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 framläges 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.

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

"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.

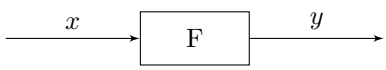


Figure 1.1: Schematic figure of a system F, mapping from input x to output y.

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.

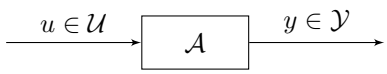


Figure 2.1: A linear dynamical system.

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 $x_1 = y$, $\ldots$, $x_k = y^{(k-1)}$, $x_{k+1} = v$, $\ldots$, $x_{k+m} = v^{m-1}$, 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.

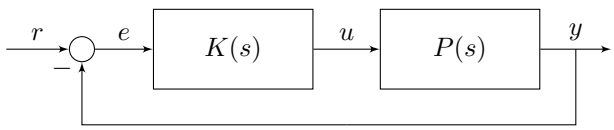


Figure 2.2: A feedback control system.

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

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

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

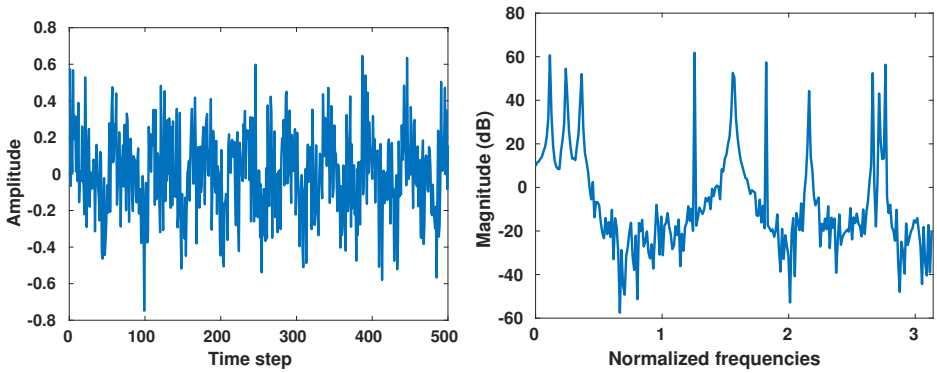
$$\Phi(e^{i\theta}) = \sum_{k=-\infty}^{\infty} c_k e^{-ik\theta}, \tag{2.2.1}$$

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].

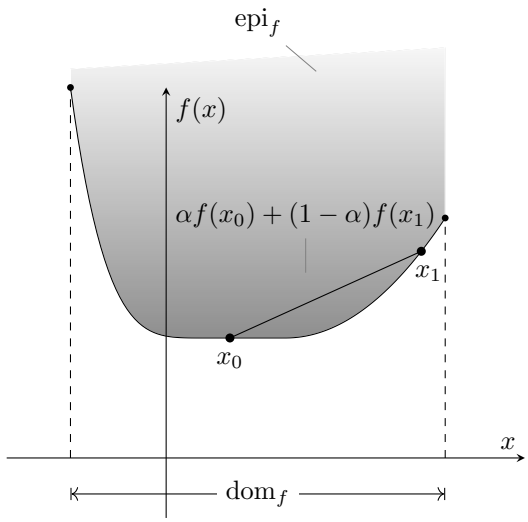


Figure 2.4: Figure illustrating the notions epi$_f$, dom$_f$, and convexity. Note also that the function is lower semi-continuous, and thus the epigraph is closed.

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

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

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].

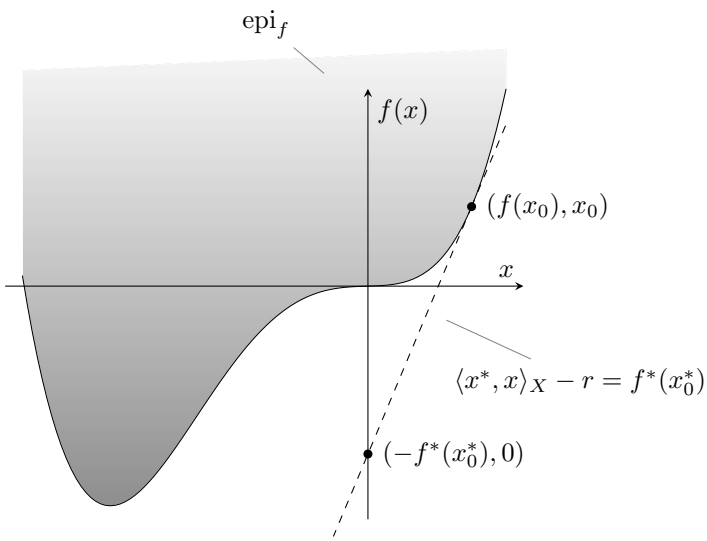


Figure 2.5: The convex conjugate defines supporting hyperplanes for the epigraph.

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

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

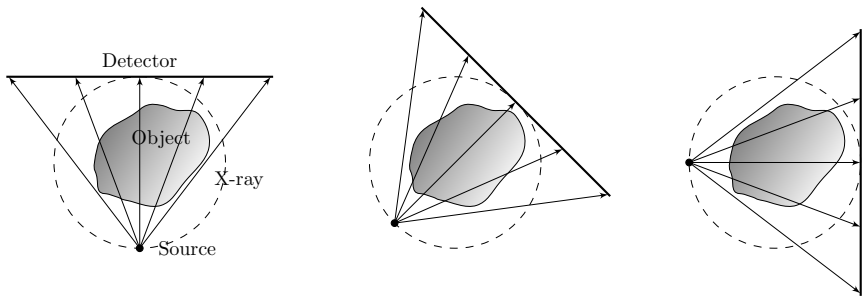
cf. [81, p. 1]. Here, I_{in} is the intensity of the incoming X-ray beam and I_{out} is the intnesity of the outgoing X-ray beam. Both of these are known: I_{in} by the design of the machine and I_{out} by measuring it using detectors. Thus, in tomographic imaging the data acquisition is normally modeled 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 $z = h > 0$.

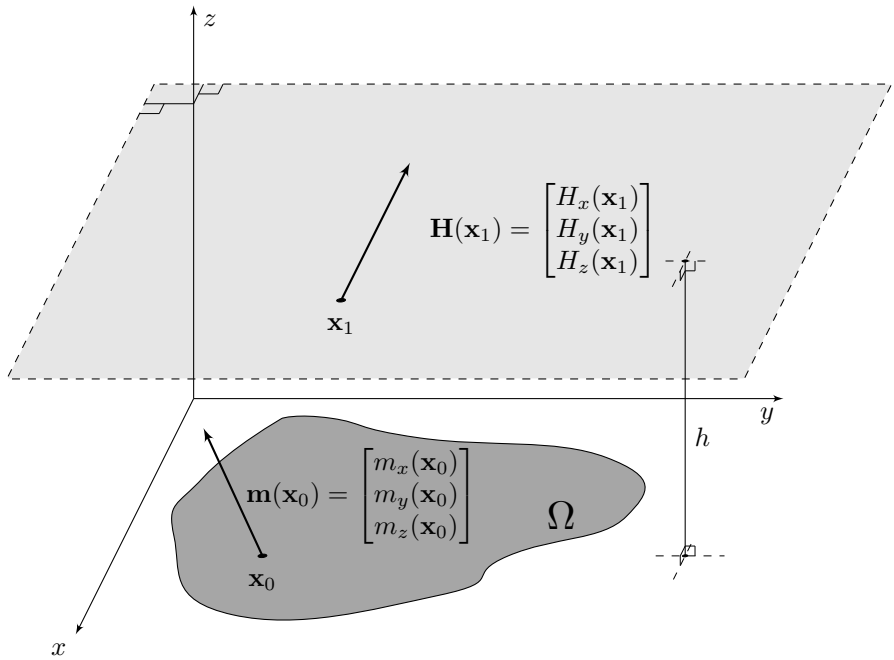


Figure 2.7: Figure illustrating the geometry of the inverse problem in magnetization. $\Omega \subset \mathbb{R}^2_{z=0} := \{(x, y, z) \in \mathbb{R}^3 \mid z = 0\}$ is the compact set on which the magnetization $\mathbf{m}$ has support. Measurements of H_z are taken at hight h above the thin specimen,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

Multidimensional rational covariance extension

by

Axel Ringh, Johan Karlsson, and Anders Lindquist

Abstract

The rational covariance extension problem (RCEP) is an important problem in systems and control occurring in such diverse fields as control, estimation, system identification, and signal and image processing, leading to many fundamental theoretical questions. In fact, this inverse problem is a key component in many identification and signal processing techniques and plays a fundamental role in prediction, analysis, and modeling of systems and signals. It is well-known that the RCEP can be reformulated as a (truncated) trigonometric moment problem subject to a rationality condition. In this paper we consider the more general multidimensional trigonometric moment problem with a similar rationality constraint. This generalization creates many interesting new mathematical questions and also provides new insights into the original one-dimensional problem. A key concept in this approach is the complete smooth parametrization of all solutions, allowing solutions to be tuned to satisfy additional design specifications without violating the complexity constraints. As an illustration of the potential of this approach we apply our results to multidimensional spectral estimation, Wiener system identification, and image compression.

Keywords: covariance extension, trigonometric moment problem, convex optimization, generalized entropy, multidimensional spectral estimation, system identification, image compression

A.1 Introduction

In this paper we consider the (truncated) multidimensional trigonometric moment problem with a certain complexity constraint. Many problems in multidimensional systems theory including realization, control, and identification problems, can be cast in this framework [6]. Other applications of this type are image processing [25] and spectral estimation in radar, sonar, and medical imaging [79].

More precisely, given a set of complex numbers $c_{\mathbf{k}}$, $\mathbf{k} \in \Lambda$, where $\mathbf{k} := (k_1, \dots, k_d)$ is a vector-valued index belonging to a specified index set $\Lambda \subset \mathbb{Z}^d$, find a nonnegative bounded measure $d\mu$ such that

$$c_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} d\mu(\boldsymbol{\theta}) \quad \text{for all } \mathbf{k} \in \Lambda, \tag{A.1.1}$$

where $\mathbb{T} := (-\pi, \pi]$, $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_d) \in \mathbb{T}^d$, and $(\mathbf{k}, \boldsymbol{\theta}) := \sum_{j=1}^{d} k_j \theta_j$ is the scalar product in $\mathbb{R}^d$. Moreover, let $e^{i\boldsymbol{\theta}} := (e^{i\theta_1}, \ldots, e^{i\theta_d})$. By the Lebesgue decomposition [75, p. 121], the measure $d\mu$ can be decomposed in a unique fashion as

$$d\mu(\boldsymbol{\theta}) = \Phi(e^{i\boldsymbol{\theta}}) dm(\boldsymbol{\theta}) + d\hat{\mu}(\boldsymbol{\theta}) \tag{A.1.2a}$$

with an absolutely continuous part Φdm with spectral density Φ and Lebesgue measure

$$dm(\boldsymbol{\theta}) := (1/2\pi)^d \prod_{j=1}^{d} d\theta_j$$

and a singular part $d\hat{\mu}$ containing, e.g., spectral lines. This is an inverse problem, which in general has infinitely many solutions if one exists. A first problem of interest to us in this paper is how to smoothly parametrize the family of all solutions that satisfy the rational complexity constraint

$$\Phi(e^{i\boldsymbol{\theta}}) = \frac{P(e^{i\boldsymbol{\theta}})}{Q(e^{i\boldsymbol{\theta}})}, \quad \text{where } P, Q \in \bar{\mathfrak{P}}_+ \backslash \{0\}, \tag{A.1.2b}$$

where $\mathfrak{P}_+$ is the convex cone of positive trigonometric polynomials

$$P(e^{i\boldsymbol{\theta}}) = \sum_{\mathbf{k} \in \Lambda} p_{\mathbf{k}} e^{-i(\mathbf{k}, \boldsymbol{\theta})} \tag{A.1.3}$$

that are positive for all $\boldsymbol{\theta} \in \mathbb{T}^d$, and $\bar{\mathfrak{P}}_+$ is its closure; $\mathfrak{P}_+$ will be called the *positive cone*. Moreover, we use the notation $\partial\mathfrak{P}_+ := \bar{\mathfrak{P}}_+ \backslash \mathfrak{P}_+$ for its boundary, i.e., the subset of nonnegative $P \in \bar{\mathfrak{P}}_+$ that are zero in at least one point. In this paper we develop a theory based on convex optimization for this problem.

For $d = 1$ and $\Lambda = \{0, 1, \ldots, n\}$ this trigonometric moment problem with complexity constraints is well understood, and it has a solution with $d\hat{\mu} = 0$ if and only if the Toeplitz matrix

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

is positive definite [55]. Such a sequence, $c_0, \ldots, c_n$, will therefore be called a positive sequence in this paper.

In his pioneering work on spectral estimation, J.P. Burg observed that among all spectral densities Φ satisfying the moment constraints

$$c_k = \int_{\mathbb{T}} e^{ik\theta} \Phi(e^{i\theta}) \frac{d\theta}{2\pi}, \quad k = 0, 1, \ldots, n, \tag{A.1.4a}$$

the one with maximal entropy

$$\int_{\mathbb{T}} \log \Phi(e^{i\theta}) \frac{d\theta}{2\pi} \tag{A.1.4b}$$

is of the form $\Phi(e^{i\theta}) = 1/Q(e^{i\theta})$, where $Q(e^{i\theta})$ is a positive trigonometric polynomial [7, 8]. Later, in 1981, R.E. Kalman posed the *rational covariance extension problem* (RCEP) [43]: given a finite covariance sequence $c_0, \ldots, c_n$, determine all infinite extensions $c_{n+1}, c_{n+2}, \ldots$ such that

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

is a positive rational function of degree bounded by $2n$. This problem, which is important in systems theory [55], is precisely a (one-dimensional) trigonometric moment problem with the complexity constraint (A.1.2b). The designation "covariance" emanates from the fact that $c_0, c_1, c_2, \ldots$, can be interpreted as the covariance lags $\mathbb{E}\{y(t+k)\overline{y(t)}\} = c_k$ of a wide-sense stationary stochastic process y with spectral density Φ.

In 1983, T.T. Georgiou [33] (also see [34]) proved that to each positive covariance sequence and positive numerator polynomial P, there exists a rational covariance extension of the sought form (A.1.2b). He also conjectured that this extension is unique and hence gives a complete parameterization of all rational extensions of degree bounded by $2n$. This conjecture was first proven in [19], where it was also shown that the complete parameterization is smooth, allowing for tuning. The proofs in [33, 34, 19] were nonconstructive, using topological methods. Later a constructive proof was given in [14, 15], leading to an approach based on convex optimization. Here Φ is obtained as the maximizer of a generalized entropy functional

$$\int_{\mathbb{T}} P(e^{i\theta}) \log \Phi(e^{i\theta}) \frac{d\theta}{2\pi} \tag{A.1.5}$$

subject to the moment conditions (A.1.4a), and the problem is reduced to solving a dual convex optimization problem. Since then, this approach have been extensively studied [35, 15, 9, 10, 27, 62, 54, 73, 71, 11, 84, 28, 64], and the approach has also been generalized to a quite complete theory for scalar moment problems [12, 16, 38, 13, 17]. Moreover a number of multivariate counterparts, i.e., when Φ is matrix-valued, have also been solved [32, 37, 65, 5, 67, 53, 83, 2].

A considerable amount of research has also been done in the area of multidimensional spectral estimation; for example, Woods [82], Ekstrom and Woods [26], Dickinson [23], and Lev-Ari *et al.* [51] to mention a few. Of special interest are also results by Lang and McClellan [49, 50, 59, 60, 48, 47], as they consider a similar entropy functional. In many of these areas it seems natural to consider rational models. Nevertheless, the multidimensional version of the RCEP has only been considered at a few instances, for the two-dimensional case in [37, 36] and in the

more general setting of moment problems with arbitrary basis functions in our recent paper [46].

The purpose of this paper is to extend the theory of rational covariance extension from the one-dimensional to the general d-dimensional case and to develop methods for multidimensional spectral estimation. In Section A.2 we summarize the main theoretical results of the paper. This includes the main theorem characterizing the optimal solutions to the weighted entropy functional, which is then proved in Section A.3. In Section A.4 we prove that under certain assumptions the problem is well-posed in the sense of Hadamard and provide comments and examples related to these assumptions. In Section A.5 we consider simultaneous matching of covariance lags and logarithmic moments, and Section A.6 is devoted to a discrete version of the problem, where the measure $d\mu$ consists of discrete point masses placed equidistantly in a discrete grid in $\mathbb{T}^d$. This is a generalization to the multidimensional case of recent results in [54] and is motivated by computational considerations. In fact, these discrete solutions provide approximations to solutions to moment problems with absolutely continuous measures and allow for fast arithmetics based on the fast Fourier transform (FFT) (cf. [71]). Sections A.7, A.8, and A.9 are devoted to three examples of how the theory can be applied; the first in system identification, the second in Wiener system identification and texture generation, and the third in image compression. Finally, the paper also contains an appendix to which some proofs are deferred in order to improve readability.

A.2 Main results

Given the moments $\{c_{\mathbf{k}}\}_{\mathbf{k}\in\Lambda}$, the problem under consideration is to find a positive measure (A.1.2) of bounded variation satisfying the moment constraint (A.1.1). Let us pause to pin down the structure of the index set Λ. In view of (A.1.1), we have $c_{-\mathbf{k}} = \bar{c}_{\mathbf{k}}$, where $\bar{}$ denotes complex conjugation. Revisiting the one-dimensional result [16, 18, 17] for moment problems with arbitrary basis functions, we observe that the theory holds also for sequences with "gaps", e.g., for a sequence $c_0, \ldots, c_{k-1}, c_{k+1}, \ldots, c_n$. As seen in [46] this observation equally applies to the multidimensional case. Therefore, we shall consider covariance sequences $\{c_{\mathbf{k}}\}_{\mathbf{k}\in\Lambda}$, where $\Lambda \subset \mathbb{Z}^d$ is any finite index set such that $0 \in \Lambda$ and $-\Lambda = \Lambda$. We will denote the cardinality of Λ by $|\Lambda|$. Further, let $n_j = \max\{k_j \,|\, \mathbf{k} \in \Lambda\}$ denote the maximum range of Λ in dimension j.

Next, given the inner product

$$\langle c, p\rangle = \sum_{\mathbf{k}\in\Lambda} c_{\mathbf{k}}\bar{p}_{\mathbf{k}},$$

we define the open convex cone

$$\mathfrak{C}_+ := \left\{c \mid \langle c, p\rangle > 0 \text{ for all } P \in \bar{\mathfrak{P}}_+ \setminus \{0\}\right\},$$

the closure of which, $\bar{\mathfrak{C}}_+$, is the dual cone of $\bar{\mathfrak{P}}_+$, with boundary $\partial\mathfrak{C}_+$.

We now extend the domain of the generalized entropy functional in (A.1.5) to multidimensional nonnegative measures of the type (A.1.2) and consider functionals

$$\mathbb{I}_P(d\mu) = \int_{\mathbb{T}^d} P(e^{i\boldsymbol{\theta}}) \log \Phi(e^{i\boldsymbol{\theta}})\, dm(\boldsymbol{\theta}), \tag{A.2.1}$$

where Φ is the absolutely continuous part of $d\mu$.[1] This functional is concave, but not strictly concave since the singular part of the measure does not influence the value. This leads to the optimization problem to maximize (A.2.1) subject to the moment constraints (A.1.1). Since the constraints are linear, this is a convex problem. However, as it is an infinite-dimensional optimization problem, it is more convenient to work with the dual problem, which has a finite number of variables but an infinite number of constraints. In fact, the dual problem amounts to minimizing

$$\mathbb{J}_P(Q) = \langle c, q \rangle - \int_{\mathbb{T}^d} P(e^{i\boldsymbol{\theta}}) \log Q(e^{i\boldsymbol{\theta}}) dm(\boldsymbol{\theta}) \tag{A.2.2}$$

over all $Q \in \bar{\mathfrak{P}}_+$, and hence $Q(e^{i\boldsymbol{\theta}}) \geq 0$ for all $\boldsymbol{\theta} \in \mathbb{T}^d$. Note that (A.2.2) takes an infinite value for $Q \equiv 0$.

Theorem A.2.1. *For every $c \in \mathfrak{C}_+$ and $P \in \bar{\mathfrak{P}}_+ \setminus \{0\}$ the functional* (A.2.2) *is strictly convex and has a unique minimizer $\hat{Q} \in \bar{\mathfrak{P}}_+ \setminus \{0\}$. Moreover, there exists a unique $\hat{c} \in \partial\mathfrak{C}_+$ and a nonnegative singular measure $d\hat{\mu}$ which has support $\operatorname{supp}(d\hat{\mu}) \subseteq \{\boldsymbol{\theta} \in \mathbb{T}^d \mid \hat{Q}(e^{i\boldsymbol{\theta}}) = 0\}$ such that*

$$c_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} \left(\frac{P}{\hat{Q}} dm + d\hat{\mu} \right) \quad \textit{for all } \mathbf{k} \in \Lambda$$

and

$$\hat{c}_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} d\hat{\mu}, \ \textit{for all } \mathbf{k} \in \Lambda.$$

For any such $d\hat{\mu}$, the measure $d\mu(\boldsymbol{\theta}) = (P(e^{i\boldsymbol{\theta}})/\hat{Q}(e^{i\boldsymbol{\theta}}))dm(\boldsymbol{\theta}) + d\hat{\mu}(\boldsymbol{\theta})$ is an optimal solution to the problem to maximize (A.2.1) *subject to the moment constraints* (A.1.1)*. Moreover, $d\hat{\mu}$ can be chosen with support in at most $|\Lambda| - 1$ points.*

Corollary A.2.2. *Let $c \in \mathfrak{C}_+$. Then, for any*

$$d\mu = \frac{P}{Q} dm, \quad P, Q \in \bar{\mathfrak{P}}_+ \setminus \{0\}$$

satisfying the moment condition (A.1.1)*, Q is the unique minimizer over $\bar{\mathfrak{P}}_+$ of the dual functional* (A.2.2)*.*

[1] Note that the absolutely continuous part is uniquely defined by the Lebesgue decomposition, and hence the functional $\mathbb{I}_P(d\mu)$ is uniquely defined. Moreover, this definition of $\mathbb{I}_P(d\mu)$ can be motivated by the fact that $\lim_{n\to\infty} \int_{\mathbb{T}^d} \log(\Phi(e^{i\boldsymbol{\theta}}) + f_n(\boldsymbol{\theta}))dm(\boldsymbol{\theta}) = \int_{\mathbb{T}^d} \log(\Phi(e^{i\boldsymbol{\theta}}))dm(\boldsymbol{\theta})$ for any log-integrable Φ and nonnegative "good kernel" $f_n(\boldsymbol{\theta})$ (see, e.g., [78, p. 48]). See also the discussion in Section A.3.

This corollary implies that, for any $c \in \mathfrak{C}_+$, any measure $d\mu$ with only absolutely continuous rational part matching c can be obtained by solving (A.2.2) for a suitable P. However, although $c \in \mathfrak{C}_+$, not all P result in an absolutely continuous solution $d\mu = (P/Q)dm$ that satisfies (A.1.1). Nevertheless, the case when this happens is of particular interest.

Corollary A.2.3. *Suppose that $d \leq 2$. Then, for any $c \in \mathfrak{C}_+$ and $P \in \mathfrak{P}_+$ there exists a $Q \in \mathfrak{P}_+$ such that $d\mu = (P/Q)dm$ satisfies* (A.1.1)*. Moreover this Q is the unique solution to the strictly convex optimization problem to minimize the dual functional* (A.2.2) *over all $Q \in \mathfrak{P}_+$.*

This result can be deduced from the early work of Lang and McClellan [49], although they do not consider rational solutions explicitly, nor parameterizations of them. Note that Corollary A.2.3 is only valid for $P \in \mathfrak{P}_+$, while Theorem A.2.1 holds for all $P \in \bar{\mathfrak{P}}_+ \setminus \{0\}$. This will be further discussed in Section A.4, where the proof of Corollary A.2.3 will also be concluded.

Covariance and cepstral matching

It follows from Theorem A.2.1 and Corollary A.2.3 that Q is completely determined by the pair (c, P). For $d = 1$ the choice $P \equiv 1$ leads to Burg's formulation (A.1.4), which has been termed the *maximum-entropy* (ME) solution. On the other hand, better dynamical range of the spectrum can be obtained by taking advantage of the extra degrees of freedom in P. Several methods for selecting P have been suggested in the one-dimensional setting. Examples are methods based on inverse problems as in [44, 30, 45], a linear-programming approach as in [9, 10], and simultaneous matching of covariances and cepstral coefficients as in [61] and independently in [9, 10, 27, 54]. Here, in the multivariate setting, we consider the selection of P based on the simultaneous matching of logarithmic moments.

We define the (real) *cepstrum* of a multidimensional spectrum as the (real) logarithm of its absolutely continuous part. The *cepstral coefficients* are the corresponding Fourier coefficients

$$\gamma_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} \log \Phi(e^{i\boldsymbol{\theta}}) dm(\boldsymbol{\theta}) \quad \text{for } \mathbf{k} \in \Lambda \setminus \{0\}. \tag{A.2.3}$$

For spectra that only have an absolutely continuous part this agrees with earlier definitions in the literature (see, e.g., [63, pp. 500-507] or [22, Chapter 6]).

Given a set of cepstral coefficients we now also enforce cepstral matching of the sought family of spectra. This means that we look for $\Phi = P/Q$ that also satisfies (A.2.3). Note that the index $\mathbf{k} = 0$ is not included in (A.2.3). In fact, for technical reasons, we shall set $\gamma_0 = 1$. Also to avoid trivial cancelations of constants in P/Q, we need to introduce the set

$$\mathfrak{P}_{+,\circ} := \{P \in \mathfrak{P}_+ \mid p_0 = 1\}.$$

Theorem A.2.4. *Let $\gamma_{\mathbf{k}}$, $\mathbf{k} \in \Lambda \setminus \{0\}$, be any sequence of complex numbers such that $\gamma_{-\mathbf{k}} = \bar{\gamma}_{\mathbf{k}}$, and set $\gamma = \{\gamma_{\mathbf{k}}\}_{\mathbf{k}\in\Lambda}$, where $\gamma_0 = 1$. Then, for $c \in \mathfrak{C}_+$, the convex optimization problem (D) to minimize*

$$\mathbb{J}(P,Q) = \langle c, q\rangle - \langle \gamma, p\rangle + \int_{\mathbb{T}^d} P \log\left(\frac{P}{Q}\right) dm \tag{A.2.4}$$

subject to $(P,Q) \in \bar{\mathfrak{P}}_{+,\circ} \times \bar{\mathfrak{P}}_+$ has an optimal solution $(\hat{P},\hat{Q})$. If such a solution belongs to $\mathfrak{P}_{+,\circ} \times \mathfrak{P}_+$, then $\hat{\Phi} = \hat{P}/\hat{Q}$ satisfies the logarithmic moment condition (A.2.3) *and $d\mu = \hat{\Phi}dm$ the moment condition* (A.1.1)*. Moreover, $\hat{\Phi}$ is also an optimal solution to the problem (P) to maximize*

$$\mathbb{I}(\Phi) = \int_{\mathbb{T}^d} \log\Phi \, dm \tag{A.2.5}$$

subject to (A.1.1) *and* (A.2.3) *for $d\mu = \Phi dm$. Finally, if $d \le 2$, then $\hat{P} \in \mathfrak{P}_{+,\circ}$ implies that $\hat{Q} \in \mathfrak{P}_+$.*

For reasons to become clear in Section A.5, the optimization problems (P) and (D) will be referred to as the primal and dual problems, respectively. A drawback with Theorem A.2.4 is that even when $d \le 2$, a solution to the dual problem can be guaranteed to have a rational spectrum that satisfies (A.1.1) and (A.2.3) only if $\hat{P} \in \mathfrak{P}_{+,\circ}$. In fact, as we shall see in Section A.5, for a solution with $\hat{P} \in \partial\mathfrak{P}_{+,\circ}$ we might have $\hat{Q} \in \partial\mathfrak{P}_+$ and hence covariance mismatch. A remedy in the case $d \le 2$ is to use the Enqvist regularization, introduced in the one-dimensional setting in [27]. This makes the optimization problem strictly convex and forces the solution $\hat{P}$ into the set $\mathfrak{P}_{+,\circ}$. In this way we obtain strict covariance matching and approximative cepstral matching. This statement will be made precise in Theorem A.5.7 in Section A.5.

The circulant covariance extension problem

In the recent paper [54], Lindquist and Picci studied, for the case $d = 1$, the situation when the underlying stochastic process $y(t)$ is periodic. For the N-periodic case, the covariance sequence must satisfy the extra condition $c_{N-k} = \bar{c}_k$, i.e., the $N \times N$ Toeplitz matrix of one period is Hermitan *circulant*. In this case, the spectral measure must be discrete with point masses at $\zeta_\ell = e^{i\ell\frac{2\pi}{N}}$, $\ell = 0, 1, \dots, N-1$, on the discrete unit circle, and instead of the moment condition (A.1.1) we have

$$c_k = \frac{1}{N}\sum_{\ell=0}^{N-1} \Phi(\zeta_\ell)\zeta_\ell^k, \tag{A.2.6}$$

which is the inverse discrete Fourier transform of the sequence $(\Phi(\zeta_\ell))$.

This was generalized to the multidimensional case in [72], where a circulant version of Theorem A.2.1 and Corollary A.2.3 was derived. For $\mathbf{N} := (N_1, \dots, N_d)$, consider the discretization of the d-dimensional torus

$$\boldsymbol{\zeta}_{\boldsymbol{\ell}} := (e^{i\ell_1 \frac{2\pi}{N_1}}, \dots, e^{i\ell_d \frac{2\pi}{N_d}}),$$

where

$$\mathbb{Z}^d_{\mathbf{N}} := \{\boldsymbol{\ell} = (\ell_1, \dots, \ell_d) \mid 0 \leq \ell_j \leq N_j - 1, j = 1, \dots, d\},$$

and define $\boldsymbol{\zeta}_{\boldsymbol{\ell}}^{\mathbf{k}} = \prod_{j=1}^d \zeta_{\ell_j}^{k_j}$. Next, let $\mathfrak{P}_+(\mathbf{N})$ be the positive cone of all trigonometric polynomials (A.1.3) such that $P(\boldsymbol{\zeta}_{\boldsymbol{\ell}}) > 0$ for all $\boldsymbol{\ell} \in \mathbb{Z}^d_{\mathbf{N}}$. Moreover, define the interior $\mathfrak{C}_+(\mathbf{N})$ of the dual cone as the set of all $\{c_{\mathbf{k}}\}_{\mathbf{k}\in\Lambda}$ such that $\langle c, p\rangle > 0$ for all $P \in \bar{\mathfrak{P}}_+(\mathbf{N}) \setminus \{0\}$. Clearly $\mathfrak{P}_+(\mathbf{N}) \supset \mathfrak{P}_+$, and hence $\mathfrak{C}_+(\mathbf{N}) \subset \mathfrak{C}_+$. Then Theorem 2 and Corollary 3 in [72] can be combined in the following theorem.

Theorem A.2.5 ([72])**.** *Suppose that $2n_j < N_j$, for $j = 1, \dots, d$, and let $c \in \mathfrak{C}_+(\mathbf{N})$ and $P \in \bar{\mathfrak{P}}_+(\mathbf{N}) \setminus \{0\}$. Then, there exists a $\hat{Q} \in \bar{\mathfrak{P}}_+(\mathbf{N}) \setminus \{0\}$ such that $\hat{Q}$ is a solution to the convex problem to minimize*[2]

$$\mathbb{J}_P^{\mathbf{N}}(Q) = \langle c, q\rangle - \frac{1}{\prod_{j=1}^d N_j} \sum_{\boldsymbol{\ell}\in\mathbb{Z}^d_{\mathbf{N}}} P(\boldsymbol{\zeta}_{\boldsymbol{\ell}}) \log Q(\boldsymbol{\zeta}_{\boldsymbol{\ell}})$$

over all $Q \in \bar{\mathfrak{P}}_+(\mathbf{N})$. Moreover, there exists a nonnegative function $\hat{\mu}$ with support $\operatorname{supp}(\hat{\mu}) = \{\boldsymbol{\zeta}_{\boldsymbol{\ell}} \,|\, \hat{Q}(\boldsymbol{\zeta}_{\boldsymbol{\ell}}) = 0,\, \boldsymbol{\ell} \in \mathbb{Z}^d_{\mathbf{N}}\}$ such that

$$c_{\mathbf{k}} = \frac{1}{\prod_{j=1}^d N_j} \sum_{\boldsymbol{\ell}\in\mathbb{Z}^d_{\mathbf{N}}} \boldsymbol{\zeta}_{\boldsymbol{\ell}}^{\mathbf{k}} \left(\frac{P(\boldsymbol{\zeta}_{\boldsymbol{\ell}})}{\hat{Q}(\boldsymbol{\zeta}_{\boldsymbol{\ell}})} + \hat{\mu}(\boldsymbol{\zeta}_{\boldsymbol{\ell}}) \right), \tag{A.2.7}$$

and the number of mass points for $\hat{\mu}$ can be chosen so that at most $|\Lambda| - 1$ points $\hat{\mu}(\boldsymbol{\zeta}_{\boldsymbol{\ell}})$ are nonzero. Finally, if $P \in \mathfrak{P}_+(\mathbf{N})$ then $\hat{Q} \in \mathfrak{P}_+(\mathbf{N})$, which is then also unique, and hence $\Phi = P/\hat{Q}$ satisfies (A.2.7) *with $\hat{\mu} \equiv 0$.*

In [54] it was shown in the one-dimensional case that as $N \to \infty$ the solution of the discrete problem, call it $\hat{Q}_N$, converges to the solution to the corresponding continuous problem, call it $\hat{Q}$. This gives a natural way to compute an approximate solution to the continuous problem using the fast computations of the discrete Fourier transform. The same also holds in higher dimensions, as seen in the following result.

Theorem A.2.6. *Suppose that $P \in \bar{\mathfrak{P}}_+ \setminus \{0\}$ and $c \in \mathfrak{C}_+$, and let $\hat{Q}$ and $\hat{Q}_{\mathbf{N}}$ be the optimal solutions of Theorems A.2.1 and A.2.5, respectively. Then*

$$\lim_{\min(\mathbf{N})\to\infty} \hat{Q}_{\mathbf{N}} = \hat{Q}$$

uniformly.

[2]Note that limits such as $P\log(Q)$ and P/Q may not be well defined in the multidimensional case, and therefore we define the expressions $P\log(Q)$ and P/Q to be zero whenever $P = 0$. This is not needed in the continuous case as the set where P is zero is of measure zero.

A.3 The multidimensional RCEP

Most of this section will be devoted to proving Theorem A.2.1. Some technical details are deferred to the appendix. Possible interpretations of P will be discussed in the end of the section together with an example showing the non-uniqueness of the measure $d\hat{\mu}$.

Proof of Theorem A.2.1

Deriving the dual problem

For a given $P \in \bar{\mathfrak{P}}_+ \setminus \{0\}$ and $c \in \mathfrak{C}_+$, consider the primal problem to maximize (A.2.1) subject to the moment constraints (A.1.1) over the set of nonnegative bounded measures, i.e., over $d\mu = \Phi dm + d\hat{\mu}$, where Φ is a nonnegative $L^1(\mathbb{T}^d)$ function and $d\hat{\mu}$ is a nonnegative singular measure. The Lagrangian of this problem becomes

$$\mathcal{L}_P(\Phi, d\hat{\mu}, Q) = \int_{\mathbb{T}^d} P \log \Phi dm + \sum_{\mathbf{k}\in\Lambda} \bar{q}_{\mathbf{k}} \left(c_{\mathbf{k}} - \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} (\Phi dm + d\hat{\mu}) \right),$$

where $\bar{q}_{\mathbf{k}}$, $\mathbf{k} \in \Lambda$, are Lagrange multipliers. Identifying $\sum_{\mathbf{k}\in\Lambda} \bar{q}_{\mathbf{k}} e^{i(\mathbf{k},\boldsymbol{\theta})}$ with the trigonometric polynomial Q, this can be simplified to

$$\mathcal{L}_P(\Phi, d\hat{\mu}, Q) = \int_{\mathbb{T}^d} P \log \Phi \, dm + \langle c, q \rangle - \int_{\mathbb{T}^d} Q\Phi dm - \int_{\mathbb{T}^d} Q d\hat{\mu}.$$

The dual function $\sup_{d\mu\geq 0} \mathcal{L}_P(\Phi, d\hat{\mu}, Q)$ is finite only if $Q \in \bar{\mathfrak{P}}_+ \setminus \{0\}$. To see this, let $Q \notin \bar{\mathfrak{P}}_+$, i.e., suppose there is a $\boldsymbol{\theta}_0 \in \mathbb{T}^d$ for which $Q(\boldsymbol{\theta}_0) < 0$. Then, by letting $\hat{\mu}(\boldsymbol{\theta}_0) \to \infty$ in the singular part $d\hat{\mu}$, we get that $\mathcal{L}_P(\Phi, d\hat{\mu}, Q) \to \infty$. Moreover, if $Q \equiv 0$ then since P is continuous and $P \not\equiv 0$ there is a small neighbourhood where $P > 0$. Letting $\Phi \to \infty$ in this neighbourhood we again have that $\mathcal{L}_P(\Phi, d\hat{\mu}, Q) \to \infty$. Hence we can restrict the multipliers to $\bar{\mathfrak{P}}_+ \setminus \{0\}$.

Now note that any pair $(\Phi, d\hat{\mu})$ maximizing $\mathcal{L}_P(\Phi, d\hat{\mu}, Q)$ must satisfy $\int_{\mathbb{T}^d} Q d\hat{\mu} = 0$, or equivalently, that the support of $d\hat{\mu}$ is contained in $\{\boldsymbol{\theta} \in \mathbb{T}^d \,|\, Q(e^{i\boldsymbol{\theta}}) = 0\}$. Otherwise letting $d\hat{\mu} = 0$ would result in a larger value of the Lagrangian.

Note that the value of the Lagrangian becomes $-\infty$ for any Φ that vanishes on a set of positive measure, and hence such a Φ cannot be optimal. Now, for any direction $\delta\Phi$ such that $\Phi + \epsilon\delta\Phi$ is a nonnegative $L^1(\mathbb{T}^d)$ function for sufficiently small $\epsilon > 0$, consider the directional derivative

$$\delta\mathcal{L}_P(\Phi, d\hat{\mu}, Q; \delta\Phi) = \lim_{\varepsilon\to 0} \frac{1}{\varepsilon} \left(\mathcal{L}_P(\Phi+\delta\Phi, d\hat{\mu}, Q) - \mathcal{L}_P(\Phi, d\hat{\mu}, Q)\right) = \int_{\mathbb{T}^d} \left(\frac{P}{\Phi} - Q\right) \delta\Phi dm.$$

For a stationary point this must be nonpositive for all feasible directions $\delta\Phi$, and in particular this holds for $\delta\Phi = \Phi \operatorname{sign}(P - Q\Phi)$ which by construction is a feasible

direction. For this direction, the constraint becomes $\int_{\mathbb{T}^d} |P - Q\Phi| dm \leq 0$, requiring that $\Phi = P/Q$ a.e., which inserted into the dual function yields

$$\sup_{d\mu \geq 0} \mathcal{L}_P(\Phi, d\hat{\mu}, Q) = \mathbb{J}_P(Q) + \int_{\mathbb{T}^d} P(\log P - 1) dm, \tag{A.3.1}$$

where the last term in (A.3.1) does not depend on Q and

$$\mathbb{J}_P(Q) = \langle c, q \rangle - \int_{\mathbb{T}^d} P \log Q \, dm. \tag{A.3.2}$$

Hence the dual problem is equivalent to minimizing $\mathbb{J}_P$ over $\bar{\mathfrak{P}}_+ \setminus \{0\}$.

Lower semicontinuity of the dual functional

For any $Q \in \mathfrak{P}_+$, $\mathbb{J}_P(Q)$ is clearly continuous. However, for $Q \in \partial\mathfrak{P}_+$, $\log Q$ will approach $-\infty$ in the points where $Q(e^{i\boldsymbol{\theta}}) = 0$, and hence we need to consider the behavior of the integral term in (A.3.2). Since P is a fixed nonnegative trigonometric polynomial, it suffices to consider the integral $\int_{\mathbb{T}^d} \log Q \, dm$. However, this integral is known as the (logarithmic) Mahler measure of the Laurent polynomial Q [58], and it is finite for all $Q \in \bar{\mathfrak{P}}_+ \setminus \{0\}$ [76, Lemma 2, p. 223]. This leads to the following lemma, the proof of which is deferred to the appendix.

Lemma A.3.1. *For any $P \in \bar{\mathfrak{P}}_+ \setminus \{0\}$ and $c \in \mathfrak{C}_+$, the functional $\mathbb{J}_P : \bar{\mathfrak{P}}_+ \setminus \{0\} \to \mathbb{R}$ is lower semicontinuous.*

The uniqueness of a solution

From the first directional derivative

$$\delta\mathbb{J}_P(Q; \delta Q) = \langle c, \delta q \rangle - \int_{\mathbb{T}^d} \frac{P}{Q} \delta Q dm$$

of the dual functional (A.3.2), we readily derive the second

$$\delta^2\mathbb{J}_P(Q; \delta Q) = \int_{\mathbb{T}^d} \frac{P}{Q^2} (\delta Q)^2 dm,$$

which is clearly nonnegative for all variations δQ. Therefore, since, in addition, the constraint set $\bar{\mathfrak{P}}_+$ is convex, the dual problem is a convex optimization problem. To see that $\mathbb{J}_P$ is actually strictly convex, note that since P is positive a.e., so is P/Q^2. Therefore, for $\delta^2\mathbb{J}_P(Q; \delta Q)$ to be zero we must have $\delta Q = 0$ a.e., which implies that it is zero everywhere since it is continuous. This implies that if there exists a solution, this solution is unique.

The existence of a solution

If we can show that $\mathbb{J}_P$ has compact sublevel sets, then $\mathbb{J}_P$ must have a minimum since it is lower semicontinuous (Lemma A.3.1).

Lemma A.3.2. *The sublevel sets $\mathbb{J}_P^{-1}(-\infty, r]$ are compact for all $r \in \mathbb{R}$.*

For the proof of Lemma A.3.2 we need the following lemma modifying Proposition 2.1 in [17] to the present setting.

Lemma A.3.3. *For a fixed $c \in \mathfrak{C}_+$, there exists an $\varepsilon > 0$ such that for every $(P, Q) \in (\bar{\mathfrak{P}}_+ \setminus \{0\}) \times (\bar{\mathfrak{P}}_+ \setminus \{0\})$*

$$\mathbb{J}_P(Q) \geq \varepsilon \|Q\|_\infty - \int_{\mathbb{T}^d} P dm \, \log \|Q\|_\infty. \tag{A.3.3}$$

Proof. Since $\langle c, q \rangle$ is a continuous function, it achieves a minimum on the compact set $\{Q \in \bar{\mathfrak{P}}_+ \setminus \{0\} \mid \|q\|_\infty = 1\}$, where $\|q\|_\infty := \max_{\mathbf{k} \in \Lambda} |q_{\mathbf{k}}|$. The minimum value κ_c must be positive since $c \in \mathfrak{C}_+$ and hence $\langle c, q \rangle > 0$ for any $q \in \bar{\mathfrak{P}}_+ \setminus \{0\}$. For any $Q \in \bar{\mathfrak{P}}_+ \setminus \{0\}$ we thus have

$$\langle c, q \rangle = \left\langle c, \frac{q}{\|q\|_\infty} \right\rangle \|q\|_\infty \geq \kappa_c \|q\|_\infty. \tag{A.3.4}$$

By Lemma A.10.1, $\|Q\|_\infty \leq |\Lambda| \|q\|_\infty$, and hence by choosing $\varepsilon \leq \kappa_c / |\Lambda|$ we get

$$\langle c, q \rangle \geq \kappa_c \|q\|_\infty \geq \frac{\kappa_c}{|\Lambda|} \|Q\|_\infty \geq \varepsilon \|Q\|_\infty. \tag{A.3.5}$$

To obtain a bound on the second term in (A.3.2), we observe that

$$\int_{\mathbb{T}^d} P \log Q dm = \int_{\mathbb{T}^d} P \log \left[\frac{Q}{\|Q\|_\infty} \right] dm + \int_{\mathbb{T}^d} P dm \log \|Q\|_\infty \leq \int_{\mathbb{T}^d} P dm \, \log \|Q\|_\infty,$$

since $Q / \|Q\|_\infty \leq 1$. Hence (A.3.3) follows. □

Proof of Lemma A.3.2. For any $r \in \mathbb{R}$, which is large enough for the sublevel set $\{Q \in \bar{\mathfrak{P}}_+ \setminus \{0\} \mid r \geq \mathbb{J}_P(Q)\}$ to be nonempty,

$$r \geq \mathbb{J}_P(Q) \geq \varepsilon \|Q\|_\infty - \int_{\mathbb{T}^d} P dm \, \log \|Q\|_\infty$$

for some $\varepsilon > 0$ (Lemma A.3.3). Comparing linear and logarithmic growth we see that the sublevel set is bounded both from above and from below. Moreover, since $\mathbb{J}_P$ is lower semicontinuous (Lemma A.3.1), the sublevel sets are also closed [75, p. 37]. Therefore they are compact. □

Existence of a singular measure

It remains to show that there exists a measure $d\hat{\mu}$ as prescribed by the theorem, and that $d\mu = (P/\hat{Q})dm + d\hat{\mu}$ is in fact an optimal solution to the primal problem to maximize (A.2.1) subject to the moment constraints (A.1.1). To this end, note that $\mathbb{J}_P$ is a closed, proper, strictly convex function with nonempty interior of the effective domain, where an example of the latter is the point $Q \equiv 1$. Now, let $\hat{q}$ be the unique minimum of $\mathbb{J}_P$, and hence the zero vector is a subgradient of $\mathbb{J}_P$ at $\hat{q}$. By Theorem 25.6 in [74] the set of subgradients of $\mathbb{J}_P$ at $\hat{q}$ can be written as

$$0 \in \partial \mathbb{J}_P(\hat{q}) = \text{closure}(\text{conv}\, S(\hat{q})) + K(\hat{q}), \tag{A.3.6}$$

where $K(\hat{q}) = \{-\hat{c}_K \,|\, \langle \hat{c}_K, q - \hat{q}\rangle \geq 0 \text{ for all } q \in \bar{\mathfrak{P}}_+ \setminus \{0\}\}$ is the normal cone, and $S(\hat{q})$ is the set of limit points of sequences of the form $(\nabla \mathbb{J}_P(q_\ell))_{\ell \in \mathbb{Z}_+}$ for which $q_\ell \in \mathfrak{P}_+$ and such that q_ℓ converges to $\hat{q}$ as $\ell \to \infty$. Let $v = (v_{\mathbf{k}})_{\mathbf{k}\in\Lambda} \in S(\hat{q})$. Then there is a sequence $(q_\ell)_{\ell\in\mathbb{Z}_+} \subset \mathfrak{P}_+$ such that $q_\ell \to \hat{q}$ and $\nabla_{\bar{q}_\ell,\mathbf{k}} \mathbb{J}_P(q_\ell) = c_{\mathbf{k}} - \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})}(P/Q_\ell)dm \to v_{\mathbf{k}}$ for $\mathbf{k} \in \Lambda$.[3] In particular the sequence $\int_{\mathbb{T}^d}(P/Q_\ell)dm$ is bounded, hence there is a subsequence of $(P/Q_\ell)dm$ that converges to a nonnegative measure in weak* [57, p. 128]. Since the corresponding nonnegative polynomials $Q_\ell \to \hat{Q}$ converge uniformly, the weak* limit of the subsequence of $(P/Q_\ell)dm$ is of the form $(P/\hat{Q})dm + d\hat{\mu}_{\text{S}}$, where $\hat{\mu}_{\text{S}}$ is a nonnegative measure with support in $\text{null}(\hat{Q})$. The linear maps $d\mu \mapsto \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} d\mu$ are closed for $\mathbf{k} \in \Lambda$, and consequently

$$S(\hat{q}) \subset \left\{ v \,\middle|\, v_{\mathbf{k}} = c_{\mathbf{k}} - \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} \left(\frac{P}{\hat{Q}} dm + d\hat{\mu}_{\text{S}} \right) \quad \text{for } \mathbf{k} \in \Lambda, \right.$$
$$\left. \text{and where } \hat{\mu}_{\text{S}} \geq 0 \text{ and } \text{supp}(\hat{\mu}_{\text{S}}) \subset \text{null}(\hat{Q}) \right\}, \tag{A.3.7}$$

which is closed and convex. Next, note that $K(\hat{q}) = \{-\hat{c}_K \,|\, \langle \hat{c}_K, \hat{q}\rangle = 0, \hat{c}_K \in \bar{\mathfrak{C}}_+\}$. Inserting this and (A.3.7) into (A.3.6) yields

$$0 = c_{\mathbf{k}} - \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} \frac{P}{\hat{Q}} dm - \underbrace{\left(\int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} d\hat{\mu}_{\text{S}} + \hat{c}_{K,\mathbf{k}} \right)}_{=:\hat{c}_{\mathbf{k}}} \quad \text{for } \mathbf{k} \in \Lambda, \tag{A.3.8}$$

for some $\hat{c} \in \bar{\mathfrak{C}}_+$ with $\langle \hat{c}, \hat{q}\rangle = 0$. Moreover, it is shown in [48] that for any $\hat{c} \in \partial\mathfrak{C}_+$ there exists a discrete nonnegative representation $d\hat{\mu}$ with support in $|\Lambda| - 1$ points that satisfies $\int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} d\hat{\mu} = \hat{c}_{\mathbf{k}}$ for $\mathbf{k} \in \Lambda$. To show that the solution is optimal also for the primal problem we observe that, for all $d\mu = \Phi dm + d\hat{\mu}$,

$$\mathbb{I}_P(\Phi) \leq \mathcal{L}_P(\Phi, d\hat{\mu}, Q) \leq \mathbb{J}_P(Q) + \int_{\mathbb{T}^d} P(\log P - 1) dm.$$

[3] Here ∇_z denotes the Wirtinger derivatives, $\nabla_z = (\partial/\partial x - i\partial/\partial y)/2$ and $\nabla_{\bar{z}} = (\partial/\partial x + i\partial/\partial y)/2$, where $z = x + i\,y$ is a complex variable [69, pp. 66-69].

Since equality holds for the feasible point $d\mu = (P/\hat{Q})dm + d\hat{\mu}$, optimality follows. This completes the proof of Theorem A.2.1.

An alternative proof of the results above on the dual problem (lower semi-continuity, uniqueness of solution, and existence of solution) can be constructed along the lines of [31, Section 5]. In the proof of that paper they use the existence of a coercive spectral density, which in our case follows from the existence of a spectral density in the exponential family [37]. Also compare this with the proofs of Theorems 5.1 and 5.2 in [46], which deals with a more general setting.

Comments and an example

In the one-dimensional case it has already been observed that P need not be confined to the cone $\widetilde{\mathfrak{P}}_+ \setminus \{0\}$ but could be a general nonnegative integrable function with zero locus of measure zero [16, 17]. This fact was implemented in [38] to interpret the functional (A.1.5) as a Kullback-Leibler pseudo-distance between P and Φ and hence with P as a Kullback-Leibler prior. In fact, maximizing (A.1.5) is equivalent to minimizing the Kullback-Leibler divergence

$$\mathbb{D}(P\|\Phi) := \int_{\mathbb{T}} P \log\left(\frac{P}{\Phi}\right) dm,$$

which is nonnegative for functions with the same total mass and equal to zero only when the functions are equal. In our present more general setting, P could be any absolutely integrable, nonnegative function for which the set $\{\boldsymbol{\theta} \in \mathbb{T}^d \mid P(e^{i\boldsymbol{\theta}}) = 0\}$ has measure zero. In this context it is also possible to interpret the functional (A.2.1) as a Kullback-Leibler distance, not only between the two functions P and Φ, but between the two measures $dp := Pdm$ and $d\mu$. Since dp is absolutely continuous with respect to $d\mu$ we obtain (cf. [70, Equation (3.1)])

$$\int_{\mathbb{T}^d} P \log\left(\frac{P}{\Phi}\right) dm = \int_{\mathbb{T}^d} \log\left(\frac{dp}{d\mu}\right) dp$$

where $(dp/d\mu) = P/\Phi$ is the Radon-Nikodym derivative.

Except in the one-dimensional case, the singular part of the measure is in general not unique. To illustrate this fact, we consider the following example in two dimensions, similar to Example 5.4 in [46], where Q has zeros along a line.

Example A.3.4. Given $\Lambda = \{(0,0), (-1,0), (1,0), (0,-1), (0,1), (-1,-1), (1,1), (-1,1), (1,-1)\}$, consider

$$\begin{aligned} P(e^{i\theta_1}, e^{i\theta_2}) &= (1 - \cos\theta_1), \\ \hat{Q}(e^{i\theta_1}, e^{i\theta_2}) &= (1 - \cos\theta_1)(2 - \cos\theta_2). \end{aligned}$$

Let c be the covariances of the spectrum $\Phi = P/\hat{Q}$, i.e., $c_{0,0} = 1/\sqrt{3}$, $c_{1,0} = 0$, $c_{0,1} = -1 + 2/\sqrt{3}$, $c_{1,1} = 0$, and $c_{-1,1} = 0$, the remaining covariances being

uniquely determined by the conjugate symmetry $c_{-\mathbf{k}} = \bar{c}_{\mathbf{k}}$. Moreover, let $\hat{c}$ be given by

$$\hat{c}_{\mathbf{k}} = \int_{\mathbb{T}^2} e^{i(\mathbf{k},\boldsymbol{\theta})} \delta(\theta_1) d\theta_1 \frac{d\theta_2}{2\pi}$$

so that $\hat{c}_{0,0} = 1$, $\hat{c}_{1,0} = 1$, $\hat{c}_{0,1} = 0$, $\hat{c}_{1,1} = 0$, and $\hat{c}_{-1,1} = 0$. Clearly $P, \hat{Q} \in \bar{\mathfrak{P}}_+ \setminus \{0\}$, and thus $c \in \mathfrak{C}_+$ since

$$\langle c, r \rangle = \sum_{\mathbf{k} \in \Lambda} c_{\mathbf{k}} \bar{r}_{\mathbf{k}} = \int_{\mathbb{T}^2} R(e^{i\boldsymbol{\theta}}) \frac{P(e^{i\boldsymbol{\theta}})}{\hat{Q}(e^{i\boldsymbol{\theta}})} dm(\boldsymbol{\theta}) > 0$$

for any $R \in \bar{\mathfrak{P}}_+ \setminus \{0\}$. In the same way,

$$\langle \hat{c}, \hat{q} \rangle = \int_{\mathbb{T}^2} \hat{Q}(e^{i\boldsymbol{\theta}}) \delta(\theta_1) dm(\boldsymbol{\theta}) = \int_{-\pi}^{\pi} (1 - \cos\theta_1) \delta(\theta_1) d\theta_1 \int_{-\pi}^{\pi} (2 - \cos\theta_2) \frac{d\theta_2}{2\pi} = 0,$$

and thus $\hat{c} \in \partial\mathfrak{C}_+$. Hence, $(\hat{Q}, \hat{c})$ is the unique pair prescribed by Theorem A.2.1 for the covariance sequence $c + \hat{c}$ and the numerator polynomial P. However, since $\hat{Q}$ is zero for $\theta_1 = 0$, any measure $d\hat{\mu}$ with support constrained to the line $\theta_1 = 0$ and mass 1 such that $\int_{\mathbb{T}^2} \cos\theta_2 d\hat{\mu} = 0$ is a solution.

A.4 Well-posedness and counter examples

The intuition behind Corollary A.2.3 is that the optimal solution $\hat{Q}$ is repelled from the boundary by the following assumption (Assumption A.4.1) whenever $P \in \mathfrak{P}_+$. Then, since the measure $d\hat{\mu}$ can only have mass in the zeros of Q, we must have $d\hat{\mu} = 0$.

Assumption A.4.1. The cone $\bar{\mathfrak{P}}_+$ has the property

$$\int_{\mathbb{T}^d} \frac{1}{Q} dm(\boldsymbol{\theta}) = \infty \quad \text{for all } Q \in \partial\mathfrak{P}_+.$$

As noted in [17], Assumption A.4.1 always holds in the one-dimensional case ($d = 1$), since the trigonometric functions are Lipschitz continuous. Using results by Georgiou [36, p. 819] it can be shown that this assumption is also valid for $d = 2$. However, Lang and McClellan [49] note that Assumption A.4.1 does not hold in general for dimensions $d \geq 3$. To see this, they consider the polynomial $Q(e^{i\boldsymbol{\theta}}) = \sum_{\ell=1}^{d} (1 - \cos\theta_\ell) \in \partial\mathfrak{P}_+$ and show that $\int_{\mathbb{T}^d} \frac{1}{Q} dx < \infty$ for $d \geq 3$. In fact, we have the following amplification of this fact, the proof of which we defer to the appendix.

Proposition A.4.2. *For $d \geq 3$, Assumption A.4.1 does not hold if the index set Λ contains at least three linearly independent vector-valued indices.*

Observe that a problem of dimension $d \geq 3$ for which Λ contains less than three linearly independent vector-valued indices trivially reduces to a problem in one or two dimensions. Hence in general we identify Assumption A.4.1 with the case $d \leq 2$. Corollary A.2.3 now follows directly from the following lemma.

Lemma A.4.3. *Let $P \in \mathfrak{P}_+$, and suppose that Assumption A.4.1 holds. Then the optimal solution $\hat{Q}$ to the problem to minimize* (A.2.2) *over all $Q \in \bar{\mathfrak{P}}_+$ belongs to $\mathfrak{P}_+$.*

Proof. Let $Q \in \partial\mathfrak{P}_+$ be arbitrary. Then, for any $\rho > 0$, $Q(e^{i\boldsymbol{\theta}}) + \rho > 0$ for all $\boldsymbol{\theta} \in \mathbb{T}^d$. Hence the functional $\mathbb{J}_P$ is also differentiable in $Q + \rho$, and the directional derivative in the direction 1 is

$$\delta\mathbb{J}_P(Q + \rho; 1) = \langle c, 1\rangle - \int_{\mathbb{T}^d} \frac{P}{Q+\rho} dm.$$

Now note that $P/(Q + \rho)$ is nonnegative in all points, that it is pointwise monotone increasing for decreasing values of ρ, and that it converges pointwise in an extended real-valued sense[4] to P/Q. Hence by Lebesgue's monotone convergence theorem [75, p. 21] we have, as $\rho \to 0$,

$$\int_{\mathbb{T}^d} \frac{P}{Q+\rho} dm \;\longrightarrow\; \int_{\mathbb{T}^d} \frac{P}{Q} dm,$$

which, since $P \in \mathfrak{P}_+$, is infinite by Assumption A.4.1. Therefore 1 is a descent direction from the point Q, and hence the optimal solution is not obtained there. Since $Q \in \partial\mathfrak{P}_+$ is arbitrary, this means that the optimal solution is not attained on the boundary, i.e., we have $\hat{Q} \in \mathfrak{P}_+$. □

It turns out that the multidimensional rational covariance extension problem for $d \leq 2$ is in fact well-posed in the sense of Hadamard, i.e., the solution depends smoothly on c and P, which is an important property when it comes to tuning of solutions to design specifications. This follows from the following generalizations to the multidimensional case of Theorems 1.3 and 1.4 in [17], proved in the appendix.

Theorem A.4.4. *Let $f^p : \mathfrak{P}_+ \to \mathfrak{C}_+$ be the map from Q to c, given component-wise by*

$$c_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} \frac{P}{Q} dm$$

for a fixed $P \in \mathfrak{P}_+$. If $d \leq 2$, f^p is a diffeomorphism.

Theorem A.4.5. *Suppose that $d \leq 2$. Let f^p be as in Theorem A.4.4, and let $c \in \mathfrak{C}_+$ be fixed. Then the function $g^c : \mathfrak{P}_+ \to \mathfrak{P}_+$ mapping P to $Q = (f^p)^{-1}(c)$ is a diffeomorphism onto its image $\mathfrak{Q}_+$.*

[4] In this case, the limit may be ∞.

By Corollary A.2.3, the unique solution $\hat{Q}$ of the dual problem belongs to the interior $\mathfrak{P}_+$ for every pair $(c, P) \in \mathfrak{C}_+ \times \mathfrak{P}_+$ if Assumption A.4.1 holds. Note that, while the more general Theorem A.2.1 holds for all $P \in \bar{\mathfrak{P}}_+ \setminus \{0\}$, Corollary A.2.3 is only valid for $P \in \mathfrak{P}_+$. The reason for this is that if $P \in \mathfrak{P}_+$ the directional derivative of $\mathbb{J}_P$ tends to $-\infty$ on the boundary by Assumption A.4.1, so a minimum is not attained there, as we just saw in the proof of Lemma A.4.3. On the other hand, if $P \in \partial\mathfrak{P}_+$, we have $\int_{\mathbb{T}^d}(P/Q)dm < \infty$ for some $Q \in \partial\mathfrak{P}_+$; take, for example, $Q = P$. More generally, the integral may not diverge if the zeros of Q belong to a subset of the zeros of P. In this case, there is no guarantee that the optimal solution is an interior point. The following simple one-dimension example illustrates this.

Example A.4.6. Consider a one-dimensional problem of degree one, i.e., with $\Lambda = \{-1, 0, 1\}$. Fix $c = (1, c_1)$, where $c_1 \in (-1, 0)$ is arbitrary. Clearly the Toeplitz matrix $T(c) = [c_{k-\ell}]_{k,\ell=0}^{n}$ is positive definite, and hence $c \in \mathfrak{C}_+$. We fix $P(e^{i\theta}) = 2 + e^{i\theta} + e^{-i\theta}$, which belongs to $\partial\mathfrak{P}_+$ since $P(e^{i\pi}) = 0$. We want to find a $Q \in \mathfrak{P}_+$ of degree at most one so that $\Phi = P/Q$ matches the covariance sequence c, i.e,

$$c_k = \int_{\mathbb{T}} e^{ik\theta}\frac{P}{Q}dm, \quad k = 0, 1. \tag{A.4.1}$$

Any such Q must have the form $Q(e^{i\theta}) = \lambda(1 - \rho e^{i\theta})(1 - \bar{\rho}e^{-i\theta})$ for some $\lambda > 0$ and $|\rho| < 1$. Now, clearly

$$\Phi(e^{i\theta}) = \lambda^{-1}\frac{2 + e^{i\theta} + e^{-i\theta}}{1 - |\rho|^2}\frac{1 - |\rho|^2}{(1 - \rho e^{i\theta})(1 - \bar{\rho}e^{-i\theta})},$$

where the second factor takes the form

$$\frac{1}{1 - \rho e^{i\theta}} + \frac{1}{1 - \bar{\rho}e^{-i\theta}} - 1 = \cdots + \bar{\rho}^2 e^{-2i\theta} + \bar{\rho}e^{-i\theta} + 1 + \rho e^{i\theta} + \rho^2 e^{2i\theta} + \cdots,$$

which implies that $c_0 = \lambda^{-1}(2 + \rho + \bar{\rho})(1 - |\rho|^2)^{-1}$ and $c_1 = \lambda^{-1}(1 + \rho)^2(1 - |\rho|^2)^{-1}$. Since $c_0 = 1$, we have $c_1 = (1+\rho)^2(2+\rho+\bar{\rho})^{-1}$, which has positive, real denominator. Then, since $c_1 < 0$, $1 + \rho$ is purely imaginary, which is impossible since $1 + \rho$ has a positive real part. Hence, there is no $Q \in \mathfrak{P}_+$ of degree at most one satisfying (A.4.1). However, for a certain $Q \in \partial\mathfrak{P}_+$, namely, $Q(e^{i\theta}) = (2 + e^{i\theta} + e^{-i\theta})/(1 + c_1)$, we obtain $d\mu = (P/Q)dm - c_1\delta(\theta - \pi)d\theta$, i.e.,

$$d\mu = (1 + c_1)dm - c_1\delta(\theta - \pi)d\theta,$$

which matches c with $-1 < c_1 < 0$. Now $\Phi = 1 + c_1$ and the singular measure $d\hat{\mu} = \delta(\theta - \pi)d\theta$ has all its mass at the zero of Q, as required by Theorem A.2.1.

In this context it is interesting to note that the covariance extension problem is usually formulated as a partial realization problem where one wants to determine an extension of the partial covariance sequence c so that

$$\Phi_+(z) = \frac{1}{2}c_0 + \sum_{k=1}^{\infty} c_k z^{-k}$$

is positive real, i.e., Φ_+ maps the unit disc to the right half of the complex plane; see, e.g., [55]. Then $\Phi_+(e^{i\theta}) + \Phi_+(e^{i\theta})^*$ is the corresponding spectral density $\Phi(e^{i\theta})$. In our example such a solution is provided by

$$\Phi_+(z) = \frac{1}{2}\left(1 + c_1 - c_1\frac{1-z}{1+z}\right) = \frac{1}{2} + c_1 z - c_1 z^2 + \cdots,$$

yielding precisely $\Phi = 1 + c_1$. The singular measure never appears in this framework.

A.5 Logarithmic moments and cepstral matching

Given $c \in \mathfrak{C}_+$, Corollary A.2.3 and Theorem A.4.5 together provide a complete smooth parameterization in terms of $P \in \mathfrak{P}_+$ of all $\Phi = P/Q$ such that $d\mu = \Phi dm$ satisfies the moment equations (A.1.1). Therefore the solution can be tuned to satisfy additional design specification by adjusting P. How to determine the best P is, however, a separate problem. Theorem A.2.4, to be proved next, extends results from the one-dimensional case to simultaneously estimate P using the cepstral coefficients and logarithmic moment matching.

Proof of Theorem A.2.4. The proof follows along the same lines as that of Theorem A.2.1. By relaxing the primal problem (P) we get the Lagrangian

$$\begin{aligned}\mathcal{L}(\Phi, P, Q) &= \int_{\mathbb{T}^d} \log\Phi\, dm + \sum_{\mathbf{k}\in\Lambda} \bar{q}_{\mathbf{k}}\left(c_{\mathbf{k}} - \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})}\Phi\, dm\right) \qquad \text{(A.5.1)}\\ &+ \sum_{\mathbf{k}\in\Lambda\setminus\{0\}} \bar{p}_{\mathbf{k}}\left(\int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})}\log\Phi\, dm - \gamma_{\mathbf{k}}\right),\end{aligned}$$

where $\bar{q}_k$ and $\bar{p}_k$ are Lagrangian multipliers. Setting $p_0 = \gamma_0 = 1$ and rearranging terms, this can be written as

$$\mathcal{L}(\Phi, P, Q) = \langle c, q\rangle - \int_{\mathbb{T}^d} Q\Phi\, dm - \langle\gamma, p\rangle + 1 + \int_{\mathbb{T}^d} P\log\Phi\, dm, \qquad \text{(A.5.2)}$$

where the first term in (A.5.1) has been incorporated in the last term of (A.5.2). As before, $\sup_{\Phi\geq 0}\mathcal{L}(\Phi, P, Q)$ is only finite if we restrict Q to $\bar{\mathfrak{P}}_+$, and similarly we need to restrict P to $\bar{\mathfrak{P}}_{+,\circ}$. Taking the directional derivative of (A.5.2) in any direction $\delta\Phi$ such that $\Phi + \epsilon\delta\Phi$ is a nonnegative $L^1(\mathbb{T}^d)$ function for all $\varepsilon \in (0, a)$ for a sufficiently small $a > 0$, we obtain

$$\delta\mathcal{L}(\Phi, P, Q; \delta\Phi) = \int_{\mathbb{T}^d} (P\frac{1}{\Phi} - Q)\delta\Phi dm.$$

For the directional derivative to be nonpositive for all feasible directions $\delta\Phi$ we need $\Phi = P/Q$ a.e. (cf. Section A.3), which inserted into (A.5.2) yields

$$\sup_{\Phi}\mathcal{L}(\Phi, P, Q) = \mathbb{J}(P, Q) + 1 - \int_{\mathbb{T}^d} P dm, \qquad \text{(A.5.3)}$$

with $\mathbb{J}(P,Q)$ given by (A.2.4). A closer look at the last term in (A.5.3) shows that

$$\int_{\mathbb{T}^d} P dm = \int_{\mathbb{T}^d} \sum_{\mathbf{k}\in\Lambda} p_{\mathbf{k}} e^{i(\mathbf{k},\boldsymbol{\theta})} dm = \sum_{\mathbf{k}\in\Lambda} p_{\mathbf{k}} \prod_{j=1}^{d} \int_{-\pi}^{\pi} e^{ik_j\theta_j} \frac{d\theta_j}{2\pi} = 1,$$

since all integrals vanish except those for $k_1 = \ldots = k_d = 0$. Consequently, $\mathbb{J}$ is precisely the dual functional (A.5.3).

Using the Wirtinger derivative to form the gradient of $\mathbb{J}$ (see, e.g., [69, pp. 66-69]), we obtain

$$\frac{\partial \mathbb{J}(P,Q)}{\partial \bar{q}_{\mathbf{k}}} = c_{\mathbf{k}} - \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} \frac{P}{Q}\, dm, \quad \mathbf{k}\in\Lambda, \tag{A.5.4a}$$

$$\frac{\partial \mathbb{J}(P,Q)}{\partial \bar{p}_{\mathbf{k}}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} \log\left(\frac{P}{Q}\right) dm - \gamma_{\mathbf{k}}, \quad \mathbf{k}\in\Lambda\setminus\{0\}. \tag{A.5.4b}$$

In deriving (A.5.4b) we used the fact that

$$\int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} dm = \prod_{j=1}^{d} \int_{-\pi}^{\pi} e^{ik_j\theta_j} \frac{d\theta_j}{2\pi} = 0, \quad \mathbf{k}\neq 0. \tag{A.5.5}$$

Therefore, if $\hat{P}\in\mathfrak{P}_{+,\circ}$ and $\hat{Q}\in\mathfrak{P}_+$, and hence the optimal solution is a stationary point of $\mathbb{J}$, then the spectrum $\Phi = \hat{P}/\hat{Q}$ fulfills both covariance matching (A.1.1) and cepstral matching (A.2.3).

The following three lemmas ensure the existence of a solution and show that the problem is in fact convex. The arguments are similar to those in the proof of Theorem A.2.1, and are given in the appendix.

Lemma A.5.1. *Given $c\in\mathfrak{C}_+$ and a sequence $\gamma = \{\gamma_{\mathbf{k}}\}_{\mathbf{k}\in\Lambda}$ with $\gamma_0 = 1$ and $\gamma_{-\mathbf{k}} = \bar{\gamma}_{\mathbf{k}}$, the functional $(P,Q)\mapsto\mathbb{J}(P,Q)$ is lower semicontinuous on $\bar{\mathfrak{P}}_{+,\circ}\times(\bar{\mathfrak{P}}_+\setminus\{0\})$.*

Lemma A.5.2. *The sublevel sets $\mathbb{J}^{-1}(-\infty,r]$ are compact.*

Lemma A.5.3. *The dual problem (D) in Theorem A.2.4 is convex on the domain $\bar{\mathfrak{P}}_{+,\circ}^{(n_1,\ldots,n_d)}\times\bar{\mathfrak{P}}_{+}^{(n_1,\ldots,n_d)}$.*

Next we show that if $\hat{Q}\in\mathfrak{P}_+$ and $\hat{P}\in\mathfrak{P}_{+,\circ}$ then $\hat{\Phi} = \hat{P}/\hat{Q}$ is also optimal for the primal problem of Theorem A.2.4. This follows by observing that $\hat{\Phi}$ is a primal feasible point and that the primal functional (A.2.5) takes the same value as the Lagrangian (A.5.1) in this point, since we have covariance and cepstral matching (cf. the proof of Theorem A.2.1). Finally, if $d\leq 2$ then $\hat{Q}\in\mathfrak{P}_+$ whenever $\hat{P}\in\mathfrak{P}_{+,\circ}$, which follows directly from Lemma A.4.3. This concludes the proof of Theorem A.2.4. □

From this proof we see that the stationarity of $\mathbb{J}(P,Q)$ in Q ensures covariance matching and the stationarity in P provides cepstral matching. Therefore we can only guarantee matching for a solution in the interior $\mathfrak{P}_{+,\circ} \times \mathfrak{P}_+$. This subtle fact was overlooked in [10, 27], where it is claimed that we also have covariance matching for $\hat{P} \in \partial\mathfrak{P}_{+,\circ}$. However, even when $d \leq 2$, we cannot guarantee that there is a solution $\hat{Q}$ belonging to the interior $\mathfrak{P}_+$ if $\hat{P} \in \partial\mathfrak{P}_{+,\circ}$. The following example illustrates this.

