomez, M.W. Hoffman, D. Pfau, T. Schaul, and N. de Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3981–3989, 2016.

[4] E. Avventi. *Spectral Moment Problems : Generalizations, Implementation and Tuning*. PhD thesis, 2011. Optimization and Systems Theory, Department of Mathematics, KTH Royal Institue of Technology.

[5] S. Banert, A. Ringh, J. Adler, J. Karlsson, and O. Öktem. Data-driven nonsmooth optimization. *arXiv preprint arXiv:1808.00946*, 2018.

[6] L. Baratchart, S. Chevillard, and J. Leblond. Silent and equivalent magnetic distributions on thin plates. *HAL-Inria preprint hal-01286117v2*, 2016. Accepted for publication in Theta Series in Advanced Mathematics.

[7] L. Baratchart, D.P. Hardin, E.A. Lima, E.B. Saff, and B.P. Weiss. Characterizing kernels of operators related to thin-plate magnetizations via generalizations of Hodge decompositions. *Inverse Problems*, 29(1):1–29, 2013.

[8] H.H. Bauschke and P.L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, New York, NY, 2011.

[9] M. Bertero, H. Lantéri, and L. Zanni. Iterative image reconstruction: A point of view. *Mathematical Methods in Biomedical Imaging and Intensity-Modulated Radiation Therapy (IMRT)*, 7:37–63, 2008.

[10] C. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, New York, NY, 2006.

[11] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen. Optimal approximation with sparsely connected deep neural networks. *arXiv preprint arXiv:1705.01714*, 2017.

[12] J.M. Borwein and A.S. Lewis. Convergence of best entropy estimates. *SIAM Journal on Optimization*, 1(2):191–205, 1991.

[13] J.M. Borwein and A.S. Lewis. Duality relationships for entropy-like minimization problems. *SIAM Journal on Control and Optimization*, 29(2):325–338, 1991.

[14] J.M. Borwein and A.S. Lewis. On the convergence of moment problems. *Transactions of the American Mathematical Society*, 325(1):249–271, 1991.

[15] J.M. Borwein and A.S. Lewis. Partially-finite programming in L_1 and the existence of maximum entropy estimates. *SIAM Journal on Optimization*, 3(2):248–267, 1993.

[16] R.W. Brown, Y.-C. N. Cheng, E.M. Haacke, M.R. Thompson, and R. Venkatesan. *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. John Wiley & Sons, New York, NY, 2014.

[17] A.M. Bruckstein, D.L. Donoho, and M Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM review*, 51(1):34–81, 2009.

[18] J.P. Burg. Maximum entropy spectral analysis. In *Proceedings of the 37th Meeting Society of Exploration Geophysicists*, 1967.

[19] J.P. Burg. *Maximum Entropy Spectral Analysis*. PhD thesis, 1975. Department of Geophysics, Stanford University.

[20] C.I. Byrnes, P. Enqvist, and A. Lindquist. Cepstral coefficients, covariance lags, and pole-zero models for finite data strings. *IEEE Transactions on Signal Processing*, 49(4):677–693, 2001.

[21] C.I. Byrnes, P. Enqvist, and A. Lindquist. Identifiability and well-posedness of shaping-filter parameterizations: A global analysis approach. *SIAM Journal on Control and Optimization*, 41(1):23–59, 2002.

[22] C.I. Byrnes, T.T. Georgiou, and A. Lindquist. A generalized entropy criterion for Nevanlinna-Pick interpolation with degree constraint. *IEEE Transactions on Automatic Control*, 46(6):822–839, 2001.

[23] C.I. Byrnes, S.V. Gusev, and A. Lindquist. A convex optimization approach to the rational covariance extension problem. *SIAM Journal on Control and Optimization*, 37(1):211–229, 1998.

[24] C.I. Byrnes, S.V. Gusev, and A. Lindquist. From finite covariance windows to modeling filters: A convex optimization approach. *SIAM Review*, 43(4):645–675, 2001.

[25] C.I. Byrnes and A. Lindquist. A convex optimization approach to generalized moment problems. In K. Hashimoto, Y. Oishi, and Y. Yamamoto, editors, *Control and Modeling of Complex Systems*, Trends in Mathematics, pages 3–21. Birkhäuser, Boston, 2003.

[26] C.I. Byrnes and A. Lindquist. The generalized moment problem with complexity constraint. *Integral Equations and Operator Theory*, 56(2):163–180, 2006.

