

CONTEMPORARY MATHEMATICS

313

Inverse Problems, Image Analysis, and Medical Imaging

AMS Special Session on
Interaction of Inverse Problems and Image Analysis
January 10–13, 2001
New Orleans, Louisiana

M. Zuhair Nashed
Otmar Scherzer
Editors



Inverse Problems, Image Analysis, and Medical Imaging

This page intentionally left blank

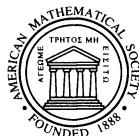
CONTEMPORARY MATHEMATICS

313

Inverse Problems, Image Analysis, and Medical Imaging

AMS Special Session on
Interaction of Inverse Problems and Image Analysis
January 10–13, 2001
New Orleans, Louisiana

M. Zuhair Nashed
Otmar Scherzer
Editors



American Mathematical Society
Providence, Rhode Island

Editorial Board

Dennis DeTurck, managing editor

Andreas Blass Andy R. Magid Michael Vogelius

2000 *Mathematics Subject Classification.* Primary 00Bxx, 45Q05, 65Jxx 65Rxx 65Kxx,
65T60, 34G20, 35A15, 41-XX, 42-XX.

Library of Congress Cataloging-in-Publication Data

AMS Special Session on Interaction of Inverse Problems and Image Analysis (2001 : New Orleans, La.)

Inverse problems, image analysis, and medical imaging : AMS Special Session on Interaction of Inverse Problems and Image Analysis, January 10–13, 2001, New Orleans, Louisiana / M. Zuhair Nashed, Otmar Scherzer, editors.

p. cm. — (Contemporary mathematics, ISSN 0271-4132 ; 313)

Includes bibliographical references.

ISBN 0-8218-2979-3 (acid-free paper)

1. Image processing—Digital techniques—Mathematical models—Congresses. 2. Image analysis—Mathematical models—Congresses. 3. Inverse problems (Differential equations)—Congresses. 4. Diagnostic imaging—Mathematical models—Congresses. I. Nashed, M. Zuhair. II. Scherzer, Otmar, 1964— III. Title. IV. Contemporary mathematics (American Mathematical Society) ; v. 313.

TA1637 .A47 2001

621.36'7—dc21

2002034298

Copying and reprinting. Material in this book may be reproduced by any means for educational and scientific purposes without fee or permission with the exception of reproduction by services that collect fees for delivery of documents and provided that the customary acknowledgment of the source is given. This consent does not extend to other kinds of copying for general distribution, for advertising or promotional purposes, or for resale. Requests for permission for commercial use of material should be addressed to the Acquisitions Department, American Mathematical Society, 201 Charles Street, Providence, Rhode Island 02904-2294, USA. Requests can also be made by e-mail to reprint-permission@ams.org.

Excluded from these provisions is material in articles for which the author holds copyright. In such cases, requests for permission to use or reprint should be addressed directly to the author(s). (Copyright ownership is indicated in the notice in the lower right-hand corner of the first page of each article.)

© 2002 by the American Mathematical Society. All rights reserved.

The American Mathematical Society retains all rights
except those granted to the United States Government.

Printed in the United States of America.

- ∞ The paper used in this book is acid-free and falls within the guidelines
established to ensure permanence and durability.
Visit the AMS home page at <http://www.ams.org/>

10 9 8 7 6 5 4 3 2 1 07 06 05 04 03 02

Contents

Preface	vii
List of Participants	ix
Regularization of nonlinear unstable operator equations by secant methods with application to gravitational sounding problem R. B. ALEXEEV AND A. B. SMIRNOVA	1
A fractal set constructed from a class of wavelet sets J. J. BENEDETTO AND S. SUMETKIJAKAN	19
Joint invariant signatures for curve recognition M. BOUTIN	37
Inpainting based on nonlinear transport and diffusion T. F. CHAN AND J. SHEN	53
Towards fast non-rigid registration U. CLARENZ, M. DROSKE, AND M. RUMPF	67
A note on wavelet-based inversion algorithms C. DE MOL AND M. DEFRISE	85
A comparison between the Wavelet-Galerkin and the Sinc-Galerkin methods in solving nonhomogeneous heat equations M. EL-GAMEL AND A. I. ZAYED	97
Fast diffusion registration B. FISCHER AND J. MODERSITZKI	117
Iterative stabilization and edge detection C. W. GROETSCH AND O. SCHERZER	129
Backprojections in tomography, spherical functions and addition formulas: a few challenges F. A. GRÜNBAUM	143
Mathematical models for 2D positron emission tomography B. A. MAIR AND J. A. ZAHNEN	153
Explicit versus implicit relative error regularization on the space of functions of bounded variation O. SCHERZER	171

Sampling methods for approximate solution of pde F. STENGER, A. R. NAGHSH-NILCHI, J. NIEBSCH, AND R. RAMLAU	199
Diffusion and regularization of vector- and matrix-valued images J. WEICKERT AND T. BROX	251
A numerically robust hybrid steepest descent method for the convexly constrained generalized inverse problems I. YAMADA, N. OGURA, AND N. SHIRAKAWA	269

Preface

This volume contains the refereed proceedings of the Special Session on Interaction of Inverse Problems and Image Analysis held at the annual meeting of the American Mathematical Society which took place in New Orleans, Louisiana, January 10-13, 2001.

The volume contains 15 papers, 14 of which are authored or coauthored by a participant at the Session. One paper is by an invited speaker who was not able to participate at the meeting.

Inverse Problems deal with determining for a given input-output system an input that produces an observed output, or of determining an input that produces a desired output. In terms of an operator T acting between say two normed spaces X and Y , the problem of solving the equation $T(x) = y$ for given data $y \in Y$ is a canonical example of an inverse problem. Typically inverse problems are *ill-posed*. Important examples of ill-posed inverse problems include integral equations of the first kind, tomography, and inverse scattering. *Signal Analysis/Processing* deals with digital representations of signals and their analog reconstructions from digital representations. *Image Analysis* deals with problems such as image recovery, enhancement, feature extraction, and motion detection. *Medical Imaging* is an important branch of *Image Science* and deals with image analysis in medical applications.

The common thread among the areas of Inverse Problems, Signal Analysis, and Image Analysis is a canonical problem of recovery of an object (function, signal, picture) from partial or indirect information about the object (often contaminated by noise). Both Inverse Problems and Imaging Science have emerged in recent years as interdisciplinary research fields with profound applications in many areas of Science, Engineering, Technology, and Medicine. Research in Inverse Problems and Image Processing has rich interactions with several areas of Mathematics, and strong links to Signal Processing, Variational Problems, Applied Harmonic Analysis and Computational Mathematics.

The goal of the Special Session on Interaction of Inverse Problems and Image Analysis was to gather a group of mathematicians and a few scientists and engineers from universities and research centers to report on recent research advances and to provide motivation for mathematicians interested in learning about the interaction of these two fields. For the latter goal, a couple of the invited 20-minute talks provided overview presentations in the two areas. This facilitated understanding of other invited talks and encouraged non-experts to attend the session. The spirit of some of the expository presentations is conveyed in several papers in this volume.

The volume contains carefully refereed and edited original research papers and a few high-level expository/survey papers to provide an overview and perspectives

on the interaction of inverse problems, image analysis, and medical imaging. In particular there are papers on the following topics : regularization of linear and nonlinear ill-posed problems, bounded variation regularization, fractal and wavelet sets, curve recognition, picture inpainting, image registration, diffusion and regularization of images, positron emission tomography, backprojection, robust hybrid steepest descent methods, sampling methods, Sinc-Galerkin and Wavelet-Galerkin methods for approximate solution of partial differential equations.

The session was highly successful and very well attended. This is the first time that a Special Session on this theme was organized at an AMS meeting. In addition to the invited speakers from universities and research laboratories in the USA, there were invited speakers from Austria, Belgium, France, Germany, and Japan.

We are grateful to the authors and coauthors of the papers in this volume for their contributions and to the referees for their meticulous reports that helped us to produce a high quality volume. The authors have graciously agreed to make the revisions recommended by the referees and the editors. We are grateful to Christine M. Thivierge, the Acquisitions Assistant for AMS, for her expert professional guidance (and also for her constructive patience). Finally, we thank the Program Committee of the American Mathematical Society for scheduling this Special Session and Prof. Dennis DeTurck and the Editorial Board of *Contemporary Mathematics* for supporting the publication of this volume. The final stage of editing this volume was completed during a visit of Otmar Scherzer to the University of Delaware; the visit was supported by grant Y-123INF of the Austrian Science Foundation (FWF).

Zuhair Nashed (University of Delaware and University of Central Florida)
Otmar Scherzer (University of Innsbruck)

List of Participants in the Special Session
Interactions of Inverse Problems and Image Analysis
New Orleans, Louisiana, January 10-13, 2001

- Akram Aldroubi
- John Benedetto
- Amin Boumenir
- Mireille Boutin
- Alfred Carasso
- Ingrid Daubechies
- Christine De Mol
- Bernd Fischer
- Charles Groetsch
- Bernard Mair
- Francois Malgouyres
- Mike Mayergoiz
- Jan Modersitzki
- Zuhair Nashed
- Stanley Osher
- Martin Rumpf
- Otmar Scherzer
- Jianhong Shen
- Alexandra Smirnova
- Frank Stenger
- Kevin Vixie
- Joachim Weickert
- David Wilson
- Isao Yamada
- Ozgar Yilmaz
- Ahmed Zayed

This page intentionally left blank

Regularization of nonlinear unstable operator equations by secant methods with application to gravitational sounding problem

Roman B. Alexeev and Alexandra B. Smirnova

ABSTRACT. A novel iteratively regularized secant-type algorithm with simultaneous updates of the operator $(F'^*(x_n)F'(x_n) + \varepsilon_n I)^{-1}$ is suggested for solving nonlinear ill-posed operator equations $F(x) = 0$, $H_1 \rightarrow H_2$, on a pair of Hilbert spaces H_1 and H_2 . A convergence theorem is proved. The stability of the process towards noise in the data is analyzed, and a stopping time is chosen so that the method converges as the noise level tends to zero. The proposed scheme is illustrated by a numerical example in which a nonlinear inverse problem of gravitational sounding is considered.

Introduction

The theme of our paper is solving nonlinear operator equations of the form:

$$(1.1) \quad F(x) = 0, \quad F : H_1 \rightarrow H_2,$$

on a pair of Hilbert spaces H_1 and H_2 . We assume here that equation (1.1) has a solution \hat{x} , not necessarily unique, and the operator F is twice Fréchet differentiable without such structural assumptions as monotonicity, invertibility of $F'(x)$ etc.

In order to avoid the ill-posed inversion of the Fréchet derivative operator $F'(x)$ various discrete and continuous methods based on a regularization are suggested. A principal point in the numerical implementation of regularized Newton's and Gauss-Newton's procedures is the computation of the operators $(F'(x) + \varepsilon I)^{-1}$ and $(F'^*(x)F'(x) + \varepsilon I)^{-1}$ respectively (see, for example, [3] or [1], [2]). This computation for certain operators requires a considerable effort in many applications. Besides it may decrease the accuracy of the approximate solution.

For finite dimensional well-posed problems, when the Jacobian $F'(\hat{x})$ is boundedly invertible, several approaches are taken in order to reduce the cost associated with the storage and inversion of $F'(x)$ in Quasi-Newton schemes. Probably the most used approach is a so called secant method: at every step of an iterative process the Jacobian $F'(x_n)$ is replaced with an approximation, obtained from $F(x_{n+1})$ and $F(x_n)$. In fact, multivariable generalizations of the secant method have also been proposed. Although they require some extra calculations of $F(x)$, they have

1991 *Mathematics Subject Classification.* 65J15, 58C15, 47H17.

Key words and phrases. nonlinear problem, regularization, Fréchet derivative, secant method.

r -order equal to the largest root of $r^{n+1} - r^n - 1 = 0$. However none of those generalizations seem to perform well in practice (see [7], p.168). For this reason the most popular secant methods are 'two-point' methods, and the best known among them is the algorithm introduced by C.Broyden [6]

$$(1.2) \quad x_{n+1} = x_n - J_n^{-1}F(x_n),$$

$$(1.3) \quad J_{n+1} = J_n + \frac{(s_n, \cdot)}{\|s_n\|^2} F(x_{n+1}),$$

as long as $F(x_n)$ and therefore $s_n := x_{n+1} - x_n$ do not vanish. Here x_0 is an initial guess, which is assumed to be sufficiently close to \hat{x} , and J_0 is some regular and sufficiently close initial approximation to $F'(\hat{x})$. Note that J_{n+1} in (1.2)-(1.3) satisfies the secant condition:

$$J_{n+1}s_n = F(x_{n+1}) - F(x_n) = \int_0^1 F'(x_n + ts_n) dt s_n \approx F'(x_n)s_n,$$

i.e. it approximates the Jacobian $F'(x_n)$ in the direction s_n . Broyden's method has a limited memory variant ([7], p. 188), which requires no explicit matrix storage for the approximate derivative. It is based on the recursion for the inverse

$$(1.4) \quad J_{n+1}^{-1} = J_n^{-1} - \frac{(s_n, J_n^{-1} \cdot)}{(s_n, J_n^{-1}(F(x_{n+1}) - F(x_n)))} J_n^{-1} F(x_{n+1}).$$

In form (1.2)-(1.3) Broyden's method can also be defined in infinite dimensional Hilbert spaces [8], [12] for continuously invertible $F'(\hat{x})$ and B_0 . In finite dimensional well-posed case, under standard assumptions, Broyden's method is superlinearly convergent. In infinite dimensional Hilbert space compactness of the operator $J_0 - F'(\hat{x})$ is additionally required for superlinear convergence.

In [10] the regularized version of Broyden's method for nonlinear ill-posed problems ($F'(\hat{x})$ is not boundedly invertible) is suggested with the regularization being done by mollifying the data and by stopping the iterative process at an appropriate index $n = N$. The convergence is analyzed under the following basic assumption:

$$(1.5) \quad F'(\hat{x}) = F'(x)R_x^{\hat{x}}, \quad \|R_x^{\hat{x}} - I\| \leq C_R\|\hat{x} - x\|, \quad \hat{x}, x \in U(x_0),$$

which means that the operator $F'(x)$ remains almost the same for all x up to a certain modification by $R_x^{\hat{x}}$.

Another approach to the construction of a regularized secant-type method was proposed in [11], where the regularization is based on the introducing of a dynamical system with the trajectory $x(t)$ starting at some initial point x_0 and converging to a solution of (1.1) as $t \rightarrow +\infty$:

$$(1.6) \quad \dot{x}(t) = -B(t)[F'^*(x(t))F(x(t)) + \varepsilon(t)(x(t) - x_0)],$$

$$(1.7) \quad \dot{B}(t) = -[(F'^*(x(t))F'(x(t)) + \varepsilon(t)I)B(t) - I],$$

$$x(0) = x_0 \in H, \quad B(0) \in L(H), \quad 0 < \varepsilon(t) \rightarrow 0 \quad \text{as } t \rightarrow +\infty.$$

In (1.6)-(1.7) the operator $(F'^*(x(t))F'(x(t)) + \varepsilon(t)I)^{-1}$ is updated continuously without actual inversion of the Fréchet derivative. It is shown in [11] that $x(t)$ converges to a solution of (1.1) at the rate $O(\varepsilon(t))$. An attractive feature of this result is the absence of assumption (1.5) as well as the absence of the assumptions about the location of the spectrum of the operator $F'(x)$. The absence of these

assumptions is made possible by a restriction on the initial approximation point x_0 . Namely, it is assumed that $x_0 - \hat{x}$ is sufficiently smooth: $x_0 - \hat{x} = F'^*(\hat{x})F'(\hat{x})w$ for some element $w \in H$.

In that paper the discrete analogs of algorithm (1.6)-(1.7) are presented (see formulas (2.4)-(2.5) and (2.46)-(2.47) below). In section 2 an auxiliary lemma is stated and proved, and the main convergence theorem is established. As a consequence of this theorem the stability of process (2.4)-(2.5) towards noise in the data is obtained in section 3. Also in this section the choice of an optimal regularization parameter (the stopping time), such that the method converges to a solution of (1.1) when the noise level tends to zero, is done. In section 4 the practically important problem of gravitational sounding [4] is considered, and method (2.4)-(2.5) is tested numerically. Based on theoretical and numerical results the recommendations on the choice of $\{\varepsilon_n\}$, λ and B_0 are given in section 2, Remark 2.5, and in section 4.

1. Numerical Algorithm and Convergence Theorem

Consider iteratively-regularized Gauss-Newton's procedure for solving nonlinear equation (1.1):

$$(2.1) \quad x_{n+1} = x_n - [F'^*(x_n)F'(x_n) + \varepsilon_n I]^{-1}[F'^*(x_n)F(x_n) + \varepsilon_n(x_n - x_0)],$$

where $\varepsilon_n > 0$, $x_0 \in H_1$ and I is an identity operator on H_1 . The reader may consult [3] (and also later publications [5], [9]) for a convergence analysis of (2.1). Problem (2.1) is equivalent to the following system

$$(2.2) \quad x_{n+1} = x_n - Q_n[F'^*(x_n)F(x_n) + \varepsilon_n(x_n - x_0)],$$

$$(2.3) \quad [F'^*(x_n)F'(x_n) + \varepsilon_n I]Q_n - I = 0, \quad Q_n \in L(H_1).$$

Solving equation (2.3) by iterations, we arrive at the regularized procedure

$$(2.4) \quad x_{n+1} = x_n - B_n[F'^*(x_n)F(x_n) + \varepsilon_n(x_n - x_0)],$$

$$(2.5) \quad B_{n+1} = [I - \lambda(F'^*(x_n)F'(x_n) + \varepsilon_n I)]B_n + \lambda I,$$

$$x_0 \in H_1, \quad B_0 \in L(H_1), \quad 0 < \varepsilon_n \rightarrow 0 \quad \text{as } n \rightarrow \infty, \quad \lambda > 0.$$

In order to establish our main convergence result we will need the following lemma. It is a discrete operator-theoretical version of the well-known Gronwall inequality. A continuous analog of this lemma was analyzed in [11].

LEMMA 2.1. *Let*

$$(2.6) \quad V_{n+1} = (I - A_n)V_n + G_n,$$

where A_n , G_n , $V_n \in L(H)$, $A_n = A_n^*$, and H is a Hilbert space. If there exists a sequence of real numbers $\{\gamma_n\}$ such that

$$(2.7) \quad (A_n h, h) \geq \gamma_n \|h\|^2$$

and

$$(2.8) \quad \|A_n\| \leq 2 - \gamma_n,$$

then

$$(2.9) \quad \|V_{n+1}\| \leq e^{-(\gamma_0 + \dots + \gamma_n)} \left\{ \|V_0\| + \sum_{k=0}^n e^{\gamma_0 + \dots + \gamma_k} \|G_k\| \right\}.$$

REMARK 2.2. If $\|A_n\| \leq 1$, then for a sequence $\{\gamma_n\}$, satisfying (2.7), condition (2.8) is fulfilled.

Proof of Lemma 2.1 By identity (2.6)

$$\begin{aligned} V_{n+1} &= (I - A_n)(I - A_{n-1})\dots(I - A_0)V_0 + \dots + (I - A_n)\dots(I - A_1)G_0 \\ &\quad + (I - A_n)\dots(I - A_2)G_1 + \dots + (I - A_n)G_{n-1} + G_n. \end{aligned}$$

From (2.7) and (2.8) it follows that $\gamma_n \leq 1$ and $\|I - A_n\| \leq 1 - \gamma_n$. Therefore

$$\begin{aligned} (2.10) \quad \|V_{n+1}\| &\leq (1 - \gamma_n)(1 - \gamma_{n-1})\dots(1 - \gamma_0)\|V_0\| + (1 - \gamma_n)\dots(1 - \gamma_1)\|G_0\| \\ &\quad + (1 - \gamma_n)\dots(1 - \gamma_2)\|G_1\| + \dots + (1 - \gamma_n)\|G_{n-1}\| + \|G_n\|. \end{aligned}$$

Applying the elementary estimate

$$(1 - x_1)\dots(1 - x_s) \leq e^{-(x_1 + \dots + x_s)}, \quad x_k \leq 1, \quad k = 1, 2, \dots, s,$$

to the right-hand side of (2.10) one gets

$$\begin{aligned} (2.11) \quad \|V_{n+1}\| &\leq e^{-(\gamma_n + \gamma_{n-1} + \dots + \gamma_0)}\|V_0\| + e^{-(\gamma_n + \dots + \gamma_1)}\|G_0\| \\ &\quad + e^{-(\gamma_n + \dots + \gamma_2)}\|G_1\| + \dots + e^{-\gamma_n}\|G_{n-1}\| + \|G_n\|. \end{aligned}$$

Inequality (2.11) is equivalent to (2.9). \square

THEOREM 2.3. Let H_1 and H_2 be Hilbert spaces, $F : H_1 \rightarrow H_2$.

1) Suppose the equation $F(x) = 0$ is solvable (not necessarily uniquely) and \hat{x} is a solution.

2) Let F be twice Fréchet differentiable in H_1 and

$$(2.12) \quad \|F'(x)\| \leq N_1, \quad \|F''(x)\| \leq N_2 \quad \text{for all } x \in H_1.$$

3) Assume that the sequence of positive numbers $\{\varepsilon_n\}$, the positive number λ , and the initial operator $B_0 \in L(H_1)$ satisfy

$$(2.13) \quad (a) \quad \varepsilon_n \rightarrow 0 \quad \text{monotonically and} \quad \frac{\varepsilon_n - \varepsilon_{n+1}}{\lambda \varepsilon_n^2} \leq b \quad \text{for some } b > 0,$$

$$(2.14) \quad (b) \quad \lambda \leq \frac{2}{N_1^2 + 2\varepsilon_0}.$$

$$(2.15) \quad \text{Write } \eta := b e^{2\lambda\varepsilon_0} + \|I - B_0(F'^*(\hat{x})F'(\hat{x}) + \varepsilon_0 I)\| e^{-\lambda\varepsilon_0},$$

$$(2.16) \quad (c) \quad (\lambda b \varepsilon_0 + 1)\eta < 1, \quad \varepsilon_0 \|B_0\| \leq e^{\lambda\varepsilon_0}.$$

4) Suppose the initial approximation x_0 is chosen so that $x_0 - \hat{x}$ can be written in the form

$$(2.17) \quad x_0 - \hat{x} = F'^*(\hat{x})F'(\hat{x})w \quad \text{with} \quad \|w\| \leq \rho,$$

and

$$(2.18) \quad (\lambda b \varepsilon_0 + 1) \left\{ \eta + 2N_1 N_2 e^{\lambda\varepsilon_0} \rho + \mu \max \left[\sqrt{2(\eta + 1)\rho}, \frac{\mu \|x_0 - \hat{x}\|}{\varepsilon_0} \right] \right\} \leq 1,$$

where

$$(2.19) \quad \mu := \sqrt{(7e^{\lambda\varepsilon_0} + 3\varepsilon_0 \|B_0\| e^{-\lambda\varepsilon_0}) N_1 N_2}.$$

Then the sequence $\{x_n\}$, defined by (2.4)-(2.5), converges to \hat{x} and

$$(2.20) \quad \|x_n - \hat{x}\| \leq \frac{1 - (\eta + 2N_1 N_2 e^{\lambda\varepsilon_0} \rho)(\lambda b \varepsilon_0 + 1)}{\mu^2(\lambda b \varepsilon_0 + 1)} \varepsilon_n.$$

REMARK 2.4. Formally in order to verify conditions (a)-(c) on the parameters of algorithm (2.4)-(2.5) one has to know the solution \hat{x} . However if one takes $B_0 = \lambda I$, conditions (a)-(c) can be verified without knowing \hat{x} . Indeed, taking (2.14) into consideration it is easy to see that $\|I - \lambda(F'^*(\hat{x})F'(\hat{x}) + \varepsilon_0 I)\| \leq 1 - \lambda\varepsilon_0$ and $\lambda\varepsilon_0 \leq 1$. Therefore (c) holds, if b satisfies $(\lambda b\varepsilon_0 + 1)(be^{2\lambda\varepsilon_0} + (1 - \lambda\varepsilon_0)e^{-\lambda\varepsilon_0}) < 1$.