Example A.5.4. Consider the one-dimensional problem with $c_0 = 2$, $c_{-1} = c_1 = 1$, and $\gamma_1 = -1$. Set

$$P(e^{i\theta}) = 1 - (e^{i\theta} + e^{-i\theta})/2 = 1 - \cos\theta,$$

and $Q = P$. Clearly P and Q belong to the boundary, since $P(e^{i0}) = Q(e^{i0}) = 0$. Moreover $\Phi = P/Q = 1$, so there is neither covariance matching nor cepstral matching. A simple calculation shows that $\partial\mathbb{J}/\partial q_0 = \partial\mathbb{J}/\partial q_1 = \partial\mathbb{J}/\partial p_1 = 1$. However, for any feasible direction $(\delta q_0, \delta q_1, \delta p_1)$ in (P,Q) we have $\mathrm{Re}\{\delta p_1\} \geq 0$ and $\mathrm{Re}\{\delta q_0 + 2\delta q_1\} \geq 0$, and hence there is no feasible descent direction from this point. Therefore we have a local minimum, which, by convexity, is also a global minimum. Consequently, we have an optimal solution on the boundary where we have neither covariance nor cepstral matching.

Remark A.5.5. From Theorem A.2.1 we know that it is possible to achieve covariance matching in this example by adding a nonnegative singular measure $d\hat{\mu}$, representing spectral lines. In fact, a similar statement can be proved for cepstral matching, namely that there exists a nonpositive measure $d\tilde{\mu}$ such that $\mathrm{supp}(d\tilde{\mu}) \subseteq \{\boldsymbol{\theta} \in \mathbb{T}^d \mid \hat{P}(\boldsymbol{\theta}) = 0\}$ and

$$\gamma_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} \left(\log(\hat{P}/\hat{Q}) dm(\boldsymbol{\theta}) - d\tilde{\mu}(\boldsymbol{\theta})\right)$$

for all $\mathbf{k} \in \Lambda \setminus \{0\}$. However, while the physical interpretation of $d\hat{\mu}$ in Theorem A.2.1 is clear, in this case it is not obvious what $d\tilde{\mu}$ represents in terms of the spectrum.

Note that the optimization problem is convex but in general not strictly convex, and hence the solution might not be unique. This is illustrated in the following example (cf. [55, Remark 12.5.7, and p. 506]).

Example A.5.6. Again consider a one-dimensional problem, this time with $c_0 = 1$, $c_{-1} = c_1 = 0$, and $\gamma_1 = 0$. Choosing

$$P(e^{i\theta}) = Q(e^{i\theta}) = 1 - \rho\cos\theta, \quad |\rho| \leq 1,$$

we obtain $\Phi = 1$, which matches the given covariances and cepstral coefficients. Therefore all P and Q of this form are stationary points of $\mathbb{J}$ and are thus optimal for the dual problem in Theorem A.2.4.

In one dimension there is strict convexity, and thus a unique solution, if and only if there is an optimal solution for which $\hat{P}$ and $\hat{Q}$ are coprime [10].

Regularizing the problem

A motivation for simultaneous covariance and cepstral matching is to obtain a rational spectrum $\Phi = P/Q$ that matches the covariances without having to provide a prior P. However, even if $d \leq 2$, the dual problem in Theorem A.2.4 cannot be guaranteed to produce such a spectrum that satisfies the covariance constraints (A.1.1). To remedy this we consider the regularization proposed by Enqvist [27], which has the objective function

$$\mathbb{J}_\lambda(P, Q) = \mathbb{J}(P, Q) - \lambda \int_{\mathbb{T}^d} \log P \, dm,$$

where $\lambda \in (0, \infty)$ is the regularization parameter.

The partial derivative with respect to $\bar{q}_\mathbf{k}$ is identical to (A.5.4a), whereas the partial derivative with respect to $\bar{p}_\mathbf{k}$ becomes

$$\frac{\partial \mathbb{J}_\lambda(P, Q)}{\partial \bar{p}_\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k}, \boldsymbol{\theta})} \left(\log\left(\frac{P}{Q}\right) - \frac{\lambda}{P} \right) dm - \gamma_\mathbf{k}.$$

By Assumption A.4.1, this gradient will be infinite for $P \in \partial\mathfrak{P}_+$, and hence the optimal solution is not on the boundary. Moreover, with this regularization, the optimization problem becomes strictly convex and we thus have a unique solution.

Theorem A.5.7. *Suppose that $d \leq 2$, and let $\gamma_\mathbf{k}$, $\mathbf{k} \in \Lambda \setminus \{0\}$, be any sequence of complex numbers such that $\gamma_{-\mathbf{k}} = \bar{\gamma}_\mathbf{k}$. Set $\gamma = \{\gamma_\mathbf{k}\}_{\mathbf{k}\in\Lambda}$, where $\gamma_0 = 1$, and let $c \in \mathfrak{C}_+$. Then for any $\lambda > 0$ there exists a unique solution $(\hat{P}, \hat{Q})$ to the strictly convex optimization problem to minimize*

$$\mathbb{J}_\lambda(P, Q) = \langle c, q \rangle - \langle \gamma, p \rangle + \int_{\mathbb{T}^d} P \log\left(\frac{P}{Q}\right) dm - \lambda \int_{\mathbb{T}^d} \log P \, dm$$

subject to $P \in \mathfrak{P}_{+,\circ}$ and $Q \in \mathfrak{P}_+$. Moreover, $\Phi = \hat{P}/\hat{Q}$ fulfills the covariance matching (A.1.1) *and approximately fulfills the cepstral matching* (A.2.3) *via*

$$\gamma_\mathbf{k} + \varepsilon_\mathbf{k} = \int_{\mathbb{T}^d} e^{i(\mathbf{k}, \boldsymbol{\theta})} \log \Phi \, dm, \quad \text{where } \varepsilon_\mathbf{k} = \lambda \int_{\mathbb{T}^d} e^{i(\mathbf{k}, \boldsymbol{\theta})} \hat{P}^{-1} dm.$$

Proof. In view of what has been said, all of the results follow from Theorem A.2.4 except the strict convexity. To prove this, we note that the second directional derivative of $\mathbb{J}_\lambda$ is given by

$$\delta^2 \mathbb{J}_\lambda(P, Q; \delta P, \delta Q) = \int_{\mathbb{T}^d} P \left(\delta P \frac{1}{P} - \delta Q \frac{1}{Q} \right)^2 dm + \int_{\mathbb{T}^d} \delta P^2 \frac{\lambda}{P^2} dm$$

(cf. the proof of Lemma A.5.3 in the appendix). Since both integrands are nonnegative, they both need to be zero almost everywhere in order for the derivative to vanish. However, since $P > 0$, this implies that $\delta P \equiv 0$ by continuity. Then the first integrand becomes $\delta Q^2 P/Q^2$ and in the same way we must thus have $\delta Q \equiv 0$. Hence $\delta^2 \mathbb{J}_\lambda(P, Q; \delta P, \delta Q) > 0$, implying uniqueness. □

A.6 The circulant problem

Theorem A.2.5 in Section A.2 can be viewed as a periodic version of Theorem A.2.1 and Corollary A.2.3, as can be seen by following the lines of [54], where the one-dimensional problem was first introduced. To this end, we introduce the discrete measure $d\nu_{\mathbf{N}}$, i.e.,

$$d\nu_{\mathbf{N}}(\boldsymbol{\theta}) = \sum_{\boldsymbol{\ell}\in\mathbb{Z}^d_{\mathbf{N}}} \delta\big(\theta_1 - \phi_1(\ell_1), \dots, \theta_d - \phi_d(\ell_d)\big) \prod_{j=1}^{d} \frac{d\theta_j}{N_j}, \tag{A.6.1}$$

where $\phi_j(\ell) := 2\pi\ell/N_j$ and δ is the multidimensional Dirac delta function. Then the moment matching condition (A.2.7) takes the form

$$c_{\mathbf{k}} = \frac{1}{\prod_{j=1}^{d} N_j} \sum_{\boldsymbol{\ell}\in\mathbb{Z}^d_{\mathbf{N}}} \zeta_{\boldsymbol{\ell}}^{\mathbf{k}} \Phi(\boldsymbol{\zeta}_{\boldsymbol{\ell}}) = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} \Phi(e^{i\boldsymbol{\theta}}) d\nu_{\mathbf{N}},$$

which is similar to (A.1.1), but where $d\nu_{\mathbf{N}}$ and dm have different mass distributions (discrete versus continuous). In fact, the main difference in the statements of Theorem A.2.5 and Theorem A.2.1 together with Corollary A.2.3 is that different measures and cones are used. In the same way, versions of Theorems A.2.4 and A.5.7 also hold in the circulant case; see [72] for details.

In connection to this it is also interesting to observe that the discrete counterpart of Assumption A.4.1,

$$\int_{\mathbb{T}^d} \frac{1}{Q} d\nu_{\mathbf{N}} = \infty \quad \text{for all } Q \in \partial\mathfrak{P}_+(\mathbf{N}), \tag{A.6.2}$$

holds for any measure $d\nu_{\mathbf{N}}$ with discrete mass distribution (see also [49]). However, if $P \in \partial\mathfrak{P}_+(\mathbf{N})$ we may still obtain solutions without covariance matching, because for any Q that is zero only in a subset of points where P is zero we will have $\int_{\mathbb{T}^d}(P/Q)d\nu_{\mathbf{N}} < \infty$ and hence the optimal solution may occur on the boundary.

Remark A.6.1. Although the measure (A.6.1) has mass in points placed in the roots of unity on the d-dimensional torus, one could also consider other mass distributions. One could place the mass points in the odd points of the roots of unity, i.e., in the points $\{e^{i(2k_j-1)\pi/N_\ell}\}_{k_j=1}^{N_\ell}$, a situation which has been studied in the one-dimensional case and which correspond to spectra of skew-periodic processes [73]. The same holds in the multidimensional setting. Also note that all dimensions do not need to have mass distributions of the same type. For example, the approach in this paper works even if the process is periodic in some of the dimensions, while nonperiodic in others.

Convergence of discrete to continuous

In [54] Lindquist and Picci proved for the one-dimensional case that when the number of mass points in the discrete measure $d\nu_{\mathbf{N}}$ in (A.6.1) goes to infinity, the

solution converges to the solution of the problem with the continuous measure dm. The same is true in higher dimensions, and the formal result is given in Theorem A.2.6 in Section A.2. In this subsection we will prove this statement. Note that we use the notation

$$\mathbb{J}_P(Q) = \langle c, q\rangle - \int_{\mathbb{T}^d} P \log Q \, dm, \tag{A.6.3a}$$

$$\mathbb{J}_P^{\mathbf{N}}(Q) = \langle c, q\rangle - \int_{\mathbb{T}^d} P \log Q \, d\nu_{\mathbf{N}}, \tag{A.6.3b}$$

to explicitly distinguish the objective functions using the continuous and the discrete measure. Moreover let $\hat{Q}$ be the minimizer of (A.6.3a), subject to $Q \in \bar{\mathfrak{P}}_+$, and $\hat{Q}_{\mathbf{N}}$ be a minimizer of (A.6.3b), subject to $Q \in \bar{\mathfrak{P}}_+(\mathbf{N})$. Before proving the theorem, we make some clarifying observations.

Remark A.6.2. We have already noted that the singular measure $d\hat{\mu}$ is not unique. However, the corresponding "rest covariance" $\hat{c}$, which $d\hat{\mu}$ matches, is unique (cf. (A.3.8)). In connection with this it is interesting to note that although this is the case, and although $\hat{Q}_{\mathbf{N}} \to \hat{Q}$, in general $\hat{c}_{\mathbf{N}} \not\to \hat{c}$. To see this, note that for a P which is positive in all points except for some irrational frequency[5] where $P = 0$, we will have $P \in \mathfrak{P}_+(\mathbf{N})$ for all $\mathbf{N}$, since this point will never belong to the grid. Thus we will have $\hat{Q}_{\mathbf{N}} \in \mathfrak{P}_+(\mathbf{N})$ and therefore $\hat{c}_N = 0$. However $P \in \partial\mathfrak{P}_+$, and therefore we can have $\hat{Q} \in \partial\mathfrak{P}_+$ and hence $\hat{c} \neq 0$. One can construct such an example based on Example A.4.6 by shifting the spectral line to an irrational frequency point.

Remark A.6.3. In connection to the previous remark, we note that in two dimensions we have $\hat{Q} \in \mathfrak{P}_+$ whenever $P \notin \partial\mathfrak{P}_+$, since Assumption A.4.1 is valid for $d = 2$. Hence there will be no singular measure. Moreover, since $\hat{Q}_{\mathbf{N}} \to \hat{Q}$ as $\min(\mathbf{N})$ goes to infinity, for large enough value of $\min(\mathbf{N})$ we must have $\hat{Q}_{\mathbf{N}} > 0$, i.e., $\hat{Q}_{\mathbf{N}} \in \mathfrak{P}_+$. Therefore $(P/\hat{Q}_{\mathbf{N}})d\nu_{\mathbf{N}}$ tends to $(P/\hat{Q})dm$ in weak*.

The first thing we need to show is that $\hat{Q}_{\mathbf{N}}$ is in fact well-defined. That this is not evident from the statement of the theorem becomes apparent when noting the following relationship among the cones of trigonometric polynomials:

$$\bar{\mathfrak{P}}_+(\mathbf{N}) \supset \bar{\mathfrak{P}}_+(2\mathbf{N}) \supset \ldots \supset \bar{\mathfrak{P}}_+.$$

For the dual cones we therefore have [57, pp. 157-158]

$$\bar{\mathfrak{C}}_+(\mathbf{N}) \subset \bar{\mathfrak{C}}_+(2\mathbf{N}) \subset \ldots \subset \bar{\mathfrak{C}}_+,$$

and thus it is not guaranteed that minimizing (A.6.3b) over $Q \in \bar{\mathfrak{P}}_+(\mathbf{N})$ has a solution for $c \in \mathfrak{C}_+$. However note that when $N_l \to \infty$ the corresponding set $\{e^{ik_l 2\pi/N_l}\}_{k_l \in \mathbb{Z}_{N_l}}$ becomes dense on the unit circle. Therefore $\bar{\mathfrak{P}}_+ = \bigcap_{\mathbf{N} \in \mathbb{Z}_+^d} \bar{\mathfrak{P}}_+(\mathbf{N})$. Using this we have the following lemma, proved in the appendix, which is a generalization to the multivariable case of Proposition 6 in [54].

[5] An irrational frequency is an angle $\lambda\pi$ for which λ is an irrational number.

Lemma A.6.4. *For any $c \in \mathfrak{C}_+$ there exists an N_0 such that $c \in \mathfrak{C}_+(\mathbf{N})$ for all* $\min(\mathbf{N}) \geq N_0$.

This shows that for each $c \in \mathfrak{C}_+$, the problem of minimizing (A.6.3b) over $Q \in \bar{\mathfrak{P}}_+(\mathbf{N})$ does in fact have a solution for large enough values of $\mathbf{N}$. Interestingly, the lemma is equivalent to $\lim_{\min(\mathbf{N})\to\infty} \mathfrak{C}_+(\mathbf{N}) = \mathfrak{C}_+$.

Proof of Theorem A.2.6. Let $\hat{Q}$ and $\hat{Q}_{\mathbf{N}}$ be as in the statement of the theorem. Choose a $c \in \mathfrak{C}_+$ and a $P \in \bar{\mathfrak{P}}_+ \setminus \{0\}$ and fix N_0 in accordance with Lemma A.6.4. Throughout the rest of this proof we only consider $\min(\mathbf{N}) \geq N_0$, which means that an optimal solution $\hat{Q}_{\mathbf{N}}$ exists. Moreover, in the proof we need the following result, which is proved in the appendix.

Lemma A.6.5. *The sequence $(\hat{Q}_{\mathbf{N}})$ is bounded in $L^\infty(\mathbb{T}^d)$.*

Since $(\hat{Q}_{\mathbf{N}})$ is bounded, there is a convergent subsequence, call it $(\hat{Q}_{\mathbf{N}})$ for convenience, converging in the $L^\infty(\mathbb{T}^d)$ norm to some function $\hat{Q}_\infty$. Since $(\hat{Q}_{\mathbf{N}})$ is a set of continuous functions, this means that the convergence is in fact uniform and hence $\hat{Q}_\infty$ is a continuous function. Now since i) the convergence is uniform, ii) $\hat{Q}_\infty$ is continuous, and iii) the grid points become dense on $\mathbb{T}^d$ as $\min(\mathbf{N})$ goes to infinity, we obtain $\hat{Q}_\infty(e^{i\boldsymbol{\theta}}) \geq 0$ for all $\boldsymbol{\theta}$, and hence $\hat{Q}_\infty$ belongs to $\bar{\mathfrak{P}}_+ \setminus \{0\}$.

It remains to show that $\hat{Q}_\infty = \hat{Q}$. This will be done by proving that $\|\hat{Q}_\infty - \hat{Q}\|_\infty \leq \varepsilon$ for all $\varepsilon > 0$. To do this, fix a $\tilde{Q} \in \mathfrak{P}_+$ and consider $\hat{Q} + \eta\tilde{Q}$, which belongs to $\mathfrak{P}_+$ for all $\eta > 0$. By simply adding and subtracting $\eta\tilde{Q}$, the triangle inequality gives

$$\|\hat{Q}_\infty - \hat{Q}\|_\infty \leq \eta\|\tilde{Q}\|_\infty + \|(\hat{Q}_\infty + \eta\tilde{Q}) - \hat{Q}\|_\infty. \tag{A.6.4}$$

We want to bound the second term. To this end, note that

$$\mathbb{J}_P(\hat{Q} + \eta\tilde{Q}) - \mathbb{J}_P(\hat{Q}) = \langle c, \eta\tilde{q}\rangle - \int_{\mathbb{T}^d} P \log\left(1 + \frac{\eta\tilde{Q}}{\hat{Q}}\right) dm,$$

and, since the integral is nonnegative, we obtain

$$\mathbb{J}_P(\hat{Q} + \eta\tilde{Q}) \leq \mathbb{J}_P(\hat{Q}) + \eta\langle c, \tilde{q}\rangle. \tag{A.6.5}$$

The same holds for $\mathbb{J}_P^{\mathbf{N}}$, i.e., $\mathbb{J}_P^{\mathbf{N}}(\hat{Q}_{\mathbf{N}} + \eta\tilde{Q}) \leq \mathbb{J}_P^{\mathbf{N}}(\hat{Q}_{\mathbf{N}}) + \eta\langle c, \tilde{q}\rangle$. By optimality we also have $\mathbb{J}_P^{\mathbf{N}}(\hat{Q}_{\mathbf{N}}) \leq \mathbb{J}_P^{\mathbf{N}}(\hat{Q} + \eta\tilde{Q}) < \infty$ for all $\eta > 0$, and hence

$$\mathbb{J}_P^{\mathbf{N}}(\hat{Q}_{\mathbf{N}} + \eta\tilde{Q}) \leq \mathbb{J}_P^{\mathbf{N}}(\hat{Q} + \eta\tilde{Q}) + \eta\langle c, \tilde{q}\rangle. \tag{A.6.6}$$

Now, since $\hat{Q}_{\mathbf{N}} + \eta\tilde{Q} \to \hat{Q}_\infty + \eta\tilde{Q} \in \mathfrak{P}_+$, we know that, for large enough values of $\min(\mathbf{N})$, we have $\hat{Q}_{\mathbf{N}} + \eta\tilde{Q} \in \mathfrak{P}_+$. Therefore, the left hand side of (A.6.6) is guaranteed to be well-defined for all values of $\min(\mathbf{N})$ larger than this value. We can thus take the limit on both sides of (A.6.6) to obtain

$$\mathbb{J}_P(\hat{Q}_\infty + \eta\tilde{Q}) \leq \mathbb{J}_P(\hat{Q} + \eta\tilde{Q}) + \eta\langle c, \tilde{q}\rangle,$$

which together with (A.6.5) yields

$$\mathbb{J}_P(\hat{Q}_\infty + \eta\tilde{Q}) \leq \mathbb{J}_P(\hat{Q}) + 2\eta\langle c, \tilde{q}\rangle. \tag{A.6.7}$$

Now consider the sets $D_\delta = \{Q \in \bar{\mathfrak{P}}_+ \mid \mathbb{J}_P(Q) \leq \mathbb{J}_P(\hat{Q}) + \delta\}$. Since the Hessian at the optimal solution is positive definite we have $\bigcap_{\delta>0} D_\delta = \{\hat{Q}\}$. Therefore, it follows from (A.6.7) that $\eta > 0$ can be chosen so that $\|(\hat{Q}_\infty + \eta\tilde{Q}) - \hat{Q}\|_\infty < \tilde{\varepsilon}$ for any $\tilde{\varepsilon} > 0$. Consequently, by selecting η sufficiently small, we may bound (A.6.4) by an arbitrary small positive number. Hence $\hat{Q}_\infty = \hat{Q}$. □

A.7 Application to system identification

The power spectrum of a signal represents the energy distribution across frequencies of the signal. For a multidimensional, discrete-time, zero-mean, and homogeneous[6] stochastic process $\{y(\mathbf{t})\}$, defined for $\mathbf{t} \in \mathbb{Z}^d$, the power spectrum is defined as the nonnegative measure $d\mu$ on $\mathbb{T}^d$ whose Fourier coefficients are the covariances

$$c_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} d\mu.$$

In one dimension the singular part of the measure represents spectral lines, and if the absolutely continuous part is also rational, $\Phi = P/Q$, one can use spectral factorization to determine the filter coefficients for an autoregressive-moving-average (ARMA) model which, when fed with white noise input, reproduces a stochastic signal with the same power distribution as Φ. Therefore the one-dimensional rational covariance extension problem can be used for system identification [55].

With the theory developed in this paper we can estimate rational spectra in higher dimensions. However spectral factorization is not in general possible when $d > 1$ [24]. For $d = 2$, Geronimo and Woerdeman have established conditions for when it is possible to factorize a given trigonometric polynomial as a sum-of-one-square [39, Theorem 1.1.1]. These include a nontrivial rank condition on a reduced matrix of Fourier coefficients, which we shall call Γ_{red}, but also gives an explicit algorithm for obtaining the factors in cases when it is possible. Nevertheless, in the following example we will illustrate how the theory could be used in the case when covariances and cepstral coefficients come from a rational, factorizable spectrum.

We consider a two-dimensional recursive filter with transfer function

$$\frac{b(e^{i\theta_1}, e^{i\theta_2})}{a(e^{i\theta_1}, e^{i\theta_2})} = \frac{\sum_{\mathbf{k}\in\Lambda_+} b_{\mathbf{k}} e^{-i(\mathbf{k},\boldsymbol{\theta})}}{\sum_{\mathbf{k}\in\Lambda_+} a_{\mathbf{k}} e^{-i(\mathbf{k},\boldsymbol{\theta})}},$$

[6] Homogeneity implies that covariances $c_{\mathbf{k}} := E\{y(\mathbf{t}+\mathbf{k})\overline{y(\mathbf{t})}\}$ are invariant with "time" $\mathbf{t} \in \mathbb{Z}^d$. From this it is also easy to see that $c_{-\mathbf{k}} = \bar{c}_{\mathbf{k}}$.

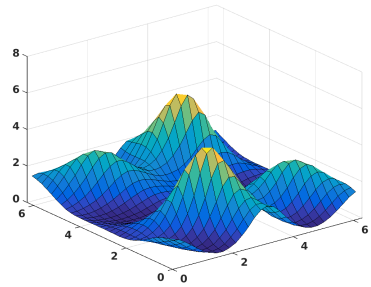


Figure A.1: The true spectrum.

where $\Lambda_+ = \{(k_1, k_2) \in \mathbb{Z}^2 \mid 0 \leq k_1 \leq 2, 0 \leq k_2 \leq 2\}$ and the coefficients are given by $b_{(k_1,k_2)} = B_{k_1+1,k_2+1}$ and $a_{(k_1,k_2)} = A_{k_1+1,k_2+1}$, where

$$B = \begin{bmatrix} 0.9589 & -0.0479 & 0.0959 \\ 0.0959 & 0.0479 & 0.0959 \\ -0.0959 & 0.0479 & 0.1918 \end{bmatrix}, \quad A = \begin{bmatrix} 1.0000 & 0.1000 & 0.0500 \\ -0.1000 & 0.0500 & -0.0500 \\ 0.2000 & -0.0500 & -0.1000 \end{bmatrix}.$$

Then the corresponding spectrum is given by

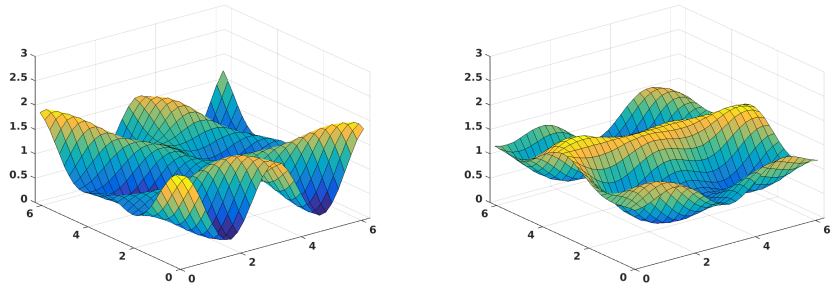
$$\Phi(e^{i\boldsymbol{\theta}}) = \Phi(e^{i\theta_1}, e^{i\theta_2}) = \frac{P(e^{i\theta_1}, e^{i\theta_2})}{Q(e^{i\theta_1}, e^{i\theta_2})} = \left| \frac{b(e^{i\theta_1}, e^{i\theta_2})}{a(e^{i\theta_1}, e^{i\theta_2})} \right|^2,$$

and hence the index set Λ of the coefficients of the trigonometric polynomials P and Q is given by $\Lambda = \{(k_1, k_2) \in \mathbb{Z}^2 \mid |k_1| \leq 2, |k_2| \leq 2\}$.

We approximate the continuous problem with a discrete one in accordance with Theorem A.2.6. The two-dimensional spectrum Φ is evaluated on a grid of size 30×30, and shown in Figure A.1. The trigonometric polynomials corresponding to the true spectrum are shown in Figure A.2. Its covariances and cepstral coefficients are computed, and a spectrum is then estimated by (unregularized) covariance and cepstral matching along the lines of Theorem A.2.4. The problem is solved numerically using CVX, a Matlab package for solving disciplined convex programming problems [41, 40], and the resulting spectrum is shown in Figure A.3a. The relative error[7] is shown in Figure A.3b. As seen from the relative error, we recover the true spectrum with good accuracy. For the ME solution, the resulting spectrum and relative error are shown in Figure A.4.

For system identification we are now interested in factorizing the two rational spectra as a sum-of-one-square, if possible. To check factorizability for the two solutions, we apply the rank condition from [39, Theorem 1.1.1], which requires

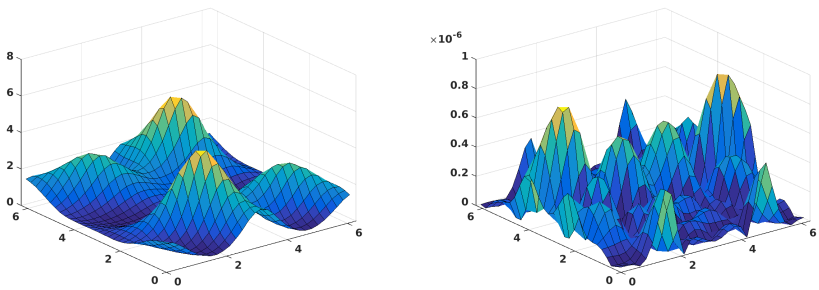
[7] We define the relative error between two functions Φ_{true} and Φ_{est} be the pointwise evaluation of $|\Phi_{\text{true}} - \Phi_{\text{est}}|/\Phi_{\text{true}}$.



(a) The true polynomial P.

(b) The true polynomial Q.

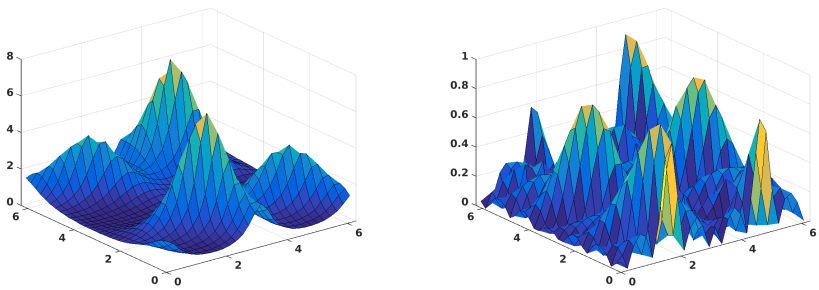
Figure A.2: The spectrum of the system



(a) Estimated spectrum.

(b) Relative error.

Figure A.3: Spectrum estimated with covariance and cepstral matching.



(a) ME-spectrum.

(b) Relative error.

Figure A.4: The ME-estimation and relative error to true spectrum.

that the corresponding submatrix $\Gamma_{\text{red}} \in \mathbb{C}^{6\times 6}$ should be of rank four in both cases. However, such a matrix is generically full rank and we have to study the singular values in order to determine the numerical rank.

To illustrate this issue, in Figure A.5 we plot the singular values of Γ_{red} for the respective polynomials. Figure A.5b shows the singular values corresponding to the solution $Q_{\text{true P}}$ computed with the true polynomial P as prior (cf. Theorem A.2.1 and Section A.3). This solution, as well as the solution obtained by covariance and cepstral matching, gives the exact spectrum back, up to numerical errors, and hence should be factorizable. For both these solutions we can also observe a significant decrease in size between the fourth and the fifth singular values in Figure A.5b. This indicates that the matrices in fact have numerical rank four, and spectral factorization is thus possible. Performing the spectral factorization on the solution with covariance and cepstral matching gives polynomials with coefficients

$$B_{\text{est}} = \begin{bmatrix} 0.9589 & -0.0479 & 0.0959 \\ 0.0959 & 0.0479 & 0.0959 \\ -0.0959 & 0.0479 & 0.1918 \end{bmatrix}, \quad A_{\text{est}} = \begin{bmatrix} 1.0000 & 0.1000 & 0.0500 \\ -0.1000 & 0.0500 & -0.0500 \\ 0.2000 & -0.0500 & -0.1000 \end{bmatrix},$$

which agree completely with the true coefficients.

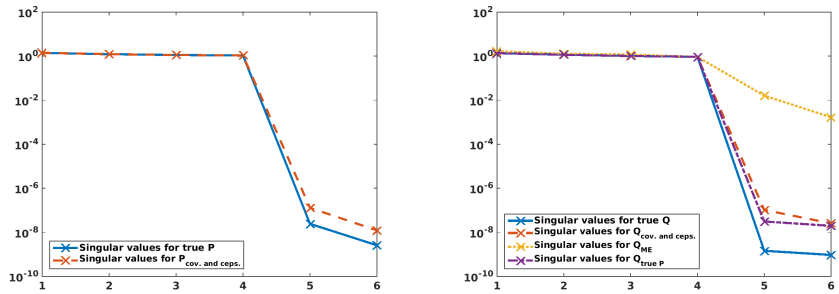
For the ME spectrum on the other hand there is no guarantee that it will be factorizable. In general there is *a priori* no reason why spectral factorization should be possible. However, in Figure A.5b we observe a decrease in size between the fourth and the fifth singular values also for the ME solution $\Phi_{\text{ME}} = 1/Q_{\text{ME}}$, although this decrease is significantly smaller than for the other polynomials. If for the moment we assume that the rank condition on Γ_{red} is actually (approximately) satisfied and apply the factorization algorithm of [39], we obtain the coefficients

$$A_{\text{ME}} = \begin{bmatrix} 1.0317 & 0.1423 & -0.0251 \\ -0.1881 & -0.0173 & -0.1252 \\ 0.2872 & -0.0570 & -0.2597 \end{bmatrix}$$

for the possible spectral factor a_{ME} of Q_{ME}. Forming the corresponding true Q, namely, $|a_{\text{ME}}|^2$, and comparing it with Q_{ME}, we obtain a relative error of up to 10% with respect to Q_{ME}. We leave the question whether this is a reasonable approximation to a future study. Note also that if the ME spectrum is factorizable, the factors are given directly from the covariances by the Geronimo and Woerdeman algorithm. However if this is not the case, rational covariance extension will still give a rational spectrum. An important open question related to this, and suggested by the above analysis, is whether the solution can be tuned by an appropriate choice of P so that the rank condition is satisfied, and hence factorization is possible.

A.8 Application to texture generation

Wiener systems form a class of nonlinear dynamical systems that consist of a linear dynamic part composed with a static nonlinearity, as in Figure A.6. They belong



(a) Singular values of Γ_{red} for different P. (b) Singular values of Γ_{red} for different Q.

Figure A.5: The singular values of the reduced covariance matrix.

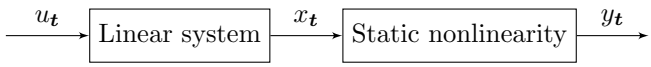


Figure A.6: A Wiener system with thresholding as static nonlinearity.

to a class of so called block-oriented systems, which has a long history [4], and applications are found in many areas of science and engineering [3]. A lot of research has been done in the area of identification of Wiener systems, see, e.g, [42] and references therein, and the area is still active [56, 80, 1].

In this example we shall use Wiener systems to model and generate textures. The idea of using dynamical systems for modeling of textures and images is not new and has been considered in, e.g., [21, 65]. The setup we present here is motivated by [29], where thresholded Gaussian random fields are used to model porous materials for design of surface structures in pharmaceutical film coatings.

To this end, we let $\{x_{\boldsymbol{t}};\, \boldsymbol{t} \in \mathbb{Z}^d\}$ be the stationary output of a linear system with Gaussian white noise input $\{u_{\boldsymbol{t}};\, \boldsymbol{t} \in \mathbb{Z}^d\}$, and let $y_{\boldsymbol{t}} = f(x_{\boldsymbol{t}})$ where f is the static nonlinearity

$$f(x) = \begin{cases} 1 & x > \tau \\ 0 & \text{otherwise,} \end{cases} \tag{A.8.1}$$

where the thresholding parameter τ is assumed to be unknown. We assume that $u_{\boldsymbol{t}}$ is a zero-mean process, and hence $x_{\boldsymbol{t}}$ is also a zero-mean Gaussian process, which we assume to be normalized $c_{\mathbf{0}} := E[x_{\boldsymbol{t}}^2] = 1$. Due to these assumptions, the output $y_{\boldsymbol{t}}$ of the static nonlinearity has mean

$$E[y_{\boldsymbol{t}}] = P(y_{\boldsymbol{t}} = 1) = 1 - P(x_{\boldsymbol{t}} \leq \tau) = 1 - \phi(\tau), \tag{A.8.2}$$

where $\phi(\tau)$ is the Gaussian cumulative distribution function

$$\phi(\tau) = \int_{-\infty}^{\tau} \frac{1}{\sqrt{2\pi}} \exp(-s^2/2) ds.$$

Covariances of thresholded Gaussian variables

Next we consider the relation between the covariances of the input $x_{\boldsymbol{t}}$ and those of the output $y_{\boldsymbol{t}}$, respectively, and use this to estimate the covariances of the process $x_{\boldsymbol{t}}$ [3, 66].

To this end, let x_1, $x_2 \in N(0,1)$ be two jointly Gaussian stochastic variables and set $y_\ell = f(x_\ell)$, for $\ell = 1, 2$, where $f : \mathbb{R} \to \mathbb{R}$ is a given function. In addition, let ρ and r be the covariances $\rho := E[x_1 x_2]$ and $r := E[y_1 y_2] - E[y_1]E[y_2]$, respectively. We are interested in the relation between ρ and r, and to this end we introduce $R := E[y_1 y_2]$. Now note that R is related to the covariance ρ via [66, Equation (21)] (see also [3, p. 32]), i.e.,

$$\frac{\partial R}{\partial \rho} = \int_{\mathbb{R}^2} \frac{\exp\left(-\frac{x_1^2 + x_2^2 - 2\rho x_1 x_2}{2(1-\rho^2)}\right)}{2\pi\sqrt{1-\rho^2}} f'(x_1) f'(x_2) dx_1 dx_2.$$

In our case $f(x)$ is given by (A.8.1), and thus $f'(x) = \delta_\tau(x)$ is a Dirac delta function at τ. Therefore

$$\frac{\partial R}{\partial \rho} = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{\tau^2}{1+\rho}\right),$$

and from this it follows that

$$R(\rho) = b + \int_0^{\rho} \frac{1}{2\pi\sqrt{1-s^2}} \exp\left(-\frac{\tau^2}{1+s}\right) ds,$$

for some constant b. In order to determine b, first note that $\rho = 0$ implies that x_1 and x_2 are uncorrelated, and hence independent, since the joint distribution is Gaussian. This in turn means that y_1 and y_2 are independent, since f is a static function, and hence we get

$$b = R(0) = E[y_1 y_2] = E[y_1]E[y_2].$$

Therefore r can be expressed as

$$\begin{aligned} r &= R(\rho) - E[y_1]E[y_2] \\ &= \int_0^{\rho} \frac{1}{2\pi\sqrt{1-s^2}} \exp\left(-\frac{\tau^2}{1+s}\right) ds. \end{aligned} \tag{A.8.3}$$

The integrand is well-defined for $-1 < \rho < 1$, and the integral converges for all values in the closed interval $[-1, 1]$. Moreover, the integrand is strictly positive on $(-1, 1)$ and by the inverse function theorem this transformation is invertible.

Estimating the linear part of the Wiener system

By using the inverse of (A.8.3) we can estimate the covariances $c_{\mathbf{k}} := E[x_{\boldsymbol{t}+\mathbf{k}}x_{\boldsymbol{t}}]$ from estimates of the covariances $r_{\mathbf{k}} := E[y_{\boldsymbol{t}+\mathbf{k}}y_{\boldsymbol{t}}] - E[y_{\boldsymbol{t}+\mathbf{k}}]E[y_{\boldsymbol{t}}]$. Note however that (A.8.3) depends on the threshold parameter τ, which is assumed to be unknown. In order to estimate τ we use (A.8.2), which gives $\tau_{\text{est}} = \phi^{-1}(1 - E[y_{\boldsymbol{t}}])$. Having estimates of the covariances $c_{\mathbf{k}}$, we can now appeal to Theorem A.2.1 in order to estimate a rational spectrum for $x_{\boldsymbol{t}}$.

Given this rational spectral density we want to recover a linear dynamical system corresponding to the spectrum. However, as was discussed in Section A.7 this is nontrivial, and we therefore resort to a heuristic and apply the factorization procedure in [39, Theorem 1.1.1] although some of the conditions required to ensure the existence of a spectral factor may not be met (cf. Section A.7)

The complete procedure for identifying the Wiener system with thresholding as static nonlinearity is summarized in Algorithm A.1.

Algorithm A.1

Input: $(y_{\boldsymbol{t}})$

1: Estimate threshold parameter: $\tau_{\text{est}} = \phi^{-1}(1 - E[y_{\boldsymbol{t}}])$
2: Estimate covariances: $r_{\mathbf{k}} := E[y_{\boldsymbol{t}+\mathbf{k}}y_{\boldsymbol{t}}] - E[y_{\boldsymbol{t}+\mathbf{k}}]E[y_{\boldsymbol{t}}]$
3: Compute covariances $c_{\mathbf{k}} := E[x_{\boldsymbol{t}+\mathbf{k}}x_{\boldsymbol{t}}]$ by using (A.8.3)
4: Estimate a rational spectrum using Theorem A.2.1
5: Apply the factorization procedure in [39, Theorem 1.1.1]

Output: τ_{est}, coefficients for the linear dynamical system

Simulation results

Next we test the procedure outlined above on simulated data. To this end, we consider the two-dimensional recursive filter with transfer function given by

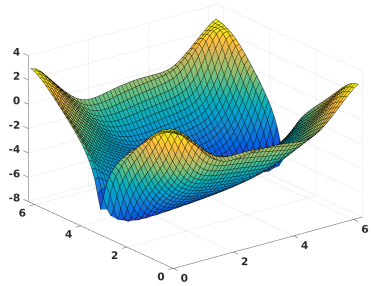
$$\frac{b(e^{i\theta_1}, e^{i\theta_2})}{a(e^{i\theta_1}, e^{i\theta_2})} = \frac{\sum_{\mathbf{k}\in\Lambda_+} b_{\mathbf{k}} e^{-i(\mathbf{k},\boldsymbol{\theta})}}{\sum_{\mathbf{k}\in\Lambda_+} a_{\mathbf{k}} e^{-i(\mathbf{k},\boldsymbol{\theta})}},$$

where $\Lambda_+ = \{(k_1, k_2) \in \mathbb{Z}^2 \mid 0 \le k_1 \le 2,\ 0 \le k_2 \le 2\}$ and the coefficients are given by $b_{(k_1,k_2)} = B_{k_1+1,k_2+1}$ and $a_{(k_1,k_2)} = A_{k_1+1,k_2+1}$, where

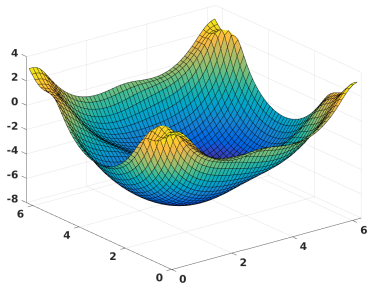
$$B = \begin{bmatrix} 0.75 & -0.2 & 0.05 \\ 0.2 & 0.3 & 0.05 \\ -0.05 & -0.05 & 0.1 \end{bmatrix}, \quad A = \begin{bmatrix} 3.6623 & -4.0222 & 0.9987 \\ -4.0939 & 4.8705 & -1.1913 \\ 1.2018 & -1.3539 & 0.2155 \end{bmatrix}.$$

The threshold parameter in (A.8.1) is set to $\tau = 0.06$.

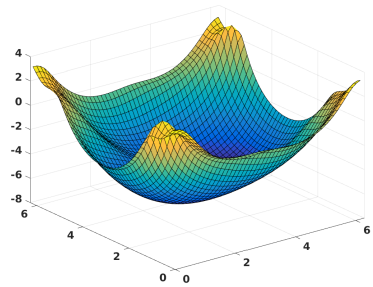
The system is simulated with Gaussian white noise as input, and 500×500 samples are taken as output. These samples are used to estimate the threshold



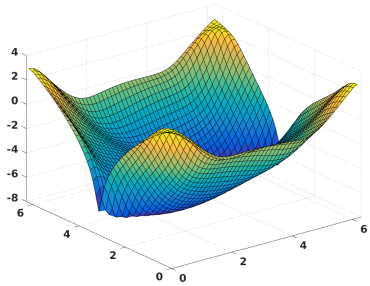
(a) True spectrum.



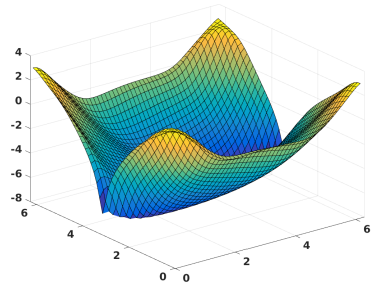
(b) Identified ME spectrum.



(c) Identified and factorized ME spectrum.



(d) Identified spectrum with true P.



(e) Identified and factorized spectrum with true P.

Figure A.7: Log-plot of the true spectrum and the identified spectra, both before and after factorization.

parameter, which gives the estimate $\tau_{\text{est}} = 0.0570$. Moreover, they are used to estimate covaraiances $r_{\mathbf{k}}$ on a grid $\Lambda = \{(k_1, k_2) \in \mathbb{Z}^2 \mid |k_1| \leq 3, |k_2| \leq 3\}$. Note that this grid Λ does not agree with the true degree of the linear system. From the estimated covariances $(r_{\mathbf{k}})$ we determine the covariances $(c_{\mathbf{k}})$, which are then used in the dual optimization problem described in Theorem A.2.1. We compute the solution with two different P, the first one being $P \equiv 1$, which corresponds to the maximum entropy (ME) solution, and the second one being $P = P_{\text{true}}$, i.e., the trigonometric polynomial corresponding to the filter b. The optimization problems are solved using the CVX toolbox in Matlab [41, 40]. The corresponding spectra obtained are shown in Figure A.7. As can be seen in the figure, using the true P gives a better agreement with the true spectrum, shown in Figure A.7a, which indicates that an appropriate tuning of p can improve the fit. Although there are methods in the literature on how to do simultaneously estimation of p and q [10, 27, 44, 72] (cf. Theorem A.2.4), the question on how to best select p is still open.

After estimating the spectra, we compute estimates of filter coefficients for the autoregressive part of the linear system, and the corresponding estimates are

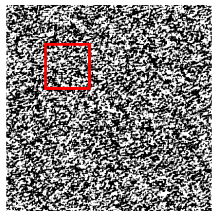
$$A_{\text{ME}} = \begin{bmatrix} 4.1270 & -3.8799 & 0.3572 & 0.2297 \\ -5.4210 & 4.0752 & 0.4412 & -0.2174 \\ 2.4057 & -0.0926 & -1.7157 & 0.1816 \\ -0.4199 & -0.6931 & 0.9018 & -0.1010 \end{bmatrix},$$

$$A_{\text{True } P} = \begin{bmatrix} 3.7207 & -4.3079 & 1.3210 & -0.0861 \\ -4.2527 & 5.4070 & -1.6585 & 0.0364 \\ 1.3381 & -1.6108 & 0.1836 & 0.2351 \\ -0.0562 & 0.0019 & 0.2183 & -0.2145 \end{bmatrix}.$$

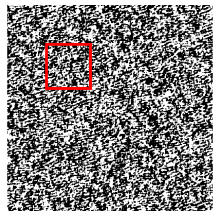
Using these filter coefficients, together with the corresponding filter coefficients for the moving-average part, we simulate the estimated Wiener system. The corresponding generated textures are shown in Figure A.8. Visually, the generated textures seem to have similar structures. However, by comparing the covariances, which are shown in Figure A.9, it can be seen that the texture generated by the filter obtained using the true p matches the higher order covariances considerably better.

A.9 Application to image compression

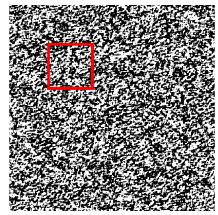
Since the expression (A.1.2b) is determined by a limited number of parameters, this approach enables compression of data. Moreover, the smoothness of the parameterization will facilitate tuning to specifications. Therefore we apply the two-dimensional circulant RCEP to compression of black-and-white images. Compression is achieved by approximating the image with a rational spectrum, thereby using fewer parameters. We compare the ME spectrum to the solution resulting from regularized covariance and cepstral matching. By choosing $n_1 \ll N_1$, $n_2 \ll N_2$,



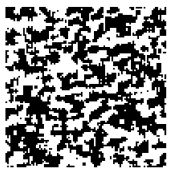
(a) Output of true system.



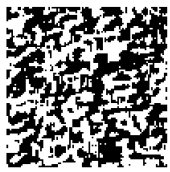
(b) Output of identified maximum entropy system.



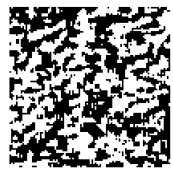
(c) Output of identified system with true p.



(d) Close-up of Figure A.8a.

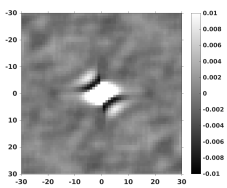


(e) Close-up of Figure A.8b.

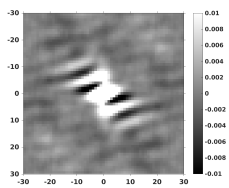


(f) Close-up of Figure A.8c.

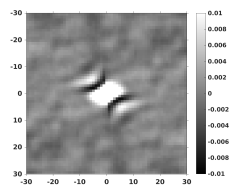
Figure A.8: Output of the true and the identified systems. Figures A.8a - A.8c show 500×500 samples, and Figures A.8d - A.8f show 100×100 samples.



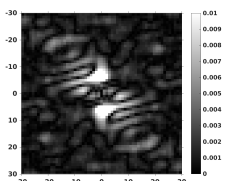
(a) Covariances $r_{\mathbf{k}}$ of texture in Figure A.8a.



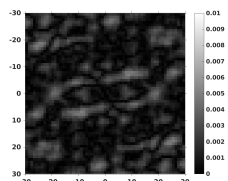
(b) Covariances $r_{\mathbf{k}}^{\mathrm{ME}}$ of texture in Figure A.8b.



(c) Covariances $r_{\mathbf{k}}^{\mathrm{True}\ P}$ of texture in Figure A.8c.



(d) Absolute error $|r_{\mathbf{k}} - r_{\mathbf{k}}^{\mathrm{ME}}|$.



(e) Absolute error $|r_{\mathbf{k}} - r_{\mathbf{k}}^{\mathrm{True}\ P}|$.

Figure A.9: Covariances and covariance errors for the textures. Here $\mathbf{k} = (k_1, k_2)$ where the x-axis corresponds to k_1 and the y-axis corresponds to k_2.

where N_1 and N_2 are the dimensions of the image, we obtain a significant reduction in the number of parameters describing the image.

A seemingly straight-forward way is to compute the covariances and cepstral coefficients directly from the image, and then use these to compute the spectrum. However, if the discrete spectrum is zero in one of the grid points, the (discrete) cepstrum is not well-defined. Hence simultaneous covariance and cepstral matching cannot be applied. Therefore we transform the image, denoted by Ψ, using $\Phi = e^{\Psi}$. Since Ψ is real, Φ is guaranteed to be real and positive for all discrete frequencies, and Ψ is obtained as $\Psi = \log \Phi$. We then compute (A.1.1) and (A.2.3) and obtain the approximant $\hat{\Phi}$ from Theorem A.5.7. Here we use the real sequences of covariances and cepstral coefficients obtained by extending the image by symmetric mirroring (i.e., using the discrete cosine transform [68, Section 4.2]). However, the covariances and cepstral coefficients of Φ can also be computed as the inverse two-dimensional FFT of e^{Ψ} and Ψ, respectively.

Moreover, note that an ME solution of the same maximum degree as a solution with a full-degree P has about half the number of parameters. To compensate for this, we let the degree of the ME solution be a factor $\sqrt{2}$ higher (rounded up), in order to get a fair comparison.

Compression of simplistic images

To better understand the different methods we first perform compression on a simple image of only black and white squares. The original image is shown in Figure A.10 and various results are shown in Figure A.11. Figure A.11a, shows that, if too few coefficients are used, the compression cannot represent the harmonics present in the image, regardless of the use of a nontrivial P. A visual assessment of the result shows that A.11e clearly outperforms A.11a, and that A.11f is still slightly better than A.11b. However A.11c and A.11d are better than A.11g and A.11h, respectively. In order to more objectively assess the quality of the two different compression methods, we also compute the MSSIM value of the compressed images. This is a measure, taking values in the interval $[0, 1]$, for evaluating quality and degradation of images, for which 1 means exact agreement [81]. A plot of the MSSIM value for compressions of different degree is shown in Figure A.12. However note that this measure does not agree completely with the visual impression of all images. Most notably, the measure gives a higher value to the grey image in Figure A.11a than the image with structure in Figure A.11e.

Compression of real images

We now apply the methods to some more realistic images. In the first example, shown in Figure A.13a, the original image is the Shepp-Logan phantom often used in medical imaging [77], of size 256×256 pixels. In Figure A.13b a compression using covariance and cepstral mathing is shown, where $n_1 + 1 = n_2 + 1 = 30$. Hence this image is described by $2 \cdot 30^2 = 1800$ parameters, compared to the original

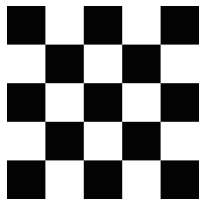
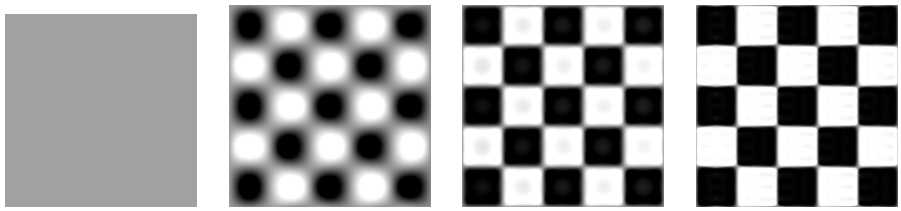


Figure A.10: A simplistic test image. Each black or white square is 128×128 pixels.

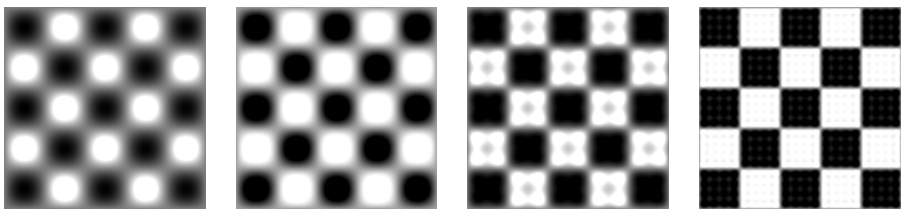


(a) Cepstral matching, $n = 4$.

(b) Cepstral matching, $n = 10$.

(c) Cepstral matching, $n = 11$.

(d) Cepstral matching, $n = 30$.



(e) ME solution, $n = 6$.

(f) ME solution, $n = 15$.

(g) ME solution, $n = 16$.

(h) ME solution, $n = 43$.

Figure A.11: Compressions of the simple image shown in Figure A.10. The top row shows compression with regularized covariance and cepstral matching, where $\lambda = 10^{-2}$, and the bottom row shows compression with the ME solution. In all cases $n_1 = n_2$, and the pair of compressions in each column have approximately the same number of parameters, namely, $n_{\text{me}} \approx \sqrt{2}\, n_{\text{ceps}}$.

$256^2 = 65536$ parameters, which corresponds to a reduction in parameters of about 97%. We also compute an ME compression, with degree $n_1+1 = n_2+1 = 45 \approx \sqrt{2}\cdot 30$ which is shown in Figure A.13c.

The second example is a compression of the classical Lenna image, often used in the image processing literature. The original image, shown in Figure A.14a, is 512×512 pixels. For regularized cepstral matching we set $n_1 + 1 = n_2 + 1 = 60$, corresponding to a compression rate of about 97%, and the result is shown in Figure A.14b. The ME compression, computed with $n_1 + 1 = n_2 + 1 = 85 \approx \sqrt{2} \cdot 60$, is shown in Figure A.14c.

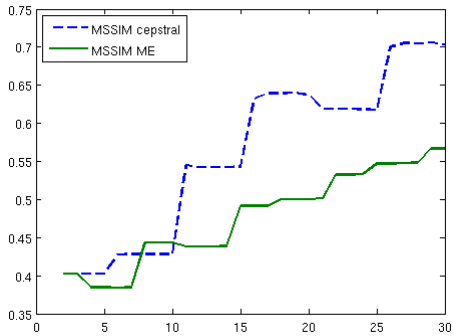


Figure A.12: MSSIM values of different compression levels, plotted against n for the compression with cepstral matching. Hence the corresponding ME compression has $\lceil\sqrt{2}n\rceil$ coefficients.

Table A.1: MSSIM values of different compression techniques on the two test images.

Shepp-Logan		Lenna	
Compression	MSSIM value	Compression	MSSIM value
Cepstral	0.8690	Cepstral	0.7451
ME	0.7044	ME	0.7489

The MSSIM values for these compressions are shown in Table A.1. They seem to agree with the visual impression. Interestingly the compression with cepstral matching is better for the Shepp-Logan phantom. However, in the Lenna image neither of the methods outperform the other. The ME compression has more ringing artifacts, but it is less blurred than the cepstral compression. We believe that this is related to the fact that if you have relatively few sharp transitions in pixel values, which is the case in Figures A.10 and A.13a, placing both poles and zero close to each other can achieve this transition efficiently and thus give better quality on the compressed image. However when this is not the case, as with the Lenna image, the trade-off between having spectral zeros or matching higher frequencies is more complex.

Similar methods have previously been used for compression of textures [21, 65], where, instead of a scalar two-dimensional moment problem, a one-dimensional vector problem is considered. Here the image is modeled by a periodic stochastic vector process rather than a two-dimensional random field, leading to a discrete vector moment problem akin to the one presented in [54]. This is connected to the circulant moment problem considered in Section A.2 and to modeling of reciprocal systems [52, 20].

(a) Original image. (b) Cepstral matching, $n = 30$ and $\lambda = 10^{-2}$. (c) ME solution, $n = 45$.

Figure A.13: Compression of the Shepp-Logan phantom with a compression rate of about 97%.

(a) Original image. (b) Cepstral matching, $n = 60$ and $\lambda = 10^{-2}$. (c) ME solution, $n = 85$.

Figure A.14: Compression of the Lenna image with a compression rate of about 97%.

A.10 Appendix

In this appendix we provide the proofs that have been deferred in the main text. Some of the proofs use general properties of multidimensional trigonometric polynomials, summarized in this lemma.

Lemma A.10.1. *For all $P \in \bar{\mathfrak{P}}_+$ we have i) $|p_{k_1\dots,k_d}| \leq p_{0\dots,0}$ and ii) $\|P\|_\infty \leq |\Lambda|\|p\|_\infty$.*

Proof. The fact that $|p_{\mathbf{k}}| = \left|\int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} P dm\right| \leq \int_{\mathbb{T}^d} |e^{i(\mathbf{k},\boldsymbol{\theta})}|\,|P| dm = p_0$ shows i). Next we note that P has $|\Lambda|$ coefficients, and hence

$$\|P\|_\infty \leq \sup_{\boldsymbol{\theta}\in\mathbb{T}^d} \sum_{\mathbf{k}\in\Lambda} |p_k||e^{i(\mathbf{k},\boldsymbol{\theta})}| = \sum_{\mathbf{k}\in\Lambda} |p_{\mathbf{k}}| \leq |\Lambda|\|p\|_\infty,$$

which proves ii). □

Proof of Lemma A.3.1. To show lower semicontinuity of

$$\mathbb{J}_P(Q) = \langle c, q\rangle - \int_{\mathbb{T}^d} P \log Q \, dm = \langle c, q\rangle + \int_{\mathbb{T}^d} -P \log Q \, dm$$

we note that $\langle c, q\rangle$ is continuous and hence only the integral needs to be considered.

Fix any $Q \in \bar{\mathfrak{P}}_+ \setminus \{0\}$. From [76, p. 223] we know that it is log-integrable. Moreover, let (Q_n) be a sequence of trigonometric polynomials in $\bar{\mathfrak{P}}_+ \setminus \{0\}$ that converges to Q in $L^\infty(\mathbb{T}^d)$. We know that Q is bounded, and, since the convergence $Q_n \to Q$ is uniform, we must have $M := \sup_n\{\max_{\boldsymbol{\theta}}[Q_n]\} < \infty$, and thus $0 \leq Q/M \leq 1$ and $0 \leq Q_n/M \leq 1$ for all n. Moreover, $\lim_{n\to\infty} -\log(Q_n/M) = -\log(Q/M)$ in an extended real-valued sense. Since $-P\log(Q_n/M) \geq 0$, by Fatou's lemma [75, p. 23], we have

$$\int_{\mathbb{T}^d} -P \log\left(\frac{Q}{M}\right) dm \leq \liminf_{n\to\infty} \int_{\mathbb{T}^d} -P \log\left(\frac{Q_n}{M}\right) dm.$$

Since (Q_n) is an arbitrary sequence, the functional is lower semicontinuous in Q. Moreover, since Q is also arbitrary it follows that $\mathbb{J}_P$ is lower semicontinuous on $\bar{\mathfrak{P}}_+ \setminus \{0\}$. □

Proof of Proposition A.4.2. Let $\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3 \in \Lambda$ be three linearly independent index vectors. First note that for the nonnegative trigonometric polynomial $Q(e^{i\boldsymbol{\theta}}) = \sum_{\ell=1}^3 (1 - (e^{i(\mathbf{k}_\ell, \boldsymbol{\theta})} + e^{-i(\mathbf{k}_\ell, \boldsymbol{\theta})})/2)$ we have $Q(e^{i\mathbf{0}}) = 0$, and hence $Q \in \partial\mathfrak{P}_+$. Next we will show that $\int_{\mathbb{T}^d} Q^{-1} dm(\boldsymbol{\theta})$ is finite. By the variable change $\boldsymbol{\phi} = A\boldsymbol{\theta}$, where $A \in \mathbb{R}^{d\times d}$ is selected to be invertible and with the ℓth row equal to $\mathbf{k}_\ell$ for $\ell = 1, 2, 3$, the integral becomes

$$\int_{\mathbb{T}^d} \frac{1}{Q} dm(\boldsymbol{\theta}) = \int_{A(\mathbb{T}^d)} \frac{\det(A)^{-1}}{\sum_{\ell=1}^3 (1 - \cos(\phi_\ell))} dm(\boldsymbol{\phi}),$$

where the set $A(\mathbb{T}^d) = \{A\boldsymbol{\theta} \mid \boldsymbol{\theta} \in \mathbb{T}^d\}$. Due to the periodicity of the integrand, the integral is bounded by

$$\kappa \int_{\mathbb{T}^3} \frac{d\phi_1 d\phi_2 d\phi_3}{\sum_{\ell=1}^3 (1 - \cos(\phi_\ell))}$$

for some constant κ that depends on A and d. This bound is finite [49, 46], and therefore the proposition follows. □

To prove Theorem A.4.4, we need the following lemma.

Lemma A.10.2. *f^p is a bijective map.*

Proof. By Corollary A.2.3, f^p is injective, since there is a unique minimizer of (A.2.2) over all $Q \in \mathfrak{P}_+$. Hence there is at most one q corresponding to a certain c, proving injectivity. Surjectivity also follows from Corollary A.2.3. We fix a $P \in \mathfrak{P}_+$ and simply note that there exists a unique solution for all $c \in \mathfrak{C}_+$, given by $q = (f^p)^{-1}(c)$. □

Proof of Theorem A.4.4. In the proof of Theorem A.2.1 we saw that for all nontrivial variations δQ, $\partial^2 \mathbb{J}_P(Q; \delta Q) > 0$. Hence

$$\frac{\partial f^p_{\mathbf{k}}}{\partial q_{\boldsymbol{\ell}}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k}-\boldsymbol{\ell},\boldsymbol{\theta})} \frac{P}{Q^2} dm = \frac{\partial^2 \mathbb{J}_P(Q)}{\partial q_{\boldsymbol{\ell}} \partial \bar{q}_{\mathbf{k}}} \tag{A.10.1}$$

is positive definite. Next, we define the map

$$\varphi^p : \mathfrak{C}_+ \times \mathfrak{P}_+ \to \{(r_{\mathbf{k}})_{\mathbf{k}\in\Lambda} \in \mathbb{C}^{|\Lambda|} \,|\, r_{-\mathbf{k}} = \bar{r}_{\mathbf{k}}, \mathbf{k} \in \Lambda\} \cong \mathbb{R}^{|\Lambda|}$$

as

$$\varphi^p_{\mathbf{k}}(c, q) = c_{\mathbf{k}} - \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} \frac{P}{Q} dm,$$

component-wise. By Corollary A.2.3, $\varphi^p(c, q) = 0$ has a unique solution for each $c \in \mathfrak{C}_+$. Since $\partial\varphi^p/\partial q = \partial f^p/\partial q$ is invertible, the implicit function theorem implies that $q = (f^p)^{-1}(c)$ is locally a $\mathcal{C}^1$ function and hence a local diffeomorphism. However, f^p is a bijection (Lemma A.10.2) and therefore a (global) diffeomorphism. □

By Theorem A.4.4, the function g^c is a well-defined map. The proof of Theorem A.4.5 now follows along the same lines.

Lemma A.10.3. *g^c is a bijective map.*

Proof. Surjectivity of g^c on the image $\mathfrak{Q}_+$ follows directly from definition. A straight-forward generalization of Lemma 2.4 in [17] shows that g^c is injective. □

Proof of Theorem A.4.5. Let the map

$$\varphi^c : \mathfrak{P}_+ \times \mathfrak{P}_+ \to \{(r_{\mathbf{k}})_{\mathbf{k}\in\Lambda} \in \mathbb{C}^{|\Lambda|} \,|\, r_{-\mathbf{k}} = \bar{r}_{\mathbf{k}}, \mathbf{k} \in \Lambda\} \cong \mathbb{R}^{|\Lambda|}$$

be given by

$$\varphi^c_{\mathbf{k}}(p, q) = c_{\mathbf{k}} - \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} \frac{P}{Q} dm.$$

The Jacobian with respect to q is the same as (A.10.1). Hence $q = g^c(p)$ is $\mathcal{C}^1$ by the implicit function theorem. Since (A.10.1) gives a positive definite Jacobian matrix,

$$\frac{\partial \varphi^c_{\mathbf{k}}}{\partial p_{\boldsymbol{\ell}}} = -\int_{\mathbb{T}^d} e^{i(\mathbf{k}-\boldsymbol{\ell},\boldsymbol{\theta})} \frac{1}{Q} dm$$

defines an invertible Jacobian. Hence $p = (g^c)^{-1}(q)$ is $\mathcal{C}^1$, so g^c is a local diffeomorphism. Since it is a bijection (Lemma A.10.3), it is a (global) diffeomorphism. □

Proof of Lemma A.5.1. For any $Q \in \bar{\mathfrak{P}}_+ \setminus \{0\}$, $\log Q$ is integrable [76, p. 223]. Since $P \in \bar{\mathfrak{P}}_{+,\circ}$, P is not the zero polynomial, hence, since $x \log x \to 0$ as $x \to 0$, $P \log P$ is integrable and in fact continuous for all $P \in \bar{\mathfrak{P}}_{+,\circ}$. Hence

$$\int_{\mathbb{T}^d} P \log P \, dm - \int_{\mathbb{T}^d} P \log Q \, dm = \int_{\mathbb{T}^d} P \log \left(\frac{P}{Q} \right) dm,$$

and therefore we can rewrite the functional $\mathbb{J}(P, Q)$ as

$$\mathbb{J}(P, Q) = \langle c, q \rangle - \langle \gamma, p \rangle + \int_{\mathbb{T}^d} P \log P \, dm - \int_{\mathbb{T}^d} P \log Q \, dm.$$

All terms in this expression are continuous, except possibly the last integral. However, following along the same lines as in the proof of Lemma A.3.1, we can apply Fatou's lemma showing that $\mathbb{J}(P, Q)$ is lower semicontinuous. □

Proof of Lemma A.5.2. To show that $\mathbb{J}$ has compact sublevel sets $\mathbb{J}^{-1}(-\infty, r]$, we proceed as in [55, p. 503] by first splitting the objective function into two parts

$$\mathbb{J}_1(P, Q) = \langle c, q \rangle - \int_{\mathbb{T}^d} P \log Q \, dm \quad \text{and} \quad \mathbb{J}_2(P) = -\langle \gamma, p \rangle + \int_{\mathbb{T}^d} P \log P \, dm.$$

The sublevel set consists of the $(P, Q) \in \bar{\mathfrak{P}}_{+,\circ} \times \bar{\mathfrak{P}}_+$ such that $r \geq \mathbb{J}_1(P, Q) + \mathbb{J}_2(P)$, and from Lemma A.3.3 we have $\mathbb{J}_1(P, Q) \geq \varepsilon \|Q\|_\infty - \log \|Q\|_\infty$, since $\int_{\mathbb{T}^d} P dm = 1$ by (A.5.5). Next we show that $\mathbb{J}_2(P)$ is bounded from below. We first note that since $P \in \bar{\mathfrak{P}}_{+,\circ}$ we have $p_0 = 1$, and thus P is bounded away from the zero polynomial. Now, since $x \log(x)$ achieves a minimum $> -\infty$ on any compact set $[0, a]$, $P \log P$ must achieve a minimum $> -\infty$ on $\mathbb{T}^d$. Calling this minimum κ_P, we have

$$\int_{\mathbb{T}^d} P \log P \, dm \geq \int_{\mathbb{T}^d} \kappa_P dm = \kappa_P.$$

To bound the term $-\langle \gamma, p \rangle$ from below we note that

$$\langle \gamma, p \rangle = \sum_{\mathbf{k} \in \Lambda} \bar{\gamma}_{\mathbf{k}} p_{\mathbf{k}} \leq \left| \sum_{\mathbf{k} \in \Lambda} \bar{\gamma}_{\mathbf{k}} p_{\mathbf{k}} \right| \leq \sum_{\mathbf{k} \in \Lambda} |\bar{\gamma}_{\mathbf{k}}| \, |p_{\mathbf{k}}| \leq \sum_{\mathbf{k} \in \Lambda} \|\gamma\|_\infty |p_{\mathbf{k}}| \leq \|\gamma\|_\infty |\Lambda| \|p\|_\infty$$

and thus $-\langle \gamma, p \rangle \geq -|\Lambda| \|\gamma\|_\infty \|p\|_\infty = -|\Lambda| \|\gamma\|_\infty$, since $\|p_\infty\| = p_0 = 1$ by Lemma A.10.1. Hence there exist some $\rho > -\infty$ such that $\mathbb{J}_2(P) \geq \rho$. From this we have that for any $(P, Q) \in \mathbb{J}^{-1}(-\infty, r]$,

$$r \geq \mathbb{J}(P, Q) \geq \varepsilon \|Q\|_\infty - \log \|Q\|_\infty + \rho.$$

This shows that the sublevel set $\mathbb{J}^{-1}(-\infty, r]$ is a subset of $\{(P, Q) \in \bar{\mathfrak{P}}_{+,\circ} \times \bar{\mathfrak{P}}_+ \mid \varepsilon \|Q\|_\infty - \log \|Q\|_\infty \leq r - \rho\}$. Since $\bar{\mathfrak{P}}_{+,\circ}$ is bounded, by comparing linear and logarithmic growth we see that this set is bounded, and thus so is also the sublevel set. As before, since it is the sublevel set of a lower semicontinuous function it will be closed, and hence it is compact. □

Proof of Lemma A.5.3. Consider the directional derivative of $\mathbb{J}$ in a point $(P,Q) \in \bar{\mathfrak{P}}_{+,\circ} \times \in \bar{\mathfrak{P}}_+$ in any direction $(\delta P, \delta Q)$ such that $P+\varepsilon\delta P \in \bar{\mathfrak{P}}_{+,\circ}$, and $Q+\varepsilon\delta Q \in \bar{\mathfrak{P}}_+$ for all $\varepsilon \in (0,a)$ for some $a > 0$. A quite straight-forward calculation yields

$$\delta\mathbb{J}(P,Q;\delta P,\delta Q) = \langle c, \delta q\rangle - \langle \gamma, \delta p\rangle + \int_{\mathbb{T}^d} \left[\delta P \log\left(\frac{P}{Q}\right) - \delta Q \frac{P}{Q}\right] dm,$$

where we have used the fact, obtained from (A.5.5), that $\int_{\mathbb{T}^d} \delta P dm = \delta p_0 = 0$, since $p_0 = 1$ is constant. Likewise, the second directional derivative becomes

$$\delta^2\mathbb{J}(P,Q;\delta P,\delta Q) = \int_{\mathbb{T}^d} P\left(\delta P \frac{1}{P} - \delta Q \frac{1}{Q}\right)^2 dm,$$

which is clearly nonnegative for all feasible directions and hence positive semidefinite. Thus the problem is convex. □

Proof of Lemma A.6.4. First note that $\mathfrak{C}_+(\mathbf{N}) \subset \mathfrak{C}_+$. To prove the lemma, it is sufficient to prove that any $c \in \mathfrak{C}_+$ belongs to $\mathfrak{C}_+(\mathbf{N})$ if $\min(\mathbf{N})$ is large enough.