[27] C.I. Byrnes, A. Lindquist, S.V. Gusev, and A.S. Matveev. A complete parameterization of all positive rational extensions of a covariance sequence. *IEEE Transactions on Automatic Control*, 40(11):1841–1857, 1995.

[28] C.-T. Chen. *Linear system theory and design*. Oxford University Press, New York, NY, 3rd edition, 1999.

[29] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2292–2300, 2013.

[30] Ph. Delsarte, Y. Genin, Y. Kamp, and P. Van Dooren. Speech modelling and the trigonometric moment problem. *Philips Journal of Research*, 37:277–92, 1982.

[31] G. Doetsch. *Introduction to the Theory and Application of the Laplace Transformation*. Springer, Berlin Heidelberg, 1974.

[32] J.C. Doyle, B.A. Francis, and A.R. Tannenbaum. *Feedback control theory*. Dover, New York, NY, 2009. Unabridged republication of original published by Macmillan Publishing, 1992.

[33] B. Dumitrescu. *Positive Trigonometric Polynomials and Signal Processing Applications*. Springer, Dordrecht, 2007.

[34] M. Dür. Copositive programming - a survey. In M. Diehl, F. Glineur, E. Jarlebring, and W. Michiels, editors, *Recent advances in optimization and its applications in engineering*, pages 3–20. Springer, 2010.

[35] P.L. Duren. *Theory of H_p spaces*. Academic press, New York, NY, 1970.

[36] H. Dym and H.P. McKean. *Fourier Series and Integrals*. Academic press, San Diego, CA, 1972.

[37] H.W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*. Kluwer Academic Publisher, 2000.

[38] P. Enqvist. *Spectral Estimation by Geometric, Topological and Optimization Methods*. PhD thesis, 2001. Optimization and Systems Theory, Department of Mathematics, KTH Royal Institue of Technology.

[39] P. Enqvist. A convex optimization approach to ARMA(n,m) model design from covariance and cepstral data. *SIAM Journal on Control and Optimization*, 43(3):1011–1036, 2004.

[40] P. Enqvist and E. Avventi. Approximative covariance interpolation with a quadratic penalty. In *IEEE Annual Conference on Decision and Control (CDC)*, pages 4275–4280. IEEE, 2007.

[41] C. Foias, H. Özbay, and A.R. Tannenbaum. *Robust control of infinite dimensional systems*. Springer, Berlin Heidelberg, 1996.

[42] A. Friedman. *Foundations of Modern Analysis*. Dover, New York, NY, 1982. Unabridged republication of original published by Holt, Rinehart and Winston, 1970.

[43] J. Garnett. *Bounded analytic functions*. Springer, New York, NY, revised 1st edition, 2007.

[44] T.T. Georgiou. *Partial Realization of Covariance Sequences*. PhD thesis, 1983. Center for Mathematical Systems Theory, Univeristy of Florida.

[45] T.T. Georgiou. Realization of power spectra from partial covariance sequences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(4):438–449, 1987.

[46] T.T. Georgiou. Relative entropy and the multivariable multidimensional moment problem. *IEEE Transactions on Information Theory*, 52(3):1052–1066, 2006.

[47] T.T. Georgiou and A. Lindquist. Kullback-Leibler approximation of spectral density functions. *IEEE Transactions on Information Theory*, 49(11):2910–2917, 2003.

[48] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, Cambridge, MA, 2016. Available at `http://www.deeplearningbook.org`.

[49] K. Gregor and Y. LeCun. Learning fast approximations of sparse coding. In *International Conference on Machine Learning (ICML)*, pages 399–406, 2010.

[50] A. Griewank and A. Walther. *Evaluating derivatives: Principles and techniques of algorithmic differentiation*. SIAM, Philadelphia, PA, 2nd edition, 2008.

[51] D.J. Griffiths. *Introduction to Electrodynamics*. Prentice-Hall, Upper Saddle River, NJ, 1999.

[52] E. Haber and L. Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34(1):1–22, 2017.

[53] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *IEEE International Conference on Computer Vision (ICCP)*, pages 1026–1034, 2015.

[54] J.W. Helton, J.A. Ball, C.R. Johnson, and J.N. Palmer. *Operator theory, analytic functions, matrices, and electrical engineering*. American Mathematical Society, Providence, RI, 1987.