REMARK 2.5. One way to choose parameters $\{\varepsilon_n\}$, λ , B_0 satisfying (a)-(c) is as follows:

1. Pick λ and ε_0 to satisfy condition (b).
2. Choose b so that $(\lambda b\varepsilon_0 + 1)(be^{2\lambda\varepsilon_0} + e^{-\lambda\varepsilon_0}) < 1$ in order to satisfy condition (c) with $B_0 = \lambda I$.
3. Take $\varepsilon_n = \frac{\varepsilon_0 c^a}{(c+n)^a}$ with any $a \in (0, 1]$ and $c > 0$ satisfying

$$(2.21) \quad \left(1 + \frac{1}{c}\right)^a \left[\left(1 + \frac{1}{c}\right)^a - 1 \right] \leq \lambda b\varepsilon_0.$$

One can check (2.21) implies that for the sequence $\{\varepsilon_n\}$ condition (a) holds.

REMARK 2.6. Condition 4) of Theorem 2.3 is not algorithmically verifiable. However some condition of this type is necessary if one works with an operator $F'(x)$, which has no continuous inverse, and if no assumptions on the spectrum of $F'(x)$ are made.

Proof of Theorem 2.3 From (2.4) one gets

$$(2.22) \quad \|x_{n+1} - \hat{x}\| = \|x_n - \hat{x} - B_n[F'^*(x_n)F(x_n) + \varepsilon_n(x_n - x_0)]\|.$$

By condition 2) of Theorem 2.3 one has

$$\begin{aligned} F'^*(x_n)F(x_n) &= F'^*(x_n)[F'(\hat{x})(x_n - \hat{x}) + R_2(x_n, \hat{x})] \\ &= F'^*(x_n)R_2(x_n, \hat{x}) + (F'(x_n) - F'(\hat{x}))^*F'(\hat{x})(x_n - \hat{x}) \\ (2.23) \quad &\quad + F'^*(\hat{x})F'(\hat{x})(x_n - \hat{x}), \end{aligned}$$

where

$$\|R_2(x_n, \hat{x})\| \leq \frac{N_2}{2} \|x_n - \hat{x}\|^2.$$

We have $\|F'(x_n) - F'(\hat{x})\| \leq N_2 \|x_n - \hat{x}\|$. Hence

$$\begin{aligned} &\|F'^*(x_n)R_2(x_n, \hat{x}) + (F'(x_n) - F'(\hat{x}))^*F'(\hat{x})(x_n - \hat{x})\| \\ (2.24) \quad &\leq \frac{3N_1N_2}{2} \|x_n - \hat{x}\|^2. \end{aligned}$$

Thus by (2.22)-(2.24) one concludes

$$\begin{aligned} &\|x_{n+1} - \hat{x}\| \leq \|x_n - \hat{x} - B_n[(F'^*(\hat{x})F'(\hat{x}) + \varepsilon_n I)(x_n - \hat{x}) \\ (2.25) \quad &\quad + \varepsilon_n(\hat{x} - x_0)]\| + \frac{3N_1N_2}{2} \|B_n\| \|x_n - \hat{x}\|^2. \end{aligned}$$

From assumption (2.17) it follows that

$$\begin{aligned} &\|B_n(\hat{x} - x_0)\| = \|B_n[F'^*(\hat{x})F'(\hat{x}) + \varepsilon_n I][F'^*(\hat{x})F'(\hat{x}) + \varepsilon_n I]^{-1}F'^*(\hat{x})F'(\hat{x})w\| \\ (2.26) \quad &\leq \rho \|B_n[F'^*(\hat{x})F'(\hat{x}) + \varepsilon_n I]\|. \end{aligned}$$

Estimates (2.25) and (2.26) imply

$$(2.27) \quad \begin{aligned} \|x_{n+1} - \hat{x}\| &\leq \|I - B_n[F'^*(\hat{x})F'(\hat{x}) + \varepsilon_n I]\| \|x_n - \hat{x}\| \\ &+ \varepsilon_n \rho \|B_n[F'^*(\hat{x})F'(\hat{x}) + \varepsilon_n I]\| + \frac{3N_1 N_2}{2} \|B_n\| \|x_n - \hat{x}\|^2. \end{aligned}$$

Let us now derive the estimates for $\|B_n\|$ and for $\|I - B_n[F'^*(\hat{x})F'(\hat{x}) + \varepsilon_n I]\|$. Since for any $h \in H_1$ $((F'^*(x_n)F'(x_n) + \varepsilon_n I)h, h) \geq \varepsilon_n \|h\|^2$, by conditions (2.13) and (2.14) one can apply Lemma 2.1 to operator sequence (2.5) with

$$V_n := B_n, \quad A_n := \lambda(F'^*(x_n)F'(x_n) + \varepsilon_n I), \quad G_n := \lambda I.$$

Inequality (2.9) yields

$$\begin{aligned} \|B_{n+1}\| &\leq e^{-\lambda(\varepsilon_0 + \dots + \varepsilon_n)} \left(\|B_0\| + \sum_{k=0}^n \lambda e^{\lambda(\varepsilon_0 + \dots + \varepsilon_k)} \right) \\ &= e^{-\lambda(\varepsilon_0 + \dots + \varepsilon_n)} \left(\|B_0\| + \sum_{k=0}^n \frac{1}{\varepsilon_{k+1}} \lambda \varepsilon_{k+1} e^{\lambda(\varepsilon_0 + \dots + \varepsilon_k)} \right). \end{aligned}$$

From monotonicity of $\{\varepsilon_n\}$ one obtains

$$(2.28) \quad \begin{aligned} \|B_{n+1}\| &\leq e^{-\lambda(\varepsilon_0 + \dots + \varepsilon_n)} \left(\|B_0\| + \frac{1}{\varepsilon_{n+1}} \sum_{k=0}^n \lambda \varepsilon_{k+1} e^{\lambda(\varepsilon_0 + \dots + \varepsilon_k)} \right) \\ &= e^{-\lambda(\varepsilon_0 + \dots + \varepsilon_n)} \left(\|B_0\| + \frac{1}{\varepsilon_{n+1}} \int_{\lambda\varepsilon_0}^{\lambda(\varepsilon_0 + \dots + \varepsilon_{n+1})} e^s ds \right) \\ &= \|B_0\| e^{-\lambda(\varepsilon_0 + \dots + \varepsilon_n)} + \frac{1}{\varepsilon_{n+1}} \left(e^{\lambda\varepsilon_{n+1}} - e^{-\lambda(\varepsilon_1 + \dots + \varepsilon_n)} \right) \\ &\leq \|B_0\| e^{-\lambda\varepsilon_0} + \frac{e^{\lambda\varepsilon_0}}{\varepsilon_{n+1}}. \end{aligned}$$

Note that $\lim_{n \rightarrow \infty} \sum_{k=0}^n \varepsilon_k = \infty$. Indeed, by condition (2.13)

$$\frac{\varepsilon_k}{\varepsilon_{k+1}} \leq \lambda b \varepsilon_{k+1} + 1 \leq \lambda b \varepsilon_0 + 1.$$

Therefore one gets

$$\int_{\varepsilon_{n+1}}^{\varepsilon_0} \frac{ds}{s^2} \leq \sum_{k=0}^n \frac{\varepsilon_k - \varepsilon_{k+1}}{\varepsilon_{k+1}^2} \leq (\lambda b \varepsilon_0 + 1)^2 \sum_{k=0}^n \frac{\varepsilon_k - \varepsilon_{k+1}}{\varepsilon_k^2} \leq \lambda b (\lambda b \varepsilon_0 + 1)^2 (n+1).$$

Thus

$$\frac{1}{\varepsilon_{n+1}} \leq \frac{1}{\varepsilon_0} + \text{const} (n+1).$$

Introduce the notation

$$(2.29) \quad \Lambda_n := I - B_n[F'^*(\hat{x})F'(\hat{x}) + \varepsilon_n I].$$

One has by (2.29) and (2.5)

$$\begin{aligned} \Lambda_{n+1} &= I - B_{n+1}[F'^*(\hat{x})F'(\hat{x}) + \varepsilon_n I] - B_{n+1}(\varepsilon_{n+1} - \varepsilon_n) \\ &= I - \{[I - \lambda(F'^*(x_n)F'(x_n) + \varepsilon_n I)]B_n + \lambda I\}[F'^*(\hat{x})F'(\hat{x}) + \varepsilon_n I] - B_{n+1}(\varepsilon_{n+1} - \varepsilon_n) \\ &= I - [I - \lambda(F'^*(x_n)F'(x_n) + \varepsilon_n I)]B_n[F'^*(\hat{x})F'(\hat{x}) + \varepsilon_n I] - \lambda[F'^*(\hat{x})F'(\hat{x}) + \varepsilon_n I] \end{aligned}$$

$$(2.30) \quad -B_{n+1}(\varepsilon_{n+1} - \varepsilon_n).$$

Identity (2.29) implies that

$$\begin{aligned} \Lambda_{n+1} &= I - [I - \lambda(F'^*(x_n)F'(x_n) + \varepsilon_n I)](I - \Lambda_n) - \lambda[F'^*(\hat{x})F'(\hat{x}) + \varepsilon_n I] - B_{n+1}(\varepsilon_{n+1} - \varepsilon_n) \\ &= [I - \lambda(F'^*(x_n)F'(x_n) + \varepsilon_n I)]\Lambda_n + \lambda(F'^*(x_n)F'(x_n) - F'^*(\hat{x})F'(\hat{x})) - B_{n+1}(\varepsilon_{n+1} - \varepsilon_n). \end{aligned}$$

Apply Lemma 2.1 once again with

$$\begin{aligned} V_n &:= \Lambda_n, \quad A_n := \lambda(F'^*(x_n)F'(x_n) + \varepsilon_n I), \\ G_n &:= \lambda(F'^*(x_n)F'(x_n) - F'^*(\hat{x})F'(\hat{x})) - B_{n+1}(\varepsilon_{n+1} - \varepsilon_n). \end{aligned}$$

Let us estimate $\|G_n\|$:

$$\|G_n\| \leq \lambda \| (F'(x_n) - F'(\hat{x}))^* F(x_n) + F'^*(\hat{x})(F'(x_n) - F'(\hat{x})) \| + (\varepsilon_n - \varepsilon_{n+1}) \|B_{n+1}\|.$$

From (2.12) one gets

$$(2.31) \quad \|G_n\| \leq 2N_1 N_2 \lambda \|x_n - \hat{x}\| + (\varepsilon_n - \varepsilon_{n+1}) \|B_{n+1}\|.$$

By condition (2.18) $(\lambda b \varepsilon_0 + 1) \left(\eta + 2N_1 N_2 e^{\lambda \varepsilon_0} \rho + \frac{\mu^2 \|x_0 - \hat{x}\|}{\varepsilon_0} \right) \leq 1$, which yields

$$(2.32) \quad \frac{\|x_0 - \hat{x}\|}{\varepsilon_0} \leq \frac{1 - (\eta + 2N_1 N_2 e^{\lambda \varepsilon_0} \rho)(\lambda b \varepsilon_0 + 1)}{\mu^2 (\lambda b \varepsilon_0 + 1)} := R.$$

Suppose by induction that

$$(2.33) \quad \frac{\|x_k - \hat{x}\|}{\varepsilon_k} \leq R \quad \forall k = 1, 2, \dots, n.$$

Under this assumption inequality (2.31) implies

$$(2.34) \quad \|G_n\| \leq 2N_1 N_2 R \lambda \varepsilon_n + (\varepsilon_n - \varepsilon_{n+1}) \|B_{n+1}\|.$$

By (2.9) and (2.34) one has:

$$(2.35) \quad \begin{aligned} \|\Lambda_{n+1}\| &\leq e^{-\lambda(\varepsilon_0 + \dots + \varepsilon_n)} \left\{ \|\Lambda_0\| + \sum_{k=0}^n e^{\lambda(\varepsilon_0 + \dots + \varepsilon_k)} \left[2N_1 N_2 R \lambda \varepsilon_k \right. \right. \\ &\quad \left. \left. + (\varepsilon_k - \varepsilon_{k+1}) \|B_{k+1}\| \right] \right\}. \end{aligned}$$

From (2.28) one gets

$$(2.36) \quad \begin{aligned} \|\Lambda_{n+1}\| &\leq e^{-\lambda(\varepsilon_0 + \dots + \varepsilon_n)} \left\{ \|\Lambda_0\| + \sum_{k=0}^n e^{\lambda(\varepsilon_0 + \dots + \varepsilon_k)} \left[2N_1 N_2 R \lambda \varepsilon_k + (\varepsilon_k - \varepsilon_{k+1}) \right. \right. \\ &\quad \left. \left. \left(\|B_0\| e^{-\lambda(\varepsilon_0 + \dots + \varepsilon_k)} + \frac{1}{\varepsilon_{k+1}} (e^{\lambda \varepsilon_{k+1}} - e^{-\lambda(\varepsilon_1 + \dots + \varepsilon_k)}) \right) \right] \right\} = e^{-\lambda(\varepsilon_0 + \dots + \varepsilon_n)} \left\{ \|\Lambda_0\| \right. \\ &\quad \left. + 2N_1 N_2 R \left[e^{\lambda \varepsilon_0} \lambda \varepsilon_0 + \sum_{k=1}^n e^{\lambda \varepsilon_k} \lambda \varepsilon_k e^{\lambda(\varepsilon_0 + \dots + \varepsilon_{k-1})} \right] + \sum_{k=0}^n \left[\frac{\varepsilon_k - \varepsilon_{k+1}}{\lambda \varepsilon_{k+1}^2} e^{\lambda \varepsilon_{k+1}} \right. \right. \\ &\quad \left. \left. \lambda \varepsilon_{k+1} e^{\lambda(\varepsilon_0 + \dots + \varepsilon_k)} + \left(\|B_0\| - \frac{e^{\lambda \varepsilon_0}}{\varepsilon_{k+1}} \right) (\varepsilon_k - \varepsilon_{k+1}) \right] \right\}. \end{aligned}$$

Since $\varepsilon_0 \|B_0\| \leq e^{\lambda \varepsilon_0}$ and $\varepsilon_k \geq \varepsilon_{k+1}$ for any $k \geq 0$, one concludes that

$$\left(\|B_0\| - \frac{e^{\lambda \varepsilon_0}}{\varepsilon_{k+1}} \right) (\varepsilon_k - \varepsilon_{k+1}) \leq 0.$$

Thus by (2.13) one obtains:

$$\begin{aligned} \|\Lambda_{n+1}\| &\leq e^{-\lambda(\varepsilon_0+\dots+\varepsilon_n)} \left\{ \|\Lambda_0\| + 2N_1N_2Re^{\lambda\varepsilon_0} \int_0^{\lambda(\varepsilon_0+\dots+\varepsilon_n)} e^s ds \right. \\ &\quad \left. + b e^{\lambda\varepsilon_0} \int_{\lambda\varepsilon_0}^{\lambda(\varepsilon_0+\dots+\varepsilon_{n+1})} e^s ds = \|\Lambda_0\|e^{-\lambda(\varepsilon_0+\dots+\varepsilon_n)} \right. \\ &\quad \left. + 2N_1N_2Re^{\lambda\varepsilon_0} \left(1 - e^{-\lambda(\varepsilon_0+\dots+\varepsilon_n)} \right) + b \left(e^{\lambda(\varepsilon_0+\varepsilon_{n+1})} - e^{\lambda(\varepsilon_0-\varepsilon_1-\dots-\varepsilon_n)} \right). \right. \end{aligned}$$

Hence

$$(2.37) \quad \|\Lambda_{n+1}\| \leq (2N_1N_2R + be^{\lambda\varepsilon_0})e^{\lambda\varepsilon_0} + \|\Lambda_0\|e^{-\lambda\varepsilon_0} := k.$$

Now denote

$$(2.38) \quad D_n := B_n [F'^*(\hat{x})F'(\hat{x}) + \varepsilon_n I].$$

By (2.29) and (2.37) $\|\Lambda_n\| = \|I - D_n\| \leq k$, therefore

$$(2.39) \quad \|D_n\| \leq k + 1.$$

From (2.29), (2.38) and (2.27) one has

$$\|x_{n+1} - \hat{x}\| \leq \|\Lambda_n\| \|x_n - \hat{x}\| + \varepsilon_n \rho \|D_n\| + \frac{3N_1N_2}{2} \|B_n\| \|x_n - \hat{x}\|^2.$$

Thus by (2.28), (2.37) and (2.39) one gets

$$\|x_{n+1} - \hat{x}\| \leq \frac{3N_1N_2(e^{\lambda\varepsilon_0} + \varepsilon_0 \|B_0\|e^{-\lambda\varepsilon_0})}{2\varepsilon_n} \|x_n - \hat{x}\|^2 + k \|x_n - \hat{x}\|$$

$$(2.40) \quad + \varepsilon_n \rho (k + 1).$$

Introduce the notation

$$(2.41) \quad \sigma_n := \frac{\|x_n - \hat{x}\|}{\varepsilon_n}.$$

Inequality (2.40) implies

$$\sigma_{n+1} \leq \frac{3N_1N_2(e^{\lambda\varepsilon_0} + \varepsilon_0 \|B_0\|e^{-\lambda\varepsilon_0})\varepsilon_n}{2\varepsilon_{n+1}} \sigma_n^2 + \frac{k\varepsilon_n}{\varepsilon_{n+1}} \sigma_n + \frac{\varepsilon_n \rho (k + 1)}{\varepsilon_{n+1}}.$$

From condition (2.13)

$$(2.42) \quad \frac{\varepsilon_n}{\varepsilon_{n+1}} \leq \lambda b \varepsilon_{n+1} + 1 \leq \lambda b \varepsilon_0 + 1.$$

Thus one gets

$$\begin{aligned} \sigma_{n+1} &\leq \frac{3N_1N_2(e^{\lambda\varepsilon_0} + \varepsilon_0 \|B_0\|e^{-\lambda\varepsilon_0})(\lambda b \varepsilon_0 + 1)}{2} \sigma_n^2 + k(\lambda b \varepsilon_0 + 1) \sigma_n \\ &\quad + \rho(k + 1)(\lambda b \varepsilon_0 + 1). \end{aligned} \tag{2.43}$$

By induction assumption (2.33), by (2.37), (2.15) and (2.19)

$$\begin{aligned} \sigma_{n+1} &\leq \left\{ \frac{3N_1N_2(e^{\lambda\varepsilon_0} + \varepsilon_0 \|B_0\|e^{-\lambda\varepsilon_0})}{2} + 2N_1N_2e^{\lambda\varepsilon_0} \right\} (\lambda b \varepsilon_0 + 1) R^2 \\ &\quad + \left\{ b e^{2\lambda\varepsilon_0} + \|\Lambda_0\| e^{-\lambda\varepsilon_0} + 2N_1N_2e^{\lambda\varepsilon_0} \rho \right\} (\lambda b \varepsilon_0 + 1) R \end{aligned}$$

$$(2.44) \quad = \frac{\mu^2(\lambda b\varepsilon_0 + 1)}{2} R^2 + (\eta + 2N_1 N_2 e^{\lambda\varepsilon_0} \rho)(\lambda b\varepsilon_0 + 1)R + (\eta + 1)\rho(\lambda b\varepsilon_0 + 1).$$

Therefore from (2.32) one obtains

$$\begin{aligned} \sigma_{n+1} - R &\leq \frac{\mu^2(\lambda b\varepsilon_0 + 1)}{2} R^2 + \left[(\eta + 2N_1 N_2 e^{\lambda\varepsilon_0} \rho)(\lambda b\varepsilon_0 + 1) - 1 \right] R + (\eta + 1)\rho(\lambda b\varepsilon_0 + 1) \\ &= \frac{\mu^2(\lambda b\varepsilon_0 + 1)[1 - (\eta + 2N_1 N_2 e^{\lambda\varepsilon_0} \rho)(\lambda b\varepsilon_0 + 1)]^2}{2\mu^4(\lambda b\varepsilon_0 + 1)^2} \\ &\quad - \frac{[1 - (\eta + 2N_1 N_2 e^{\lambda\varepsilon_0} \rho)(\lambda b\varepsilon_0 + 1)]^2}{\mu^2(\lambda b\varepsilon_0 + 1)} + (\eta + 1)\rho(\lambda b\varepsilon_0 + 1) \\ (2.45) \quad &= -\frac{[1 - (\eta + 2N_1 N_2 e^{\lambda\varepsilon_0} \rho)(\lambda b\varepsilon_0 + 1)]^2 - 2\mu^2(\eta + 1)\rho(\lambda b\varepsilon_0 + 1)^2}{2\mu^2(\lambda b\varepsilon_0 + 1)} \leq 0, \end{aligned}$$

since by (2.18) it follows that

$$\mu\sqrt{2(\eta + 1)\rho}(\lambda b\varepsilon_0 + 1) \leq 1 - (\eta + 2N_1 N_2 e^{\lambda\varepsilon_0} \rho)(\lambda b\varepsilon_0 + 1).$$

Estimate (2.45) yields $\sigma_{n+1} \leq R$.

Inequality (2.20) follows from (2.41) and (2.32). This completes the proof. \square

COROLLARY 2.7. Consider the discretization of continuous algorithm (1.6)-(1.7) in the following form:

$$(2.46) \quad u_{n+1} = u_n - T_n[F'^*(u_n)F(u_n) + \varepsilon_n(u_n - u_0)],$$

$$(2.47) \quad T_{n+1} = [I - \lambda(F'^*(u_{n+1})F'(u_{n+1}) + \varepsilon_n I)]T_n + \lambda I,$$

$$u_0 \in H_1, \quad T_0 \in L(H_1), \quad 0 < \varepsilon_n \rightarrow 0 \quad \text{as } n \rightarrow \infty, \quad \lambda > 0.$$

Suppose assumptions 1)-3) of Theorem 2.3 with the operator B_0 being replaced by T_0 are satisfied, the initial approximation u_0 is chosen so that $u_0 - \hat{x}$ admits the representation

$$(2.48) \quad u_0 - \hat{x} = F'^*(\hat{x})F'(\hat{x})v \quad \text{with } \|v\| \leq \rho,$$

and $\|u_n - \hat{x}\|$ and ρ are sufficiently small.