Let $c \in \mathfrak{C}_+$. From (A.3.4) there exists $\kappa_c > 0$ such that

$$\langle c, p\rangle \geq \kappa_c \|p\|_\infty \quad \text{for all } p \in \bar{\mathfrak{P}}_+. \tag{A.10.2}$$

We want to show that $\langle c, \hat{p}\rangle > 0$ for any $\hat{p} \in \bar{\mathfrak{P}}_+(\mathbf{N}) \setminus \{0\}$. Without loss of generality we may take $\|\hat{p}\|_\infty = 1$. Then $|\partial\hat{P}(e^{i\boldsymbol{\theta}})/\partial\theta_j| \leq \sum_{\mathbf{k}\in\Lambda} |k_j|$, and, since $\hat{P}(e^{i\boldsymbol{\theta}}) \geq 0$ in $\boldsymbol{\theta} \in \mathbb{T}_{\mathbf{N}}$, it follows that $\hat{P}(e^{i\boldsymbol{\theta}}) \geq -\pi\Delta/\min(\mathbf{N})$, where $\Delta = \sum_{\mathbf{k}\in\Lambda} \|\mathbf{k}\|_1$. Therefore $\tilde{P} := \hat{P} + \pi\Delta/\min(\mathbf{N}) \in \bar{\mathfrak{P}}_+$, and by using (A.10.2) we get

$$\begin{aligned}\langle c, \hat{p}\rangle &= \langle c, \hat{p}\rangle + c_0\frac{\pi\Delta}{\min(\mathbf{N})} - c_0\frac{\pi\Delta}{\min(\mathbf{N})} = \langle c, \tilde{p}\rangle - c_0\frac{\pi\Delta}{\min(\mathbf{N})} \\ &\geq \kappa_c\|\tilde{p}\|_\infty - c_0\frac{\pi\Delta}{\min(\mathbf{N})} = \kappa_c\left(\|\hat{p}\|_\infty + \frac{\pi\Delta}{\min(\mathbf{N})}\right) - c_0\frac{\pi\Delta}{\min(\mathbf{N})}. \end{aligned} \tag{A.10.3}$$

The last equality follows from the fact that by Lemma A.10.1 we have that $\|\tilde{p}\|_\infty = \tilde{p}_0 = p_0 + \pi\Delta/\min(\mathbf{N}) = \|\tilde{p}\|_\infty + \pi\Delta/\min(\mathbf{N})$. Now we note that $\kappa_c \leq c_0$, since $P \equiv 1$ is a feasible point in the minimization defining κ_c (see the paragraph above (A.3.4)). Therefore, if $\kappa_c = c_0$, then by (A.10.3) we have that $\langle c, \hat{p}\rangle > 0$. If instead $\kappa_c < c_0$, then by selecting $\min(\mathbf{N}) > \pi\Delta(1 - c_0/\kappa_c)$, we obtain $\langle c, \hat{p}\rangle > \kappa_c\|\hat{p}\|_\infty > 0$. In any case, since $\hat{p} \in \bar{\mathfrak{P}}_+(\mathbf{N}) \setminus \{0\}$ is arbitrary, it follows that $c \in \mathfrak{C}_+(\mathbf{N})$. □

Proof of Lemma A.6.5. For a fixed $\tilde{Q} \in \mathfrak{P}_+$ we have $\lim_{\min(\mathbf{N})\to\infty} \mathbb{J}_P^{\mathbf{N}}(\tilde{Q}) = \mathbb{J}_P(\tilde{Q})$, since the sums in (A.6.3b) are Riemann sums converging to (A.6.3a). Hence we can define $L := \sup_N \mathbb{J}_P^{\mathbf{N}}(\tilde{Q}) < \infty$. Also, by optimality, $\infty > \mathbb{J}_P^{\mathbf{N}}(\tilde{Q}) \geq \mathbb{J}_P^{\mathbf{N}}(\hat{Q}_{\mathbf{N}})$ for all values of $\mathbf{N}$ and also $\infty > \mathbb{J}_P(\tilde{Q}) \geq \mathbb{J}_P(\hat{Q})$. Using this and Lemma A.3.3 we obtain

$$L \geq \mathbb{J}_P^{\mathbf{N}}(\tilde{Q}) \geq \mathbb{J}_P^{\mathbf{N}}(\hat{Q}_{\mathbf{N}}) \geq \varepsilon_{\mathbf{N}}\|\hat{Q}_{\mathbf{N}}\|_\infty - \|P\|_1\|\log(\hat{Q}_{\mathbf{N}})\|_\infty$$

for all values of $\mathbf{N}$. In accordance with (A.3.5), we can choose $\varepsilon_{\mathbf{N}} := \kappa_c^{\mathbf{N}}/|\Lambda|$, where $\kappa_c^{\mathbf{N}}$ is the minimum value of $\langle c, q_{\mathbf{N}} \rangle$ on the compact set $\{Q \in \bar{\mathfrak{P}}_+(\mathbf{N}) \mid \|q\|_\infty = 1\}$. If we can show $\kappa_c := \inf_{\mathbf{N}} \kappa_c^{\mathbf{N}} > 0$, we can choose $\varepsilon := \kappa_c/|\Lambda| \leq \varepsilon_{\mathbf{N}}$ for all $\mathbf{N}$, so that

$$L \geq \varepsilon \|\hat{Q}_{\mathbf{N}}\|_\infty - \|P\|_1 \|\log(\hat{Q}_{\mathbf{N}})\|_\infty.$$

Then comparing linear and logarithmic growth this implies that $(\hat{Q}_{\mathbf{N}})$ is bounded.

To show that $\kappa_c > 0$ first note that for every finite value of $\min(\mathbf{N})$ we have $\kappa_c^{\mathbf{N}} > 0$. Now assume $\inf_{\mathbf{N}} \kappa_c^{\mathbf{N}} = 0$. Then there must exist a sequence $(q_{\mathbf{N}}^\star)$ such that $\langle c, q_{\mathbf{N}}^\star \rangle \to 0$ as $\min(\mathbf{N}) \to \infty$, where $q_{\mathbf{N}}^\star \in \bar{\mathfrak{P}}_+(\mathbf{N})$ and $\|q_{\mathbf{N}}^\star\|_\infty = 1$. Now, since every $q_{\mathbf{N}}^\star$ is a vector in $\mathbb{C}^{|\Lambda|}$, the constraint $\|q\|_\infty = 1$ defines a compact set. Hence there is a subsequence, also indexed with $\mathbf{N}$, so that $q^\star := \lim_{\min(\mathbf{N})\to\infty} q_{\mathbf{N}}^\star$ is well-defined and $\|q^\star\|_\infty = 1$. Then $\langle c, q^\star \rangle = 0$. However, since $c \in \mathfrak{C}_+$ and $q^\star \in \bar{\mathfrak{P}}_+$, this implies that $q^\star = 0$, which contradicts $\|q^\star\|_\infty = 1$. Hence $\kappa_c > 0$, as claimed. □

References

[1] M.R. Abdalmoaty and H. Hjalmarsson. A simulated maximum likelihood method for estimation of stochastic wiener systems. In *IEEE Annual Conference on Decision and Control (CDC)*, pages 3060–3065. IEEE, 2016.

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

[3] J.S. Bendat. *Nonlinear systems techniques and applications*. Wiley, 1998.

[4] S.A. Billings. Identification of nonlinear systems-a survey. *IEE Proceedings D-Control Theory and Applications*, 127(6):272–285, 1980.

[5] A. Blomqvist, A. Lindquist, and R. Nagamune. Matrix-valued Nevanlinna-Pick interpolation with complexity constraint: An optimization approach. *IEEE Transactions on Automatic Control*, 48(12):2172–2190, 2003.

[6] N.K. Bose. *Multidimensional Systems Theory and Applications*. Kluwer Academic Publishers, second edition, 2003.

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

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

[9] 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.

[10] 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.

[11] C.I. Byrnes, T.T. Georgiou, and A. Lindquist. A new approach to spectral estimation: A tunable high-resolution spectral estimator. *IEEE Transactions on Signal Processing*, 48(11):3189–3205, 2000.

[12] 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.

[13] C.I. Byrnes, T.T. Georgiou, A. Lindquist, and A. Megretski. Generalized interpolation in H^∞ with a complexity constraint. *Transactions of the American Mathematical Society*, 358(3):965–987, 2006.

[14] 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.

[15] 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.

[16] 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.

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

[18] C.I. Byrnes and A. Lindquist. The moment problem for rational measures: Convexity in the spirit of Krein. In *Modern Analysis and Application: Mark Krein Centenary Conference, Vol. I: Operator Theory and Related Topics*, volume 190 of *Operator Theory Advances and Applications*, pages 157–169. Birkhäuser, 2009.

[19] 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.

[20] F. P. Carli, A. Ferrante, M. Pavon, and G. Picci. A maximum entropy solution of the covariance extension problem for reciprocal processes. *Automatic Control, IEEE Transactions on*, 56(9):1999–2012, 2011.

[21] A. Chiuso, A. Ferrante, and G. Picci. Reciprocal realization and modeling of textured images. In *IEEE Annual Conference on Decision and Control (CDC), and European Control Conference (ECC)*, pages 6059–6064. IEEE, 2005.

[22] J.R. Deller, J.G. Proakis, and J.H.L. Hansen. *Discrete-time processing of speech signals*. IEEE Press, Piscataway, N.Y., 2000.

[23] B. Dickinson. Two-dimensional markov spectrum estimates need not exist. *IEEE Transactions on Information Theory*, 26(1):120–121, 1980.

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

[25] M.P. Ekstrom. *Digital image processing techniques*. Academic Press, 1984.

[26] M.P. Ekstrom and J.W. Woods. Two-dimensional spectral factorization with applications in recursive digital filtering. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(2):115–128, 1976.

[27] 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.

[28] 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.

[29] S. Eriksson Barman. Gaussian random field based models for the porous structure of pharmaceutical film coatings. In *Acta Stereologica [En ligne], Proceedings ICSIA, 14th ICSIA abstracts*, 2015. `http://popups.ulg.ac.be/0351-580X/index.php?id=3775`.

[30] G. Fanizza. *Modeling and Model Reduction by Analytic Interpolation and Optimization*. PhD thesis, 2008. Optimization and Systems Theory, Department of Mathematics, KTH Royal Institue of Technology.

[31] A. Ferrante, M. Pavon, and F. Ramponi. Further results on the Byrnes-Georgiou-Lindquist generalized moment problem. In A. Chiuso, S. Pinzoni, and A. Ferrante, editors, *Modeling, Estimation and Control*, pages 73–83. Springer, 2007.

[32] A. Ferrante, M. Pavon, and F. Ramponi. Hellinger versus Kullback-Leibler multivariable spectrum approximation. *IEEE Transactions on Automatic Control*, 53(4):954–967, 2008.

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

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

[35] T.T. Georgiou. The interpolation problem with a degree constraint. *IEEE Transactions on Automatic Control*, 44(3):631–635, 1999.

[36] T.T. Georgiou. Solution of the general moment problem via a one-parameter imbedding. *IEEE Transactions on Automatic Control*, 50(6):811–826, 2005.

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

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

[39] J. S. Geronimo and H. J. Woerdeman. Positive extensions, Fejér-Riesz factorization and autoregressive filters in two variables. *Annals of Mathematics*, 160(3):839–906, 2004.

[40] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, volume 371 of *Lecture Notes in Control and Information Sciences*, pages 95–110. Springer-Verlag, London, 2008.

[41] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, Mar. 2014.

[42] W. Greblicki. Nonparametric identification of wiener systems. *IEEE Transactions on information theory*, 38(5):1487–1493, 1992.

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

[44] J. Karlsson, T.T. Georgiou, and A. Lindquist. The inverse problem of analytic interpolation with degree constraint and weight selection for control synthesis. *IEEE Transactions on Automatic Control*, 55(2):405–418, 2010.

[45] J. Karlsson and A. Lindquist. Stability-preserving rational approximation subject to interpolation constraints. *IEEE Transactions on Automatic Control*, 53(7):1724–1730, 2008.

[46] J. Karlsson, A. Lindquist, and A. Ringh. The multidimensional moment problem with complexity constraint. *Integral Equations and Operator Theory*, 84(3):395–418, 2016.

[47] S.W. Lang and J.H. McClellan. Spectral estimation for sensor arrays. In *Proceedings of the First ASSP Workshop on Spectral Estimation*, pages 3.2.1–3.2.7, 1981.

[48] S.W. Lang and J.H. McClellan. The extension of Pisarenko's method to multiple dimensions. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 125–128, 1982.

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

[50] 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.

[51] H. Lev-Ari, S. Parker, and T. Kailath. Multidimensional maximum-entropy covariance extension. *IEEE Transactions on Information Theory*, 35(3):497–508, 1989.

[52] B.C. Levy, R. Frezza, and A.J. Krener. Modeling and estimation of discrete-time gaussian reciprocal processes. *IEEE Transactions on Automatic Control*, 35(9):1013–1023, 1990.

[53] A. Lindquist, C. Masiero, and G. Picci. On the multivariate circulant rational covariance extension problem. In *IEEE Annual Conference on Decision and Control (CDC)*, pages 7155–7161. IEEE, 2013.

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

[55] A. Lindquist and G. Picci. *Linear stochastic systems: A geometric approach to modeling, estimation and identification.* Springer, Berlin Heidelberg, 2015.

[56] F. Lindsten, T.B. Schön, and M.I. Jordan. Bayesian semiparametric Wiener system identification. *Automatica*, 49(7):2053–2063, 2013.

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

[58] K. Mahler. On some inequalities for polynomials in several variables. *Journal of the London Mathematical Society*, 1(1):341–344, 1962.

[59] J.H. McClellan and S.W. Lang. Mulit-dimensional MEM spectral estimation. In *Proceedings of the Institute of Acoustics "Spectral Analysis and its Use in Underwater Acoustics": Underwater Acoustics Group Conference, Imperial College, London, 29-30 April 1982*, pages 10.1–10.8, 1982.

[60] J.H. McClellan and S.W. Lang. Duality for multidimensional MEM spectral analysis. *Communications, Radar and Signal Processing, IEE Proceedings F*, 130(3):230–235, April 1983.

[61] B.R. Musicus and A.M. Kabel. Maximum entropy pole-zero estimation. Technical Report 510, Research Laboratory of Electronics, Massachusetts Institute of Technology, August 1985.

[62] H.I. Nurdin. New results on the rational covariance extension problem with degree constraint. *Systems & Control Letters*, 55(7):530 – 537, 2006.

[63] A.V. Oppenheim and R.W. Schafer. *Digital Signal Processing.* Prentice-Hall, New Jerseys, 1975.

[64] M. Pavon and A. Ferrante. On the geometry of maximum entropy problems. *SIAM Review*, 55(3):415–439, 2013.

[65] G. Picci and F.P. Carli. Modelling and simulation of images by reciprocal processes. In *Tenth international conference on Computer Modelling and Simulation, UKSIM*, pages 513–518, 2008.

[66] R. Price. A useful theorem for nonlinear devices having gaussian inputs. *IRE Transactions on Information Theory*, 4(2):69–72, 1958.

[67] F. Ramponi, A. Ferrante, and M. Pavon. A globally convergent matricial algorithm for multivariate spectral estimation. *IEEE Transactions on Automatic Control*, 54(10):2376–2388, 2009.

[68] K.R. Rao and P. Yip. *Discrete cosine transform: Algorithms, advantages, applications*. Academic press, San Diego, C.A., 1990.

[69] R. Remmert. *Theory of complex functions*. Graduate texts in mathematics. Springer, New York, NY, 1991. Translation of: Funktionentheorie I. 2nd ed.

[70] A. Rényi. On measures of entropy and information. In *Fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 547–561, 1961.

[71] A. Ringh and J. Karlsson. A fast solver for the circulant rational covariance extension problem. In *European Control Conference (ECC)*, pages 727–733, July 2015.

[72] 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.

[73] A. Ringh and A. Lindquist. Spectral estimation of periodic and skew periodic random signals and approximation of spectral densities. In *33rd Chinese Control Conference (CCC)*, pages 5322–5327, 2014.

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

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

[76] A. Schinzel. *Polynomials with special regard to reducibility*. Cambridge University Press, 2000.

[77] L.A Shepp and B.F. Logan. The Fourier reconstruction of a head section. *IEEE Transactions on Nuclear Science*, 21(3):21–43, 1974.

[78] E.M. Stein and R. Shakarchi. *Fourier analysis: An introduction*. Princeton University Press, Princeton, NJ, 2003.

[79] P. Stoica and R. Moses. *Introduction to Spectral Analysis*. Prentice-Hall, Upper Saddle River, NJ, 1997.

[80] B. Wahlberg, J. Welsh, and L. Ljung. Identification of stochastic wiener systems using indirect inference. *IFAC-PapersOnLine*, 48(28):620 – 625, 2015. 17th IFAC Symposium on System Identification SYSID 2015.

[81] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.

[82] J.W. Woods. Two-dimensional Markov spectral estimation. *IEEE Transactions on Information Theory*, 22(5):552–559, 1976.

[83] M. Zorzi. A new family of high-resolution multivariate spectral estimators. *IEEE Transactions on Automatic Control*, 59(4):892–904, 2014.

[84] M. Zorzi. Rational approximations of spectral densities based on the alpha divergence. *Mathematics of Control, Signals, and Systems*, 26(2):259–278, 2014.

Paper B

Multidimensional rational covariance extension with approximate covariance matching

Multidimensional rational covariance extension with approximate covariance mathcing

by

Axel Ringh, Johan Karlsson, and Anders Lindquist

Abstract

In our companion paper [A. Ringh, J. Karlsson, and A. Lindquist, SIAM J. Control Optim., 54 (2016), pp. 1950–1982] we discussed the multidimensional rational covariance extension problem (RCEP), which has important applications in image processing, and spectral estimation in radar, sonar, and medical imaging. This is an inverse problem where a power spectrum with a rational absolutely continuous part is reconstructed from a finite set of moments. However, in most applications these moments are determined from observed data and are therefore only approximate, and the RCEP may not have a solution. In this paper we extend the results of our companion paper to handle approximate covariance matching. We consider two problems, one with a soft constraint and the other one with a hard constraint, and show that they are connected via a homeomorphism. We also demonstrate that the problems are well-posed and illustrate the theory by examples in spectral estimation and texture generation.

Keywords: approximate covariance extension, trigonometric moment problem, convex optimization, multidimensional spectral estimation, texture generation

B.1 Introduction

Trigonometric moment problems are ubiquitous in systems and control, such as spectral estimation, signal processing, system identification, image processing and remote sensing [5, 20, 59]. In the (truncated) multidimensional trigonometric moment problem we seek a nonnegative measure $d\mu$ on $\mathbb{T}^d$ satisfying the moment equation

$$c_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} d\mu(\boldsymbol{\theta}) \quad \text{for all } \mathbf{k} \in \Lambda, \tag{B.1.1}$$

where $\mathbb{T} := (-\pi, \pi]$, $\boldsymbol{\theta} := (\theta_1, \dots, \theta_d) \in \mathbb{T}^d$, and $(\mathbf{k}, \boldsymbol{\theta}) := \sum_{j=1}^d k_j \theta_j$ is the scalar product in $\mathbb{R}^d$. Here $\Lambda \subset \mathbb{Z}^d$ is a finite index set satisfying $0 \in \Lambda$ and $-\Lambda = \Lambda$. A necessary condition for (B.1.1) to have a solution is that the sequence

$$c := [c_{\mathbf{k}} \mid \mathbf{k} := (k_1, \dots, k_d) \in \Lambda] \tag{B.1.2}$$

satisfy the symmetry condition $c_{-\mathbf{k}} = \bar{c}_{\mathbf{k}}$, where $^-$ denotes the complex conjugate. The space of sequences (B.1.2) with this symmetry will be denoted by $\mathfrak{C}$ and will be represented by vectors c, formed by ordering the coefficient in some prescribed manner, e.g., lexicographical. Note that $\mathfrak{C}$ is isomorphic to $\mathbb{R}^{|\Lambda|}$, where $|\Lambda|$ is the cardinality of Λ. However, as we shall see below, not all $c \in \mathfrak{C}$ are *bona fide* moments for nonnegative measures $d\mu$.

In many of the applications mentioned above there is a natural complexity constraint prescribed by design specifications. In the context of finite-dimensional systems these constraints often arise in the requirement that transfer functions be rational. This leads to the *rational covariance extension problem* (RCEP), which has been studied in various degrees of generality in [25, 26, 36, 52, 53] and can be posed as follows.

Define $e^{i\boldsymbol{\theta}} := (e^{i\theta_1}, \dots, e^{i\theta_d})$ and let

$$d\mu(\boldsymbol{\theta}) = \Phi(e^{i\boldsymbol{\theta}})dm(\boldsymbol{\theta}) + d\nu(\boldsymbol{\theta}), \tag{B.1.3a}$$

be the (unique) Lebesgue decomposition of $d\mu$ (see, e.g., [56, p. 121]), where

$$dm(\boldsymbol{\theta}) := (1/2\pi)^d \prod_{j=1}^{d} d\theta_j$$

is the (normalized) Lebesgue measure and $d\nu$ is a singular measure. Then given a $c \in \mathfrak{C}$, we are interested in parameterizing solutions to (B.1.1) such that the absolutely continuous part of the measure (B.1.3a) takes the form

$$\Phi(e^{i\boldsymbol{\theta}}) = \frac{P(e^{i\boldsymbol{\theta}})}{Q(e^{i\boldsymbol{\theta}})}, \quad p, q \in \bar{\mathfrak{P}}_+ \backslash \{0\}, \tag{B.1.3b}$$

where $\bar{\mathfrak{P}}_+$ is the closure of the convex cone $\mathfrak{P}_+$ of the coefficients $p \in \mathfrak{C}$ corresponding to trigonometric polynomials

$$P(e^{i\boldsymbol{\theta}}) = \sum_{\mathbf{k}\in\Lambda} p_{\mathbf{k}} e^{-i(\mathbf{k},\boldsymbol{\theta})}, \quad p_{-\mathbf{k}} = \bar{p}_{\mathbf{k}} \tag{B.1.4}$$

that are positive for all $\boldsymbol{\theta} \in \mathbb{T}^d$.

The reason for referring to this problem as a rational covariance extension problem is that the numbers (B.1.2) correspond to covariances $c_{\mathbf{k}} := \mathbb{E}\{y(\boldsymbol{t}+\mathbf{k})\overline{y(\boldsymbol{t})}\}$ of a discrete-time, zero-mean, and homogeneous[1] stochastic process $\{y(\boldsymbol{t});\, \boldsymbol{t} \in \mathbb{Z}^d\}$. The corresponding power spectrum, representing the energy distribution across frequencies, is defined as the nonnegative measure $d\mu$ on $\mathbb{T}^d$ whose Fourier coefficients are the covariances (B.1.2). A scalar version of this problem ($d = 1$) was first posed by Kalman [34] and has been extensively studied and solved in the literature [24, 12, 6, 21, 48, 13, 41, 7, 61, 47]. It has been generalized to more general scalar

[1] Homogeneity generalizes stationarity in the case $d = 1$.

moment problems [8, 27, 9] and to the multidimensional setting [26, 25, 53, 52, 36]. Also worth mentioning here is work by Lang and McClellan [39, 40, 45, 46, 38, 37] considering the multidimensional maximum entropy problem, which hence has certain overlap with the above literature.

The multidimensional RCEP posed above has a solution if and only if $c \in \mathfrak{C}_+$, where $\mathfrak{C}_+$ is the open convex cone

$$\mathfrak{C}_+ := \left\{ c \mid \langle c, p \rangle > 0, \quad \text{for all } p \in \bar{\mathfrak{P}}_+ \setminus \{0\} \right\},$$

where $\langle c, p \rangle := \sum_{\mathbf{k} \in \Lambda} c_{\mathbf{k}} \bar{p}_{\mathbf{k}}$ is the inner product in $\mathfrak{C}$ (Theorem B.2.4). However, the covariances $[c_{\mathbf{k}} \mid \mathbf{k} := (k_1, \ldots, k_d) \in \Lambda]$ are generally determined from statistical data. Therefore the condition $c \in \mathfrak{C}_+$ may not be satisfied, and testing this condition is difficult in the multidimensional case. Therefore, we may want to find a positive measure $d\mu$ and a corresponding $r \in \mathfrak{C}_+$, namely

$$r_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} d\mu(\boldsymbol{\theta}), \quad \mathbf{k} \in \Lambda, \tag{B.1.5}$$

so that r is close to c in some norm, e.g., the Euclidean norm $\| \cdot \|_2$. This is an ill-posed inverse problem which in general has an infinite number of solutions $d\mu$. As we already mentioned, we are interested in rational solutions (B.1.3), and to obtain such solutions we use regularization as in [53]. Hence, we seek a $d\mu$ that minimizes

$$\lambda \mathbb{D}(Pdm, d\mu) + \frac{1}{2} \|r - c\|_2^2$$

subject to (B.1.5), where $\lambda > 0$ is a regularization parameter and

$$\mathbb{D}(Pdm, d\mu) := \int_{\mathbb{T}^d} \left(P \log \frac{P}{\Phi} dm + d\mu - Pdm \right) \tag{B.1.6}$$

is the nomalized Kullback-Leibler divergence [33, Chp. 4] [15, 61]. As will be explained in Section B.2, $\mathbb{D}(Pdm, d\mu)$ is always nonnegative and has the property $\mathbb{D}(Pdm, Pdm) = 0$.

In this paper we shall consider a more general problem in the spirit of [22]. To this end, for any Hermitian, positive definite matrix M, we define the weighted vector norm $\|x\|_M := (x^* M x)^{1/2}$, where * denotes the conjugate transpose, and consider the problem

$$\begin{aligned} \min_{d\mu \geq 0,\, r} \quad & \mathbb{D}(Pdm, d\mu) + \frac{1}{2} \|r - c\|_{W^{-1}}^2 \\ \text{subject to} \quad & r_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} d\mu(\boldsymbol{\theta}), \quad \mathbf{k} \in \Lambda, \end{aligned} \tag{B.1.7}$$

which is the same as the problem above with $W = \lambda I$. However, since the space $\mathfrak{C}$ has a certain symmetry, we will also limit the matrices W^{-1} in the weighted

norm to respect this symmetry. In particular, this means that, in addition to being Hermitian positive definite, we will assume that W^{-1} maps $\mathfrak{C}$ into $\mathfrak{C}$. The latter condition corresponds to W^{-1} being Hermitian centrosymmetric with respect to the index set Λ, i.e., $[W^{-1}]_{-k,-\ell} = \overline{[W^{-1}]_{k,\ell}}$ for all $k, \ell \in \Lambda$. This will be assumed throughout the rest of this paper, and we shall refer to such W as a *weight matrix*.

Using the same principle as in [57], we shall also consider the problem of minimizing $\mathbb{D}(Pdm, d\mu)$ subject to (B.1.5) and the hard constraint

$$\|r - c\|^2 \leq \lambda\,. \tag{B.1.8}$$

Since (B.1.5) are *bona fide* moments and hence $r \in \mathfrak{C}_+$, while $c \notin \mathfrak{C}_+$ in general, this problem will not have a solution if the distance from c to $\mathfrak{C}_+$ is greater than $\sqrt{\lambda}$. Hence the choice of λ must be made with some care. Analogously with the *rational covariance extension with soft constraints* in (B.1.7), we shall consider the more general problem

$$\begin{aligned} \min_{d\mu \geq 0,\, r} \quad & \mathbb{D}(Pdm, d\mu) \\ \text{subject to} \quad & r_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} d\mu(\boldsymbol{\theta}), \quad \mathbf{k} \in \Lambda, \\ & \|r - c\|^2_{W^{-1}} \leq 1, \end{aligned} \tag{B.1.9}$$

which we shall refer to as the *rational covariance extension problem with hard constraints*. Again, this problem reduces to the simpler problem by setting $W = \lambda I$.

As we shall see, the soft-constrained problem (B.1.7) always has a solution, while the hard-constrained problem (B.1.9) may fail to have a solution for some weight matrices W. However, in Section B.7 we show that the two problems are in fact equivalent in the sense that whenever (B.1.9) has a solution there is a corresponding W in (B.1.7) that gives the same solution, and any solution of (B.1.7) can also be obtained from (B.1.9) by a suitable choice of W. The reason for considering both formulations is that one formulation might be more suitable than the other for the particular application at hand. For example, an absolute error estimate for the covariances is more naturally incorporated in the formulation with hard constraints. A possible choice of the weight matrix W in either formulation would be the covariance matrix of the estimated moments, as suggested in [22]. This corresponds to the Mahalanobis distance and could be a natural way to incorporate uncertainty of the covariance estimates in the spectral estimation procedure.

Previous work in this direction can be found in [58, 22, 10, 57, 35], where [58, 35, 10] consider the problem of selecting an appropriate covariances sequence to match in a given confidence region. The two approximation problems considered here are similar to the ones considered in [57] and [22], respectively. (For more details, also see [3, Chp. B].)

We begin in Section B.2 by reviewing the regular multidimensional RCEP for exact covariance matching in a broader perspective. In Section B.3 we present our main results on approximate rational covariance extension with soft constraints,

and in Section B.4 we show that the dual solution is well-posed. In Section B.5 we investigate conditions under which there are solutions without a singular part. The approximate rational covariance extension with hard constraints is considered in Section B.6, and in Section B.7 we establish a homeomorphism between the weight matrices in the two problems, showing that the problems are actually equivalent when solutions exist. We also show that under certain conditions the homeomorphism can be extended to hold between all sets of parameters, allowing us to carry over results from the soft-constrained setting to the hard-constrained one. In Section B.8 we discuss the properties of various covariance estimators, in Section B.9 we give a two-dimensional example from spectral estimation, and in Section B.10 we apply our theory to system identification and texture reconstruction. Some of the results of this paper were announced in [54] without proofs.

B.2 Rational covariance extension with exact matching

The trigonometric moment problem of determining a positive measure $d\mu$ satisfying (B.1.1) is an inverse problem that has a solution if and only if $c \in \bar{\mathfrak{C}}_+$ [36, Thm. 2.3], where $\bar{\mathfrak{C}}_+$ is the closure of $\mathfrak{C}_+$, and then in general it has infinitely many solutions. However, the nature of possible rational solutions (B.1.3) will depend on the location of c in $\bar{\mathfrak{C}}_+$. To clarify this point we need the following lemma.

Lemma B.2.1. $\bar{\mathfrak{P}}_+ \setminus \{0\} \subset \mathfrak{C}_+$.

Proof. Obviously the inner product $\langle q, p\rangle := \sum_{\mathbf{k}\in\Lambda} q_{\mathbf{k}}\bar{p}_{\mathbf{k}}$ can be expressed in the integral form

$$\langle q, p\rangle = \int_{\mathbb{T}^d} Q(e^{i\boldsymbol{\theta}})\overline{P(e^{i\boldsymbol{\theta}})}dm(\boldsymbol{\theta}), \tag{B.2.1}$$

and therefore $\langle q, p\rangle > 0$ for all $q, p \in \bar{\mathfrak{P}}_+ \setminus \{0\}$, as P and Q can have zeros only on sets of measure zero. Hence the statement of the lemma follows. □

Therefore, under certain particular conditions, the multidimensional RCEP has a very simple solution with a polynomial spectral density, namely

$$d\mu = P(e^{i\boldsymbol{\theta}})dm(\boldsymbol{\theta}), \quad p \in \bar{\mathfrak{P}}_+\setminus\{0\}. \tag{B.2.2}$$

Proposition B.2.2. *The multidimensional RCEP has a unique polynomial solution* (B.2.2) *if and only if* $c \in \bar{\mathfrak{P}}_+\setminus\{0\}$*, namely* $P = C$*, where*

$$C(e^{i\boldsymbol{\theta}}) := \sum_{\mathbf{k}\in\Lambda} c_{\mathbf{k}} e^{-i(\mathbf{k},\boldsymbol{\theta})}.$$

The proof of Proposition B.2.2 is immediate by noting that any such C is a *bona fide* spectral density and noting that $c_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})}C(e^{i\boldsymbol{\theta}})dm(\boldsymbol{\theta})$.

As seen from the following result presented in [36, Sec. 6], the other extreme occurs for $c \in \partial\mathfrak{C}_+ := \bar{\mathfrak{C}}_+ \setminus \mathfrak{C}_+$, when only singular solutions exist.

Proposition B.2.3. *For any $c \in \partial\mathfrak{C}_+$ there is a solution $d\mu$ of* (B.1.1) *with support in at most $|\Lambda| - 1$ points. There is no solution with an absolutely continuous part Φdm.*

However, for any $c \in \mathfrak{C}_+$, there is a rational solution (B.1.3) parametrized by $p \in \bar{\mathfrak{P}}_+ \backslash \{0\}$, as demonstrated in [53] by considering a primal-dual pair of convex optimization problems. In that paper the primal problem is a weighted maximum entropy problem, but as also noted in [53, Sec. 3.2], it is equivalent to

$$\begin{aligned} \min_{d\mu \geq 0} \quad & \int_{\mathbb{T}^d} P \log \frac{P}{\Phi} dm(\boldsymbol{\theta}) \\ \text{subject to} \quad & c_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} d\mu(\boldsymbol{\theta}), \quad \mathbf{k} \in \Lambda, \end{aligned} \tag{B.2.3}$$

where Φdm is the absolutely continuous part of $d\mu$. This amounts to minimizing the (regular) Kullback-Leibler divergence between Pdm and $d\mu$, subject to $d\mu$ matching the given data [27, 53]. In the present case of exact covariance matching, this problem is equivalent to minimizing (B.1.6) subject to (B.1.1), since P is fixed and the total mass of $d\mu$ is determined by the 0:th moment $c_{\mathbf{0}} = \int_{\mathbb{T}^d} d\mu$. Hence both $\int_{\mathbb{T}^d} d\mu$ and $\int_{\mathbb{T}^d} Pdm$ are constants in this case. Hence problem (B.1.7) and problem (B.1.9) are natural extensions of (B.2.3) for the case where the covariance sequence is not known exactly.

The primal problem (B.2.3) is a problem in infinite dimensions, but with a finite number of constraints. The dual to this problem will then have a finite number of variables but an infinite number of constraints and is given by

$$\min_{q \in \bar{\mathfrak{P}}_+} \quad \langle c, q \rangle - \int_{\mathbb{T}^d} P \log Q \, dm(\boldsymbol{\theta}). \tag{B.2.4}$$

In particular, Theorem 2.1 in [53], based on corresponding analysis in [36], reads as follows.

Theorem B.2.4. *Problem* (B.2.3) *has a solution if and only if $c \in \mathfrak{C}_+$. For every $c \in \mathfrak{C}_+$ and $p \in \bar{\mathfrak{P}}_+ \setminus \{0\}$ the functional in* (B.2.4) *is strictly convex and has a unique minimizer $\hat{q} \in \bar{\mathfrak{P}}_+ \setminus \{0\}$. Moreover, there exists a unique $\hat{c} \in \partial\mathfrak{C}_+$ and a (not necessarily unique) nonnegative singular measure $d\hat{\nu}$ with support*

$$\operatorname{supp}(d\hat{\nu}) \subseteq \{\boldsymbol{\theta} \in \mathbb{T}^d \mid \hat{Q}(e^{i\boldsymbol{\theta}}) = 0\} \tag{B.2.5}$$

such that

$$c_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} \left(\frac{P}{\hat{Q}} dm + d\hat{\nu} \right), \quad \mathbf{k} \in \Lambda, \tag{B.2.6a}$$

$$\hat{c}_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} d\hat{\nu}, \quad \mathbf{k} \in \Lambda. \tag{B.2.6b}$$

For any such $d\hat{\nu}$, the measure

$$d\hat{\mu}(\boldsymbol{\theta}) = \frac{P(e^{i\boldsymbol{\theta}})}{\hat{Q}(e^{i\boldsymbol{\theta}})} dm(\boldsymbol{\theta}) + d\hat{\nu}(\boldsymbol{\theta}) \tag{B.2.7}$$

is an optimal solution to the problem (B.2.3)*. Moreover, $d\hat{\nu}$ can be chosen with support in at most $|\Lambda| - 1$ points, where $|\Lambda|$ is the cardinality of the index set Λ.*

If $c \in \partial\mathfrak{C}_+$, only a singular measure with finite support would match the moment condition (Proposition B.2.3). In this case, the problem (B.2.3) makes no sense, since any feasible solution has infinite objective value.

In [36] we also derived the KKT conditions

$$\hat{q} \in \bar{\mathfrak{P}}_+, \quad \hat{c} \in \partial\mathfrak{C}_+, \quad \langle \hat{c}, \hat{q} \rangle = 0 \tag{B.2.8a}$$

$$c_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} \frac{P}{\hat{Q}} dm + \hat{c}_{\mathbf{k}}, \quad \mathbf{k} \in \Lambda, \tag{B.2.8b}$$

which are necessary and sufficient for optimality of the primal and dual problems.

Since (B.2.3) is an inverse problem, we are interested in how the solution depends on the parameters of the problem. From Propositions 7.3 and 7.4 in [36] we have the following result.

Proposition B.2.5. *Let c, p and $\hat{q}$ be as in Theorem B.2.4. Then the map $(c, p) \mapsto \hat{q}$ is continuous.*

To get a full description of well-posedness of the solution we would like to extend this continuity result to the map $(c, p) \mapsto (\hat{q}, \hat{c})$. However, such a generalization is only possible under certain conditions. The following result is a consequence of Proposition B.2.5 and [53, Cor. 2.3].

Proposition B.2.6. *Let c, p, $\hat{q}$ and $\hat{c}$ be as in Theorem B.2.4. Then, for $d \leq 2$ and all $(c, p) \in \mathfrak{C}_+ \times \mathfrak{P}_+$, the mapping $(c, p) \to (\hat{q}, \hat{c})$ is continuous.*

Corollary 2.3 in [53] actually ensures that $\hat{c} = 0$ for $d \leq 2$ and $p \in \mathfrak{P}_+$. However, in Section B.4 we present a generalization of Proposition B.2.6 to cases with $d \geq 3$, where then $\hat{c}$ may be nonzero. (The proof of this generalization can be found in [55].) Here we shall also consider an example where continuity fails when p belongs to the boundary $\partial\mathfrak{P}_+ := \bar{\mathfrak{P}}_+ \setminus \mathfrak{P}_+$, i.e., the corresponding nonnegative trigonometric polynomial $P(e^{i\boldsymbol{\theta}})$ is zero in at least one point.

B.3 Approximate covariance extension with soft constraints

To handle the case with noisy covariance data, when c may not even belong to $\mathfrak{C}_+$, we relax the exact covariance matching constraint (B.1.1) in the primal problem (B.2.3) to obtain the problem (B.1.7). In this case it is natural to reformulate the objective function in (B.2.3) to include a term that also accounts for changes in the

total mass of $d\mu$. Consequently, we have exchanged the objective function in (B.2.3) by the normalized Kullback-Leibler divergence (B.1.6) plus a term that ensures approximate data matching.

Using the normalized Kullback-Leibler divergence, as proposed in [33, Chp. 4] [15, 61], is an advantage in the approximate covariance matching problem since this divergence is always nonnegative, precisely as is the case for probability densities. To see this, observe that, in view of the basic inequality $x - 1 \geq \log x$,

$$\begin{aligned}\mathbb{D}(Pdm, d\mu) &= \int_{\mathbb{T}^d} \left(P \left(-\log \frac{\Phi}{P} \right) dm + d\mu - Pdm \right) \\ &\geq \int_{\mathbb{T}^d} \left(P(1 - \frac{\Phi}{P})dm + \Phi dm - Pdm \right) + \int_{\mathbb{T}^d} d\nu \geq 0,\end{aligned}$$

since $d\nu$ is a nonnegative measure. Moreover, $\mathbb{D}(Pdm, Pdm) = 0$, as can be seen by taking $d\mu = Pdm$ in (B.1.6).

The problem under consideration is to find a nonnegative measure $d\mu = \Phi dm + d\nu$ minimizing

$$\mathbb{D}(Pdm, d\mu) + \frac{1}{2}\|r - c\|^2_{W^{-1}}$$

subject to (B.1.5). To derive the dual of this problem we consider the corresponding maximization problem and form the Lagrangian

$$\begin{aligned}\mathcal{L}(\Phi, d\nu, r, q) = & -\mathbb{D}(Pdm, d\mu) - \frac{1}{2}\|r - c\|^2_{W^{-1}} + \sum_{\mathbf{k}\in\Lambda} \bar{q}_{\mathbf{k}} \Big(r_{\mathbf{k}} - \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} d\mu(\boldsymbol{\theta}) \Big) \\ = & -\mathbb{D}(Pdm, d\mu) - \frac{1}{2}\|r - c\|^2_{W^{-1}} + \langle r, q \rangle - \int_{\mathbb{T}^d} Q d\mu \,,\end{aligned}$$

where $q := [q_{\mathbf{k}} \mid \mathbf{k} := (k_1, \ldots, k_d) \in \Lambda]$ are Lagrange multipliers and Q is the corresponding trigonometric polynomial (B.1.4). However,

$$\mathbb{D}(Pdm, d\mu) = \int_{\mathbb{T}^d} P(\log P - 1)dm - \int_{\mathbb{T}^d} P \log \Phi dm + r_{\mathbf{0}} \,, \tag{B.3.1}$$

and therefore

$$\begin{aligned}\mathcal{L}(\Phi, d\nu, r, q) = & \int_{\mathbb{T}^d} P \log \Phi dm - \int_{\mathbb{T}^d} Q\Phi dm - \int_{\mathbb{T}^d} Q d\nu - \int_{\mathbb{T}^d} P(\log P - 1)dm \\ & + \langle r, q - e \rangle - \frac{1}{2}\|r - c\|^2_{W^{-1}} \,,\end{aligned} \tag{B.3.2}$$

where $e := [e_{\mathbf{k}}]_{\mathbf{k}\in\Lambda}$, $e_{\mathbf{0}} = 1$ and $e_{\mathbf{k}} = 0$ for $\mathbf{k} \in \Lambda \setminus \{\mathbf{0}\}$, and hence $r_{\mathbf{0}} = \langle r, e \rangle$.

In deriving the dual functional

$$\varphi(q) = \sup_{\Phi\geq 0, d\nu \geq 0, r} \mathcal{L}(\Phi, d\nu, r, q),$$

to be minimized, we only need to consider $q \in \bar{\mathfrak{P}}_+ \setminus \{0\}$, as φ will take infinite values for $q \notin \bar{\mathfrak{P}}_+$. In fact, following along the lines of [53, p. 1957], we note that if $Q(e^{i\boldsymbol{\theta}_0}) < 0$, then (B.3.2) will tend to infinity when $\nu(\boldsymbol{\theta}_0) \to \infty$. Moreover, since $p \in \bar{\mathfrak{P}}_+ \setminus \{0\}$, there is a neighborhood where $P(e^{i\boldsymbol{\theta}}) > 0$; letting Φ tend to infinity in this neighborhood, (B.3.2) will tend to infinity if $Q \equiv 0$. We also note that the nonnegative function Φ can only be zero on a set of measure zero; otherwise, the first term in (B.3.2) will be $-\infty$.

The directional derivative[2] of the Lagrangian (B.3.2) in any feasible direction $\delta\Phi$, i.e., any direction $\delta\Phi$ such that $\Phi + \varepsilon\delta\Phi \geq 0$ for sufficiently small $\varepsilon > 0$, is easily seen to be

$$\delta\mathcal{L}(\Phi, d\nu, r, q; \delta\Phi) = \int_{\mathbb{T}^d} \left(\frac{P}{\Phi} - Q \right) \delta\Phi dm.$$

In particular, the direction $\delta\Phi := \Phi \operatorname{sign}(P - Q\Phi)$ is feasible since $(1 \pm \varepsilon)\Phi \geq 0$ for $0 < \varepsilon < 1$. Therefore, any maximizing Φ must satisfy $\int_{\mathbb{T}^d} |P - Q\Phi| dm \leq 0$ and hence (B.1.3b). Moreover, a maximizing choice of $d\nu$ will require that

$$\int_{\mathbb{T}^d} Q d\,\nu = 0, \tag{B.3.3}$$

as this nonnegative term can be made zero by the simple choice $d\nu \equiv 0$, and consequently (B.2.5) must hold. Finally, the directional derivative

$$\delta\mathcal{L}(\Phi, d\nu, r, q; \delta r) = \langle \delta r, q - e + W^{-1}(r - c) \rangle$$

is zero for all $\delta r \in \mathfrak{C}$ if

$$r = c + W(q - e). \tag{B.3.4}$$

Inserting this together with (B.1.3b) and (B.3.3) into (B.3.2) then yields the dual functional

$$\varphi(q) = \langle c, q \rangle - \int_{\mathbb{T}^d} P \log Q \, dm + \frac{1}{2} \|q - e\|_W^2 - c_0.$$

Consequently, the dual of the (primal) optimization problem (B.1.7) is equivalent to

$$\min_{q \in \bar{\mathfrak{P}}_+} \quad \langle c, q \rangle - \int_{\mathbb{T}^d} P \log Q \, dm + \frac{1}{2} \|q - e\|_W^2. \tag{B.3.5}$$

Theorem B.3.1. *For every $p \in \bar{\mathfrak{P}}_+ \setminus \{0\}$ the functional in* (B.3.5) *is strictly convex and has a unique minimizer $\hat{q} \in \bar{\mathfrak{P}}_+ \setminus \{0\}$. Moreover, there exists a unique $\hat{r} \in \mathfrak{C}_+$, a unique $\hat{c} \in \partial\mathfrak{C}_+$ and a (not necessarily unique) nonnegative singular measure $d\hat{\nu}$ with support*

$$\operatorname{supp}(d\hat{\nu}) \subseteq \{\boldsymbol{\theta} \in \mathbb{T}^d \mid \hat{Q}(e^{i\boldsymbol{\theta}}) = 0\} \tag{B.3.6}$$

[2] Formally, the Gâteaux differential [44].

such that

$$\hat{r}_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} \left(\frac{P}{\hat{Q}} dm + d\hat{\nu} \right) \quad \textit{for all } \mathbf{k} \in \Lambda, \tag{B.3.7a}$$

$$\hat{c}_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} d\hat{\nu} \quad \textit{for all } \mathbf{k} \in \Lambda\,, \tag{B.3.7b}$$

and the measure

$$d\hat{\mu}(\boldsymbol{\theta}) = \frac{P(e^{i\boldsymbol{\theta}})}{\hat{Q}(e^{i\boldsymbol{\theta}})} dm(\boldsymbol{\theta}) + d\hat{\nu}(\boldsymbol{\theta}) \tag{B.3.8}$$

is an optimal solution to the primal problem (B.1.7)*. Moreover, $d\hat{\nu}$ can be chosen with support in at most $|\Lambda| - 1$ points.*

Proof. The objective functional $\mathbb{J}$ of the dual problem (B.3.5) can be written as the sum of two terms, namely

$$\mathbb{J}_1(q) = \langle \tilde{c}, q \rangle - \int_{\mathbb{T}^d} P \log(Q) dm \quad \text{and} \quad \mathbb{J}_2(q) = \langle c - \tilde{c}, q \rangle + \frac{1}{2} \|q - e\|_W^2\,,$$

where $\tilde{c} \in \mathfrak{C}_+$. The functional $\mathbb{J}_1$ is strictly convex (Theorem B.2.4), and trivially the same holds for $\mathbb{J}_2$ since it is a positive definite quadratic form. Consequently, $\mathbb{J} = \mathbb{J}_1 + \mathbb{J}_2$ is strictly convex, as claimed. Moreover, $\mathbb{J}_1$ is lower semicontinuous [53, Lem. 3.1] with compact sublevel sets $\mathbb{J}_1^{-1}(-\infty, \rho]$ [53, Lem. 3.2]. Likewise, $\mathbb{J}_2$ is continuous with compact sublevel sets. Therefore, $\mathbb{J}$ is lower semicontinuous with compact sublevel sets and therefore has a minimum $\hat{q}$, which must be unique by strict convexity.

In view of (B.3.4), the optimal value of r is given by

$$\hat{r} = c + W(\hat{q} - e) \tag{B.3.9}$$

and is hence unique. Since therefore the linear term $c + W(q - e)$ in the gradient of $\mathbb{J}$ takes the value $\hat{r}$ at the optimal point, the analysis in [53, Sec. 3.1.5] applies with obvious modifications, showing that there is a $\hat{c} \in \bar{\mathfrak{C}}_+$, which then must be unique, such that

$$\hat{r}_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} \frac{P}{\hat{Q}} dm + \hat{c}_{\mathbf{k}}.$$

Moreover, there is a discrete measure $d\hat{\nu}$ with support in at most $|\Lambda| - 1$ points such that (B.3.7b) holds; see, e.g., [36, Prop. 2.4]. Then (B.3.7a) holds as well. In view of (B.3.3),

$$\langle \hat{c}, \hat{q} \rangle = \int_{\mathbb{T}^d} \hat{Q} d\hat{\nu} = 0, \tag{B.3.10}$$

and consequently $\hat{c} \in \partial \mathfrak{C}_+$, and the support of $d\hat{\nu}$ must satisfy (B.3.6).

Finally, let r be given in terms of $d\mu$ by (B.1.5), and let $\mathbb{I}(d\mu)$ be the corresponding primal functional in (B.1.7). Then, for any such $d\mu$,

$$\mathbb{I}(d\mu) = \mathcal{L}(\Phi, d\nu, r, \hat{q}) \leq \mathcal{L}(\hat{\Phi}, d\hat{\nu}, \hat{r}, \hat{q}) = \mathbb{I}(d\hat{\mu}),$$

and hence $d\hat{\mu}$ is an optimal solution to the primal problem (B.1.7), as claimed. □

We collect the KKT conditions in the following corollary.

Corollary B.3.2. *The conditions*

$$\hat{q} \in \bar{\mathfrak{P}}_+, \quad \hat{c} \in \partial \mathfrak{C}_+, \quad \langle \hat{c}, \hat{q} \rangle = 0 \tag{B.3.11a}$$

$$\hat{r}_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} \frac{P}{\hat{Q}} dm + \hat{c}_{\mathbf{k}}, \quad \mathbf{k} \in \Lambda \tag{B.3.11b}$$

$$\hat{r} - c = W(\hat{q} - e). \tag{B.3.11c}$$

are necessary and sufficient conditions for optimality of the dual pair (B.1.7) *and* (B.3.5) *of optimization problems.*

B.4 On the well-posedness of the soft-constrained problem

In the previous sections we have shown that the primal and dual optimization problems are well-defined. Next we investigate the well-posedness of the primal problem as an inverse problem. Thus, we first establish the continuity of the solutions $\hat{q}$ in terms of the parameters W, c, and p.

Continuity of $\hat{q}$ with respect to c, p and W

We start considering the continuity of the optimal solution with respect to the parameters. The parameter set of interest is

$$\mathcal{P} = \{(c, p, W) \mid c \in \mathfrak{C}, p \in \bar{\mathfrak{P}}_+ \setminus \{0\}, W > 0\}. \tag{B.4.1}$$

Theorem B.4.1. *Let*

$$\mathbb{J}_{c,p,W}(q) = \langle c, q \rangle - \int_{\mathbb{T}^d} P \log Q \, dm + \frac{1}{2} \|q - e\|_W^2. \tag{B.4.2}$$

Then the map $(c, p, W) \mapsto \hat{q} := \arg\min_{q \in \bar{\mathfrak{P}}_+} \mathbb{J}_{c,p,W}(q)$ *is continuous on* $\mathcal{P}$.

Proof. Following the procedure in [36, Prop. 7.3] we use the continuity of the optimal value (Lemma B.12.1) to show the continuity of the optimal solution. To this end, let $(c^{(k)}, p^{(k)}, W^{(k)})$ be a sequence of parameters in $\mathcal{P}$ converging to $(c, p, W) \in \mathcal{P}$ as $k \to \infty$. Moreover, defining $\mathbb{J}_k(q) := \mathbb{J}_{c^{(k)},p^{(k)},W^{(k)}}(q)$ and $\mathbb{J}(q) := \mathbb{J}_{c,p,W}(q)$ for simplicity of notation, let $\hat{q}_k = \arg\min_{q \in \bar{\mathfrak{P}}_+} \mathbb{J}_k(q)$ and $\hat{q} = \arg\min_{q \in \bar{\mathfrak{P}}_+} \mathbb{J}(q)$. By Lemma B.12.1, $(\hat{q}_k)$ is bounded, and hence there is a subsequence, which for simplicity we also call $(\hat{q}_k)$, converging to a limit q_∞. If we can show that $q_\infty = \hat{q}$, then the theorem follows. To this end, choosing a $q_0 \in \mathfrak{P}_+$, we have

$$\begin{aligned}
\mathbb{J}_k(\hat{q}_k) &= \mathbb{J}_k(\hat{q}_k + \varepsilon q_0) - \langle c^{(k)}, \varepsilon q_0 \rangle + \int_{\mathbb{T}^d} P^{(k)} \log \left(\frac{\hat{Q}_k + \varepsilon Q_0}{\hat{Q}_k} \right) dm \\
&\quad + \frac{1}{2} \|\hat{q}_k - e\|_{W^{(k)}}^2 - \frac{1}{2} \|\hat{q}_k + \varepsilon q_0 - e\|_{W^{(k)}}^2 \\
&\geq \mathbb{J}_k(\hat{q}_k + \varepsilon q_0) - \langle c^{(k)}, \varepsilon q_0 \rangle + \frac{1}{2} \|\hat{q}_k - e\|_{W^{(k)}}^2 - \frac{1}{2} \|\hat{q}_k + \varepsilon q_0 - e\|_{W^{(k)}}^2.
\end{aligned}$$

Consequently, by Lemma B.12.1,

$$\mathbb{J}(\hat{q}) = \lim_{k\to\infty} \mathbb{J}_k(\hat{q}_k) \geq \lim_{k\to\infty} \mathbb{J}_k(\hat{q}_k+\varepsilon q_0) - \varepsilon\langle c^{(k)}, q_0\rangle + \frac{1}{2}\|\hat{q}_k - e\|^2_{W^{(k)}} - \frac{1}{2}\|\hat{q}_k+\varepsilon q_0 - e\|^2_{W^{(k)}}.$$

However $\hat{q}_k + \varepsilon q_0 \in \mathfrak{P}_+$, and, since $(c, p, W, q) \mapsto \mathbb{J}_{c,p,W}(q)$ is continuous in $\mathcal{P} \times \mathfrak{P}_+$, we obtain

$$\begin{aligned}\mathbb{J}(\hat{q}) &\geq \lim_{k\to\infty} \left(\mathbb{J}_k(\hat{q}_k + \varepsilon q_0) - \varepsilon\langle c^{(k)}, q_0\rangle + \frac{1}{2}\|\hat{q}_k - e\|^2_{W^{(k)}} - \frac{1}{2}\|\hat{q}_k + \varepsilon q_0 - e\|^2_{W^{(k)}} \right) \\ &= \mathbb{J}(q_\infty + \varepsilon q_0) - \varepsilon\langle c, q_0\rangle + \frac{1}{2}\|q_\infty - e\|^2_W - \frac{1}{2}\|q_\infty + \varepsilon q_0 - e\|^2_W. \qquad \text{(B.4.3)}\end{aligned}$$

Letting $\varepsilon \to 0$ in (B.4.3), we obtain the inequality $\mathbb{J}(\hat{q}) \geq \mathbb{J}(q_\infty)$. By strict convexity of $\mathbb{J}$ the optimal solution is unique, and hence $\hat{q} = q_\infty$. □

Continuity of $\hat{c}$ with respect to $\hat{q}$

We have now established the continuity from (c, p, W) to $\hat{q}$. In the same way as in Proposition B.2.6 we are also interested in the continuity of the map $(c, p, W) \mapsto (\hat{q}, \hat{c})$. This would follow if we could show that the map from $\hat{q}$ to $\hat{c}$ is continuous. From the KKT condition (B.3.11c), it is seen that $\hat{r}$ is continuous in c, W, and $\hat{q}$. In view of (B.3.11b), i.e.,

$$\hat{r}_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} \frac{P}{\hat{Q}} dm + \hat{c}_{\mathbf{k}}, \quad \mathbf{k} \in \Lambda$$

the continuity of $\hat{c}$ would follow if $\int_{\mathbb{T}^d} P\hat{Q}^{-1} dm$ is continuous in $(p, \hat{q})$ whenever it is finite. If $p \in \mathfrak{P}_+$, this follows from the continuity of the map $\hat{q} \mapsto \hat{Q}^{-1}$ in $L_1(\mathbb{T}^d)$. For the case $d \leq 2$, this is trivial since if $\int_{\mathbb{T}^d} \hat{Q}^{-1} dm$ is finite, then $\hat{q} \in \mathfrak{P}_+$ and $\hat{Q}$ is bounded away from zero (cf. Proposition B.2.6). However, for the case $d > 2$ the optimal $\hat{q}$ may belong to the boundary $\partial\mathfrak{P}_+$, i.e., $\hat{Q}$ might be zero in some point. The following proposition shows the L_1 continuity of $\hat{q} \mapsto \hat{Q}^{-1}$ for certain cases.

Proposition B.4.2. *For $d \geq 3$, let $\hat{q} \in \bar{\mathfrak{P}}_+$ and suppose that the Hessian $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \hat{Q}$ is positive definite in each point where $\hat{Q}$ is zero. Then $\hat{Q}^{-1} \in L_1(\mathbb{T}^d)$ and the mapping from the coefficient vector $q \in \bar{\mathfrak{P}}_+$ to Q^{-1} is L_1 continuous in the point $\hat{q}$.*

The proof of this proposition is given in [55]. From Propositions B.4.2 and B.2.6 the following continuity result follows directly.

Corollary B.4.3. *For all $c \in \mathfrak{C}, p \in \mathfrak{P}_+, W > 0$, the mapping $(c, p, W) \to (\hat{q}, \hat{c})$ is continuous in any point (c, p, W) for which the Hessian $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \hat{Q}$ is positive definite in each point where $\hat{Q}$ is zero.*

The condition $p \in \mathfrak{P}_+$ is needed since we may have pole-zero cancellations in $P/\hat{Q}$ when $p \in \partial\mathfrak{P}_+$, and then $\int_{\mathbb{T}^d} P/\hat{Q} dm$ may be finite even if $\hat{Q}^{-1} \notin L_1(\mathbb{T}^d)$. The following example shows that this may lead to discontinuities in the map $p \mapsto \hat{c}$ (cf. Example 3.8 in [36]).

Example B.4.4. Let

$$c = \begin{bmatrix}1\\3\\1\end{bmatrix} = \begin{bmatrix}0\\2\\0\end{bmatrix} + \begin{bmatrix}1\\1\\1\end{bmatrix} = \int_{-\pi}^{\pi} \begin{bmatrix}e^{-i\theta}\\1\\e^{i\theta}\end{bmatrix} (2dm + d\nu_0)\,,$$

where $dm = d\theta/2\pi$ and $d\nu_0$ is the singular measure $\delta_0(\theta)d\theta$ with support in $\theta = 0$. Since $d\mu := 2dm + d\nu_0$ is positive, $c \in \bar{\mathfrak{C}}_+$. Moreover, since

$$T_c = \begin{bmatrix}3 & 1\\1 & 3\end{bmatrix} > 0$$

we have that $c \in \mathfrak{C}_+$ (see, e.g., [41, p. 2853]). Thus we know [53, Cor. 2.3] that for each $p \in \mathfrak{P}_+$ we have a unique $\hat{q} \in \mathfrak{P}_+$ such that $P/\hat{Q}$ matches c, and hence $\hat{c} = 0$. However, for $p = 2(-1, 2, -1)'$ we have that $\hat{q} = (-1, 2, -1)'$ and $\hat{c} = (1, 1, 1)'$ (Theorem B.2.4). Then, for the sequence (p_k), where $p_k = 2(-1, 2 + 1/k, -1) \in \mathfrak{P}_+$, we have $\hat{c}_k = 0$, so

$$\lim_{k\to\infty} \hat{c}_k = \lim_{k\to\infty} \begin{bmatrix}0\\0\\0\end{bmatrix} \neq \begin{bmatrix}1\\1\\1\end{bmatrix},$$

which shows that the mapping $p \to \hat{c}$ is not continuous.

B.5 Tuning to avoid a singular part

In many situations we prefer solutions where there is no singular measure $d\nu$ in the optimal solution. An interesting question is therefore for what prior P and weight W do we obtain $d\hat{\nu} = 0$? The following result provides a sufficient condition.

Proposition B.5.1. *Let $c \in \mathfrak{C}$ and let p be the Fourier coefficients of the prior P. If the weight satisfies*[3]

$$\|W^{-1/2}\|_{2,1} < \|c - p\|_{W^{-1}}^{-1}, \tag{B.5.1}$$

then the optimal solution of (B.1.7) *is of the form*

$$d\hat{\mu} = (P/\hat{Q})dm,$$

i.e., the singular part $d\hat{\nu}$ vanishes.

Remark B.5.2. Note that for a scalar weight, $W = \lambda I$, the bound (B.5.1) simplifies to

$$\lambda > |\Lambda|^{1/2}\|c - p\|_2, \tag{B.5.2}$$

where $|\Lambda|$ is the cardinality of index set Λ.

For the proof of Proposition B.5.1 we need the following lemma.

[3]Here $\|A\|_{2,1} = \max_{c\neq 0} \|Ac\|_1/\|c\|_2$ denotes the subordinate (induced) matrix norm.

Lemma B.5.3. *Condition* (B.5.1) *implies*

$$\|W^{-1}(\hat{r}-c)\|_1 < 1, \tag{B.5.3}$$

where $\hat{r}$ *is the optimal value of* r *in problem* (B.1.7).

Proof. Let

$$\mathbb{I}(d\mu, r) := \mathbb{D}(Pdm, d\mu) + \frac{1}{2}\|r-c\|^2_{W^{-1}} \tag{B.5.4}$$

be the cost function of problem (B.1.7), and let $(d\hat{\mu}, \hat{r})$ be the optimal solution. Clearly, $\mathbb{I}(Pdm, p) \geq \mathbb{I}(d\hat{\mu}, \hat{r})$, and consequently

$$\|\hat{r}-c\|_{W^{-1}} \leq \|p-c\|_{W^{-1}},$$

since $\mathbb{D}(Pdm, d\hat{\mu}) \geq 0$ and $\mathbb{D}(Pdm, Pdm) = 0$. Therefore,

$$\begin{aligned}\|W^{-1}(\hat{r}-c)\|_1 &\leq \|W^{-1/2}\|_{2,1}\|W^{-1/2}(\hat{r}-c)\|_2 \\ &= \|W^{-1/2}\|_{2,1}\|\hat{r}-c\|_{W^{-1}} \\ &\leq \|W^{-1/2}\|_{2,1}\|p-c\|_{W^{-1}},\end{aligned}$$

which is less than one by (B.5.1). Hence (B.5.1) implies (B.5.3). □

Proof of Proposition B.5.1. Suppose the optimal solution has a nonzero singular part $d\hat{\nu}$, and form the directional derivative of (B.5.4) at $(d\hat{\mu}, \hat{r})$ in the direction $-d\hat{\nu}$. Then Φ in (B.1.3a) does not vary, and

$$\delta\mathbb{I}(d\hat{\mu}, \hat{r}; -d\hat{\nu}, \delta r) = -\int_{\mathbb{T}^d} d\hat{\nu} + \delta r^* W^{-1}(\hat{r}-c),$$

where

$$\delta r_{\mathbf{k}} = -\int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} d\hat{\nu}.$$

Then $|\delta r_{\mathbf{k}}| \leq \int d\hat{\nu}$ for all $\mathbf{k} \in \Lambda$, and hence

$$|\delta r^* W^{-1}(\hat{r}-c)| \leq \|W^{-1}(\hat{r}-c)\|_1 \int_{\mathbb{T}^d} d\hat{\nu} < \int_{\mathbb{T}^d} d\hat{\nu},$$

by (B.5.3) (Lemma B.5.3). Consequently,

$$\delta\mathbb{I}(d\hat{\mu}, \hat{r}; -d\hat{\nu}, \delta r) < 0$$

whenever $d\hat{\nu} \neq 0$, which contradicts optimality. Hence $d\hat{\nu}$ must be zero. □

The condition of Proposition B.5.1 is just sufficient and is in general conservative. To illustrate this, we consider a simple one-dimensional example ($d = 1$).

Example B.5.4. Consider a covariance sequence $(1, c_1)$, where $c_1 \neq 0$, and a prior $P(e^{i\theta}) = 1 - \cos\theta$, and set $W = \lambda I$. Then, since

$$c = \begin{pmatrix} c_1 \\ 1 \\ c_1 \end{pmatrix} \quad \text{and} \quad p = \begin{pmatrix} -1/2 \\ 1 \\ -1/2 \end{pmatrix},$$

the sufficient condition (B.5.2) for an absolutely continuous solution is

$$\lambda > \sqrt{\tfrac{3}{2}}\,|1 + 2c_1|. \tag{B.5.5}$$

We want to investigate how restrictive this condition is.

Clearly we will have a singular part if and only if $\hat{Q} = q_0 P$, in which case we have

$$\hat{q} = q_0 \begin{pmatrix} -1/2 \\ 1 \\ -1/2 \end{pmatrix} \quad \text{and} \quad \hat{c} = \beta \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

for some $\beta > 0$. In fact, it follows from $\langle \hat{c}, \hat{q} \rangle = 0$ in (B.3.11a) that $\hat{c}_1 = \hat{c}_0$. Moreover, (B.3.11b) and (B.3.11c) yield

$$\begin{aligned} \hat{r} &= \int \frac{P}{\hat{Q}} \begin{pmatrix} e^{i\theta} \\ 1 \\ e^{-i\theta} \end{pmatrix} dm + \hat{c} = \begin{pmatrix} \beta \\ \beta + 1/q_0 \\ \beta \end{pmatrix}, \\ c &= \hat{r} - \lambda(\hat{q} - e) = \begin{pmatrix} \beta + \lambda q_0/2, \\ \beta + 1/q_0 - \lambda q_0 + \lambda \\ \beta + \lambda q_0/2 \end{pmatrix}. \end{aligned}$$

By eliminating β, we get

$$c_1 = 1 - \frac{1}{q_0} + \frac{3}{2} q_0 \lambda - \lambda,$$

and solving for q_0 yields

$$q_0 = \frac{\lambda + c_1 - 1 + (6\lambda + (\lambda + c_1 - 1)^2)^{1/2}}{3\lambda}$$

(note that $\lambda > 0$ and $q_0 > 0$). Again, using (B.3.11c) we have

$$\begin{aligned} \beta &= c_1 - \lambda q_0/2 \\ &= c_1 - \frac{1}{6}\left(\lambda + c_1 - 1 + (6\lambda + (\lambda + c_1 - 1)^2)^{1/2}\right). \end{aligned}$$

We are interested in λ for which $\beta > 0$, i.e.,

$$6c_1 - (\lambda + c_1 - 1) > (6\lambda + (\lambda + c_1 - 1)^2)^{1/2}, \tag{B.5.6}$$

which is equivalent to the two conditions

$$1 + 5c_1 > \lambda \tag{B.5.7a}$$

$$2c_1(1 + 2c_1) > \lambda(1 + 2c_1), \tag{B.5.7b}$$

which could be seen by noting that the left member of (B.5.6) must be positive and then squaring both sides. To find out whether this has a solution we consider three cases, namely $c_1 < -1/2$, $-1/2 < c_1 < 0$, and $c_1 > 0$. For $c_1 < -1/2$, condition (B.5.7) becomes $2c_1 < \lambda < 1 + 5c_1$, which is impossible since $1 + 5c_1 < 2c_1$ in this region. Condition (B.5.7) cannot be satisfied when $-1/2 < c_1 < 0$ because then λ would be negative, which contradicts $\lambda > 0$. When $c_1 > 0$, condition (B.5.7) is satisfied if and only if $\lambda < 2c_1$.

Consequently, there is no singular part if either c_1 is negative or

$$\lambda \geq 2c_1.$$

This shows that the condition (B.5.5) is not tight.

B.6 Covariance extension with hard constraints

The alternative optimization problem (B.1.9) amounts to minimizing $\mathbb{D}(Pdm, d\mu)$ subject to the hard constraint $\|r - c\|^2_{W^{-1}} \leq 1$, where $r_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} d\mu$. Hard constraints of this type were used in [57] in the context of entropy maximization. In general the data $c \notin \bar{\mathfrak{C}}_+$, whereas, by definition, $r \in \bar{\mathfrak{C}}_+$. Consequently, a necessary condition for the existence of a solution is that $\bar{\mathfrak{C}}_+$ and the strictly convex set

$$\mathfrak{S}_W = \{r \mid \|r - c\|^2_{W^{-1}} \leq 1\} \tag{B.6.1}$$

have a nonempty intersection. In the case that $\mathfrak{S}_W \cap \bar{\mathfrak{C}}_+ \subset \partial\mathfrak{C}_+$, this intersection only contains one point [44, Sec. 3.12]. In this case, any solution to the moment problem contains only a singular part (Proposition B.2.3), and then the primal problem (B.1.9) has a unique feasible point r, but the objective function is infinite. Moreover, $\mathbb{D}(Pdm, d\mu) \geq 0$ is strictly convex with $\mathbb{D}(Pdm, Pdm) = 0$, so if $p \in \mathfrak{S}_W$, then (B.1.9) has the trivial unique optimal solution $d\hat{\mu} = Pdm$, and $\hat{r} = p$. The remaining case, $p \notin \mathfrak{S}_W \cap \mathfrak{C}_+ \neq \emptyset$, needs further analysis.

To this end, setting $d\mu = \Phi dm + d\nu$, we consider the Lagrangian

$$\begin{aligned}
\mathcal{L}(\Phi, d\nu, r, q, \gamma) = & -\mathbb{D}(Pdm, d\mu) + \sum_{\mathbf{k}\in\Lambda} \bar{q}_{\mathbf{k}} \left(r_{\mathbf{k}} - \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} d\mu(\boldsymbol{\theta}) \right) \\
& + \gamma \left(1 - \|r - c\|^2_{W^{-1}}\right) \\
= & -\mathbb{D}(Pdm, d\mu) + \langle r, q \rangle - \int_{\mathbb{T}^d} Q d\mu + \gamma \left(1 - \|r - c\|^2_{W^{-1}}\right),
\end{aligned}$$

where $\gamma \geq 0$. Therefore, in view of (B.3.1),

$$\mathcal{L}(\Phi, d\nu, r, q, \gamma) = \int_{\mathbb{T}^d} P \log \Phi dm - \int_{\mathbb{T}^d} Q\Phi dm - \int_{\mathbb{T}^d} Q d\nu - \int_{\mathbb{T}^d} P(\log P - 1) dm + \langle r, q - e\rangle + \gamma \left(1 - \|r - c\|^2_{W^{-1}}\right), \tag{B.6.2}$$

where, as before, $e := [e_{\mathbf{k}}]_{\mathbf{k}\in\Lambda}$, $e_{\mathbf{0}} = 1$ and $e_{\mathbf{k}} = 0$ for $\mathbf{k} \in \Lambda \setminus \{\mathbf{0}\}$, and hence $r_{\mathbf{0}} = \langle r, e\rangle$. This Lagrangian differs from that in (B.3.2) only in the last term, which does not depend on Φ. Therefore, in deriving the dual functional

$$\varphi(q, \gamma) = \sup_{\Phi\geq 0, d\nu \geq 0, r} \mathcal{L}(\Phi, d\nu, r, q, \gamma),$$

we only need to consider $q \in \bar{\mathfrak{P}}_+ \setminus \{0\}$, and a first variation in Φ yields (B.1.3b) and (B.3.3). The directional derivative

$$\delta\mathcal{L}(\Phi, d\nu, r, q, \gamma; \delta r) = q - e + 2\gamma W^{-1}(r - c)$$

is zero for

$$r = c + \frac{1}{2\gamma} W(q - e). \tag{B.6.3}$$

Thus inserting (B.1.3b) and (B.3.3) and (B.6.3) into (B.6.2) yields the dual functional

$$\varphi(q, \gamma) = \langle c, q\rangle - \int_{\mathbb{T}^d} P \log Q \, dm + \frac{1}{4\gamma}\|q - e\|^2_W + \gamma - c_{\mathbf{0}} \tag{B.6.4}$$

to be minimized over all $q \in \bar{\mathfrak{P}}_+ \setminus \{0\}$ and $\gamma \geq 0$. Since $\frac{d\varphi}{d\gamma} = -\frac{1}{4\gamma^2}\|q - e\|^2_W + 1$, there is a stationary point

$$\gamma = \frac{1}{2}\|q - e\|_W \tag{B.6.5}$$

that is nonnegative as required.