[55] K. Hoffman. *Banach Spaces of Analytic Functions*. Dover, New York, 2007. Unabridged republication of the Dover reprint from 1988 of the original published by Prentice-Hall, 1962.

[56] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pages 448–456, 2015.

[57] O. Kallenberg. *Foundations of modern probability*. Springer, New York, NY, 1997.

[58] R.E. Kalman. Realization of covariance sequences. In *Toeplitz memorial conference*, 1981. Tel Aviv, Israel.

[59] J. Karlsson and A. Ringh. Generalized Sinkhorn iterations for regularizing inverse problems using optimal mass transport. *SIAM Journal on Imaging Sciences*, 10(4):1935–1962, 2017.

[60] H.K. Khalil. *Nonlinear systems*. Prentice Hall, Upper Saddle River, NJ, 3 edition, 2000.

[61] A. Kirsch. *An introduction to the mathematical theory of inverse problems*. Springer, New York, NY, 2nd edition, 2011.

[62] A. Krizhevsky, I. Sutskever, and G.E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

[63] Y.A. Kuznetsov. *Elements of applied bifurcation theory*. Springer, New York, NY, 3rd edition, 2013.

[64] S.W. Lang and J.H. McClellan. Multidimensional MEM spectral estimation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 30(6):880–887, 1982.

[65] S.W. Lang and J.H. McClellan. Spectral estimation for sensor arrays. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 31(2):349–358, 1983.

[66] J.B. Lasserre. *Moments, positive polynomials and their applications*. Imperial College Press, London, 2009.

[67] Q.V. Le, M.' A. Ranzato, R. Monga, M. Devin, K. Chen, G.S. Corrado, D. Jeff, and A.Y. Ng. Building high-level features using large scale unsupervised learning. In *International Conference on Machine Learning (ICML)*, pages 81–88, 2012.

[68] J.M. Lee. *Introduction to Smooth Manifolds*. Springer, New York, NY, 2nd edition, 2013.

[69] K. Li and J. Malik. Learning to optimize. In *International Conference on Learning Representations (ICLR)*, 2017.

[70] Q. Li, L. Chen, C. Tai, and E. Weinan. Maximum principle based algorithms for deep learning. *The Journal of Machine Learning Research*, 18(1):5998–6026, 2017.

[71] A. Lindquist and G. Picci. The circulant rational covariance extension problem: The complete solution. *IEEE Transactions on Automatic Control*, 58(11):2848–2861, 2013.

[72] A. Lindquist and G. Picci. *Linear Stochastic Systems*. Springer, Berlin Heidelberg, 2015.

[73] L. Ljung. *System identification: Theory for the user*. Prentice Hall, Upper Saddle River, NJ, 2nd edition, 1999.

[74] L. Ljung and J. Sjöberg. A system identification perspective on neural nets. In *IEEE Workshop on Neural Networks for Signal Processing*, pages 423–435. IEEE, 1992.

[75] D.G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, New York, NY, 1969.

[76] D.G. Luenberger. *Introduction to dynamic systems. Theory, models, and applications*. John Wiley & Sons, New York, NY, 1979.

[77] T.L. Magnanti. Fenchel and Lagrange duality are equivalent. *Mathematical Programming*, 7(1):253–258, 1974.

[78] J. Markoff. How many computers to identify a cat? 16,000. *The New York Times*, page B1, 26th of June 2012. Retrieved 29 October 2018 online via `https://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html`.

[79] L.R. Mead and N. Papanicolaou. Maximum entropy in the problem of moments. *Journal of Mathematical Physics*, 25(8):2404–2417, 1984.

[80] P.A. Midgley and R.E. Dunin-Borkowski. Electron tomography and holography in materials science. *Nature materials*, 8(4):271, 2009.

[81] F. Natterer. *The Mathematics of Computerized Tomography*. SIAM, Philadelphia, PA, 2001.

[82] F. Natterer and F. Wübbeling. *Mathematical Methods in Image Reconstruction*. SIAM, Philadelphia, PA, 2001.

[83] N.K. Nikolski. *Operators, Functions, and Systems: An Easy Reading. Volume I: Hardy, Hankel, and Toeplitz*. American Mathematical Society, Boston, MA, 2002.

[84] O. Öktem. Mathematics of electron tomography. In O. Scherzer, editor, *Handbook of Mathematical Methods in Imaging*, pages 937–1031. Springer, New York, NY, 2015.