Then the sequence $\{u_n\}$, defined by (2.46)-(2.47), converges to \hat{x} and

$$\|u_n - \hat{x}\| = O(\varepsilon_n) \quad \text{as } n \rightarrow \infty.$$

Indeed, by Lemma 2.1

$$\begin{aligned} \|T_{n+1}\| &\leq \|T_0\| e^{-\lambda(\varepsilon_0 + \dots + \varepsilon_n)} + \frac{1}{\varepsilon_{n+1}} \left(e^{\lambda\varepsilon_{n+1}} - e^{-\lambda(\varepsilon_1 + \dots + \varepsilon_n)} \right) \\ (2.49) \quad &\leq \|T_0\| \varepsilon^{-\lambda\varepsilon_0} + \frac{e^{\lambda\varepsilon_0}}{\varepsilon_{n+1}}. \end{aligned}$$

Define $W_n := I - T_n[F'^*(\hat{x})F'(\hat{x}) + \varepsilon_n I]$, then by Lemma 2.1 one has

$$\begin{aligned} \|W_{n+1}\| &\leq e^{-\lambda(\varepsilon_0 + \dots + \varepsilon_n)} \left\{ \|W_0\| + \sum_{k=0}^n e^{\lambda(\varepsilon_0 + \dots + \varepsilon_k)} [2N_1 N_2 \lambda \|u_{k+1} - \hat{x}\| \right. \\ (2.50) \quad &\quad \left. + (\varepsilon_k - \varepsilon_{k+1}) \|T_{k+1}\|] \right\}. \end{aligned}$$

Let the inequality $\|u_0 - \hat{x}\| \leq \hat{R}\varepsilon_0$ be fulfilled for some $\hat{R} > 0$. If one assumes by induction that

$$(2.51) \quad \|u_k - \hat{x}\| \leq \hat{R}\varepsilon_k, \quad \forall k = 1, 2, \dots, n,$$

then one gets

$$\begin{aligned} \sum_{k=0}^n \lambda \|u_{k+1} - \hat{x}\| e^{\lambda(\varepsilon_0 + \dots + \varepsilon_k)} &= \lambda \|u_{n+1} - \hat{x}\| e^{\lambda(\varepsilon_0 + \dots + \varepsilon_n)} + \sum_{k=0}^{n-1} \frac{\|u_{k+1} - \hat{x}\|}{\varepsilon_{k+1}} \\ \lambda \varepsilon_{k+1} e^{\lambda(\varepsilon_0 + \dots + \varepsilon_k)} &\leq \lambda \|u_{n+1} - \hat{x}\| e^{\lambda(\varepsilon_0 + \dots + \varepsilon_n)} + \hat{R} \left(e^{\lambda(\varepsilon_0 + \dots + \varepsilon_n)} - e^{\lambda\varepsilon_0} \right). \end{aligned}$$

Thus

$$\begin{aligned} \|W_{n+1}\| &\leq 2N_1 N_2 (\hat{R} + \lambda \|u_{n+1} - \hat{x}\|) + b e^{2\lambda\varepsilon_0} \\ (2.52) \quad + \|W_0\| e^{-\lambda\varepsilon_0} &:= \hat{k} + 2N_1 N_2 \lambda \|u_{n+1} - \hat{x}\|. \end{aligned}$$

Hence for $\hat{\sigma}_n := \frac{\|u_n - \hat{x}\|}{\varepsilon_n}$ one obtains

$$\begin{aligned} \hat{\sigma}_{n+1} &\leq \frac{3N_1 N_2 (e^{\lambda\varepsilon_0} + \varepsilon_0 \|T_0\| e^{-\lambda\varepsilon_0}) \varepsilon_n}{2\varepsilon_{n+1}} \hat{\sigma}_n^2 + \frac{\hat{k}\varepsilon_n}{\varepsilon_{n+1}} \hat{\sigma}_n + \frac{\varepsilon_n \rho(\hat{k} + 1)}{\varepsilon_{n+1}} \\ (2.53) \quad + 2N_1 N_2 \lambda \varepsilon_n (\rho + \hat{\sigma}_n) \hat{\sigma}_{n+1}. \end{aligned}$$

By analyzing estimate (2.53) one can find the values of \hat{R} and ρ so that (2.51) and (2.53) imply $\hat{\sigma}_{n+1} \leq \hat{R}$. The existence of such values follows from conditions (2.15) and (2.16) of Theorem 2.3 with $B_0 = T_0$. \square

PROPOSITION 2.8. Condition that $x_0 - \hat{x}$ can be written as $x_0 - \hat{x} = F'^*(\hat{x})F'(\hat{x})w$ with $\|w\| \leq \rho$ is equivalent to

$$(2.54) \quad \sup_{\varepsilon > 0} \|(F'^*(\hat{x})F'(\hat{x}) + \varepsilon I)^{-1}(x_0 - \hat{x})\| \leq \rho.$$

Indeed, denote $A := F'^*(\hat{x})F'(\hat{x})$. Clearly, if $x_0 - \hat{x} = Aw$ then (2.54) holds. Conversely, let us assume (2.54). Write $x_0 - \hat{x} = y_1 + y_2$, where $y_1 \in \text{Ker } A$, $y_2 \in \overline{\text{Ran } A}$. Note $H_1 = \text{Ker } A \oplus \overline{\text{Ran } A}$. Choose a sequence of positive numbers $\{\alpha_n\}$ so that $\alpha_n \rightarrow 0$ and

$$(2.55) \quad (A + \alpha_n I)^{-1}(x_0 - \hat{x}) \rightarrow w \quad \text{as } n \rightarrow \infty, \quad w \in H_1, \quad \|w\| \leq \rho,$$

in the weak topology of H_1 . Since for all $n \geq 1$ $\text{Ker } A$ and $\overline{\text{Ran } A}$ are reducing subspaces for $(A + \alpha_n I)^{-1}$, the sequence $(A + \alpha_n I)^{-1}y_1$ is also weakly convergent. But $(A + \alpha_n I)^{-1}y_1 = \frac{1}{\alpha_n}y_1$, therefore $y_1 = 0$. Next, from (2.55) one concludes that $A(A + \alpha_n I)^{-1}(x_0 - \hat{x})$ converges in the norm to the projection of $x_0 - \hat{x}$ onto $\overline{\text{Ran } A}$, which is equal to $x_0 - \hat{x}$. Hence $x_0 - \hat{x} = Aw$ with $\|w\| \leq \rho$. \square

2. Stability Analysis

In this section we assume that the operator F is given by its approximation $\tilde{F} : H_1 \rightarrow H_2$ such that

$$(3.1) \quad \|\tilde{F}(\hat{x})\| \leq \delta_1.$$

We also assume that source-type condition (2.17) is replaced with

$$(3.2) \quad x_0 - \hat{x} = F'^*(\hat{x})F'(\hat{x})w + \zeta \quad \text{with} \quad \|w\| \leq \rho, \quad \|\zeta\| \leq \delta_2,$$

and that

$$(3.3) \quad \|(F'^*(\hat{x})F'(\hat{x}) - \tilde{F}'^*(\hat{x})\tilde{F}'(\hat{x}))w\| \leq \delta_3.$$

Consider the iteratively regularized algorithm

$$(3.4) \quad \tilde{x}_{n+1} = \tilde{x}_n - \tilde{B}_n[\tilde{F}'^*(\tilde{x}_n)\tilde{F}'(\tilde{x}_n) + \varepsilon_n(\tilde{x}_n - x_0)],$$

$$(3.5) \quad \tilde{B}_{n+1} = [I - \lambda(\tilde{F}'^*(\tilde{x}_n)\tilde{F}'(\tilde{x}_n) + \varepsilon_n I)]\tilde{B}_n + \lambda I,$$

$$\tilde{x}_0 = x_0 \in H_1, \quad \tilde{B}_0 = B_0 \in L(H_1), \quad 0 < \varepsilon_n \rightarrow 0 \quad \text{as } n \rightarrow \infty, \quad \lambda > 0.$$

A convergence analysis of (3.4)-(3.5) is provided in the following theorem.

THEOREM 3.1. *Let H_1 and H_2 be Hilbert spaces, F and $\tilde{F} : H_1 \rightarrow H_2$.*

1) *Suppose the equation $F(x) = 0$ is solvable (not necessarily uniquely), \hat{x} is a solution and (3.1) is satisfied.*

2) *Let F and \tilde{F} be twice Fréchet differentiable in H_1 and*

$$(3.6) \quad \|\tilde{F}'(x)\| \leq N_1, \quad \|\tilde{F}''(x)\| \leq N_2 \quad \text{for all } x \in H_1.$$

3) *Assume that the sequence of positive numbers $\{\varepsilon_n\}$, the positive number λ , and the initial operator $B_0 \in L(H_1)$ satisfy*

$$(3.7) \quad (a) \quad \varepsilon_n \rightarrow 0 \quad \text{monotonically and} \quad \frac{\varepsilon_n - \varepsilon_{n+1}}{\lambda \varepsilon_n^2} \leq b \quad \text{for some } b > 0,$$

$$(3.8) \quad (b) \quad \lambda \leq \frac{2}{N_1^2 + 2\varepsilon_0}.$$

$$(3.9) \quad \text{Write } \tilde{\eta} := 2 \left\{ b e^{2\lambda\varepsilon_0} + \|I - B_0(\tilde{F}'^*(\hat{x})\tilde{F}'(\hat{x}) + \varepsilon_0 I)\| e^{-\lambda\varepsilon_0} \right\},$$

$$(3.10) \quad (c) \quad (\lambda b \varepsilon_0 + 1)\tilde{\eta} < 1, \quad \varepsilon_0 \|B_0\| \leq e^{\lambda\varepsilon_0}.$$

4) *Suppose the initial approximation x_0 is chosen so that (3.2), (3.3) hold and*

$$(3.11) \quad (\lambda b \varepsilon_0 + 1) \left\{ \tilde{\eta} + 4N_1 N_2 e^{\lambda\varepsilon_0} \rho + \tilde{\mu} \max \left[\sqrt{2(\tilde{\eta} + 2)\rho}, \frac{\tilde{\mu} \|x_0 - \hat{x}\|}{\varepsilon_0} \right] \right\} \leq 1,$$

where

$$(3.12) \quad \tilde{\mu} := \sqrt{(11e^{\lambda\varepsilon_0} + 3\varepsilon_0 \|B_0\| e^{-\lambda\varepsilon_0}) N_1 N_2}.$$

Suppose that \tilde{n} is the minimal n with $\max \left\{ \frac{2(\delta_2 + \delta_3)}{\rho \varepsilon_n}, \frac{\delta_1}{\rho \varepsilon_n^{3/2}}, \frac{N_2 \delta_1}{\varepsilon_n} \right\} > 1$.

Then for any $n \leq \tilde{n}$

$$(3.13) \quad 1) \quad \|\tilde{x}_n - \hat{x}\| \leq \frac{1 - (\tilde{\eta} + 4N_1 N_2 e^{\lambda\varepsilon_0} \rho)(\lambda b \varepsilon_0 + 1)}{\tilde{\mu}^2 (\lambda b \varepsilon_0 + 1)} \varepsilon_n,$$

and

$$(3.14) \quad 2) \quad \|\tilde{x}_{\tilde{n}} - \hat{x}\| = O(\delta_1^{2/3} + \delta_2 + \delta_3),$$

where $\{\tilde{x}_n\}$ is defined by (3.4)-(3.5).

REMARK 3.2. The stability analysis of the iteratively regularized Gauss-Newton procedure under assumptions (3.1)-(3.3) was done by A.Bakushinsky in [3].

REMARK 3.3. The error in the operator F may be due to one or a combination of the following sources:

1. A physical phenomenon is modeled approximately by a nonlinear operator equation $\tilde{F}(x) = 0$, $\tilde{F} : H_1 \rightarrow H_2$.

2. The operator \tilde{F} is computed from measured data.

3. \tilde{F} is a numerical approximation of F , i.e. one can take projections $P_n : H_1 \rightarrow H_1^{(n)}$ and $Q_m : H_2 \rightarrow H_2^{(m)}$ onto finite dimensional subspaces and define $\tilde{F} = Q_m F P_n$.

REMARK 3.4. Algorithm (2.46)-(2.47) under conditions (3.1)-(3.3) can be treated similarly.

Proof of Theorem 3.1 From (3.2) it follows that

$$\begin{aligned} \tilde{B}_n(\hat{x} - x_0) &= \tilde{B}_n[\tilde{F}'^*(\hat{x})\tilde{F}'(\hat{x}) + \varepsilon_n I][\tilde{F}'^*(\hat{x})\tilde{F}'(\hat{x}) + \varepsilon_n I]^{-1} \{(F'^*(\hat{x})F'(\hat{x}) \\ (3.15) \quad &- \tilde{F}'^*(\hat{x})\tilde{F}'(\hat{x}))w + \tilde{F}'^*(\hat{x})\tilde{F}'(\hat{x})w + \zeta\}. \end{aligned}$$

Applying the representation

$$\begin{aligned} \tilde{B}_n\tilde{F}'^*(x_n)\tilde{F}(\hat{x}) &= \tilde{B}_n[\tilde{F}'^*(\hat{x})\tilde{F}'(\hat{x}) + \varepsilon_n I][\tilde{F}'^*(\hat{x})\tilde{F}'(\hat{x}) + \varepsilon_n I]^{-1} \{\tilde{F}'^*(\hat{x}) \\ (3.16) \quad &+ (\tilde{F}'(\tilde{x}_n) - \tilde{F}'(\hat{x}))^*\} \tilde{F}(\hat{x}), \end{aligned}$$

and the estimate

$$(3.17) \quad \|\tilde{F}'^*(\hat{x})\tilde{F}'(\hat{x}) + \varepsilon_n I\|^{-1}\tilde{F}'^*(\hat{x}) \leq \frac{1}{2\sqrt{\varepsilon_n}},$$

under assumptions 1) and 2) of Theorem 3.1 and conditions (3.1)-(3.3) one has

$$\begin{aligned} \|\tilde{x}_{n+1} - \hat{x}\| &\leq \|I - \tilde{B}_n[\tilde{F}'^*(\hat{x})\tilde{F}'(\hat{x}) + \varepsilon_n I]\| \|\tilde{x}_n - \hat{x}\| \\ &+ \varepsilon_n \|\tilde{B}_n[\tilde{F}'^*(\hat{x})\tilde{F}'(\hat{x}) + \varepsilon_n I]\| \left(\rho + \frac{\delta_2 + \delta_3}{\varepsilon_n} \right) + \frac{3N_1 N_2}{2} \|\tilde{B}_n\| \|\tilde{x}_n - \hat{x}\|^2 \\ (3.18) \quad &+ \|\tilde{B}_n[\tilde{F}'^*(\hat{x})\tilde{F}'(\hat{x}) + \varepsilon_n I]\| \left(\frac{1}{2\sqrt{\varepsilon_n}} + \frac{N_2 \|\tilde{x}_n - \hat{x}\|}{\varepsilon_n} \right) \delta_1. \end{aligned}$$

Using the same argument as in the proof of Theorem 2.3 one can easily check that

$$\|\tilde{B}_{n+1}\| \leq \|B_0\| e^{-\lambda(\varepsilon_0 + \dots + \varepsilon_n)} + \frac{1}{\varepsilon_{n+1}} \left(e^{\lambda\varepsilon_{n+1}} - e^{-\lambda(\varepsilon_1 + \dots + \varepsilon_n)} \right) \leq \|B_0\| e^{-\lambda\varepsilon_0} + \frac{e^{\lambda\varepsilon_0}}{\varepsilon_{n+1}}.$$

By (3.11) one obtains

$$(3.19) \quad \frac{\|x_0 - \hat{x}\|}{\varepsilon_0} \leq \frac{1 - (\tilde{\eta} + 4N_1 N_2 e^{\lambda\varepsilon_0} \rho)(\lambda b \varepsilon_0 + 1)}{\tilde{\mu}^2(\lambda b \varepsilon_0 + 1)} := \tilde{R}.$$

If one assumes that

$$(3.20) \quad \|\tilde{x}_k - \hat{x}\| \leq \tilde{R} \varepsilon_k \quad \text{for any } k = 1, 2, \dots, n,$$

then one gets

$$(3.21) \quad \|\tilde{\Lambda}_{n+1}\| \leq (2N_1 N_2 \tilde{R} + b e^{\lambda\varepsilon_0}) e^{\lambda\varepsilon_0} + \|\tilde{\Lambda}_0\| e^{-\lambda\varepsilon_0} := \tilde{k},$$

where

$$(3.22) \quad \tilde{\Lambda}_n := I - \tilde{B}_n[\tilde{F}'^*(\hat{x})\tilde{F}'(\hat{x}) + \varepsilon_n I].$$

Inequality (3.21) yields

$$(3.23) \quad \|\tilde{B}_n[\tilde{F}'^*(\hat{x})\tilde{F}'(\hat{x}) + \varepsilon_n I]\| \leq \tilde{k} + 1.$$

Now by (3.18) for $\tilde{\sigma}_n := \frac{\|\hat{x}_n - \hat{x}\|}{\varepsilon_n}$ one has

$$\tilde{\sigma}_{n+1} \leq \frac{3N_1 N_2 (e^{\lambda\varepsilon_0} + \varepsilon_0 \|B_0\| e^{-\lambda\varepsilon_0})(\lambda b\varepsilon_0 + 1)}{2} \tilde{\sigma}_n^2 + \left[\tilde{k} \left(1 + \frac{N_2 \delta_1}{\varepsilon_n} \right) + \frac{N_2 \delta_1}{\varepsilon_n} \right]$$

$$(3.24) \quad (\lambda b\varepsilon_0 + 1) \tilde{\sigma}_n + (\tilde{k} + 1)(\lambda b\varepsilon_0 + 1) \left(\rho + \frac{\delta_2 + \delta_3}{\varepsilon_n} + \frac{\delta_1}{2\varepsilon_n^{3/2}} \right).$$

From (3.20), (3.21) one concludes from (3.24) that

$$\begin{aligned} \tilde{\sigma}_{n+1} &\leq \left\{ \frac{3N_1 N_2 (e^{\lambda\varepsilon_0} + \varepsilon_0 \|B_0\| e^{-\lambda\varepsilon_0})}{2} + 2N_1 N_2 e^{\lambda\varepsilon_0} \left(1 + \frac{N_2 \delta_1}{\varepsilon_n} \right) \right\} (\lambda b\varepsilon_0 + 1) \tilde{R}^2 \\ &\quad + \left\{ \left(b e^{2\lambda\varepsilon_0} + \|\tilde{\Lambda}_0\| e^{-\lambda\varepsilon_0} \right) \left(1 + \frac{N_2 \delta_1}{\varepsilon_n} \right) \right. \\ &\quad \left. + 2N_1 N_2 e^{\lambda\varepsilon_0} \left(\rho + \frac{\delta_2 + \delta_3}{\varepsilon_n} + \frac{\delta_1}{2\varepsilon_n^{3/2}} \right) \right\} (\lambda b\varepsilon_0 + 1) \tilde{R} \\ &\quad + \left\{ b e^{2\lambda\varepsilon_0} + \|\tilde{\Lambda}_0\| e^{-\lambda\varepsilon_0} + 1 \right\} (\lambda b\varepsilon_0 + 1) \rho \left(1 + \frac{\delta_2 + \delta_3}{\rho\varepsilon_n} + \frac{\delta_1}{2\rho\varepsilon_n^{3/2}} \right). \end{aligned}$$

Suppose that \tilde{n} is the minimal value of n , for which

$$(3.25) \quad \max \left\{ \frac{2(\delta_2 + \delta_3)}{\rho\varepsilon_n}, \frac{\delta_1}{\rho\varepsilon_n^{3/2}}, \frac{N_2 \delta_1}{\varepsilon_n} \right\} > 1.$$

Then from (3.9), (3.12) and (3.19) by direct calculations one has

$$\tilde{\sigma}_{n+1} \leq \frac{\tilde{\mu}^2(\lambda b\varepsilon_0 + 1)}{2} \tilde{R}^2 + (\tilde{\eta} + 4N_1 N_2 e^{\lambda\varepsilon_0} \rho)(\lambda b\varepsilon_0 + 1) \tilde{R} + (\tilde{\eta} + 2)\rho(\lambda b\varepsilon_0 + 1) \leq \tilde{R}$$

for any $n < \tilde{n}$. Thus (3.13) holds. Inequalities (3.13) and (3.25) imply (3.14). \square

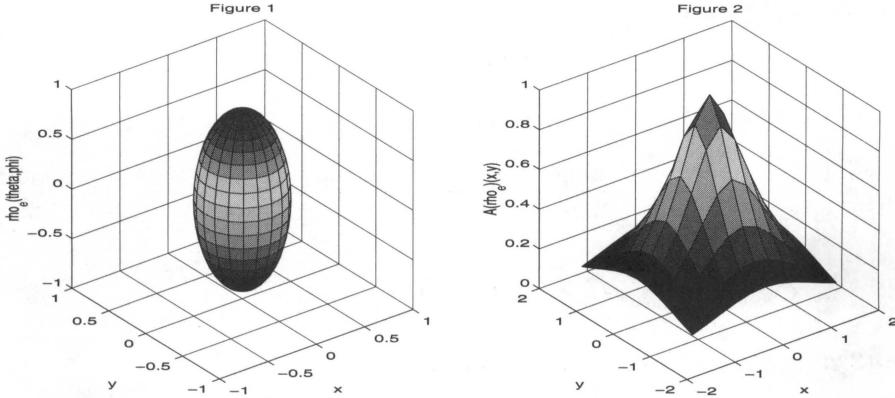
REMARK 3.5. One of the main assumptions of the paper is that for all $x \in H_1$ $\|F'(x)\| \leq N_1$, $\|F''(x)\| \leq N_2$ (or $\|\tilde{F}'(x)\| \leq N_1$, $\|\tilde{F}''(x)\| \leq N_2$). The second condition is often fulfilled in practical applications. But if one considers e.g. the operator of autoconvolution, then the first estimate holds only in a bounded region. The structure of the proofs for convergence and stability of the proposed methods is always that it is shown

$$\frac{\|x_n - \hat{x}\|}{\varepsilon_n} \leq R$$

holds and, because ε_n is monotone decreasing, all iterates will stay in a bounded neighborhood of the solution \hat{x} ,

$$\|x_n - \hat{x}\| \leq \varepsilon_0 R.$$

Therefore it is actually sufficient for the first estimate $\|F'(x)\| \leq N_1$ to hold in a sufficiently large neighborhood of \hat{x} .



3. Numerical Aspects

To understand the numerical aspects, related to the above approach, the practically important problem of gravitational sounding (see [4]) was considered. Let T be a sub-surface homogeneous body with density $\omega_1 \neq \omega_0$, where ω_0 is the density of surrounding uniform medium. The potential V of the gravitational field outside T is given by the following triple integral:

$$(4.1) \quad V(x, y, z) = \nu \int \int_T \int \frac{d\xi d\kappa d\zeta}{\sqrt{(x - \xi)^2 + (y - \kappa)^2 + (z - \zeta)^2}},$$

where $\nu = \nu_0(\omega_1 - \omega_0)$, ν_0 is a gravitational constant. For the z -component of the gravitational field on the level $z = 0$ one has

$$(4.2) \quad -\frac{\partial V(x, y, 0)}{\partial z} = \nu \int \int_T \int \frac{\zeta d\xi d\kappa d\zeta}{r_{|z=0}^{3/2}}.$$

Here $r = \sqrt{(x - \xi)^2 + (y - \kappa)^2 + (z - \zeta)^2}$ is the boundary of T . Equation (4.2) sets the correspondence between the function r to be defined and the experimentally measurable values $g(x, y) := -\frac{\partial V(x, y, 0)}{\partial z}$. If T is a convex body and a point inside T is known *a priori*, then $r = \rho(\phi, \theta)$ in spherical coordinates satisfies the nonlinear integral equation of the first kind:

$$(4.3) \quad A(\rho) = g(x, y)$$

with

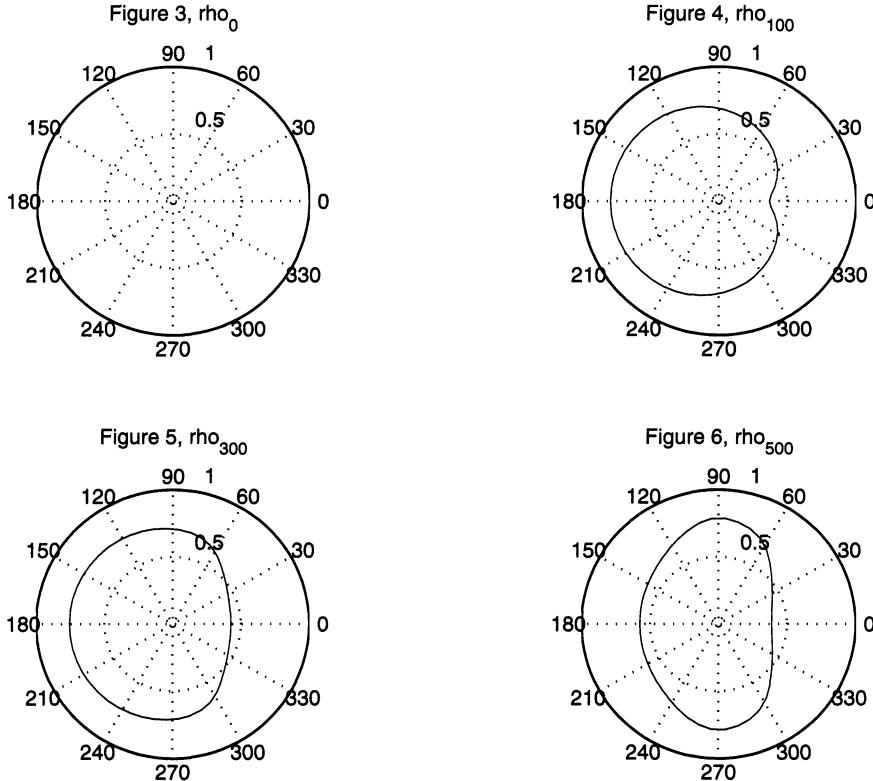
$$A(\rho) := \int_0^{2\pi} \int_0^\pi \int_0^{\rho(\phi, \theta)} \frac{\sin \theta (H - r \cos \theta) r^2 dr d\theta d\phi}{[(x - r \sin \theta \cos \phi)^2 + (y - r \sin \theta \sin \phi)^2 + (H - r \cos \theta)^2]^{3/2}}.$$

Let A act between the pair of Hilbert spaces H_1 and H_2 , where $H_1 = H^1(0, 2\pi) \times (0, \pi)$ or $L^2(0, 2\pi) \times (0, \pi)$ and $H_2 = L^2(-2, 2) \times (-2, 2)$. The aim of the numerical experiment was to solve equation (4.3) using regularized algorithm (2.4)-(2.5):

$$(4.4) \quad \rho_{n+1} = \rho_n - B_n[A'^*(\rho_n)(A(\rho_n) - g) + \varepsilon_n(\rho_n - \rho_0)],$$

$$(4.5) \quad B_{n+1} = [I - \lambda(A'^*(\rho_n)A'(\rho_n) + \varepsilon_n I)]B_n + \lambda I,$$

$$\rho_0 \in H_1, \quad B_0 \in L(H_1), \quad 0 < \varepsilon_n \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty, \quad \lambda > 0.$$



The gravity strength anomaly $g(x, y)$ (see Figure 2) was chosen as the solution to the direct problem for the model surface $\frac{x^2}{(0.4)^2} + \frac{y^2}{(0.4)^2} + \frac{z^2}{(0.9)^2} = 1$, illustrated on Figure 1. Then the approximate solutions to the inverse problem were computed for two initial functions: $\rho_0(\phi, \theta) \equiv 1$, $\rho_0(\phi, \theta) \equiv 0.6$. The parameters of the iterative process were as follows:

$$H = 1.5, \quad \varepsilon_n = 0.01/(1+n), \quad \lambda = 0.05, \quad B_0 = 0.05I.$$

The experiment was conducted in the presence of noise, whose maximum value was 6% of the maximum value of $g(x, y)$.