For $\gamma = 0$, considering (B.6.2) we see that the supremum defining $\varphi(q, 0)$ is ∞ if $q \neq e$. However, for $\gamma = 0$ and $q = e$, then what remains in (B.6.2) is only $-\mathbb{D}(Pdm, d\mu)$, and the supremum is thus 0 and attained for $d\mu = Pdm$. This shows that (B.6.5) is optimal to (B.6.4) for all $\gamma \geq 0$, and inserting (B.6.5) into (B.6.4) and removing the constant term $-c_{\mathbf{0}}$ we obtain the modified dual functional

$$\mathbb{J}(q) = \langle c, q\rangle - \int_{\mathbb{T}^d} P \log Q \, dm + \|q - e\|_W. \tag{B.6.6}$$

Moreover, combining (B.6.3) and (B.6.5), we obtain

$$\|r - c\|_{W^{-1}} = 1, \tag{B.6.7}$$

which also follows from strict convexity of $\mathbb{D}(Pdm, d\mu)$ and that $p \notin \mathfrak{S}_W$.

Theorem B.6.1. *Suppose that $p \in \bar{\mathfrak{P}}_+ \setminus \{0\}$, $p \notin \mathfrak{S}_W$, and $\mathfrak{S}_W \cap \mathfrak{C}_+ \neq \emptyset$. Then the modified dual problem*

$$\min_{q \in \bar{\mathfrak{P}}_+} \mathbb{J}(q) \tag{B.6.8}$$

has a unique solution $\hat{q} \in \bar{\mathfrak{P}}_+ \setminus \{0\}$. Moreover, there exists a unique $\hat{r} \in \mathfrak{C}_+$, a unique $\hat{c} \in \partial\mathfrak{C}_+$, and a (not necessarily unique) nonnegative singular measure $d\hat{\nu}$ with support

$$\operatorname{supp}(d\hat{\nu}) \subseteq \{\boldsymbol{\theta} \in \mathbb{T}^d \mid \hat{Q}(e^{i\boldsymbol{\theta}}) = 0\} \tag{B.6.9}$$

such that

$$\hat{r}_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} \left(\frac{P}{\hat{Q}} dm + d\hat{\nu} \right) \quad \text{for all } \mathbf{k} \in \Lambda, \tag{B.6.10a}$$

$$\hat{c}_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} d\hat{\nu} \quad \text{for all } \mathbf{k} \in \Lambda\,, \tag{B.6.10b}$$

and the measure

$$d\hat{\mu}(\boldsymbol{\theta}) = \frac{P(e^{i\boldsymbol{\theta}})}{\hat{Q}(e^{i\boldsymbol{\theta}})} dm(\boldsymbol{\theta}) + d\hat{\nu}(\boldsymbol{\theta}) \tag{B.6.11}$$

is an optimal solution to the primal problem (B.1.9)*. Moreover,*

$$\|\hat{r} - c\|_{W^{-1}} = 1\,, \tag{B.6.12}$$

and $d\hat{\nu}$ can be chosen with support in at most $|\Lambda| - 1$ points.

If $p \in \mathfrak{S}_W$, the unique optimal solution is $d\hat{\mu} = P dm$, and then $\hat{r} = p$. If $\mathfrak{S}_W \cap \bar{\mathfrak{C}}_+ \subset \partial\mathfrak{C}_+$, any solution to the moment problem will have only a singular part. Finally, if $\mathfrak{S}_W \cap \bar{\mathfrak{C}}_+ = \emptyset$, then the problem (B.1.9) *will have no solution.*

Proof. We begin by showing that the functional $\mathbb{J}$ has a minimum under the stated conditions. To this end, we first establish that the functional $\mathbb{J}$ has compact sublevel sets $\mathbb{J}^{-1}(-\infty, \rho]$, i.e., $\|q\|_\infty$ is bounded for all q such that $\mathbb{J}(q) \leq \rho$, where ρ is sufficiently large for the sublevel set to be nonempty. To this end, the functional (B.6.8) can be decomposed as

$$\mathbb{J}(q) = h(q) + \tilde{h}(q) - \int_{\mathbb{T}^d} P \log Q \, dm,$$

where $h(q) := \langle c, q \rangle + \|q\|_W$ and $\tilde{h}(q) := \|q - e\|_W - \|q\|_W$. By the reverse triangle inequality, $|\tilde{h}(q)| \leq \|q - e - q\|_W = \|e\|_W$, and thus $\tilde{h}(q)$ is bounded for all $q \in \bar{\mathfrak{P}}_+$. The integral term will tend to $-\infty$ as $\|q\|_\infty \to \infty$. Therefore, we need to have the term $h(q)$ to tend to $+\infty$ as $\|q\|_\infty \to \infty$, in which case we can appeal to the fact that linear growth is faster than logarithmic growth. However, if $c \notin \bar{\mathfrak{C}}_+$, as is generally assumed, there is a $q \in \bar{\mathfrak{P}}_+$ such that $\langle c, q \rangle < 0$, so we need to ensure that the positive term $\|q\|_W$ dominates.

Let $\tilde{r} \in \mathfrak{S}_W \cap \mathfrak{C}_+ \neq \emptyset$. Then, by Theorem B.2.4, there is a positive measure $d\tilde{\mu} = \tilde{\Phi} dm + d\tilde{\nu}$ with a nonzero $\tilde{\Phi}$ such that

$$\tilde{r} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} d\tilde{\mu}\,,$$

and $\tilde{r}$ satisfies the constraints in the primal problem (B.1.9). Consequently,

$$\varphi(q,\gamma) \geq \mathcal{L}(\tilde{\Phi}, d\tilde{\nu}, \tilde{r}, q, \gamma) \geq -\mathbb{D}(Pdm, d\tilde{\mu})$$

for all $q \in \bar{\mathfrak{P}}_+$ and $\gamma \geq 0$, which in particular implies that

$$\mathbb{J}(q) \geq -\mathbb{D}(Pdm, d\tilde{\mu}) \quad \text{for all } q \in \bar{\mathfrak{P}}_+. \tag{B.6.13}$$

Now, if there is a $q \in \bar{\mathfrak{P}}_+ \setminus \{0\}$ such that $h(q) \leq 0$, then $\mathbb{J}(\lambda q) \to -\infty$ as $\lambda \to \infty$, which contradicts (B.6.13). Therefore, $h(q) > 0$ for all $q \in \bar{\mathfrak{P}}_+ \setminus \{0\}$. Then, since h is continuous, it has a minimum $\varepsilon > 0$ on the compact set $K := \{q \in \bar{\mathfrak{P}}_+ \mid \|q\|_\infty = 1\}$. Therefore,

$$h(q) = \left(\left\langle c, \frac{q}{\|q\|_\infty} \right\rangle + \left\| \frac{q}{\|q\|_\infty} \right\|_W \right) \|q\|_\infty \geq \varepsilon \|q\|_\infty \geq \frac{\varepsilon}{|\Lambda|} \|Q\|_\infty,$$

since $\|Q\|_\infty \leq |\Lambda| \|q\|_\infty$ [53, Lem. A.1]. Likewise,

$$\begin{aligned} \int_{\mathbb{T}^d} P \log Q dm &= \int_{\mathbb{T}^d} P \log \left[\frac{Q}{\|Q\|_\infty} \right] dm + \int_{\mathbb{T}^d} P \log \|Q\|_\infty dm \\ &\leq \int_{\mathbb{T}^d} P \log \|Q\|_\infty dm\,, \end{aligned}$$

since $Q/\|Q\|_\infty \leq 1$. Hence

$$\rho \geq \mathbb{J}(q) \geq \frac{\varepsilon}{|\Lambda|} \|Q\|_\infty - \int_{\mathbb{T}^d} P \log \|Q\|_\infty dm - \|e\|_W\,. \tag{B.6.14}$$

Comparing linear and logarithmic growth we see that the sublevel set is bounded from above and below. Moreover, a trivial modification of [53, Lem. 3.1] shows that $\mathbb{J}$ is lower semi-continuous, and hence $\mathbb{J}^{-1}(-\infty, \rho]$ is compact. Consequently, the problem (B.6.8) has an optimal solution $\hat{q}$.

We now want to show that $\hat{q} \neq e$ since, in view of (B.6.5), this also means that $\hat{\gamma} > 0$, i.e., we have strict complementarity between the Lagrangian multiplier γ and the constraint $\|r - c\|^2_{W^{-1}} \leq 1$. To this end, we first note that by the assumption $p \notin \mathfrak{S}_W$, we have that $\|c - p\|^2_{W^{-1}} > 1$, and thus also that

$$\|c - p\|^2_{W^{-1}} > \|c - p\|_{W^{-1}}. \tag{B.6.15}$$

Now, consider the point $\tilde{q} = e + \varepsilon W^{-1}(c - p)$, i.e., a small perturbation around $q = e$. For $|\varepsilon|$ small enough, we also have that $\tilde{q} \in \mathfrak{P}_+$. Moreover, in this point the unmodified dual functional $\mathbb{J}(q) - c_{\mathbf{0}}$ takes the value

$$\mathbb{J}(\tilde{q}) - c_{\mathbf{0}} = \varepsilon \langle c, W^{-1}(c - p) \rangle - \int_{\mathbb{T}^d} P \log \tilde{Q} dm + |\varepsilon| \|(c - p)\|_{W^{-1}}.$$

Analyzing the middle term further, if we let the vector of basis functions $[e^{i(\mathbf{k},\boldsymbol{\theta})}]_{\mathbf{k}\in\Lambda}$ be ordered in the same way as elements in $\mathfrak{C}$, then any trigonometric polynomial Q can be written as $Q(e^{i\boldsymbol{\theta}}) = \langle [e^{i(\mathbf{k},\boldsymbol{\theta})}]_{\mathbf{k}\in\Lambda}, q\rangle$. Now, by a series expansion of the logarithm we get that

$$\begin{aligned}\int_{\mathbb{T}^d} P\log\tilde{Q}dm &= \int_{\mathbb{T}^d} P\left(\varepsilon\left\langle [e^{i(\mathbf{k},\boldsymbol{\theta})}]_{\mathbf{k}\in\Lambda}, W^{-1}(c-p)\right\rangle - \mathcal{O}(|\varepsilon|^2)\right)dm \\ &= \varepsilon\langle p, W^{-1}(c-p)\rangle - \mathcal{O}(|\varepsilon|^2),\end{aligned}$$

since P has finite total mass, and by moving the integration into each component of $[e^{i(\mathbf{k},\boldsymbol{\theta})}]_{\mathbf{k}\in\Lambda}$ and using that the complex exponentials are orthogonal. This gives

$$\begin{aligned}\mathbb{J}(\tilde{q}) - c_{\mathbf{0}} &= \varepsilon\langle c, W^{-1}(c-p)\rangle - \varepsilon\langle p, W^{-1}(c-p)\rangle + \mathcal{O}(|\varepsilon|^2) + \|c-p\|_{W^{-1}} \\ &= \varepsilon\|c-p\|^2_{W^{-1}} + |\varepsilon|\|c-p\|_{W^{-1}} + \mathcal{O}(|\varepsilon|^2) \\ &= \varepsilon\left(\|c-p\|^2_{W^{-1}} + \operatorname{sign}(\varepsilon)\|c-p\|_{W^{-1}}\right) + \mathcal{O}(|\varepsilon|^2) \\ &\leq \delta\varepsilon + \mathcal{O}(|\varepsilon|^2)\end{aligned}$$

for some $\delta > 0$, where the last inequality follows from (B.6.15). Thus, for $\varepsilon < 0$ with $|\varepsilon|$ sufficiently small, we have that $\mathbb{J}(\tilde{q}) - c_{\mathbf{0}} < 0$. Since $\mathbb{J}(e) - c_{\mathbf{0}} = 0$, this shows that $q = e$ is not optimal to (B.6.8).

Next we show that $\hat{q}$ is unique. For this we return to the original dual problem to find a minimum of (B.6.4). The solution $\hat{q}$ is a minimizer of $\varphi(q,\hat{\gamma})$, where

$$\hat{\gamma} = \frac{1}{2}\|\hat{q}-e\|_W\,,$$

and $\mathbb{J}(\hat{q}) = \varphi(\hat{q},\hat{\gamma}) + c_{\mathbf{0}}$. To show that φ is strictly convex, we form the Hessian

$$H = \begin{bmatrix}\int_{\mathbb{T}^d} P/Q^2 dm & 0\\ 0 & 0\end{bmatrix} + \frac{1}{2\gamma^3}\begin{bmatrix}\gamma^2 W & -\gamma(q-e)^* W\\ -\gamma W(q-e) & (q-e)^* W(q-e)\end{bmatrix}$$

and the quadratic form

$$\begin{bmatrix}x\\ \xi\end{bmatrix}^* H \begin{bmatrix}x\\ \xi\end{bmatrix} = x^*\left(\int_{\mathbb{T}^d} P/Q^2 dm\right)x + \frac{1}{2\gamma^3}[\gamma x - \xi(q-e)]^* W[\gamma x - \xi(q-e)],$$

which is positive for all nonzero (x,ξ), since $(q-e)\neq 0$ and $\gamma > 0$. Consequently, φ has a unique minimizer $(\hat{q},\hat{\gamma})$, where $\hat{q}$ is the unique minimizer of $\mathbb{J}$.

It follows from (B.6.3) and (B.6.5) that

$$\hat{r} = c + \frac{W(\hat{q}-e)}{\|\hat{q}-e\|_W}\,, \tag{B.6.16}$$

which consequently is unique. Moreover, $h(\hat{q}) = \langle \hat{r}, \hat{q}\rangle - \hat{r}_{\mathbf{0}}$, and hence we can follow along the same lines as the proof of Theorem B.3.1 to show that there is a unique

$\hat{c} \in \partial\mathfrak{C}_+$ such that $\langle \hat{c}, \hat{q} \rangle = 0$ and a positive discrete measure $d\hat{\nu}$ with support in $|\Lambda| - 1$ points so that (B.6.9) and (B.6.10) hold. Next, let $\mathbb{I}(d\mu) = -\mathbb{D}(Pdm, d\mu)$ be the primal functional in (B.1.9), where $d\mu$ is restricted to the set of positive measures $d\mu := \Phi dm + d\nu$ such that r, given by (B.1.5), satisfies the constraint $\|r - c\|_W \leq 1$. In view of (B.6.12),

$$\mathbb{I}(d\mu) = \mathcal{L}(\Phi, d\nu, r, \hat{q}, \hat{\gamma}) \leq \mathcal{L}(\hat{\Phi}, d\hat{\nu}, \hat{r}, \hat{q}, \hat{\gamma}) = \mathbb{I}(d\hat{\mu})$$

for any such $d\mu$, and hence $d\hat{\mu}$ is an optimal solution to the primal problem (B.1.9). Finally, the cases $p \in \mathfrak{S}_W$, $\mathfrak{S}_W \cap \bar{\mathfrak{C}}_+ \subset \partial\mathfrak{C}_+$, and $\mathfrak{S}_W \cap \bar{\mathfrak{C}}_+ = \emptyset$ have already been discussed above. □

Corollary B.6.2. *Suppose that $p \in \bar{\mathfrak{P}}_+ \setminus \{0\}$ and $\mathfrak{S}_W \cap \mathfrak{C}_+ \neq \emptyset$. The KKT conditions*

$$\hat{q} \in \bar{\mathfrak{P}}_+, \quad \hat{c} \in \partial\mathfrak{C}_+, \quad \langle \hat{c}, \hat{q} \rangle = 0 \tag{B.6.17a}$$

$$\hat{r}_{\mathbf{k}} = \int_{\mathbb{T}^d} e^{i(\mathbf{k},\boldsymbol{\theta})} \frac{P}{\hat{Q}} dm + \hat{c}_{\mathbf{k}}, \quad \mathbf{k} \in \Lambda \tag{B.6.17b}$$

$$(\hat{r} - c)\|\hat{q} - e\|_W = W(\hat{q} - e), \quad \hat{r} \in \mathfrak{S}_W \tag{B.6.17c}$$

are necessary and sufficient conditions for optimality of the dual pair (B.1.9) *and* (B.6.8) *of optimization problems.*

The corollary follows by noting that if $p \in \mathfrak{S}_W$, then we obtain the trivial solution $\hat{q} = e$, which corresponds to the primal optimal solution $d\hat{\mu} = Pdm$.

Proposition B.6.3. *The condition*

$$W > cc^* \tag{B.6.18}$$

is sufficient for the pair (B.1.9) *and* (B.6.8) *of dual problems to have optimal solutions.*

Proof. If $W > cc^*$, then $(q - e)^* W (q - e) \geq \langle c, q - e \rangle^2$ with equality only for $q = e$. Hence, if $q \neq e$, then $\|q - e\|_W > |\langle c, q - e \rangle|$, i.e.,

$$h(q) + \tilde{h}(q) - c_{\mathbf{0}} = \langle c, q - e \rangle + \|q - e\|_W > 0$$

for all $q \in \bar{\mathfrak{P}}_+$ except $q = e$. Considering $\mathbb{J}(q) = h(q) + \tilde{h}(q) - c_{\mathbf{0}} - \int_{\mathbb{T}^d} P \log Q \, dm + c_{\mathbf{0}}$, we then proceed as in the proof of Theorem B.6.1. □

Remark B.6.4. Condition (B.6.18) guarantees that $0 \in \mathrm{int}(\mathfrak{S}_W)$ and hence in particular that $\mathfrak{S}_W \cap \mathfrak{C}_+ \neq \emptyset$ as required in Theorem B.6.1. To see this, note that $0 \in \bar{\mathfrak{C}}_+$ and that $r = 0$ satisfies the hard constraint in (B.1.9) if $c^* W^{-1} c \leq 0$. However, since $W > cc^*$, there is a $W_0 > 0$ such that $W = W_0 + cc^*$. Then the well-known matrix inversion lemma (see, e.g., [42, p. 746]) yields

$$(W_0 + cc^*)^{-1} = W_0^{-1} - W_0^{-1} c (1 + c^* W_0^{-1} c)^{-1} c^* W_0^{-1},$$

and therefore

$$c^*W^{-1}c = c^*W_0^{-1}c - c^*W_0^{-1}c(1 + c^*W_0^{-1}c)^{-1}c^*W_0^{-1}c = \frac{c^*W_0^{-1}c}{1 + c^*W_0^{-1}c} < 1,$$

which establishes that $0 \in \operatorname{int}(\mathfrak{S}_W)$. However, for $\mathfrak{S}_W \cap \mathfrak{C}_+$ to be nonempty, $r = 0$ need not be contained in this set. Hence, condition (B.6.18) is not necessary, although it is easily testable. In fact, this provides an alternative proof of Proposition B.6.3.

B.7 On the equivalence between the two problems

Clearly $\mathfrak{S}_W \cap \mathfrak{C}_+$ is always nonempty if $c \in \mathfrak{C}_+$. Then both the problem (B.1.7) with soft constraints and the problem (B.1.9) with hard constraints have a solution for any choice of W. On the other hand, if $c \notin \mathfrak{C}_+$, the problem with soft constraints will always have a solution, while the problem with hard constraints may fail to have one for certain choices of W. However, if the weight matrix in the hard-constrained problem – let us denote it W_{hard} – is chosen in the set $\mathcal{W} := \{W > 0 \mid \mathfrak{S}_W \cap \mathfrak{C}_+ \neq \emptyset, p \notin \mathfrak{S}_W\}$, then it can be seen from Corollaries B.3.2 and B.6.2 that we obtain exactly the same solution $\hat{q}$ in the soft-constrained problem by choosing

$$W_{\text{soft}} = W_{\text{hard}}/\|\hat{q} - e\|_{W_{\text{hard}}}. \tag{B.7.1}$$

We note that (B.7.1) can be written as $W_{\text{hard}} = \alpha W_{\text{soft}}$, where $\alpha := \|\hat{q} - e\|_{W_{\text{hard}}}$. Therefore, substituting W_{hard} in (B.7.1), we obtain

$$W_{\text{soft}} = \frac{\alpha W_{\text{soft}}}{\|\hat{q} - e\|_{\alpha W_{\text{soft}}}} = \alpha^{1/2}\frac{W_{\text{soft}}}{\|\hat{q} - e\|_{W_{\text{soft}}}},$$

which yields $\alpha = \|\hat{q} - e\|^2_{W_{\text{soft}}}$. Hence the inverse of (B.7.1) is given by

$$W_{\text{hard}} = W_{\text{soft}}\|\hat{q} - e\|^2_{W_{\text{soft}}}. \tag{B.7.2}$$

By Theorem B.4.1 $\hat{q}$ is continuous in W_{soft}, and hence, by (B.7.2), the corresponding W_{hard} varies continuously with W_{soft}. In fact, this can be strengthened to a homeomorphism between the two weight matrices.

Theorem B.7.1. *The map* (B.7.1) *is a homeomorphism between $\mathcal{W}$ and the space of all (Hermitian positive definite) weight matrices, and the inverse is given by* (B.7.2).

Proof. By [11, Lem. 2.3], a continuous map between two spaces of the same dimension is a homeomorphism if and only if it is injective and proper, i.e., the preimage of any compact set is compact. To see that $\mathcal{W}$ is open, we observe that $\mathfrak{S}_W$ is continuous in W and that $\mathfrak{C}_+$ is an open set. As noted above, the map (B.7.2) – let us call it f – is continuous and also injective, as it can be inverted. Hence it only remains to show that f is proper. To this end, we take a compact set $K \subset \mathcal{W}$ and

show that $f^{-1}(K)$ is also compact. There are two ways this could fail. First, the preimage could contain a singular semi-definite matrix. However, this is impossible by (B.7.2), since $\|\hat{q}\|_\infty$ is bounded for $W_{\text{hard}} \in K$ (Lemma B.12.2) and a nonzero scaling of a singular matrix cannot be nonsingular. Second, $\|W_{\text{soft}}\|_F$ could tend to infinity. However, this is also impossible. To see this, we first show that there is a $\kappa > 0$ such that $\|p - r\|_{W_{\text{hard}}^{-1}} \geq \kappa$ for all $r \in \mathfrak{S}_{W_{\text{hard}}}$ and all $W_{\text{hard}} \in K$. To this end, we observe that the minimum of $\|p - r\|_{W^{-1}}$ over all $W \in K$ and r satisfying the constraint $\|r - c\|_{W^{-1}} \leq 1$ is bounded by

$$\kappa := \min_{W \in K} \|p - c\|_{W^{-1}} - 1$$

by the triangle inequality $\|p-r\|_{W^{-1}} \geq \|p-c\|_{W^{-1}} - \|c-r\|_{W^{-1}} \geq \|p-c\|_{W^{-1}} - 1$. The minimum is attained since K is compact, and positive since $p \notin \bigcup_{W \in K} \mathfrak{S}_W$. Now, from Corollary B.6.2 we see that $\hat{q} = e$ if and only if $\hat{r} = p$. The map from $\hat{q} \mapsto \hat{r}$ is continuous in $q = e$. In fact, $\hat{Q}$ is uniformly positive in a neighborhood of e and hence the corresponding $\hat{c} = 0$. Due to this continuity, if $\hat{q} \to e$, then $\hat{r} \to p$, which cannot happen since $\|p - r\|_{W^{-1}} \geq \kappa$ for all $W \in K$. Thus, since $\|\hat{q} - e\|_W$ is bounded away from zero, the preimage $f^{-1}(K)$ of K is bounded. Finally, consider a convergent sequence (W_k) in $f^{-1}(K)$ converging to a limit W_∞. Since the sequence is bounded and cannot converge to a singular matrix, we must have $W_\infty > 0$, i.e., $W_\infty \in f^{-1}(\mathcal{W})$. By continuity, $f(W_k)$ tends to the limit $f(W_\infty)$, which must belong to K since it is compact. Hence the preimage W_∞ must belong to $f^{-1}(K)$. Therefore, $f^{-1}(K)$ is compact as claimed. □

It is illustrative to consider the simple case when $W = \lambda I$. Then the two maps (B.7.1) and (B.7.2) become

$$\begin{aligned} \lambda_{\text{soft}} &= \frac{\sqrt{\lambda_{\text{hard}}}}{\|\hat{q} - e\|_2}, \\ \lambda_{\text{hard}} &= \lambda_{\text{soft}}^2 \|\hat{q} - e\|_2^2. \end{aligned} \tag{B.7.3}$$

Whereas the range of λ_{soft} is the semi-infinite interval $(0, \infty)$, for the homeomorphism to hold λ_{hard} is confined to

$$\lambda_{\min} < \lambda < \lambda_{\max},$$

where $\lambda_{\min}$ is the distance from c to the cone $\bar{\mathfrak{C}}_+$ and $\lambda_{\max} = \|c - p\|$. When $\lambda_{\text{soft}} \to \infty$, $\lambda_{\text{hard}} \to \lambda_{\max}$ and $\hat{q} \to e$. If $\lambda_{\text{hard}} \geq \lambda_{\max}$, then the corresponding problem has the trivial unique solution $\hat{q} = e$, corresponding to the primal solution $d\hat{\mu} = P dm$.

Note that Theorem B.7.1 implies that some continuity results in one of the problems can be automatically transferred to the other problem. In particular, we have the following result.

Theorem B.7.2. *Let*

$$\mathbb{J}_W(q) = \langle c, q \rangle - \int_{\mathbb{T}^d} P \log Q \, dm + \|q - e\|_W. \tag{B.7.4}$$

Then the map $W \mapsto \hat{q} := \arg\min_{q \in \bar{\mathfrak{P}}_+} \mathbb{J}_W(q)$ *is continuous.*

Proof. The theorem follows by noting that $W \mapsto \hat{q} := \arg\min_{q \in \bar{\mathfrak{P}}_+} \mathbb{J}_W(q)$ can be seen as a composition of two continuous maps, namely the one in Theorem B.4.1 and the one in Theorem B.7.1. □

Next we shall also vary c and p, and to this end we introduce a more explicit notation for $\mathfrak{S}_W$ and $\mathcal{W}$, namely $\mathfrak{S}_{c,W} = \mathfrak{S}_W$ in (B.6.1) and

$$\mathcal{W}_{c,p} := \{W > 0 \mid \mathfrak{S}_{c,W} \cap \mathfrak{C}_+ \neq \emptyset, p \notin \mathfrak{S}_{c,W}\}.$$

Then the corresponding set of parameters (B.4.1) for the problem with hard constraints is given by

$$\mathcal{P}_{\text{hard}} = \{(c, p, W) \mid c \in \mathfrak{C}, p \in \bar{\mathfrak{P}}_+ \setminus \{0\}, W \in \mathcal{W}_{c,p}\}, \tag{B.7.5}$$

the interior of which is

$$\text{int}(\mathcal{P}_{\text{hard}}) = \{(c, p, W) \mid c \in \mathfrak{C}, p \in \mathfrak{P}_+, W \in \mathcal{W}_{c,p}\}.$$

Theorem B.7.1 can now be modified accordingly to yield the following theorem, the proof of which is deferred to the appendix.

Theorem B.7.3. *Let the map* $(c, p, W_{\text{hard}}) \mapsto W_{\text{soft}}$ *be given by* (B.7.1) *and the map* $(c, p, W_{\text{soft}}) \mapsto W_{\text{hard}}$ *by* (B.7.2)*. Then the map that sends* $(c, p, W_{\text{hard}}) \in \text{int}(\mathcal{P}_{\text{hard}})$ *to* $(c, p, W_{\text{soft}}) \in \text{int}(\mathcal{P})$ *is a homeomorphism.*

Note that this theorem is not a strict amplification of Theorem B.7.1 as we have given up the possibility for p to be on the boundary $\partial\mathfrak{P}_+$. The same is true for the following modification of Theorem B.7.2.

Theorem B.7.4. *Let* $\mathbb{J}_{c,p,W}(q)$ *be as in* (B.7.4)*. Then the map* $(c, p, W) \mapsto \hat{q} := \arg\min_{q \in \bar{\mathfrak{P}}_+} \mathbb{J}_{c,p,W}(q)$ *is continuous on* $\text{int}(\mathcal{P}_{\text{hard}})$.

Proof. The theorem follows immediately by noting that $(c, p, W_{\text{hard}}) \mapsto \hat{q}$ can be seen as a composition of two continuous maps, namely $(c, p, W_{\text{hard}}) \mapsto (c, p, W_{\text{soft}})$ of Theorem B.7.3 and $(c, p, W_{\text{soft}}) \mapsto \hat{q}$ of Theorem B.4.1. □

Theorem B.7.4 is a counterpart of Theorem B.4.1 for the problem with hard constraints, except that p is restricted to the interior $\mathfrak{P}_+$. It should be possible to extend the result to hold for all $p \in \bar{\mathfrak{P}}_+ \setminus \{0\}$ via a direct proof along the lines of the proof of Theorem B.4.1.

B.8 Estimating covariances from data

For a scalar stationary stochastic process $\{y(t);\, t \in \mathbb{Z}\}$, it is well-known that the biased covariance estimate

$$c_k = \frac{1}{N} \sum_{t=0}^{N-k-1} y_t \bar{y}_{t+k},$$

based on an observation record $\{y_t\}_{t=0}^{N-1}$, yields a positive definite Toeplitz matrix, which is equivalent to $c \in \mathfrak{C}_+$ [2, pp. 13-14]. In fact, these estimates correspond to the ones obtained from the periodogram estimate of the spectrum (see, e.g., [59, Sec. 2.2]). On the other hand, the Toeplitz matrix of the unbiased estimate

$$c_k = \frac{1}{N-k} \sum_{t=0}^{N-k-1} y_t \bar{y}_{t+k}$$

is in general not positive definite.

The same holds in higher dimensions ($d > 1$) where the observation record is $\{y_{\boldsymbol{t}}\}_{\boldsymbol{t} \in \mathbb{Z}_N^d}$ with

$$\mathbb{Z}_N^d = \{(\ell_1, \ldots, \ell_d) \,|\, 0 \le \ell_j \le N_j - 1, j = 1, \ldots, d\}.$$

The unbiased estimate is then given by

$$c_{\mathbf{k}} = \frac{1}{\prod_{j=1}^d (N_j - |k_j|)} \sum_{\boldsymbol{t} \in \mathbb{Z}_N^d} y_{\boldsymbol{t}} \bar{y}_{\boldsymbol{t}+\mathbf{k}}, \tag{B.8.1}$$

and the biased estimate by

$$c_{\mathbf{k}} = \frac{1}{\prod_{j=1}^d N_j} \sum_{\boldsymbol{t} \in \mathbb{Z}_N^d} y_{\boldsymbol{t}} \bar{y}_{\boldsymbol{t}+\mathbf{k}}, \tag{B.8.2}$$

where we define $y_{\boldsymbol{t}} = 0$ for $\boldsymbol{t} \notin \mathbb{Z}_N^d$. The sequence of unbiased covariance estimates does not in general belong to $\mathfrak{C}_+$, but the biased covariance estimates yields $c \in \mathfrak{C}_+$ also in the multidimensional setting. In fact, this can be seen by noting that the biased estimate corresponds to the Fourier coefficients of the periodogram [18, Sec. 6.5.1], i.e., if the estimates $c_{\mathbf{k}}$ are given by (B.8.2), then

$$\Phi_{\text{periodogram}}(\boldsymbol{\theta}) := \frac{1}{\prod_{j=1}^d N_j} \Big| \sum_{\boldsymbol{t} \in \mathbb{Z}_N^d} y_{\boldsymbol{t}} e^{i(\boldsymbol{t}, \boldsymbol{\theta})} \Big|^2 = \sum_{\mathbf{k} \in \mathbb{Z}_N^d - \mathbb{Z}_N^d} c_{\mathbf{k}} e^{-i(\mathbf{k}, \boldsymbol{\theta})}, \tag{B.8.3}$$

where $\mathbb{Z}_N^d - \mathbb{Z}_N^d$ denotes the Minkowski set difference. This leads to the following lemma.

Lemma B.8.1. *Given the observed data* $\{y_{\boldsymbol{t}}\}_{\boldsymbol{t}\in\mathbb{Z}_N^d}$*, let* $\{c_{\mathbf{k}}\}_{\mathbf{k}\in\Lambda}$ *be given by* (B.8.2)*. Then* $c \in \mathfrak{C}_+$*.*

Proof. Given $\{y_{\boldsymbol{t}}\}_{\boldsymbol{t}\in\mathbb{Z}_N^d}$, let $c = \{c_{\mathbf{k}}\}_{\mathbf{k}\in\mathbb{Z}_N^d}$, where $c_{\mathbf{k}}$ be given by (B.8.2). In view of (B.2.1) and (B.8.3) we have

$$\langle c, p\rangle = \int_{\mathbb{T}^d} \frac{1}{\prod_{j=1}^d N_j} \Big| \sum_{\boldsymbol{t}\in\mathbb{Z}_N^d} y_{\boldsymbol{t}} e^{i(\boldsymbol{t},\boldsymbol{\theta})} \Big|^2 P(e^{i\boldsymbol{\theta}}) dm(\boldsymbol{\theta}),$$

which is positive for all $p \in \bar{\mathfrak{P}}_+ \setminus \{0\}$. Consequently, $c \in \mathfrak{C}_+$. □

An advantage of the approximate procedures to the RCEP is that they can also be used for cases where the biased estimate is not available, e.g., where the covariance is estimated from snapshots.

B.9 Application to spectral estimation

As long as we use the biased estimate (B.8.2), we may apply exact covariance matching as outlined in Section B.2, whereas in general approximate covariance matching will be required for unbiased covariance estimates. However, as will be seen in the following example, approximate covariance matching may sometimes be better even if $c \in \mathfrak{C}_+$.

In this application it is easy to determine a bound on the acceptable error in the covariance matching, so we use the procedure with hard constraints. Given data generated from a two-dimensional stochastic system, we test three different procedures, namely (i) using the biased estimate and exact matching, (ii) using the biased estimate and the approximate matching (B.1.9), and (iii) using the unbiased estimate and the approximate matching (B.1.9). The procedures are then evaluated by checking the size of the error between the matched covariances and the true ones from the dynamical system.

An example

Let $y_{(t_1,t_2)}$ be the steady-state output of a two-dimensional recursive filter driven by a white noise input $u_{(t_1,t_2)}$. Let the transfer function of the recursive filter be

$$\frac{b(e^{i\theta_1}, e^{i\theta_2})}{a(e^{i\theta_1}, e^{i\theta_2})} = \frac{\sum_{\mathbf{k}\in\Lambda_+} b_{\mathbf{k}} e^{-i(\mathbf{k},\boldsymbol{\theta})}}{\sum_{\mathbf{k}\in\Lambda_+} a_{\mathbf{k}} e^{-i(\mathbf{k},\boldsymbol{\theta})}},$$

where $\Lambda_+ = \{(k_1, k_2) \in \mathbb{Z}^2 \mid 0 \le k_1 \le 2, 0 \le k_2 \le 2\}$ and the coefficients are given by $b_{(k_1,k_2)} = B_{k_1+1,k_2+1}$ and $a_{(k_1,k_2)} = A_{k_1+1,k_2+1}$, where

$$B = \begin{bmatrix} 0.9 & -0.2 & 0.05 \\ 0.2 & 0.3 & 0.05 \\ -0.05 & -0.05 & 0.1 \end{bmatrix}, \; A = \begin{bmatrix} 1 & 0.1 & 0.1 \\ -0.2 & 0.2 & -0.1 \\ 0.4 & -0.1 & -0.2 \end{bmatrix}.$$

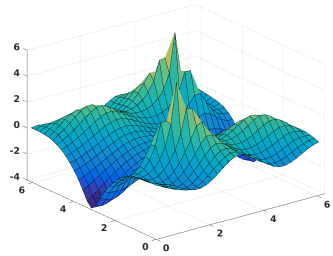


Figure B.1: Log-plot of the original spectrum.

The spectral density Φ of $y_{(t_1,t_2)}$, which is shown in Figure B.1 and is similar to the one considered in [53, Sec. 7], is given by

$$\Phi(e^{i\theta_1}, e^{i\theta_2}) = \frac{P(e^{i\theta_1}, e^{i\theta_2})}{Q(e^{i\theta_1}, e^{i\theta_2})} = \left| \frac{b(e^{i\theta_1}, e^{i\theta_2})}{a(e^{i\theta_1}, e^{i\theta_2})} \right|^2 ,$$

and hence the index set Λ of the coefficients of the trigonometric polynomials P and Q is given by $\Lambda = \Lambda_+ - \Lambda_+ = \{(k_1, k_2) \in \mathbb{Z}^2 \mid |k_1| \leq 2, |k_2| \leq 2\}$. Using this example, we perform two different simulation studies.

First simulation study

The system was simulated for 500 time steps along each dimension, starting from $y_{(t_1,t_2)} = u_{(t_1,t_2)} = 0$ whenever either $t_1 < 0$ or $t_2 < 0$. Then covariances were estimated from the 9×9 last samples, using both the biased and the unbiased estimators. With this covariance data we investigate the three procedures (i), (ii), and (iii) described above. In each case, both the maximum entropy (ME) solutions and solutions with the true numerator are computed.[4] The weighting matrix is taken to be $W = \lambda I$, where λ is $\lambda_{\text{biased}} := \|c_{\text{true}} - c_{\text{biased}}\|_2^2$ in procedure (ii) and $\lambda_{\text{unbiased}} := \|c_{\text{true}} - c_{\text{unbiased}}\|_2^2$ in procedure (iii).[5] The norm of the error[6] between the matched covariances and the true ones, $\|\hat{r} - c_{\text{true}}\|_2$, is shown in Table. B.1. The means and standard deviations are computed over the 100 runs.

The biased covariance estimates belong to the cone $\mathfrak{C}_+$ (Lemma B.8.1), and therefore procedure (i) can be used. The corresponding error in Table B.1 is the statistical error in estimating the covariance. This error is quite large because of a

[4] ME: $P \equiv 1$. True numerator: $P = P_{\text{true}}$.

[5] Note that this is the smallest λ for which the true covariance sequence belongs to the uncertainty set $\{r \mid \|r - c\|_2^2 \leq \lambda\}$.

[6] Here we use the norm of the covariance estimation error as the measure of fit. However, note that this is not the only way to compare accuracy of the different methods. The reason for this choice is that comparing the accuracy of the spectral estimates is not straightforward since it depends on the selected metric or distortion measure.

Table B.1: Norm differences $\|\hat{r} - c_{\text{true}}\|_2$ for different solutions in the first simulation setup.

	Mean	Std.
Biased, exact matching	3.2374	1.7944
Biased, approximate matching, ME-solution	3.7886	1.3274
Biased, approximate matching, using true P	3.8152	1.6509
Unbiased, approximate matching, ME-solution	3.2575	1.4721
Unbiased, approximate matching, using true P	3.2811	1.7787

short data record. Using approximate covariance matching in this case seems to give a worse match. However, approximate matching of the unbiased covariances gives as good a fit as exact matching of the biased ones.

Second simulation study

In this simulation, the setup is the same as the previous one, except that the simulation data has been discarded if the *unbiased* estimate belongs to $\bar{\mathfrak{C}}_+$. To obtain 100 such data sets, 414 simulations of the system were needed. (As a comparison, in the previous experiment 23 out of the 100 runs resulted in an unbiased estimate outside $\bar{\mathfrak{C}}_+$.) Again, the norm of the error between matched covariances and the true ones is shown in Table B.2, and the means and standard deviations are computed over the 100 runs.

As before, the biased covariance estimates belong to the cone $\mathfrak{C}_+$, and therefore procedure (i) can be used. Comparing this with the results from procedure (ii) suggests that there may be an advantage not to enforce exact matching, although we know that the data belongs to the cone. Regarding procedure (iii), we know that the unbiased covariance estimates do not belong to the cone $\bar{\mathfrak{C}}_+$, and hence we need to use approximate covariance matching. In this example, this procedure turns out to give the smallest estimation error.

B.10 Application to system identification and texture reconstruction

Next we apply the theory of this paper to texture generation via Wiener system identification. Wiener systems form a class of nonlinear dynamical systems consisting of a linear dynamic part composed with a static nonlinearity as illustrated in Figure B.2. This is a subclass of so-called block-oriented systems [4], and Wiener system identification is a well-researched area (see, e.g., [32] and references therein) that

Table B.2: Norm differences $\|\hat{r} - c_{\text{true}}\|_2$ for different solutions in the second setup, where all unbiased estimate are outside $\bar{\mathfrak{C}}_+$.

	Mean	Std.
Biased, exact matching	2.9245	2.2528
Biased, approximate matching, ME-solution	1.9087	1.1324
Biased, approximate matching, using true P	1.8532	1.1904
Unbiased, approximate matching, ME-solution	1.5018	0.6601
Unbiased, approximate matching, using true P	1.4451	0.7296

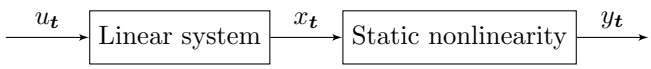


Figure B.2: A Wiener system with thresholding as static nonlinearity.

is still very active [43, 60, 1]. Here, we use Wiener systems to model and generate textures.

Using dynamical systems for modeling images and textures is not new and has been considered in, e.g., [14, 50]. The setup presented here is motivated by [23], where thresholded Gaussian random fields are used to model porous materials for design of surface structures in pharmaceutical film coatings. Hence we let the static nonlinearity, call it f, be a thresholding with unknown thresholding parameter τ. In our previous work [55] we applied exact covariance matching to such a problem. However, in general there is no guarantee that the estimated covariance sequence c belongs to the cone $\mathfrak{C}_+$. Consequently, here we shall use approximate covariance matching instead.

The Wiener system identification can be separated into two parts. We start by identifying the nonlinear part. Using the notation of Figure B.2, let $\{u_{\boldsymbol{t}};\, \boldsymbol{t} \in \mathbb{Z}^d\}$ be a zero-mean Gaussian white noise input, and let $\{x_{\boldsymbol{t}};\, \boldsymbol{t} \in \mathbb{Z}^d\}$ be the stationary output of the linear system, which we assume to be normalized so that $c_{\mathbf{0}} := \mathbb{E}[x_{\boldsymbol{t}}^2] = 1$. Moreover, let $y_{\boldsymbol{t}} = f(x_{\boldsymbol{t}})$, where f is the static nonlinearity

$$f(x) = \begin{cases} 1 & x > \tau \\ 0 & \text{otherwise} \end{cases} \tag{B.10.1}$$

with unknown thresholding parameter τ. Since $\mathbb{E}[y_{\boldsymbol{t}}] = 1 - \phi(\tau)$, where $\phi(\tau)$ is the Gaussian cumulative distribution function, an estimate of τ is given by $\tau_{\text{est}} = \phi^{-1}(1 - \mathbb{E}[y_{\boldsymbol{t}}])$.

Now, let $c_{\mathbf{k}}^x := \mathbb{E}[x_{\boldsymbol{t}+\mathbf{k}} x_{\boldsymbol{t}}]$ be the covariances of $x_{\boldsymbol{t}}$, and let $c_{\mathbf{k}}^y := \mathbb{E}[y_{\boldsymbol{t}+\mathbf{k}} y_{\boldsymbol{t}}] - \mathbb{E}[y_{\boldsymbol{t}+\mathbf{k}}]\,\mathbb{E}[y_{\boldsymbol{t}}]$ be the covariances of $y_{\boldsymbol{t}}$. As was explained in [55], by using results

from [51] one can obtain a relation between $c_{\mathbf{k}}^y$ and $c_{\mathbf{k}}^x$, given by

$$c_{\mathbf{k}}^y = \int_0^{c_{\mathbf{k}}^x} \frac{1}{2\pi\sqrt{1-s^2}} \exp\left(-\frac{\tau^2}{1+s}\right) ds. \tag{B.10.2}$$

This is an invertible map, which we compute numerically, and given τ_{est} we can thus get estimates of the covariances $c_{\mathbf{k}}^x$ from estimates of the covariances $c_{\mathbf{k}}^y$. However, even if c^y is a biased estimate so that $c^y \in \mathfrak{C}_+$, c^x may not be a *bona fide* covariance sequence.

Identifying the linear system

Solving (B.1.7) or (B.1.9) for a given sequence of covariance estimates c, we obtain an estimate of the absolutely continuous part of the power spectrum Φ of that process. In the case $d = 1$, $\Phi = P/Q$ can be factorized as

$$\Phi(e^{i\theta}) = \frac{P(e^{i\theta})}{Q(e^{i\theta})} = \frac{|b(e^{i\theta})|^2}{|a(e^{i\theta})|^2},$$

which provides a transfer function of a corresponding linear system, which fed by a white noise input will produce an autoregressive-moving-average (ARMA) process with an output signal with precisely the power distribution Φ in steady state. For $d \geq 2$, a spectral factorization of this kind is not possible in general [19], but instead there is always a factorization as a sum-of-several-squares [17, 29],

$$\Phi(e^{i\boldsymbol{\theta}}) = \frac{P(e^{i\boldsymbol{\theta}})}{Q(e^{i\boldsymbol{\theta}})} = \frac{\sum_{k=1}^{\ell} |b_k(e^{i\boldsymbol{\theta}})|^2}{\sum_{k=1}^{m} |a_k(e^{i\boldsymbol{\theta}})|^2},$$

the interpretation of which in terms of a dynamical system is unclear when $m, \ell > 1$. Therefore we resort to a heuristic and apply the factorization procedure in [28, Thm. 1.1.1], although some of the conditions required to ensure the existence of a spectral factor may not be met. (See [53, Sec. 7] for a more detailed discussion.)

Simulation results

The method, which is summarized in Algorithm B.1, is tested on some textures from the Outex database [49] (available online from `http://www.outex.oulu.fi/`). These textures are color images and have thus been converted to binary textures by first converting them to black-and-white and then thresholding them.[7] Three such textures are shown in Figure B.3a through B.3c.

[7] The algorithm has been implemented and tested in Matlab, version R2015b. The textures have been normalized to account for light inhomogeneities using a reference image available in the database. The conversion from color images to black-and-white images was done with the built-in function `rgb2gray`, and the threshold level was set to the mean value of the maximum and minimum pixel values in the black-and-white image.

Algorithm B.1

Input: $(y_{\boldsymbol{t}})$

1: Estimate threshold parameter: $\tau_{\text{est}} = \phi^{-1}(1 - E[y_{\boldsymbol{t}}])$
2: Estimate covariances: $c^y_{\mathbf{k}} := E[y_{\boldsymbol{t}+\mathbf{k}} y_{\boldsymbol{t}}] - E[y_{\boldsymbol{t}+\mathbf{k}}]E[y_{\boldsymbol{t}}]$
3: Compute covariances $c^x_{\mathbf{k}} := E[x_{\boldsymbol{t}+\mathbf{k}} x_{\boldsymbol{t}}]$ by using (B.10.2)
4: Estimate a rational spectrum using Theorem B.3.1 or B.6.1
5: Apply the factorization procedure in [28, Thm. 1.1.1]

Output: τ_{est}, coefficients for the linear dynamical system

In this example there is no natural bound on the error, so we use the problem with soft constraints, for which we choose the weight $W = \lambda I$ with $\lambda = 0.01$ for all data sets. Moreover, we do maximum-entropy reconstructions, i.e., we set the prior to $P \equiv 1$. The optimization problems are then solved by first discretizing the grid $\mathbb{T}^2$, in this case in 50×50 points (cf. [53, Thm. 2.6]), and solving the corresponding problems using the CVX toolbox [31, 30]. The reconstructions are shown in Figures B.3d - B.3f. Each reconstruction seems to provide a reasonable visual representation of the structure of the corresponding original. This is especially the case for the second texture.

B.11 Conclusions

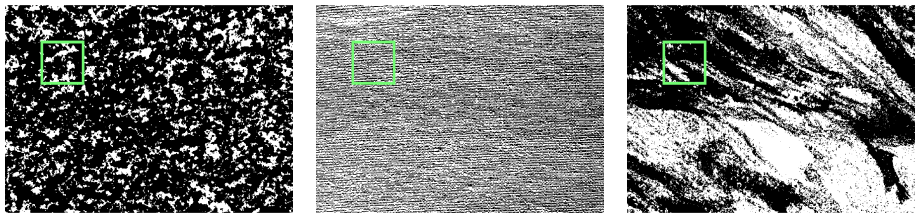
In this work we extend the results of our previous paper [53] on the multidimensional RCEP to allow for approximate covariance matching. We have provided two formulations to this problem, and we have shown that they are connected via a homeomorphism. In both formulations we have used weighted 2-norms to quantify the mismatch of the estimated covariances. However, we expect that by suitable modifications of the proofs similar results can be derived for other norms since all norms have directional derivatives in each point [16, p. 49].

These results provide a procedure for multidimensional spectral estimation, but in order to obtain a complete theory for multidimensional system identification and realization theory there are still some open problems, such as spectral factorization and interpretations in terms of multidimensional stochastic systems, as briefly discussed in Section B.10.

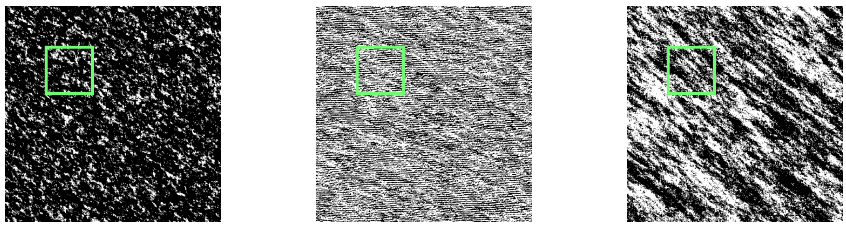
B.12 Appendix

Let $B_\rho(x^{(0)})$ denote the closed ball $\{x \in X \mid \|x - x^{(0)}\|_X \leq \rho\}$, where X is either a set of vectors or a set of matrices, depending on the context. The norm $\|\cdot\|_X$ is the Euclidean norm for vectors and the Frobenius norm for matrices.

Lemma B.12.1. *Let $\mathcal{P}$ be given by* (B.4.1) *and $\mathbb{J}_{c,p,W}$ by* (B.4.2). *Furthermore, let $\hat{q} := \min_{q \in \bar{\mathfrak{P}}_+} \mathbb{J}_{c,p,W}(q)$. Then the map $(c, p, W) \mapsto \mathbb{J}_{c,p,W}(\hat{q})$ is continuous for*

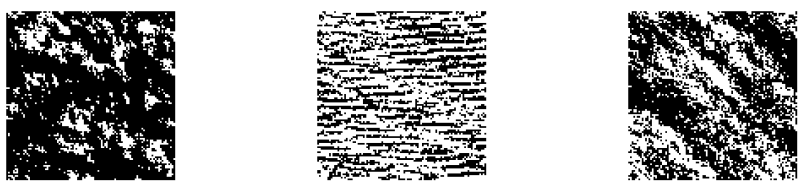


(a) First texture. (b) Second texture. (c) Third texture.



(d) Reconstruction of B.3a. (e) Reconstruction of B.3b. (f) Reconstruction of B.3c.

(g) Close-up of B.3a. (h) Close-up of B.3b. (i) Close-up of B.3c.



(j) Close-up of B.3d. (k) Close-up of B.3e. (l) Close-up of B.3f.

Figure B.3: In Figures B.3a - B.3c three different binary textures, of size 1200×900 pixels, are shown. These are obtained from the textures granite001-inca-100dpi-00, paper010-inca-100dpi-00, and plastic008-inca-100dpi-00 in the Outex database, respectively. The textures in Figures B.3a - B.3c are used as input (y_t) to Algorithm B.1 and in Figures B.3d - B.3f the corresponding reconstructed textures of size 500×500 are shown. In Figures B.3g - B.3l close-ups of size 100×100 are shown of the original and reconstructed textures (areas marked in Figures B.3a - B.3f).

$(c, p, W) \in \mathcal{P}$. *Moreover, for any compact* $K \subset \mathcal{P}$ *the corresponding set of optimal solutions* $\hat{q}$ *is bounded.*

Proof. The proof follows along the lines of Lemma 7.2 and Proposition 7.4 in [36]. Let $(c^{(0)}, p^{(0)}, W^{(0)}) \in \mathcal{P}$ be arbitrary and let

$$\tilde{B}_\rho(c^{(0)}, p^{(0)}, W^{(0)}) := B_\rho(c^{(0)}) \times \left(B_\rho(p^{(0)}) \cap \bar{\mathfrak{P}}_+\right) \times B_\rho(W^{(0)}),$$

where $\rho > 0$ is chosen so that $\tilde{B}_\rho(c^{(0)}, p^{(0)}, W^{(0)}) \subset \mathcal{P}$, i.e., $\rho < \|p^{(0)}\|_2$ and $W > 0$ for all $\|W - W^{(0)}\|_F \leq \rho$. First we will show that the minimizer $\hat{q}_{c,p,W}$ of $\mathbb{J}_{c,p,W}$ is bounded for all $(c, p, W) \in \tilde{B}_\rho(c^{(0)}, p^{(0)}, W^{(0)})$. To this end, note that by optimality

$$\mathbb{J}_{c,p,W}(\hat{q}_{c,p,W}) \leq \mathbb{J}_{c,p,W}(e) = \langle c, e\rangle - \int_{\mathbb{T}^d} P \log 1 \, dm + \frac{1}{2}\|e - e\|_W^2 = c_{\mathbf{0}},$$

and hence $\mathbb{J}_{c,p,W}(p)$ is bounded from above on the compact set $\tilde{B}_\rho(c^{(0)}, p^{(0)}, W^{(0)})$. Consequently, by using the same inequality as in the proof of [36, Lem. 7.1], we see that

$$c_{\mathbf{0}} \geq \mathbb{J}_{c,p,W}(\hat{q}_{c,p,W}) \geq \langle c, \hat{q}_{c,p,W}\rangle - \|P\|_1 \log \|\hat{Q}_{c,p,W}\|_\infty + \frac{1}{2}\|\hat{q}_{c,p,W} - e\|_W^2.$$

Due to norm equivalence between $\|Q\|_\infty$ and $\|q\|_W$, and since the quadratic term is dominating, the norm of $\hat{q}_{c,p,W}$ is bounded in the set $\tilde{B}_\rho(c^{(0)}, p^{(0)}, W^{(0)})$.

Now, let $K \subset \mathcal{P}$ be compact. We want to show that $\hat{q}$ is bounded on K. Assume it is not. Then let $(c^{(k)}, p^{(k)}, W^{(k)}) \in K$ be a sequence with $\|\hat{q}_k\| \to \infty$. Since K is compact, there is a converging subsequence $(c^{(k)}, p^{(k)}, W^{(k)}) \to (c, p, W) \in K$ with $\|\hat{q}_k\| \to \infty$. Since $(c, p, W) \in K$, there is a $\rho > 0$ such that $\tilde{B}_\rho(c, p, W) \subset \mathcal{P}$. However, all but finitely many points $(c^{(k)}, p^{(k)}, W^{(k)})$ belong to $\tilde{B}_\rho(c, p, W)$, and since $\hat{q}_k$ is bounded for all $(c^{(k)}, p^{(k)}, W^{(k)}) \in \tilde{B}_\rho(c, p, W)$, we cannot have $\|\hat{q}_k\| \to \infty$.

Next, let $(c^{(1)}, p^{(1)}, W^{(1)}), (c^{(2)}, p^{(2)}, W^{(2)}) \in \tilde{B}_\rho(c^{(0)}, p^{(0)}, W^{(0)})$ and let $\hat{q}_1, \hat{q}_2 \in \bar{\mathfrak{P}}_+$ be the unique minimizers of $\mathbb{J}_{c^{(1)},p^{(1)},W^{(1)}}$ and $\mathbb{J}_{c^{(2)},p^{(2)},W^{(2)}}$, respectively. Choose a $q_0 \in \mathfrak{P}_+$, and note that Q_0 is strictly positive and bounded. By optimality,

$$\mathbb{J}_{c^{(1)},p^{(1)},W^{(1)}}(\hat{q}_1) \leq \mathbb{J}_{c^{(1)},p^{(1)},W^{(1)}}(\hat{q}_2 + \varepsilon q_0) \tag{B.12.1a}$$

$$\mathbb{J}_{c^{(2)},p^{(2)},W^{(2)}}(\hat{q}_2) \leq \mathbb{J}_{c^{(2)},p^{(2)},W^{(2)}}(\hat{q}_1 + \varepsilon q_0) \tag{B.12.1b}$$

for all $\varepsilon > 0$. Hence, if we can show that, for any $\delta > 0$, there are an $\varepsilon > 0$ and a $\tilde{\rho} > 0$ such that

$$|\mathbb{J}_{c^{(2)},p^{(2)},W^{(2)}}(\hat{q}_1 + \varepsilon q_0) - \mathbb{J}_{c^{(1)},p^{(1)},W^{(1)}}(\hat{q}_1)| \leq \delta \tag{B.12.2a}$$

$$|\mathbb{J}_{c^{(1)},p^{(1)},W^{(1)}}(\hat{q}_2 + \varepsilon q_0) - \mathbb{J}_{c^{(2)},p^{(2)},W^{(2)}}(\hat{q}_2)| \leq \delta \tag{B.12.2b}$$

hold whenever $\|c^{(1)} - c^{(2)}\|_2 \leq \tilde{\rho}$, $\|p^{(1)} - p^{(2)}\|_2 \leq \tilde{\rho}$ and $\|W^{(1)} - W^{(2)}\|_F \leq \tilde{\rho}$, then this would imply that

$$\mathbb{J}_{c^{(2)},p^{(2)},W^{(2)}}(\hat{q}_2) - \delta \leq \mathbb{J}_{c^{(1)},p^{(1)},W^{(1)}}(\hat{q}_1) \leq \mathbb{J}_{c^{(2)},p^{(2)},W^{(2)}}(\hat{q}_2) + \delta,$$

showing that the optimal value is continuous in $c^{(1)}, p^{(1)}, W^{(1)}$. The lower bound is obtained by using (B.12.2a) and (B.12.1b), and the upper bound is obtained from (B.12.1a) and (B.12.2b). To prove (B.12.2a), we note that

$$\begin{aligned}
&|\mathbb{J}_{c^{(2)},p^{(2)},W^{(2)}}(\hat{q}_1+\varepsilon q_0)-\mathbb{J}_{c^{(1)},p^{(1)},W^{(1)}}(\hat{q}_1)| \\
&\quad=\Bigg|\langle c^{(2)}-c^{(1)},\hat{q}_1\rangle+\langle c^{(2)},\varepsilon q_0\rangle-\int_{\mathbb{T}^d}P^{(1)}\log\left(1+\frac{\varepsilon Q_0}{\hat{Q}_1}\right)dm \\
&\qquad-\int_{\mathbb{T}^d}(P^{(2)}-P^{(1)})\log\left(\hat{Q}_1+\varepsilon Q_0\right)dm+\tfrac{1}{2}\|\hat{q}_1+\varepsilon q_0-e\|^2_{W^{(2)}}-\tfrac{1}{2}\|\hat{q}_1-e\|^2_{W^{(1)}}\Bigg| \\
&\quad\le\|c^{(2)}-c^{(1)}\|_2\|\hat{q}_1\|_2+\varepsilon\left(\langle c^{(2)},q_0\rangle+\int_{\mathbb{T}^d}P^{(1)}\frac{Q_0}{\hat{Q}_1}dm\right) \\
&\qquad+\|P^{(2)}-P^{(1)}\|_\infty\|\log(\hat{Q}_1+\varepsilon Q_0)\|_1+\tfrac{1}{2}\left|\,\|\hat{q}_1+\varepsilon q_0-e\|^2_{W^{(2)}}-\|\hat{q}_1-e\|^2_{W^{(2)}}\right| \\
&\qquad+\tfrac{1}{2}\left|\,\|\hat{q}_1-e\|^2_{W^{(2)}}-\|\hat{q}_1-e\|^2_{W^{(1)}}\right|. \qquad \text{(B.12.3)}
\end{aligned}$$

Next we observe that

$$0\le\int_{\mathbb{T}^d}P^{(1)}\frac{Q_0}{\hat{Q}_1}dm=\langle\hat{r}_1-\hat{c}_1,q_0\rangle\le\langle c_1+W^{(1)}(\hat{q}_1-e),q_0\rangle$$

by the KKT conditions (B.3.11) and the fact that $q_0\in\mathfrak{P}_+$, $\hat{c}_1\in\bar{\mathfrak{C}}_+$. Hence ε can be selected small enough for the second and fourth terms in (B.12.3) each to be bounded by $\delta/5$ for any $(c^{(1)},p^{(1)},W^{(1)}),(c^{(2)},p^{(2)},W^{(2)})\in B_\rho(c^{(0)},p^{(0)},W^{(0)})$. Each of the remaining terms can now be bounded by $\delta/5$ by selecting $\tilde{\rho}$ sufficiently small. Hence (B.12.2a) follows. This also proves (B.12.2b). □

Lemma B.12.2. *Let $\mathcal{P}_{\text{hard}}$ be given by* (B.7.5) *and $\mathbb{J}_{c,p,W}$ by* (B.6.6)*. Furthermore, let $\hat{q}:=\min_{q\in\bar{\mathfrak{P}}_+}\mathbb{J}_{c,p,W}(q)$. Then, for any compact $K\subset\mathcal{P}_{\text{hard}}$, the corresponding set of optimal solutions $\hat{q}$ is bounded.*

Proof. The proof follows closely thatof the corresponding part of Lemma B.12.1. Let $(c^{(0)},p^{(0)},W^{(0)})\in\mathcal{P}_{\text{hard}}$ be arbitrary, and let

$$\tilde{B}_\rho(c^{(0)},p^{(0)},W^{(0)}):=B_\rho(c^{(0)})\times\left(B_\rho(p^{(0)})\cap\bar{\mathfrak{P}}_+\right)\times B_\rho(W^{(0)}),$$

where $\rho>0$ is chosen so that $\tilde{B}_\rho(c^{(0)},p^{(0)},W^{(0)})\subset\mathcal{P}_{\text{hard}}$. To see that the minimizer $\hat{q}_{c,p,W}$ of $\mathbb{J}_{c,p,W}$ is bounded for all $(c,p,W)\in\tilde{B}_\rho(c^{(0)},p^{(0)},W^{(0)})$, first note that by optimality

$$\mathbb{J}_{c,p,W}(\hat{q}_{c,p,W})\le\mathbb{J}_{c,p,W}(e)=\langle c,e\rangle-\int_{\mathbb{T}^d}P\log 1\,dm+\|e-e\|_W=c_{\mathbf{0}},$$

and hence $\mathbb{J}_{c,p,W}(p)$ is bounded from above on the compact set $\tilde{B}_\rho(c^{(0)},p^{(0)},W^{(0)})$.

Now let $h(q):=\langle c,q\rangle+\|q\|_W$, as in the proof of Theorem B.6.1. Following the same line of argument as in that proof, we see that $h(q)>0$ for all $q\in\bar{\mathfrak{P}}_+$ and

$(c,p,W) \in \tilde{B}_\rho(c^{(0)},p^{(0)},W^{(0)})$. Since h is continuous in the arguments (q,c,p,W), it has a minimum $\varepsilon > 0$ on the compact set of tuples (q,c,p,W) such that $q \in \bar{\mathfrak{P}}_+ \setminus \{0\}$, $\|q\|_\infty = 1$, and $(c,p,W) \in \tilde{B}_\rho(c^{(0)},p^{(0)},W^{(0)})$ hold. Thus the second half of inequality (B.6.14) still holds, i.e.,

$$\mathbb{J}_{c,p,W}(q) \geq \frac{\varepsilon}{|\Lambda|}\|Q\|_\infty - \int_{\mathbb{T}^d} P \log \|Q\|_\infty dm - \kappa\|e\|_\infty \tag{B.12.4}$$

for all q, where $\kappa > 0$ is a constant that follows from norm equivalence. This is true in particular for $\hat{q}_{c,p,W}$, and thus

$$c_{\mathbf{0}} \geq \mathbb{J}_{c,p,W}(\hat{q}_{c,p,W}) \geq \frac{\varepsilon}{|\Lambda|}\|\hat{Q}_{c,p,W}\|_\infty - \int_{\mathbb{T}^d} P \log \|\hat{Q}_{c,p,W}\|_\infty dm - \varepsilon\|e\|_\infty.$$

Since the linear growth dominates the logarithmic growth, the norm of $\hat{q}_{c,p,W}$ is bounded on the set $\tilde{B}_\rho(c^{(0)},p^{(0)},W^{(0)})$. The proof now follows *verbatim* from the argument in the second paragraph in the proof of Lemma B.12.1. □

Proof of Theorem B.7.3. This is a modification of the proof of Theorem B.7.1, again utilizing [11, Lem. 2.3], where we replace the map f defined by $\mathcal{W} \ni W_{\text{hard}} \mapsto W_{\text{soft}} \in \{W \mid W > 0\}$ and redefine it with the map $\text{int}(\mathcal{P}_{\text{hard}}) \ni (c,p,W_{\text{hard}}) \mapsto (c,p,W_{\text{soft}}) \in \text{int}(\mathcal{P})$. To show that f is a homeomorphism we need to show that the map is proper. To this end, we take a compact set $K \subset \text{int}(\mathcal{P}_{\text{hard}})$ and show that $f^{-1}(K)$ is also compact. Again, there are two ways this could fail. First, the preimage could contain a singular semidefinite matrix. However, this is impossible by (B.7.2) since $\|\hat{q}\|_\infty$ is bounded for $(c,p,W_{\text{hard}}) \in K$ (Lemma B.12.2) and a nonzero scaling of a singular matrix cannot be nonsingular. Second, $\|W_{\text{soft}}\|_F$ could tend to infinity. However, this is also impossible. To see this, we first show that there is a $\kappa > 0$ such that $\|p-r\|_{W_{\text{hard}}^{-1}} \geq \kappa$ for all $r \in \mathfrak{S}_{c,W_{\text{hard}}}$ and all $(c,p,W_{\text{hard}}) \in K$. Again, using the triangle inequality $\|p-r\|_{W_{\text{hard}}^{-1}} \geq \|p-c\|_{W_{\text{hard}}^{-1}} - \|c-r\|_{W_{\text{hard}}^{-1}}$, we observe that the minimum of $\|p-r\|_{W_{\text{hard}}^{-1}}$ over all $(c,p,W_{\text{hard}}) \in K$ and r satisfying the constraint $\|r-c\|_{W_{\text{hard}}^{-1}} \leq 1$ is bounded by

$$\kappa := \min_{(c,p,W_{\text{hard}}) \in K} \|p-c\|_{W_{\text{hard}}^{-1}} - 1.$$

Note that the minimum is attained since K is compact, and positive since $p \notin \bigcup_{(c,W_{\text{hard}}) \in K} \mathfrak{S}_{c,W}$. The remaining part of the proof now follows with minor modifications from the proof of Theorem B.7.1 by noting that $\hat{q}$ is bounded away from e, and hence the preimage $f^{-1}(K)$ is bounded. Therefore, the limit of a sequence in the preimage must belong to $f^{-1}(K)$, and hence $f^{-1}(K)$ is compact as claimed. □

References

[1] M.R. Abdalmoaty and H. Hjalmarsson. A simulated maximum likelihood method for estimation of stochastic wiener systems. In *IEEE Annual Conference on Decision and Control (CDC)*, pages 3060–3065. IEEE, 2016.

[2] N. Ahiezer and M. Krein. *Some questions in the theory of moments*, volume 2 of *Translations of mathematical monographs*. American Mathematical Society, Providence, R.I., 1962.

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

[4] S.A. Billings. Identification of nonlinear systems-a survey. *IEE Proceedings D-Control Theory and Applications*, 127(6):272–285, 1980.

[5] N.K. Bose. *Multidimensional Systems Theory and Applications*. Kluwer Academic Publishers, second edition, 2003.

[6] 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.

[7] C.I. Byrnes, T.T. Georgiou, and A. Lindquist. A new approach to spectral estimation: A tunable high-resolution spectral estimator. *IEEE Transactions on Signal Processing*, 48(11):3189–3205, 2000.

[8] 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.

[9] C.I. Byrnes, T.T. Georgiou, A. Lindquist, and A. Megretski. Generalized interpolation in H^∞ with a complexity constraint. *Transactions of the American Mathematical Society*, 358(3):965–987, 2006.

[10] C.I. Byrnes and A. Lindquist. The uncertain generalized moment problem with complexity constraint. In W. Kang, C. Borges, and M. Xiao, editors, *New Trends in Nonlinear Dynamics and Control and their Applications*, volume 295 of *Lecture Notes in Control and Information Science*, pages 267–278. Springer Berlin Heidelberg, 2003.

[11] C.I. Byrnes and A. Lindquist. Interior point solutions of variational problems and global inverse function theorems. *International Journal of Robust and Nonlinear Control*, 17(5-6):463–481, 2007.

[12] 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.

[13] F.P. Carli and T.T. Georgiou. On the covariance completion problem under a circulant structure. *IEEE Transactions on Automatic Control*, 56(4):918–922, 2011.

[14] A. Chiuso, A. Ferrante, and G. Picci. Reciprocal realization and modeling of textured images. In *IEEE Annual Conference on Decision and Control (CDC), and European Control Conference (ECC)*, pages 6059–6064. IEEE, 2005.

[15] I Csiszár. Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. *The Annals of Statistics*, 19(4):pp. 2032–2066, 1991.

[16] K. Deimling. *Nonlinear functional analysis*. Springer, Berlin Heidelberg, 1985.

[17] M.A. Dritschel. On factorization of trigonometric polynomials. *Integral Equations and Operator Theory*, 49(1):11–42, 2004.

[18] D.E. Dudgeon and R.M. Mersereau. *Multidimensional digital signal processing*. Prentice-Hall, Englewood Cliffs, NJ, 1984.

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

[20] M.P. Ekstrom. *Digital image processing techniques*. Academic Press, 1984.

[21] 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.

[22] 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.

[23] S. Eriksson Barman. Gaussian random field based models for the porous structure of pharmaceutical film coatings. In *Acta Stereologica [En ligne], Proceedings ICSIA, 14th ICSIA abstracts*, 2015. `http://popups.ulg.ac.be/0351-580X/index.php?id=3775`.

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

[25] T.T. Georgiou. Solution of the general moment problem via a one-parameter imbedding. *IEEE Transactions on Automatic Control*, 50(6):811–826, 2005.

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

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

[28] J. S. Geronimo and H. J. Woerdeman. Positive extensions, Fejér-Riesz factorization and autoregressive filters in two variables. *Annals of Mathematics*, 160(3):839–906, 2004.

[29] J.S. Geronimo and M.-J. Lai. Factorization of multivariate positive laurent polynomials. *Journal of Approximation Theory*, 139(1):327–345, 2006.

[30] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, volume 371 of *Lecture Notes in Control and Information Sciences*, pages 95–110. Springer-Verlag, London, 2008.

[31] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, Mar. 2014.

[32] W. Greblicki. Nonparametric identification of wiener systems. *IEEE Transactions on information theory*, 38(5):1487–1493, 1992.

[33] H. Gzyl. *The method of maximum entropy*. World Scientific, Singapore, 1995.

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

[35] J. Karlsson, P. Enqvist, and A. Gattami. Confidence assessment for spectral estimation based on estimated covariances. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4343–4347, 2016.

[36] J. Karlsson, A. Lindquist, and A. Ringh. The multidimensional moment problem with complexity constraint. *Integral Equations and Operator Theory*, 84(3):395–418, 2016.

[37] S.W. Lang and J.H. McClellan. Spectral estimation for sensor arrays. In *Proceedings of the First ASSP Workshop on Spectral Estimation*, pages 3.2.1–3.2.7, 1981.

[38] S.W. Lang and J.H. McClellan. The extension of Pisarenko's method to multiple dimensions. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 125–128, 1982.

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

[40] 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.