[85] M. Olivi. The Laplace transform in control theory. In J.-D. Fournier, J. Grimm, J. Leblond, and J.R. Partington, editors, *Harmonic Analysis and Rational Approximation*, pages 193–209. Springer, 2006.

[86] A.V. Oppenheim, A.S. Willsky, and S.H. Nawab. *Signals and systems*. Prentice-Hall, Upper Saddle River, NJ, 2nd edition, 1997.

[87] V. Papyan, Y. Romano, and M. Elad. Convolutional neural networks analyzed via convolutional sparse coding. *The Journal of Machine Learning Research*, 18(1):2887–2938, 2017.

[88] V. Papyan, Y. Romano, J. Sulam, and M. Elad. Theoretical foundations of deep learning via sparse representations: A multilayer sparse model and its connection to convolutional neural networks. *IEEE Signal Processing Magazine*, 35(4):72–89, 2018.

[89] J.R. Partington. *Interpolation, identification, and sampling*. Clarendon, Oxford, 1997.

[90] T. Qi, J. Zhu, and J. Chen. Fundamental bounds on delay margin: When is a delay system stabilizable? In *Chinese Control Conference (CCC)*, pages 6006–6013. IEEE, 2014.

[91] T. Qi, J. Zhu, and J. Chen. Fundamental limits on uncertain delays: When is a delay system stabilizable by LTI controllers? *IEEE Transactions on Automatic Control*, 62(3):1314–1328, 2017.

[92] K.J. Åström. *Introduction to stochastic control theory*. Dover, Mineola, NY, 2006. Unabridged republication of original published by Academic Press, 1970.

[93] K.J. Åström and R.M. Murray. *Feedback systems*. Princeton university press, Princeton, NJ, 2008.

[94] A. Ringh, J. Karlsson, and A. Lindquist. The multidimensional circulant rational covariance extension problem: Solutions and applications in image compression. In *IEEE Annual Conference on Decision and Control (CDC)*, pages 5320–5327. IEEE, 2015.

[95] A. Ringh, J. Karlsson, and A. Lindquist. Multidimensional rational covariance extension with applications to spectral estimation and image compression. *SIAM Journal on Control and Optimization*, 54(4):1950–1982, 2016.

[96] R.T. Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, NJ, 1970.

[97] W. Rudin. *Principles of mathematical analysis*. McGraw-Hill, New York, NY, 3rd edition, 1976.

[98] W. Rudin. *Real and complex analysis*. McGraw-Hill, New York, NY, 1987.

[99] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533, 1986.

[100] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[101] L. Ruthotto and E. Haber. Deep neural networks motivated by partial differential equations. *arXiv preprint arXiv:1804.04272*, 2018.

[102] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[103] J.-P. Schott and J.H. McClellan. Maximum entropy power spectrum estimation with uncertainty in correlation measurements. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(2):410–418, 1984.

[104] T. Söderström and P. Stoica. *System Identification*. Prentice Hall International, Hemel Hempstead, Hertfordshire, 1989.

[105] P. Stoica and R.L. Moses. *Spectral analysis of signals*. Prentice-Hall, Upper Saddle River, NJ, 2005.

[106] A. Villani. Another note on the inclusion $L^p(\mu) \subset L^q(\mu)$. *The American Mathematical Monthly*, 92(7):485–487, 1985.

[107] C. Villani. *Optimal transport: Old and new*. Springer, Berlin Heidelberg, 2008.

[108] A. Vretblad. *Fourier Analysis and Its Applications*. Springer, New York, NY, 2003.

"Scientists have calculated that the chance of anything so patently absurd actually existing are millions to one. But magicians have calculated that million-to-one chances crop up nine times out of ten."

— Terry Pratchett

"In actual fact we should recognise the general principle that a lack of information cannot be remedied by any mathematical trickery."

— Cornelius Lanczos

Part II: Research Papers

Paper A

Multidimensional rational covariance extension

Paper B

Multidimensional rational covariance extension with approximate covariance matching

Paper C

Lower bounds on the maximum delay margin by analytic interpolation

Paper D

Generalized Sinkhorn iterations for regularizing inverse problems using optimal mass transport

Paper E

Learning to solve inverse problems using Wasserstein loss

Paper F

Data-driven nonsmooth optimization