One can see on Figures 4, 5 and 6 the intersections of the approximate solutions, obtained after 100, 300 and 500 iterations, and the xz -plane for $\rho_0 \equiv 1$. After 500 iterations the discrepancy $\sigma := \|A(\rho_{500}) - g\|_{L^2}$ was equal to $9.73 \cdot 10^{-5}$ and the relative error $\Delta := \|\rho_{500} - \rho_{\text{exact}}\|/\|\rho_{\text{exact}}\|_{L^2}$ was equal to $9.22 \cdot 10^{-3}$.

Figures 8, 9 and 10 show the intersections of the approximate solutions ρ_{100} , ρ_{300} and ρ_{500} and the xz -plane for $\rho_0 \equiv 0.6$. In this case $\sigma = 7.32 \cdot 10^{-5}$ and $\Delta = 6.45 \cdot 10^{-3}$.

Table 1 represents the dependence of the discrepancies on the parameters ε_0 and λ after 300 iterations for $\rho_0 \equiv 0.6$. Table 2 represents the dependence of the relative errors on the parameters ε_0 and λ after 300 iterations for $\rho_0 \equiv 0.6$.

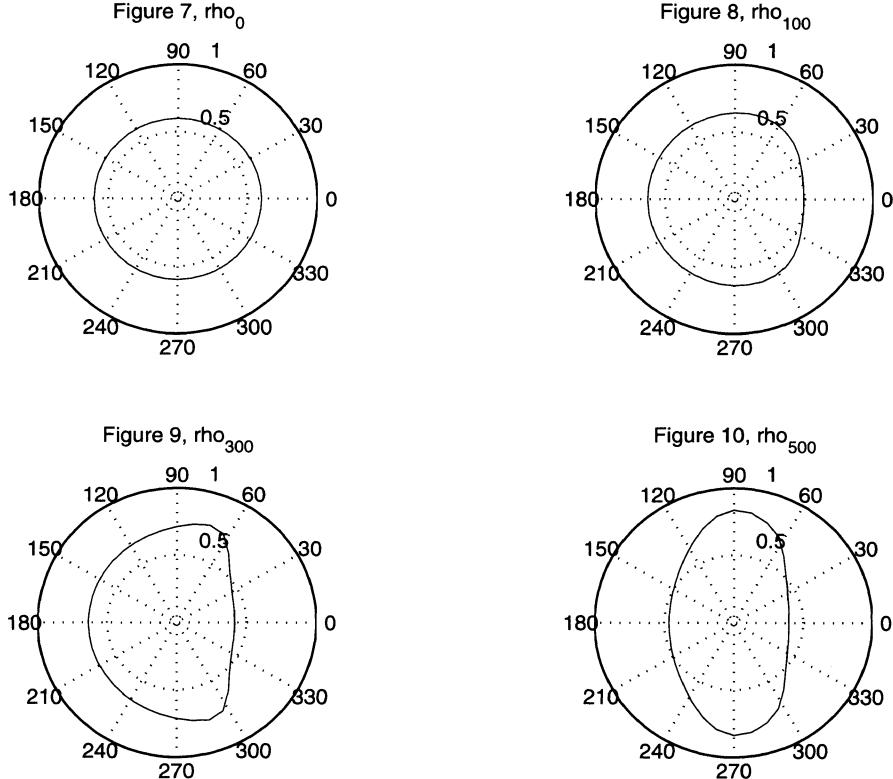


Table 1

$\sigma := \ A(\rho_{300}) - g\ _{L_2}$				
$B_0 = \lambda I, \quad \varepsilon_n = \varepsilon_0/(1+n), \quad \rho_0 \equiv 0.6, \quad N = 300$				
λ	ε_0	0.001	0.01	0.05
0.025		$3.48 \cdot 10^{-4}$	$2.62 \cdot 10^{-4}$	$3.85 \cdot 10^{-4}$
0.05		$3.28 \cdot 10^{-4}$	$2.59 \cdot 10^{-2}$	$3.79 \cdot 10^{-4}$
0.075		$2.69 \cdot 10^{-4}$	$2.59 \cdot 10^{-4}$	$3.75 \cdot 10^{-4}$
0.1		$2.51 \cdot 10^{-4}$	$2.48 \cdot 10^{-4}$	$2.76 \cdot 10^{-4}$

Table 2

$\Delta := \ \rho_{300} - \rho_{\text{exact}}\ / \ \rho_{\text{exact}}\ _{L^2}$				
$B_0 = \lambda I, \quad \varepsilon_n = \varepsilon_0/(1+n), \quad \rho_0 \equiv 0.6, \quad N = 300$				
λ	ε_0	0.001	0.01	0.05
0.025		$4.03 \cdot 10^{-2}$	$3.41 \cdot 10^{-2}$	$4.07 \cdot 10^{-2}$
0.05		$3.29 \cdot 10^{-2}$	$3.27 \cdot 10^{-2}$	$3.41 \cdot 10^{-2}$
0.075		$3.12 \cdot 10^{-2}$	$2.87 \cdot 10^{-2}$	$3.28 \cdot 10^{-2}$
0.1		$2.25 \cdot 10^{-2}$	$2.17 \cdot 10^{-2}$	$3.08 \cdot 10^{-2}$

REMARK 4.1.

- Other regularizing sequences of the form $\varepsilon_n = \varepsilon_0/(1+n)^a$, $a \in (0, 1]$, were also considered. The best numerical results were obtained with $a = 1$.
- For the problem investigated an appropriate range of possible values of ε_0 appears to be from 0.001 to 0.1; for larger values the accuracy is lower, and for smaller values the process does not converge.
- Clearly for larger values of λ the iterative sequence converges faster, and after 300 steps the most accurate iterative solutions were obtained for $\lambda = 0.1$. However for smaller values of λ the process is more stable and allows one to get better accuracy after 500 iterations.
- The number of iterations (300-500) in the application example is quite high for a Newton type method, but the algorithm is nevertheless efficient, since the effort per step is small due to the update strategy for the operator $(F'^*(x_n)F'(x_n) + \varepsilon_n I)^{-1}$. Also as it was mentioned above the number of iterations can be reduced if one takes larger values of λ .

References

- [1] Airapetyan, R.G., Ramm A.G. and Smirnova, A.B. [1999] *Continuous analog of Gauss-Newton method*, Math. Models and Meth. in Appl. Sci. **9**, N3, 463–474.
- [2] Airapetyan, R.G., Ramm, A.G. and Smirnova, A.B. [2000] *Continuous methods for solving nonlinear ill-posed problems*, Operator theory and its applications, Amer. Math.Soc., Providence RI, Fields Inst. Commun. **25**, 111-137.
- [3] A.B. Bakushinskii, A.B. [1993] *Iterative methods for nonlinear operator equations without regularity. New approach*, Dokl. Russian Acad. Sci. **330**, 282-284.
- [4] Bakushinskii, A.B. and Goncharskii, A.V. [1994] *Ill-Posed Problems: Theory and Applications*, Kluwer, Dordrecht.
- [5] Blaschke, B., Neubauer, A. and Scherzer O. [1997] *On convergence rates for the iteratively regularized Gauss-Newton method*, IMA J. Num. Anal. **17**, 421-436.
- [6] Broyden, C.G. [1965] *A class of methods for solving nonlinear simultaneous equations*, Math. Comp. **19**, 577-593.
- [7] Dennis, J.E., Schnabel, R.B. [1983] *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, New Jersey.
- [8] Griewank, A. [1984] *Rates of convergence for secant methods on nonlinear problems in Hilbert spaces*, Lecture notes in Mathematics **1230**.
- [9] Hohage, T. [1997] *Logarithmic convergence rates of the iteratively regularized Gauss-Newton method for an inverse potential and inverse scattering problem*, Inverse problems **13**, 1279-1299.
- [10] Kaltenbacher, B. [1998] *On Broyden's method for the regularization of nonlinear ill-posed problems*, Numer. Funct. Anal. and Optimiz. **19**, 807-833.
- [11] Ramm, A. G. and Smirnova, A.B., Continuous regularized Gauss-Newton-type algorithm for nonlinear ill-posed equations with simultaneous updates of inverse derivative, (to appear).
- [12] Saschs, E.W. [1986] Broyden's method in Hilbert space, Mathematical programming **35**, 71-82.

DEPARTMENT OF MATHEMATICS AND STATISTICS, GEORGIA STATE UNIVERSITY, ATLANTA, GA 30303-3083, USA

E-mail address: alexeev@mathstat.gsu.edu

DEPARTMENT OF MATHEMATICS AND STATISTICS, GEORGIA STATE UNIVERSITY, ATLANTA, GA 30303-3083, USA

E-mail address: smirn@mathstat.gsu.edu

This page intentionally left blank

A Fractal Set Constructed from a Class of Wavelet Sets

John J. Benedetto and Songkiat Sumetkijakan

1. Introduction

1.1. Background. A function ψ in $L^2(\mathbb{R}^d)$ is called a *single dyadic orthonormal wavelet* if the family of functions $\{\psi_{j,k}(\cdot) \equiv 2^{jd/2}\psi(2^j \cdot - k) : j \in \mathbb{Z}, k \in \mathbb{Z}^d\}$, is an orthonormal basis for $L^2(\mathbb{R}^d)$. More generally, a collection of functions $\psi^1, \psi^2, \dots, \psi^M$ in $L^2(\mathbb{R}^d)$ is called a *wavelet collection* if the dyadic dilations and integer translations of the ψ^i 's,

$$\left\{ \psi_{j,k}^i(\cdot) \equiv 2^{jd/2}\psi^i(2^j \cdot - k) : j \in \mathbb{Z}, k \in \mathbb{Z}^d, i = 1, \dots, M \right\},$$

form an orthonormal basis for $L^2(\mathbb{R}^d)$.

The first wavelet in L^2 is the Haar wavelet $\psi = \mathbf{1}_{[0,\frac{1}{2})} - \mathbf{1}_{[\frac{1}{2},1)}$ discovered by Haar [Haa10] in 1909. The existence of smooth wavelets had not been established until March 1987 when Daubechies [Dau88] showed that for any $n \in \mathbb{N}$ there exists a compactly supported orthonormal wavelet which is n -times continuously differentiable. This remarkable achievement was proved at about the same time as Mallat and Meyer's discovery of a method for constructing orthonormal wavelets from a family of subspaces of $L^2(\mathbb{R}^d)$ satisfying certain properties, e.g., see [Mal85], [Mal89], [Mey86], [Mey87], [Mey89], [Mey90]. This family is called a *multiresolution analysis*(MRA); and Meyer [Mey87] first observed that Daubechies' theorem can be formulated in this context. It is now known, e.g., see [Aus95], [Gri95], [Wan], that if $\{\psi^1, \psi^2, \dots, \psi^M\}$ is a wavelet collection associated with an MRA, then M must be $2^d - 1$. Further, it was proved by Auscher [Aus95] that every wavelet collection $\{\psi^1, \psi^2, \dots, \psi^M\}$ for $L^2(\mathbb{R}^d)$, whose members satisfy a weak smoothness and decay condition on the Fourier transform side, must come from an MRA. Because of these results, and notwithstanding Journé's celebrated example of a non-MRA wavelet basis for $L^2(\mathbb{R})$, e.g., [Dau92], there was some question during the mid-1990s about the existence of single dyadic orthonormal wavelets in $L^2(\mathbb{R}^d)$.

The existence of single dyadic orthonormal wavelets in $L^2(\mathbb{R}^d)$, $d > 1$, was proved using operator algebra methods by Dai and Larson, e.g., see [DLS97]. The

2000 *Mathematics Subject Classification*. Primary 42C40, 28A80.

The first named author gratefully acknowledges DARPA Grant MDA-972011003 and the General Research Board of the University of Maryland.

wavelets constructed in [DLS97] are inverse Fourier transforms of characteristic functions of certain measurable sets in $\widehat{\mathbb{R}}^d$. These sets are called wavelet sets. A *wavelet set* in $\widehat{\mathbb{R}}^d$ is defined as a measurable set K such that the characteristic function $\mathbf{1}_K$ is the Fourier transform $\widehat{\psi}$ of some dyadic orthonormal wavelet ψ in $L^2(\mathbb{R}^d)$. Soardi and Weiland [SW98] later constructed several wavelet sets in $\widehat{\mathbb{R}}^2$; there are also examples from the same period due to Zakharov [Zak96]. A necessary and sufficient condition for the existence of a wavelet set K in $\widehat{\mathbb{R}}^d$ is given by the tiling properties:

$$(1 \text{ a}) \quad \{K + k : k \in \mathbb{Z}^d\} \text{ is a partition for } \widehat{\mathbb{R}}^d \text{ a.e.,}$$

$$(1 \text{ b}) \quad \{2^j K : j \in \mathbb{Z}\} \text{ is a partition for } \widehat{\mathbb{R}}^d \text{ a.e.,}$$

e.g., see [BL99], [BL01].

For comparison, it should be pointed out that a complete characterization of single orthonormal wavelets ψ with $\|\psi\|_2 = 1$ is given by the equations,

$$(2 \text{ a}) \quad \sum_{j \in \mathbb{Z}} \left| \widehat{\psi}(2^j \xi) \right|^2 = 1 \quad \text{a.e. } \xi \in \widehat{\mathbb{R}}^d,$$

$$(2 \text{ b}) \quad \sum_{j=0}^{\infty} \widehat{\psi}(2^j \xi) \overline{\widehat{\psi}(2^j(\xi + k))} = 0 \quad \text{a.e. } \xi \in \widehat{\mathbb{R}}^d, \quad k \in \mathbb{Z}^d \setminus 2\mathbb{Z}^d,$$

see, e.g., section 7.1 in [HW96].

With regard to the techniques herein and those of [BL01], we note the ground-breaking results found in [GM92], [GH94], and [LW96], which implement similar concepts. As such we emphasize that these important papers deal essentially with MRAs, whereas our wavelets can *not* be obtained from MRAs.

1.2. Contents. In section 2, we review a general construction of wavelet sets in $\widehat{\mathbb{R}}^d$, introduced in [BL01], which we call the *neighborhood-mapping method*. We then give a reformulation of this construction when such wavelet sets are contained in $[-1, 1]^d$, and point out a self-similarity property of these sets. This is the content of section 3. Section 4 explains how a fractal set \mathcal{B} arises from this class of wavelet sets, and there is a characterization of \mathcal{B} in equations (9), (11), and Theorem 8. This will be used to prove a result concerning the generality of the neighborhood-mapping method in section 5. A brief analysis of another construction method of wavelet sets in $\widehat{\mathbb{R}}^d$, due to Baggett, Medina, and Merrill [BMM99], is given in section 6.

1.3. Notation and definitions. We shall use the standard notation and results from harmonic analysis and fractal theory as found in [Ben97], [SW71], and [Fal90]. For completeness we provide the following definitions which are used in the sequel.

DEFINITION 1. A map $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is called a *contraction* if there is a positive number $c < 1$ such that

$$\|\varphi(x) - \varphi(y)\| \leq c \|x - y\| \quad \text{for all } x, y \in \mathbb{R}^d$$

We say that φ is a *similarity* with *ratio* $c < 1$ if $\|\varphi(x) - \varphi(y)\| = c \|x - y\|$ for all $x, y \in \mathbb{R}^d$.

A set F that is invariant under a finite collection of similarities $\varphi_1, \dots, \varphi_n$, that is

$$F = \bigcup_{i=1}^n \varphi_i(F),$$

is called a *self-similar set*.

DEFINITION 2. a. For any non-empty set U in \mathbb{R}^d , the *diameter* of U is defined as

$$|U| = \sup \{|x - y| : x, y \in U\}.$$

Let F be a subset of \mathbb{R}^d . If $\{U_i\}$ is a countable collection of sets of diameter at most δ that cover F , i.e., $F \subseteq \bigcup_{i=1}^{\infty} U_i$ with $0 < |U_i| \leq \delta$ for each i , we say that $\{U_i\}$ is a δ -cover of F .

b. For a subset $F \subseteq \mathbb{R}^d$, a non-negative number s , and any $\delta > 0$, we define

$$\mathcal{H}_\delta^s(F) = \inf \left\{ \sum_{i=1}^{\infty} |U_i|^s : \{U_i\}_{i=1}^{\infty} \text{ is a } \delta\text{-cover of } F \right\}.$$

Note that $\mathcal{H}_\delta^s(F)$ is increasing with δ , but decreasing with s , when the other two variables are fixed. This gives rise to a well-defined measure,

$$\mathcal{H}^s(F) = \lim_{\delta \rightarrow 0} \mathcal{H}_\delta^s(F),$$

taking values in $[0, \infty]$, called the *s-dimensional Hausdorff measure* of F .

Moreover if $s < t$ and $\{U_i\}_{i=1}^{\infty}$ is a δ -cover of F , we have $\mathcal{H}_\delta^t(F) \leq \delta^{t-s} \mathcal{H}_\delta^s(F)$. Therefore,

- if $\mathcal{H}^s(F) < \infty$, then $\mathcal{H}^t(F) = 0$ for all $t > s$, and
- if $\mathcal{H}^t(F) > 0$, then $\mathcal{H}^s(F) = \infty$ for all $s < t$.

c. The *Hausdorff dimension* of F is

$$\dim_H(F) = \inf \{s : \mathcal{H}^s(F) = 0\} = \sup \{s : \mathcal{H}^s(F) = \infty\}.$$

A set F whose Hausdorff dimension is not an integer is said to be a *fractal set*.

Throughout the paper, especially in section 3, we denote the unit cube $[-\frac{1}{2}, \frac{1}{2}]^d$ by Q . The set $\{-1, 0, 1\}^d \setminus \{(0, \dots, 0)\}$ will be denoted by Λ .

2. A general construction of wavelet sets

A general construction of wavelet sets in \mathbb{R}^d was introduced by Benedetto and Leon [BL01] in 1997. They begin with a fixed number N and a neighborhood of the origin $K_0 \subseteq [-N, N]^d$, which is τ -congruent to the unit cube $Q \equiv [-\frac{1}{2}, \frac{1}{2}]^d$. This means that there exist countable measurable partitions $\{P_m : m \in \mathbb{Z}\}$ and $\{Q_m : m \in \mathbb{Z}\}$ of K_0 and Q , respectively, and a sequence $\{n_m : m \in \mathbb{Z}\} \subseteq \mathbb{Z}^d$ such that $P_m = Q_m + n_m$ for all $m \in \mathbb{Z}$. K_0 is then successively transformed by a given measurable injective integer-translated map,

$$\begin{aligned} T : K_0 &\rightarrow \overline{[-2N, 2N]^d \setminus [-N, N]^d} \\ x &\mapsto x + n_x \quad \text{for some } n_x \in \mathbb{Z}^d, \end{aligned}$$

in the following way.

We first define

$$A_0 = K_0 \cap \left[\bigcup_{j \geq 1} 2^{-j} K_0 \right] \quad \text{and} \quad K_1 = [K_0 \setminus A_0] \cup T(A_0).$$

Note that

$$K_1^- \equiv K_0 \setminus A_0 \subseteq K_0 \subseteq [-N, N]^d,$$

and

$$K_1^+ \equiv T(A_0) \subseteq \overline{[-2N, 2N]^d \setminus [-N, N]^d}.$$

Generally, for a given $K_n = K_n^- \cup K_n^+$, where K_n^- is contained in $[-N, N]^d$ and K_n^+ is contained in the “annulus” $\overline{[-2N, 2N]^d \setminus [-N, N]^d}$, $n \geq 1$, we define

$$A_n = K_n \cap \left[\bigcup_{j \geq 1} 2^{-j} K_n \right]$$

and

$$K_{n+1} = [K_n^- \setminus A_n] \cup [K_n^+ \cup T(A_n)] \equiv K_{n+1}^- \cup K_{n+1}^+.$$

Thus,

$$K_{n+1} = \left[K_0 \setminus \bigcup_{k=0}^n A_k \right] \cup \left[\bigcup_{k=0}^n T(A_k) \right].$$

Finally, we define

$$(3) \quad K = \left[K_0 \setminus \bigcup_{k=0}^{\infty} A_k \right] \cup \left[\bigcup_{k=0}^{\infty} T(A_k) \right].$$

It is not difficult to show that K satisfies tiling conditions in (1 a) and (1 b), and hence K is a wavelet set. This is the construction in [BL01].

REMARK 3. Even though the definition of A_n is intuitively clear, there are some unnecessary set computations involved. For the sake of computational efficiency, we shall show that

$$(4) \quad A_n = \left[K_0 \setminus \bigcup_{k=0}^{n-1} A_k \right] \cap \left[\bigcup_{j \geq 1} 2^{-j} T(A_{n-1}) \right].$$

Indeed, from the decomposition of K_n into K_n^- inside the cube $[-N, N]^d$ and K_n^+ outside the open cube $(-N, N)^d$, but still inside $[-2N, 2N]^d$, it is clear that

$$A_n = K_n^- \cap \left[\bigcup_{j \geq 1} 2^{-j} K_n \right] \text{ a.e.}$$

Moreover, $K_n^- \cap \left[\bigcup_{j \geq 1} 2^{-j} K_n^- \right]$ is empty. In fact, if $x \in K_n^- \subseteq K_0 \setminus A_0$, then $x \notin 2^{-j} K_0$ for all $j \geq 1$, and so x cannot be in the union $\bigcup_{j \geq 1} 2^{-j} K_n^-$. Therefore,

for each $n \geq 1$, we have

$$(5) \quad \begin{aligned} A_n &= K_n^- \cap \left[\bigcup_{j \geq 1} 2^{-j} K_n^+ \right] \\ &= \left[K_0 \setminus \bigcup_{k=0}^{n-1} A_k \right] \cap \left[\bigcup_{j \geq 1} 2^{-j} \left(\bigcup_{k=0}^{n-1} T(A_k) \right) \right]. \end{aligned}$$

Note that the difference between (5) and the right side of (4) can be rewritten as

$$\left[K_0 \setminus \bigcup_{k=0}^{n-2} A_k \right] \cap \left[\bigcup_{j \geq 1} 2^{-j} \left(\bigcup_{k=0}^{n-2} T(A_k) \right) \right] \setminus A_{n-1},$$

which, because of (5), is empty.

3. Wavelet sets in $[-1, 1]^d$

It is a consequence of the geometrical characterization in (1) that a wavelet set can not be contained in a closed proper sub-cube of $I \equiv [-1, 1]^d$. In fact, we have the following result.

PROPOSITION 4. *For any $\alpha < 1$, a wavelet set K can not be contained in $[-\alpha, \alpha]^d$.*

PROOF. Suppose that $K \subseteq [-\alpha, \alpha]^d$ is a wavelet set. By (1), the integral translates of K will cover $\widehat{\mathbb{R}}^d$, i.e.,

$$\dot{\bigcup}_{k \in \mathbb{Z}^d} (K + k) = \widehat{\mathbb{R}}^d,$$

where $\dot{\bigcup}$ designates disjoint union.