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

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

[43] F. Lindsten, T.B. Schön, and M.I. Jordan. Bayesian semiparametric Wiener system identification. *Automatica*, 49(7):2053–2063, 2013.

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

[45] J.H. McClellan and S.W. Lang. Mulit-dimensional MEM spectral estimation. In *Proceedings of the Institute of Acoustics "Spectral Analysis and its Use in Underwater Acoustics": Underwater Acoustics Group Conference, Imperial College, London, 29-30 April 1982*, pages 10.1–10.8, 1982.

[46] J.H. McClellan and S.W. Lang. Duality for multidimensional MEM spectral analysis. *Communications, Radar and Signal Processing, IEE Proceedings F*, 130(3):230–235, April 1983.

[47] B.R. Musicus and A.M. Kabel. Maximum entropy pole-zero estimation. Technical Report 510, Research Laboratory of Electronics, Massachusetts Institute of Technology, August 1985.

[48] H.I. Nurdin. New results on the rational covariance extension problem with degree constraint. *Systems & Control Letters*, 55(7):530 – 537, 2006.

[49] T. Ojala, T. Maenpaa, M. Pietikainen, J. Viertola, J. Kyllonen, and S. Huovinen. Outex - New framework for empirical evaluation of texture analysis algorithms. In *International Conference on Pattern Recognition*, volume 1, pages 701–706. IEEE, 2002.

[50] G. Picci and F.P. Carli. Modelling and simulation of images by reciprocal processes. In *Tenth international conference on Computer Modelling and Simulation, UKSIM*, pages 513–518, 2008.

[51] R. Price. A useful theorem for nonlinear devices having gaussian inputs. *IRE Transactions on Information Theory*, 4(2):69–72, 1958.

[52] 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.

[53] 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.

[54] A. Ringh, J. Karlsson, and A. Lindquist. Multidimensional rational covariance extension with approximate covariance matching. In *Proceedings of the 22nd International Symposium on Mathematical Theory of Networks and Systems*, pages 457–460, 2016.

[55] 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.

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

[57] 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.

[58] C.R. Shankwitz and T.T. Georgiou. On the maximum entropy method for interval covariance sequences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(10):1815–1817, 1990.

[59] P. Stoica and R. Moses. *Introduction to Spectral Analysis*. Prentice-Hall, Upper Saddle River, NJ, 1997.

[60] B. Wahlberg, J. Welsh, and L. Ljung. Identification of stochastic wiener systems using indirect inference. *IFAC-PapersOnLine*, 48(28):620 – 625, 2015. 17th IFAC Symposium on System Identification SYSID 2015.

[61] M. Zorzi. Rational approximations of spectral densities based on the alpha divergence. *Mathematics of Control, Signals, and Systems*, 26(2):259–278, 2014.

Paper C

Lower bounds on the maximum delay margin by analytic interpolation

Lower bounds on the maximum delay margin by analytic interpolation

by

Axel Ringh, Johan Karlsson, and Anders Lindquist

Abstract

We study the delay margin problem in the context of recent works by T. Qi, J. Zhu, and J. Chen, where a sufficient condition for the maximal delay margin is formulated in terms of an interpolation problem obtained after introducing a rational approximation. Instead we omit the approximation step and solve the same problem directly using techniques from function theory and analytic interpolation. Furthermore, we introduce a constant shift in the domain of the interpolation problem. In this way we are able to improve on their lower bound for the maximum delay margin.

Keywords: delay systems, robust control, Nevanlinna-Pick interpolation, stability of linear dynamical systems

C.1 Introduction

Time delays are ubiquitous in linear time invariant (LTI) systems, especially in networks, and may occur through communication delay, computational delay or physical transport delay. Consequently, systems with delay have been the subject of much study in systems and control; see, e.g., [11, 22, 8] and references therein.

This paper is devoted to the achievable delay margin in unstable control systems with time delay, a topic that has been studied in various contexts in, e.g., [26, 4, 14, 17, 23, 1, 24, 25, 15, 16]. This problem is related to the gain margin and phase margin problems in robust control [5], [20], but the delay margin problem is more complicated, and many unsolved problems remain. Loosely speaking, we are looking for the largest time delay $\tau_{\max}$ such that there exists an LTI controller that stabilizes the time delay system for each delay in the interval $[0, \tau_{\max})$. In general this is an unsolved problem, and results have been confined to obtaining upper and lower bounds for $\tau_{\max}$. In [23] upper bounds for some simple systems are presented, but in general they are not tight. Methods for finding lower bounds based on different methods have been proposed, e.g, using robust control [26, 14], integral quadratic constrains [17] (see also [21]), and analytic interpolation [24, 25].

Our present paper builds on the approach in [24], [25], which formulates a sufficient condition for the maximum delay margin in terms of an interpolation problem with a real weight and obtains a lower bound using a rational approximation of the weight. In the present paper we instead reformulate the interpolation problem

as an infinite dimensional analytic interpolation problem and solve it directly using techniques from function theory and complex analysis. This is related to work on discrete time systems in [18, 19]; methods that can also be used for control design and implementation. In addition, by introducing a constant shift, we show that the lower bound can be further improved. In this short paper we concentrate on the delay margin itself and leave a deeper study of control implementation to a future paper.

The outline of the paper is as follows. In Section C.2 we define the delay margin problem and describe the results in [23], [24], [25]. In Section C.3 we modify the approach of [24], [25] to obtain better lower bounds and provide an algorithm for this. This method is then improved in Section C.4 by a simple shift of the corresponding complementary sensitivity function. Section C.5 is devoted to some numerical simulations. To facilitate comparison with the results in [25] we use some of the same systems as there. In Section C.6 we provide a succinct discussion of control implementation, and in Section C.7 we discuss some possible future directions of research.

C.2 The delay margin problem

Let $P(s)$ be the transfer function of a continuous-time, finite-dimensional, single-input-single-output LTI system, and consider the feedback control system depicted in Figure C.1. Here $e^{-\tau s}$ is a delay, and $K(s)$ is a feedback controller in the class

$$\mathcal{F}(\mathcal{H}_\infty) := \left\{ \frac{N(s)}{D(s)} \;\middle|\; N, D \in \mathcal{H}_\infty(\mathbb{C}_+) \text{ and } D(s) \not\equiv 0 \right\},$$

where $\mathbb{C}_+$ denotes the open right half plane, and $\mathcal{H}_\infty(\mathbb{C}_+)$ denote the Hardy space of bounded analytic functions on $\mathbb{C}_+$; see, e.g., [7]. The basic problem in control theory is to find a $K(s)$ in this quotient field that stabilizes the closed loop system for a class of systems.

Let us first consider the standard problem without delay ($\tau = 0$). The closed loop system is stable if

$$1 + P(s)K(s) \neq 0 \quad \text{for all } s \in \bar{\mathbb{C}}_+, \tag{C.2.1}$$

where $\bar{\mathbb{C}}_+$ is the closed right half plane. This is equivalent to that the sensitivity function

$$S(s) := (1 + P(s)K(s))^{-1}$$

belongs to $\mathcal{H}_\infty$, which in turn is equivalent to $T \in \mathcal{H}_\infty$, where

$$T(s) := 1 - S(s) = P(s)K(s)\big(1 + P(s)K(s)\big)^{-1}$$

is the complementary sensitivity function [5]. The feedback system is *internally stable* if, in addition, there is no pole-zero cancellation between P and K in $\bar{\mathbb{C}}_+$ [5,

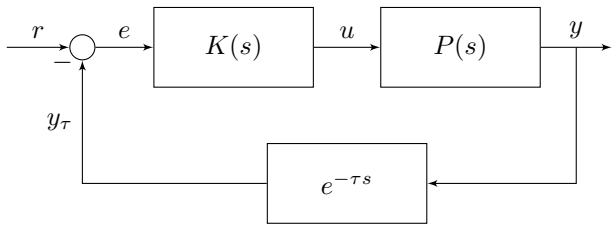


Figure C.1: Block diagram representation of an LTI system with time delay.

pp. 35-36]. Assuming for simplicity that the poles and zeros are distinct, this is equivalent to the interpolation conditions[1]

$$T(p_j) = 1, \quad j = 1, \dots, n, \tag{C.2.2a}$$
$$T(z_j) = 0, \quad j = 1, \dots, m, \tag{C.2.2b}$$

where $p_1, \dots, p_n$ are the unstable poles and $z_1, \dots, z_m$ the nonminimum phase zeros of P, respectively; see, e.g., [27], [12, Chapters 2 and 7]. In the sequel we shall simply say that K stabilizes P when all these conditions are satisfied.

If K stabilizes P, by continuity it also stabilizes $Pe^{-\tau s}$ for sufficiently small $\tau > 0$. The question is how large τ can be while retaining internal stability. Following [23] we define the *delay margin* for a given controller K as

$$DM(P, K) := \sup_{\tau \geq 0} \ \tau \quad \text{such that } K \text{ stabilizes } Pe^{-ts} \text{ for } t \in [0, \tau],$$

and the *maximum delay margin* for a plant P as

$$\tau_{\max} = DM(P) := \sup_{K \in \mathcal{F}(\mathcal{H}_\infty)} DM(P, K).$$

This means that $\tau_{\max}$ is the largest value such that for any $\bar{\tau} < \tau_{\max}$ there exists a controller K that stabilizes the plant P for all τ in the interval $[0, \bar{\tau}]$. If the plant P is stable we trivially have $\tau_{\max} = \infty$, since $K \equiv 0$ stabilizes it, and thus we shall only consider unstable plants.

To determine $\tau_{\max}$ is in general a hitherto unsolved problem, but work has been done to obtain lower and upper bounds.

Upper bounds for maximum delay margin

In [23] it was shown that for any strictly proper real-rational plant P with unstable poles in $re^{\pm i\theta}$, $r > 0$ and $\theta \in [0, \pi/2]$, there is an upper bound $\bar{\tau}$ for $\tau_{\max}$ given by

$$\bar{\tau} = \frac{1}{r}\Big(\pi \sin(\theta) + 2 \max\{\cos(\theta),\ \theta \sin(\theta)\}\Big) \tag{C.2.3}$$

[1] If the poles and zeros are not distinct the interpolation conditions need to be imposed with multiplicity [27].

[23, Thm. 7, 9 and 11]. Moreover, this upper bound is in fact shown to be tight in the special cases of either exactly one real unstable pole or exactly two conjugate unstable poles. These results are the first that show that there is an upper bound on the achievable delay margin when using LTI controllers, and they describe a region for the delay where stabilization is not possible. However, the provided bounds of the maximum delay margin are in general not tight, and have lately also been improved upon in [15, 16].

Lower bounds for maximum delay margin

To ensure stability we are in general more interested in a lower bound $\bar{\tau} \leq \tau_{\max}$. This problem is considered in the recent papers [24, 25], where an approach based on analytic interpolation and rational approximations is taken. The starting point is that (C.2.1) can be written

$$1 + T(s)(e^{-\tau s} - 1) \neq 0 \quad \text{for } s \in \bar{\mathbb{C}}_+, \tag{C.2.4}$$

where T is the complementary sensitivity function. A sufficient condition for (C.2.4) to hold for all τ on an interval $[0, \bar{\tau}]$ is that

$$\sup_{\tau \in [0,\bar{\tau}]} \inf_{\substack{T \in \mathcal{H}_\infty \\ \text{subject to (C.2.2)}}} \|T(s)(e^{-\tau s} - 1)\|_{\mathcal{H}_\infty} < 1. \tag{C.2.5}$$

Now, since sup inf $\leq$ inf sup, this condition holds whenever

$$\inf_{\substack{T \in \mathcal{H}_\infty \\ \text{subject to (C.2.2)}}} \|T(i\omega)\phi_{\bar{\tau}}(\omega)\|_{L_\infty} < 1, \tag{C.2.6}$$

where

$$\begin{aligned} \phi_{\bar{\tau}}(\omega) &= \sup_{\tau \in [0,\bar{\tau}]} |e^{-i\tau\omega} - 1| \\ &= \begin{cases} 2\left|\sin(\frac{\bar{\tau}\omega}{2})\right| & \text{for } |\omega\bar{\tau}| \leq \pi \\ 2 & \text{for } |\omega\bar{\tau}| > \pi. \end{cases} \end{aligned} \tag{C.2.7}$$

In [25] the function $\phi_{\bar{\tau}}$ is approximated by the magnitude of a rational function $w_{\bar{\tau}}$ such that $\phi_{\bar{\tau}}(\omega) \leq |w_{\bar{\tau}}(i\omega)|$ for all ω. Using this approximation and the interpolation conditions on T for internal stability the authors derive an algorithm for computing the largest $\bar{\tau}$ for which (C.2.6) holds. This thus gives a lower bound for the maximum delay margin.

C.3 Formulating and solving (C.2.6) using analytic interpolation

In this section we will solve the problem (C.2.6) directly using analytic interpolation without resorting to approximation of $\phi_{\bar{\tau}}(\omega)$ via rational functions. Continuing in

the manner of [25] we note that (C.2.6), the sufficient condition for the closed loop system to be internally stable for all $\tau \in [0, \bar{\tau}]$, holds if there exists a $T(s) \in \mathcal{H}_\infty(\mathbb{C}_+)$ such that

$$\|T(i\omega)\phi_{\bar{\tau}}(\omega)\|_{L_\infty} < 1 \text{ and } \begin{cases} T(p_j) = 1, \ j = 1, \dots, n, \\ T(z_j) = 0, \ j = 1, \dots, m. \end{cases} \tag{C.3.1}$$

Next, we may replace $\phi_{\bar{\tau}}$ by the outer function $W_{\bar{\tau}} \in \mathcal{H}_\infty(\mathbb{C}_+)$ with the same magnitude as $\phi_{\bar{\tau}}$ on $i\mathbb{R}$ [13, p. 133], and we arrive at the equivalent problem

$$\|TW_{\bar{\tau}}\|_{\mathcal{H}_\infty} < 1 \text{ and } \begin{cases} T(p_j) = 1, \ j = 1, \dots, n, \\ T(z_j) = 0, \ j = 1, \dots, m, \end{cases} \tag{C.3.2}$$

where

$$W_{\bar{\tau}}(s) = \exp\left[\frac{1}{\pi}\int_{-\infty}^{\infty} \log\left(\phi_{\bar{\tau}}(\omega)\right) \frac{\omega s + i}{\omega + is} \frac{1}{1+\omega^2}\, d\omega\right]. \tag{C.3.3}$$

Observing that $W_{\bar{\tau}}$ is outer, and setting $\tilde{T} := TW_{\bar{\tau}}$, (C.3.2) is seen to be equivalent to

$$\|\tilde{T}\|_{\mathcal{H}_\infty} < 1 \text{ and } \begin{cases} \tilde{T}(p_j) = W_{\bar{\tau}}(p_j), & j = 1, \dots, n, \\ \tilde{T}(z_j) = 0, & j = 1, \dots, m, \end{cases} \tag{C.3.4}$$

and thus the only way the weight enters is through the values of the outer function $W_{\bar{\tau}}$ at the pole locations p_j [18, Section 4.C] (cf. [19]). Since $W_{\bar{\tau}}$ is outer, no unstable poles or nonminimum-phase zeros have been added in $\mathbb{C}_+$.

Hence we have reduced the problem to determining whether there exists a $\tilde{T} \in \mathcal{H}_\infty$ such that (C.3.4) holds. The values $W_{\bar{\tau}}(p_j)$, $j = 1, \dots, n$, can be computed from (C.3.3) by numerical integration. Then setting

$$v := [p_1, \dots, p_n, z_1, \dots, z_m] \tag{C.3.5a}$$

$$w := [W_{\bar{\tau}}(p_1), \dots, W_{\bar{\tau}}(p_n), 0, \dots, 0], \tag{C.3.5b}$$

the interpolation problem (C.3.4) is solvable if and only if the corresponding Pick matrix

$$\text{Pick}(v, w) := \left[\tfrac{1 - w_j \bar{w}_k}{v_j + \bar{v}_k}\right]_{j,k=1}^{n+m} \tag{C.3.6}$$

is positive definite; see, e.g., [5, pp. 151-152]. In case the poles and zeros are not distinct, (C.3.6) needs to be replaced by a more general criterion, e.g., using the input-to-state framework [3, 10] as in [2].

We have thus shown that for a given $\bar{\tau}$, the problem (C.2.6) has a solution if and only if the Pick matrix (C.3.6) with interpolation values (C.3.5) is positive definite. Moreover, if (C.2.6) has a solution for some $\bar{\tau}$ then clearly it has a solution for any smaller value, since $\phi_{\bar{\tau}}(\omega)$ is point-wise nondecreasing in $\bar{\tau}$. Therefore the optimal $\bar{\tau}$ can be computed using the bisection algorithm, iteratively testing feasibility of (C.2.6). The method is summarized in Algorithm C.1. Note that by (C.2.3) we have

Algorithm C.1 Lower bound on maximum delay margin

Input: Unstable poles p_j, $j = 1, \ldots, n$, and nonminimum phase zeros z_j, $j = 1, \ldots, m$, of the plant P.

$\tau_- = 0$.
$\tau_+ = 2\pi/\max_j(|p_j|)$,
while $\tau_+ - \tau_- >$ `tol` **do**
  $\tau_{\text{mid}} = (\tau_+ + \tau_-)/2$
  Compute new interpolation values $W_{\tau_{\text{mid}}}(p_j)$
  if Pick matrix (C.3.6) with values (C.3.5) is positive definite **then**
    $\tau_- = \tau_{\text{mid}}$
  else
    $\tau_+ = \tau_{\text{mid}}$
  end if
end while
$\bar{\tau} = \tau_-$

Output: $\bar{\tau}$, lower bound on maximum delay margin

$2\pi/\max_j(|p_j|) \geq \tau_{\max}$, which gives a valid choice for the initial upper bound in the bisection algorithm.

The improvement of this method over that in [25] depends on how well the magnitude of the fifth-order approximation $w_{6\tau}(i\omega)$ used in [25] fits $\phi_{\bar{\tau}}(\omega)$ for $\omega \in \mathbb{R}$. To illustrate this, the relative error for $\bar{\tau} = 1$ is shown in Figure C.2. In this particular case only a minor improvement in the lower bound is expected.

However, our formulation of the problem allows for adding further constraints to the interpolation problem. This can be done in order to shape the sensitivity function, similarly to what has been done for discrete time systems in [18]. In the current setting this can be achieved by letting ϕ_{design} be the modulus on the imaginary axis of the designed weight function and by considering $\|T(i\omega)\phi_{\max}(\omega)\|_{L_\infty} < 1$ in (C.3.1) instead, where $\phi_{\max}(\omega) = \max\{\phi_{\bar{\tau}}(\omega), \phi_{\text{design}}(\omega)\}$.

C.4 Improving the lower bound using a constant shift

Consider the constraint $\|T(i\omega)\phi_{\bar{\tau}}(\omega)\|_{L_\infty} < 1$ in (C.3.1). For each ω the image of the complementary sensitivity function, $T(i\omega)$, is confined to a ball centered at the origin and with radius $|\phi_{\bar{\tau}}(\omega)^{-1}|$. However, choosing the center of the ball at the origin is quite arbitrary, and by instead carefully selecting the center elsewhere, we may improve the estimate of the lower bound. To this end, let $T = \hat{T} + w_0$ where $w_0 \in \mathbb{C}$. The condition (C.2.4) can then be written

$$\hat{T}(s)\big(e^{-\tau s} - 1\big) \neq -1 + w_0 - w_0 e^{-\tau s}. \tag{C.4.1}$$

Here the right hand side is an $\mathcal{H}_\infty$ function, and it is nonzero in all of $\bar{\mathbb{C}}_+$ if and only if $\Re(w_0) < 1/2$, as can be seen from Lemma C.8.1 in the appendix. Consequently,

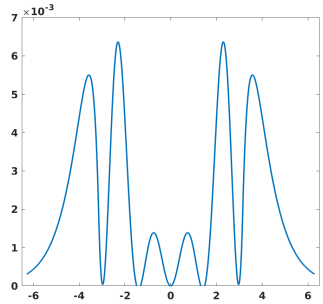


Figure C.2: Relative error between $\phi_{\bar{\tau}}$ and the magnitude of fifth-order approximation $w_{6\tau}$ in [25], for $\bar{\tau} = 1$. The relative error is given point-wise by $\big(|w_{6\tau}(i\omega)| - \phi_{\bar{\tau}}(\omega)\big)/\phi_{\bar{\tau}}(\omega)$.

for $\Re(w_0) < 1/2$, the inverse is an $\mathcal{H}_\infty$ function and thus (C.4.1) can be written as

$$\hat{T}(s)\frac{e^{-\tau s} - 1}{1 - w_0 + w_0 e^{-\tau s}} \neq -1. \tag{C.4.2}$$

Hence we need modify the function $\phi_{\bar{\tau}}$ in Section C.3 to read

$$\phi_{\bar{\tau}}(\omega) := \sup_{\tau \in [0,\bar{\tau}]} \left| \frac{e^{-\tau i\omega} - 1}{1 - w_0 + w_0 e^{-\tau i\omega}} \right|,$$

which reduces to (C.2.7) when $w_0 = 0$. Then using the same argument as before, we see that

$$\|\hat{T}(i\omega)\phi_{\bar{\tau}}(\omega)\|_{L_\infty} < 1$$

is a sufficient condition for (C.4.2) to hold.

As shown in the appendix, $\phi_{\bar{\tau}}(\omega)$ can be determined in closed form, i.e.,

$$\phi_{\bar{\tau}}(\omega)^{-1} = \begin{cases} 0.5 - \Re(w_0), & \omega \geq \bar{\omega}_+, \\ |0.5 - i0.5\cot(\omega\bar{\tau}/2) - w_0|, & \bar{\omega}_+ > \omega > \bar{\omega}_-, \\ 0.5 - \Re(w_0), & \omega \leq \bar{\omega}_-, \end{cases} \tag{C.4.3}$$

where $\bar{\omega}_+$ and $\bar{\omega}_-$ are defined as follows: first define

$$\bar{\omega} := \frac{2}{\bar{\tau}} \cot^{-1}(-2 \cdot \Im(w_0)),$$

where we set $\cot^{-1}(0) = \pi/2$. Moreover, note that $\bar{\omega} \neq 0$ for any finite w_0. Next, define $\bar{\omega}_+$ and $\bar{\omega}_-$ by first setting $\bar{\omega}_+ = \bar{\omega}$ if $\bar{\omega} > 0$ or $\bar{\omega}_- = \bar{\omega}$ if $\bar{\omega} < 0$ and then defining the remaining variable via

$$\bar{\omega}_+ = \bar{\omega}_- + 2\pi/\bar{\tau}.$$

Following the procedure in Section C.3 we define, via the representation (C.3.3), an outer function $W_{\bar{\tau}}(s)$ with the property $|W_{\bar{\tau}}(i\omega)| = \phi_{\bar{\tau}}(\omega)$ for all points on the imaginary axis. Consequently, we are left with the problem to find a $\hat{T}$ such that

$$\|\hat{T}W_{\bar{\tau}}\|_{\mathcal{H}_\infty} < 1 \text{ and } \begin{cases} \hat{T}(p_j) = 1 - w_0, & j = 1, \dots, n, \\ \hat{T}(z_j) = -w_0, & j = 1, \dots, m, \end{cases}$$

which, in turn, is equivalent to

$$\|\tilde{T}\|_{\mathcal{H}_\infty} < 1 \text{ and } \begin{cases} \tilde{T}(p_j) = (1 - w_0)W_{\bar{\tau}}(p_j), & j = 1, \dots, n, \\ \tilde{T}(z_j) = -w_0 W_{\bar{\tau}}(z_j), & j = 1, \dots, m. \end{cases}$$

In the same manner as in Section C.3 we can then determine feasibility by checking whether the corresponding Pick matrix (C.3.6) is positive definite. A refined algorithm for computing a lower bound for the maximum delay margin is thus obtained by suitable changes in Algorithm C.1.

C.5 Numerical example

In this section we investigate the performance of the method proposed in Section C.4 on some examples. To facilitate comparison with the results of [25] we consider the various SISO-systems given in [25, Ex.1].

Systems with one unstable pole and one nonminimum phase zero

We begin with the system [25, Eq. (41)], i.e.,

$$P(s) = \frac{s - z}{s - p}, \tag{C.5.1}$$

where $z, p > 0$. As in [25] we set $z = 2$ and compute an estimate for the delay margin for different values of p in the interval $[0.3, 4]$. Results are shown in Figure C.3. From this we can see that with $w_0 = -10$ we get a considerable improvement over the bound in [25] in the region $p < z = 2$, and in this case we get close to the theoretical bound from [23] (which is tight in this region). However, with $w_0 = -10$ our method seems to perform worse than [25] in the region $p > z = 2$. On the other hand, in this region the value $w_0 = 0.35$ achieves some improvement. Note that the true stability margin is, to the best of our knowledge, still unknown in this region.

The system [25, Eq. (42)], given by

$$P(s) = 0.1\frac{(0.1s - 1)(s + 0.1659)}{(s - 0.1081)(s^2 + 0.2981s + 0.06281)}, \tag{C.5.2}$$

has similar characteristics as the previous example, with one unstable pole ($p = 0.1081$) and nonminimim phase zero ($z = 10$). Also in this case our method gives a considerable improvement over [25] when w_0 is selected to be negative, and as w_0 tends to $-\infty$ our bound seems to approach the theoretical bound $2/0.1081 - 2/10 \approx 18.3$ from [23]; see Figure C.4.

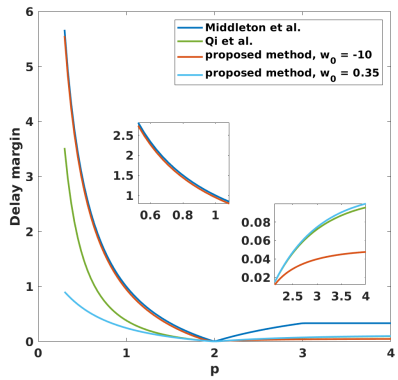


Figure C.3: Results for the example in (C.5.1).

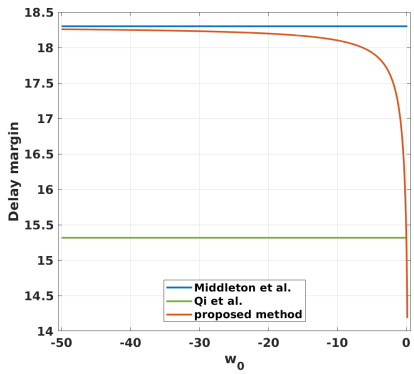


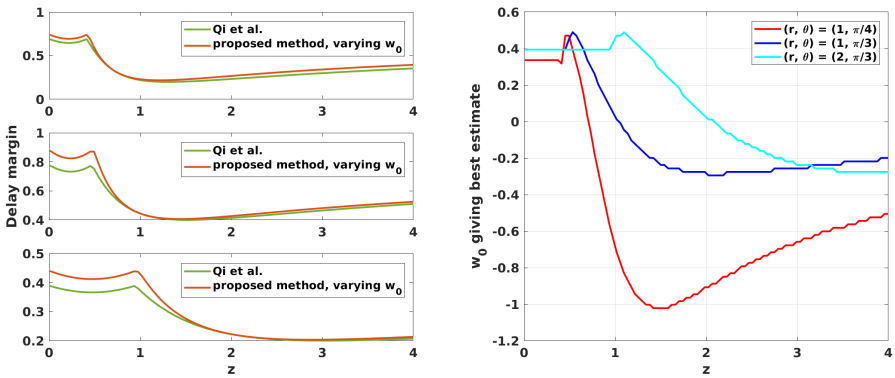
Figure C.4: Results for the example in (C.5.2), with w_0 real. When w_0 goes to $-\infty$ we seem to get arbitrarily close to the result by Middleton et al. [23], while for $w_0 > 0$ the bound deteriorate quickly.

System with two unstable real poles

Next we consider the system [25, Eq. (40)], given by

$$P(s) = \frac{1}{(s - p_1)(s - p_2)}.$$

In this case p_1 is fixed to 0.2, and the delay margin computed for different values of $p_2 \in [0.1, 3]$. Then for values of $w_0 \in [-10, 0.5)$ only minor improvements over the result in [25] are achieved; for the corresponding optimal choice of w_0, the improvements are between 0.19% and 2.9% depending on p_2.



(a) Estimates of the delay margin for the cases, from top to bottom, $(r,\theta) = (1,\pi/4)$, $(1,\pi/3)$, and $(r,\theta) = (2,\pi/3)$.

(b) Best choice of w_0 as function of the zero position z.

Figure C.5: Reults for the example in (C.5.3).

System with conjugate pair of complex poles

Finally we consider the system [25, Eq. (45)], which has a pair of unstable complex poles and a nonminimum phase zero. This system is given by

$$P(s) = \frac{s-z}{(s-re^{i\theta})(s-re^{-i\theta})}, \tag{C.5.3}$$

and we compute an estimate of the delay margin for three fixed values of the pair (r,θ), namely for $(r,\theta) = (1,\pi/4)$, $(r,\theta) = (1,\pi/3)$, and $(r,\theta) = (2,\pi/3)$. Moreover, for these values of (r,θ) we vary z in $[0.01, 4]$ and for each value of z we investigate all values of $w_0 \in [-1.5, 0.5)$ (with steps 0.02) to find the w_0 that maximizes the estimated delay margin. Results are shown in Figure C.5, where Figure C.5a shows the estimated delay margin and Figure C.5b shows the corresponding best value of w_0. The proposed method gives significantly improved bounds in some regions, for example when $\theta = \pi/3$ and z is small compared to r.

C.6 On the control implementation

There are certain problems with the implementation of the stabilizing controller that need attention. The complementary sensitivity function is given by

$$T(s) = \tilde{T}(s)W_{\bar{\tau}}(s)^{-1} + w_0. \tag{C.6.1}$$

Indeed, since $W_{\bar{\tau}}$ is outer, it is nonzero in $\mathbb{C}_+$, and hence it can be inverted there. However, since $W_{\bar{\tau}}(0) = 0$, T typically has a pole in $s = 0$, and therefore the closed

loop system may not be stable (cf. [5, p. 36]). This can be rectified by replacing $\phi_{\bar{\tau}}$ by

$$\phi_{\bar{\tau},\varepsilon}(\omega) = \max(\varepsilon, \phi_{\bar{\tau}}(\omega))$$

for $\varepsilon > 0$. This will give a stable system and, by continuity, as $\varepsilon \to 0$ we can obtain a maximum delay margin estimate arbitrary close to $\bar{\tau}$.

Selecting $\bar{\tau}$ to be the supremum for which (C.2.6) holds gives rise to a singular Pick matrix (C.3.6) and a unique solution $\tilde{T}$ which is a Blaschke product [9, pp. 5-9], so $\|\tilde{T}\|_{\mathcal{H}_\infty} = 1$. Such a solution will not satisfy (C.2.6) and thus may not have delay margin $\bar{\tau}$. However, for any $\bar{\tau}$ smaller than the supremum the Pick matrix is positive definite and the analytic interpolation problem (e.g., (C.3.4)) has infinitely many rational solutions [3], [6]. We must now choose such a solution appropriately so that the stabilizing controller

$$K = P^{-1}(\tilde{T} + (1 - w_0)W_{\bar{\tau}})^{-1}(\tilde{T} + w_0 W_{\bar{\tau}}), \tag{C.6.2}$$

is a rational function and thus can be implemented by a finite-dimensional system. Hence, unlike the approach in [24, 25], an approximation may be needed to design the controller. Again, methods similar to the ones presented in [18] can be used to obtain such an approximation, but details are left for a forthcoming paper.

C.7 Conclusions and future directions

In this work we build on the approach in [24], [25] for computing a lower bound for the maximum delay margin of a system. We introduce a parameter that can be tuned to improve the bounds, and in numerical examples we can in some cases come (arbitrarily) close to the true upper bound. Subsequent work will focus on why this is the case, but also on how to tune the method and how to construct implementable controllers; the latter by following along the lines of [3], [6], [18].

Acknowledgement

We would like to thank Jie Chen for introducing us to the problem and for helpful discussions. We would also like to thank the referees for useful suggestions and comments.

C.8 Appendix

A bound on $\Re(w_0)$

Lemma C.8.1. *For $\tau > 0$ the function $h(s) = -1 + w_0 - w_0 e^{-\tau s}$ is nonzero in $\bar{\mathbb{C}}_+$ if and only if $\Re(w_0) < 1/2$.*

Proof. Suppose $\tau > 0$. If $\Re(w_0) < 0$, $h(s)$ is trivially nonzero for all $s \in \bar{\mathbb{C}}_+$. Consequently we need only consider the case $\Re(w_0) \geq 0$. Then $\{w_0 e^{-\tau s} \mid s \in \bar{\mathbb{C}}_+\} = |w_0|\bar{\mathbb{D}}$, where $\bar{\mathbb{D}}$ is the closed unit disc $\{s \in \mathbb{C} \mid |s| \leq 1\}$. Therefore $h(s)$ is nonzero if and only if $1 - w_0 \notin |w_0|\bar{\mathbb{D}}$, which is true if and only if $|1 - w_0| > |w_0|$ which in turn is true if and only if $\Re(w_0) < 1/2$. □

Computing $\phi_{\bar{\tau}}(\omega)$

Since $\sup_x f(x) = \inf_x 1/f(x)$ we have that

$$\phi_{\bar{\tau}}(\omega)^{-1} = \inf_{\tau \in [0,\bar{\tau}]} |w_0 - g(\omega, \tau)|,$$

where $g(\omega, \tau) := (1 - e^{-i\omega\tau})^{-1}$. Introducing the set

$$A_{\bar{\tau}}(\omega) := \big\{ g(\omega, \tau) \mid \tau \in [0, \bar{\tau}] \big\}$$

for each ω,

$$\phi_{\bar{\tau}}(\omega)^{-1} = \operatorname{dist}\Big(w_0 \ , \ A_{\bar{\tau}}(\omega) \Big),$$

where $\operatorname{dist}(s_1, C) := \inf_{s_2 \in C} |s_1 - s_2|$ denotes the distance between a point and a set. Next we note that

$$g(\omega, \tau) = \frac{1}{2} - \frac{i}{2} \frac{\sin(\omega\tau)}{1 - \cos(\omega\tau)} = \frac{1}{2} - \frac{i}{2} \cot\left(\frac{\omega\tau}{2} \right),$$

so $\Im\big(g(\omega, \tau)\big)$ is a monotone increasing function of the product $\omega\tau$ in any interval $(0, 2\pi) + k \cdot 2\pi$, $k \in \mathbb{Z}$. Moreover,

$$g(\omega, \tau) - w_0 = \left[\frac{1}{2} - \Re(w_0) \right] + i \left[\Im\big(g(\omega, \tau)\big) - \Im(w_0) \right],$$

where the real part is positive since we need $\Re(w_0) < 1/2$ by Lemma C.8.1. Therefore $|g(\omega, \tau) - w_0|$ will take a minimum value when $\big|\Im\big(g(\omega, \tau)\big) - \Im(w_0)\big|$ is as small as possible. For a fixed $\omega \geq 0$, consider three cases. First, if $\omega \leq 2\pi/\bar{\tau}$ and if $\Im\big(g(\omega, \bar{\tau})\big) \geq \Im(w_0)$, then, since $\Im\big(g(\omega, \tau)\big)$ is monotone increasing in τ, $1/2 - \Im(w_0) \in A_{\bar{\tau}}(\omega)$ and $|g(\omega, \tau) - w_0| \geq \frac{1}{2} - \Re(w_0)$, and hence

$$\operatorname{dist}\Big(w_0 \ , \ A_{\bar{\tau}}(\omega) \Big) = \frac{1}{2} - \Re(w_0).$$

Second, if $\omega > 2\pi/\bar{\tau}$ the argument can be reduced to the above one by noticing that $\Im\big(g(\omega, \tau)\big)$ is 2π-periodic in $\omega\tau$ and that $[0, 2\pi] \subset \{\omega\tau \mid \tau \in [0, \bar{\tau}]\}$. Third, if $\Im\big(g(\omega, \bar{\tau})\big) < \Im(w_0)$, then the minimum will be obtained for $\tau = \bar{\tau}$, so

$$\operatorname{dist}\Big(w_0 \ , \ A_{\bar{\tau}}(\omega) \Big) = |w_0 - g(\omega, \bar{\tau})|.$$

In the same manner we obtain the analogous results for negative ω. Now define $\bar{\omega}_+ \in (0, 2\pi/\bar{\tau})$ to be the value of ω for which $\Im\big(g(\bar{\omega}_+, \bar{\tau})\big) = \Im(w_0)$, and let $\bar{\omega}_- \in (-2\pi/\bar{\tau}, 0)$ be the corresponding negative value. These are the frequencies at which $\phi_{\bar{\tau}}^{-1}$ changes form. Moreover, they can be computed by using $\bar{\omega}$ as in Section C.4.

References

[1] I. Alterman and L. Mirkin. On the robustness of sampled-data systems to uncertainty in continuous-time delays. *IEEE Transactions on Automatic Control*, 56(3):686–692, 2011.

[2] A. Blomqvist and R. Nagamune. Optimization-based computation of analytic interpolants of bounded complexity. *Systems & Control Letters*, 54(9):855–864, 2005.

[3] 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.

[4] J. Chen, G. Gu, and C.N. Nett. A new method for computing delay margins for stability of linear delay systems. *Systems & Control Letters*, 26(2):107–117, 1995.

[5] J.C. Doyle, B.A. Francis, and A.R. Tannenbaum. *Feedback control theory.* Macmillan, 1992.

[6] G. Fanizza, J. Karlsson, A. Lindquist, and R. Nagamune. Passivity-preserving model reduction by analytic interpolation. *Linear Algebra and its Applications*, 425(2-3):608–633, 2007.

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

[8] E. Fridman. *Introduction to time-delay systems.* Birkhäuser, Basel, 2014.

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

[10] T.T. Georgiou. The structure of state covariances and its relation to the power spectrum of the input. *IEEE Transactions on Automatic Control*, 47(7):1056–1066, 2002.

[11] K. Gu, J. Chen, and V.L. Kharitonov. *Stability of time-delay systems.* Birkhäuser, Boston, MA, 2003.

[12] J.W. Helton and O. Merino. *Classical Control Using H^∞ Methods: Theory, Optimization, and Design.* Society for Industrial and Applied Mathematics, 1998.

[13] K. Hoffman. *Banach Spaces of Analytic Functions.* Prentice-Hall, Inc., New Jersey, 1962.

[14] Y.-P. Huang and K. Zhou. Robust stability of uncertain time-delay systems. *IEEE Transactions on Automatic Control*, 45(11):2169–2173, 2000.

[15] P. Ju and H. Zhang. Further results on the achievable delay margin using LTI control. *IEEE Transactions on Automatic Control*, 61(10):3134–3139, 2016.

[16] P. Ju and H. Zhang. Achievable delay margin using LTI control for plants with unstable complex poles. *Science China Information Sciences*, 61(9):092203, 2018.

[17] C.-Y. Kao and A. Rantzer. Stability analysis of systems with uncertain time-varying delays. *Automatica*, 43(6):959–970, 2007.

[18] J. Karlsson, T.T. Georgiou, and A. Lindquist. The inverse problem of analytic interpolation with degree constraint and weight selection for control synthesis. *IEEE Transactions on Automatic Control*, 55(2):405–418, 2010.

[19] J. Karlsson and A. Lindquist. On degree-constrained analytic interpolation with interpolation points close to the boundary. *IEEE Transactions on Automatic Control*, 54(6):1412–1418, 2009.

[20] P. Khargonekar and A. Tannenbaum. Non-Euclidian metrics and the robust stabilization of systems with parameter uncertainty. *IEEE Transactions on Automatic Control*, 30(10):1005–1013, 1985.

[21] A. Megretski and A. Rantzer. System analysis via integral quadratic constraints. *IEEE Transactions on Automatic Control*, 42(6):819–830, 1997.

[22] W. Michiels and S.-I. Niculescu. *Stability and stabilization of time-delay systems: An eigenvalue-based approach.* SIAM, Philadelphia, PA, 2007.

[23] R.H. Middleton and D.E. Miller. On the achievable delay margin using LTI control for unstable plants. *IEEE Transactions on Automatic Control*, 52(7):1194–1207, 2007.

[24] 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.

[25] 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.

[26] Z.-Q. Wang, P. Lundström, and S. Skogestad. Representation of uncertain time delays in the H_∞ framework. *International Journal of Control*, 59(3):627–638, 1994.

[27] D.C. Youla, J.J. Bongiorno Jr, and C.N. Lu. Single-loop feedback-stabilization of linear multivariable dynamical plants. *Automatica*, 10(2):159–173, 1974.

Paper D

Generalized Sinkhorn iterations for regularizing inverse problems using optimal mass transport

Generalized Sinkhorn iterations for regularizing inverse problems using optimal mass transport

by

Johan Karlsson, and Axel Ringh

Abstract

The optimal mass transport problem gives a geometric framework for optimal allocation and has recently gained significant interest in application areas such as signal processing, image processing, and computer vision. Even though it can be formulated as a linear programming problem, it is in many cases intractable for large problems due to the vast number of variables. A recent development addressing this builds on an approximation with an entropic barrier term and solves the resulting optimization problem using Sinkhorn iterations. In this work we extend this methodology to a class of inverse problems. In particular we show that Sinkhorn-type iterations can be used to compute the proximal operator of the transport problem for large problems. A splitting framework is then used to solve inverse problems where the optimal mass transport cost is used for incorporating a priori information. We illustrate the method on problems in computerized tomography. In particular we consider a limited-angle computerized tomography problem, where a priori information is used to compensate for missing measurements.

Keywords: optimal mass transport, Sinkhorn iterations, convex optimization, proximal methods, variable splitting, inverse problems, medical imaging

D.1 Introduction

The optimal mass transport problem provides a useful framework that can be utilized in many contexts, ranging from optimal allocation of resources to applications in imaging and machine learning. In this work we extend this framework and consider optimization problems where the objective function contains an optimal mass transport cost. This includes several inverse problems of interest, for example to model deformations in the underlying object.

The optimal mass transport problem is sometimes referred to as the Monge-Kantorovich transportation problem and was originally formulated by Gaspard Monge in 1781 [67] for transport of soil for construction of forts and roads in order to minimize the transport expenses. He described the problem as follows: “divide two equal volumes into infinitesimal particles and associate them one to another so that the sum of the path lengths multiplied by the volumes of the particles be

minimum possible" [45]. The second founder of the field is the mathematician and economist Leonid Kantorovich, who made major advances to the area and as part of this reformulated the problem as a convex optimization problem along with a dual framework. He later received the Nobel Memorial Prize in Economic Sciences for his contributions to the theory of optimum allocation of resources [66]. For an introduction and an overview of the optimal mass transport problem, see, e.g., [67]. During the last few decades the approach has gained much interest in several application fields, such as image processing [39, 38, 43], signal processing [34, 37, 41], computer vision, and machine learning [57, 50, 18, 59].

In our setting the optimal transportation cost is used as a distance for comparing objects and incorporating a priori information. An important property of this distance is that it does not only compare objects point by point, as standard L^p metrics, but instead quantifies the length with which that mass is moved. This property makes the distance natural for quantifying uncertainty and modeling deformations [41, 46, 47]. More specifically geodesics (in, e.g., the associated Wasserstein-2 metric [67]) preserve "lumpiness," and when linking objects via geodesics of the metric there is a natural deformation between the objects. Such a property appears highly desirable in tracking moving objects and integrating data from a variety of sources (see, e.g., [41]). This is in contrast to the fade-in-fade-out effect that occurs when linking objects in standard metrics (e.g., the L^2 metric).

Although the optimal mass transport problem has many desirable properties it also has drawbacks. Monge's formulation is a nonconvex optimization problem and the Kantorovich formulation results in large-scale optimization problems that are intractable to solve with standard methods even for modest size transportation problems. A recent development addressing this computational problem builds on adding an entropic barrier term and solving the resulting optimization problem using the so-called Sinkhorn iterations [28]. This allows for computing an approximate solution of large transportation problems and has proved useful for many problems where no computationally feasible method previously exists. Examples include computing multi-marginal optimal transport problems and barycenters (centroids) [7] and sampling from multivariate probability distribution [40]. For quadratic cost functions this approach can also be seen as the solution to a Schrödinger bridge problem [21].

In this work we build on these methods and consider variational problems that contain an optimal mass transport cost. In particular, we focus on problems of the form

$$\min_{\mu_{\text{est}}} \quad T_\epsilon(\mu_0, \mu_{\text{est}}) + g(\mu_{\text{est}}), \tag{D.1.1}$$

where T_ϵ is the entropy regularized optimal mass transport cost, and g both quantifies data mismatch and can contain other regularization terms. Typically μ_0 is a given prior and the minimizing argument μ_{est} is the sought reconstruction. This formulation allows us to model deformations, to address problems with unbalanced masses (cf. [37]), and to solve gradient flow problems [8, 55] appearing in the Jordan-

Kinderlehrer-Otto framework [42]. A common technique for solving optimization problems with several additive terms is to utilize the proximal point method [56] together with a variable splitting technique [4, 14, 31, 48]. These methods typically utilize the proximal operator and provide a powerful computational framework for solving composite optimization problems. Similar frameworks have previously been used for solving optimization problems that include a transportation cost. In [8] a fluid dynamics formulation [6] is considered that can be used to compute the proximal operator, and in [55] an entropic proximal operator [64] is used for solving problems of the form (D.1.1). In this paper we propose a new fast iterative computational method for computing the proximal operator of $T_\epsilon(\mu_0, \cdot)$ based on Sinkhorn iterations. This allows us to use splitting methods, such as Douglas-Rachford-type methods [32, 13], in order to solve large scale inverse problems of interest in medical imaging.

The paper is structured as follows. In Section D.2 we describe relevant background material on optimal transport and Sinkhorn iterations, and we also review the proximal point algorithm and variable splitting in optimization. Section D.3 considers the dual problem of (D.1.1) and we introduce generalized Sinkhorn iterations which can be used for computing (D.1.1) efficiently for several cost functions g. In particular the proximal operator of $T_\epsilon(\mu_0, \cdot)$ is computed using iterations with the same computational cost as Sinkhorn iterations. In Section D.4 we describe how a Douglas-Rachford-type splitting technique can be applied for solving a general class of large scale problems on the form (D.1.1), and in Section D.5 we apply these techniques to inverse problems using an optimal transport prior, namely two reconstruction problems in computerized tomography. In Section D.6 we discuss conclusions and further directions. Finally, the paper has four appendices containing deferred proofs, connections with the previous work [55], and details regarding the numerical simulations.

Notation

Next, we briefly introduce some notation. Most operations in this paper are defined elementwise. In particular we use $\odot$, $./$, exp, and log to denote elementwise multiplication, division, exponential function, and logarithm function, respectively. We also use $\leq$ ($<$) to denote elementwise inequality (strict). Let $\mathcal{I}_S(\mu)$ be the indicator function of the set S, i.e., $\mathcal{I}_S(\mu) = 0$ if $\mu \in S$ and $\mathcal{I}_S(\mu) = +\infty$ otherwise. Finally, let $\mathbf{1}_n$ denote the $n \times 1$ (column) vector of ones.

D.2 Background

The optimal mass transport problem and entropy regularization

Monge's formulation of the optimal mass transport problem is set up as follows. Given two nonnegative functions, μ_0 and μ_1, of the same mass, defined on a compact

set $X \subset \mathbb{R}^d$, one seeks the transport function $\phi : X \to X$ that minimizes the transportation cost

$$\int_X c(\phi(x), x)\mu_0(x)dx$$

and that is a mass preserving map from μ_0 to μ_1, i.e.,

$$\int_{x\in A} \mu_1(x)dx = \int_{\phi(x)\in A} \mu_0(x)dx \quad \text{for all } A \subset X.$$

Here $c(x_0, x_1) : X \times X \to \mathbb{R}_+$ is a cost function that describes the cost for transporting a unit mass from x_0 to x_1. It should be noted that this optimization problem is not convex and the formulation is not symmetric with respect to functions μ_0 and μ_1.

In the Kantorovich formulation one instead introduces a transference plan, $M : X \times X \to \mathbb{R}_+$, which characterizes the mass which is moved from x_0 to x_1. This construction generalizes to general nonnegative measures $dM \in \mathcal{M}_+(X \times X)$ and allows for transference plans where the mass in one point in μ_0 is transported to a set of points in μ_1. The resulting optimization problem is convex, and the cost is given by

$$\begin{aligned} T(\mu_0, \mu_1) = \min_{dM\in\mathcal{M}_+(X\times X)} & \int_{(x_0,x_1)\in X\times X} c(x_0, x_1)dM(x_0, x_1) \\ \text{subject to } \mu_0(x_0)dx_0 &= \int_{x_1\in X} dM(x_0, x_1), \\ \mu_1(x_1)dx_1 &= \int_{x_0\in X} dM(x_0, x_1). \end{aligned} \tag{D.2.1}$$

The optimal mass transport cost is not necessarily a metric. However, if X is a separable metric space with metric d and we let $c(x_0, x_1) = d(x_0, x_1)^p$, where $p \geq 1$, then $T(\mu_0, \mu_1)^{1/p}$ is a metric on the set of nonnegative measures on X with fixed mass. This is the so-called Wasserstein metrics.[1] Moreover, $T(\mu_0, \mu_1)$ is weak* continuous on this set [67, Thm. 6.9]. Although the optimal mass transport is only defined for functions (measures) of the same mass, it can also be extended to handle measures with unbalanced masses [37] (see also [22] for recent developments).

In this paper we consider the discrete version of the Kantorovich formulation (D.2.1)

$$\begin{aligned} T(\mu_0, \mu_1) = \min_{M\geq 0} \quad & \operatorname{Tr}(C^T M) \\ \text{subject to} \quad & \mu_0 = M\mathbf{1}_{n_1} \\ & \mu_1 = M^T\mathbf{1}_{n_0}. \end{aligned} \tag{D.2.2}$$

[1] The Wasserstein metric was, as many other concepts in this field, first defined by Kantorovich, and the naming is therefore somewhat controversial; see [66] and [67, pp. 106-107].

In this setting the mass distributions are represented by two vectors $\mu_0 \in \mathbb{R}_+^{n_0}$ and $\mu_1 \in \mathbb{R}_+^{n_1}$, where the element $[\mu_k]_i$ corresponds to the mass in the point $x_{(k,i)} \in X$ for $i = 1, \ldots, n_k$ and $k = 0, 1$. A transference plan is represented by a matrix $M \in \mathbb{R}_+^{n_0 \times n_1}$ where the value $m_{ij} := [M]_{ij}$ denotes the amount of mass transported from point $x_{(0,i)}$ to $x_{(1,j)}$. Such a plan is a feasible transference plan from μ_0 to μ_1 if the row sums of M are μ_0 and the column sums of M are μ_1. The associated cost of a transference plan is $\sum_{i=1}^{n_0} \sum_{j=1}^{n_1} c_{ij} m_{ij} = \mathrm{Tr}(C^T M)$, where $[C]_{ij} = c_{ij} = c(x_{(0,i)}, x_{(1,j)})$ is the transportation cost from $x_{(0,i)}$ to $x_{(1,j)}$.

Even though (D.2.2) is a convex optimization problem, it is in many cases computationally infeasible due to the vast number of variables. The number of variables is $n_0 n_1$, and if one seeks to solve the optimal transport problem between two 256×256 images, this results in more than $4 \cdot 10^9$ variables.

One recent method to approximate the optimal transport problem, which allows for addressing large problems, is to introduce an entropic regularizing term. This was proposed in [28], where the following optimization problem is considered:

$$T_\epsilon(\mu_0, \mu_1) = \min_{M \geq 0} \quad \mathrm{Tr}(C^T M) + \epsilon D(M) \tag{D.2.3a}$$

$$\text{subject to } \mu_0 = M \mathbf{1}_{n_1} \tag{D.2.3b}$$

$$\mu_1 = M^T \mathbf{1}_{n_0}, \tag{D.2.3c}$$

where $D(M) = \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} (m_{ij} \log(m_{ij}) - m_{ij} + 1)$ is a normalized entropy term [27]. This type of regularization is sometimes denoted entropic proximal [64] and has previously been considered explicitly for linear programming [35]. Also worth noting is that $D(M)$ is nonnegative and equal to zero if and only if $M = \mathbf{1}_{n_0} \mathbf{1}_{n_1}^T$.

A particularly nice feature with this problem is that any optimal solution belongs to an a priori known structure parameterized by $n_0 + n_1$ variables via diagonal scaling (see (D.2.6) below). This can be seen by relaxing the equality constraints and considering the Lagrange function

$$\begin{aligned} L(M, \lambda_0, \lambda_1) \quad = \quad & \mathrm{Tr}(C^T M) + \epsilon D(M) \\ & + \lambda_0^T (\mu_0 - M \mathbf{1}_{n_1}) + \lambda_1^T (\mu_1 - M^T \mathbf{1}_{n_0}). \end{aligned} \tag{D.2.4}$$

For given dual variables, $\lambda_0 \in \mathbb{R}^{n_0}$ and $\lambda_1 \in \mathbb{R}^{n_1}$, the minimum m_{ij} is obtained at

$$0 = \frac{\partial L(M, \lambda_0, \lambda_1)}{\partial m_{ij}} = c_{ij} + \epsilon \log(m_{ij}) - \lambda_0(i) - \lambda_1(j) \tag{D.2.5}$$

which can be expressed explicitly as $m_{ij} = e^{\lambda_0(i)/\epsilon} e^{-c_{ij}/\epsilon} e^{\lambda_1(j)/\epsilon}$, or equivalently the solution is of the form

$$M = \mathrm{diag}(u_0) K \mathrm{diag}(u_1) \tag{D.2.6}$$

where $K = \exp(-C/\epsilon)$, $u_0 = \exp(\lambda_0/\epsilon)$, and $u_1 = \exp(\lambda_1/\epsilon)$. A theorem by Sinkhorn [63] states that for any matrix K with positive elements, there exist diagonal

matrices $\mathrm{diag}(u_0)$ and $\mathrm{diag}(u_1)$ with $u_0, u_1 > 0$ such that $M = \mathrm{diag}(u_0)K\mathrm{diag}(u_1)$ has prescribed row sums and columns sums (i.e., M satisfies (D.2.3b) and (D.2.3c)). Furthermore, the vectors u_0 and u_1 may be found by Sinkhorn iterations, i.e., alternatively solving (D.2.3b) for u_0 and (D.2.3c) for u_1:

$$\mathrm{diag}(u_0)K\mathrm{diag}(u_1)\mathbf{1} = \mu_0 \quad \Rightarrow \quad u_0 = \mu_0./(Ku_1) \tag{D.2.7a}$$

$$\mathrm{diag}(u_1)K^T\mathrm{diag}(u_0)\mathbf{1} = \mu_1 \quad \Rightarrow \quad u_1 = \mu_1./(K^Tu_0). \tag{D.2.7b}$$

The main computational bottlenecks in each iteration are the multiplications Ku_1 and K^Tu_0, and the iterations are therefore highly computationally efficient, in particular for cases where the matrix K has a structure which can be exploited (see the discussion in Section D.3). Furthermore the convergence rate is linear [36] (cf. [20] for generalization to positive functions).

Recently, in [7], it was shown that the same iterative procedure for solving (D.2.3) can also be recovered using Bregman projections [16] or Bregman-Dykstra iterations [5, 15]. This approach was also used to solve several related problems, such as for computing barycenters, multimarginal optimal transport problems, and tomographic reconstruction [7]. It has also been shown that the Sinkhorn iterations can be interpreted as block-coordinate ascent maximization of the dual problem via a generalization of the Bregman-Dykstra iterations [55, Prop. 3.1]. In Section D.3 we derive this result using Lagrange duality. This observation opens up for enlarging the set of optimization problems that fit into this framework and may also result in new algorithms adopted to the dual optimization problem.

Variable splitting in convex optimization

Variable splitting is a technique for solving variational problems where the objective function is the sum of several terms that are simple in some sense (see, e.g., [4, 24, 31]). One of the more common algorithms for variable splitting is ADMM [14], but there are plenty of other algorithms, such as primal-dual forward-backward splitting algorithms [12, 19, 62], primal-dual forward-backward-forward splitting algorithms [25], and Douglas-Rachford-type algorithms [13, 32]. For a good overview see [48]. In this work we will explicitly consider variable splitting using a Douglas-Rachford-type algorithm presented in [13], but in order to better understand how the algorithm works we will first have a look at the proximal point algorithm and basic Douglas-Rachford variable splitting.

The basic idea behind many of the splitting techniques mentioned above springs from the so-called proximal point algorithm for maximally monotone operators [56]. An operator $S : H \to H$, where H is a real Hilbert space with the inner product $\langle \cdot, \cdot \rangle$, is called monotone if

$$\langle z - z', w - w' \rangle \geq 0 \quad \text{for all } z, z' \in H, w \in S(z), \text{ and } w' \in S(z'),$$

and maximally monotone if in addition the graph of S,

$$\{(z, w) \in H \times H \mid w \in S(z)\},$$

is not properly contained in the graph of any other monotone operator. The interest in such operators stems from the fact that the subdifferential of a proper, convex, and lower semi-continuous function f, denoted by ∂f, is a maximally monotone operator [56]. Moreover, a global minimizer of such f is any point z so that $0 \in \partial f(z)$, which we denote by $z \in \mathrm{zer}(\partial f)$. For a maximally monotone operator S and any scalar $\sigma > 0$ the operator $(I - \sigma S)^{-1}$ is called the resolvent operator or proximal mapping. The proximal point algorithm is a fixed-point iteration of the resolvent operator,

$$z^{k+1} = (I - \sigma S)^{-1}(z^k),$$

and if $\mathrm{zer}(S) \neq \emptyset$ then z^k converges weakly to a point $z^\infty \in \mathrm{zer}(S)$ [56]. For the case that $S = \partial f$ the resolvent operator is called the proximal operator and is given by

$$(I - \sigma \partial f)^{-1}(z) = \mathrm{Prox}_f^\sigma(z) := \arg\min_{z'} \left\{ f(z') + \frac{1}{2\sigma} \|z' - z\|_2^2 \right\}. \tag{D.2.8}$$

Hence, fixed-point iterations of the form

$$z^{k+1} = \arg\min_{z'} \left\{ f(z') + \frac{1}{2\sigma} \|z' - z^k\|_2^2 \right\}$$

generates a sequence that converges weakly to a global minimizer of f. The parameter σ determines the weighting between f and the squared norm in H, and can be interpreted as a step length.

When the function to be minimized is a sum of several terms, then the resolvent operator of $S = A + B$, i.e., $(I - \sigma(A + B))^{-1}$, can be approximated in terms of the operators A and B and their resolvent operators $(I - \sigma A)^{-1}$ and $(I - \sigma B)^{-1}$ [31]. This can give rise to fast schemes when the proximal operator of each term in the sum can be computed efficiently. One specific such algorithm, which is globally convergent, is the Douglas-Rachford splitting algorithm [32]. In Section D.4 we will use the splitting algorithm presented in [13], which extends this framework, in order to address a fairly general class of inverse problems.

D.3 The dual problem and generalized Sinkhorn iterations

In this section we will see that the Sinkhorn iteration is identical to block-coordinate ascent of the corresponding dual problem. Further, we will show that this procedure can also be applied to a set of inverse problems where the transportation cost is used as a regularizing term, and, in particular, for computing the proximal operator of $T_\epsilon(\mu_0, \cdot)$.

Sinkhorn iterations and the dual problem

The Lagrangian dual of (D.2.3) is defined as the minimum of (D.2.4) over all $M \geq 0$ [10, p. 485]. In our case this can be obtained by noting that

$$\begin{aligned}
&\min_{M \geq 0} L(M, \lambda_0, \lambda_1) = L(M^*, \lambda_0, \lambda_1) \\
&= \lambda_0^T \mu_0 + \lambda_1^T \mu_1 + \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \left(m_{ij}^* (c_{ij} + \epsilon \log m_{ij}^* - \lambda_0(i) - \lambda_1(j)) + 1 - m_{ij}^* \right) \\
&= \lambda_0^T \mu_0 + \lambda_1^T \mu_1 + \epsilon \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} (1 - m_{ij}^*) \\
&= \lambda_0^T \mu_0 + \lambda_1^T \mu_1 - \epsilon \exp(\lambda_0^T / \epsilon) \exp(-C/\epsilon) \exp(\lambda_1 / \epsilon) + \epsilon n_0 n_1,
\end{aligned}$$

where the optimal solution $M^* = [m_{ij}^*]_{ij}$ is specified by (D.2.5), which is also used in the third and fourth equalities. This gives the following expression for the dual problem, a result which can also be found in [29, Sec. 5].

Proposition D.3.1 ([29])**.** *A Lagrange dual of* (D.2.3) *is given by*

$$\max_{\lambda_0, \lambda_1} \lambda_0^T \mu_0 + \lambda_1^T \mu_1 - \epsilon \exp(\lambda_0^T / \epsilon) \exp(-C/\epsilon) \exp(\lambda_1 / \epsilon) + \epsilon n_0 n_1. \tag{D.3.1}$$

Note the resemblance between the entropy relaxed dual formulation (D.3.1) and the dual of the optimal transport problem (D.2.2) [67]:

$$\max_{\lambda_0, \lambda_1} \quad \mu_0^T \lambda_0 + \mu_1^T \lambda_1 \tag{D.3.2a}$$

$$\text{subject to} \quad \lambda_0 \mathbf{1}_{n_1}^T + \mathbf{1}_{n_0} \lambda_1^T \leq C. \tag{D.3.2b}$$

The difference is that the inequality constraint (D.3.2b) is exchanged for the penalty term

$$-\epsilon \exp(\lambda_0^T / \epsilon) \exp(-C/\epsilon) \exp(\lambda_1 / \epsilon) \tag{D.3.3}$$

in the objective function of (D.3.1). As $\epsilon \to 0$, the value of the barrier term (D.3.3) goes to 0 if the constraint (D.3.2b) is satisfied and to $-\infty$ otherwise.

Next, consider maximizing the dual objective (D.3.1) with respect to λ_0 for a fixed λ_1. This is attained by setting the gradient of the objective function in (D.3.1) with respect to λ_0 equal to zero, hence λ_0 satisfies

$$\mu_0 = \exp(\lambda_0 / \epsilon) \odot \left(\exp(-C/\epsilon) \exp(\lambda_1 / \epsilon) \right).$$

This is identical to the update formula (D.2.7a) for $u_0 = \exp(\lambda_0 / \epsilon)$, corresponding to the Sinkhorn iterations, where as before $u_1 = \exp(\lambda_1 / \epsilon)$. By symmetry, maximizing λ_1 for a fixed λ_0 gives a corresponding expression which is identical to (D.2.7b). Hence the Sinkhorn iterations corresponds to block-coordinate ascent in the dual problem, i.e., iteratively maximizing the objective in (D.3.1) with respect to λ_0 while keeping λ_1 fixed, and vice versa.

Corollary D.3.2 ([55])**.** *The Sinkhorn iteration scheme* (D.2.7) *is a block-coordinate ascent algorithm of the dual problem* (D.3.1)*.*

This was previously observed in [55, Sec. 3.2]. As we will see next, block-coordinate ascent of the dual problem results in fast Sinkhorn-type iterations for several different problems.

Generalized Sinkhorn iterations

Let us go back to the optimization problem (D.1.1) that contains an optimal mass transport cost

$$\min_{\mu_{\text{est}}} \quad T_\epsilon(\mu_0, \mu_{\text{est}}) + g(\mu_{\text{est}}), \tag{D.3.4}$$

where μ_0 is a prior, and g is a term that could include other regularization terms and data mismatch. In order to guarantee that this problem has a solution and is convex, we introduce the following assumption.

Assumption D.3.3. Let g be a proper, convex, and lower semi-continuous function that is finite in at least one point with mass equal to μ_0, i.e., $g(\mu_{\text{est}}) < \infty$ for some μ_{est} with $\sum_{i=1}^{n_0} \mu_0(i) = \sum_{j=1}^{n_1} \mu_{\text{est}}(j)$.

The first part of this assumption is to make the problem convex, and the second part is imposed so that (D.3.4) has a feasible solution. Moreover, note that $T_\epsilon(\mu_0, \mu_{\text{est}}) < \infty$ restricts μ_{est} to a compact set, which guarantees the existence of an optimal solution. Using the definition of the optimal transport cost, the problem (D.3.4) can equivalently be formulated as

$$\begin{aligned} \min_{M \geq 0, \mu_{\text{est}}} \ & \operatorname{Tr}(C^T M) + \epsilon D(M) + g(\mu_{\text{est}}) \\ \text{subject to } & \mu_0 = M \mathbf{1}_{n_1} \\ & \mu_{\text{est}} = M^T \mathbf{1}_{n_0}. \end{aligned} \tag{D.3.5}$$

The Lagrangian dual problem of (D.3.5) can be obtained using the same steps as the derivation of Proposition D.3.1. See Appendix D.7 for details. Results similar to Proposition D.3.4 are also obtained in the recent preprints [23, Thm. 1] and [60].

Proposition D.3.4. *Let $\mu_0 > 0$ be given, and let g satisfy Assumption D.3.3. Then the Lagrange dual of* (D.3.5) *is given by*

$$\max_{\lambda_0, \lambda_1} \lambda_0^T \mu_0 - g^*(-\lambda_1) - \epsilon \exp(\lambda_0^T/\epsilon) \exp(-C/\epsilon) \exp(\lambda_1/\epsilon) + \epsilon n_0 n_1 \tag{D.3.6}$$

and strong duality holds.

The only difference between (D.3.6) and (D.3.1) is that the term $\lambda_1^T \mu_1$ is replaced by $-g^*(-\lambda_1)$, where g^* denotes the dual (or Fenchel) conjugate functional

$$g^*(\lambda) = \sup_{\mu} \left(\lambda^T \mu - g(\mu) \right).$$

Clearly, Proposition D.3.1 is a special case of Proposition D.3.4, and the optimization problem $T_\epsilon(\mu_0, \mu_1)$ in (D.2.3) is recovered from (D.3.4) if μ_{est} is fixed to μ_1, i.e., $g(\mu_{\text{est}}) = \mathcal{I}_{\mu_1}(\mu_{\text{est}})$. Since (D.3.6) is a dual problem, the objective function is concave [10, p. 486], but not necessarily strictly concave (e.g., as in the case (D.3.1)). Moreover, Assumption D.3.3 assures strong duality between (D.3.5) and (D.3.6) (see the proof in Appendix D.7). As for the standard optimal mass transport problem, we now consider block-coordinate ascent to compute an optimal solution. The corresponding optimality conditions are given in the following lemma and are obtained by noting that the optimum is only achieved when zero is a (sub)gradient of (D.3.6) [10, pp. 711-712].

Lemma D.3.5. *For a fixed λ_1, then λ_0 is the maximizing vector of* (D.3.6) *if*

$$\mu_0 = \exp(\lambda_0/\epsilon) \odot (\exp(-C/\epsilon) \exp(\lambda_1/\epsilon)). \tag{D.3.7a}$$

Similarly, for a fixed λ_0, then λ_1 is the maximizing vector of (D.3.6) *if*

$$0 \in \partial g^*(-\lambda_1) - \exp(\lambda_1/\epsilon) \odot \left(\exp(-C^T/\epsilon) \exp(\lambda_0/\epsilon) \right). \tag{D.3.7b}$$

When (D.3.7) can be computed efficiently, block-coordinate ascent could give rise to a fast computational method for solving (D.3.6). In particular this is the case if (D.3.7b) can be solved element by element, i.e., in $\mathcal{O}(n)$ (excluding matrix-vector multiplication $\exp(-C^T/\epsilon) \exp(\lambda_0/\epsilon)$). This is true for several cases of interest.

Example D.3.6. $g(\mu) = \mathcal{I}_{\{\mu_1\}}(\mu)$. This corresponds to the optimal mass transport problem (D.2.3)

Example D.3.7. $g(\mu) = \|\mu - \mu_1\|_1$.

Example D.3.8. $g(\mu) = \frac{1}{2}\|\mu - \mu_1\|_2^2$.

Example D.3.9. $g(\mu) = \frac{1}{2}\|A\mu - \mu_1\|_2^2$ where A^*A is diagonal and invertible.

Example D.3.8 is of particular importance, since this corresponds to computing the proximal operator of the transportation cost, and will be addressed in detail in the next subsection. Note that Lemma D.3.5 can also be expressed in terms of the entropic proximal operator [55]; see Appendix D.9 for details.

Sinkhorn-type iterations for evaluating the proximal operator

The proximal point algorithm [56] and splitting methods [31] are extensively used in optimization, and a key tool is the computation of the proximal operator (D.2.8) (see Sections D.2 and D.4). In order to use these kinds of methods for solving problems of

the form (D.3.4) we consider the proximal operator of the entropy regularized mass transport cost and propose a Sinkhorn-type algorithm for computing the proximal operator of the transportation cost

$$\operatorname{Prox}^{\sigma}_{T_\epsilon(\mu_0,\cdot)}(\mu_1) = \arg\min_{\mu_{\text{est}}} T_\epsilon(\mu_0, \mu_{\text{est}}) + \frac{1}{2\sigma}\|\mu_{\text{est}} - \mu_1\|_2^2.$$

First note that this can be identified with the optimization problem (D.3.5) where the data fitting term and the corresponding conjugate functional are

$$g(\mu) = \frac{1}{2\sigma}\|\mu - \mu_1\|_2^2, \qquad g^*(\lambda) = \lambda^T\left(\mu_1 + \frac{\sigma}{2}\lambda\right).$$

The dual problem is then given by

$$\max_{\lambda_0,\lambda_1} \lambda_0^T\mu_0 + \lambda_1^T(\mu_1 - \frac{\sigma}{2}\lambda_1) - \epsilon\exp(\lambda_0^T/\epsilon)\exp(-C/\epsilon)\exp(\lambda_1/\epsilon) + \epsilon n_0 n_1. \quad \text{(D.3.8)}$$

The optimality conditions corresponding to Lemma D.3.5 lead to the equations (see the proof of Theorem D.3.10 in Appendix D.8 for the derivation)

$$\lambda_0 = \epsilon\log\left(\mu_0./(K\exp(\lambda_1/\epsilon))\right), \quad \text{(D.3.9a)}$$

$$\lambda_1 = \frac{\mu_1}{\sigma} - \epsilon\omega\left(\frac{\mu_1}{\sigma\epsilon} + \log\left(K^T\exp(\lambda_0/\epsilon)\right) - \log(\sigma\epsilon)\right). \quad \text{(D.3.9b)}$$

Here ω denotes the elementwise Wright ω function,[2] i.e., the function mapping $x \in \mathbb{R}$ to $\omega(x) \in \mathbb{R}_+$ for which $x = \log(\omega(x)) + \omega(x)$ [26]. The first equation can be identified with the first update equation (D.2.7a) in the Sinkhorn iteration. Note that the bottlenecks in the iterations (D.3.9) are the multiplications with K and K^T. All other operations are elementwise and can hence be computed in $\mathcal{O}(n)$, where $n = \max(n_0, n_1)$. The full algorithm is presented in Algorithm D.1. This leads to one of our main results.

Theorem D.3.10. *The variables (λ_0, λ_1) in Algorithm D.1 converge to the optimal solution of the dual problem* (D.3.8)*. Furthermore, the convergence rate is locally q-linear.*

Proof. See Appendix D.8. □

This proof is based on the duality (i.e., Proposition D.3.4 and Lemma D.3.5). The algorithm could also be derived directly using Bregman projections [16], similarly to the derivation of the Sinkhorn iteration in [7]. Some remarks are in order regarding the computation of the iterations for the Sinkhorn-type iterations.

[2] Our implementation uses $\omega(x) = W(e^x)$ for $x \in \mathbb{R}$, where W is the Lambert W function [26].

Remark D.3.11. The bottlenecks in the iterations (D.3.9) are the multiplications with the matrices K and K^T. All other operations are elementwise. In many cases of interest the structures of K can be exploited for fast computations. In particular when the mass points (pixel/voxel locations) $x_{(0,i)} = x_{(1,i)}$ are on a regular grid and the cost function is translation invariant, e.g., as in our application example (see (D.10.1)), where the cost function only depends on the distance between the grid points. Then the matrices C and K are multilevel Toeplitz-block-Toeplitz and the multiplication can be performed in $\mathcal{O}(n \log(n))$ using the fast Fourier transform (FFT) (see, e.g., [49]).