Observe that, for each $i = 1, \dots, d$, the union of all translates $K + (k_1, \dots, k_d)$ with $k_i \neq 0$ leaves out the band $B_i \equiv \{(x_1, x_2, \dots, x_d) \in \mathbb{R}^d : \alpha - 1 < x_i < 1 - \alpha\}$. That is, for each $i = 1, 2, \dots, d$,

$$\dot{\bigcup}_{k_i \neq 0} (K + k) \cap B_i = \emptyset.$$

Therefore,

$$\dot{\bigcup}_{k \neq (0, \dots, 0)} (K + k) \cap B = \emptyset,$$

where $B \equiv \bigcap_{i=1}^d B_i = (\alpha - 1, 1 - \alpha)^d$. This clearly implies that $B \subseteq K$ and that $\emptyset \neq B \subseteq K \cap 2K$, a contradiction to the second condition in (1). \square

To construct wavelet sets K by the neighborhood-mapping method in section 2, we note that N must always be at least $\frac{1}{2}$ since K_0 has to be τ -congruent to $[-\frac{1}{2}, \frac{1}{2}]^d$ and contained in $[-N, N]^d$. Further, if $K \subseteq [-1, 1]^d$, then it is also necessary that $N = \frac{1}{2}$. In fact, it is impossible to map a point in $[-N, N]^d \setminus [-\frac{1}{2}, \frac{1}{2}]^d$ into $[-1, 1]^d \setminus [-N, N]^d$ by \mathbb{Z}^d -translation; and this is a requirement for the map T . Hence, the only possible K_0 in this case is $Q \equiv [-\frac{1}{2}, \frac{1}{2}]^d$, and the function T is a mapping from Q to $2Q \setminus \overline{Q}$ defined by $T(x) = x + n_x$ for some $n_x \in \Lambda \equiv \{-1, 0, 1\}^d \setminus \{(0, \dots, 0)\}$.

In this case, the A_i s can be obtained by successively computing $\frac{1}{2}T$ of $\frac{1}{2}Q$, i.e.,

$$(6) \quad \begin{aligned} A_0 &= \frac{1}{2}Q, \\ A_1 &= \frac{1}{2}T\left(\frac{1}{2}Q\right), \\ A_2 &= \frac{1}{2}T\left(\frac{1}{2}T\left(\frac{1}{2}Q\right)\right), \\ &\vdots \\ A_k &= \frac{1}{2}S^k(Q), \end{aligned}$$

where $S(x) \equiv T\left(\frac{1}{2}x\right)$. First, it is clear by definition that $A_0 = \frac{1}{2}Q = \left[-\frac{1}{4}, \frac{1}{4}\right]^d$. Since $2^{-j}T(E) \subseteq \left[-\frac{1}{4}, \frac{1}{4}\right]^d$ for all $j \geq 2$, $E \subseteq Q$, we also have

$$A_n = \left[K_0 \setminus \bigcup_{k=0}^{n-1} A_k \right] \cap \left[\frac{1}{2}T(A_{n-1}) \right].$$

It is then left to show that $\frac{1}{2}T(A_{n-1})$ does not intersect $\bigcup_{k=0}^{n-1} A_k$, so that $A_n = \frac{1}{2}T(A_{n-1})$. This in turn follows from the limitation of the map T , that it can only translate by at most 1 both horizontally and vertically, and an induction argument showing that each A_n is contained in a “square annulus” $B_n = [-b_n, b_n]^2 \setminus [-a_n, a_n]^2$, where $b_n - a_n = 2^{-n-2}$ and the B_n s are mutually disjoint.

Therefore, combining (3) and (6), we obtain

$$(7) \quad K = \left[Q \setminus \bigcup_{k=0}^{\infty} \frac{1}{2}S^k(Q) \right] \cup \left[\bigcup_{k=0}^{\infty} \frac{1}{2}T\left(\frac{1}{2}S^k(Q)\right) \right].$$

When T is regular enough, e.g., $|\cup_{k \in \Lambda} \partial \Omega_k| = 0$ where $\Omega_k \equiv \{x \in Q : T(x) = x + n_x\}$, we also have

$$K = \left[Q \setminus \bigcup_{k=0}^{\infty} \frac{1}{2}\overline{S^k(Q)} \right] \cup \left[\bigcup_{k=1}^{\infty} \overline{S^k(Q)} \right] \quad \text{a.e.}$$

In light of this reformulation of the wavelet set K in the case $K_0 = Q$, and because of our approach in section 4, we set

$$(8) \quad M \equiv \bigcup_{k=1}^{\infty} \overline{S^k(Q)} \subseteq \overline{2Q \setminus Q},$$

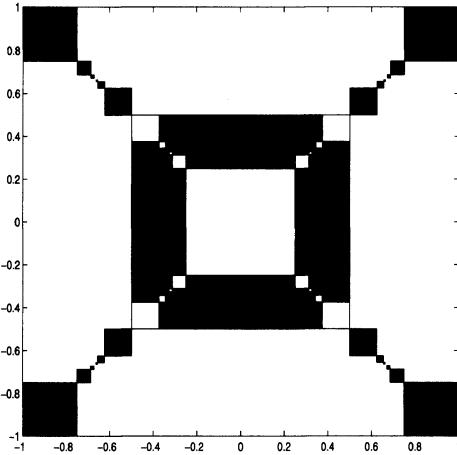
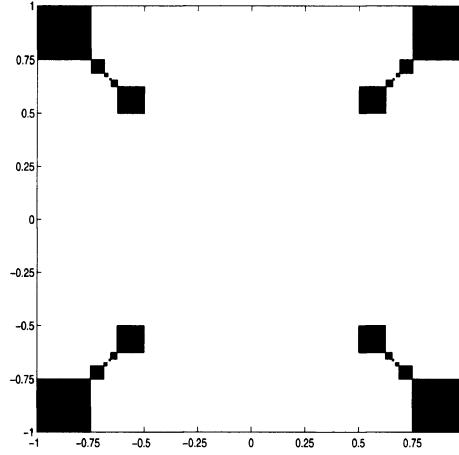
where, of course, S and M both depend on the map T .

EXAMPLE 5. Consider the map $T_1 : \left[-\frac{1}{2}, \frac{1}{2}\right]^2 \rightarrow [-1, 1]^2 \setminus \left[-\frac{1}{2}, \frac{1}{2}\right]^2$ defined by

$$T_1(x, y) = (x, y) - (\operatorname{sgn} x, \operatorname{sgn} y);$$

and write $S_1 \equiv S_{T_1}$ and $M_1 \equiv M_{T_1}$.

Note that using this map T_1 , the wavelet set constructed by the above method, shown in Figure 1, is a generalization to \mathbb{R}^2 of the wavelet set $[-1, -\frac{1}{2}] \cup [\frac{1}{2}, 1]$ associated with the Shannon or Littlewood-Paley wavelet. Observe also that the map $S_1 \equiv T_1\left(\frac{1}{2}\cdot\right)$ is a similarity map from one quadrant to the opposite with ratio $\frac{1}{2}$ when restricted to each quadrant. The resulting set M_1 in Figure 2 is “self-similar” at the four limit points $F_1 \equiv \{(\pm\frac{2}{3}, \pm\frac{2}{3})\}$ in the sense that there is a compact

FIGURE 1. The set K_7 FIGURE 2. The set $\bigcup_{k=1}^6 \overline{S_1^k(Q)}$

subset P_1 , viz., $P_1 = \left([-1, -\frac{3}{4}] \cup [\frac{3}{4}, 1]\right)^2$ consisting of the four largest squares in the outer corners of M_1 , such that

$$M_1 = S_1(M_1) \cup P_1 \quad \text{and} \quad F_1 = S_1(F_1).$$

4. A fractal set arising from the neighborhood-mapping construction

Since wavelet sets have to satisfy the translation and dilation tiling conditions in (1 a) and (1 b), it is not surprising that all known examples of wavelet sets appear to be fractal-like or self-similar. However, such examples (in $\widehat{\mathbb{R}}^d$) can never be fractal sets. The reason is that they always have measure 1, and hence Hausdorff dimension d . Besides, due to the dyadic nature of the neighborhood-mapping method, the constructed wavelet sets, which we call the *neighborhood-mapping wavelet sets*, do not have fractal boundaries. Nevertheless, we shall construct a fractal (self-affine) set $\mathcal{B} \subseteq \widehat{\mathbb{R}}^2$ arising from all of the neighborhood-mapping wavelet sets in $[-1, 1]^2$.

We first define a set function \mathcal{T} mapping a given set $E \subseteq Q$ to the union

$$\mathcal{T}(E) \equiv \bigcup \{T(E) | T : Q \rightarrow 2Q \setminus Q \text{ is a measurable 1-1 } \mathbb{Z}\text{-translated map}\},$$

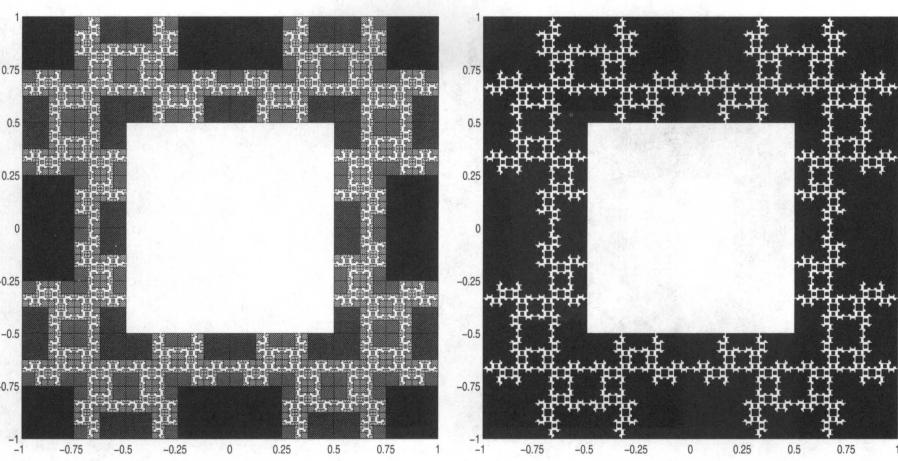
and define

$$\begin{aligned} \mathcal{S}(E) &\equiv \mathcal{T}\left(\frac{1}{2}E\right), \\ \mathcal{M} &\equiv \bigcup_{k=1}^{\infty} \mathcal{S}^k(Q). \end{aligned}$$

Observe that since T maps $[-\frac{1}{2}, \frac{1}{2}]^2$ into $[-1, 1]^2 \setminus [-\frac{1}{2}, \frac{1}{2}]^2$ by integer translation, then $T(x, y) = (x - \varepsilon_1 \operatorname{sgn} x, y - \varepsilon_2 \operatorname{sgn} y)$ for some $(\varepsilon_1, \varepsilon_2) \in \{0, 1\}^2 \setminus \{(0, 0)\}$. Thus,

$$\mathcal{T}(E) = \{(x - \varepsilon_1 \operatorname{sgn} x, y - \varepsilon_2 \operatorname{sgn} y) : (x, y) \in E, (\varepsilon_1, \varepsilon_2) \in \{0, 1\}^2 \setminus \{(0, 0)\}\}.$$

Clearly, for any such map T , its associated set $M = M_T$ defined in (8), and any set $E \subseteq Q$, we have the inclusions $T(E) \subseteq \mathcal{T}(E)$ and $M_T \subseteq \mathcal{M}$. Figure 3 shows an approximation of \mathcal{M} .

FIGURE 3. The set $\bigcup_{k=1}^6 \mathcal{S}^k(Q) \subseteq \mathcal{M}$

Moreover, the limit points of M lie outside the set \mathcal{M} , i.e., inside the white jagged line $\mathcal{B} \equiv \overline{(2Q \setminus Q)} \setminus \mathcal{M}$ in Figure 3. It can be shown that \mathcal{B} is a fractal set with Hausdorff dimension $\log 3 / \log 2$. In fact, the set \mathcal{B} can be viewed as twelve self-similar sets of the same dimension. Consider, for instance, the set $\mathcal{B}' = \mathcal{B} \cap [\frac{1}{2}, 1]^2 - (\frac{1}{2}, \frac{1}{2})$. We shall illustrate, see Figures 4 and 5, and also prove that

$$(9) \quad \mathcal{B}' = \bigcup_{i=1}^3 \varphi_i(\mathcal{B}'),$$

where

$$\begin{aligned} \varphi_1(x, y) &= \frac{1}{2} \left(\frac{1}{2} - y, \frac{1}{2} - x \right) = \begin{pmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \frac{1}{4} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ \varphi_2(x, y) &= \frac{1}{2} \left(\frac{1}{2} + y, \frac{1}{2} - x \right) = \begin{pmatrix} 0 & \frac{1}{2} \\ -\frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \frac{1}{4} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ \varphi_3(x, y) &= \frac{1}{2} \left(\frac{1}{2} - y, \frac{1}{2} + x \right) = \begin{pmatrix} 0 & -\frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \frac{1}{4} \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \end{aligned}$$

Since φ_1, φ_2 and φ_3 are similarity maps with ratio $1/2$ and the disjoint union $\bigcup_{i=1}^3 \varphi_i \left((0, \frac{1}{2})^2 \right)$ is contained in $(0, \frac{1}{2})^2$, we conclude by the Hutchinson Theorem(Theorem 6) that the Hausdorff dimension of \mathcal{B}' is $\log 3 / \log 2$. Therefore $\dim_H(\mathcal{B}) = \log 3 / \log 2$ since \mathcal{B} is the union of twelve reflections of $\mathcal{B} \cap [\frac{1}{2}, 1]^2$.

THEOREM 6. *Let φ_i be contractions on \mathbb{R}^d with ratios $c_i < 1$ for each $i = 1, \dots, m$. Then there exists a unique non-empty compact set F that satisfies*

$$F = \bigcup_{i=1}^m \varphi_i(F).$$

Furthermore, suppose each φ_i is a similarity with ratio $c_i < 1$, and the φ_i s satisfy the open set condition, which says that there exists a non-empty bounded open set V such that

$$\forall i = 1, \dots, m, \quad \varphi_i(V) \subseteq V$$

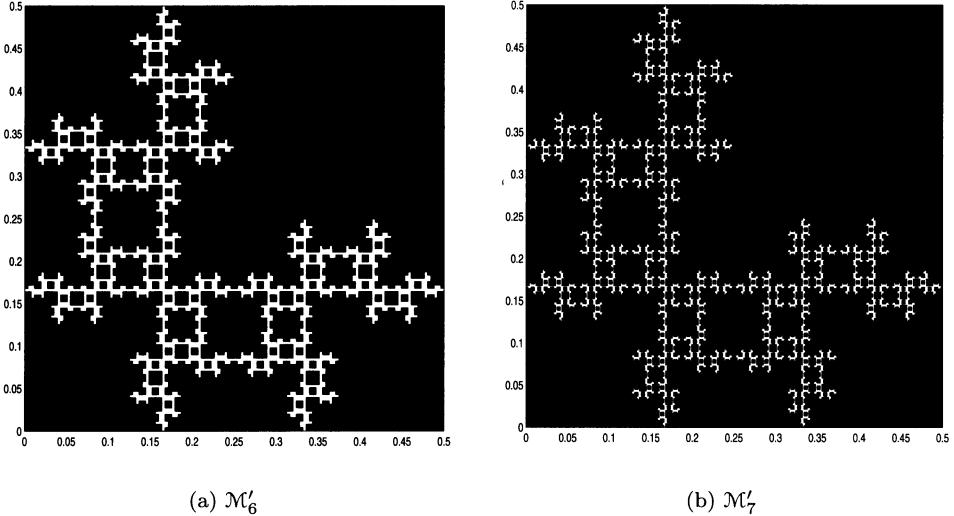


FIGURE 4. $\mathcal{M}'_n = \bigcup_{k=1}^n \mathcal{S}^k(Q) \cap \left[\frac{1}{2}, 1\right]^2 - \left(\frac{1}{2}, \frac{1}{2}\right)$

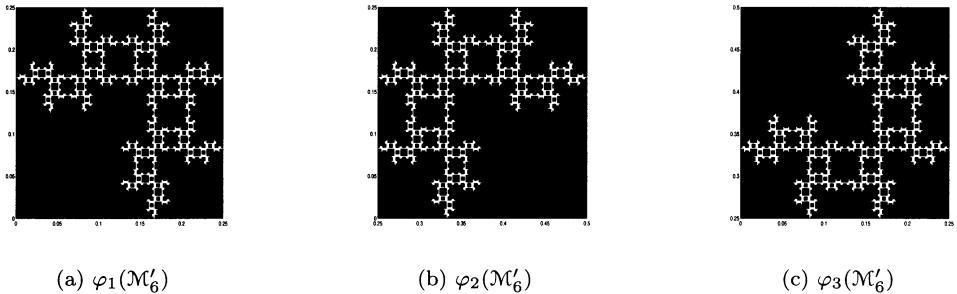


FIGURE 5. The images of \mathcal{M}'_6 under each map $\varphi_i, i = 1, 2, 3$

and

$$\forall i \neq j, \quad \varphi_i(V) \cap \varphi_j(V) = \emptyset.$$

Then the Hausdorff dimension of the invariant set F is s , where s is given by

$$\sum_{i=1}^m c_i^s = 1.$$

In order to prove (9) and to prove that \mathcal{B} is the union of certain reflections of $\mathcal{B} \cap \left[\frac{1}{2}, 1\right]^2$, a characterization of \mathcal{M} , and hence \mathcal{B} , is required. Since this involves binary representations, we set up the following conventions. For a given point $x \in [-1, 1]$, let $\pm.x^{(1)}x^{(2)}\dots$ denote its binary representation. When ambiguity arises, we pick the representation with infinitely many 1's. For example, $\frac{1}{4}$ is written as .00111... instead of .01. The following lemma characterizes all points in $\mathcal{M} \equiv \bigcup_{k=1}^{\infty} \mathcal{S}^k(Q)$.

LEMMA 7. Let (x, y) be a point in the “square annulus” $\overline{2Q \setminus Q}$. Then (x, y) belongs to \mathcal{M} if and only if there exists $k \in \mathbb{N}$ such that $x^{(k)} = x^{(k+1)}$ and $y^{(k)} = y^{(k+1)}$.

PROOF. By induction on $k \in \mathbb{N}$, we prove the stronger statement that

$$(10) \quad (x, y) \in \mathcal{S}^k(Q) \quad \text{if and only if} \quad x^{(k)} = x^{(k+1)} \text{ and } y^{(k)} = y^{(k+1)}.$$

First, let us observe that if $x^{(k)} = x^{(k+1)}$, then $(x + n)^{(k)} = (x + n)^{(k+1)}$ for any $n \in \mathbb{N}$. If $(x, y) \in \mathcal{S}\left([- \frac{1}{2}, \frac{1}{2}]^2\right) = \mathcal{T}\left([- \frac{1}{4}, \frac{1}{4}]^2\right)$, then $x^{(1)} = x^{(2)}$ and $y^{(1)} = y^{(2)}$ because the first two digits of both x and y are zeros when $(x, y) \in [- \frac{1}{4}, \frac{1}{4}]^2$ and \mathcal{T} translates only by integers. Conversely, if $x^{(1)} = x^{(2)}$ and $y^{(1)} = y^{(2)}$, then $(x, y) - (x^{(1)} \operatorname{sgn} x, y^{(1)} \operatorname{sgn} y)$ lies in $[- \frac{1}{4}, \frac{1}{4}]^2$ and so $(x, y) \in \mathcal{T}(\frac{1}{2}Q)$.

Now let us assume (10) and let $(u, v) \in \mathcal{S}^{k+1}(Q)$. So $(u, v) = \frac{1}{2}(x, y) + (a, b)$ for some $(a, b) \in \Lambda$ and $(x, y) \in \mathcal{S}^k(Q)$. Hence by the “only if” part of (10), $(\frac{x}{2})^{(k+1)} = x^{(k)} = x^{(k+1)} = (\frac{x}{2})^{(k+2)}$ and similarly for y . Therefore, $u^{(k+1)} = u^{(k+2)}$ and $v^{(k+1)} = v^{(k+2)}$.

To prove the “if” part of (P_{k+1}) , assume $u^{(k+1)} = u^{(k+2)}$ and $v^{(k+1)} = v^{(k+2)}$. Since $(u, v) \in [-1, 1]^2$, there is $(a, b) \in \Lambda$ such that $(u, v) - (a, b)$ is in $[- \frac{1}{2}, \frac{1}{2}]^2$. It is then clear that $(2(u - a))^{(k)} = (u - a)^{(k+1)} = (u - a)^{(k+2)} = (2(u - a))^{(k+1)}$ and the similar statement holds for v , hence $2[(u, v) - (a, b)] \in \mathcal{S}^k(Q)$. Therefore $(u, v) \in \frac{1}{2}\mathcal{S}^k(Q) + (a, b) \subseteq \mathcal{S}^{k+1}(Q)$. The proof is complete. \square

To sum up, we have proved that

$$(11) \quad \mathcal{M} \equiv \bigcup_{k=1}^{\infty} \mathcal{S}^k(Q) = \bigcup_{k=1}^{\infty} \left\{ (x, y) \in \overline{2Q \setminus Q} : x^{(k)} = x^{(k+1)} \text{ and } y^{(k)} = y^{(k+1)} \right\}$$

We are now ready to prove (9). It follows from Lemma 7 that $(x, y) \in \mathcal{B}'$ if and only if $x^{(1)} = y^{(1)} = 0$, $x^{(2)} \neq 1$ or $y^{(2)} \neq 1$ and for all $k \geq 2$, $x^{(k)} \neq x^{(k+1)}$ or $y^{(k)} \neq y^{(k+1)}$. For brevity, let us consider only the set equation

$$(12) \quad \mathcal{B}' \cap \left[0, \frac{1}{4}\right]^2 = \varphi_1(\mathcal{B}').$$

When $(x, y) \in \mathcal{B}'$ and $\varphi_1(x, y) \equiv (x_1, y_1)$, we have $x_1^{(k)} = y_1^{(k)} = 0$ for $k = 1, 2$ and $x_1^{(k)} = 1 - x^{(k-1)}$, $y_1^{(k)} = 1 - y^{(k-1)}$ for all $k \geq 3$. Therefore, for $k \geq 2$ $x_1^{(k)} \neq x_1^{(k+1)}$ or $y_1^{(k)} \neq y_1^{(k+1)}$ with the first two digits of both x_1 and y_1 vanish and hence $(x_1, y_1) \in \mathcal{B}' \cap [0, \frac{1}{4}]^2$. The converse arguments are also valid and the equality (12) is proved. The other two cases, e.g., $\mathcal{B}' \cap ([\frac{1}{4}, \frac{1}{2}] \times [0, \frac{1}{4}]) = \varphi_2(\mathcal{B}')$, are similar and the proof will be omitted.

THEOREM 8. Let $T : Q \rightarrow 2Q \setminus Q$ be measurable injective integer-translated map and $S = T(\frac{1}{2}\cdot)$. The limit points of $M = \bigcup_{k=1}^{\infty} \overline{\mathcal{S}^k(Q)}$ are contained in the set $\overline{2Q \setminus Q} \setminus \mathcal{M}$.

PROOF. Let $(x, y) \in Q$. Then for each $k \in \mathbb{N}_0$ there exists $\varepsilon_k = (\zeta_k, \eta_k)$ in Λ such that

$$\begin{aligned} S(x, y) &= \frac{1}{2}(x, y) + \varepsilon_0 \\ S^2(x, y) &= \frac{1}{2^2}(x, y) + \frac{\varepsilon_0}{2} + \varepsilon_1 \\ &\vdots \\ S^{n+1}(x, y) &= \frac{(x, y)}{2^{n+1}} + \frac{\varepsilon_0}{2^n} + \frac{\varepsilon_1}{2^{n-1}} + \cdots + \varepsilon_n. \end{aligned}$$

Since S maps $[-1, 1]^2$ into $[-1, 1]^2 \setminus [-\frac{1}{2}, \frac{1}{2}]^2$, there must exist strictly increasing sequences of integers

$$\{p : \zeta_p \neq 0\} = \{p_1 < p_2 < \dots\}, \quad \{q : \eta_q \neq 0\} = \{q_1 < q_2 < \dots\}$$

and, for a given $n \in \mathbb{N}$, unique indices $i = i(n)$ and $j = j(n)$ for which p_i is the largest p less than or equal to n such that $\zeta_p \neq 0$ and q_j is the largest q less than or equal to n such that $\eta_q \neq 0$. That is, $p_i \leq n < p_{i+1}$ and $q_j \leq n < q_{j+1}$. So

$$\begin{aligned} (\tilde{a}_n, \tilde{b}_n) &\equiv S^{n+1}(x, y) \\ &= \left(\frac{x}{2^{n+1}} + \frac{\zeta_{p_1}}{2^{n-p_1}} + \cdots + \frac{\zeta_{p_i}}{2^{n-p_i}}, \frac{y}{2^{n+1}} + \frac{\eta_{q_1}}{2^{n-q_1}} + \cdots + \frac{\eta_{q_j}}{2^{n-q_j}} \right). \end{aligned}$$

Furthermore, $\zeta_{p_k} = (-1)^k \operatorname{sgn} x$, $\eta_{q_k} = (-1)^k \operatorname{sgn} y$ and $\{p : \zeta_p \neq 0\} \cup \{q : \eta_q \neq 0\} = \{0, 1, 2, \dots\} \equiv \mathbb{N}_0$ because (ζ_k, η_k) can never be $(0, 0)$. For brevity, we suppose $x, y \geq 0$. So

$$a_n \equiv \tilde{a}_n - \frac{x}{2^{n+1}} = -\frac{1}{2^{n-p_1}} + \frac{1}{2^{n-p_2}} + \cdots + \frac{(-1)^i}{2^{n-p_i}}$$

and

$$b_n \equiv \tilde{b}_n - \frac{y}{2^{n+1}} = -\frac{1}{2^{n-q_1}} + \frac{1}{2^{n-q_2}} + \cdots + \frac{(-1)^j}{2^{n-q_j}}.$$

In binary representation,

$$a_n \equiv \tilde{a}_n - \frac{x_0}{2^{n+1}} = \begin{cases} +\underbrace{0 \dots 0}_{n-p_i} \underbrace{1 \dots 1}_{p_i-p_{i-1}} \dots \dots \underbrace{0 \dots 01 \dots 1}_{p_3-p_2 \ p_2-p_1} & \text{if } i \text{ is even} \\ -\underbrace{0 \dots 0}_{n-p_i} \underbrace{1 \dots 1}_{p_i-p_{i-1}} \dots \dots \underbrace{1 \dots 10 \dots 0}_1 & \text{if } i \text{ is odd} \end{cases}$$

and

$$b_n \equiv \tilde{b}_n - \frac{y_0}{2^{n+1}} = \begin{cases} +\underbrace{0 \dots 0}_{n-q_j} \underbrace{1 \dots 1}_{q_j-q_{j-1}} \dots \dots \underbrace{0 \dots 01 \dots 1}_{q_3-q_2 \ q_2-q_1} & \text{if } j \text{ is even} \\ -\underbrace{0 \dots 0}_{n-q_j} \underbrace{1 \dots 1}_{q_j-q_{j-1}} \dots \dots \underbrace{1 \dots 10 \dots 0}_1 & \text{if } j \text{ is odd} \end{cases}$$

Let (a, b) be a limit point of $\{S^{n+1}(x, y)\}_{n \in \mathbb{N}}$ and hence of $\{(a_n, b_n)\}_{n \in \mathbb{N}}$. Clearly, $(a, b) \in \overline{2Q \setminus Q}$. We will prove that $(a, b) \notin \mathcal{M}$ by showing that for each $k \in \mathbb{N}$, either $a^{(k)} \neq a^{(k+1)}$ or $b^{(k)} \neq b^{(k+1)}$.

For any $N \in \mathbb{N}$, there exists an integer $n > N$ for which

$$|a_n - a| \leq \frac{1}{2^N} \text{ and } |b_n - b| \leq \frac{1}{2^N}.$$

This means exactly that the first N digits of a_n coincide with those of a . Of course, the same statement also holds for b_n and b . This gives rise to a sequence $\{n_l\}_{l=1}^\infty$ of strictly increasing integers such that $|a_{n_{l+1}} - a| \leq \frac{1}{2^{n_l}}$ for all l .

It then suffices to prove that for each integer $n = 2, 3, \dots$,

$$(M_n) \quad \forall k = 1, 2, \dots, n-1, \text{ either } a_n^{(k)} \neq a_n^{(k+1)} \text{ or } b_n^{(k)} \neq b_n^{(k+1)}.$$

When $n = 2$, we only have to show that $a_2^{(1)} \neq a_2^{(2)}$ or $b_2^{(1)} \neq b_2^{(2)}$. In fact, it can be checked that if $p_m = 1$ for some m , then $a_2^{(1)} \neq a_2^{(2)}$. Since either p_m or q_m must be 1 for some m , the case $n = 2$ is done. Now assume (M_n) with $p_{i(n)} \leq n < p_{i(n)+1}$ and $q_{j(n)} \leq n < q_{j(n)+1}$. When $p_{i(n)+1} = n+1$, one has $i(n+1) = i(n) + 1$ and

$$\begin{aligned} n+1-p_{i(n+1)} &= 0, \\ p_{i(n+1)}-p_{i(n+1)-1} &= n+1-p_{i(n)}, \\ p_{i(n+1)-1}-p_{i(n+1)-2} &= p_{i(n)}-p_{i(n)-1}, \\ &\vdots \end{aligned}$$

Thus a_{n+1} equals the “reverse” ($0 \rightarrow 1$ and $1 \rightarrow 0$) of a_n shifted one place to the right with 1 as the first digit provided that $p_{i(n)+1} = n+1$. If, in addition, $p_{i(n)} = n$ then the second digit will be 0. Since either $p_{i(n)+1} = n+1$ or $q_{j(n)+1} = n+1$, (M_{n+1}) clearly holds for $k = 2, \dots, n$. (M_{n+1}) is also the case for $k = 1$ because either $p_{i(n)}$ or $q_{j(n)}$ has to be n . \square

5. Generality of the neighborhood-mapping construction

It is natural to investigate on how general the neighborhood-mapping construction of wavelet sets is. In fact, the algorithm can construct every wavelet set in the annulus $[-1, 1]^2 \setminus [-\frac{1}{4}, \frac{1}{4}]^2$. We prove the following theorem.

THEOREM 9. *Every wavelet set in $[-1, 1]^2 \setminus [-\frac{1}{4}, \frac{1}{4}]^2$ can be constructed by means of the neighborhood-mapping method.*

PROOF. Let K be a wavelet set contained in $[-1, 1]^2 \setminus [-\frac{1}{4}, \frac{1}{4}]^2$. Define the integer-translated map $T : [-\frac{1}{2}, \frac{1}{2}]^2 \rightarrow [-1, 1]^2 \setminus [-\frac{1}{2}, \frac{1}{2}]^2$ by

$$T(x) = \begin{cases} x - n & \text{if } x \in K + n, \\ x - \operatorname{sgn}(x) & \text{if } x \in K. \end{cases}$$

T is well-defined almost everywhere because $\{K + n : n \in \mathbb{Z}^d\}$ is a partition of \mathbb{R}^d a.e. Since the domain of T is a unit cube, it is clearly injective. Observe that for almost every $x \in Q = [-\frac{1}{2}, \frac{1}{2}]^2$,

$$x \in K \text{ if and only if } T(x) \notin K$$

and for almost every $x \in 2Q \setminus Q$,

$$x \in K \text{ if and only if } \frac{1}{2}x \notin K.$$

From the neighborhood mapping construction with $S = T(\frac{1}{2}\cdot)$

$$\tilde{K} \equiv \left[Q \setminus \bigcup_{k=0}^{\infty} \frac{1}{2}S^k(Q) \right] \dot{\cup} \left[\bigcup_{k=0}^{\infty} T\left(\frac{1}{2}S^k(Q)\right) \right]$$

is a wavelet set. In order to prove that $K = \tilde{K}$ a.e., it suffices to show that

$$(13) \quad K \cap Q \subseteq Q \setminus \bigcup_{k=0}^{\infty} \frac{1}{2} S^k(Q) \quad \text{a.e.}$$

and

$$(14) \quad K \setminus Q \subseteq \bigcup_{k=1}^{\infty} S^k(Q) \quad \text{a.e.}$$

In fact, if $K \subseteq \tilde{K}$ and $|K| = |\tilde{K}| = 1$ then $K = \tilde{K}$ a.e.

Since $\frac{1}{2} S^k(Q) \subseteq Q$, it is easy to check that (13) is equivalent to saying that $\bigcup_{k=0}^{\infty} (K \cap \frac{1}{2} S^k(Q)) = \emptyset$ a.e. We show that $K \cap \frac{1}{2} S^k(Q) = \emptyset$ a.e. by induction on k . The case $k = 0$ is clear by the assumption. If $\frac{1}{2} S^k(Q) \cap K = \emptyset$ a.e., then $S^{k+1}(Q) = T(\frac{1}{2} S^k(Q)) \subseteq K$ a.e. and so $\frac{1}{2} S^{k+1}(Q) \cap K = \emptyset$ a.e.

In order to prove (14), it is sufficient to show, for almost everywhere, that $\bigcup_{k=1}^{\infty} (K \setminus Q) \cap S^k(Q) = (K \setminus Q) \cap (\bigcup_{k=1}^{\infty} S^k(Q))$ is a subset of $\bigcup_{k=1}^{\infty} S^k(Q)$ because $(K \setminus Q) \setminus (\bigcup_{k=1}^{\infty} S^k(Q)) \subseteq 2Q \setminus Q \setminus \mathcal{M} = \emptyset$. We first make $\bigcup_{k=1}^{\infty} S^k(Q)$ a disjoint union by letting

$$Q_1 = S(Q),$$

$$Q_2 = S^2(Q) \setminus S(Q),$$

⋮

$$Q_k = S^k(Q) \setminus \bigcup_{l < k} S^l(Q),$$

and then show by induction on k that $(K \setminus Q) \cap Q_k \subseteq S^k(Q)$ a.e.

Let $(x, y) \in (K \setminus Q) \cap S(Q)$. Then $x^{(1)} = x^{(2)}$ and $y^{(1)} = y^{(2)}$. From the surjectivity of T onto $(2Q \setminus Q) \cap K = K \setminus Q$, $(x, y) = T(\frac{1}{2}(u, v))$ for some $(u, v) \in 2Q$. Since the first two binary digits of $u/2$ and $v/2$ are equal, i.e., $(u/2)^{(1)} = (u/2)^{(2)}$ and $(v/2)^{(1)} = (v/2)^{(2)}$, and $-\frac{1}{2} \leq \frac{u}{2}, \frac{v}{2} \leq \frac{1}{2}$, the first two digits of both $u/2$ and $v/2$ must be zero. Therefore $-\frac{1}{4} \leq \frac{u}{2}, \frac{v}{2} \leq \frac{1}{4}$ and $(x, y) \in T(\frac{1}{2}Q) = S(Q)$.

Assume that $(K \setminus Q) \cap Q_k \subseteq S^k(Q)$ a.e. and let $(x, y) \in (K \setminus Q) \cap Q_{k+1}$ be given. By the definition of Q_{k+1} , $k+1$ is the first (smallest) i such that $x^{(i)} = x^{(i+1)}$, $y^{(i)} = y^{(i+1)}$, and there exists $(u, v) \in 2Q$ for which $(x, y) = T(\frac{1}{2}(u, v))$. In fact $(u, v) \in K$ since K tiles \mathbb{R}^2 by both integer translation and dyadic dilation and $K \subseteq [-1, 1]^2 \setminus [-\frac{1}{4}, \frac{1}{4}]^2$. Besides $(u, v) \in Q_k$. To verify that $(u, v) \notin Q = [-\frac{1}{2}, \frac{1}{2}]^2$, one observes that (x, y) not being in $Q_1 = S(Q)$ implies that $(u, v) = 2T^{-1}(x, y) \notin Q$. Thus $(u, v) \in (K \setminus Q) \cap Q_k$ which is contained in $S^k(Q)$ by the assumption. Consequently $(x, y) = T(\frac{1}{2}(u, v)) \in S^{k+1}(Q)$ as desired. \square

COROLLARY 10. *For every wavelet set K in $[-1, 1]^2$,*

$$|K \cap [-c, c]^2| > 0 \quad \text{for all } c > \frac{1}{4}.$$

PROOF. If $|K \cap [-\frac{1}{4}, \frac{1}{4}]^2| > 0$, then we are done. Otherwise, by Theorem 9, the wavelet set K can be constructed by the neighborhood-mapping method. Due to the property of the map T , it can be shown that A_2 is the only A_n that intersects the “square annulus” $[-\frac{5}{16}, \frac{5}{16}] \setminus [-\frac{1}{4}, \frac{1}{4}]$. Again, since T is \mathbb{Z}^2 -translated map, it

is not possible for A_2 to cover $[-c, c]^2 \setminus [-\frac{1}{4}, \frac{1}{4}]^2$ for any $c > \frac{1}{4}$. This completes the proof. \square

6. Analysis on Baggett, Medina, and Merrill's wavelet set construction

In [BMM99], Baggett, Medina, and Merrill introduced a general construction of wavelet sets and subspace wavelet sets. For brevity, they wrote $x \equiv y$ if $x - y = n$ for some $n \in \mathbb{Z}^d$ and introduced the notion of *complementary pair*.

DEFINITION 11. Let $E \subseteq \widehat{\mathbb{R}}^d$ be invariant under multiplication by $1/2$, i.e., $\frac{1}{2}E \subseteq E$. By a complementary pair for E , we mean a pair (R, R') of measurable injective maps $R : Q \rightarrow \widehat{\mathbb{R}}^d$ and $R' : E \rightarrow E$ satisfying

- (1) $R(Q) \subseteq E$ and $R'(E) \subseteq E \setminus R(Q)$
- (2) $2R(x) \equiv x$ and $2R'(x) \equiv x$ a.e.
- (3) $E = \bigcup_{j \geq 0} R'^j(R(Q))$.

In their paper [BMM99], Baggett et al. have proven the following results.

- (A) If (R, R') is a complementary pair for a set $E \subseteq \widehat{\mathbb{R}}^d$, then $W = 2E \setminus E$ is a *subspace* wavelet set. Moreover, if, in addition, $\bigcup_{j \in \mathbb{Z}} 2^j(E)$ is a neighborhood of the origin, then W is a wavelet set for $L^2(\mathbb{R}^d)$. (Note that $\bigcup_{j \in \mathbb{Z}} 2^j(E) = \bigcup_{j \in \mathbb{Z}} 2^j(W)$.)
- (B) If W is a *subspace* wavelet set, then there exists a complementary pair (R, R') for the set $E = \bigcup_{j < 0} 2^j(W)$. In fact, R' can always be taken equal to $1/2$.
- (C) If a map $R : Q \rightarrow \widehat{\mathbb{R}}^d$ satisfies $2R(x) \equiv x$ a.e., then there exists a set E and a map $R' : E \rightarrow E \setminus R(Q)$ such that (R, R') is a complementary pair for E . (The set E and the map R' are by no means uniquely determined by R . Therefore the resulting wavelet set $W = 2E \setminus E$ is not unique.)

Since $E \subseteq 2E$, we see in (A) and (B) that $W = 2E \setminus E$ if and only if $E = \bigcup_{j < 0} 2^j(W)$.

In part (C), not every such a map R will result in a wavelet set for L^2 of the whole of \mathbb{R}^d . For instance, let us take

$$R(x) = \begin{cases} (x+1)/2 & \text{if } x \geq 0 \\ (x-1)/2 & \text{if } x < 0. \end{cases}$$

Then $R'(x)$ can be defined as $\frac{1}{2}x$, and

$$E = \bigcup_{j \geq 0} \frac{1}{2^j} R(Q) = \bigcup_{j \geq 0} \frac{1}{2^j} \left(\left[-\frac{3}{4}, -\frac{1}{2} \right] \cup \left[\frac{1}{2}, \frac{3}{4} \right] \right).$$

The set $W = 2E \setminus E = \left[-\frac{3}{2}, -1 \right] \cup \left[1, \frac{3}{2} \right]$ is only a proper subspace wavelet set.

Let us first consider how the neighborhood-mapping construction of wavelet sets in $[-1, 1]^d$ fits into this construction. In this case, we have a measurable injective \mathbb{Z}^d -translated map $T : Q \rightarrow 2Q \setminus Q$ and $S(x) = T(\frac{1}{2}x)$. Then the set

$$K = \left[Q \setminus \bigcup_{k=0}^{\infty} \frac{1}{2} S^k(Q) \right] \cup \left[\bigcup_{k=1}^{\infty} S^k(Q) \right]$$

was shown to be a wavelet set. If we let $E = \bigcup_{k \geq 0} \frac{1}{2} S^k(Q)$, it is clear that

$$2E \setminus E = \left[Q \cup \bigcup_{k \geq 1} S^k(Q) \right] \setminus \left[\bigcup_{k \geq 0} \frac{1}{2} S^k(Q) \right] = K.$$

Moreover E can be written as the union $\bigcup_{j \geq 0} R'^j(R(Q))$ where $R'(x) = \frac{1}{2}T(x)$ and $R(x) = \frac{1}{2}x$. Since, in addition, $R(Q) = \frac{1}{2}Q \subseteq E$ and $R'(E) = E \setminus \frac{1}{2}Q$ the pair $(R, R') = (\frac{1}{2}, \frac{1}{2}T)$ is a complementary pair for E . But E has yet another complementary pair. Define $\tilde{R} : Q \rightarrow E$ by letting

$$\tilde{R}(x) = \begin{cases} T(x)/2 & \text{if } x \in \bigcup_{k=0}^{\infty} \frac{1}{2} S^k(Q) \\ x/2 & \text{otherwise.} \end{cases}$$

Then it is easy to check that $(\tilde{R}, \frac{1}{2})$ is a complementary pair for E .

For more general wavelet sets K constructed by the neighborhood-mapping method, we let T_0 be the injective \mathbb{Z}^d -translated function that maps the unit cube Q onto the neighborhood K_0 of the origin used in the set up of the construction. The existence of T_0 follows from the condition that K_0 must be τ -congruent to Q . We then define

$$E = \bigcup_{j \geq 1} 2^{-j} K = \bigcup_{j \geq 1} 2^{-j} \left[\left(K_0 \setminus \bigcup_{k=0}^{\infty} A_k \right) \cup T \left(\bigcup_{k=0}^{\infty} A_k \right) \right]$$

and $\check{R} : K_0 \rightarrow E$ as

$$(15) \quad \check{R}(x) = \begin{cases} T(x)/2 & \text{if } x \in \bigcup_{k=0}^{\infty} A_k \\ x/2 & \text{otherwise.} \end{cases}$$

It can then be shown that $(\check{R} \circ T_0, \frac{1}{2})$ is a complementary pair for E . In fact

$$\begin{aligned} \check{R} \circ T_0(Q) &= \check{R}(K_0) = \frac{1}{2} \left[\left(K_0 \setminus \bigcup_{k=0}^{\infty} A_k \right) \cup T \left(\bigcup_{k=0}^{\infty} A_k \right) \right] \subseteq E, \\ \frac{1}{2} E &= E \setminus R(Q), \end{aligned}$$

and

$$\bigcup_{j \geq 0} 2^{-j} R(Q) = \bigcup_{j \geq 1} 2^{-j} K = E.$$

It should be mentioned that the above complementary pair corresponding to the wavelet set K constructed by neighborhood-mapping method comprises of the simple map $1/2$ and the somewhat complicated map \check{R} . The definition of \check{R} in (15) essentially depends on the very wavelet set K we would like to construct. It is clear that the Baggett et al. wavelet set construction is theoretically more general than the neighborhood-mapping method. One way to see this is by considering the boundedness of wavelet sets. There are examples of unbounded wavelet sets, e.g., [DLS98], but the neighborhood-mapping wavelet sets are always in the cube $[-2N, 2N]^d$. In practice while a neighborhood-mapping wavelet set K is constructed by a couple of simple objects, i.e. a neighborhood K_0 of 0 and a map $T : K_0 \rightarrow [-2N, 2N]^d \setminus [-N, N]^d$ satisfying some properties, we are not able to find a simpler complementary pair that depends on some simple entities.

Finally, it is worth noting that if we set

$$E_n = \bigcup_{j \geq 1} 2^{-j} \left[\left(K_0 \setminus \bigcup_{k=0}^n A_k \right) \cup T \left(\bigcup_{k=0}^n A_k \right) \right],$$

then

$$2E_n \setminus E_n = \left(K_0 \setminus \bigcup_{k=0}^{n+1} A_k \right) \cup T \left(\bigcup_{k=0}^n A_k \right).$$

References

- [Aus95] Pascal Auscher. Solution of two problems on wavelets. *J. Geom. Anal.*, 5(2):181–236, 1995.
- [Ben97] John J. Benedetto. *Harmonic Analysis and Applications*. CRC Press, Inc., Boca Raton, Florida, 1997.
- [BL99] John J. Benedetto and Manuel T. Leon. The construction of multiple dyadic minimally supported frequency wavelets on \mathbb{R}^d . In *The Functional and Harmonic Analysis of Wavelets and Frames (San Antonio, TX, 1999)*, pages 43–74. Amer. Math. Soc., Providence, RI, 1999.
- [BL01] John J. Benedetto and Manuel T. Leon. The construction of single wavelets in d-dimensions. *J. Geom. Anal.*, 11(1):1–15, 2001.
- [BMM99] Lawrence W. Baggett, Herbert A. Medina, and Kathy D. Merrill. Generalized multi-resolution analyses and a construction procedure for all wavelet sets in \mathbb{R}^n . *J. Fourier Anal. Appl.*, 5(6):563–573, 1999.
- [Dau88] Ingrid Daubechies. Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, 41(7):909–996, 1988.
- [Dau92] Ingrid Daubechies. *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992.
- [DLS97] Xingde Dai, David R. Larson, and Darrin M. Speegle. Wavelet sets in \mathbb{R}^n . *J. Fourier Anal. Appl.*, 3(4):451–456, 1997.
- [DLS98] Xingde Dai, David R. Larson, and Darrin M. Speegle. Wavelet sets in \mathbb{R}^n . II. In *Wavelets, multiwavelets, and their applications (San Diego, CA, 1997)*, pages 15–40. Amer. Math. Soc., Providence, RI, 1998.
- [Fal90] Kenneth Falconer. *Fractal geometry*. John Wiley & Sons Ltd., Chichester, 1990. Mathematical foundations and applications.
- [GH94] Karlheinz Gröchenig and Andrew Haas. Self-similar lattice tilings. *J. Fourier Anal. Appl.*, 1(2):131–170, 1994.
- [GM92] K. Gröchenig and W. R. Madych. Multiresolution analysis, Haar bases, and self-similar tilings of \mathbb{R}^n . *IEEE Trans. Inform. Theory*, 38(2, part 2):556–568, 1992.
- [Gri95] Gustaf Gripenberg. A necessary and sufficient condition for the existence of a father wavelet. *Studia Math.*, 114(3):207–226, 1995.
- [Haa10] Alfred Haar. Zur Theorie der orthogonalen Funktionen-Systeme. *Math. Ann.*, 69:331–371, 1910.
- [HW96] Eugenio Hernández and Guido Weiss. *A first course on wavelets*. CRC Press, Boca Raton, FL, 1996.
- [LW96] Jeffrey C. Lagarias and Yang Wang. Integral self-affine tiles in \mathbb{R}^n . I. Standard and nonstandard digit sets. *J. London Math. Soc.* (2), 54(1):161–179, 1996.
- [Mal85] Stéphane G. Mallat. An efficient image representation for multiscale analysis. Technical report, GRASP Laboratory, Dept. of Computer and Information Science, U. of Pennsylvania, Philadelphia, (about 1985).
- [Mal89] Stéphane G. Mallat. Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$. *Trans. Amer. Math. Soc.*, 315(1):69–87, September 1989.
- [Mey86] Yves Meyer. Ondelettes, fonctions splines et analyses graduées. Lectures given at the University of Torino, Italy, 1986.
- [Mey87] Yves Meyer. Principe d'incertitude, bases hilbertiennes et algèbres d'opérateurs. *Astérisque*, (145-146):4, 209–223, 1987. Séminaire Bourbaki, Vol. 1985/86.