Remark D.3.12. In order for (D.2.3) to approximate the optimal mass transportation problem (D.2.2) it is desirable to use a small ϵ. The entropy regularization has a smoothing effect on the transference plan M, and a too large value of ϵ may thus result in an undesirable solution to the variational problem (D.3.4). However, as $\epsilon \to 0$ the problem becomes increasingly ill-conditioned and the convergence becomes slower [28]. To handle the ill-conditioning one can stabilize the computations using logarithmic reparameterizations. One such approach is described in [20], where the variables $\log(u_0), \log(u_1)$ are used, together with appropriate normalization and truncation of the variables, to compute the Sinkhorn iterations (D.2.7). Another approach to handling the ill-conditioning is described in [60], where a different logarithmic reparameterization is used together with an adaptive scheme for scaling of both ϵ and the discretization grid (cf. [23]). However, note that these approaches are not compatible with utilizing FFT computations for the matrix-vector products by exploiting the Toeplitz-block-Toeplitz structure in C (and K). Due to this fact we have not used this type of stabilization.

Algorithm D.1 Generalized Sinkhorn algorithm for evaluating the proximal operator of $T_\epsilon(\mu_0, \cdot)$.

Input: ϵ, C, λ_0, μ_0, μ_1

$K = \exp(-C/\epsilon)$

while Not converged **do**

  $\lambda_0 \leftarrow \epsilon \log\left(\mu_0 ./ (K \exp(\lambda_1/\epsilon))\right)$

  $\lambda_1 \leftarrow \frac{\mu_1}{\sigma} - \epsilon\omega\left(\frac{\mu_1}{\sigma\epsilon} + \log\left(K^T \exp(\lambda_0/\epsilon)\right) - \log(\sigma\epsilon)\right)$

end while

Output: $\mu_{\text{est}} \leftarrow \exp(\lambda_1/\epsilon) \odot (K^T \exp(\lambda_0/\epsilon))$

D.4 Inverse problems with optimal mass transport priors

In this section we use the splitting framework [13] to formulate and solve inverse problems. This is a generalization of the Douglas-Rachford algorithm [32]. In particular this allows us to address large scale problems of the form (D.3.4), since the

proximal operator of the regularized transport problem can be computed efficiently using generalized Sinkhorn iterations.

The theory in [13] provides a general framework that allows for solving a large set of convex optimization problems. Here we consider problems of the form

$$\inf_{z \in H} f(z) + \sum_{i=1}^{m} g_i(L_i z - r_i) \tag{D.4.1}$$

where H is a Hilbert space, $f : H \to \bar{\mathbb{R}}$ is proper, convex, and lower semi-continuous, $g_i : G_i \to \bar{\mathbb{R}}$, where G_i is a Hilbert space and g_i is proper, convex, and lower semi-continuous, and $L_i : H \to G_i$ is a nonzero bounded linear operator, which is a special case of the structure considered in [13].

The problem (D.4.1) can be solved by the iterative algorithm [13, Eq. (3.6)], in which we only need to evaluate the proximal operators $\operatorname{Prox}_f^\tau$ and $\operatorname{Prox}_{g_i^*}^{\sigma_i}$ for $i = 1, \dots, m$. For reference, this simplified version of the algorithm is shown in Algorithm D.2. Note that by Moreau decomposition we have that $\operatorname{Prox}_f^\tau(x) = x - \tau \operatorname{Prox}_{f^*}^{1/\tau}(x/\tau)$ [4, Thm. 14.3], and therefore Algorithm D.2 can be applied as long as either the proximal operators of the functionals f and $\{g_i\}_{i=1}^m$ or the proximal operators of their Fenchel conjugates can be evaluated in an efficient way.

In the following examples we restrict the discussion to the finite-dimensional setting where the underlying set X is a d-dimensional regular rectangular grid with points x_i for $i = 1, \dots n$, and hence the corresponding Hilbert space is $H = \mathbb{R}^n$. Let $A : \mathbb{R}^n \to \mathbb{R}^m$ be a linear operator (matrix) representing measurements, and let $\nabla : \mathbb{R}^n \to \mathbb{R}^{d \times n}$ be a discrete gradient operator[3] based on the grid X. Furthermore, for $Y = (y_1, \dots, y_n) \in \mathbb{R}^{d \times n}$ we let $\|Y\|_{2,1} = \sum_{i=1}^n \|y_i\|_2$ be the isotropic ℓ_1-norm (sometimes called group ℓ_1-norm). With this notation $\|\nabla \mu\|_{2,1}$ is the isotropic total variation (TV) of μ, which is often used as a convex surrogate for the support of the gradient [58, 17]. This terminology allows us to set up a series of optimization problems for addressing inverse problems.

Optimal mass transport priors using variable splitting

Next, we use optimal mass transport for incorporating a priori information. We consider a particular case of (D.3.4) for achieving this and formulate the reconstruction problem as follows:

$$\begin{aligned} \min_{\mu_{\text{est}}} \quad & \gamma T_\epsilon(\mu_0, \mu_{\text{est}}) + \|\nabla \mu_{\text{est}}\|_{2,1} \\ \text{subject to} \quad & \|A\mu_{\text{est}} - b\|_2 \leq \kappa. \end{aligned} \tag{D.4.2}$$

Here μ_0 is a prior, κ quantifies the allowed measurement error, and γ determines the trade off between the optimal transport prior and the TV-regularization. Since

[3] In each dimension of the regular rectangular grid a forward-difference is applied and the boundary of the domain is padded with zeros, i.e., $(\nabla \mu)_{j,i}$ is the forward-difference along the j-axis in the grid point x_i. See Appendix D.10 for more details.

Algorithm D.2 Douglas-Rachford-type primal-dual algorithm [13].

Input: $\tau, (\sigma_i)_{i=1}^m, (\lambda_n)_{n\geq 1}$, such that $\sum_{n=1}^{\infty} \lambda_n(2-\lambda_n)=\infty$ and $\tau \sum_{i=1}^m \sigma_i \|L_i\|^2 < 4$.

$n = 0$
$p_{1,0} = w_{1,0} = z_{1,0} = 0$
$p_{2,i,0} = w_{2,i,0} = z_{2,i,0} = 0$
while Not converged **do**
  $n \leftarrow n + 1$
  $p_{1,n} \leftarrow \text{Prox}_f^{\tau}\left(x_n - \frac{\tau}{2} \sum_{i=1}^m L_i^* v_{i,n}\right)$
  $w_{1,n} \leftarrow 2p_{1,n} - x_n$
  for i = 1, ..., m **do**
    $p_{2,i,n} \leftarrow \text{Prox}_{g_i^*}^{\sigma_i}\left(v_{i,n} + \frac{\sigma_i}{2} L_i w_{1,n} - \sigma_i r_i\right)$
    $w_{2,i,n} \leftarrow 2p_{2,i,n} - v_{i,n}$
  end for
  $z_{1,n} \leftarrow w_{1,n} - \frac{\tau}{2} \sum_{i=1}^m L_i^* w_{2,i,n}$
  $x_{n+1} \leftarrow x_n + \lambda_n(z_{1,n} - p_{1,n})$
  for i = 1, ..., m **do**
    $z_{2,i,n} \leftarrow w_{2,i,n} + \frac{\sigma_i}{2} L_i(2z_{1,n} - w_{1,n})$
    $v_{i,n+1} \leftarrow v_{i,n} + \lambda_n(2z_{2,i,n} - p_{2,i,n})$
  end for
end while

Output: $x^* = \text{Prox}_f^{\tau}\left(x_n - \frac{\tau}{2} \sum_{i=1}^m L_i^* v_{i,n}\right)$

we can compute the proximal operator of $T_\epsilon(\mu_0, \mu)$ in an efficient way, this problem can be solved by, e.g., making a variable splitting according to

$$\begin{aligned} &f(\cdot) = \gamma T_\epsilon(\mu_0, \cdot), \\ &g_1(\cdot) = \|\cdot\|_{2,1}, \qquad L_1 = \nabla, \quad r_1 = 0, \\ &g_2(\cdot) = \mathcal{I}_{\mathcal{B}_m(\kappa)}(\cdot), \quad L_2 = A, \quad r_2 = b, \end{aligned}$$

where $\mathcal{B}_m(\kappa) = \{\hat{b} \in \mathbb{R}^m : \|\hat{b}\|_2 \leq \kappa\}$ is the κ-ball in $\mathbb{R}^m$. We apply Algorithm D.2 for solving this problem and use Algorithm D.1 for computing the proximal operator of $T_\epsilon(\mu_0, \cdot)$. Explicit expressions for the proximal operators of the Fenchel duals of g_1 and g_2 can be computed; see, e.g., [48] for details.

Remark D.4.1. An alternative first-order method to solve optimization problems with one or more transportation costs is considered in [30]. The authors use a dual forward-backward (proximal-gradient) scheme where a key component is the evaluation of the gradient of the dual conjugate functional of $T_\epsilon(\mu_0, \cdot)$. However, our problem (D.4.2) contains two terms in addition to the optimal mass transport cost and does not directly fit into this framework. Since our method builds on the splitting framework in [13], it allows for an arbitrary number of cost terms (see (D.4.1)).

Formulating standard inverse problems using variable splitting

Given a linear forward operator A, many common regularization methods for solving inverse problems can be formulated as optimization problems of the form (D.4.1). Hence they can also be solved using variable splitting. We will use this to compare the proposed reconstruction method with two other approaches, first with an approach using TV-regularization [17], and second with an approach where we use a priori information with respect to the standard ℓ_2-norm. We formulate and solve both these problems using Algorithm D.2.

First, we consider the TV-regularization problem

$$\begin{aligned} \min_{\mu_{\text{est}} \geq 0} \quad & \|\nabla \mu_{\text{est}}\|_{2,1} \\ \text{subject to} \quad & \|A\mu_{\text{est}} - b\|_2 \leq \kappa \end{aligned} \tag{D.4.3}$$

and formulate it in the setting of (D.4.1) by defining

$$\begin{aligned} &f(\cdot) = \mathcal{I}_{\mathbb{R}^n_+}(\cdot), \\ &g_1(\cdot) = \|\cdot\|_{2,1}, \qquad L_1 = \nabla, \quad r_1 = 0, \\ &g_2(\cdot) = \mathcal{I}_{\mathcal{B}_{m(\kappa)}}(\cdot), \quad L_2 = A, \quad r_2 = b. \end{aligned}$$

The positivity constraint is handled by f, the TV-regularization by g_1 and the data matching by g_2. All functions needed in Algorithm D.2 can be explicitly computed [48].

In (D.4.3) there is no explicit notion of prior information and reconstruction is entirely based on the data and an implicit assumption on the sparsity of the gradient. One way to explicitly incorporate prior information in the problem is to add an ℓ_2-norm term that penalizes deviations from the given prior. This leads to the optimization problem

$$\begin{aligned} \min_{\mu_{\text{est}} \geq 0} \quad & \gamma\|\mu_{\text{est}} - \mu_0\|_2^2 + \|\nabla \mu_{\text{est}}\|_{2,1} \\ \text{subject to} \quad & \|A\mu_{\text{est}} - b\|_2 \leq \kappa \end{aligned} \tag{D.4.4}$$

which we formulate in the setting of (D.4.1) by defining

$$\begin{aligned} &f(\cdot) = \mathcal{I}_{\mathbb{R}^n_+}(\cdot), \\ &g_1(\cdot) = \gamma\|\cdot\|_2^2, \qquad L_1 = I, \quad r_1 = \mu_0 \\ &g_2(\cdot) = \|\cdot\|_{2,1}, \qquad L_2 = \nabla, \quad r_2 = 0 \\ &g_3(\cdot) = \mathcal{I}_{\mathcal{B}_{m(\kappa)}}(\cdot), \quad L_3 = A, \quad r_3 = b. \end{aligned}$$

The positivity constraint is handled by f, the ℓ_2 prior by g_1, the TV-regularization by g_2 and the data matching by g_3. The parameter γ determines the trade off between the ℓ_2 prior and the TV-regularization. Also here, all functions needed in Algorithm D.2 can be computed explicitly [48].

D.5 Application in computerized tomography

Computerized tomography (CT) is an imaging modality that is frequently used in many areas, especially in medical imaging (see, e.g., the monographs [9, 44, 52, 53]). In CT the object is probed with X-rays, and since different materials attenuate X-rays to different degrees, the intensities of the incoming and outgoing X-rays contain information of the material content and distribution. In the simplest case, where the attenuation is assumed to be energy independent and where scatter and nonlinear effects are ignored, one gets the equation [53, Chp. 3]

$$\int_L \mu_{\text{true}}(x)dx = \log\left(\frac{I_0}{I}\right).$$

Here $\mu_{\text{true}}(x)$ is the attenuation in the point x, L is the line along which the X-ray beam travels through the object, and I_0 and I are the the incoming and outgoing intensities. By taking several measurements along different lines L, one seeks to reconstruct the attenuation map μ_{true}.

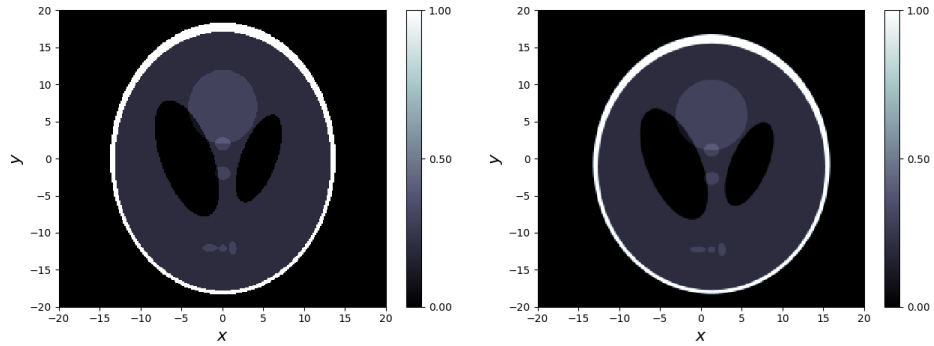
A set of measurements thus corresponds to the line integral of μ_{true} along a limited set of lines, and the corresponding operator that maps μ_{true} to the line integrals is called a *ray transform* or a *partial Radon transform*. Let A be the partial Radon transform operator, i.e., the operator such that $A(\mu)$ gives the line integral of μ along certain lines. This is a linear operator, and we consider the inverse problem of recovering μ_{true} from measurements

$$b = A(\mu_{\text{true}}) + \text{noise}.$$

However, this is an ill-posed inverse problem [33, p. 40]. In particular, the problem is severely ill-posed if the set of measurements is small or limited to certain angles, and hence regularization is needed to obtain an estimate μ_{est} of μ_{true}. One way to obtain such a μ_{est} is to formulate variational problems akin to the ones in Section D.4. In this section we consider CT problems, such as image reconstruction from limited-angle measurements, and use optimal mass transport to incorporate prior information to compensate for missing measurements. We also compare this method with standard reconstruction techniques. Tomography problems with transport priors have previously been considered in [1, 7], but in a less general setting. We compare and discuss the details in Remark D.5.2 in the end of this section.

Numerical simulations

To this end, consider the Shepp-Logan phantom in Figure D.1a [61] and the hand image in Figure D.4a [3]. Assume that the deformed images in Figure D.1b and Figure D.4b has been reconstructed previously from a detailed CT scan of the patient (where the deformation is due to, e.g., motion or breathing). By using the deformed images D.1b and D.4b as prior information we want to reconstruct the images D.1a



(a) (b)

Figure D.1: Figure showing (a) the Shepp-Logan phantom, (b) the deformed Shepp-Logan prior used in the first example. Gray scale values are shown to the right of each image.

and D.4a from relatively few measurements of the latter. For the reconstruction of the Shepp-Logan image we consider a scenario where the set of angles belongs to a limited interval (see the next paragraph). For the reconstruction of the hand image phantom we consider uniform spacing across all angles. In both examples we compare the reconstructions obtained by solving the variational problems (D.4.2), (D.4.3), and (D.4.4), as well as a standard filtered backprojection reconstruction [52].

In these examples, the images have a resolution of 256×256 pixels and we compute the data from the phantoms in Figures D.1a and D.4a. For the Shepp-Logan example we let data be collected from 30 equidistant angles in the interval $[\pi/4, 3\pi/4]$, and for the hand example from 15 equidistant angles in the interval $[0, \pi]$. In both cases the data is the line integrals from 350 parallel lines for each angle. On each data set white Gaussian noise is added (5% and 3%, respectively). The corresponding optimization problems (D.4.2), (D.4.3), and (D.4.4) are solved[4] with the Douglas-Rachford-type algorithm (Algorithm D.2) from [13], where the functions are split according to the description in Section D.4. The exact parameter values are given in Appendix D.10. Since the grid is regular and the cost in the transportation cost is spatially invariant, we can use the FFT for the computations of $T_\epsilon(\mu_0, \cdot)$ and the corresponding proximal operator, and thus we do not need to explicitly store the cost matrix or the final transference plan (see Remarks D.3.11 and D.5.1). These examples have been implemented and solved using ODL[5] [2], which is a python library for fast prototyping focusing on inverse problems. The ray transform computations are performed by the GPU-accelerated version of

[4] A fixed number of 10 000 Douglas-Rachford iterations are computed in each reconstruction.
[5] Open source code, available from `https://github.com/odlgroup/odl`

ASTRA[6] [54, 65]. The code used for these examples is available online from `http://www.math.kth.se/~aringh/Research/research.html`.

Remark D.5.1. These inverse problems are highly underdetermined and ill-posed. The total number of pixels is $256^2 = 65\,536$, but the number of data points in the examples are only $350{\cdot}30 = 10\,500$ and $350{\cdot}15 = 5\,250$, respectively. Also note that solving the corresponding optimal transport problems explicitly would amount to matrices of sizes $256^2 \times 256^2$, which means solving linear programs with over $4 \cdot 10^9$ variables.

Reconstruction of Shepp-Logan image

The reconstructions are shown in Figure D.2. Both the filtered backprojection reconstruction in Figure D.2a and the TV-reconstruction in Figure D.2b suffer from artifacts and severe vertical blurring due to poor vertical resolution resulting from the limited angle measurements. Figure D.2c shows the reconstruction with ℓ_2-prior. Some details are visible; however, these are at the same locations as in the prior and do not adjust according to the measurements of the phantom. Considerable artifacts also appear in this reconstruction, typically as fade-in-fade-out effects where the prior and the data do not match. The fade-in-fade-out effect that often occurs when using strong metrics for regularization is illustrated in Figure D.3. Selecting a low value γ (Figure D.3a) results in a reconstruction close to the TV-reconstruction, and selecting a large value γ gives a reconstruction close to the prior (Figure D.3c). By selecting a medium value γ one gets a reconstruction that preserves many of the details found in the prior; however, they remain at the same position as in the prior and hence are not adjusted to account for the measurement information.

The reconstruction with optimal mass transport prior is shown in Figure D.2d. Some blurring occurs, especially in the top and the bottom of the image; however, the overall shape is better preserved compared to the other reconstructions. Fine details are not visible, but the major features are better estimated compared to the TV- and ℓ_2-reconstructions. This example illustrates how one can improve the reconstruction by incorporating prior information, but without the fade-in-fade-out effects that typically occur when using a strong metric such as ℓ_2 for regularization.

Reconstruction of hand image

The reconstructions based on the phantom and the prior in Figure D.4 are shown in Figure D.5, and we obtain similar results similar to those in the previous example. The filtered backprojection reconstruction in Figure D.5a is quite fragmented, and although some details are visible there is also a considerable amount of noise and artifacts throughout the image. The TV-reconstruction, shown in Figure D.5b, is smeared, and few details are visible. Moreover, in the reconstruction with ℓ_2 prior, shown in Figure D.5c, artifacts similar to those in the Shepp-Logan example

[6] Open source code, available from `https://github.com/astra-toolbox/astra-toolbox`

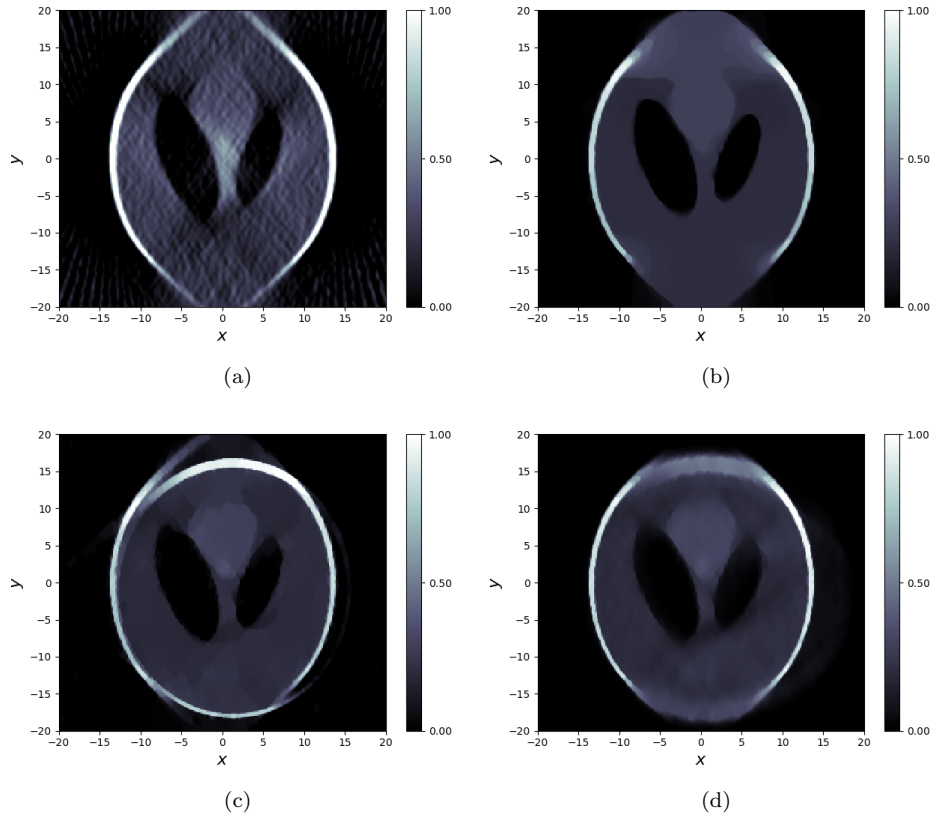
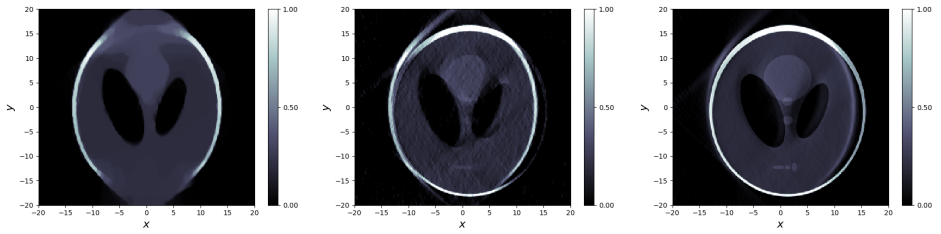
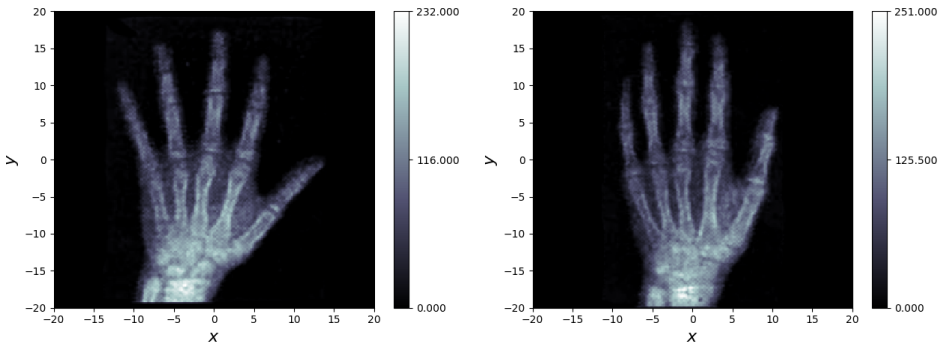


Figure D.2: Reconstructions using different methods. (a) Filtered backprojection, (b) reconstruction using TV-regularization, (c) reconstruction with ℓ_2^2-prior and TV-regularization ($\gamma = 10$), and (d) reconstruction with optimal transport prior and TV-regularization ($\gamma = 4$).



(a) (b) (c)

Figure D.3: Reconstructions using ℓ_2 prior with different regularization parameters: (a) $\gamma = 1$, (b) $\gamma = 100$, and (c) $\gamma = 10\,000$.



(a) (b)

Figure D.4: Figure showing (a) the hand image used as phantom, and (b) the hand image used as prior in the second example. Gray scale values are shown to the right of each image.

are present. Again, they stem from a mismatch between the prior and the data. Note especially that the thumb, but also the index and middle fingers, almost look dislocated or broken in the reconstruction. The reconstruction with an optimal mass transport prior is shown in Figure D.5d. Also in this case the fine details are not visible. Note, however, that details are more visible in D.5d, compared to the TV-reconstruction D.5b, and that the reconstruction in D.5d does not suffer from the the same kind of artifacts as the ℓ_2-regularized reconstruction in D.5c.

To illustrate the effect of the optimal mass transport prior on the final reconstruction, we also include Figure D.6. Figure D.6a shows the prior D.4b with certain areas marked, and Figure D.6b shows how the mass from these areas in D.6a are transported to D.6c. For reference the optimal mass transport reconstruction from D.5d is shown in Figure D.6c. By comparing the images one can see that the thumb

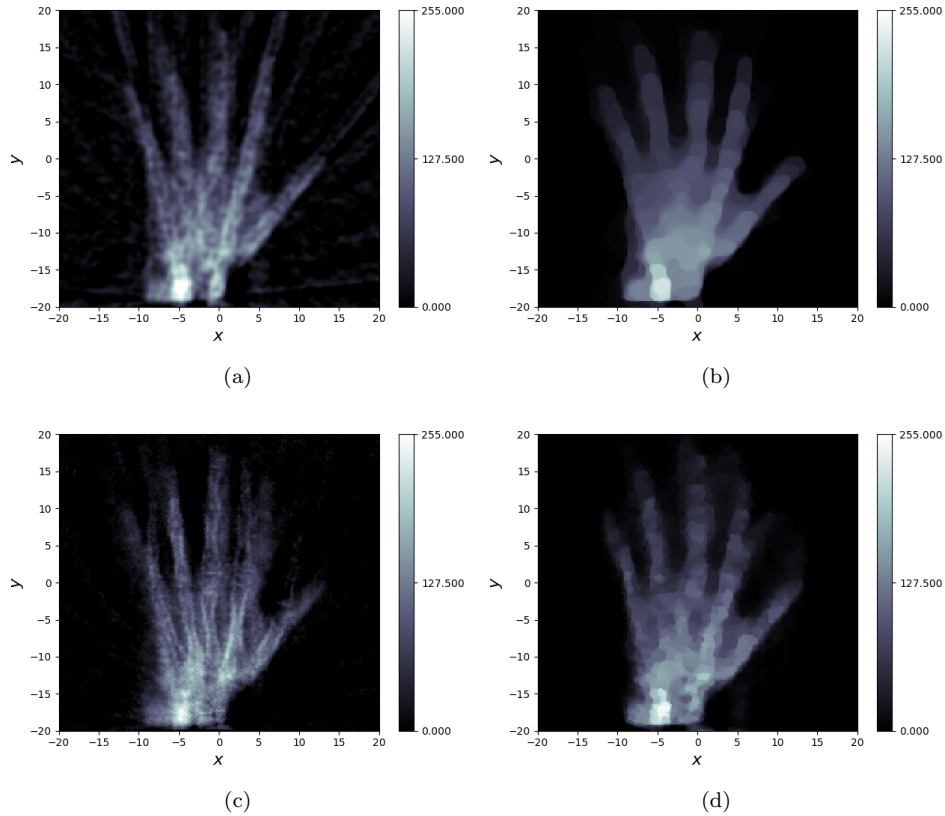
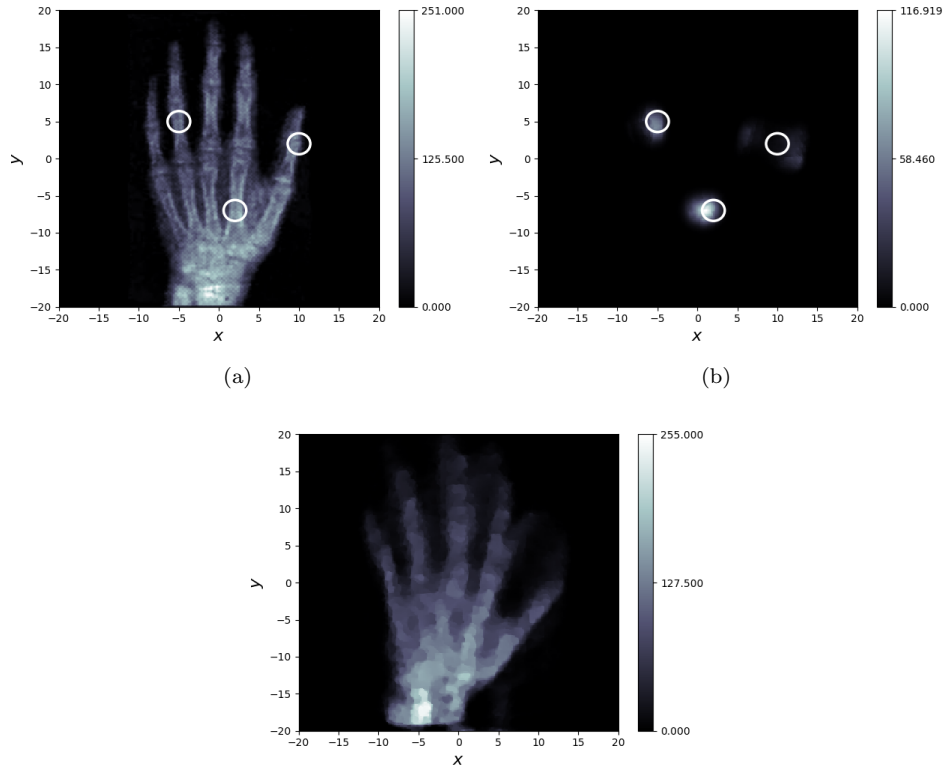


Figure D.5: Reconstructions using different methods. (a) Filtered backprojection, (b) reconstruction using TV-regularization, (c) reconstruction with ℓ_2^2-prior and TV-regularization, and (d) reconstruction with optimal transport prior and TV-regularization.



(a) (b)

(c)

Figure D.6: Visualizing how the mass from certain regions are transported: (a) the prior and three regions that are considered, (b) how the mass from certain regions in the prior is transported to the reconstruction, and (c) the reconstruction using an optimal mass transport prior.

in the prior is to a large extent transported to the location of the thumb in the phantom, which is what the optimal mass transport is intended to do. However, a fraction of the mass from the thumb is also transported to the index finger, giving rise to some artifacts. One can also note that there is a certain "halo-effect" around each region, especially for the region on the thumb. This is most likely due to the computational regularization $\epsilon D(M)$ in (D.2.3), which forces the optimal M to be strictly positive. It would therefore be desirable to precondition or in other ways improve the numerical conditioning of the steps in Algorithm D.1, to allow for the computation of the proximal operator for lower values of ϵ.

Remark D.5.2. As mentioned in the beginning of this section, the papers [1, 7] also consider tomography problems with optimal mass transport priors. In both these papers, the optimization problems are of the form

$$\min_{\mu_{\text{est}}} T(\mu_0, \mu_{\text{est}}) + \sum_{\ell=1}^{L} \gamma_\ell \tilde{T}(P_{\theta_\ell}\mu_{\text{est}}, b_\ell),$$

where T and $\tilde{T}$ are transportation costs and P_{θ_ℓ} denotes the projection along the angle θ_ℓ (i.e., a partial Radon transform). In our formulation we use a hard data matching constraint based on the L_2-norm (see (D.4.2)) which models Gaussian noise. A computational approach is provided in [7] that also builds on Sinkhorn iterations (Bregman projections). However, the computational procedure builds on the fact that the adjoint $P_{\theta_\ell}^T$ commutes with elementwise functions, which is a result of the implementation of the partial ray transform P_{θ_ℓ} using nearest neighbor interpolation.[7] On the other hand, our method allows for an arbitrary (linear) forward operator.

D.6 Concluding remarks and further directions

In this work we have considered computational methods for solving inverse problems containing entropy regularized optimal mass transport terms. First, using a dual framework we have generalized the Sinkhorn iteration, thereby making it applicable to a set of optimization problems. In particular, the corresponding proximal operator of the entropy regularized optimal mass transport cost is computed efficiently. Next, we use this to address a large class of inverse problems using variable splitting. In particular we use a Douglas-Rachford-type method to solve two problems in CT where prior information is incorporated using optimal mass transport.

Interestingly, both the Sinkhorn iterations and the proposed approach for computing the proximal operator are identical to coordinate ascent of the dual problem. For these problems the coordinate ascent step can be computed explicitly by (D.2.7) or (D.3.9), respectively. In this setting the hard constraint (D.3.2b) is replaced by

[7] Implementing the partial ray transform P_{θ_ℓ} using nearest neighbor interpolation results in a matrix representation where each column is an elementary unit vector.

the soft constraint (D.3.3). The Hessian of the barrier term is given explicitly by

$$-\frac{1}{\epsilon}\begin{pmatrix} \operatorname{diag}(u_0 \odot (Ku_1)) & \operatorname{diag}(u_0)K\operatorname{diag}(u_1) \\ \operatorname{diag}(u_1)K^T\operatorname{diag}(u_0) & \operatorname{diag}(u_1 \odot (K^T u_0)) \end{pmatrix}$$

and multiplication of this can be computed quickly (cf. Remark D.3.11). This opens up for accelerating convergence using, e.g., quasi-Newton or parallel tangent methods [51].

The methodology presented in this paper naturally extends to problems with several optimal mass transport terms. In particular this would be useful for cases where estimates are not guaranteed to have the same mass or where some parts are not positive. An example of such a problem is the computation of a centroid (barycenter) from noisy measurements (cf. [23]), i.e.,

$$\min_{\mu_\ell,\, \ell=0,\dots,L} \sum_{\ell=1}^{L} \left(T_\epsilon(\mu_0, \mu_\ell) + \frac{1}{2\sigma_\ell}\|\mu_\ell - \nu_\ell\|_2^2 \right). \tag{D.6.1}$$

This problem has applications in clustering and will be considered in future work.

D.7 Appendix 1: Proof of Proposition D.3.4

We seek the dual problem of the following optimization problem

$$\begin{aligned} \min_{M\geq 0, \mu_{\text{est}}} \quad & \operatorname{Tr}(C^T M) + \epsilon D(M) + g(\mu_{\text{est}}) \\ \text{subject to} \quad & \mu_0 = M\mathbf{1}_{n_1} \\ & \mu_{\text{est}} = M^T \mathbf{1}_{n_0}. \end{aligned} \tag{D.7.1}$$

Lagrange relaxation gives the Lagrangian

$$\begin{aligned} L(M, \mu_{\text{est}}, \lambda_0, \lambda_1) &= \operatorname{Tr}(C^T M) + \epsilon D(M) + g(\mu_{\text{est}}) \\ &\quad + \lambda_0^T(\mu_0 - M\mathbf{1}_{n_1}) + \lambda_1^T(\mu_{\text{est}} - M^T\mathbf{1}_{n_0}) \\ &= \operatorname{Tr}((C - \mathbf{1}_{n_1}\lambda_0^T - \lambda_1\mathbf{1}_{n_0}^T)^T M) + \epsilon D(M) \\ &\quad + \lambda_0^T\mu_0 - ((-\lambda_1)^T\mu_{\text{est}} - g(\mu_{\text{est}})). \end{aligned}$$

First note that by definition $g^*(-\lambda_1) = \max_{\mu_{\text{est}}}(-\lambda_1)^T\mu_{\text{est}} - g(\mu_{\text{est}})$, and hence the last term equals $-g^*(-\lambda_1)$ for the minimizing μ_{est}. Next, the minimizing M is unique and satisfies

$$\epsilon \log(m_{ij}) + c_{ij} - \lambda_0(i) - \lambda_1(j) = 0$$

or, equivalently,

$$M = \operatorname{diag}(\exp(\lambda_0/\epsilon))\exp(-C/\epsilon)\operatorname{diag}(\exp(\lambda_1/\epsilon)).$$

By plugging these into the Lagrangian we obtain

$$\min_{M,\mu_{\text{est}}} L(M, \mu_{\text{est}}, \lambda_0, \lambda_1) = \epsilon \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} (1 - m_{ij}) + \lambda_0^T \mu_0 - g^*(-\lambda_1)$$
$$= \lambda_0^T \mu_0 - g^*(-\lambda_1) - \epsilon \exp(\lambda_0^T/\epsilon) \exp(-C/\epsilon) \exp(\lambda_1/\epsilon) + \epsilon n_0 n_1,$$

which is the objective function of the dual problem.

Moreover, by Assumption D.3.3 and the fact that $T_\epsilon(\mu_0, \mu_{\text{est}}) < \infty$ restricts μ_{est} to a compact set, there exists an optimal solution to (D.7.1). To see this, note that if μ_{est} is a point such that $g(\mu_{\text{est}}) < \infty$ and $\sum_{i=1}^{n_0} \mu_0(i) = \sum_{j=1}^{n_1} \mu_{\text{est}}(j)$, then $M = \mu_0 {\mu_{\text{est}}}^T$ is an inner point where the objective function takes a finite value. From this it also follows that the minimum of (D.7.1) is finite. Therefore by [10, Prop. 5.3.2] strong duality holds. □

D.8 Appendix 2: Proof of Theorem D.3.10

First we would like to show that Algorithm D.1 corresponds to coordinate ascent of the dual problem, i.e., that steps 3 and 4 correspond to the maximization of λ_0 and λ_1, respectively. This follows directly for step 3 since it is identical to (D.3.7a) in Lemma D.3.5. Note that step 3 is the Sinkhorn iterate (D.2.7a) (cf. Corollary D.3.2).

Next, consider the condition (D.3.7b) in Lemma D.3.5, from which it follows that the minimizing λ_1 satisfies

$$\mu_1 - \sigma\lambda_1 = \exp(\lambda_1/\epsilon) \odot \left(K^T u_0\right),$$

where we let $K = \exp(-C/\epsilon)$ and $u_0 = \exp(\lambda_0/\epsilon)$. Taking the logarithm and adding $(\mu_1 - \sigma\lambda_1)/(\sigma\epsilon) - \log(\sigma\epsilon)$ to each side, we get

$$\frac{\mu_1 - \sigma\lambda_1}{\sigma\epsilon} + \log\left(\frac{\mu_1 - \sigma\lambda_1}{\sigma\epsilon}\right) = \frac{\mu_1}{\sigma\epsilon} + \log\left(\frac{K^T u_0}{\sigma\epsilon}\right),$$

or, equivalently,

$$\frac{\mu_1 - \sigma\lambda_1}{\sigma\epsilon} = \omega\left(\frac{\mu_1}{\sigma\epsilon} + \log\left(\frac{K^T u_0}{\sigma\epsilon}\right)\right), \tag{D.8.1}$$

where ω denotes the elementwise Wright omega function, i.e., the function mapping $x \in \mathbb{R}$ to $\omega(x) \in \mathbb{R}_+$ for which $x = \log(\omega(x)) + \omega(x)$ [26]. The function is well defined as a function $\mathbb{R} \to \mathbb{R}_+$ which is the domain and range of interest in our case. This expression of λ_1 is equivalent to step 4 (and (D.3.9b)), and we have thus shown that Algorithm D.1 is a coordinate ascent algorithm for the dual problem (D.3.8).

Next, note that (D.3.8) is a strictly convex optimization problem. This follows since the Hessian

$$-\frac{1}{\epsilon}\begin{pmatrix} \operatorname{diag}(u_0 \odot (K u_1)) & \operatorname{diag}(u_0) K \operatorname{diag}(u_1) \\ \operatorname{diag}(u_1) K^T \operatorname{diag}(u_0) & \operatorname{diag}(u_1 \odot (K^T u_0)) \end{pmatrix} - \begin{pmatrix} 0 & 0 \\ 0 & \sigma I \end{pmatrix} \tag{D.8.2}$$

is strictly negative definite. This can be seen by noting that the first term is diagonally dominant and hence negative semidefinite. Since $\sigma > 0$, then any zero eigenvector of (D.8.2) can only have nonzero elements in the first block. However, the $(1,1)$-block of (D.8.2), i.e., $-\text{diag}(u_0 \odot (Ku_1))/\epsilon$, is negative definite, and hence no zero eigenvalue exists and (D.8.2) is strictly negative definite.

As seen above, the optimization problem is strictly convex, and hence there is a unique stationary point which is also the unique maximum. To show convergence, note that the objective function in (D.3.8) is continuously differentiable, and hence any limit point of the coordinate ascent is stationary [10, Prop. 2.7.1]. Further, since the superlevel sets of (D.3.8) are bounded and there is a unique stationary point, Algorithm D.1 converges to unique maximum. Finally, locally linear convergence follows from [11, Thm. 2.2] since the optimization problem is strictly convex. □

Remark D.8.1. For future reference, we note that (D.8.1) can be rewritten as

$$\begin{aligned}\frac{\lambda_1}{\epsilon} &= \frac{\mu_1}{\sigma\epsilon} - \omega\left(\frac{\mu_1}{\sigma\epsilon} + \log\left(\frac{K^T u_0}{\sigma\epsilon}\right)\right) \\ &= \frac{\mu_1}{\sigma\epsilon} - \left(\frac{\mu_1}{\sigma\epsilon} + \log\left(\frac{K^T u_0}{\sigma\epsilon}\right) - \log\left[\omega\left(\frac{\mu_1}{\sigma\epsilon} + \log\left(\frac{K^T u_0}{\sigma\epsilon}\right)\right)\right]\right) \qquad \text{(D.8.3)} \\ &= \log\left(\frac{\sigma\epsilon\omega\left(\frac{\mu_1}{\sigma\epsilon} + \log\left(\frac{K^T u_0}{\sigma\epsilon}\right)\right)}{K^T u_0}\right),\end{aligned}$$

where we in the second equality use the definition $\omega(x) = x - \log(\omega(x))$. This expression can alternatively be used in Algorithm D.1. However, our experience is that (D.8.1) is better conditioned than (D.8.3).

D.9 Appendix 3: Connection with method based on Dykstra's algorithm

We would like to thank the reviewers for pointing out [55]. In fact, Algorithm D.1 can be seen as a (nontrivial) special case of the iterations in [55, Prop. 3.3]. We will here provide a separate derivation leading to a simplified but equivalent version of this algorithm. The algorithm [55, Prop. 3.3] also addresses the optimization problem (D.3.5) and can be used when the entropic proximal, defined by

$$\text{Prox}_{\sigma g}^{\overline{\text{KL}}}(z) := \arg\min_x \ \sigma g(x) + D(x|z),$$

where $D(x|z) = \sum_{i=1}^n (x_i \log(x_i/z_i) - x_i + z_i)$, is fast to compute. It can be noted that computing the entropic proximal is equivalent to

$$\begin{aligned}\text{Prox}_{\sigma g}^{\overline{\text{KL}}}(z) := \ \arg\min_{x,x'} \quad & g(x) + \frac{1}{\sigma} D(x'|z) \\ \text{subject to} \quad & x = x',\end{aligned}$$

which has a Lagrange dual that is given by (cf. the proof of Proposition D.3.4)

$$\max_{\lambda} \; -g^*(-\lambda) - \exp(\sigma\lambda^T)z/\sigma + \mathbf{1}^T z/\sigma$$

and where the optima of the primal and dual problems relate as $\sigma\lambda = \log(x./z)$. The maximizing argument λ is specified by

$$0 \in \partial g^*(-\lambda) - \exp(\sigma\lambda) \odot z/\sigma.$$

Noting that this condition is equivalent to (D.3.7b) with $z/\sigma = \exp(-C^T/\epsilon)\exp(\lambda_0/\epsilon)$ and $\epsilon = 1/\sigma$, the dual update of λ_1 can be written as

$$\exp(\lambda_1/\epsilon) = \frac{\overline{\mathrm{Prox}^{\mathrm{KL}}_{\epsilon^{-1}g}}(\exp(-C^T/\epsilon)\exp(\lambda_0/\epsilon))}{\exp(-C^T/\epsilon)\exp(\lambda_0/\epsilon)} = \frac{\overline{\mathrm{Prox}^{\mathrm{KL}}_{\epsilon^{-1}g}}(K^T u_0)}{K^T u_0}, \qquad \text{(D.9.1)}$$

where as before $K = \exp(-C/\epsilon)$ and $u_i = \exp(\lambda_i/\epsilon)$, for $i = 0, 1$. Using this we can state the simplified, but equivalent, version of [55, Prop. 3.3], shown in Algorithm D.3. The algorithm is equivalent in the sense that if $K = \xi^T$, then it holds that $b^{(2\ell)} = b^{(2\ell-1)} = u_0^{(\ell)}$ and $a^{(2\ell+1)} = a^{(2\ell)} = u_1^{(\ell)}$ for $\ell \geq 1$. Similar algorithms have also been derived in the recent preprints [23, 60].

Algorithm D.3 Simplified version of [55, Prop. 3.3].

Input: ϵ, C, λ_0, μ_0, g

1: $K = \exp(-C/\epsilon)$ and $u_1 = \mathbf{1}$
2: $\ell = 0$
3: **while** Not converged **do**
4: $\quad \ell \leftarrow \ell + 1$
5: $\quad u_0^{(\ell)} \leftarrow \mu_0./(K u_1^{(\ell-1)})$
6: $\quad u_1^{(\ell)} \leftarrow \dfrac{\overline{\mathrm{Prox}^{\mathrm{KL}}_{\epsilon^{-1}g}}(K^T u_0^{(\ell)})}{K^T u_0^{(\ell)}}$
7: **end while**

Output: $\mu_{\text{est}} \leftarrow u_1^{(\ell)} \odot (K^T u_0^{(\ell)})$

The case which is a main focus for this paper is $g(\cdot) = \frac{1}{2\sigma}\|\cdot - \mu_1\|_2^2$, and comparing the expressions (D.9.1) and (D.8.3) indicates that

$$\overline{\mathrm{Prox}^{\mathrm{KL}}_{\epsilon^{-1}g}}(K^T u_0) = \sigma\epsilon\, \omega\left(\frac{\mu_1}{\sigma\epsilon} + \log\left(\frac{K^T u_0}{\sigma\epsilon}\right)\right).$$

This can in fact be verified by direct computations along the lines of the proof of Theorem D.3.10.

D.10 Appendix 4: Parameters in the numerical examples

The problem is set up with noise level 5% in the Shepp-Logan example and 3% in the hand example. The parameter κ is selected to be 120% of the norm of the noise. That is, in the Shepp logan example white noise is generated and normalized so that $\|b - A\mu_{\text{true}}\|_2 / \|A\mu_{\text{true}}\|_2 = 0.05$ and $\kappa = 1.2 \cdot 0.05 \cdot \|A\mu_{\text{true}}\|_2$. This ensures that the true image μ_{true} belongs to the feasible region $\{\mu : \|b - A\mu\|_2 \leq \kappa\}$. The cost function in the optimal mass transport distance is given by

$$c(x_1, x_2) = \min(\|x_1 - x_2\|_2, 20)^2, \tag{D.10.1}$$

where the truncation at 20 is done in order to improve the conditioning of the computations. Further, the proximal operator of the optimal mass transport functional is computed using 200 generalized Sinkhorn iterations. Each optimization problem is solved using 10 000 iterations in the Douglas-Rachford algorithm. The step size parameters σ_i are set to be $\sigma_i = (\tau \|L_i\|_{\text{op}}^2)^{-1}$ where $\|L\|_{\text{op}} = \sup_{\|x\|_2 \leq 1} \|Lx\|_2$ is the operator norm (approximated using the power iterations in ODL [2]). The remaining parameters are selected according to Table D.1. The gradient operator used in the TV-terms is the default `Gradient` operator in ODL, which pads the boundaries with zeros and applies a forward-difference in each dimension (see the documentation `http://odlgroup.github.io/odl/`). Other options are available, e.g., zero-order-hold padding or periodic boundary conditions, as well as backward- or central-difference. In our case, zero-padding is not an issue since both phantoms are zero along the boundary; see Figures D.1a and D.4a. For the filtered backprojection reconstruction we use the ODL-implementation with a Hann filter with filter parameter 0.7 for the Shepp-Logan example and 0.5 for the hand example.

Table D.1: Parameter values for the variational problems and reconstruction algorithms. A $\star$ means that the parameter is not used in this problem.

Reconstruction example		Parameters in optimization problem		Parameters in Algorithm D.2	
Phantom	Objective function	γ	ϵ	τ	λ
Shepp-Logan	TV	$\star$	$\star$	0.05	1
	TV + L_2^2	$\{1, 10, 100, 10\,000\}$	$\star$	0.05	1
	TV + OMT	4	1	5	1.8
Hand	TV	$\star$	$\star$	$1/\sqrt{2}$	1
	TV + L_2^2	10	$\star$	$1/\sqrt{2}$	1
	TV + OMT	4	1.5	$500\sqrt{2}$	1.8

Acknowledgements

The authors would like to thank Yongxin Chen and Tryphon Georgiou for input and fruitful discussions on the optimal mass transport problem, and Krister Svanberg for insightful discussion on optimization. The authors would also like to thank Jonas Adler and Ozan Öktem for input on the example and the implementation in ODL. The figures in Figure D.4a and Figure D.4b have generously been provided by Klas Modin and can be found in [3]. The authors would also like to thank the anonymous referees for their valuable feedback and for bringing [55, 30, 60] to their attention.

References

[1] I. Abraham, R. Abraham, M. Bergounioux, and G. Carlier. Tomographic reconstruction from a few views: A multi-marginal optimal transport approach. *Applied Mathematics & Optimization*, 75(1):55–73, 2017.

[2] J. Adler, H. Kohr, and O. Öktem. Odl 0.6.0, April 2017.

[3] M. Bauer, S. Joshi, and K. Modin. Diffeomorphic density matching by optimal information transport. *SIAM Journal on Imaging Sciences*, 8(3):1718–1751, 2015.

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

[5] H.H. Bauschke and A.S. Lewis. Dykstras algorithm with Bregman projections: A convergence proof. *Optimization*, 48(4):409–427, 2000.

[6] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[7] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[8] J.-D. Benamou, G. Carlier, and M. Laborde. An augmented Lagrangian approach to Wasserstein gradient flows and applications. *ESAIM: Proceedings and Surveys*, 54:1–17, 2016.

[9] M. Bertero and P. Boccacci. *Introduction to inverse problems in imaging*. IOP Publishing, 1998.

[10] D. Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.

[11] J.C. Bezdek, R.J. Hathaway, R.E. Howard, C.A. Wilson, and M.P. Windham. Local convergence analysis of a grouped variable version of coordinate descent. *Journal of Optimization Theory and Applications*, 54(3):471–477, 1987.

[12] R.I. Boţ and E.R. Csetnek. On the convergence rate of a forward-backward type primal-dual splitting algorithm for convex optimization problems. *Optimization*, 64(1):5–23, 2015.

[13] R.I. Boţ and C. Hendrich. A Douglas–Rachford type primal-dual method for solving inclusions with mixtures of composite and parallel-sum type monotone operators. *SIAM Journal on Optimization*, 23(4):2541–2565, 2013.

[14] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, Jan. 2011.

[15] J.P. Boyle and R.L. Dykstra. A method for finding projections onto the intersection of convex sets in Hilbert spaces. In *Advances in order restricted statistical inference*, pages 28–47. Springer, 1986.

[16] L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[17] E. Candes and J. Romberg. Signal recovery from random projections. In *Electronic Imaging 2005*, pages 76–86. International Society for Optics and Photonics, 2005.

[18] F. Carli, L. Ning, and T.T. Georgiou. Convex clustering via optimal mass transport. *arXiv preprint arXiv:1307.5459*, 2013.

[19] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[20] Y. Chen, T.T. Georgiou, and M. Pavon. Entropic and displacement interpolation: A computational approach using the Hilbert metric. *SIAM Journal on Applied Mathematics*, 76(6):2375–2396, 2016.

[21] Y. Chen, T.T. Georgiou, and M. Pavon. On the relation between optimal transport and Schrödinger bridges: A stochastic control viewpoint. *Journal of Optimization Theory and Applications*, 169(2):671–691, 2016.

[22] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Unbalanced optimal transport: Geometry and Kantorovich formulation. *arXiv preprint arXiv:1508.05216*, 2015.

[23] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Scaling algorithms for unbalanced transport problems. *arXiv preprint arXiv:1607.05816*, 2016.

[24] P.L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

[25] P.L. Combettes and J.-C. Pesquet. Primal-dual splitting algorithm for solving inclusions with mixtures of composite, lipschitzian, and parallel-sum type monotone operators. *Set-Valued and variational analysis*, 20(2):307–330, 2012.

[26] R.M. Corless and D.J. Jeffrey. The Wright ω function. In *Artificial intelligence, automated reasoning, and symbolic computation*, pages 76–89. Springer, 2002.

[27] I Csiszár. Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. *The Annals of Statistics*, 19(4):pp. 2032–2066, 1991.

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

[29] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning (ICML)*, pages 685–693, 2014.

[30] M. Cuturi and G. Peyré. A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, pages 320–343, 2016.

[31] J. Eckstein. *Splitting methods for monotone operators with applications to parallel optimization.* PhD thesis, 1989. Department of Civil Engineering, Massachusetts Institute of Technology.

[32] J. Eckstein and D.P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, 1992.

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

[34] B. Engquist and B.D. Froese. Application of the Wasserstein metric to seismic signals. *Communications in Mathematical Sciences*, 12(5), 2014.

[35] S.-C. Fang. An unconstrained convex programming view of linear programming. *Zeitschrift für Operations Research*, 36(2):149–161, 1992.

[36] J. Franklin and J. Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its applications*, 114:717–735, 1989.

[37] T.T. Georgiou, J. Karlsson, and M.S. Takyar. Metrics for power spectra: An axiomatic approach. *IEEE Transactions on Signal Processing*, 57(3):859–867, 2009.

[38] F. De Goes, D. Cohen-Steiner, P. Alliez, and M. Desbrun. An optimal transport approach to robust reconstruction and simplification of 2d shapes. *Computer Graphics Forum*, 30(5):1593–1602, 2011.

[39] S. Haker, L. Zhu, A. Tannenbaum, and S. Angenent. Optimal mass transport for registration and warping. *International Journal of computer vision*, 60(3):225–240, 2004.

[40] P. Jacob, F. Lindsten, and T. Schön. Coupling of particle filters. *arXiv preprint arXiv:1606.01156*, 2016.

[41] X. Jiang, Z.-Q. Luo, and T.T. Georgiou. Geometric methods for spectral analysis. *IEEE Transactions on Signal Processing*, 60(3):1064–1074, 2012.

[42] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker-Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[43] T. Kaijser. Computing the Kantorovich distance for images. *Journal of Mathematical Imaging and Vision*, 9(2):173–191, 1998.

[44] A.C Kak and M Slaney. *Principles of Computerized Tomographic Imaging*. Society for Industrial and Applied Mathematics, 2001.

[45] L. Kantorovich. On a problem of Monge. *Journal of Mathematical Sciences*, 133(4):1383–1383, 2006.

[46] J. Karlsson and T.T Georgiou. Uncertainty bounds for spectral estimation. *IEEE Transactions on Automatic Control*, 58(7):1659–1673, 2013.

[47] J. Karlsson and L. Ning. On robustness of ℓ_1-regularization methods for spectral estimation. In *IEEE Annual Conference on Decision and Control (CDC)*, pages 1767–1773. IEEE, 2014.

[48] N. Komodakis and J.-C. Pesquet. Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems. *IEEE Signal Processing Magazine*, 32(6):31–54, 2015.

[49] D. Lee. Fast multiplication of a recursive block toeplitz matrix by a vector and its application. *Journal of Complexity*, 2(4):295–305, 1986.

[50] H. Ling and K. Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE transactions on pattern analysis and machine intelligence*, 29(5):840–853, 2007.

[51] D.G. Luenberger. *Linear and nonlinear programming*. Addison-Wesley, 2 edition, 1984.

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

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

[54] W.J. Palenstijn, K.J. Batenburg, and J. Sijbers. Performance improvements for iterative electron tomography reconstruction using graphics processing units (GPUs). *Journal of structural biology*, 176(2):250–253, 2011.

[55] G. Peyré. Entropic approximation of Wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

[56] R.T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

[57] Y. Rubner, C. Tomasi, and L. Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

[58] L.I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

[59] A. Sadeghian, D. Lim, J. Karlsson, and J. Li. Automatic target recognition using discrimination based on optimal transport. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2604–2608. IEEE, 2015.

[60] B. Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *arXiv preprint arXiv:1610.06519*, 2016.

[61] L.A Shepp and B.F. Logan. The Fourier reconstruction of a head section. *IEEE Transactions on Nuclear Science*, 21(3):21–43, 1974.

[62] E.Y. Sidky, H.J. Jakob, and P. Xiaochuan. Convex optimization problem prototyping for image reconstruction in computed tomography with the Chambolle–Pock algorithm. *Physics in medicine and biology*, 57(10):3065, 2012.

[63] R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *American Mathematical Monthly*, pages 402–405, 1967.

[64] M. Teboulle. Entropic proximal mappings with applications to nonlinear programming. *Mathematics of Operations Research*, 17(3):670–690, 1992.

[65] W. van Aarle, W.J. Palenstijn, J. De Beenhouwer, T. Altantzis, S. Bals, K.J. Batenburg, and J. Sijbers. The ASTRA Toolbox: A platform for advanced algorithm development in electron tomography. *Ultramicroscopy*, 157:35–47, 2015.

[66] A.M. Vershik. Long history of the Monge-Kantorovich transportation problem. *The Mathematical Intelligencer*, 35(4):1–9, 2013.

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

Paper E

Learning to solve inverse problems using Wasserstein loss

Learning to solve inverse problems using Wasserstein loss

by

Jonas Adler, Axel Ringh, Ozan Öktem, and Johan Karlsson

Abstract

We propose using the Wasserstein loss for training in inverse problems. In particular, we consider a learned primal-dual reconstruction scheme for ill-posed inverse problems using the Wasserstein distance as loss function in the learning. This is motivated by miss-alignments in training data, which when using standard mean squared error loss could severely degrade reconstruction quality. We prove that training with the Wasserstein loss gives a reconstruction operator that correctly compensates for miss-alignments in certain cases, whereas training with the mean squared error gives a smeared reconstruction. Moreover, we demonstrate these effects by training a reconstruction algorithm using both mean squared error and optimal transport loss for a problem in computerized tomography.

Keywords: machine learning, optimal mass transport, Wasserstein distance, inverse problems, computed tomography

E.1 Introduction

In inverse problems the goal is to determine model parameters, henceforth called signal, from indirect noisy observations. Example of such problems arise in many different fields in science and engineering, e.g., in X-ray Computed Tomography (CT) [36], electron tomography [37], and magnetic resonance imaging [11]. Machine learning has recently also been applied in this area, especially in imaging applications. Using supervised machine-learning to solve inverse problems in imaging requires training data where ground truth images are paired with corresponding noisy indirect observations. The learning provides a mapping that associates observations to corresponding images. However, in several applications there are difficulties in obtaining the ground truth, e.g., in many cases it may have undergone a distortion. For example, a recent study showed that MRI images may be distorted by up to 4 mm due to, e.g., inhomogeneities in the main magnetic field [47]. If these images are used for training, the learned MRI reconstruction will suffer in quality. Similar geometric inaccuracies arise in several other imaging modalities, such as Cone Beam CT and full waveform inversion in seismic imaging.

This work provides a scheme for learning a reconstruction method for an ill-posed inverse problem with a Wasserstein loss by leveraging upon recent advances

in efficient solutions of optimal transport [14, 28, 44] and learned iterative schemes for inverse problems [5]. The proposed method is demonstrated on a computed tomography example, where we show a significant improvement compared to training the same network using mean squared error loss. In particular, using the Wasserstein loss instead of standard mean squared error gives a result that is more robust against potential miss-alignment in training data.

E.2 Background

Inverse problems

Formalizing the notion of an inverse problem, our goal is to reconstruct an estimate of the signal $f_{\text{true}} \in X$ from noisy indirect measurements (data) $g \in Y$ assuming

$$g = \mathcal{A}(f_{\text{true}}) + \delta g. \tag{E.2.1}$$

In the above, the sets X and Y are called the reconstruction and data space, respectively. Both are typically Hilbert or Banach spaces. Moreover $\mathcal{A}\colon X \to Y$ denotes the forward operator, which models how a given signal gives rise to data in absence of noise. Finally, $\delta g \in Y$ is the noise component of data. Many inverse problems of interest are ill-posed, meaning that there is no uniques solution to (E.2.1) and hence there is no inverse to $\mathcal{A}$. Typically reconstructions of f_{true} are sensitive to the data and small errors gets amplified. One way to mitigate these effects is to use regularization [16].

Variational regularization In variational regularization one formulates the reconstruction problem as an optimization problem. To this end, one introduces a data discrepancy functional $f \mapsto \mathcal{L}(\mathcal{A}(f), g)$, where $\mathcal{L} : Y \times Y \to \mathbb{R}$, that quantifies the miss-fit in data space, and a regularization functional $S\colon X \to \mathbb{R}$ that encodes a priori information about f_{true} by penalizing undesirable solutions. For a given $g \in Y$, this gives an optimization problem of the form

$$\min_{f \in X} \ \mathcal{L}(\mathcal{A}(f), g) + \lambda S(f). \tag{E.2.2}$$

Here, λ acts as a trade-off parameter between the data discrepancy and regularization functional. In many cases $\mathcal{L}$ is taken to be the negative data log-likelihood, e.g., $\mathcal{L}(\mathcal{A}(f), g) = \| \mathcal{A}(f) - g\|_2^2$ in the case of additive white Gaussian noise. Moreover, a typical choice for regularization functional is total variation (TV) regularization, $S(f) = \|\nabla f\|_1$ [43]. Such regularizers typically give rise to large scale and non-differentiable optimization problems, which requires advanced optimization algorithms. These methods are typically based on the proximal point algorithm and operator splitting [7] and commonly used examples are FISTA [8], the Primal-Dual Hybrid-Gradient algorithm [12], and ADMM [10].

Learning for inverse problems In many applications, and so also for some inverse problems, data driven approaches have shown dramatic improvements over the state-of-the-art [30]. Using supervised learning to solve an inverse problem amounts to finding a parametrized operator $\mathcal{A}_{\boldsymbol{\Theta}}^{\dagger}: Y \to X$ where the parameters $\boldsymbol{\Theta}$ are selected so that

$$g = \mathcal{A}(f_{\text{true}}) + \delta g \implies \mathcal{A}_{\boldsymbol{\Theta}}^{\dagger}(g) \approx f_{\text{true}}.$$

For inverse problems in image processing, such as denoising and deblurring, we have $Y = X$ and it is possible to apply a wide range of widely studied machine learning techniques, such as neural networks with various architectures, including fully convolutional networks [24] and denoising auto-encoders [49].

However, in more complicated inverse problems as in tomography, the data and reconstruction spaces are very different, e.g., their dimension after discretization may differ. For this reason, learning a mapping from Y to X becomes nontrivial, and classical architectures that map, e.g., images to images using convolutional networks cannot be applied as-is. One solution is to use fully-connected layers as in [40] for very small scale tomographic reconstruction problems. A major disadvantage with such a fully learned approach is that the parameters space has to be very high dimensional in order to be able to learn both the prior and the data model, which often renders it infeasible due to training time and lack of training data.

A more successful approach is to first apply some crude reconstruction operator $\mathcal{A}^{\dagger}: Y \to X$ and then use machine learning to post process the result. This separates the learning from the complications of mapping between spaces since the operator $\mathcal{A}^{\dagger}$ can be applied off-line, prior to training. Such an approach has been demonstrated for tomographic reconstruction in [41, 48]. Its drawback for ill-posed inverse problems is that information is typically lost by using $\mathcal{A}^{\dagger}$, and this information cannot be recovered by post processing.

Finally, another approach is to incorporate the forward operator $\mathcal{A}$ and its adjoint $\mathcal{A}^*: Y \to X$ into the neural network. In these learned iterative schemes, classical neural networks are interlaced with applications of the forward and backward operator, thus allowing for the learned reconstruction operator to work directly from data without having to learn the data model. For example, in [50] an ADMM-like scheme for Fourier inversion is learned and [42] consider solving inverse problems typically arising in image restoration by a learned gradient-descent scheme. In [4] this later approach is shown to be applicable to large scale tomographic inversion. Finally, in [5] they apply learning in both spaces X and Y, yielding a Learned Primal-Dual scheme, and show that it outperforms learned post-processing for reconstruction of medical CT images.

Loss functions for learning Once the $\boldsymbol{\Theta}$ parametrization of $\mathcal{A}_{\boldsymbol{\Theta}}^{\dagger}$ is set, the parameters are typically chosen by minimization of some loss functional L. Without doubt, the most common loss function is the mean squared error, also called $\mathcal{L}_2$ loss,

given by

$$L(\Theta) = \mathbb{E}_{\mathsf{f},\mathsf{g}}\left[\| \mathcal{A}_\Theta^\dagger(\mathsf{g}) - \mathsf{f}\|_2^2\right]. \tag{E.2.3}$$

It has however been noted that it is sub-optimal for imaging, and a range of other loss functions have been investigated. These include the classical ℓ_p norms and the structural similarity index (SSIM) [51], as well as more complex losses such as perceptual losses [26] and adversarial networks [33].

Recently, optimal mass transport has also been considered as loss function, e.g., for classification [18], generative models [6, 19], and system identification for stochastic differential equations [23]. In this work we consider using optimal transport for training a reconstruction scheme for ill-posed inverse problems.

Optimal mass transport and Sinkhorn iterations

In optimal mass transport the aim is to transform one distribution into another by moving the mass in a way that minimizes the cost of the movement. For an introduction and overview of the topic, see, e.g., the monograph [46]. Lately, the area has attracted a lot of research [14, 15, 13] with applications to, e.g., signal processing [22, 20, 25, 17], image processing [31], and inverse problems [9, 35, 28, 2].

The optimal mass transport problem can be formulated as follows: let $\Omega \subset \mathbb{R}^d$, and let μ_0 and μ_1 be two measures, defined on Ω, with the same total mass. Given a cost $c : \Omega \times \Omega \to \mathbb{R}_+$ that describes the cost for transporting a unit mass from one point to another, find a (mass preserving) transference plan M that is as cheap as possible. Here, the transference plan characterizes how to move the mass of μ_0 in order to deform it into μ_1. Letting the transference plan be a nonnegative measure dM on the space $\Omega \times \Omega$ yields a linear programming problem in the space of measures:

$$\begin{aligned} T(\mu_0, \mu_1) = \quad & \min_{dM \geq 0} \quad \int_{(x_0,x_1)\in\Omega\times\Omega} c(x_0, x_1) dM(x_0, x_1) \\ & \text{subject to} \quad \mu_0(x_0)dx_0 = \int_{x_1\in\Omega} dM(x_0, x_1), \\ & \qquad\qquad\quad \mu_1(x_1)dx_1 = \int_{x_0\in\Omega} dM(x_0, x_1). \end{aligned} \tag{E.2.4}$$

Moreover, under suitable conditions one can define the Wasserstein metrics W_p using T. This is done by taking $c(x_0, x_1) = d(x_0, x_1)^p$, for $p \geq 1$ and where d is a metric on Ω, and defining $W_p(\mu_0, \mu_1) := T(\mu_0, \mu_1)^{1/p}$ [46, Definition 6.1]. As the name indicates, W_p is a metric on the set of nonnegative measures on Ω with fixed mass [46, Theorem 6.9], and T is weak* continuous on this set. One important property is that T (and thus also W_p) does not only compare objects point by point, as standard L^p metrics, but instead quantifies how the mass is moved. This makes optimal transport natural for quantifying uncertainty and modelling deformations [25, 27].

One way to solve the optimal transport problem in applications is to discretize Ω and solve the corresponding finite-dimensional linear programming problem. In this setting the two measures are represented by point masses on the discretization grid, i.e., by two vectors $\mu_0, \mu_1 \in \mathbb{R}^n_+$ where the element $[\mu_k]_i$ corresponds to the mass in the point $x_{(i)} \in \Omega$ for $i = 1, \dots, n$ and $k = 0, 1$. Moreover, a transference plan is represented by a matrix $M \in \mathbb{R}^{n\times n}_+$ where the value $m_{ij} := [M]_{ij}$ denotes the amount of mass transported from point $x_{(i)}$ to $x_{(j)}$. The associated cost of a transference plan is $\sum_{i,j=1}^n c_{ij} m_{ij} = \mathrm{Tr}(C^T M)$, where $[C]_{ij} = c_{ij} = c(x_{(i)}, x_{(j)})$ is the transportation cost from $x_{(i)}$ to $x_{(j)}$, and by discretizing the constraints we get that M is a feasible transference plan from μ_0 to μ_1 if the row sums of M is μ_0 and the column sums of M is μ_1. The discrete version of (E.2.4) thus takes the form

$$\begin{aligned} T(\mu_0, \mu_1) = \quad & \min_{M \geq 0} \quad \mathrm{Tr}(C^T M) \\ & \text{subject to} \quad \mu_0 = M\mathbf{1}_n, \quad \mu_1 = M^T \mathbf{1}_n, \end{aligned} \tag{E.2.5}$$

where $M \geq 0$ denotes element-wise non-negativity of the matrix. However, even though (E.2.5) is a linear programming problem it is in many cases computationally infeasible due to the vast number of variables. Since $M \in \mathbb{R}^{n\times n}_+$ the number of variables is n^2, and thus if one seek to solve the optimal transport problem between two 512×512 images this results in more than $6 \cdot 10^{10}$ variables.

One approach for addressing this problem was proposed in [14] and introduces an entropic regularizing term $D(M) = \sum_{i,j=1}^n (m_{ij} \log(m_{ij}) - m_{ij} + 1)$ for approximating the transference plan, resulting in the perturbed optimal transport problem

$$\begin{aligned} & \min_{M \geq 0} \quad \mathrm{Tr}(C^T M) + \varepsilon D(M) \\ & \text{subject to} \quad \mu_0 = M\mathbf{1}_n, \quad \mu_1 = M^T \mathbf{1}_n. \end{aligned} \tag{E.2.6}$$

One can show that an optimal solution to (E.2.6) is of the form

$$M = \mathrm{diag}(u) K \mathrm{diag}(v),$$

where $K = \exp(-C/\varepsilon)$ (point-wise exponential) is known, and $u, v \in \mathbb{R}^n_+$ are unknown. This shows that the solution is parameterized by only $2n$ variables. Moreover, the two vectors can be computed iteratively by so called Sinkhorn iterations, i.e., alternatingly compute u and v that matches μ_0 and μ_1 respectively. This is summarizied in Algorithm E.1 where $\odot$ denotes elementwise multiplication and ./ elementwise division. The procedure has been shown to have a linear convergence rate, see [14] and references therein.

Algorithm E.1 Sinkhorn iterations for computing entropy-regularized optimal transport [14]

Input $C, \varepsilon, \mu_0, \mu_1$
initialize $v_0 > 0$ and $K = \exp(-C/\varepsilon)$
for $i = 1, \dots, N$ **do**
  $u_i \leftarrow \mu_0 ./ (K v_{i-1})$
  $v_i \leftarrow \mu_1 ./ (K^T u_i)$
end for
Return $u_N^T (K \odot C) v_N$

Moreover, when the underlying cost $c(x_0, x_1)$ is translation invariant the discretized cost matrix C, and thus also the transformation K, gets a Toeplitz-block-Toeplitz structure. This structure can be used in order to compute Kv and $K^T u$ efficiently using the fast Fourier transform in $\mathcal{O}(n \log n)$, instead of naive matrix-vector multiplication in $\mathcal{O}(n^2)$ [28] (see also [44] for connection to the convolutional Wasserstein-2 distance). This is crucial for applications in imaging since otherwise, for images of size 512×512 pixels one would have to explicitly store and multiply with matrices of size 262144×262144.