- [Mey89] Yves Meyer. Ondelettes, filtres miroirs en quadrature et traitement numérique de l'image. *Gaz. Math.*, (40):31–42, 1989.
- [Mey90] Yves Meyer. *Ondelettes et Opérateurs*. Hermann, Paris, 1990.
- [SW71] Elias M. Stein and Guido Weiss. *Introduction to Fourier analysis on Euclidean spaces*. Princeton University Press, Princeton, N.J., 1971. Princeton Mathematical Series, No. 32.
- [SW98] Paolo M. Soardi and David Weiland. Single wavelets in n -dimensions. *J. Fourier Anal. Appl.*, 4(3):299–315, 1998.
- [Wan] X. Wang. *The study of wavelets from the properties of their Fourier transforms*. PhD thesis, Washington University, St. Louis, MO.
- [Zak96] V. Zakharov. Nonseparable multidimensional Littlewood-Paley like wavelet bases. Preprint, 1996.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF MARYLAND, COLLEGE PARK, MARYLAND
20742

E-mail address: jjb@math.umd.edu
URL: <http://www.math.umd.edu/~jjb>

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF MARYLAND, COLLEGE PARK, MARYLAND
20742

E-mail address: sxs@math.umd.edu
URL: <http://www.math.umd.edu/~sxs>

This page intentionally left blank

Joint Invariant Signatures for Curve Recognition

Mireille Boutin

ABSTRACT. We present a general method for recognizing curves modulo a Lie group action. This method is based on moving frames and consists in constructing a joint invariant signature which is robust and of minimal dimension. The example of planar curve recognition modulo equi-affine transformations is discussed in details.

1. Introduction

This paper is the sequel to a warm up paper on polygon recognition [MB01]. We are interested in curve recognition modulo Lie group actions, which is an important problem in computer vision. Very often, objects are represented by the boundary of their projection onto a plane (picture). Generally, the position and orientation of the curve on the picture are irrelevant informations for the purpose of identifying the object. Depending on how the picture is taken, other types of (Lie group) transformations represent irrelevant variations in the shape of the curve, e.g. equi-affine, affine, similarity and projective transformations. The problem of recognizing the object thus reduces to recognizing curves modulo a Lie group action.

One solution to this problem was proposed by Calabi et al. [COSTH] and consists in constructing a signature parameterized by some well chosen differential invariants. The differential invariant signature of a curve is a representative for the equivalence class of this curve. In other words, two curves are in the same equivalence class if and only if their signature is the same. For planar curve recognition up to rotation and translation, Calabi et al. proposed the use of a signature curve parameterized by (κ, κ_s) , where κ is the Euclidean curvature and κ_s its derivative with respect to arc-length. They also suggested a signature for curve recognition modulo equi-affine transformations. Although these differential invariant signatures are promising, one problem is their sensitivity to noise and round off errors. To improve the results obtained on discrete data, the use of numerically invariant approximations for these differential signature curves [COSTH, MB00] was suggested.

Yet another type of signature was proposed by Peter Olver in [O01] as a solution to the problem of noise sensitivity. Instead of using differential invariants

1991 *Mathematics Subject Classification.* Primary 53A55; Secondary 68T45.

Key words and phrases. curve recognition, invariants.

This work was supported in parts by NSF grants KDI BCS-9980091 and 0074276.

to parameterize a one-dimensional signature, the author proposed parameterizing a higher dimensional signature with a different type of invariants called *joint invariants*. Denote by $M^{\times(n)}$ the Cartesian product of n copies of a manifold $M^{\times(n)} = M \times \dots \times M$ (n -times). The action of G on M can be prolonged unto $M^{\times(n)}$ by setting $g \cdot (z_1, \dots, z_n) = (g \cdot z_1, \dots, g \cdot z_n)$, for all $z_1, \dots, z_n \in M$ and all $g \in G$. A joint invariant is an invariant of the prolonged action of G on $M^{\times(n)}$. More precisely, it is a real valued function $J : M^{\times(n)} \rightarrow \mathbb{R}$ which remains unchanged under the simultaneous action of G on the n points of M , in other words $J(g \cdot z_1, \dots, g \cdot z_n) = J(z_1, \dots, z_n)$, for all $z_1, \dots, z_n \in M$ and all $g \in G$. In general, joint invariants are more noise resistant than differential invariants simply because they do not involve any derivative. The simplest joint invariant of the action of the Euclidean group on \mathbb{R}^2 is the distance between two points. In fact, any other Euclidean joint invariant is a function of distances. The simplest joint invariant of the action of the equi-affine Lie group on the plane is the signed area of the triangle spanned by three ordered points. As Élie Cartan has shown [C], the solution to many equivalence problems lies in the functional relationships between some well chosen invariants. In many cases, we can see this relationship simply by plotting the invariants. For example, by plotting the Euclidean distances between enough points on a curve, one is able to completely characterize the curve modulo rotations, translations and reflections. In fact, the six pairwise Euclidean distances defined by four points on the curve are enough for this purpose. A four-dimensional signature in \mathbb{R}^6 can thus be constructed by evaluating these six distances on all possible four points on the curve.

However, there is no a priori reason to believe that the signature of a curve should be more than one-dimensional. In fact, a one-dimensional signature is quite desirable as it insures a minimal complexity order for the recognition algorithm. Also, there is no apparent reason why the signature of a curve in an m -dimensional manifold M should be parameterized by more than m invariants. For example, since a planar curve is specified by two quantities, we see no reason why its signature should be specified by more than two quantities. Indeed, we can show that, if some slight regularity conditions hold, then one can obtain a certain set of m joint invariants using the moving frame method, and use them to parameterize a signature curve. For simplicity, we shall restrict ourselves to the case of a Lie group action on a two-dimensional smooth (Hausdorff) manifold M^2 , although the theoretical results contained in [MB01] allow for a generalization to Lie group actions on any smooth (Hausdorff) manifold.

An interesting fact is that a signature parameterized by joint invariants no longer requires the curve to be differentiable, as opposed to the differential invariant signatures. This observation led us to consider using joint invariant signatures for recognizing curves which are not necessarily smooth (e. g. polygons). In fact, discontinuities in the derivatives provide possible choices for the distinguished points called landmarks, which are the key to constructing a simple signature. Some other (more robust) possibilities for landmarks will also be discussed. In a sense, we trade off high dimensionality for the use of landmarks. We believe that, with a good choice of landmarks, this will yield efficient and noise resistant curve recognition algorithms.

2. Curve Recognition up to Orientation and Area Preserving Affine Transformations

2.1. Curve Segment Recognition Modulo $SA(2)$. Let us start with a simple example of practical interest. Consider the action of the equi-affine group $SA(2)$ on $\{z \in \mathbb{R}^2\}$ given by

$$\bar{z} = Az + b,$$

where $A \in SL(2, \mathbb{R})$ and $b \in \mathbb{R}^2$.

Let C be a planar curve segment. Assume that this curve segment can be parameterized as $C = \{\alpha(t) \in \mathbb{R}^2 | 0 \leq t \leq 1\}$ with $\alpha(t)$ a continuous and C^1 map such that $\alpha(0) \neq \alpha(1)$ and $|\alpha'(t)| \neq 0$, for all $t \in [0, 1]$. For simplicity, we exclude the cases where the curve segment self-intersects although these can be treated by a similar method. The parameterization is introduced solely to simplify the exposition; it is not relevant for we are concerned with curves obtained from images. In other words, we are interested in the graph of $\alpha(t)$.

We want to determine whether two given curve segments belong to the same orbit under the action of $SA(2)$. The case where C is a straight line segments is trivial: any two straight lines are equivalent modulo $SA(2)$. When C is not a straight line, we take the two end points $\alpha(0)$ and $\alpha(1)$ of C , and label them ζ_1 and ζ_2 respectively. These two points serve as landmarks. Observe that the order of the labeling is arbitrary, since it depends on our choice of parameterization.

For reasons to be given later, this case requires the use of a minimum of three landmarks. Denote by $\Delta(z_1, z_2, z_3) = \frac{1}{2}(z_3 - z_1) \times (z_2 - z_1)$ the signed area of the triangle with vertices z_1, z_2, z_3 and let $f : C \rightarrow \mathbb{R}$ be the function $f(z) = |\Delta(\zeta_1, \zeta_2, z)|$. If f reaches a local maximum in a unique point on C , then we let the third landmark ξ be this point. Observe that if two curves segments C and \bar{C} on which f reaches a local maximum in a unique point are equivalent modulo $SA(2)$, then their landmarks ζ_1, ξ, ζ_2 and $\bar{\zeta}_1, \bar{\xi}, \bar{\zeta}_2$ respectively are also equivalent modulo $SA(2)$ (up to the choice of ordering of the end points). This is what we call *equivariance up to order reversion* of the landmarks. It provides us with an easy equivalence test since the equivalence class of three distinct non-collinear points in the plane is characterized by their signed area (a well known fact in affine geometry).

LEMMA 2.1. *Let $C = \{\alpha(t)\}$ and $\bar{C} = \{\bar{\alpha}(t)\}$ be two curve segments on which f reaches a unique local maximum at ξ and $\bar{\xi}$ respectively. If C and \bar{C} are equivalent under $SA(2)$, then*

$$\Delta(\alpha(0), \xi, \alpha(1)) = \pm \Delta(\bar{\alpha}(0), \bar{\xi}, \bar{\alpha}(1)).$$

In general, f may reach a local maximum in many points, including whole segments, on a curve. Since the property of *being a local maximum of f* is preserved under the action of $SA(2)$, local maxima can be used to simplify the search for equivalent curve segments. For example, if f reaches a local maximum only at a finite number of points $\xi_1, \xi_2, \dots, \xi_k$ (labeled in increasing order according to the parameterization), then one can reject any other curve segment \bar{C} for which the preimage of the local maxima of f ordered according to the parameterization of C neither belong to the equivalence class of $(\xi_1, \xi_2, \dots, \xi_k)$ nor of $(\xi_k, \dots, \xi_2, \xi_1)$. In particular, if this preimage is not finite or if it is finite but has a different cardinality, then it must be rejected. It is an easy task to test whether two strings of equal length belong to the same equivalence class, as we now explain.

2. Curve Recognition up to Orientation and Area Preserving Affine Transformations

2.1. Curve Segment Recognition Modulo $SA(2)$. Let us start with a simple example of practical interest. Consider the action of the equi-affine group $SA(2)$ on $\{z \in \mathbb{R}^2\}$ given by

$$\bar{z} = Az + b,$$

where $A \in SL(2, \mathbb{R})$ and $b \in \mathbb{R}^2$.

Let C be a planar curve segment. Assume that this curve segment can be parameterized as $C = \{\alpha(t) \in \mathbb{R}^2 | 0 \leq t \leq 1\}$ with $\alpha(t)$ a continuous and C^1 map such that $\alpha(0) \neq \alpha(1)$ and $|\alpha'(t)| \neq 0$, for all $t \in [0, 1]$. For simplicity, we exclude the cases where the curve segment self-intersects although these can be treated by a similar method. The parameterization is introduced solely to simplify the exposition; it is not relevant for we are concerned with curves obtained from images. In other words, we are interested in the graph of $\alpha(t)$.

We want to determine whether two given curve segments belong to the same orbit under the action of $SA(2)$. The case where C is a straight line segments is trivial: any two straight lines are equivalent modulo $SA(2)$. When C is not a straight line, we take the two end points $\alpha(0)$ and $\alpha(1)$ of C , and label them ζ_1 and ζ_2 respectively. These two points serve as landmarks. Observe that the order of the labeling is arbitrary, since it depends on our choice of parameterization.

For reasons to be given later, this case requires the use of a minimum of three landmarks. Denote by $\Delta(z_1, z_2, z_3) = \frac{1}{2}(z_3 - z_1) \times (z_2 - z_1)$ the signed area of the triangle with vertices z_1, z_2, z_3 and let $f : C \rightarrow \mathbb{R}$ be the function $f(z) = |\Delta(\zeta_1, \zeta_2, z)|$. If f reaches a local maximum in a unique point on C , then we let the third landmark ξ be this point. Observe that if two curves segments C and \bar{C} on which f reaches a local maximum in a unique point are equivalent modulo $SA(2)$, then their landmarks ζ_1, ξ, ζ_2 and $\bar{\zeta}_1, \bar{\xi}, \bar{\zeta}_2$ respectively are also equivalent modulo $SA(2)$ (up to the choice of ordering of the end points). This is what we call *equivariance up to order reversion* of the landmarks. It provides us with an easy equivalence test since the equivalence class of three distinct non-collinear points in the plane is characterized by their signed area (a well known fact in affine geometry).

LEMMA 2.1. *Let $C = \{\alpha(t)\}$ and $\bar{C} = \{\bar{\alpha}(t)\}$ be two curve segments on which f reaches a unique local maximum at ξ and $\bar{\xi}$ respectively. If C and \bar{C} are equivalent under $SA(2)$, then*

$$\Delta(\alpha(0), \xi, \alpha(1)) = \pm \Delta(\bar{\alpha}(0), \bar{\xi}, \bar{\alpha}(1)).$$

In general, f may reach a local maximum in many points, including whole segments, on a curve. Since the property of *being a local maximum* of f is preserved under the action of $SA(2)$, local maxima can be used to simplify the search for equivalent curve segments. For example, if f reaches a local maximum only at a finite number of points $\xi_1, \xi_2, \dots, \xi_k$ (labeled in increasing order according to the parameterization), then one can reject any other curve segment \bar{C} for which the preimage of the local maxima of f ordered according to the parameterization of C neither belong to the equivalence class of $(\xi_1, \xi_2, \dots, \xi_k)$ nor of $(\xi_k, \dots, \xi_2, \xi_1)$. In particular, if this preimage is not finite or if it is finite but has a different cardinality, then it must be rejected. It is an easy task to test whether two strings of equal length belong to the same equivalence class, as we now explain.

THEOREM 2.5. *Two curve segments $C = \{\alpha(t)\}$ and $\bar{C} = \{\bar{\alpha}(t)\}$ on which f reaches a unique local maximum are equivalent if and only if either $(\Sigma_+(C), \Sigma_-(C)) = (\Sigma_+(\bar{C}), \Sigma_-(\bar{C}))$ or $(\Sigma_-(C), \Sigma_+(C)) = (\Sigma_+(\bar{C}), \Sigma_-(\bar{C}))$.*

PROOF. Observe that the signature is a function solely of signed areas, which are invariants of the equi-affine group action on the plane. Moreover, except for the choice of orientation, our procedure to determine the landmarks will consistently lead to the same three points modulo $SA(2)$. So the signature of two equivalent curve segments will be the same up to a permutation of its two curve components.

Now, suppose we are given two curve segments C and \bar{C} with the same signature up to a permutation of the two curve components. We can assume that $(\Sigma_+(C), \Sigma_-(C)) = (\Sigma_+(\bar{C}), \Sigma_-(\bar{C}))$ by reparameterizing one of the curve with $1-t$ instead of t , thus reversing the direction of travel, if necessary. We then have $\Sigma_+(C) = \Sigma_+(\bar{C})$. Consider the end points of these curve segments. One of them lies on the y axis and the other lies on the x axis. Observe that ζ_1, ζ_2 and ξ_+ must be distinct, therefore none of the end points of the signature actually lies at the origin. The end point lying on the y axis corresponds to the first points ζ_1 and $\bar{\zeta}_1$ of each curve segment. By looking at the second component of this end point, we know the value of $\Delta(\zeta_1, \xi_+, \zeta_2)$ and $\Delta(\bar{\zeta}_1, \bar{\xi}_+, \bar{\zeta}_2)$, which must be equal by assumption. It is a well known fact in affine geometry that three distinct non-collinear points in \mathbb{R}^2 can be mapped unto each other using an equi-affine transformation if and only if their signed area is the same. So there exists $g \in SA(2)$ such that $g \cdot (\zeta_1, \xi_+, \zeta_2) = (\bar{\zeta}_1, \bar{\xi}_+, \bar{\zeta}_2)$.

Observe that the invariants I_1 and I_2 are such that given three distinct points z_1, z_2 and z_3 , the value of $I_1(z_1, z_2, z_3, z_4)$ and $I_2(z_1, z_2, z_3, z_4)$ uniquely determines z_4 , provided that $I_1(z_1, z_2, z_3, z_4) \neq 0$. This means that all points $z \in C$ such that $\Delta(\xi_+, \zeta_2, z) \neq 0$ are mapped to corresponding points $\bar{z} \in \bar{C}$ by g . By continuity of the curve segments and since the parameterization has nonzero speed, we conclude that $g \cdot C = \bar{C}$. \square

Note that, although only three landmarks are used in our method, one could certainly construct a signature that uses more than three landmarks, as we shall do for the case of closed curves.

We took the picture of a leaf and segmented it (see Figure 1) in order to test the noise resistance of our method. We considered the boundaries of the left and right sides of the leaf as two curve segments. The left side curve was flipped before being compared to the right side curve. Note that no preprocessing was done on either curves. The end points of the leaf area were used as landmarks and, for each curve, a third landmark was found by taking the point on the boundary spanning the triangle of maximal area, $A_{max}(\text{right})$ and $A_{max}(\text{left})$ respectively. These three landmarks are represented as the vertices of a triangle on Figure 2. The signature of both segments is displayed in Figure 3 together with a third signature obtained by multiplying the signature of the left side of the leaf by $\frac{A_{max}(\text{right})}{A_{max}(\text{left})}$. Despite significant local variations in the shapes of the sides, the signature of the right side curve and the rescaled signature of the left side curve are surprisingly similar.

2.2. Closed Curve Recognition Modulo $SA(2)$. Let C be a closed, non-self-intersecting planar curve parameterized by $\{\alpha(t) \in \mathbb{R}^2 | 0 \leq t \leq 1\}$ with α continuous and piecewise differentiable, $\alpha(0) = \alpha(1)$ and $\alpha'(t) \neq 0$ on the differentiable pieces.



FIGURE 1. Segmented Leaf.

The first thing to do is to choose a minimum of three landmarks. Again, there are many ways to do this. For example, consider the affine curvature κ at every point of the curve, which in local coordinates $(x, y) \in \mathbb{R}^2$ with $y = u(x)$ is given by $\kappa(x) = \frac{3u_{xx}u_{xxxx}-5u_{xxx}^2}{9u_{xx}^{8/3}}$. Observe that κ is not defined at inflection points of the curve. Define the map $g : C \rightarrow \mathbb{R}$ by $g(z) = |\kappa(z)|$. One possibility for the landmarks is to take all the points on C where the function g either is not defined or reaches a local maximum.

For a closed (regular) and C^4 curve, one can show that there exist at least three points on the curve where the affine curvature reaches a local maximum [G]. The case where the affine curvature is constant corresponds to ellipses and can be treated separately. Indeed two ellipses can be mapped to each other by an affine transformation if and only if their area is the same. If g reaches a local max on a C^4 curve at a finite number of points $\{\zeta_1, \dots, \zeta_k\}$, $k \geq 3$, then these points can be used as landmarks. One can also include the endpoints of the intervals on which g reaches a local max, if such an interval exists.

If the curve is merely piecewise C^4 , then one can include the points where the affine curvature doesn't exist or is not continuous in the set of landmarks. If doing so yields less than three landmarks, then one can divide up the differentiable pieces of the curve into intervals of equal affine arc-length and use the boundary points of these intervals as landmarks. Adding the points where g reaches a local maximum is also a possibility.

A different approach to defining landmarks would be to use the affine skeleton [BSTG] of the curve C . The affine skeleton being an equivariant structure, we could, for example, define landmarks directly on the skeleton and use these as landmarks for the curve. One could think of using end points or junction points

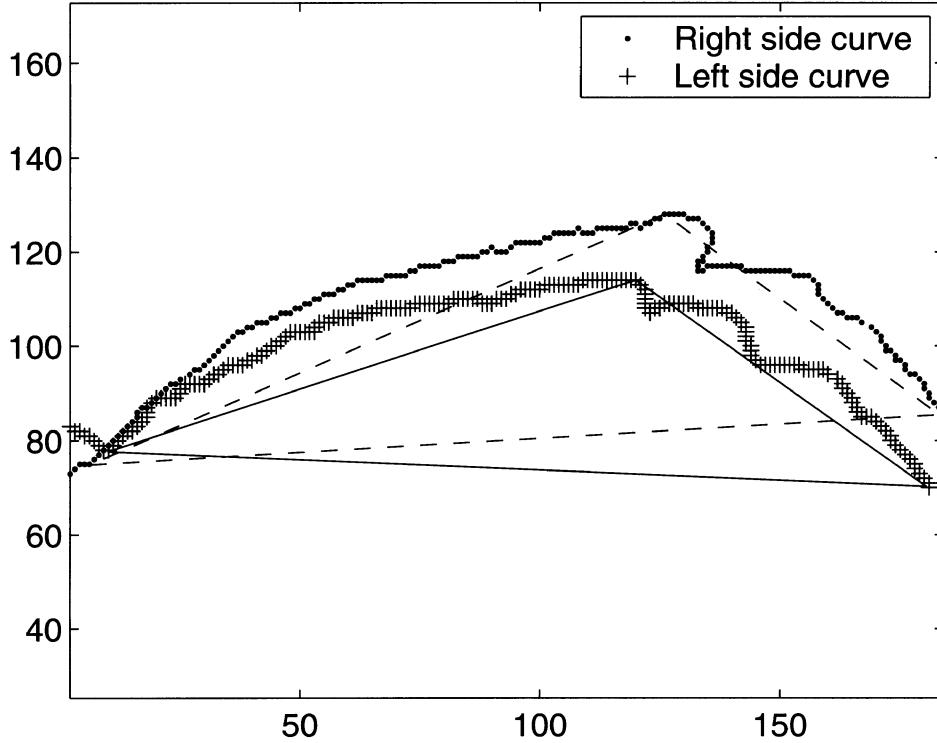


FIGURE 2. The Two Side Curves of the Leaf.

of the skeleton, which are characteristics that are preserved under the action of $SA(2)$. In any event, the essential is to obtain a string of at least three (robust) landmarks and the technique used to achieve this goal matters little.

Let $C = \{\alpha(t)\}$ be a closed curve with k landmarks $(\zeta_1, \dots, \zeta_k) \in (\mathbb{R}^2)^{\times(k)}$. Let \mathbb{Z}_k act on $(\mathbb{R}^2)^{\times(k)}$ by cyclically permuting the order of the k landmarks. Let \mathbb{Z}_2 act on $(\mathbb{R}^2)^{\times(k)}$ by reversing the order of the k landmarks. Consider $\mathbb{H}_k = < \mathbb{Z}_k, \mathbb{Z}_2 >$, the group of transformations generated by the action of \mathbb{Z}_k and \mathbb{Z}_2 on $(\mathbb{R}^2)^{\times(k)}$, and the point $\langle \zeta_1, \dots, \zeta_k \rangle \in (\mathbb{R}^2)^{\times(k)} \bmod \mathbb{H}_k$. We identify this point with a polygon $P = P(C)$ in \mathbb{R}^2 . To simplify the exposition, we set $\zeta_{k+1} = \zeta_1$, $\zeta_{k+2} = \zeta_2$, and so on. By construction, the polygon $P(C)$ associated to a curve C is equivariant, in the sense that if $\bar{C} = g \cdot C$, then $P(\bar{C}) = g \cdot P(C)$. The equivalence of two polygons can be checked in $O(k)$ steps using the map S defined in subsection 2.1. In fact, we have the following simple equivalence test, which can be proved by the same arguments as the ones used in the proof of 2.5.