The formulations described above are only defined for measures μ_0 and μ_1 with the same total mass. However, they can also be extended to handle measures with unbalanced masses [20, 13]. This can be done by also allowing for adding or subtracting mass in the two marginals, e.g., defining the cost via the optimization problem

$$T^\kappa(\mu_0, \mu_1) := \min_{\nu_0, \nu_1 \geq 0} T(\nu_0, \nu_1) + \kappa \sum_{i=0}^{1} \|\nu_i - \mu_i\|_1 \tag{E.2.7}$$

where κ is the cost of adding or subtracting a unit mass. Here ν_i are nonnegative measures on Ω. As with the standard optimal transport problem, under suitable conditions $T^\kappa(\cdot, \cdot)^{1/p}$ defines a metric if $c(x_0, x_1) = d(x_0, x_1)^p$ where d is a metric on Ω, and $p \geq 1$ (cf. [20]). Moreover, (E.2.7) can be formulated as a optimal mass transport problem where the domain Ω has been extended with an additional point. We can thus apply Sinkhorn-iterations to approximate the solution of (E.2.7), see Appendix E.7 for details.

E.3 Learning a reconstruction operator using Wasserstein loss

As mentioned in the introduction, when training data comes from real applications there might be geometric distortions between the data g and the ground truth f. In this work we propose to use optimal transport as loss function to train a reconstruction operator, i.e., to select the parameters as

$$\Theta^* \in \arg\min_{\Theta} \mathbb{E}_{\mathsf{f},\mathsf{g}} \left[T^\kappa(\mathcal{A}_\Theta^\dagger(\mathsf{g}), \mathsf{f}) \right]. \tag{E.3.1}$$

This should give better results when data g is not aligned with the ground truth f.

The above claim is motivated by considering optimal reconstructions in a simplified setting comparing $\mathcal{L}_2$ and Wasserstein losses. In the ideal case, training the network with the $\mathcal{L}_2$ loss (E.2.3) will result in a perfect reconstruction composed with a convolution that "smears" the reconstruction over the area of possible miss-alignment. On the other hand since optimal mass transport does not only compare objects point-wise, the network will learn a perfect reconstruction combined with a movement of the object to the average miss-alignment (in the ideal case and for a point mass f). These statements are made more precise in the following propositions, whose proofs are deferred to Appendix E.6.

Proposition E.3.1. *Let $g \in \mathcal{L}_2(\mathbb{R}^n)$, let τ be a $\mathbb{R}^n$-valued random variable with probability measure $dP(t)$, and let $g_\tau(x) := g(x - \tau)$. Then there exists a function $f \in \mathcal{L}_2(\mathbb{R}^n)$ that minimizes $\mathbb{E}_\tau\big[\|f - g_\tau\|_2^2\big]$, and this f has the form*

$$f(x) = (dP * g)(x) := \int_{\mathbb{R}^n} g(x - t)dP(t).$$

Proposition E.3.2. *Let τ be a $\mathbb{R}^n$-valued random variable with probability measure $dP(t)$, and $\delta_\tau(x) := \delta(x - \tau)$ where δ denotes the Dirac delta distribution on $\mathbb{R}^n$. Then*

$$\mathbb{E}_\tau\big[T(\delta_\tau, \mu)\big] = \int_{\mathbb{R}^n} \Big(\underbrace{\int_{\mathbb{R}^n} c(t, x)dP(t)}_{:=F(x)} \Big) d\mu(x)$$

whenever μ is a positive measure with unit mass, and $\mathbb{E}_\tau\big[T(\delta_\tau, \mu)\big] = \infty$ otherwise. Moreover, finding a μ that minimizes $\mu \mapsto \mathbb{E}_\tau\big[T(\delta_\tau, \mu)\big]$ is equivalent to finding a global minimizer to $x \mapsto F(x)$. In particular, if (i) the probability measure dP is symmetric around its mean, (ii) the underlying cost c is of the form $c(t, x) = d(x-t)$, where d is convex and symmetric, then $\mu(x) = \delta\big(x - \mathbb{E}[\tau]\big)$ is an optimal solution. Furthermore, if d is also strictly convex, then this is the unique minimizer.

To illustrate Propositions E.3.1 and E.3.2 we consider the following example.

Example E.3.3. Let τ be uniformly distributed on $[-1, 1]$, and let $c(x_0, x_1) = (x_0 - x_1)^2$. This gives

$$F(x) = \frac{1}{2}\int_{-1}^{1} (x - t)^2 dt = \frac{1}{3} + x^2,$$

which has minimum $x = 0$, and hence the (unique) minimizer to $\mathbb{E}_\tau\big[T(\delta_\tau, \mu)\big]$ is $\mu(x) = \delta(x)$. For the $\mathcal{L}_2$ case with the uniform distribution, the minimizer of $\mathbb{E}_\tau\big[\|f - g_\tau\|_2^2\big]$ is the smoothed function $g * \frac{1}{2}\chi_{[-1,1]}$.

The most common choice of distance c is to use the squared norm $c(x_0, x_1) = \|x_0 - x_1\|^2$, as in the previous example. In this case the result of Proposition E.3.2 can be strengthened, as shown in the following example.

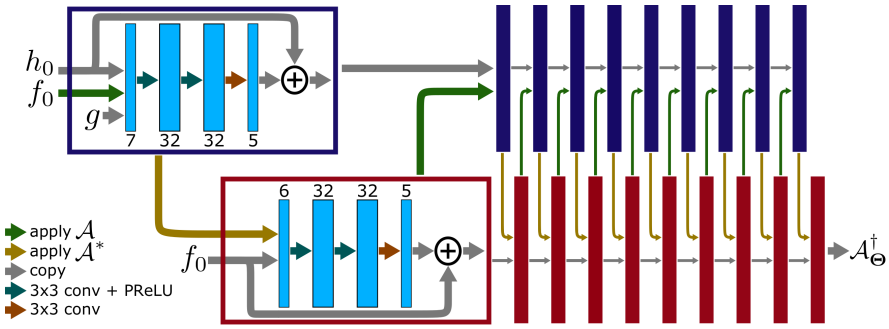


Figure E.1: Network architecture used to solve the inverse problem. Dual and primal iterates are in blue and red boxes, respectively. Several arrows pointing to the same box indicates concatenation. The initial values f_0, h_0 enter from the left, while the data g is supplied to the dual iterates.

Example E.3.4. Let τ be a $\mathbb{R}^n$-valued random variable with probability measure $dP(t)$ with finite first and second moments, and let $c(x_0, x_1) = \|x_0 - x_1\|^2$. This gives

$$F(x) = \int_{\mathbb{R}^n} (x - t)^2 dP(t) = x^2 - 2x\,\mathbb{E}[\tau] + \mathbb{E}[\tau^2],$$

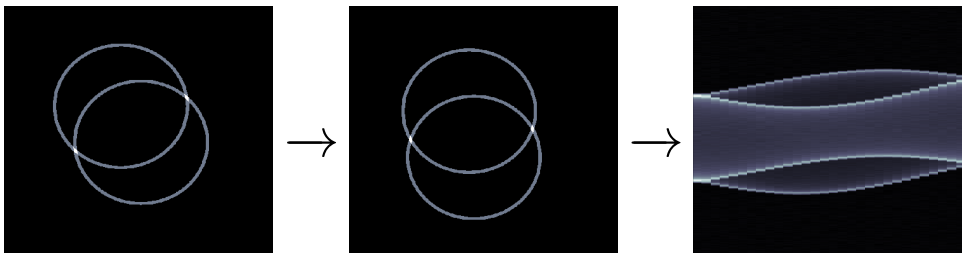
which has a unique global minimum in $x = \mathbb{E}[\tau]$ and hence the unique global minimizer to $\mathbb{E}_\tau\big[T(\delta_\tau, \mu)\big]$ is given by $\mu(x) = \delta\big(x - \mathbb{E}[\tau]\big)$.

E.4 Implementation and evaluation

We use the recently proposed learned primal-dual structure in [5] for learning a reconstruction operator $\mathcal{A}_\Theta^\dagger$ for solving the inverse problem in (E.2.1). In this algorithm, a sequence of small blocks work alternatingly in the data (dual) space Y and the reconstruction (primal) space X and are connected using the forward operator $\mathcal{A}$ and its adjoint $\mathcal{A}^*$. The algorithm works with any differentiable operator $\mathcal{A}$, but we state the version for linear operators for simplicity in Algorithm E.2. A specific instance of this network is also shown in graphical format in Figure E.1.

We compare a learned reconstruction operator of this form when trained using $\mathcal{L}_2$ loss (E.2.3) and using optimal transport loss (E.3.1). Moreover, the evaluation is done on a problem similar to the evaluation problem in [5, 4], i.e., on a problem in computed tomography. More specifically, training is done on data that consists of randomly generated circles on a domain of 512×512 pixels, and the forward operator $\mathcal{A}$ is the ray transform [36]. The ray transform is a linear operator that maps a function to a set of line integrals, i.e., for given $f \in X$ and any line ℓ it is defined by

$$\mathcal{A}(f)(\ell) = \int_\ell f ds$$



(a) Phantom. (b) Translated phantom. (c) Data.

Figure E.2: Example of data generation process used for training and validation, where E.2a shows an example phantom, E.2b is the phantom with a random translation and E.2c is the data (sinogram) corresponding to E.2b with additive white noise on top. Pairs of the form $(g_i, f_i) =$ (E.2c, E.2a) is what is used in the training.

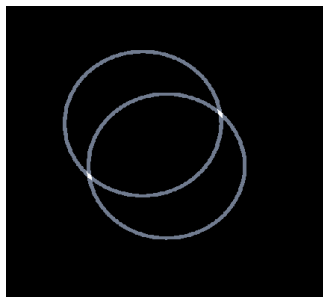
where ds is the line measure. The adjoint of this operator is given by the back-projection [34].

What makes this an ill-posed problem is that the data acquisition is done from only 30 views with 727 parallel lines. Moreover, the source of noise is two-fold in this set-up: (i) the pairs (g_i, f_i) of data sets and phantoms are not aligned, meaning that the data is computed from a phantom with a random change in position. This random change is independent for the different circles, and for each circle it is a shift which is uniformly distributed over $[-40, 40]$ pixels, both in up-down and left-right direction. (ii) on the data computed from the shifted phantom, 5% additive Gaussian noise was added. For an example, see Figure E.2.

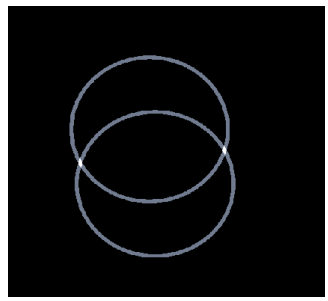
We use the optimal mass transport distance (E.2.7), which allows for unbalanced marginals, and compute it with Sinkhorn iterations. The underlying transportation cost is

$$c(x_1, x_2) = \left(1 - e^{-\|x_1 - x_2\|^4/80^4}\right)$$

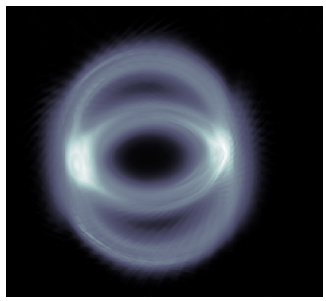
and the cost for adding and subtracting mass is $\kappa = 1$. The cost function c was chosen since it heavily penalizes large movements, while not diverging to infinity which causes numerical instabilities. Moreover, $c(x_1, x_2)^{1/4}$ is in fact a metric on $\mathbb{R}^2$ (see Appendix E.8) and thus $W_4(\mu_0, \mu_1) := T(\mu_0, \mu_1)^{1/4}$ gives rise to a Wasserstein metric on the space of images, where $T(\mu_0, \mu_1)$ is the optimal mass transport distance with the transport cost $c(x_1, x_2)$. Since this cost is translation invariant, the matrix-vector multiplications Ku and $K^T v$ can be done with fast Fourier transform, as mentioned in Section E.2. We used 10 Sinkhorn iterations with entropy regularization $\varepsilon = 10^{-3}$, to approximate the optimal mass transport. Automatic differentiation in TensorFlow was used to back-propagate the result during training.



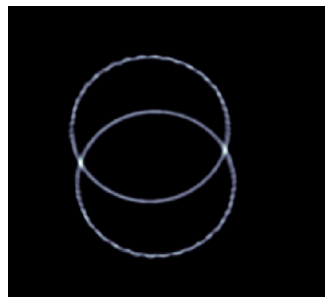
(a) Phantom.



(b) Translated phantom used for generating data.



(c) Reslut after training with mean squared error loss.



(d) Result after training with Wasserstein loss.

Figure E.3: In E.3a we show the validation phantom, which was generated from the same training set but not used in training, in E.3b the translated phantom from which the validation data was computed, in E.3c a reconstruction with the neural network trained using mean squared error loss (E.2.3), and in E.3d a reconstruction with the neural network trained using Wasserstein loss (E.3.1).

Algorithm E.2 Learned Primal-Dual reconstruction algorithm

Initialize $f_0 \in X^{N_{\text{primal}}}, h_0 \in U^{N_{\text{dual}}}$
for $i = 1, \dots, I$ **do**
  $h_i \leftarrow \Gamma_{\boldsymbol{\Theta}_i^d}\big(h_{i-1}, \mathcal{A}(f_{i-1}^{(2)}), g\big)$
  $f_i \leftarrow \Lambda_{\boldsymbol{\Theta}_i^p}\big(f_{i-1}, \mathcal{A}^*(h_i^{(1)})\big)$
end for
$\mathcal{A}_\Theta^\dagger(g) := f_I^{(1)}$

The reconstruction method in algorithm E.2 was implemented using ODL [3], ASTRA [38, 45], and TensorFlow [1]. We used the reference implementation from [5] with default parameters, i.e., the number of blocks in the primal and

dual space was $I = 10$, and the number of primal and dual variables was set to $N_{\text{primal}} = N_{\text{dual}} = 5$. Moreover, the blocks used a residual structure and had three layers of 3×3 convolutions with 32 filters. PReLU nonlinearities were used. Thus, this corresponds to a residual CNN with convolutional depth of $10 \cdot 2 \cdot 3 = 60$, as shown in graphical format in Figure E.1. We used zero initial values, $f_0 = h_0 = 0$.

The training also followed [5] closely. In particular, we used $2 \cdot 10^4$ batches of size 1, using the ADAM optimizer [29] with default values except for $\beta_2 = 0.99$. The learning rate (step length) used was cosine annealing [32] with initial step length 10^{-3}. Moreover, in order to improve training stability we performed gradient norm clipping [39] with norms limited to 1. The convolution parameters were initialized using Xavier initialization [21], and all biases were initialized to zero. The training took approximately 3 hours using a single Titan X GPU. The source code used to replicate these experiments are available online.[1].

Results are presented in Figure E.3. As can be seen, the reconstruction using $\mathcal{L}_2$ loss "smears" the reconstruction, in this case to an extent where the shape is hard to recover. On the other hand, the reconstruction using the Wasserstein loss retains the over-all global shape of the object, although relative and exact positions of the circles are not recovered. These results are qualitatively in line with the results in the simplified setting of section E.3, where the optimal $\mathcal{L}_2$ reconstructions are smeared (proposition E.3.1) and where the optimal Wasserstein reconstructions are sharp (proposition E.3.2). This suggests that training with a Wasserstein loss can be useful when there are miss-alignments in the training data.

E.5 Conclusions and future work

In this work we have considered using Wasserstein loss to train a neural network for solving ill-posed inverse problems in imaging where data is not aligned with the ground truth. We give a theoretical motivation for why this should give better results compared to standard mean squared error loss, and demonstrate it on a problem in computed tomography. In the future, we hope that this method can be applied to other inverse problems and to other problems in imaging such as segmentation.

E.6 Appendix 1: Deferred proofs

Proof of Proposition E.3.1. To show that $f(x) = (dP * g)(x) \in \mathcal{L}_2(\mathbb{R}^n)$ minimizes $\mathbb{E}_\tau[\|f - g_\tau\|_2^2]$ we note that

$$\begin{aligned}\mathbb{E}_\tau\left[\|f - g_\tau\|_2^2\right] &= \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} (f(x) - g_t(x))^2 dx \right) dP(t) \\ &= \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} (f(x) - g_t(x))^2 dP(t) \right) dx\end{aligned}$$

[1] https://github.com/adler-j/wasserstein_inverse_problems

by Fubini's theorem. Next, by expanding the expression using that $\int_{\mathbb{R}^n} dP(t) = 1$ and completing the square, this can be written as

$$\mathbb{E}_\tau\left[\|f - g_\tau\|_2^2\right] = \int_{\mathbb{R}^n} \left(f(x) - \int_{\mathbb{R}^n} g_t(x) dP(t)\right)^2 dx + c,$$

where c is a constant. Using this it follows that the minimizing f is of the form

$$f(x) = \int_{\mathbb{R}^n} g_t(x) dP(t) = (dP * g)(x).$$

To see that $f \in \mathcal{L}_2(\mathbb{R}^n)$ we note that, by using Fubini's theorem, we have

$$\begin{aligned}
\|f\|_2^2 &= \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} g_t(x) dP(t)\right)^2 dx \\
&= \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} g_s(x) g_t(x) dx\right) dP(s) dP(t) \\
&\leq \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \left(\frac{1}{2} \int_{\mathbb{R}^n} g_s(x)^2 + g_t(x)^2 dx\right) dP(s) dP(t) \\
&= \|g\|_2^2 < \infty
\end{aligned}$$

where the first inequality is the arithmetic-geometric mean inequality. This completes the proof. □

Proof of Proposition E.3.2. We consider finding the marginal μ that minimizes the expectation $\mathbb{E}_\tau\big[T(\delta_\tau, \mu)\big]$. Without loss of generality we assume that τ is zero-mean, since otherwise we simply consider $\tilde{\tau} = \tau - \mathbb{E}[\tau]$, where $\tilde{\tau}$ is a zero-mean random variable, and shift the coordinate system. First we note that $T(\delta_t, \mu)$ is only finite for nonegative measures μ with total mass 1, and hence $\mathbb{E}_\tau\big[T(\delta_\tau, \mu)\big]$ is only finite for such measures. Second, for such a μ we have

$$T(\delta_t, \mu) = \int_{\mathbb{R}^n} c(t, x) d\mu(x),$$

since one needs to transport all mass in μ into the point t where δ_t has its mass. Using this and expanding the expression for the expectation gives that

$$\begin{aligned}
\mathbb{E}_\tau\big[T(\delta_\tau, \mu)\big] &= \int_{\mathbb{R}^n} T(\delta_t, \mu) dP(t) \\
&= \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} c(t, x) dP(t)\right) d\mu(x),
\end{aligned}$$

where we have used Fubini's theorem in the last step. From this we note that the optimal μ have support only in the global minimas of the function $F(x) := \int_{\mathbb{R}^n} c(t, x) dP(t)$. This completes the first half of the statement.

To prove the second half of the statement, since $c(t,x) = d(x-t)$ we have that

$$F(x) = \int_{\mathbb{R}^n} d(x-t)dP(t).$$

We now note that since dP and d are symmetric we must have that F is also symmetric, i.e., that $F(-x) = F(x)$. However, this means that for any $x \in \mathbb{R}^n$ we have that

$$\begin{aligned} F(x) &= \frac{1}{2}F(x) + \frac{1}{2}F(-x) \\ &= \frac{1}{2}\int_{\mathbb{R}^n} d(x-t)dP(t) + \frac{1}{2}\int_{\mathbb{R}^n} d(-x-t)dP(t) \\ &= \int_{\mathbb{R}^n} \left(\frac{1}{2}d(x-t) + \frac{1}{2}d(-x-t)\right) dP(t) \\ &\geq \int_{\mathbb{R}^n} d(t)dP(t) = F(0), \end{aligned}$$

where the inequality follows from convexity and symmetry of d. This shows that $F(x)$ has a global minimum in $x = 0$, and hence by the first part of the proof it follows that an optimal solution is $\mu(x) = \delta(x)$. Now, if d is strictly convex the inequality is strict for $x \neq 0$, which shows that the optimal solution is unique. This completes the proof. □

E.7 Appendix 2: OMT for unbalanced marginals via Sinkhorn iterations

Recall the Kantorovich formulation of the optimal mass transport problem (E.2.4)

$$\begin{aligned} T(\mu_0,\mu_1) = \quad & \min_{dM\geq 0} \quad \int_{(x_0,x_1)\in\Omega\times\Omega} c(x_0,x_1)dM(x_0,x_1) \\ & \text{subject to} \quad \mu_0(x_0)dx_0 = \int_{x_1\in\Omega} dM(x_0,x_1), \\ & \qquad\qquad\quad \mu_1(x_1)dx_1 = \int_{x_0\in\Omega} dM(x_0,x_1). \end{aligned}$$

This is only well-defined if the two marginals have the same total mass, but can be extended to handle the case when the two marginals have different mass [20, 13]. One such formulation is

$$T^\kappa(\mu_0,\mu_1) := \min_{\nu_0,\nu_1\geq 0} T(\nu_0,\nu_1) + \kappa\sum_{i=0}^{1}\|\nu_i - \mu_i\|_1,$$

which in the discrete setting this becomes

$$T^{\kappa}(\mu_0,\mu_1) := \min_{\nu_0,\nu_1,M\geq 0} \operatorname{Tr}(C^T M) + \kappa \sum_{i=0}^{1} \|\nu_i - \mu_i\|_1 \quad \text{subject to} \quad \nu_0 = M\mathbf{1}_n,\ \nu_1 = M^T\mathbf{1}_n. \tag{E.7.1}$$

Here we will show how Sinkhorn iterations can be used to compute an approximate solution to the problem (E.7.1). First note that without loss of generality we can always assume that $\|\mu_0\|_1 \leq \|\mu_1\|_1$, due to the symmetry. One can then show the following lemma.

Lemma E.7.1. *Let the cost matrix be $C = [c_{ij}]_{ij}$ where $c_{ij} \geq 0$ and $c_{ii} = 0$. Then there is an optimal solution such that ν_0^*, ν_1^* to* (E.7.1) *such that $\nu_0^* = \mu_0$. Moreover, assume that $\|\mu_0\|_1 \leq \|\mu_1\|_1$. If $\kappa \geq \frac{1}{2}\max_{ij} c_{ij}$ then there is an optimal solution such that $\nu_0^* = \mu_0$ and*

$$\nu_1^* \leq \mu_1. \tag{E.7.2}$$

Proof. Let ν_0^*, ν_1^*, M^* denote an optimal solution to (E.7.1), and let m_i^* be the ith row of M^* and $\tilde{m}_j^*$ be the jth column. Also note that feasibility of (E.7.1) implies that $\nu_0^* = M^*\mathbf{1}_n$ and $\nu_1^* = (M^*)^T\mathbf{1}_n$ are determined by the transport plan M^* and have the same total mass.

To show the first statement, first assume that there is point $x_{(i)}$ such that $\nu_0^*(i) < \mu_0(i)$. A new transport plan can be obtained as $M = M^* + e_i e_i^T(\mu_0(i) - \nu_0^*(i))$, with associated marginals $\nu_\ell = \nu_\ell^* + e_i(\mu_0(i) - \nu_0^*(i))$. Since $c_{ii} = 0$, the objective value of this transport plan is less than or equal to that of M^* and hence also optimal.

Next, assume that there is point $x_{(i)}$ such that $\nu_0^*(i) > \mu_0(i)$. In this case let M be th transport plan where $m_i = m_i^* \cdot \mu_0(i)/\nu_0^*(i)$ and $m_\ell = m_\ell^*$ for $\ell \neq i$. Since $c_{ij} \geq 0$ again the objective value of this transport plan is less than or equal to that of M^* and hence also optimal.

Repeating these two steps we can modify the transport plan so that $\nu_0 = \mu_0$ without increasing the objective value. This proves the first statement.

To prove the second statement we can, without loss of generality, take $\nu_0^* = \mu_0$. By the assumption $\|\mu_0\|_1 \leq \|\mu_1\|_1$ this means that $\|\nu_1^*\|_1 \leq \|\mu_1\|_1$. Now, assume that the inequality (E.7.2) is violated. Then there must be two points $x_{(i)}$ and $x_{(j)}$ so that $\mu_1(i) < \nu_1^*(i)$ and $\nu_1^*(j) < \mu_1(j)$. In this case consider the transport plan M given by $\tilde{m}_i = (1-\alpha)\tilde{m}_i^*$, $\tilde{m}_j = \tilde{m}_j^* + \alpha\tilde{m}_i^*$, and $\tilde{m}_\ell = \tilde{m}_\ell^*$ otherwise, where $\alpha = \min\{\nu_1^*(i) - \mu_1(i), \mu_1(j) - \nu_1^*(j)\}/\|\tilde{m}_i^*\|_1$. This gives at most an increase in the transport cost of $\alpha \max_{ij} c_{ij}\|\tilde{m}_i^*\|_1$ which is less than or equal to $2\alpha\kappa\|\tilde{m}_i^*\|_1$, the latter corresponding to the decrease in the second term of the cost function. Since $\|\nu_1^*\|_1 \leq \|\mu_1\|_1$ we can repeat this process without increasing the objective function until the inequality (E.7.2) is satisfied. □

Based on lemma E.7.1 we can introduce a modified optimal transport problem that is equivalent to (E.7.1) but where the second term in the cost function is explicitly removed by incorporating it into C and M. To this end, let $m_{n+1} \in \mathbb{R}_+^{1\times n}$ and define

$$\tilde{M} = \begin{bmatrix} M \\ m_{n+1} \end{bmatrix},\ \tilde{C} = \begin{bmatrix} C \\ \kappa \mathbf{1}_n^T \end{bmatrix},\ \tilde{\mu_0} = \begin{bmatrix} \mu_0 \\ (\mu_1 - \mu_0)^T \mathbf{1}_n \end{bmatrix}.$$

Here m_{n+1}^T corresponds to the difference $\mu_1 - \nu_1$, which by the inequality (E.7.2) in lemma E.7.1 has a nonnegative optimal solution. We can thus write (E.7.1) as

$$\begin{aligned} \tilde{T}^\kappa(\mu_0, \mu_1) := \min_{\tilde{M}\geq 0}\ & \operatorname{Tr}(\tilde{C}^T \tilde{M}) \\ \text{subject to}\quad & \tilde{\mu_0} = \tilde{M}\mathbf{1}_{n+1},\ \mu_1 = \tilde{M}^T \mathbf{1}_n. \end{aligned}$$

This is an optimal transport problem where we have added an extra mass in μ_0, corresponding to the difference in total mass, from which it costs κ to move the mass to any other point in μ_1. We can now make an entropy-regularization of this problem, which gives and optimal solution $\tilde{M} = \operatorname{diag}(\tilde{u})\tilde{K}\operatorname{diag}(v)$, where $\tilde{K} = \exp(-\tilde{C}/\varepsilon)$ and $\tilde{u} = [u,\ \hat{u}]^T$ for $\hat{u}$ scalar. As before, this solution can be obtained via the Sinkhorn iterations, which takes the form

$$\begin{aligned} \tilde{u}_i &\leftarrow \tilde{\mu_0}./(\tilde{K}v_{i-1}) = \begin{bmatrix} \mu_0./(Kv_{i-1}) \\ (\mu_1 - \mu_0)^T\mathbf{1}_n/(\mathbf{1}_n v)\exp(\kappa/\varepsilon) \end{bmatrix} \\ v_i &\leftarrow \mu_1./(\tilde{K}^T\tilde{u}_i) = \mu_1./(K^T u_i + \exp(-\kappa/\varepsilon)\mathbf{1}_n\hat{u}). \end{aligned}$$

From this we note that we can still utilize the FFT-based methods for fast computations.

Remark E.7.2. Note that one can also get a symmetric formulation by using the following redefinitions instead

$$\tilde{M} = \begin{bmatrix} M & \tilde{m}_{n+1} \\ m_{n+1} & m_{n+1,n+1} \end{bmatrix},\quad \tilde{C} = \begin{bmatrix} C & \kappa\mathbf{1}_n \\ \kappa\mathbf{1}_n^T & 0 \end{bmatrix},$$

and

$$\tilde{\mu_0} = \begin{bmatrix} \mu_0 \\ \mu_1^T\mathbf{1}_n \end{bmatrix}, \tilde{\mu_1} = \begin{bmatrix} \mu_1 \\ \mu_0^T\mathbf{1}_n \end{bmatrix},$$

E.8 Appendix 3: Metric property of the cost function

This appendix is dedicated to proving the following lemma.

Lemma E.8.1. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^m$. Then*

$$d(x_1, x_2) = \left(1 - e^{-\|x_1 - x_2\|^n}\right)^{\frac{1}{n}}.$$

is a metric on $\mathbb{R}^m$ for $n \geq 1$.

Proof. It is easily seen that $d(x_1, x_2)$ is symmetric, nonnegative, and equal to zero if only if $x_1 = x_2$. Thus we only need to verify that the triangle inequality holds. To this end we note that if

$$\left(1 - e^{-(a+b)^n}\right)^{\frac{1}{n}} \leq \left(1 - e^{-a^n}\right)^{\frac{1}{n}} + \left(1 - e^{-b^n}\right)^{\frac{1}{n}}, \forall\, a, b \geq 0, \tag{E.8.1}$$

for all $n \geq 1$, then by taking $a = \|x_1 - x_2\|$, $b = \|x_2 - x_3\|$, and using the triangle inequality for the norm $\|\cdot\|$ we have that

$$\begin{aligned} d(x_1, x_3) &= \left(1 - e^{-\|x_1 - x_3\|^n}\right)^{\frac{1}{n}} \\ &\leq \left(1 - e^{-(\|x_1 - x_2\| + \|x_2 - x_3\|)^n}\right)^{\frac{1}{n}} \\ &\leq \left(1 - e^{-\|x_1 - x_2\|^n}\right)^{\frac{1}{n}} + \left(1 - e^{-\|x_2 - x_3\|^n}\right)^{\frac{1}{n}} \\ &= d(x_1, x_2) + d(x_2, x_3). \end{aligned}$$

Therefore we will show that (E.8.1) holds for all $n \geq 1$, and to do so we will

(i) show that if a function $g : \mathbb{R}_+ \to \mathbb{R}_+$ fulfills $g(0) = 0$, $g(x)' \geq 0$, $g''(x) \leq 0$ for all $x \in \mathbb{R}_+$, then $g(x_1 + x_2) \leq g(x_1) + g(x_2)$,

(ii) show that for $x \geq 0$ the map $x \mapsto (1 - e^{-x^n})^{\frac{1}{n}}$ fulfills the assumptions in (i) for any $n \geq 1$.

To show (i) we note that

$$\begin{aligned} g(x_1 + x_2) &= \int_0^{x_1 + x_2} g'(t)dt = \int_0^{x_1} g'(t)dt + \int_{x_1}^{x_1 + x_2} g'(t)dt \\ &\leq \int_0^{x_1} g'(t)dt + \int_0^{x_2} g'(t)dt = g(x_1) + g(x_2), \end{aligned}$$

where the inequality uses that $g'(t) \geq g'(x + t)$ for any $x, t \geq 0$ since $g''(x) \leq 0$ for all $x \geq 0$.

To show (ii), let $g(x) := (1 - e^{-x^n})^{\frac{1}{n}}$ and observe that $g(0) = 0$. Differentiating g twice gives

$$\begin{aligned} g'(x) &= \frac{e^{-x^n}(1 - e^{-x^n})^{\frac{1}{n}} x^{n-1}}{1 - e^{-x^n}} \\ g''(x) &= -\frac{(1 - e^{-x^n})^{\frac{1}{n}} x^{n-2}}{(e^{x^n} - 1)^2} \\ &\quad \cdot \big(\underbrace{n e^{x^n} x^n - x^n + e^{x^n} - n e^{x^n} + n - 1}_{=: h_n(x^n)} \big). \end{aligned}$$

For $x \geq 0$ we see that $g'(x) \geq 0$ for all $n \geq 1$. Moreover, for $x \geq 0$ we see that $g''(x) \leq 0$ for all $x \geq 0$ and for all $n \geq 1$ if and only if $h_n(x^n) \geq 0$. With the change

of variable $x^n = y$, we thus want to show that $h_n(y) \geq 0$ for all $y \geq 0$ and all $n \geq 1$. To see this we note that $h_n(0) = 0$ and that

$$h_n'(y) = ne^y y + e^y - 1 \geq 0 \quad \text{for all } y \geq 0 \text{ and } n \geq 1.$$

This shows (ii), and hence completes the proof. □

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] I. Abraham, R. Abraham, M. Bergounioux, and G. Carlier. Tomographic reconstruction from a few views: A multi-marginal optimal transport approach. *Applied Mathematics & Optimization*, 75(1):55–73, 2017.

[3] J. Adler, H. Kohr, and O. Öktem. Odl 0.6.0, April 2017.

[4] J. Adler and O. Öktem. Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Problems*, 33(12):124007, 2017.

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

[6] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

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

[8] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[9] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[10] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, Jan. 2011.

[11] 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.

[12] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[13] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Unbalanced optimal transport: Geometry and Kantorovich formulation. *arXiv preprint arXiv:1508.05216*, 2015.

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

[15] M. Cuturi and G. Peyré. A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, pages 320–343, 2016.

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

[17] B. Engquist and B.D. Froese. Application of the Wasserstein metric to seismic signals. *Communications in Mathematical Sciences*, 12(5), 2014.

[18] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T.A. Poggio. Learning with a wasserstein loss. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2053–2061, 2015.

[19] A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with Sinkhorn divergences. *arXiv preprint arXiv:1706.00292*, 2017.

[20] T.T. Georgiou, J. Karlsson, and M.S. Takyar. Metrics for power spectra: An axiomatic approach. *IEEE Transactions on Signal Processing*, 57(3):859–867, 2009.

[21] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[22] S. Haker, L. Zhu, A. Tannenbaum, and S. Angenent. Optimal mass transport for registration and warping. *International Journal of computer vision*, 60(3):225–240, 2004.

[23] T. Hashimoto, D. Gifford, and T. Jaakkola. Learning population-level diffusions with generative RNNs. In *International Conference on Machine Learning (ICML)*, pages 2417–2426, 2016.

[24] V. Jain and S. Seung. Natural image denoising with convolutional networks. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 769–776, 2009.

[25] X. Jiang, Z.-Q. Luo, and T.T. Georgiou. Geometric methods for spectral analysis. *IEEE Transactions on Signal Processing*, 60(3):1064–1074, 2012.

[26] J. Johnson, A. Alahi, and L. Fei-Fei. *Perceptual Losses for Real-Time Style Transfer and Super-Resolution*, pages 694–711. Springer International Publishing, Cham, 2016.

[27] J. Karlsson and T.T Georgiou. Uncertainty bounds for spectral estimation. *IEEE Transactions on Automatic Control*, 58(7):1659–1673, 2013.

[28] 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.

[29] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[30] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436—444, 2015.

[31] J. Lellmann, D.A. Lorenz, C. Schönlieb, and T. Valkonen. Imaging with Kantorovich–Rubinstein discrepancy. *SIAM Journal on Imaging Sciences*, 7(4):2833–2859, 2014.

[32] I. Loshchilov and F. Hutter. SGDR: Stochastic gradient descent with restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[33] M. Mardani, E. Gong, J.Y. Cheng, S. Vasanawala, G. Zaharchuk, M.T. Alley, N. Thakur, S. Han, W.J. Dally, J.M. Pauly, and L. Xing. Deep generative adversarial networks for compressed sensing automates MRI. *CoRR*, abs/1706.00051, 2017.

[34] A. Markoe. *Analytic Tomography*. Encyclopedia of mathematics and its applications. Cambridge University Press, New York, NY, 2006.

[35] L. Métivier, R. Brossier, Q. Merigot, E. Oudet, and J. Virieux. An optimal transport approach for seismic tomography: Application to 3d full waveform inversion. *Inverse Problems*, 32(11):115008, 2016.

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

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

[38] W.J. Palenstijn, K.J. Batenburg, and J. Sijbers. Performance improvements for iterative electron tomography reconstruction using graphics processing units (GPUs). *Journal of structural biology*, 176(2):250–253, 2011.

[39] R. Pascanu, T. Mikolov, and Y. Bengio. Understanding the exploding gradient problem. *CoRR, abs/1211.5063*, 2012.

[40] P. Paschalis, N. D. Giokaris, A. Karabarbounis, G. K. Loudos, D. Maintas, C. N. Papanicolas, V. Spanoudaki, Ch. Tsoumpas, and E. Stiliaris. Tomographic image reconstruction using artificial neural networks. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 527(1–2):211–215, 2004.

[41] D. M. Pelt and K. J. Batenburg. Fast tomographic reconstruction from limited data using artificial neural networks. *IEEE Transactions on Image Processing*, 22(12):5238–5251, 2013.

[42] P. Putzky and M. Welling. Recurrent inference machines for solving inverse problems. *arXiv preprint arXiv:1706.04008*, 2017.

[43] L.I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

[44] J. Solomon, F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66, 2015.

[45] W. van Aarle, W.J. Palenstijn, J. Cant, E. Janssens, F. Bleichrodt, A. Dabravolski, J. De Beenhouwer, K.J. Batenburg, and J. Sijbers. Fast and flexible X-ray tomography using the ASTRA toolbox. *Optics express*, 24(22):25129–25147, 2016.

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

[47] A. Walker, G. Liney, P. Metcalfe, and L. Holloway. MRI distortion: Considerations for MRI based radiotherapy treatment planning. *Australasian Physical & Engineering Sciences in Medicine*, 37(1):103–113, Mar 2014.

[48] T. Würfl, F. C. Ghesu, V. Christlein, and A. Maier. Deep learning computed tomography. In S. Ourselin, L. Joskowicz, M. Sabuncu, G. Unal, and W. Wells, editors, *MICCAI 2016: Medical Image Computing and Computer-Assisted Intervention*, volume 9902 of *Lecture Notes in Computer Science*, pages 432–440. Springer-Verlag, 2016.

[49] J. Xie, L. Xu, and E. Chen. Image denoising and inpainting with deep neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 341–349, 2012.

[50] Y. Yang, J. Sun, H. Li, and Z. Xu. Deep ADMM-Net for compressive sensing MRI. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 29, pages 10–18, 2016.

[51] H. Zhao, O. Gallo, I. Frosio, and J. Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, March 2017.

Paper F

Data-driven nonsmooth optimization

Data-driven nonsmooth optimization

by

Sebastian Banert, Axel Ringh, Jonas Adler,
Johan Karlsson, and Ozan Öktem

Abstract

In this work, we consider methods for solving large-scale optimization problems with a possibly nonsmooth objective function. 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. This scheme contains many modern primal-dual first-order solvers like the Douglas-Rachford and hybrid gradient methods as special cases. Moreover, we show convergence to an optimal point for a new method which also belongs to this class. Next, we interpret the generic scheme as a neural network and use unsupervised training to learn the best set of parameters for a specific class of objective functions while imposing a fixed number of iterations. In contrast to other approaches of "learning to optimize", we present an approach which learns parameters only in the set of convergent schemes. As use cases, we consider optimization problems arising in tomographic reconstruction and image deconvolution, and in particular a family of total variation regularization problems.

Keywords: convex optimization, proximal algorithms, monotone operators, machine learning, inverse problems, computed tomography

F.1 Introduction

Many problems in science and engineering can be formulated as convex optimization problems which then need to be solved accurately and efficiently. In this paper we focus on methods for solving such problems, namely of the form

$$\min_{x\in\mathcal{X}}\Big[F(x)+\sum_{i=1}^{m}G_i(L_ix)\Big]. \tag{F.1.1}$$

Here, $L_i\colon \mathcal{X}\to\mathcal{Y}_i$, $i=1,\dots,m$, are linear operators, where $\mathcal{X},\mathcal{Y}_1,\dots,\mathcal{Y}_m$ are Hilbert spaces, and $F\colon \mathcal{X}\to\overline{\mathbb{R}}$ and $G_i\colon \mathcal{Y}_i\to\overline{\mathbb{R}}$, $i=1,\dots,m$, are proper, convex and lower semicontinuous functions. This class of optimization problems appears for example in variational regularization of inverse problems in imaging, such as X-ray computed tomography (CT) [40, 39], magnetic resonance imaging (MRI) [17], and electron tomography [41].

A key challenge is to handle the computational burden. In imaging, and especially so for three-dimensional imaging, the resulting optimization problem is very high-dimensional even after clever digitization and might involve more than one billion variables. Moreover, many regularizers that are popular in imaging (see Section F.5), like those associated with sparsity, result in a nonsmooth objective function. These issues prevent usage of variational methods in time-critical applications, such as medical imaging in a clinical setting. Modern methods which aim at overcoming these obstacles are typically based on the proximal point algorithm [45] and operator splitting techniques, see e.g., [24, 11, 19, 15, 20, 21, 28, 14, 13, 32, 33, 9] and references therein.

The main objective of the paper is to offer a computationally tractable approach for minimizing large-scale nondifferentiable, convex functions. The key idea is to "learn" how to optimize from training data, resulting in an iterative scheme that is optimal given a fixed number of steps, while its convergence properties can be analyzed. We will make this precise in Section F.4.

Similar ideas have been proposed previously in [26, 34, 7], but these approaches are either limited to specific classes of iterative schemes, like gradient-descent-like schemes [34, 7] that are not applicable for nonsmooth optimization, or specialized to a specific class of regularizers as in [26], which limits the possible choices of regularizers and forward operators. The approach taken here leverages upon these ideas and yields a general framework for learning optimization algorithms that are applicable to solving optimization problems of the type (F.1.1), inspired by the proximal-type methods mentioned above.

A key feature is to present a general formulation that includes several existing algorithms, among them the primal-dual hybrid gradient (PDHG) algorithm (also called the Chambolle–Pock algorithm) [19] and the primal-dual Douglas–Rachford algorithm [14] as a special case. This means that the learning can be done in a space of schemes that includes these solvers as special cases. Moreover, from the proposed parametrization we also derive a new optimization algorithm. We demonstrate the performance of a solver based on this general formulation by training in an unsupervised manner for two inverse problems: image reconstruction in CT and deconvolution, both through TV regularization. In particular, we present a method to learn the parameters of a convergent solver and demonstrate the improvement to the ad-hoc parameter choice. Moreover, empirical results indicate that by using additional parameters we can achieve improved performance.

The paper is organized as follows: In Section F.2 we recall elements of monotone operator theory and convex optimization, while setting up the notation. In Section F.3, we present and analyze a new solver for monotone inclusions, which we also specialize to convex optimization problems of the form (F.1.1). Section F.4 deals with the notion of "learning" an optimization solver, and in Section F.5 we present numerical experiments for variational regularization of inverse problems in imaging.

F.2 Background

Solving optimization problems of the type in (F.1.1) are often addressed using *variable splitting* techniques, which work well if the different terms are "simple" [9, 20, 23]. To keep the discussion as general as possible and since it does not add complexity to the proof of convergence, we will carry it out for monotone inclusions instead of convex optimization problems. The following subsections present necessary background material on monotone operators, convex optimization, and variable splitting.

Fundamental notions

Let $\mathcal{H}$ be a real Hilbert space with the inner product $\langle \cdot, \cdot \rangle$. We denote convergence in norm (or strong convergence) and weak convergence by $\to$ and $\rightharpoonup$, respectively. A set-valued operator $S \colon \mathcal{H} \rightrightarrows \mathcal{H}$ is *monotone* if

$$\langle z - z', w - w' \rangle \geq 0 \quad \text{for all } z, z' \in \mathcal{H},\ w \in S(z), \text{ and } w' \in S(z').$$

A monotone operator S is called *maximally monotone* if, in addition, the graph of S, defined by $\operatorname{graph}(S) \coloneqq \{(z, w) \in \mathcal{H} \times \mathcal{H} \mid w \in S(z)\}$, is not properly contained in the graph of any other monotone operator, i.e.,

$$(z, w) \in \operatorname{graph}(S) \iff \langle z - z', w - w' \rangle \geq 0 \text{ for all } (z', w') \in \operatorname{graph}(S).$$

A monotone operator is called *strongly monotone* if there exists a $\mu > 0$ such that

$$\langle z - z', w - w' \rangle \geq \mu \left\| z - z' \right\|^2 \quad \text{for all } z, z' \in \mathcal{H},\ w \in S(z), \text{ and } w' \in S(z').$$

Next, for any scalar $\sigma > 0$, the operator $J_S^\sigma = (\mathrm{Id} + \sigma S)^{-1}$ is called the *resolvent operator* or *proximal mapping* [45]. It can be shown that J_S^σ is a single-valued operator $\mathcal{H} \to \mathcal{H}$ [9, Proposition 23.8]. Note that an efficient routine to evaluate J_S^σ for all $\sigma > 0$ also enables to evaluate the resolvent operator of S^{-1} via

$$J_{S^{-1}}^\sigma(z) = z - \sigma J_S^{1/\sigma}(z/\sigma) \tag{F.2.1}$$

for $z \in \mathcal{H}$ (see [9, Proposition 23.20]).

A *maximally monotone inclusion problem* is defined as the problem of finding a point $z \in \mathcal{H}$ such that $0 \in S(z)$, which we henceforth denote $z \in \operatorname{zer}(S)$. In fact, it is easily seen that $z \in \operatorname{zer}(S)$ is equivalent with z being a fixed-point for the resolvent operator, i.e., $z = J_S^\sigma(z)$.

One reason for the interest in maximally monotone inclusion problems is that the *subdifferential* ∂F of a proper, convex and lower semicontinuous function $F \colon \mathcal{H} \to \overline{\mathbb{R}}$ is a maximally monotone operator [38]. Here, $\partial F : \mathcal{H} \rightrightarrows \mathcal{H}$ is defined to be

$$\partial F(x) \coloneqq \{y \in \mathcal{H} \mid \forall \tilde{x} \in \mathcal{H} \colon F(\tilde{x}) \geq F(x) + \langle y, \tilde{x} - x \rangle\}$$

if $F(x) \in \mathbb{R}$ and $\partial F(x) = \emptyset$ if $F(x) \in \{\pm\infty\}$. Moreover, the subdifferential at any minimizer of such a function contains zero, so F can be minimized by solving a maximally monotone inclusion problem [9, Theorem 16.3]. Note that we do not distinguish between local and global minimizers, since any local minimizer of a convex function is global [9, Proposition 11.4].

Remark F.2.1. A continuous linear operator $A\colon \mathcal{H} \to \mathcal{H}$ of a Hilbert space $\mathcal{H}$ into itself is maximally monotone if and only if it is accretive, i.e., if $\langle x, Ax\rangle \geq 0$ for all $x \in \mathcal{H}$ [9, Corollary 20.28, see also Definition 2.23], and it is the subdifferential ∂f of a function $f\colon \mathcal{H} \to \overline{\mathbb{R}}$ if and only if it is additionally symmetric [8, Proposition 2.51]. In particular the Volterra integral operator [10, Example 4.4]

$$(Af)(t) = \int_0^t f(s)\,\mathrm{d}s$$

and its inverse are maximally monotone, but not the subdifferential of a proper, convex and lower semicontinuous function.

For $F\colon \mathcal{H} \to \overline{\mathbb{R}}$, the Fenchel dual (convex conjugate) function $F^*\colon \mathcal{H} \to \overline{\mathbb{R}}$ is defined by [9, Chapter 13]

$$F^*(y) \coloneqq \sup_{x\in\mathcal{H}} \Big[\langle x, y\rangle - F(x)\Big] \quad \text{for } y \in \mathcal{H}.$$

If F is proper, convex and lower semicontinuous, then $\partial F^* = (\partial F)^{-1}$ [9, Corollary 16.30].

The *proximal point algorithm* is a fixed-point iterative scheme for solving the maximally monotone inclusion problem. It is given by repeatedly applying the resolvent operator:

$$z^{k+1} = J_S^\sigma\big(z^k\big).$$

It can now be shown that if $\operatorname{zer}(S) \neq \emptyset$ then z^k converges weakly to a point $z^\infty \in \operatorname{zer}(S)$ [45] for all starting points $z^0 \in \mathcal{H}$. The special case when $S \coloneqq \partial F$, i.e., the case of the resolvent of a subdifferential of F, is called the *proximal operator*. One can express the proximal as [38]

$$J_{\partial F}^\sigma(x) = \operatorname{Prox}_F^\sigma(x) = \operatorname*{arg\,min}_{x'\in\mathcal{H}} \left\{ F(x') + \frac{1}{2\sigma}\left\|x' - x\right\|^2 \right\}.$$

To see this, we simply note that if x' is a minimizing argument then

$$0 \in \partial F(x') + \frac{1}{\sigma}(x' - x) \quad \iff \quad x' = J_{\partial F}^\sigma(x).$$

It is thus interesting to note that the fixed-point iteration

$$x^{k+1} = \operatorname{Prox}_F^\sigma\left(x^k\right) = \operatorname*{arg\,min}_{x'\in\mathcal{H}} \left\{ F(x') + \frac{1}{2\sigma}\left\|x' - x^k\right\|^2 \right\}$$

generates a sequence $\left(x^k\right)$ that converges weakly to a minimizer of F. In this setting, the parameter σ can be interpreted as a step length. This can give rise to methods for solving the optimization problems if the proximal operator can be efficiently computed, e.g., through a closed-form expression. Note that (F.2.1) gives a method to obtain the proximal points of F^* from those of F, namely

$$\operatorname{Prox}_{F^*}^{\tau}(x) = x - \tau \operatorname{Prox}_{F}^{1/\tau}(x/\tau) \qquad \text{for all } \tau > 0.$$

Sometimes the resolvent of the maximally monotone operator S is not easy to evaluate, but S is of the form $S = A + B$ where A and B are maximally monotone and the resolvents of A and B can be evaluated efficiently. One may then consider approximating J_{A+B}^{σ} with J_A^{σ} and J_B^{σ} (splitting) [23]. An example when this arises is in convex minimization of an objective that is a sum of two (or more) functions $F+G$, like in (F.1.1). In these cases it is often not possible to compute a closed-form expression for the proximal operator $\operatorname{Prox}_{F+G}^{\sigma}$. Such problems can be addressed using operator splitting techniques that allow for solving the problem by only evaluating $\operatorname{Prox}_F^{\sigma}$ and $\operatorname{Prox}_G^{\sigma}$ [20].[1]

Convex optimization

Next, we will consider duality and optimality conditions for the problem (F.1.1). To simplify the notation, we consider the case $m = 1$ in (F.1.1), i.e., let $\mathcal{X}$ and $\mathcal{Y}$ be two Hilbert spaces and consider the model problem

$$\min_{x \in \mathcal{X}} \Big[F(x) + G(Lx) \Big], \tag{F.2.2}$$

where $L \colon \mathcal{X} \to \mathcal{Y}$ is a continuous linear operator and $F \colon \mathcal{X} \to \overline{\mathbb{R}}$ and $G : \mathcal{Y} \to \overline{\mathbb{R}}$ are proper, convex and lower semicontinuous functions. Note that (F.1.1) is recovered by setting

$$G(y) \coloneqq \sum_{i=1}^{m} G_i(y_i) \quad \text{for } y = (y_1, \ldots, y_m) \in \mathcal{Y} \coloneqq \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_m \tag{F.2.3}$$

and $Lx \coloneqq (L_1 x, \ldots, L_m x)$ for $x \in \mathcal{X}$ in (F.2.2).

The dual formulation of the primal problem (F.2.2) is

$$\max_{y \in \mathcal{Y}} \Big[-F^*(L^* y) - G^*(-y) \Big]. \tag{F.2.4}$$

Under suitable conditions the two optimization problems (F.2.2) and (F.2.4) have the same optimal value [9, Chapter 15.3]. Also note that, since both F and G are proper, convex and lower semicontinuous functions, $F^{**} = F$ and $G^{**} = G$ by

[1] In optimization, this operator splitting is sometimes referred to as *variable splitting*. The reason for this can be understood by comparing equations (F.2.2) and (F.2.7) below.

the Fenchel–Moreau theorem [9, Theorem 13.37]. Hence, the following primal-dual formulation

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \mathcal{L}(x;y) \quad \text{with} \quad \mathcal{L}(x;y) \coloneqq \langle Lx, y\rangle + F(x) - G^*(y) \tag{F.2.5}$$

(the mapping $\mathcal{L}(\cdot;\cdot)$ is called the *Lagrangian*) is equivalent to the primal problem.[2] In fact, under suitable assumptions it can be shown that if $(\bar{x},\bar{y})$ is a saddle point to (F.2.5), then $\bar{x}$ is a solution to the primal problem (F.2.2) and $\bar{y}$ is a solution to the dual problem (F.2.4) [9, Proposition 19.20].

A necessary optimality condition for the primal-dual formulation (F.2.5) is that the corresponding point $(\bar{x},\bar{y}) \in \mathcal{X}\times\mathcal{Y}$ be stationarity with respect to both variables, i.e., that

$$L\bar{x} \in \partial G^*(\bar{y}) \quad \text{and} \quad -L^*\bar{y} \in \partial F(\bar{x}). \tag{F.2.6}$$

For later use we note that the first of these conditions can be reformulated as

$$\begin{aligned} L\bar{x} \in \partial G^*(\bar{y}) \quad &\iff \quad \bar{y} + \sigma L\bar{x} \in \bar{y} + \sigma\partial G^*(\bar{y}) = (I + \sigma\partial G^*)(\bar{y}) \\ &\iff \quad \bar{y} = J^\sigma_{\partial G^*}(\bar{y} + \sigma L\bar{x}) = \operatorname{Prox}^\sigma_{G^*}(\bar{y} + \sigma L\bar{x}), \end{aligned}$$

and the second as

$$\begin{aligned} -L^*\bar{y} \in \partial F(\bar{x}) \quad &\iff \quad \bar{x} - \tau L^*\bar{y} \in \bar{x} + \tau\partial F(\bar{x}) = (I + \tau\partial F)(\bar{x}) \\ &\iff \quad \bar{x} = J^\tau_{\partial F}(\bar{x} - \tau L^*\bar{y}) = \operatorname{Prox}^\tau_F(\bar{x} - \tau L^*\bar{y}). \end{aligned}$$

Therefore, an equivalent condition to (F.2.6) is

$$\bar{y} = \operatorname{Prox}^\sigma_{G^*}(\bar{y} + \sigma L\bar{x}) \quad \text{and} \quad \bar{x} = \operatorname{Prox}^\tau_F(\bar{x} - \tau L^*\bar{y}). \tag{F.2.7}$$

Two splitting algorithms

As mentioned before, there are many different splitting methods available to solve problems of the form (F.1.1). For ease of reference, we here mention two popular choices. The first one, given in (F.2.8), is PDHG [19]

$$\begin{aligned} y_{n+1} &= \operatorname{Prox}^\sigma_{G^*}(y_n + \sigma L v_n), \\ x_{n+1} &= \operatorname{Prox}^\tau_F(x_n - \tau L^* y_{n+1}), \\ v_{n+1} &= x_{n+1} + \theta(x_{n+1} - x_n). \end{aligned} \tag{F.2.8}$$

The second one is the Douglas–Rachford type primal-dual algorithm [14], presented in (F.2.9)

$$\begin{aligned} p_n &= \operatorname{Prox}^\tau_F(x_n - \tau L^* y_n), \\ x_{n+1} &= x_n + \lambda_n(p_n - x_n), \\ q_n &= \operatorname{Prox}^\sigma_{G^*}(y_n + \sigma L(2p_n - x_n)), \\ y_{n+1} &= y_n + \lambda_n(q_n - y_n). \end{aligned} \tag{F.2.9}$$

[2] To see this, note that $\max_{y\in\mathcal{Y}}\left[\langle Lx, y\rangle - G^*(y)\right] = G^{**}(Lx) = G(Lx)$.

F.3 A new family of optimization solvers

In this section we introduce a new family of optimization algorithms and prove convergence for a subfamily. For ease of notation we will consider the simplified optimization problem (F.2.2), but results easily extend to the general case (F.1.1).

To this end, consider the two algorithms (F.2.8) and (F.2.9). Note that they can both be written as

$$q_n = \operatorname{Prox}^{\sigma}_{G^*}\left(b_{12}y_n + b_{11}L(c_{11}p_{n-1} + c_{12}x_{n-1})\right), \tag{F.3.1a}$$
$$y_{n+1} = a_{21}q_n + a_{22}y_n, \tag{F.3.1b}$$
$$p_n = \operatorname{Prox}^{\tau}_{F}\left(d_{12}x_n + d_{11}L^*(a_{11}q_n + a_{12}y_n)\right), \tag{F.3.1c}$$
$$x_{n+1} = c_{21}p_n + c_{22}x_n, \tag{F.3.1d}$$

for suitable values of the coefficients. More precisely, the PDHG algorithm (F.2.8) is obtained by setting

$$\begin{aligned} &a_{11} = 1 && a_{12} = 0 && a_{21} = 1 && a_{22} = 0 && b_{11} = \sigma && b_{12} = 1 \\ &c_{11} = 1+\theta && c_{12} = -\theta && c_{21} = 1 && c_{22} = 0 && d_{11} = -\tau && d_{12} = 1 \end{aligned}$$

and the Douglas-Rachford algorithm (F.2.9) by setting

$$\begin{aligned} &a_{11} = \lambda_n && a_{12} = 1-\lambda_n && a_{21} = \lambda_n && a_{22} = 1-\lambda_n && b_{11} = \sigma && b_{12} = 1 \\ &c_{11} = 2 && c_{12} = -1 && c_{21} = \lambda_n && c_{22} = 1-\lambda_n && d_{11} = -\tau && d_{12} = 1. \end{aligned}$$

We now go on to analyze the scheme (F.3.1). To state our results as generally as possible, we formulate them for a monotone inclusion problem that in particular specializes to the optimality conditions in (F.2.6) when the operators are subdifferentials. The monotone inclusion problem we seek to solve reads as follows: Let $\mathcal{X}$ and $\mathcal{Y}$ be two (not necessarily finite-dimensional) Hilbert spaces, and let $L\colon \mathcal{X} \to \mathcal{Y}$ be a continuous linear operator. Let $A : \mathcal{X} \rightrightarrows \mathcal{X}$ and $B : \mathcal{Y} \rightrightarrows \mathcal{Y}$ be maximally monotone operators. Find a pair $(\bar{x}, \bar{y}) \in \mathcal{X} \times \mathcal{Y}$ such that

$$L\bar{x} \in B^{-1}\bar{y} \quad \text{and} \quad -L^*\bar{y} \in A\bar{x}. \tag{F.3.2}$$

In this setting, the scheme (F.3.1) generalizes to

$$q_n = J^{\sigma}_{B^{-1}}(b_{12}y_n + b_{11}L(c_{11}p_{n-1} + c_{12}x_{n-1})), \tag{F.3.3a}$$
$$y_{n+1} = a_{21}q_n + a_{22}y_n, \tag{F.3.3b}$$
$$p_n = J^{\tau}_{A}(d_{12}x_n + d_{11}L^*(a_{11}q_n + a_{12}y_n)), \tag{F.3.3c}$$
$$x_{n+1} = c_{21}p_n + c_{22}x_n. \tag{F.3.3d}$$

We first note that if $a_{21} = 0$ or $c_{21} = 0$ the update for either y_{n+1} or x_{n+1} becomes trivial, and the algorithm will not be globally convergent to a point fulfilling (F.3.2) in general. Henceforth we will therefore assume that a_{21} and c_{21} are not equal to 0, unless the opposite is explicitly stated.

Fixed-point analysis

In this section, we give necessary and sufficient conditions for the solution set of (F.3.2) and the fixed point set of (F.3.3) to coincide for any choice of A, B, and L. To this end, let $(\bar{q}, \bar{y}, \bar{p}, \bar{x}) \in \mathcal{Y} \times \mathcal{Y} \times \mathcal{X} \times \mathcal{X}$ be a fixed point of the iterative scheme (F.3.3) and note that (F.3.3b) and (F.3.3d) gives

$$\bar{q} = \frac{1 - a_{22}}{a_{21}} \bar{y} \quad \text{and} \quad \bar{p} = \frac{1 - c_{22}}{c_{21}} \bar{x}.$$

Using this, we further get that

$$\begin{aligned} \frac{1 - a_{22}}{a_{21}} \bar{y} &= J^{\sigma}_{B^{-1}} \left(b_{12} \bar{y} + b_{11} L \left(c_{11} \frac{1 - c_{22}}{c_{21}} \bar{x} + c_{12} \bar{x} \right) \right) \\ \frac{1 - c_{22}}{c_{21}} \bar{x} &= J^{\tau}_{A} \left(d_{12} \bar{x} + d_{11} L^* \left(a_{11} \frac{1 - a_{22}}{a_{21}} \bar{y} + a_{12} \bar{y} \right) \right) \end{aligned}$$

The conditions in (F.3.2) can now be re-phrased as

$$\bar{y} = J^{\sigma}_{B^{-1}}(\bar{y} + \sigma L \bar{x}) \quad \text{and} \quad \bar{x} = J^{\tau}_{A}(\bar{x} - \tau L^* \bar{y}),$$

and combining the above two equations yields

$$\begin{aligned} a_{21} + a_{22} = 1, \quad b_{12} = 1, \quad b_{11}(c_{11} + c_{12}) = \sigma, \\ c_{21} + c_{22} = 1, \quad d_{12} = 1, \quad d_{11}(a_{11} + a_{12}) = -\tau. \end{aligned} \tag{F.3.4}$$

The conditions in (F.3.4) are necessary and sufficient, however, due to the linearity of L, the algorithm does not change if we agree to the normalization

$$\begin{aligned} b_{11} &= \sigma, & c_{11} + c_{12} &= 1, \\ d_{11} &= -\tau, & a_{11} + a_{12} &= 1. \end{aligned}$$

If we fix all these conditions, the iteration (F.3.3) takes the form

$$q_n = J^{\sigma}_{B^{-1}}(y_n + \sigma L(x_{n-1} + c_{11}(p_{n-1} - x_{n-1}))), \tag{F.3.5a}$$
$$y_{n+1} = y_n + a_{21}(q_n - y_n), \tag{F.3.5b}$$
$$p_n = J^{\tau}_{A}(x_n - \tau L^*(y_n + a_{11}(q_n - y_n))), \tag{F.3.5c}$$
$$x_{n+1} = x_n + c_{21}(p_n - x_n). \tag{F.3.5d}$$

Convergence analysis

The following theorem gives sufficient conditions for the weak convergence of the sequence (x_n, y_n) generated by (F.3.5) to a point that satisfies (F.3.2), i.e., a point that solves the monotone inclusion problem.

Theorem F.3.1. *Assume that there is a point that satisfies* (F.3.2)*, i.e., the monotone inclusion problem has a solution. Moreover, let*

$$a_{11} = a_{21} \quad \text{and} \quad c_{11} = 1 + \frac{c_{21}}{a_{21}}. \tag{F.3.6}$$

Assume furthermore that $0 < a_{21} < 2$, $0 < c_{21} < 2$ *and*

$$\sigma\tau \left\|L\right\|^2 < \frac{a_{21}^2(2-a_{21})(2-c_{21})}{\left(a_{21}+c_{21}-a_{21}c_{21}\right)^2} \quad \text{with } \sigma, \tau > 0. \tag{F.3.7}$$

Finally, let (q_n, y_n, p_n, x_n) *be the sequence generated by scheme* (F.3.5)*. Then the following holds:*

(a) $\sum_{n\geq 0} \left\|x_n - p_n\right\|^2 < +\infty$ *and* $\sum_{n\geq 0} \left\|x_n - x_{n+1}\right\|^2 < +\infty$.

(b) $\sum_{n\geq 0} \left\|y_n - q_n\right\|^2 < +\infty$ *and* $\sum_{n\geq 0} \left\|y_n - y_{n+1}\right\|^2 < +\infty$.

(c) The sequence $(x_n, y_n)_n$ *converges weakly to a point that satisfies* (F.3.2)*.*

(d) If A *is strongly monotone, then there is a unique* $\bar{x} \in \mathcal{X}$ *such that all solutions of* (F.3.2) *are of the form* $(\bar{x}, y)$ *with some* $y \in \mathcal{Y}$*. Moreover,* $\sum_{n=1}^{\infty} \left\|p_n - \bar{x}\right\|^2 < +\infty$*, in particular* $p_n \to \bar{x}$ *strongly.*
If B^{-1} *is strongly monotone, then there is a unique* $\bar{y} \in \mathcal{Y}$ *such that all solutions of* (F.3.2) *are of the form* $(x, \bar{y})$ *with some* $x \in \mathcal{X}$*. Moreover,* $\sum_{n=1}^{\infty} \left\|q_{n+1} - \bar{y}\right\|^2 < +\infty$*, in particular* $q_n \to \bar{y}$ *strongly.*

By rewriting with (F.3.6), the iteration (F.3.5) takes the following form:

Algorithm F.3.2. *Choose parameters* $\sigma > 0$, $\tau > 0$ *and* $a_{21} \in \mathbb{R}$, $c_{21} \in \mathbb{R}$ *and starting points* $x_0 \in \mathcal{X}$, $x_1 \in \mathcal{X}$, $p_0 \in \mathcal{X}$, $y_1 \in \mathcal{Y}$*. For all* $n = 1, 2, \ldots$*, calculate*

$$q_n = J^{\sigma}_{B^{-1}}\left(y_n + \sigma L\left(p_{n-1} + \frac{c_{21}}{a_{21}}(p_{n-1} - x_{n-1})\right)\right), \tag{F.3.8a}$$

$$y_{n+1} = y_n + a_{21}(q_n - y_n), \tag{F.3.8b}$$

$$p_n = J^{\tau}_{A}(x_n - \tau L^* y_{n+1}), \tag{F.3.8c}$$

$$x_{n+1} = x_n + c_{21}(p_n - x_n). \tag{F.3.8d}$$

Then, $x_n \rightharpoonup \bar{x}$, $p_n \rightharpoonup \bar{x}$, $y_n \rightharpoonup \bar{y}$, *and* $q_n \rightharpoonup \bar{y}$, *where* $(\bar{x}, \bar{y})$ *is a solution of* (F.3.2)*, provided that* $0 < a_{21} < 2$, $0 < c_{21} < 2$ *and* (F.3.7) *are satisfied.*

The remainder of the convergence analysis will therefore refer to scheme (F.3.8). The proof of Theorem F.3.1 rests upon a number of technical results and is given in the subsection below. An immediate corollary is the convergence of the primal-dual Douglas–Rachford method with constant relaxation [14].

Corollary F.3.3. *Let $\sigma\tau \|L\| < 1$ and $0 < \lambda < 2$. Then, for the iteration*

$$\begin{aligned} q_n &= J^{\sigma}_{B^{-1}}(y_n + \sigma L(2p_{n-1} - x_{n-1})), \\ y_{n+1} &= y_n + \lambda(q_n - y_n), \\ p_n &= J^{\tau}_{A}(x_n - \tau L^* y_{n+1}), \\ x_{n+1} &= x_n + \lambda(p_n - x_n), \end{aligned}$$

the sequence $(x_n, y_n)_n$ converges weakly to a point that satisfies (F.3.2).

Proof. Set $a_{21} = c_{21} = \lambda$ in Theorem F.3.1 and observe that (F.3.7) reduces to $\sigma\tau \|L\|^2 < 1$. □

Proof of Theorem F.3.1

For the proof, we define notions of distance Q_1 and Q_2 on the space $\mathcal{X} \times \mathcal{Y}$ of pairs of primal and dual variables (Lemma F.3.5). Next, we show that the distance (in terms of Q_1) between the iterates and the set of solutions of (F.3.2) decreases (Proposition F.3.6). This property is also known as *Fejér monotonicity* [9, Chapter 5]. Proposition F.3.7 improves the statement of Proposition F.3.6 for strongly monotone operators. The proof of Theorem F.3.1 is completed by showing that any weak sequential cluster point of the iteration sequence is a solution to (F.3.2).

We start with some simple inequalities between real numbers. In particular, Lemma F.3.4 (a) shows that we do not divide by zero in (F.3.7).

Lemma F.3.4. *Let $0 < a_{21} < 2$ and $0 < c_{21} < 2$. Then*

(a) $a_{21} + c_{21} > a_{21}c_{21}$ *and*

(b) $\dfrac{a_{21}c_{21}(2 - a_{21})(2 - c_{21})}{(a_{21} + c_{21} - a_{21}c_{21})^2} \leq 1.$

Proof. By assumption, $a_{21}(2 - a_{21}) > 0$, i.e., $a_{21} > \frac{1}{2}a_{21}^2$, and the same holds for c_{21}. Therefore,

$$a_{21} + c_{21} > \frac{1}{2}a_{21}^2 + \frac{1}{2}c_{21}^2 \geq a_{21}c_{21},$$

whence (a).

For (b), use the inequality $2a_{21}c_{21} \leq a_{21}^2 + c_{21}^2$ in

$$\begin{aligned} a_{21}c_{21}(2 - a_{21})(2 - c_{21}) &= 4a_{21}c_{21} - 2a_{21}^2c_{21} - 2a_{21}c_{21}^2 + a_{21}^2c_{21}^2 \\ &\leq a_{21}^2 + c_{21}^2 + 2a_{21}c_{21} - 2a_{21}^2c_{21} - 2a_{21}c_{21}^2 + a_{21}^2c_{21}^2 \\ &= (a_{21} + c_{21} - a_{21}c_{21})^2. \end{aligned}$$

□

Lemma F.3.5. *Define the quadratic forms $Q_1, Q_2 \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ by*

$$Q_1(x,y) = \frac{1}{2\tau c_{21}} \|x\|^2 + \frac{1}{2\sigma a_{21}} \|y\|^2 - \frac{1}{a_{21}} \langle y, Lx \rangle,$$
$$Q_2(x,y) = \frac{2-c_{21}}{2\tau} \|x\|^2 + \frac{2-a_{21}}{2\sigma} \|y\|^2 - \frac{a_{21}+c_{21}-a_{21}c_{21}}{a_{21}} \langle y, Lx \rangle$$

for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Under the assumptions in Theorem F.3.1, there exist $C_1, C_2, D_1, D_2 > 0$ such that

$$Q_i(x,y) \geq C_i \|x\|^2 \quad \text{and} \quad Q_i(x,y) \geq D_i \|y\|^2$$

for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $i = 1, 2$.

Proof. We can rewrite

$$Q_1(x,y) = \frac{1}{2\sigma a_{21}} \|y - \sigma Lx\|^2 + \frac{1}{2\tau c_{21}} \|x\|^2 - \frac{\sigma}{2a_{21}} \|Lx\|^2,$$
$$Q_1(x,y) = \frac{1}{2\tau c_{21}} \left\| x - \frac{c_{21}\tau}{a_{21}} L^* y \right\|^2 + \frac{1}{2\sigma a_{21}} \|y\|^2 - \frac{c_{21}\tau}{2a_{21}^2} \|L^* y\|^2$$

and

$$Q_2(x,y) = \frac{2-a_{21}}{2\sigma} \left\| y - \frac{\sigma(a_{21}+c_{21}-a_{21}c_{21})}{a_{21}(2-a_{21})} Lx \right\|^2 + \frac{2-c_{21}}{2\tau} \|x\|^2 - \frac{\sigma(a_{21}+c_{21}-a_{21}c_{21})^2}{2a_{21}^2(2-a_{21})} \|Lx\|^2,$$
$$Q_2(x,y) = \frac{2-c_{21}}{2\tau} \left\| x - \frac{\tau(a_{21}+c_{21}-a_{21}c_{21})}{a_{21}(2-c_{21})} L^* y \right\|^2 + \frac{2-a_{21}}{2\sigma} \|y\|^2 - \frac{\tau(a_{21}+c_{21}-a_{21}c_{21})^2}{2a_{21}^2(2-c_{21})} \|L^* y\|^2.$$

From this, the assertion of the lemma is clear with the quantities

$$C_1 = \frac{1}{2\tau c_{21}} - \frac{\sigma}{2a_{21}} \|L\|^2 = \frac{a_{21} - c_{21}\sigma\tau \|L\|^2}{2\tau a_{21} c_{21}},$$
$$D_1 = \frac{1}{2\sigma a_{21}} - \frac{c_{21}\tau}{2a_{21}^2} \|L\|^2 = \frac{a_{21} - c_{21}\sigma\tau \|L\|^2}{2\sigma a_{21}^2},$$
$$C_2 = \frac{2-c_{21}}{2\tau} - \frac{\sigma(a_{21}+c_{21}-a_{21}c_{21})^2}{2a_{21}^2(2-a_{21})} \|L\|^2 = \frac{a_{21}^2(2-a_{21})(2-c_{21}) - (a_{21}+c_{21}-a_{21}c_{21})^2 \sigma\tau \|L\|^2}{2\tau a_{21}^2(2-a_{21})},$$

$$D_2 = \frac{2-a_{21}}{2\sigma} - \frac{\tau(a_{21}+c_{21}-a_{21}c_{21})^2}{2a_{21}^2(2-c_{21})}\|L\|^2$$
$$= \frac{a_{21}^2(2-a_{21})(2-c_{21}) - (a_{21}+c_{21}-a_{21}c_{21})^2\sigma\tau\|L\|^2}{2\sigma a_{21}^2(2-c_{21})}$$

provided that the numerators are positive, i.e.,

$$\sigma\tau\|L\|^2 < \min\left\{\frac{a_{21}}{c_{21}}, \frac{a_{21}^2(2-a_{21})(2-c_{21})}{(a_{21}+c_{21}-a_{21}c_{21})^2}\right\}.$$

Now, by Lemma F.3.4, the minimum is always attained by the second value, and positivity is guaranteed by (F.3.7). □

Proposition F.3.6. *Define Q_1 and Q_2 as in Lemma F.3.5, let $(\bar{x},\bar{y}) \in \mathcal{X}\times\mathcal{Y}$ satisfy* (F.3.2)*, and let the sequence (q_n, y_n, p_n, x_n) be generated by scheme* (F.3.8)*. Under the assumptions in Theorem F.3.1, we have for all $n \geq 1$*

$$Q_1(x_{n+1}-\bar{x}, y_{n+2}-\bar{y}) - Q_1(x_n-\bar{x}, y_{n+1}-\bar{y}) \leq -Q_2(p_n-x_n, q_{n+1}-y_{n+1}).$$

Proof. Let $(\bar{x},\bar{y})$ satisfy (F.3.2). Then

$$\begin{aligned}
&Q_1(x_{n+1}-\bar{x}, y_{n+2}-\bar{y}) - Q_1(x_n-\bar{x}, y_{n+1}-\bar{y})\\
&= \frac{1}{2\tau c_{21}}\Big(\|x_{n+1}-\bar{x}\|^2 - \|x_n-\bar{x}\|^2\Big)\\
&\quad + \frac{1}{2\sigma a_{21}}\Big(\|y_{n+2}-\bar{y}\|^2 - \|y_{n+1}-\bar{y}\|^2\Big)\\
&\quad + \frac{1}{a_{21}}(\langle y_{n+1}-\bar{y}, Lx_n - L\bar{x}\rangle - \langle y_{n+2}-\bar{y}, Lx_{n+1}-L\bar{x}\rangle)\\
&= \frac{1}{2\tau c_{21}}\Big(\|x_n-\bar{x}+c_{21}(p_n-x_n)\|^2 - \|x_n-\bar{x}\|^2\Big)\\
&\quad + \frac{1}{2\sigma a_{21}}\Big(\|y_{n+1}-\bar{y}+a_{21}(q_{n+1}-y_{n+1})\|^2 - \|y_{n+1}-\bar{y}\|^2\Big)\\
&\quad + \frac{1}{a_{21}}\Big(\langle y_{n+1}-\bar{y}, Lx_n-L\bar{x}\rangle\\
&\quad - \langle y_{n+1}-\bar{y}+a_{21}(q_{n+1}-y_{n+1}), Lx_n-L\bar{x}+c_{21}(Lp_n-Lx_n)\rangle\Big)\\
&= \frac{c_{21}}{2\tau}\|p_n-x_n\|^2 + \frac{1}{\tau}\langle x_n-\bar{x}, p_n-x_n\rangle + \frac{a_{21}}{2\sigma}\|q_{n+1}-y_{n+1}\|^2\\
&\quad + \frac{1}{\sigma}\langle y_{n+1}-\bar{y}, q_{n+1}-y_{n+1}\rangle + \frac{c_{21}}{a_{21}}\langle \bar{y}-y_{n+1}, Lp_n-Lx_n\rangle\\
&\quad + \langle q_{n+1}-y_{n+1}, L\bar{x}-Lx_n\rangle + c_{21}\langle y_{n+1}-q_{n+1}, Lp_n-Lx_n\rangle. \qquad \text{(F.3.9)}
\end{aligned}$$

To estimate the above, we use the monotonicity of the operator B^{-1} together with the inclusions $L\bar{x} \in B^{-1}\bar{y}$ from (F.3.2) and

$$\frac{y_{n+1}-q_{n+1}}{\sigma} + Lp_n + \frac{c_{21}}{a_{21}}(Lp_n - Lx_n) \in B^{-1}q_{n+1}, \qquad \text{(F.3.10)}$$

which is a reformulation of (F.3.8a) with n replaced by $n+1$. This yields the inequality

$$\begin{aligned} 0 &\leq \left\langle \frac{y_{n+1}-q_{n+1}}{\sigma} + Lp_n + \frac{c_{21}}{a_{21}}(Lp_n - Lx_n) - L\bar{x}, q_{n+1} - \bar{y} \right\rangle \\ &= \frac{1}{\sigma}\langle y_{n+1} - q_{n+1}, q_{n+1} - \bar{y}\rangle + \langle Lp_n - L\bar{x}, q_{n+1} - \bar{y}\rangle \\ &\quad + \frac{c_{21}}{a_{21}}\langle Lp_n - Lx_n, q_{n+1} - \bar{y}\rangle \end{aligned} \tag{F.3.11}$$

Analogously, we can rewrite (F.3.8c) as

$$\frac{x_n - p_n}{\tau} - L^* y_{n+1} \in Ap_n. \tag{F.3.12}$$

The monotonicity of A together with the inclusion $-L^*\bar{y} \in A\bar{x}$ from (F.3.2) now yields

$$\begin{aligned} 0 &\leq \left\langle \frac{x_n - p_n}{\tau} - L^* y_{n+1} + L^*\bar{y}, p_n - \bar{x} \right\rangle \\ &= \frac{1}{\tau}\langle x_n - p_n, p_n - \bar{x}\rangle + \langle \bar{y} - y_{n+1}, Lp_n - L\bar{x}\rangle. \end{aligned} \tag{F.3.13}$$

Adding (F.3.11) and (F.3.13) yields

$$\begin{aligned} 0 \leq \frac{1}{\sigma}\langle y_{n+1} - q_{n+1}, q_{n+1} - \bar{y}\rangle + \langle Lp_n - L\bar{x}, q_{n+1} - y_{n+1}\rangle \\ + \frac{c_{21}}{a_{21}}\langle Lp_n - Lx_n, q_{n+1} - \bar{y}\rangle + \frac{1}{\tau}\langle x_n - p_n, p_n - \bar{x}\rangle, \end{aligned} \tag{F.3.14}$$

which, combined with (F.3.9), gives

$$\begin{aligned} &Q_1(x_{n+1} - \bar{x}, y_{n+2} - \bar{y}) - Q_1(x_n - \bar{x}, y_{n+1} - \bar{y}) \\ &\leq \frac{c_{21}}{2\tau}\|p_n - x_n\|^2 + \frac{1}{\tau}\langle x_n - \bar{x}, p_n - x_n\rangle + \frac{a_{21}}{2\sigma}\|q_{n+1} - y_{n+1}\|^2 \\ &\quad + \frac{1}{\sigma}\langle y_{n+1} - \bar{y}, q_{n+1} - y_{n+1}\rangle + \frac{c_{21}}{a_{21}}\langle \bar{y} - y_{n+1}, Lp_n - Lx_n\rangle \\ &\quad + \langle q_{n+1} - y_{n+1}, L\bar{x} - Lx_n\rangle + c_{21}\langle y_{n+1} - q_{n+1}, Lp_n - Lx_n\rangle \\ &\quad + \frac{1}{\sigma}\langle y_{n+1} - q_{n+1}, q_{n+1} - \bar{y}\rangle + \langle Lp_n - L\bar{x}, q_{n+1} - y_{n+1}\rangle \\ &\quad + \frac{c_{21}}{a_{21}}\langle Lp_n - Lx_n, q_{n+1} - \bar{y}\rangle + \frac{1}{\tau}\langle x_n - p_n, p_n - \bar{x}\rangle \\ &= \left(\frac{c_{21}}{2\tau} - \frac{1}{\tau}\right)\|p_n - x_n\|^2 + \left(\frac{a_{21}}{2\sigma} - \frac{1}{\sigma}\right)\|q_{n+1} - y_{n+1}\|^2 \\ &\quad + \left(\frac{c_{21}}{a_{21}} + 1 - c_{21}\right)\langle q_{n+1} - y_{n+1}, Lp_n - Lx_n\rangle \\ &= -Q_2(p_n - x_n, q_{n+1} - y_{n+1}). \end{aligned} \tag{F.3.15}$$

This concludes the proof. □

Proposition F.3.7. *Let Q_1 and Q_2 be defined as in Lemma F.3.5 and assume the conditions stated in Theorem F.3.1 hold.*

1. *If A is μ_1-strongly monotone for some $\mu_1 > 0$, then*

$$Q_1(x_{n+1} - \bar{x}, y_{n+2} - \bar{y}) - Q_1(x_n - \bar{x}, y_{n+1} - \bar{y}) + \mu_1 \|p_n - \bar{x}\|^2 \\ \leq -Q_2(p_n - x_n, q_{n+1} - y_{n+1}).$$

2. *If B^{-1} is μ_2-strongly monotone for some $\mu_2 > 0$, then*

$$Q_1(x_{n+1} - \bar{x}, y_{n+2} - \bar{y}) - Q_1(x_n - \bar{x}, y_{n+1} - \bar{y}) + \mu_2 \|q_{n+1} - \bar{y}\|^2 \\ \leq -Q_2(p_n - x_n, q_{n+1} - y_{n+1}).$$

Proof. If A is μ_1-strongly monotone, we obtain from (F.3.12) and $-L^*\bar{y} \in A\bar{x}$ (F.3.2) the estimation

$$\mu_1 \|\bar{x} - p_n\|^2 \leq \left\langle \frac{x_n - p_n}{\tau} - L^* y_{n+1} + L^* \bar{y}, p_n - \bar{x} \right\rangle,$$

which is a sharpened version of (F.3.13). By modifying (F.3.14) and (F.3.15) accordingly, we get the assumption. The case of a strongly monotone B^{-1} is analogously shown by improving (F.3.11). □

Having stated and proved the necessary estimations, we are now ready to prove Theorem F.3.1.

Proof of Theorem F.3.1. Let $(\bar{x}, \bar{y})$ satisfy (F.3.2). By Proposition F.3.6, we get the estimation

$$Q_1(x_{n+1} - \bar{x}, y_{n+2} - \bar{y}) - Q_1(x_n - \bar{x}, y_{n+1} - \bar{y}) \leq -Q_2(p_n - x_n, q_{n+1} - y_{n+1}).$$

Considering Lemma F.3.5, we see that the real sequence

$$(Q_1(x_n - \bar{x}, y_{n+1} - \bar{y}))_{n \geq 1}$$

is monotonically nonincreasing and therefore has a limit for each primal-dual solution $(\bar{x}, \bar{y})$. Furthermore, for all $N \geq 1$,

$$Q_1(x_N - \bar{x}, y_{N+1} - \bar{y}) - Q_1(x_0 - \bar{x}, y_1 - \bar{y}) \\ \leq -\sum_{n=0}^{N-1} Q_2(p_n - x_n, q_{n+1} - y_{n+1}).$$

By Lemma F.3.5, we have $Q_1(x_N - \bar{x}, y_{N+1} - \bar{y}) \geq 0$ and

$$\begin{aligned} Q_1(x_0 - \bar{x}, y_1 - \bar{y}) &\geq \sum_{n=0}^{N-1} Q_2(p_n - x_n, q_{n+1} - y_{n+1}) \\ &\geq \sum_{n=0}^{N-1} C_2 \left\| p_n - x_n \right\|^2 \end{aligned}$$

as well as

$$Q_1(x_0 - \bar{x}, y_1 - \bar{y}) \geq \sum_{n=0}^{N-1} D_2 \left\| q_{n+1} - y_{n+1} \right\|^2 .$$

Since this holds for arbitrary $N \geq 1$, this proves parts (a) and (b) of the theorem.

On the other hand, we have

$$Q_1(x_0 - \bar{x}, y_1 - \bar{y}) \geq Q_1(x_N - \bar{x}, y_{N+1} - \bar{y}) \geq C_1 \left\| x_N - \bar{x} \right\|^2$$

and

$$Q_1(x_0 - \bar{x}, y_1 - \bar{y}) \geq Q_1(x_N - \bar{x}, y_{N+1} - \bar{y}) \geq D_1 \left\| y_{N+1} - \bar{y} \right\|^2$$

for all $N \geq 1$, so the sequences $(x_n)_n$ and (y_n) are bounded in $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $(n_k)_k$ be a subsequence with $x_{n_k} \rightharpoonup x_\infty \in \mathcal{X}$ and $y_{n_k+1} \rightharpoonup y_\infty \in \mathcal{Y}$. By (F.3.12) and (F.3.10), we obtain

$$\begin{aligned} \frac{x_{n_k} - p_{n_k}}{\tau} - L^* y_{n_k+1} &\in A p_{n_k}, \\ \frac{y_{n_k+1} - q_{n_k+1}}{\sigma} + L p_{n_k} + \frac{c_{21}}{a_{21}} (L p_{n_k} - L x_{n_k}) &\in B^{-1} q_{n_k+1}. \end{aligned}$$

Now apply [6, Proposition 2.4] with

$$\begin{aligned} a_k &= p_{n_k}, \\ a_k^* &= \frac{x_{n_k} - p_{n_k}}{\tau} - L^* y_{n_k+1}, \\ b_k &= \frac{y_{n_k+1} - q_{n_k+1}}{\sigma} + L p_{n_k} + \frac{c_{21}}{a_{21}} (L p_{n_k} - L x_{n_k}), \\ b_k^* &= q_{n_k+1} \end{aligned}$$

and observe that

$$\begin{aligned} a_k &= x_{n_k} + (p_{n_k} - x_{n_k}) \rightharpoonup x_\infty, \\ b_k^* &= y_{n_k+1} + (q_{n_k+1} - y_{n_k+1}) \rightharpoonup y_\infty, \\ a_k^* + L^* b_k^* &= \frac{x_{n_k} - p_{n_k}}{\tau} + L^* (q_{n_k+1} - y_{n_k+1}) \to 0, \\ L a_k - b_k &= -\frac{y_{n_k+1} - q_{n_k+1}}{\sigma} - \frac{c_{21}}{a_{21}} (L p_{n_k} - L x_{n_k}) \to 0 \end{aligned}$$

because parts (a) and (b) imply that $x_{n_k} - p_{n_k} \to 0$ and $y_{n_k+1} - q_{n_k+1} \to 0$ as $k \to +\infty$. This gives $Lx_\infty \in B^{-1}y_\infty$ and $-L^*y_\infty \in Ax_\infty$, i.e., (x_∞, y_∞) satisfies (F.3.2). Since the choice of the weakly convergent subsequence was arbitrary, each weak sequential cluster point satisfies (F.3.2). Claim (c) now follows from [9, Lemma 2.47] applied to the norm $\sqrt{Q_1(\cdot)}$ on the product space $\mathcal{X} \times \mathcal{Y}$ and to the solution set of (F.3.2).

Now assume that A is μ_1-strongly monotone for some $\mu_1 > 0$. By Proposition F.3.7, we get the estimation

$$Q_1(x_{n+1} - \bar{x}, y_{n+2} - \bar{y}) - Q_1(x_n - \bar{x}, y_{n+1} - \bar{y}) + \mu_1 \|p_n - \bar{x}\|^2 \\ \leq -Q_2(p_n - x_n, q_{n+1} - y_{n+1})$$

for all $n \geq 0$. Choose $N \geq 1$ and sum up this inequality for $n = 0, \ldots, N-1$ to obtain

$$Q_1(x_N - \bar{x}, y_{N+1} - \bar{y}) - Q_1(x_0 - \bar{x}, y_1 - \bar{y}) + \mu_1 \sum_{n=0}^{N-1} \|p_n - \bar{x}\|^2 \\ \leq -\sum_{n=0}^{N-1} Q_2(p_n - x_n, q_{n+1} - y_{n+1})$$

Since the terms $Q_1(x_N - \bar{x}, y_{N+1} - \bar{y})$ and $\sum_{n=0}^{N-1} Q_2(p_n - x_n, q_{n+1} - y_{n+1})$ are non-negative by Lemma F.3.5, we obtain

$$\mu_1 \sum_{n=0}^{N-1} \|p_n - \bar{x}\|^2 \leq Q_1(x_0 - \bar{x}, y_1 - \bar{y}).$$

Analogously, one gets

$$\mu_2 \sum_{n=0}^{N-1} \|q_{n+1} - \bar{y}\|^2 \leq Q_1(x_0 - \bar{x}, y_1 - \bar{y}),$$

if B^{-1} is μ_2-strongly monotone, and since N is arbitrary, both sums

$$\sum_{n=0}^{\infty} \|p_n - \bar{x}\|^2 \qquad \text{and} \sum_{n=0}^{\infty} \|q_{n+1} - \bar{y}\|^2$$

are finite in the respective cases. The uniqueness of the point $\bar{x}$ under the assumption of strong monotonicity of A holds by the fact that we have shown $p_n \to \bar{x}$ for *any* solution $(\bar{x}, \bar{y})$ of (F.3.2). An analogous argument for $\bar{y}$ concludes the proof of Claim (d). □

Remark F.3.8. We were not able to show the weak convergence of PDHG (F.2.8) for $\theta \neq 1$ with this proof method. Indeed, by a straightforward calculation it can be shown that from Fejér monotonicity with respect to any quadratic form of the sequence $(x_n, y_{n+1})_n$ the conditions (F.3.6) can be derived, which implies $\theta = 1$.

Application to convex optimization

In this section, we specialize the scheme (F.3.8) to the case where the monotone operators A and B are subdifferentials ∂F and ∂G of proper, convex and lower semicontinuous functions $F : \mathcal{X} \to \overline{\mathbb{R}}$ and $G : \mathcal{Y} \to \overline{\mathbb{R}}$, respectively. Algorithm F.3.2 then reads as follows:

Algorithm F.3.9. *Choose parameters $\sigma > 0$, $\tau > 0$ and $a_{21} \in \mathbb{R}$, $c_{21} \in \mathbb{R}$ and starting points $x_0 \in \mathcal{X}$, $x_1 \in \mathcal{X}$, $p_0 \in \mathcal{X}$, $y_1 \in \mathcal{Y}$. For all $n = 1, 2, \ldots$, calculate*

$$q_n = \operatorname{Prox}^{\sigma}_{G^*}\left(y_n + \sigma L\left(p_{n-1} + \frac{c_{21}}{a_{21}}(p_{n-1} - x_{n-1})\right)\right), \tag{F.3.16a}$$

$$y_{n+1} = y_n + a_{21}(q_n - y_n), \tag{F.3.16b}$$

$$p_n = \operatorname{Prox}^{\tau}_{F}(x_n - \tau L^* y_{n+1}), \tag{F.3.16c}$$

$$x_{n+1} = x_n + c_{21}(p_n - x_n). \tag{F.3.16d}$$

Then, $x_n \rightharpoonup \bar{x}$, $p_n \rightharpoonup \bar{x}$, $y_n \rightharpoonup \bar{y}$, and $q_n \rightharpoonup \bar{y}$, where $(\bar{x}, \bar{y})$ is a solution of (F.2.6)*, provided that $0 < a_{21} < 2$, $0 < c_{21} < 2$, and* (F.3.7) *are satisfied.*

In this case, it is possible to get estimations for the Lagrangian, which is defined in (F.2.5).

Theorem F.3.10. *Given the assumptions in Theorem F.3.1, let $F : \mathcal{X} \to \overline{\mathbb{R}}$ and $G : \mathcal{Y} \to \overline{\mathbb{R}}$ be two proper, convex and lower semicontinuous functions. Let $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ be arbitrary. Then the sequence $(q_n, y_n, p_n x_n)$ generated by* (F.3.16) *satisfies*

$$\min_{n=0,\ldots,N-1}\left(\mathcal{L}(p_n; y) - \mathcal{L}(x; q_{n+1})\right) \leq \frac{1}{N} Q_1(x_0 - x, y_1 - y),$$

$$\mathcal{L}\left(\frac{1}{N}\sum_{n=0}^{N-1} p_n; y\right) - \mathcal{L}\left(x; \frac{1}{N}\sum_{n=0}^{N-1} q_{n+1}\right) \leq \frac{1}{N} Q_1(x_0 - x, y_1 - y).$$

This theorem is proved using the following proposition, which bounds the Lagrangian in terms of the quadratic forms defined in Lemma F.3.5.

Proposition F.3.11. *Given the assumptions in Theorem F.3.1, let $F : \mathcal{X} \to \overline{\mathbb{R}}$ and $G : \mathcal{Y} \to \overline{\mathbb{R}}$ be two proper, convex and lower semicontinuous functions. Let $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ be arbitrary. Then the sequence (q_n, y_n, p_n, x_n) generated by* (F.3.16) *satisfies*

$$\begin{aligned}\mathcal{L}(p_n; y) - \mathcal{L}(x; q_{n+1}) \leq{}& Q_1(x_n - x, y_{n+1} - y) - Q_1(x_{n+1} - x, y_{n+2} - y) \\ &- Q_2(p_n - x_n, q_{n+1} - y_{n+1})\end{aligned}$$

for all $n \geq 1$, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

Proof. Since $B^{-1} = \partial G^*$ and $A = \partial F$, the inclusions (F.3.10) and (F.3.12) provide certain subgradients, which imply the inequalities

$$\begin{aligned} G^*(y) &\geq G^*(q_{n+1}) + \frac{1}{\sigma}\langle y_{n+1} - q_{n+1}, y - q_{n+1}\rangle + \langle Lp_n, y - q_{n+1}\rangle \\ &\quad + \frac{c_{21}}{a_{21}}\langle Lp_n - Lx_n, y - q_{n+1}\rangle, \\ F(x) &\geq F(p_n) + \frac{1}{\tau}\langle x_n - p_n, x - p_n\rangle - \langle L^* y_{n+1}, x - p_n\rangle. \end{aligned}$$

Therefore, we have

$$\begin{aligned} &\mathcal{L}(p_n; y) - \mathcal{L}(x; q_{n+1}) \\ &\quad = \langle Lp_n, y\rangle + F(p_n) - G^*(y) - \langle Lx, q_{n+1}\rangle - F(x) + G^*(q_{n+1}) \\ &\quad \leq \frac{1}{\tau}\langle x_n - p_n, p_n - x\rangle + \langle Lp_n - Lx, q_{n+1} - y_{n+1}\rangle \\ &\qquad + \frac{1}{\sigma}\langle y_{n+1} - q_{n+1}, q_{n+1} - y\rangle + \frac{c_{21}}{a_{21}}\langle Lp_n - Lx_n, q_{n+1} - y\rangle. \end{aligned}$$

The right-hand side is now (except for the replacement of $\bar{x}$ and $\bar{y}$ by x and y, respectively) equal to the one in (F.3.14), and one easily checks by an analogous calculation, that it equals the expression in the assertion. □

Proof of Theorem F.3.10. By summing the inequality in Proposition F.3.11 for $n = 0, \ldots, N-1$ and dividing by N for some $N \geq 1$, we get

$$\frac{1}{N}\sum_{n=0}^{N-1}\left(\mathcal{L}(p_n; y) - \mathcal{L}(x; q_{n+1})\right) \leq \frac{1}{N} Q_1(x_0 - x, y_1 - y)$$

for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, where we dropped nonpositive terms on the right-hand side.

We have two possibilities to further estimate the left-hand side: First, we notice that it is the arithmetic mean of numbers, which is always greater than the minimum, i.e.,

$$\frac{1}{N}\sum_{n=0}^{N-1}\left(\mathcal{L}(p_n; y) - \mathcal{L}(x; q_{n+1})\right) \geq \min_{n=0,\ldots,N-1}\left(\mathcal{L}(p_n; y) - \mathcal{L}(x; q_{n+1})\right).$$

On the other hand, the Lagrangian is convex in its first and concave in its second component, so

$$\frac{1}{N}\sum_{n=0}^{N-1}\left(\mathcal{L}(p_n; y) - \mathcal{L}(x; q_{n+1})\right) \geq \mathcal{L}\left(\frac{1}{N}\sum_{n=0}^{N-1} p_n; y\right) - \mathcal{L}\left(x; \frac{1}{N}\sum_{n=0}^{N-1} q_{n+1}\right).$$

□

F.4 Learning an optimization solver

Most optimization problems are solved using iterative methods, akin to the ones presented in Sections F.2 and F.3. However, the number of iterations it takes in order for the algorithm to converge is in general hard to predict, which creates problems in time-critical applications. In these situations one could instead consider only doing a predefined fixed number n of iterations. A natural question that arises in response to this is: *what parameter values in the optimization solver give the best improvement of the objective function in n iterations*? This question leads to a meta-optimization over optimization solvers. Moreover, in general we are not only interested in optimizing one single cost function, but rather a (potentially infinite) family $\{F_\theta\}_{\theta\in\Theta}$ of cost functions, each with a minimizer $\bar{x}_\theta$. Hence, to make the question precise one needs to specify which family of optimization solvers one is considering, which is the family of cost functions of interested, and what is meant with "best improvement".

One such question was raised in [22], where the authors consider the worst-case performance $\sup_{\theta\in\Theta}\big[F_\theta(x_n) - F_\theta(\bar{x}_\theta)\big]$ of gradient-based algorithms over the set of continuously differentiable functions with Lipschitz-continuous gradients, and with a uniform upper bound on the Lipschitz constants. Subsequent work along the same lines can be found in [30, 47].

The idea of optimizing over optimization solvers has also been considered from a machine learning perspective. This has for example been done using *reinforcement learning* [34], and using *unsupervised learning* [26, 7]. In the latter category, one looks for algorithm parameters which minimize the expected value of the difference in objective function value,

$$\mathbb{E}_\theta\left[F_\theta(x_n) - F_\theta(\bar{x}_\theta)\right] = \mathbb{E}_\theta\left[F_\theta(x_n)\right] - \mathbb{E}_\theta\left[F_\theta(\bar{x}_\theta)\right] \tag{F.4.1}$$

where Θ is endowed with a probability measure and x_n is the output of the algorithm after n iterations. However, optimizing (F.4.1) with respect to the parameters of the method is independent of the optimal points $\{\bar{x}_\theta\}_{\theta\in\Theta}$, thus, this translates into unsupervised learning, i.e., the cost function $\mathbb{E}_\theta\left[F_\theta(x_n)\right]$ does not depend on $\bar{x}_\theta$. In this setting, [7] restricts attention to an architecture that operates individually on each coordinate of x. This is done in order to limit the number of parameters in the algorithm, which otherwise would grow exponentially with the dimension of x. To overcome this, we use an approach similar to [26], where the network architecture is inspired by modern first-order optimization solvers for nonsmooth problems, as presented in Sections F.2 and F.3. Similar ideas have also recently been explored for supervised learning in inverse problems in [49, 3, 4, 36, 44, 5, 27].

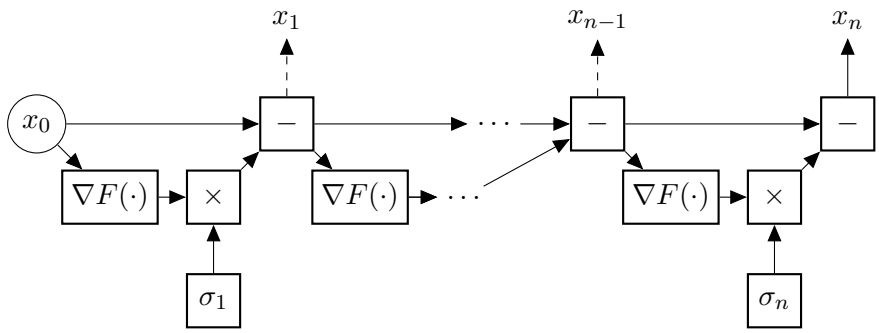


Figure F.1: Gradient descent.

Unrolled gradient descent as a neural network

Before we define the architecture considered in this work, we first present an illustrative example. To this end, consider the optimization problem

$$\min_x \quad F(x).$$

We assume that F is smooth, which means that the problem can be solved using a standard gradient descent algorithm, i.e., by performing the updates

$$x_k = x_{k-1} - \sigma_k \nabla F(x_k).$$

The gradient descent algorithm contains a set of parameters that need to be selected, namely the step length for each iteration, σ_k. This is normally done via the Goldstein rule or backtracking line search (Armijo rule) [12], which under suitable conditions ensures convergence to the optimal point $\bar{x}$.

However, if we only run the algorithm for a fixed number n of steps, the gradient descent algorithm can be seen as a *feedforward neural network*, as shown in Figure F.1. Each layer in the network performs the computation $x_{k-1} - \sigma_k \nabla F(x_{k-1})$ and the parameters of the network are $[\sigma_1, \ldots, \sigma_n]$. Moreover, if the step length is fixed to be the same in all iterations, i.e., $\sigma_1 = \ldots = \sigma_n = \sigma$ for some σ, the gradient descent algorithm can in fact be interpreted as a *recurrent neural network*. In both cases, for a given family $\{F_\theta\}_{\theta \in \Theta}$ of cost functions the network parameter(s) can be trained (optimized) by minimizing $\mathbb{E}_\theta \left[F_\theta(x_n)\right]$, where x_n is the output of the network in Figure F.1. For simple cases this can be done analytically.

Example F.4.1. Consider the family $(F_b)_b$ of functions $F_b : \mathbb{R}^n \to \mathbb{R}$ given by $F_b(x) = \frac{1}{2} x^\top A x - b^\top x$, where $A \in \mathbb{R}^{n \times n}$ is a (fixed) symmetric and positive definite matrix. The minimum of F_b is given by $\bar{x}_b = A^{-1} b$. Denote by Λ_σ the result of taking a gradient step of length $\sigma > 0$, i.e.,

$$\Lambda_\sigma(x) = x - \sigma \nabla F_b(x) = x - \sigma(Ax - b), \qquad x \in \mathbb{R}^n.$$

Let $x_0 \in \mathbb{R}^n$ be an arbitrary starting point of the iteration. This gives

$$\begin{aligned} F_b(\Lambda_\sigma(x_0)) &= F_b(x_0 - \sigma(Ax_0 - b)) \\ &= \frac{1}{2}(x_0 - \sigma(Ax_0 - b))^\top A(x_0 - \sigma(Ax_0 - b)) - b^\top(x_0 - \sigma(Ax_0 - b)) \\ &= \frac{\sigma^2}{2}(Ax_0 - b)^\top A(Ax_0 - b) - \sigma \|Ax_0 - b\|^2 + F_b(x_0). \end{aligned}$$

Let b be a random variable distributed according to $\mathsf{b} \sim \mathcal{P}$ for some probability distribution $\mathcal{P}$ with finite first and second moments. Finding a σ that minimizes the expectation

$$\begin{aligned} \mathbb{E}_{\mathsf{b}\sim\mathcal{P}}\Big[F_{\mathsf{b}}(\Lambda_\sigma(x_0))\Big] = \frac{\sigma^2}{2}\,\mathbb{E}_{\mathsf{b}\sim\mathcal{P}}\Big[(Ax_0 - \mathsf{b})^\top A(Ax_0 - \mathsf{b})\Big] \\ - \sigma\,\mathbb{E}_{\mathsf{b}\sim\mathcal{P}}\Big[\|Ax_0 - \mathsf{b}\|^2\Big] + \mathbb{E}_{\mathsf{b}\sim\mathcal{P}}\Big[F_{\mathsf{b}}(x_0)\Big], \end{aligned}$$

is a quadratic problem in one variable, and the optimal value of σ is thus

$$\begin{aligned} \sigma &= \frac{\mathbb{E}_{\mathsf{b}\sim\mathcal{P}}\Big[\|Ax_0 - \mathsf{b}\|^2\Big]}{\mathbb{E}_{\mathsf{b}\sim\mathcal{P}}\Big[(Ax_0 - \mathsf{b})^\top A(Ax_0 - \mathsf{b})\Big]} \\ &= \frac{\|Ax_0\|^2 - 2(Ax_0)^\top\,\mathbb{E}_{\mathsf{b}\sim\mathcal{P}}\big[\mathsf{b}\big] + \mathbb{E}_{\mathsf{b}\sim\mathcal{P}}\Big[\|\mathsf{b}\|^2\Big]}{x_0^\top A^3 x_0 - 2(A^2x_0)^\top\,\mathbb{E}_{\mathsf{b}\sim\mathcal{P}}\big[\mathsf{b}\big] + \mathbb{E}_{\mathsf{b}\sim\mathcal{P}}\Big[\mathsf{b}^\top A\mathsf{b}\Big]}. \end{aligned}$$

In some particular cases this expression can be simplified. For example if $A = I$, then $\sigma = 1$ as expected. Or if $x_0 = 0$, then $\sigma = \mathbb{E}_{\mathsf{b}\sim\mathcal{P}}[\|\mathsf{b}\|^2] / \mathbb{E}_{\mathsf{b}\sim\mathcal{P}}[\mathsf{b}^\top A\mathsf{b}]$.

Parametrizing a family of optimization algorithms

Similarly to the considerations for the unrolled gradient descent scheme above, for a fixed number of iterations one can consider the optimization algorithms (F.2.8), (F.2.9) and (F.3.16) as neural networks, where the variables we want to train are the parameters of the optimization methods. Optimizing these parameters with respect to the constraints corresponding to each algorithm is effectively trying to find optimal parameters for the corresponding algorithm for a given family of cost functions. However, if one only intends to do a finite number of iterations one could also remove this constraint, and thereby enlarge the space of schemes one is optimizing over.

As noted in Section F.3, all of the above mentioned optimization algorithms can be written on the form (F.3.1). That means that optimizing over the parameters in (F.3.1) can be seen as optimizing over a space of schemes that includes all three algorithms. Now, introducing the intermediate states $w_n = a_{11}q_n + a_{12}y_n$ and

$v_{n+1} = c_{11}p_n + c_{12}x_n$, and the 2×2 matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$, the scheme (F.3.1) can be written as

$$\begin{aligned} \begin{bmatrix} w_n \\ y_{n+1} \end{bmatrix} &= (\boldsymbol{A} \otimes \mathrm{Id}) \ \operatorname{diag}(\operatorname{Prox}_{G^*}^{\sigma}, \mathrm{Id}) \ (\boldsymbol{B} \otimes \mathrm{Id}) \ \begin{bmatrix} Lv_n \\ y_n \end{bmatrix} \\ \begin{bmatrix} v_{n+1} \\ x_{n+1} \end{bmatrix} &= (\boldsymbol{C} \otimes \mathrm{Id}) \ \operatorname{diag}(\operatorname{Prox}_{F}^{\tau}, \mathrm{Id}) \ (\boldsymbol{D} \otimes \mathrm{Id}) \ \begin{bmatrix} L^* w_n \\ x_n \end{bmatrix}, \end{aligned} \tag{F.4.2}$$

where the parameters of the scheme are the elements of the matrices. Here, by $\otimes$ we denote the Kronecker product, and by $\operatorname{diag}(A, B, \ldots)$ we denote the diagonal operator with the operators $A, B, \ldots$ on the diagonal. Connecting this with the previous optimization algorithms, the PDHG algorithm (F.2.8) is obtained by setting

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} \sigma & 1 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{C} = \begin{bmatrix} 1+\theta & -\theta \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} -\tau & 1 \\ 0 & 1 \end{bmatrix},$$

the primal-dual Douglas-Rachford algorithm (F.2.9) by taking

$$\boldsymbol{A} = \begin{bmatrix} \lambda_n & 1-\lambda_n \\ \lambda_n & 1-\lambda_n \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} \sigma & 1 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{C} = \begin{bmatrix} 2 & -1 \\ \lambda_n & 1-\lambda_n \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} -\tau & 1 \\ 0 & 1 \end{bmatrix},$$

and the proposed algorithm from Section F.3 by setting

$$\boldsymbol{A} = \begin{bmatrix} a_{21} & 1-a_{21} \\ a_{21} & 1-a_{21} \end{bmatrix}, \ \boldsymbol{B} = \begin{bmatrix} \sigma & 1 \\ 0 & 1 \end{bmatrix}, \ \boldsymbol{C} = \begin{bmatrix} 1+\frac{c_{21}}{a_{21}} & -\frac{c_{21}}{a_{21}} \\ c_{21} & 1-c_{21} \end{bmatrix}, \ \boldsymbol{D} = \begin{bmatrix} -\tau & 1 \\ 0 & 1 \end{bmatrix}.$$

Considering (F.4.2) as a neural network, the structure can easily be extended in order to incorporate more memory in the network. In this work we assume that the computationally expensive part of the algorithm is the evaluation of the operator L and its adjoint, which is typically the case in inverse problems in imaging, e.g., in three-dimensional CT [39, 40]. Therefore, the extension presented here thus keeps one evaluation L and one evaluation of L^* in each iteration.

To this end, let N be the number of primal variables $x^1, \ldots, x^N \in \mathcal{X}$ and M be the number of dual variables $y^1, \ldots, y^M \in \mathcal{Y}$. Introducing the four sequences of matrices $\boldsymbol{A}_n, \boldsymbol{B}_n \in \mathbb{R}^{M \times M}$ and $\boldsymbol{C}_n, \boldsymbol{D}_n \in \mathbb{R}^{N \times N}$, the iterations in (F.4.2) can be extended to yield the following algorithm.

Algorithm F.4.2. *Choose parameters $\boldsymbol{A}_n, \boldsymbol{B}_n \in \mathbb{R}^{M \times M}$ and $\boldsymbol{C}_n, \boldsymbol{D}_n \in \mathbb{R}^{N \times N}$, stepsizes $\sigma, \tau > 0$, and starting points $x_0^1, \ldots, x_0^N \in \mathcal{X}$, $y_0^2, \ldots, y_0^M \in \mathcal{Y}$. For all*

$n = 1, 2, \ldots,$ *calculate*

$$\begin{bmatrix} y_{n+1}^1 \\ y_{n+1}^2 \\ \vdots \\ y_{n+1}^M \end{bmatrix} = (\boldsymbol{A}_n \otimes \mathrm{Id}) \ \mathrm{diag}(\mathrm{Prox}_{G^*}^{\sigma}, \mathrm{Id}^{M-1}) \ (\boldsymbol{B}_n \otimes \mathrm{Id}) \begin{bmatrix} Lx_n^1 \\ y_n^2 \\ \vdots \\ y_n^M \end{bmatrix},$$
$$\begin{bmatrix} x_{n+1}^1 \\ x_{n+1}^2 \\ \vdots \\ x_{n+1}^N \end{bmatrix} = (\boldsymbol{C}_n \otimes \mathrm{Id}) \ \mathrm{diag}(\mathrm{Prox}_{F}^{\tau}, \mathrm{Id}^{N-1}) \ (\boldsymbol{D}_n \otimes \mathrm{Id}) \begin{bmatrix} L^* y_{n+1}^1 \\ x_n^2 \\ \vdots \\ x_n^N \end{bmatrix}.$$

Remark F.4.3. For the more general formulation of (F.1.1), more specialized network architectures than the one resulting from the choice (F.2.3) are possible, which handle the dual spaces separately instead of using the same stepsize σ and matrices $\boldsymbol{A}_n$ and $\boldsymbol{B}_n$ for all of them. An alternative network in the spirit of, e.g., [13, Theorem 2], to solve (F.1.1) reads as follows.

Algorithm F.4.4. *Choose parameters* $\boldsymbol{A}_{n,i}, \boldsymbol{B}_{n,i} \in \mathbb{R}^{M \times M}$*, for* $i = 1, \ldots, m$*, and* $\boldsymbol{C}_n, \boldsymbol{D}_n \in \mathbb{R}^{N \times N}$*, stepsizes* $\sigma_1, \ldots, \sigma_m, \tau > 0$*, and starting points* $x_0^1, \ldots, x_0^N \in \mathcal{X}$*,* $y_{0,i}^2, \ldots, y_{0,i}^M \in \mathcal{Y}_i$*,* $i = 1, \ldots, m$*. For all* $n = 1, 2, \ldots,$ *calculate*

$$\begin{bmatrix} y_{n+1,i}^1 \\ y_{n+1,i}^2 \\ \vdots \\ y_{n+1,i}^M \end{bmatrix} = (\boldsymbol{A}_{n,i} \otimes \mathrm{Id}) \ \mathrm{diag}(\mathrm{Prox}_{G_i^*}^{\sigma_i}, \mathrm{Id}^{M-1}) \ (\boldsymbol{B}_{n,i} \otimes \mathrm{Id}) \begin{bmatrix} L_i x_n^1 \\ y_{n,i}^2 \\ \vdots \\ y_{n,i}^M \end{bmatrix}, \quad i = 1, \ldots, m,$$
$$\begin{bmatrix} x_{n+1}^1 \\ x_{n+1}^2 \\ \vdots \\ x_{n+1}^N \end{bmatrix} = (\boldsymbol{C}_n \otimes \mathrm{Id}) \ \mathrm{diag}(\mathrm{Prox}_{F}^{\tau}, \mathrm{Id}^{N-1}) \ (\boldsymbol{D}_n \otimes \mathrm{Id}) \begin{bmatrix} \sum_{i=1}^m L_i^* y_{n+1,i}^1 \\ x_n^2 \\ \vdots \\ x_n^N \end{bmatrix}. \tag{F.4.3}$$

Extension to forward-backward-forward methods

Some methods in the literature, so called forward-backward-forward methods, include an extra evaluation of the operator and its adjoint per iteration, see, e.g., [21, 14, 16]. However, since the evaluation of the linear operator is assumed to be the expensive part in our setting, if we allow for two iterations in our framework to complete one iteration in such a framework, our proposed algorithm contains, e.g., [16, Equation (3.1)]. Letting $\cdot$ denote an element that can take any value, one such set of matrices

is given by

$$\boldsymbol{A}_{2n} = \begin{bmatrix} 0 & 0 & 1 \\ \cdot & \cdot & \cdot \\ 1 & 0 & 0 \end{bmatrix}, \qquad \boldsymbol{B}_{2n} = \begin{bmatrix} \gamma_n & 0 & 1 \\ \cdot & \cdot & \cdot \\ 0 & 0 & 1 \end{bmatrix},$$
$$\boldsymbol{C}_{2n} = \begin{bmatrix} 1 & 0 & -1 \\ \cdot & \cdot & \cdot \\ 1 & 1 & 0 \end{bmatrix}, \qquad \boldsymbol{D}_{2n} = \begin{bmatrix} -\gamma_n & 0 & 1 \\ \gamma_n & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

for the even iterations and

$$\boldsymbol{A}_{2n+1} = \begin{bmatrix} 0 & 1 & 0 \\ \cdot & \cdot & \cdot \\ 0 & 0 & 1 \end{bmatrix}, \qquad \boldsymbol{B}_{2n+1} = \begin{bmatrix} \cdot & \cdot & \cdot \\ 0 & 0 & 1 \\ \gamma_n & 0 & 1 \end{bmatrix},$$
$$\boldsymbol{C}_{2n+1} = \begin{bmatrix} 0 & 0 & 1 \\ \cdot & \cdot & \cdot \\ 0 & 0 & 1 \end{bmatrix}, \qquad \boldsymbol{D}_{2n+1} = \begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ -\gamma_n & 0 & 1 \end{bmatrix},$$

for the odd iterations.

Remark F.4.5. Other forward-backward-forward methods have been proposed in the literature, some of which are general enough to include the PDHG as a special case [28], or both the PDHG and the Douglas-Rachford algorithm as special cases [33]. However, these methods include a step-length computation in their updates. This computation involves evaluating the norm of current iterates, which is not possible to achieve by only doing the linear operations we propose. Of course, allowing the matrix elements to be nonlinear functions of the states would allow us to incorporate also these methods, however, that is beyond the scope of this article.

F.5 Application to inverse problems and numerical experiments

As we briefly outline next, optimization problems of the type in (F.1.1) arise when solving ill-posed inverse problems by means of variational regularization.

The goal in an inverse problem is to recover parameters characterizing a system under investigation from indirect observations. This can be formalized as the task of estimating (reconstructing) model parameters, henceforth called signal, $f_{\text{true}} \in X$ from indirect observations (data) $g \in Y$ where

$$g = \mathcal{A}\left(f_{\text{true}}\right) + \delta g. \tag{F.5.1}$$

In the above, X and Y are typically Hilbert or Banach spaces, and $\mathcal{A}\colon X \to Y$ (forward operator) models how a given signal gives rise to data in absence of noise. Furthermore, $\delta g \in Y$ is a single sample of a Y-valued random element that represents the noise component of data.

A natural approach for solving (F.5.1) is to minimize a function $R\colon X \to \mathbb{R}$ (data discrepancy functional) that quantifies the miss-fit in data space. Since this function needs to incorporate the aforementioned forward operator $\mathcal{A}$ and the data g, it is often of the form

$$R(f) \coloneqq \mathcal{L}\left(\mathcal{A}\left(f\right), g\right) \quad \text{for some } \mathcal{L} : Y \times Y \to \mathbb{R}.$$

If $\mathcal{L}$ is the negative data log-likelihood, then minimizing $f \mapsto R(f)$ corresponds to finding a maximum likelihood solution to (F.5.1).

However, finding a minimizer to R is an ill-posed problem, meaning that a solution (if it exists) is discontinuous with respect to the data g. Variational regularization addresses this issue by introducing an additional function $S\colon X \to \overline{\mathbb{R}}$ (regularization functional) that encodes a priori information about f_{true} and penalizes undesirable solutions [25]. This results in an optimization problem

$$\min_{f \in X} \Big[\lambda R(f) + S(f) \Big],$$

which from a statistical perspective can be interpreted as trying to find a maximum a posteriori estimate [29]. A common choice of regularization functional, especially for inverse problems in imaging, is the total variation (TV) regularization $S(f) \coloneqq \|\nabla f\|_1$, but several more advanced regularizers have also been suggested in the literature, typically exploiting some kind of sparsity using an L_1-like norm [18].

In this section, we consider an inverse problem in computerized tomography. To this end, let $\mathcal{A}$ be the *Radon transform* and consider TV regularization. This means that we are interested in minimizing

$$H_b(x) = \|\mathcal{A}\left(x\right) - b\|_2^2 + \lambda \left\|\nabla x\right\|_1, \tag{F.5.2}$$

i.e., a family of objective functions that is parametrized by the data b. This means that we can apply the ideas from Section F.4 on learning an optimization solver.

Implementation and specifications of the training

We train and evaluate several of the algorithms described in this article on a clinically realistic data set, namely simulated data from human abdomen CT scans as provided by Mayo Clinic for the AAPM Low Dose CT Grand Challenge [37]. Examples of two-dimensional phantoms from this data set are given in Figure F.2. Throughout all examples, the size of the image x is 512×512 pixels, and the regularization parameter $\lambda > 0$ is fixed. The Radon transform $\mathcal{A}$ used in this example is sampled according to a fan-beam geometry [39] and the data is generated by applying $\mathcal{A}$ to the phantoms and then adding 5% white Gaussian noise. Examples of such data (sinograms) are also shown in Figure F.2.

Problem (F.5.2) is obtained from (F.1.1) by setting $F(x) \coloneqq 0$ for all x,

$$L_1 x \coloneqq \mathcal{A}\left(x\right), \qquad L_2 x \coloneqq \nabla x,$$
$$G_1(y_1) \coloneqq \|y_1 - b\|_2^2, \qquad G_2(y_2) \coloneqq \|y_2\|_1,$$

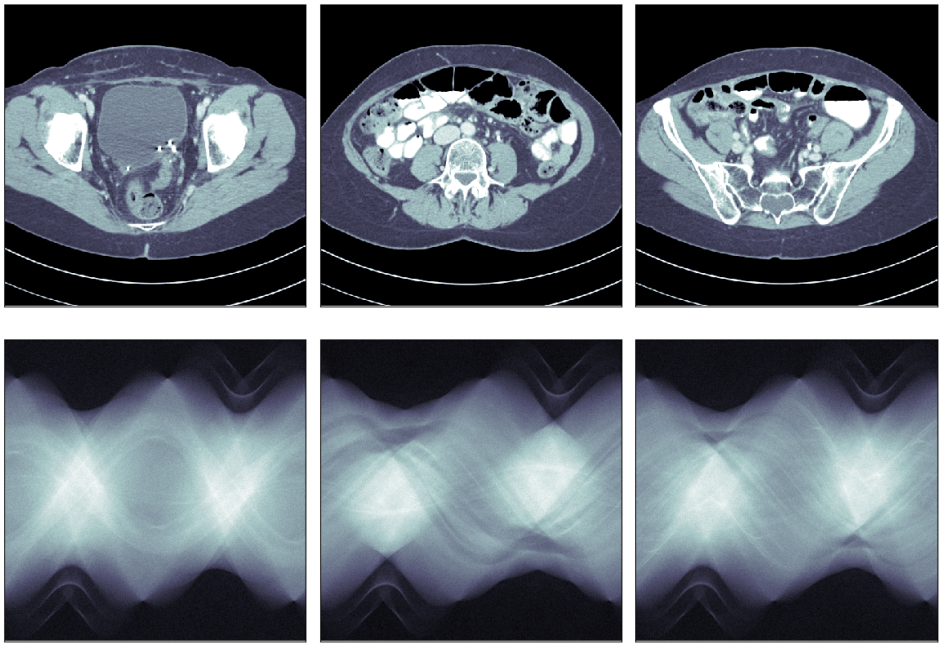


Figure F.2: The top row shows three examples of phantoms used for generating data. These phantoms take values between $[0.0, 3.25]$, but all images are shown using a window set to $[0.8, 1.2]$ in order to enhance contrast of clinically more relevant details. The lower row shows corresponding simulated, noisy sinograms.

and all the proximal operators are implemented in ODL [2]. If not stated otherwise, we use (F.2.3) to reduce (F.1.1) to (F.2.2).

For each algorithm, the number of unrolled iterations, corresponding to the depth of the network, was set to $n_{\max} = 10$, and all evaluations have been done with this depth. However, in order to heuristically induce better stability of the general schemes, we have trained using a stochastic depth as follows: In each step of the training, the depth of the network has been set to the outcome of the heavy-tailed random variable $\mathsf{n}_{\max} = \min\left[\texttt{round}(8 + \mathsf{Z}),\, 100\right]$, where Z is the exponential of a Gaussian random variable with standard deviation 1.25 and mean value $\log(2) - 1.25^2/2$, so that $\mathbb{E}[\mathsf{Z}] = 2$. The limitation to 100 iterations is due to limits in computational resources.

In order to improve stability and generalization properties of the trained networks, we have normalized the operators before training, i.e., rescaled them so that $\|\mathcal{A}\|_2 = \|\nabla\|_2 = 1$. For the same reasons, we have used the zero vector as initial guess for all networks. Training has been done using the *Adam* solver [31], with standard parameter values except for $\beta_2 = 0.99$. Moreover, we have used gradient clipping to

limit the norm of the gradient of the training cost function (F.4.1) to be less than or equal to one [43]. As step length (learning rate) we have used a cosine annealing scheme [35], i.e., a step length which in step t takes the value

$$\eta_t = \frac{\eta_0}{2}\left(1 + \cos\left(\pi \frac{t}{t_{rmmax}}\right)\right),$$

where the initial step length η_0 has been set to 10^{-3}. We have trained for $t_{\max} = 100\,000$ steps and have used 9 out of 10 phantoms from the AAPM Low Dose CT Grand Challenge for training and one for evaluation.

All algorithms have been implemented using ODL [2], the GPU accelerated version of ASTRA [42, 48], and Tensorflow [1]. The source code to replicate the experiments is available online, where the weights of the trained networks are also explicitly given.[3] We have used this setup to train the following methods.

PDHG method. This corresponds to optimal selection of the parameters θ, τ, and σ for the PDHG method (F.2.8) on the family of cost functions (F.5.2). In order to achieve this, we need to enforce the constraints $\theta \in [0, 1]$ and $\sigma\tau \|L\|^2 < 1$. This has been done implicitly by a change of variables, namely by

$$\theta = \frac{\mathrm{e}^{s_1}}{1 + \mathrm{e}^{s_1}}, \qquad \tau = \frac{1}{\|L\|} \cdot \frac{\mathrm{e}^{s_2 + s_3}}{1 + \mathrm{e}^{s_2}}, \qquad \sigma = \frac{1}{\|L\|} \cdot \frac{\mathrm{e}^{s_2 - s_3}}{1 + \mathrm{e}^{s_2}} \tag{F.5.3}$$

with $s_1, s_2, s_3 \in \mathbb{R}$. Here, s_2 determines how close the parameters σ and τ are to the constraint $\sigma\tau \|L\|^2 < 1$, while s_3 determines the trade-off between τ and σ.

PDHG method without constraints on the parameters. Here we train the same parameters θ, τ, σ as in the PDHG method. However, we do not make the change of variables (F.5.3), therefore, no constraints on θ, τ, and σ are enforced in the training. This means that the resulting scheme might not correspond to a globally convergent optimization algorithm.

Proposed method from Section F.3. This corresponds to optimal parameter selection for the method (F.3.16) on the family of cost functions (F.5.2). To adhere to the constraints in the assumptions in Theorem F.3.1, we have used the same kind of variable change as in (F.5.3), namely

$$a_{21} = \frac{2\mathrm{e}^{s_1}}{1 + \mathrm{e}^{s_1}}, \quad c_{21} = \frac{2\mathrm{e}^{s_2}}{1 + \mathrm{e}^{s_2}}, \quad \sigma = \frac{K}{\|L\|} \cdot \frac{\mathrm{e}^{s_3 - s_4}}{1 + \mathrm{e}^{s_3}}, \quad \tau = \frac{K}{\|L\|} \cdot \frac{\mathrm{e}^{s_3 + s_4}}{1 + \mathrm{e}^{s_3}}$$

with $s_1, \ldots, s_4 \in \mathbb{R}$, where $K = \frac{a_{21}^2(2 - a_{21})(2 - c_{21})}{(a_{21} + c_{21} - a_{21}c_{21})^2}$, as in (F.3.7).

[3] `https://github.com/aringh/data-driven_nonsmooth_optimization`

Table F.1: Loss function values for the CT reconstruction after 10 iterations. The values given are of the form $\frac{1}{100}\sum_{i=1}^{100} H_{b_i}(x_{10}) - H_{b_i}(x_i^*)$, i.e., the difference of the obtained objective function value and an estimate of the true minimum objective function value $H_{b_i}(x_i^*)$ corresponding to data b_i, averaged over 100 samples.

Method	Loss function values
PDHG with parameters from [46]	109.93
Trained PDHG with constraints on parameters	82.381
Trained solver (F.3.16)	24.183
Trained PDHG without constraints on parameters	27.761
Trained scheme of type (F.4.3) with $N = M = 2$	20.024
Trained scheme of type (F.4.3) with $N = M = 3$	**14.905**

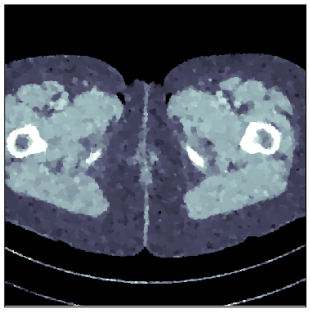
Parametrization proposed in Section F.4. Here, we have trained schemes of the form (F.4.3). We have done this for constant sequences of matrices $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{B}_1$, $\boldsymbol{B}_2$, $\boldsymbol{C}$, and $\boldsymbol{D}$. We restricted ourselves to the sizes $N = M = 2$ and $N = M = 3$.

Performance of the trained methods

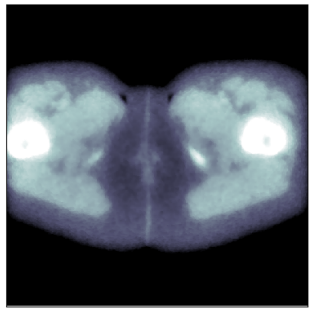
To obtain an estimation of the true optimal value of (F.5.2), we have run 1 000 iterations of PDHG with parameters as in [46]. In Table F.1 we show the difference between the obtained objective function value and the minimal objective function value, averaged over 100 samples. As can be seen, the scheme proposed in Section F.4 with $N = M = 3$ performs best at 10 iterations. Moreover, a general trend seems to be that more parameters in the algorithms improve the performance. Finally, the results from one specific phantom are presented as reconstructions in Figure F.3. Note that the reconstruction by PDHG with parameters as in [46] is left out due to the page layout.

Generalization to other iteration numbers

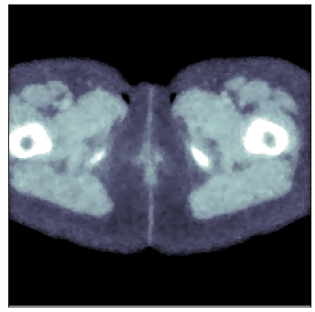
Figure F.4 shows the objective function value (F.5.2) as a function of the iteration number, i.e., how well the learned algorithms generalize to iteration numbers they are not trained for. For the trained, convergent solvers, the objective function value keeps decreasing as expected. Furthermore, the solver proposed in (F.3.16) performs better than the others also when the number of iterations are increased, but poorer in the beginning. For the other schemes, it can be noted that, while training more parameters seems to increase the performance after 10 iterations, it also seems to decrease the generalizability of the algorithm with respect to an increase in the number of iterations.



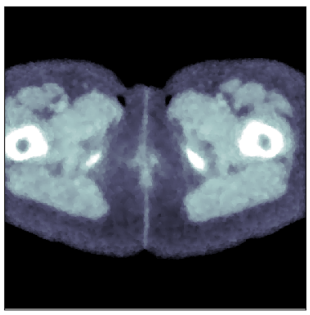
(a) TV reconstruction.



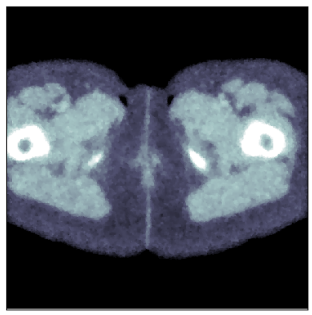
(b) Trained PDHG with constraints on parameters.



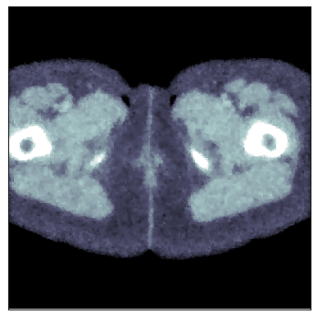
(c) Trained solver (F.3.16).



(d) Trained PDHG without constraints on parameters.



(e) Trained scheme of type (F.4.3) with $N = M = 2$.



(f) Trained scheme of type (F.4.3) with $N = M = 3$.

Figure F.3: Reconstruction with data from a phantom that was not used in the training. The TV reconstruction, to which they should be compared, is shown in F.3a. All reconstructions use 10 steps. The phantom takes values between $[0.0, 2.33]$, but all images are shown using a window set to $[0.8, 1.2]$ in order to enhance contrast of clinically more relevant details.

Generalization to deblurring

Next, we investigate the generalizability of the trained networks to other optimization problems by replacing the forward operator $\mathcal{A}$ in (F.5.2) with a convolution. This corresponds to another TV problem in imaging, namely image deblurring.

Clearly, the trained networks that correspond to optimization solvers with convergence guarantees can be applied to other convex optimization problems. (Note that we still normalize the operators to have operator norm one so that the assumptions in Theorem F.3.1 do not change.) However, nothing guarantees that parameters that give fast convergence on one type of problems will also give fast convergence on another one.

Two example images are shown in Figure F.5. The images in Figure F.5d–F.5f,

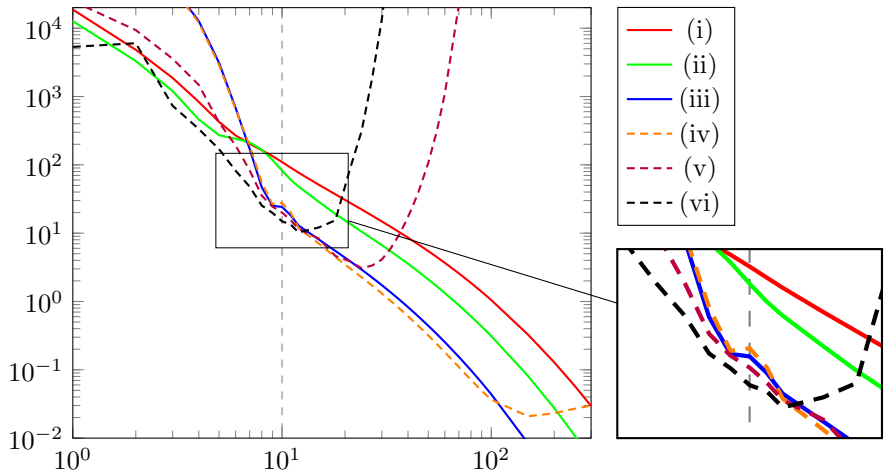


Figure F.4: The figure shows the values $\frac{1}{100}\sum_{i=1}^{100} H_{b_i}(x_n) - H_{b_i}(x_i^*)$, where $H_{b_i}(x_i^*)$ is an estimate of the true minimum objective function value corresponding to data b_i, of several reconstruction methods as a function of the iteration number n. Solid lines are real optimization solvers, dotted lines are schemes that might not converge to the true optimal solution. (i) PDHG with parameters as in [46], (ii) PDHG with trained parameters with constraints, (iii) proposed solver (F.3.16) with trained parameters, (iv) PDHG with trained free parameters, (v) proposed scheme (F.4.3) with $N = M = 2$, and (vi) proposed scheme (F.4.3) with $N = M = 3$.

corresponding to the "Raccoon" test image, are of size 1024×768 and use a different regularization parameter. Blurring has been done with Gaussian kernels. For the "Ascent" test image, the kernel has a standard deviation of approximately three pixels in each direction, whereas for the "Raccoon" test image, the kernel has a standard deviation of approximately four pixels in the up-down and six pixels in the left-right direction. As for the sinograms in the CT example, 5% white noise has been added to the blurred images. Again, to obtain an estimation of the true optimal value of we have run 1 000 iterations of PDHG with parameters as in [46]. For each algorithm, the difference between the obtained objective function value and minimal objective function value is presented in Table F.2, and the deblurred images are shown in Figures F.6 and F.7. Again, the reconstruction by PDHG with parameters as in [46] is left out due to the page layout.

The method with $N = M = 3$ does not generalize well. However, the method with $N = M = 2$ generalizes, and the optimization algorithm from Section F.3, with trained parameters, is one of the best on these two test problems.

(a) "Ascent" test image. (b) Blurred, noisy image. (c) TV reconstruction.

(d) "Raccoon" test image. (e) Blurred noisy image. (f) TV reconstruction.

Figure F.5: Example images used for the deblurring problem in Section F.5.

Table F.2: Loss function values for the deblurring problem in Section F.5. Here, $H_{b_i}(x_i^*)$ is an estimate of the true minimum objective function value corresponding to data b_i.

Method	$H_{b_i}(x_{10}) - H_{b_i}(x_i^*)$	
	Ascent	Raccoon
PDHG with parameters from [46]	5.514	11.475
Trained PDHG with constraints on parameters	4.256	8.5126
Trained solver (F.3.16)	**2.173**	4.5898
Trained PDHG without constraints on parameters	2.204	**4.4790**
Trained scheme of type (F.4.3) with $N = M = 2$	3.514	9.9139
Trained scheme of type (F.4.3) with $N = M = 3$	208.37	873.33

(a) TV reconstruction.

(b) Trained PDHG with constraints on parameters.

(c) Trained solver (F.3.16).

(d) Trained PDHG without constraints on parameters.

(e) Trained scheme of type (F.4.3) with $N = M = 2$.

(f) Trained scheme of type (F.4.3) with $N = M = 3$.

Figure F.6: Reconstructions with the trained algorithms for the "Ascent" image.

F.6 Conclusions and future work

In this work, we have first proposed a new solver for maximally monotone inclusion problems and proved convergence guarantees. In particular, we have also proposed a new convergent primal-dual proximal solver for convex optimization problems. Further, we have investigated new aspects of learning an optimization solver. This is particularly relevant in inverse problems where one can parametrize the objective function by data, leaving the other parts unchanged. This can, in fact, also be interpreted as learning a pseudo-inverse of the forward operator in an unsupervised fashion. Moreover, the framework admits enforcing convergence and stability properties in the learning. We should emphasize that this implies a form of generalizability to other data, and even other forward operators, since the scheme cannot diverge.

There are several different directions in which the work from this article can be extended: Regarding the optimization perspective, one could investigate whether (F.3.8) can be further relaxed to introduce more free parameters while retaining

(a) TV reconstruction.

(b) Trained PDHG with constraints on parameters.

(c) Trained solver (F.3.16).

(d) Trained PDHG without constraints on parameters.

(e) Trained scheme of type (F.4.3) with $N = M = 2$.

(f) Trained scheme of type (F.4.3) with $N = M = 3$.

Figure F.7: Reconstructions with the trained algorithms for the "Raccoon" image.

convergence, e.g. by relaxing (F.3.6) or letting parameters vary in each iteration.

Also from a machine learning perspective, there are aspects to be further investigated:

- Since accelerated first-order algorithms like FISTA [11] can be parametrized by (F.4.3), does the learning result in a scheme with $\mathcal{O}(1/n^2)$ convergence rate for the objective function values when trained for n iterations?

- Our numerical experiments suggest that training without "convergence constraints" gives the network more freedom and thereby improves accuracy. However, the resulting schemes seem to be unstable beyond the fixed number of iterates used for training. Is it true that, in general, convergence cannot be enforced by training alone?

- Is it possible to state and prove a time accuracy trade-off theorem, i.e., to estimate the error between the trained solver and the true solution to the optimization? If so, which properties of the underlying family of objective functions (training data) does this require?

Acknowledgments

The authors thank Dr. Cynthia McCollough, the Mayo Clinic, and the American Association of Physicists in Medicine for providing the data necessary for performing comparison using a human phantom.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] J. Adler, H. Kohr, and O. Öktem. Odl 0.6.0, April 2017.

[3] J. Adler and O. Öktem. Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Problems*, 33(12):124007, 2017.

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

[5] J. Adler, A. Ringh, O. Öktem, and J. Karlsson. Learning to solve inverse problems using wasserstein loss. *arXiv preprint arXiv:1710.10898*, 2017.

[6] A. Alotaibi, P.L. Combettes, and N. Shahzad. Solving coupled composite monotone inclusions by successive Fejér approximations of their Kuhn–Tucker set. *SIAM Journal on Optimization*, 24(4):2076–2095, 2014.

[7] 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.

[8] V. Barbu and T. Precupanu. *Convexity and optimization in Banach spaces*. Springer, Dordrecht, 4th edition, 2012.

[9] H.H. Bauschke and P.L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, New York, NY, 2nd edition, 2017.

[10] H.H. Bauschke, X. Wang, and L. Yao. Examples of discontinuous maximal monotone linear operators and the solution to a recent problem posed by b.f. svaiter. *Journal of Mathematical Analysis and Applications*, 370(1):224–241, 2010.

[11] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[12] D. Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.

[13] R.I. Boţ and E.R. Csetnek. On the convergence rate of a forward-backward type primal-dual splitting algorithm for convex optimization problems. *Optimization*, 64(1):5–23, 2015.

[14] R.I. Boţ and C. Hendrich. A Douglas–Rachford type primal-dual method for solving inclusions with mixtures of composite and parallel-sum type monotone operators. *SIAM Journal on Optimization*, 23(4):2541–2565, 2013.

[15] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, Jan. 2011.

[16] L.M. Briceño-Arias and P.L. Combettes. A monotone+skew splitting model for composite monotone inclusions in duality. *SIAM Journal on Optimization*, 21(4):1230–1250, 2011.

[17] 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.

[18] 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.

[19] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[20] P.L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

[21] P.L. Combettes and J.-C. Pesquet. Primal-dual splitting algorithm for solving inclusions with mixtures of composite, lipschitzian, and parallel-sum type monotone operators. *Set-Valued and variational analysis*, 20(2):307–330, 2012.

[22] Y. Drori and M. Teboulle. Performance of first-order methods for smooth convex minimization: A novel approach. *Mathematical Programming*, 145(1-2):451–482, 2014.

[23] J. Eckstein. *Splitting methods for monotone operators with applications to parallel optimization*. PhD thesis, 1989. Department of Civil Engineering, Massachusetts Institute of Technology.

[24] J. Eckstein and D.P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, 1992.

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

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

[27] K. Hammernik, T. Klatzer, E. Kobler, M.P. Recht, D.K. Sodickson, T. Pock, and F. Knoll. Learning a variational network for reconstruction of accelerated mri data. *Magnetic resonance in medicine*, 79(6):3055–3071, 2018.

[28] B. He and X. Yuan. Convergence analysis of primal-dual algorithms for a saddle-point problem: From contraction perspective. *SIAM Journal on Imaging Sciences*, 5(1):119–149, 2012.

[29] J. Kaipio and E. Somersalo. *Statistical and computational inverse problems*, volume 160 of *Applied Mathematical Sciences*. Springer, New York, 2005.

[30] D. Kim and J.A. Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical programming*, 159(1-2):81–107, 2016.

[31] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[32] S. Ko, D. Yu, and J.-H. Won. On a class of first-order primal-dual algorithms for composite convex minimization problems. *arXiv preprint arXiv:1702.06234*, 2017.

[33] P. Latafat and P. Patrinos. Asymmetric forward–backward–adjoint splitting for solving monotone inclusions involving three operators. *Computational Optimization and Applications*, 68(1):57–93, 2017.

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

[35] I. Loshchilov and F. Hutter. SGDR: Stochastic gradient descent with restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[36] M. Mardani, E. Gong, J.Y. Cheng, S. Vasanawala, G. Zaharchuk, M. Alley, N. Thakur, S. Han, W. Dally, J.M. Pauly, and L. Xing. Deep generative adversarial networks for compressed sensing automates mri. *arXiv preprint arXiv:1706.00051*, 2017.

[37] C. McCollough. TU-FG-207A-04: Overview of the low dose CT grand challenge. *Medical physics*, 43(6Part35):3759–3760, 2016.

[38] J.-J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

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

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

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

[42] W.J. Palenstijn, K.J. Batenburg, and J. Sijbers. Performance improvements for iterative electron tomography reconstruction using graphics processing units (GPUs). *Journal of structural biology*, 176(2):250–253, 2011.

[43] R. Pascanu, T. Mikolov, and Y. Bengio. Understanding the exploding gradient problem. *CoRR, abs/1211.5063*, 2012.

[44] P. Putzky and M. Welling. Recurrent inference machines for solving inverse problems. *arXiv preprint arXiv:1706.04008*, 2017.

[45] R.T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

[46] E.Y. Sidky, H.J. Jakob, and P. Xiaochuan. Convex optimization problem prototyping for image reconstruction in computed tomography with the Chambolle–Pock algorithm. *Physics in medicine and biology*, 57(10):3065, 2012.

[47] A.B. Taylor, J.M. Hendrickx, and F. Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1-2):307–345, 2017.

[48] W. van Aarle, W.J. Palenstijn, J. Cant, E. Janssens, F. Bleichrodt, A. Dabravolski, J. De Beenhouwer, K.J. Batenburg, and J. Sijbers. Fast and flexible x-ray tomography using the astra toolbox. *Optics express*, 24(22):25129–25147, 2016.

[49] Y. Yang, J. Sun, H. Li, and Z. Xu. Deep ADMM-Net for compressive sensing MRI. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 29, pages 10–18, 2016.