THEOREM 2.6. [MB01] *Let $k \geq 3$ and let $P = \langle p_1, \dots, p_k \rangle$ and $\bar{P} = \langle \bar{p}_1, \dots, \bar{p}_k \rangle$ be two polygons with no three consecutive points lying on a straight line. Then $P \equiv \bar{P} \bmod SA(2)$ if and only if*

$$\begin{aligned} S(p_1, p_2, \dots, p_k) &= S(\bar{p}_1, \dots, \bar{p}_k) \quad \bmod \mathbb{Z}_k \\ \text{or } S(p_k, \dots, p_2, p_1) &= S(\bar{p}_1, \dots, \bar{p}_k) \quad \bmod \mathbb{Z}_k. \end{aligned}$$

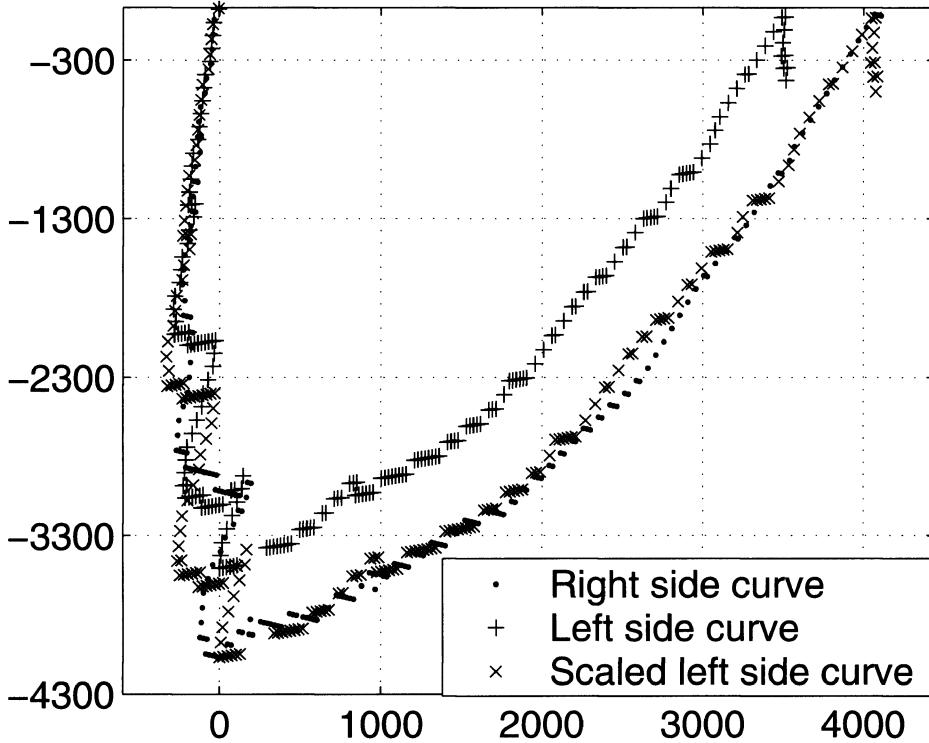


FIGURE 3. Equi-affine Invariant Signatures of the Two Side Curves of the Leaf.

If comparison of the polygons doesn't rule out equivalence, then we are ready to perform the ultimate equivalence test. For $i = 1, \dots, k$, let $\Sigma_{i+}(C)$ be the planar curve segment parameterized by $I_1(\zeta_i, \zeta_{i+1}, \zeta_{i+2}, z)$ and $I_2(\zeta_i, \zeta_{i+1}, \zeta_{i+2}, z)$, for $z \in C$ going from ζ_{i+3} to ζ_{i+2} . Consider $\Sigma_+(C)$, the string of k curves

$$\Sigma_+(C) = (\Sigma_{1+}, \Sigma_{2+}, \dots, \Sigma_{k+}).$$

Observe that the signature for polygons used in Theorem 2.6 corresponds to the first point of each curve segment Σ_{i+} . Similarly, let Σ_{i-} be the planar curve segment parameterized by $I_1(\zeta_i, \zeta_{i-1}, \zeta_{i-2}, z)$ and $I_2(\zeta_i, \zeta_{i-1}, \zeta_{i-2}, z)$ for $z \in C$ going from ζ_{i-3} to ζ_{i-2} . We define the signature $\Sigma(C)$ of a closed curve $C = \{\alpha(t)\}$ as the pair of curves $\Sigma(C) = (\Sigma_+(C), \Sigma_-(C))$. As a corollary of 2.5, we have the following theorem.

THEOREM 2.7. *Two closed simple curves with a string of $k \geq 3$ landmarks are equivalent under SA(2) if and only if either $\Sigma_+(C) \equiv \Sigma_+(\bar{C}) \pmod{\mathbb{Z}_k}$ or $\Sigma_-(C) \equiv \Sigma_-(\bar{C}) \pmod{\mathbb{Z}_k}$.*

3. Generalization to Other Lie Groups

3.1. Theoretical Foundations. Let G be a δ -dimensional Lie group acting on an m -dimensional smooth (Hausdorff) manifold M . In this paper, we assume that $M = M^2$ is a two-dimensional manifold, although the theory developed in

[MB01] applies to manifolds of any dimension. Our recipe to build a signature curve uses two ingredients: 1) a set of two suitable invariants, 2) a set of distinguished points called landmarks. Let us first talk about the landmarks.

The only property that we require of the landmarks is that they be equivalent on equivalent curves. This is what we call *equivariance of the landmarks* which we now define precisely. Let \mathcal{L} be a rule that associates to every curve C a finite string of points $\mathcal{L}(C) = (\xi_1, \dots, \xi_k) \in (M^2)^{\times(k)}$.

DEFINITION 3.1. We say that \mathcal{L} is *equivariant* if

$$\mathcal{L}(g \cdot C) = g \cdot \mathcal{L}(C), \text{ for all } g \in G.$$

We say that \mathcal{L} is *equivariant up to H_k* , for some subgroup of permutations $H_k \subset S_k$, if for any $\bar{C} = g \cdot C$ for some $g \in G$, there exists $h \in H_k$ such that

$$\mathcal{L}(\bar{C}) = g \cdot h \cdot \mathcal{L}(C).$$

DEFINITION 3.2. If \mathcal{L} is an equivariant rule, C a curve segment and if $\mathcal{L}(C) = (\xi_1, \dots, \xi_k)$, then we say that the points $\{\xi_1, \dots, \xi_k\}$ are (equivariant) landmarks for C .

The freedom in the choice of landmarks comes from the infinitely many possible choices of equivariant rules \mathcal{L} . In practical applications, the choice of \mathcal{L} should be made in such a way to insure robustness of the landmarks. For example, for the third landmark on planar curves under the action of $SA(2)$, we could have decided to take the first (or last) global maximum of the function f , but for noisy data representing curves with more than one local max, this would not yield robust landmarks. We could also have taken the first point encountered where the affine invariant curvature reaches a maximum or doesn't exist, but in general, the more derivatives the rule \mathcal{L} relies on, the more sensitive the results. So a perhaps better possibility is to pick the point that lies at half the total affine arc-length of the curve segment. There are many other possibilities for choosing landmarks and performance of any one type of landmarks will always depend on the type of curve segments under consideration.

The next ingredient on our list are the invariants that will parameterize the signature. We impose the condition that only two invariants should be used, so as to not increase the complexity of the recognition algorithm unnecessarily.

DEFINITION 3.3. We say that the group action is *semi-regular* if all the orbits have the same dimension. If, in addition, for all $z \in M$, there exists a collection of arbitrarily small neighborhoods whose intersection with the orbit through z is connected, then we say that the action is *regular*.

Most of our results are based on the following theorem due to Frobenius.

THEOREM 3.4. [F] *If G acts on an open set $O \subset M$ semi-regularly with s dimensional orbits, then $\forall x_0 \in O$ there exist $m - s$ functionally independent local invariants I_1, \dots, I_{m-s} defined on a neighborhood U of x_0 such that any other local invariant I defined near x_0 is a function $I = f(I_1, \dots, I_{m-s})$. If the action of G is regular, then the local invariants can be taken to be invariants in a neighborhood of x_0 , and two points $x_1, x_2 \in U$ are in the same orbit if and only if $I_i(x_1) = I_i(x_2)$, $\forall i = 1, \dots, m - s$.*

The set $\{I_1, \dots, I_{m-s}\}$ is often called a *complete fundamental set of invariants*. The moving frame normalization method, as explained in detail in [O99] Chapter 8, is an algorithm to construct a complete fundamental set of invariants. The idea

is to build a right-equivariant map $\rho : M \rightarrow G$ called *moving frame*. By right-equivariance, we mean that $\rho(g \cdot x) = \rho(x)g^{-1}$, for all $g \in G$ and all $x \in M$. A moving frame can be constructed following a step-by-step method. Given local coordinates $x = (x_1, \dots, x_m)$ in a neighborhood of $x_0 \in M$, we consider the m coordinates of the vector $\rho(x) \cdot x$, which contain a complete fundamental set of invariants define in a neighborhood of x_0 .

In the following, we shall make use of this interesting property of the maximal orbit dimension.

LEMMA 3.5. [MB02] *Let s_n denote the maximal orbit dimension of the prolonged action of G on $M^{\times(n)}$. If $s_n = s_{n+1}$, then $s_{n+j} = s_n$, for all $j \in \mathbb{N}$.*

DEFINITION 3.6. We say that a group action of G on a manifold M is *effective* if the subgroup $\{g \in G | g \cdot z = z, \forall z \in M\} = \{e\}$ is trivial. We say that a group action of G on a manifold M is *effective on subsets* if for all open subset $U \subset M$ the subgroup $\{g \in G | g \cdot z = z, \forall z \in U\} = \{e\}$ is trivial.

Our method also relies strongly on the following important fact.

THEOREM 3.7. [MB02] *There exists a minimal n_0 such that G acts with δ -dimensional orbits on a dense and open subset \mathcal{D} of $M^{\times(n_0)}$ if and only if G acts (locally) effectively on subsets of M .*

A general method for constructing suitable invariants to characterize polygons in an m -dimensional manifold modulo a Lie group action is explained in details in [MB01]. It turns out that the same invariants are suitable for characterizing curves. We shall summarize these results here for the particular cases of Lie group actions on a two-dimensional manifold M^2 .

We are assuming for the rest of this paper that G acts effectively on subsets of M^2 so that n_0 exists. Observe that the action of G on the (open) set of points belonging to orbits of maximal dimensions is (at least) semi-regular. So Theorem 3.4 guarantees the existence of fundamental invariants in a neighborhood of any such points, and we can construct these invariants using the moving frame normalization method.

Let n^* be the minimum integer n such that the maximal orbit dimension s_n of the action of prolonged action of G on $(M^2)^{\times(n)}$ is strictly smaller than $2n$, the dimension of $(M^2)^{\times(n)}$. In other words, n^* is the minimal number of points on which non-trivial joint invariants can depend (by Theorem 3.4). Observe that $n_0 \geq n^* - 1$. By considering sets of fundamental invariants on open subsets of $(M^2)^{\times(n^*)}$, $(M^2)^{\times(n^*+1)}$, ..., $(M^2)^{\times(n_0)}$, $(M^2)^{\times(n_0+1)}$ successively, we can cook up a set of two invariants I_1 and I_2 that are suitable for parameterizing a signature curve. Recall that δ denotes the dimension of G .

LEMMA 3.8. *Let s_n denote the orbit dimension of the prolonged Lie group action on $(M^2)^{\times(n)}$. Then the following relations hold:*

$$\begin{aligned} s_n &= 2n, \text{ for all } n < n^*, \\ s_n &= s_{n-1} + 1, \text{ for all } n_0 \geq n \geq n^*, \\ s_n &= s_{n-1}, \text{ for all } n > n_0, \\ (\text{and therefore}) \quad n_0 &= \delta + 1 - n^*. \end{aligned}$$

PROOF. By definition of n^* , we have $s_n = 2n$, for all $n > n^*$ and $s_{n^*} < 2n^*$. By Lemma 3.5, $s_n - s_{n-1} = 1$ or 2 , whenever $n \leq n_0$, while $s_n - s_{n-1} = 0$, whenever $n > n_0$. However, it turns out that $s_n - s_{n-1} < 2$ for any $n \geq n^*$. Here is why.

- For $n = n^*$. If $s_{n^*} - s_{n^*-1} = 2$, then $s_{n^*} = 2n^*$ which contradicts the definition of n^* .
- For all $n > n^*$. The number of n -point joint invariant $\#_n$ is $\#_n - \#_{n-1} \geq 1$, since if $I(x_1, \dots, x_{n^*})$ is an n^* -point joint invariant, then $\bar{I}(x_1, \dots, x_n) := I(x_{n-n^*+1}, \dots, x_n)$ can be used as a fundamental n -point joint invariant.

Therefore we have $s_{n^*-1} = 2(n^* - 1)$, $s_n - s_{n-1} = 1$, for all $n_0 \geq n \geq n^*$ and finally $s_n - s_{n-1} = 0$, for all $n > n_0$. We conclude that $\delta = 2(n^* - 1) + (n_0 - n^* + 1) = n^* + n_0 - 1$. \square

3.2. The algorithm. For simplicity, we divide the exposition of our algorithm into two distinct cases.

Case 1: $n_0 = n^* - 1$. Observe that, by Lemma 3.8, this case requires the dimension of the group δ to be even. It includes the similarity Lie group $SIM(2)$ generated by dilations, rotations and translations in the plane, which is discussed in [MB01]. The distinguishing property of such group actions is that there are exactly two fundamental invariants defined in some neighborhood U of any point $x_0 \in (M^2)^{\times(n^*)}$ belonging to an orbit of maximal dimension $s_{n^*} = \delta$, as proved by Theorem 3.4 and Lemma 3.8. These invariants can be obtained using the moving frame normalization method.

LEMMA 3.9. *Assume $n_0 = n^* - 1$. Let $x_0 \in (M^2)^{\times(n_0+1)}$ be a point belonging to an orbit of maximal dimension $s_{n_0+1} = \delta$ and let $J_1, J_2 : U \subset (M^2)^{\times(n_0+1)} \rightarrow \mathbb{R}$ be a complete fundamental set of invariants defined in a neighborhood of x_0 . There exists a neighborhood $\bar{U} \subset U$ of x_0 such that on $\{(z_1, \dots, z_{n_0+1}) \in \bar{U}\}$, the last point z_{n_0+1} can be expressed as a function*

$$z_{n_0+1} = f(z_1, \dots, z_{n_0}, J_1(z_1, \dots, z_{n_0+1}), J_2(z_1, \dots, z_{n_0+1})).$$

PROOF. Consider the 2-by-2($n_0 + 1$) Jacobian matrix

$$\frac{\partial(J_1, J_2)}{\partial(z_1, \dots, z_{n_0+1})}$$

which, by functional independence of J_1 and J_2 , has rank two on an open and dense subset of U . Since there are no n_0 -point joint invariants, the 2-by-2 sub-matrix

$$\frac{\partial(J_1, J_2)}{\partial z_{n_0+1}}$$

must also have rank two on an open and dense subset of U and the conclusion follows by the implicit function theorem. \square

Assuming that g acts transitively on the restriction

$$\Pi_{1, \dots, n_0} \bar{U} = \{(z_1, \dots, z_{n_0}) \in (M^2)^{\times(n_0)} | \exists z_{n_0+1} \text{ s.t. } (z_1, \dots, z_{n_0+1}) \in \bar{U}\},$$

then we can simply use $I_1 = J_1$ and $I_2 = J_2$ to parameterize a signature for curve segments.

Given a curve segment $C = \{\alpha(t)\}$ with $n_0 = n^* - 1$ landmarks $(\zeta_1, \dots, \zeta_{n_0})$ in the direction of the parameterization and the same number of landmarks $(\lambda_1, \dots, \lambda_{n_0})$ in the other direction, we write these landmarks as $[(\zeta_1, \dots, \zeta_{n_0}), (\lambda_1, \dots, \lambda_{n_0})]$. We define the joint invariant signature of $\Sigma(C)$ of the curve segment C as the pair of planar curves

$$\Sigma(C) = (\Sigma_+(C), \Sigma_-(C))$$

where $\Sigma_+(C)$ is the curve parameterized by $I_1(\zeta_1, \dots, \zeta_{n_0}, \alpha(t))$ and $I_2(\zeta_1, \dots, \zeta_{n_0}, \alpha(1-t))$, while $\Sigma_-(C)$ is the curve parameterized by $I_1(\lambda_1, \dots, \lambda_{n_0}, \alpha(1-t))$ and $I_2(\lambda_1, \dots, \lambda_{n_0}, \alpha(1-t))$, for $0 \leq t \leq 1$. Let \mathbb{Z}_2 act on the signature $\Sigma(C)$ by permuting its two curve components.

THEOREM 3.10 (For curve segment recognition). *Let $\bar{U} \subset$ be an open set as described in Lemma 3.9. Assume that G acts transitively on the restriction $\Pi_{1, \dots, n_0} \bar{U}$. Consider two curve segments C and \bar{C} with landmarks $[(\zeta_1, \dots, \zeta_{n_0}), (\lambda_1, \dots, \lambda_{n_0})]$ and $[(\bar{\zeta}_1, \dots, \bar{\zeta}_{n_0}), (\bar{\lambda}_1, \dots, \bar{\lambda}_{n_0})]$ respectively. If, for all $0 \leq t \leq 1$, we have $(\zeta_1, \dots, \zeta_{n_0}, \alpha(t)), (\bar{\zeta}_1, \dots, \bar{\zeta}_{n_0}, \bar{\alpha}(t)) \in \bar{U}$, as well as $(\lambda_1, \dots, \lambda_{n_0}, \alpha(1-t)), (\bar{\lambda}_1, \dots, \bar{\lambda}_{n_0}, \bar{\alpha}(1-t)) \in \bar{U}$. Then $C \equiv \bar{C} \pmod{G}$ if and only if $\Sigma(C) \equiv \Sigma(\bar{C}) \pmod{\mathbb{Z}_2}$.*

PROOF. Invariance of the signature follows from the fact that it is parameterized by invariants and that the landmarks are equivariant up to a permutation of the ζ and λ components corresponding to reversing the direction of the parameterization.

To prove that the signature of C completely characterizes the equivalence class of C , observe that, by transitivity of the action of G on Ω , there always exists $g \in G$ such that $g \cdot (\zeta_1, \dots, \zeta_{n_0}) = (\bar{\zeta}_1, \dots, \bar{\zeta}_{n_0})$. Assuming that $\Sigma_+(C) = \Sigma_+(\bar{C})$ (reparameterize \bar{C} if necessary), then by Lemma 3.9, $g \cdot C = \bar{C}$. \square

We can also use $I_1 = J_1$ and $I_2 = J_2$ to parameterize a joint-invariant signature $\Sigma(C)$ for a closed curve C with $k \geq n_0 = n^* - 1$ landmarks $(\zeta_1, \dots, \zeta_k)$. Setting $(\zeta_{k+1}, \dots, \zeta_{2k}) = (\zeta_1, \dots, \zeta_k)$, we let $\Sigma(C) = (\Sigma_+(C), \Sigma_-(C))$ where

$$\Sigma_+(C) = (\Sigma_{1+}(C), \dots, \Sigma_{k+}(C))$$

is a string of k curves $\Sigma_{i+}(C)$ parameterized by $I_1(\zeta_i, \dots, \zeta_{n_0+i-1}, z)$ and $I_2(\zeta_i, \dots, \zeta_{n_0+i-1}, z)$ with $z \in C$ going from ζ_{n_0+i} to ζ_{n_0+i-1} and $\Sigma_-(C) = \Sigma_+(\{\alpha(1-t)\})$. As a corollary of 3.10, we have the following theorem.

THEOREM 3.11 (For closed curve recognition). *Let $\bar{U} \subset$ be an open set as described in Lemma 3.9. Assume that G acts transitively on the restriction $\Pi_{1, \dots, n_0} \bar{U}$. Consider two closed simple curves C and \bar{C} with $k \geq n_0$ landmarks $(\zeta_1, \dots, \zeta_k)$ and $(\bar{\zeta}_1, \dots, \bar{\zeta}_k)$ respectively. Setting $(\zeta_{k+1}, \dots, \zeta_{2k}) = (\zeta_1, \dots, \zeta_k)$ and $(\bar{\zeta}_{k+1}, \dots, \bar{\zeta}_{2k}) = (\bar{\zeta}_1, \dots, \bar{\zeta}_k)$, assume that, for all $i = 1, \dots, k$ and for all $z \in C$ between ζ_{n_0+i} and ζ_{n_0+i-1} we have $(\zeta_i, \dots, \zeta_{n_0+i-1}, z) \in \bar{U}$. Then $C \equiv \bar{C} \pmod{G}$ if and only if either $\Sigma_+(C) = \Sigma_+(\bar{C}) \pmod{\mathbb{Z}_k}$ or $\Sigma_-(C) = \Sigma_-(\bar{C}) \pmod{\mathbb{Z}_k}$.*

Case 2: $n_0 > n^* - 1$. In addition to the equi-affine group action on the plane discussed previously, this case includes the action of the (full/special) Euclidean group acting on the plane and the action of $SL(2)$ on the Poincaré half-plane which were discussed in [MB01]. It is characterized by the fact that there exists a single fundamental invariant J defined in some neighborhood of any point $x_0 \in (M^2)^{\times(n^*)}$ belonging to an orbit of maximal dimension $s_{n^*} \leq \delta$. Again, this invariant can be obtained following the moving frame normalization method.

LEMMA 3.12. *Assuming that $n_0 > n^* - 1$, let (z_1, \dots, z_{n_0}) be local coordinates for $(M^2)^{\times(n_0)}$ and let $\{J\}$ be a complete fundamental set of invariants defined in a neighborhood of a point $x_0 \in (M^2)^{\times(n_0)}$ belonging to an orbit of maximal dimension. There exists an open set $U_{n_0} \subset (M^2)^{\times(n_0)}$ such that*

$$\{J(z_i, \dots, z_{n^*+i-1})\}_{i=1}^{n_0-n^*+1}$$

is a complete fundamental set of invariants on U_{n_0} .

PROOF. Let $J_i = J(z_i, \dots, z_{n^*+i-1})$, for $i = 1, \dots, n_0 - n^* + 1$. By Lemma 3.8, the maximal orbit dimension $s_{n_0} = \delta = n_0 + n^* - 1$. So, by Theorem 3.4, there are $2n_0 - s_{n_0} = 2n_0 - (n_0 + n^* - 1) = n_0 - n^* + 1$ fundamental invariants in a neighborhood of x_0 . The conclusion follows by observing that the Jacobian matrix

$$\frac{\partial J_1, \dots, J_{n_0-n^*+1}}{\partial (z_1, \dots, z_{n_0})}.$$

has rank $n_0 - n^* + 1$ on an open and dense subset of an open subset U_{n_0} of $(M^2)^{\times(n_0)}$. \square

LEMMA 3.13. *Assuming that $n_0 \geq n^*$, let $\{J\}$ be a complete fundamental set of invariants defined in a neighborhood of a point $x_0 \in (\mathbb{R}^2)^{\times(n_0)}$ belonging to an orbit of maximal dimension. There exists an open set $U_{n_0+1} \subset (\mathbb{R}^2)^{\times(n_0+1)}$ and an invariant $H : U_{n_0+1} \rightarrow \mathbb{R}$ such that*

$$\{J(z_i, \dots, z_{n^*+i-1})\}_{i=1}^{n_0-n^*+2} \cup \{H(z_1, \dots, z_{n_0+1})\}$$

is a complete set of fundamental invariants on U_{n_0+1} , while

$$\{J(z_i, \dots, z_{n^*+i-1})\}_{i=1}^{n_0-n^*+1}$$

is a complete fundamental set of invariants on the restriction

$$\Pi_{1, \dots, n_0} U_{n_0+1} = \{(z_1, \dots, z_{n_0}) \in (M^2)^{\times(n_0)} \mid \exists z_{n_0+1} \text{ s.t. } (z_1, \dots, z_{n_0+1}) \in U_{n_0+1}\}.$$

Moreover, we can choose U_{n_0+1} such that, on U_{n_0+1} , the last point z_{n_0+1} is a function

$$z_{n_0+1} = f(z_1, \dots, z_{n_0}, J(z_{n_0-n^*+2}, \dots, z_{n_0+1}), H(z_1, \dots, z_{n_0+1})).$$

PROOF. By Lemma 3.12, there exists $U_{n_0} \subset (M^2)^{\times(n_0)}$ on which the set $\{J(z_i, \dots, z_{n^*+i-1})\}_{i=1}^{n_0-n^*+1}$ is a complete fundamental set of invariants.

Let $J_i(z_1, \dots, z_{n_0+1}) = J(z_i, \dots, z_{n^*+i-1})$. Observe that there exists an open subset of $(M^2)^{\times(n_0+1)}$ on which the rank of the Jacobian matrix

$$\frac{\partial (J_1, \dots, J_{n_0-n^*+2})}{\partial (z_1, \dots, z_{n_0+1})}$$

is equal to $n_0 - n^* + 2$. This open subset can be taken as $\bar{U}_{n_0} \times U_1$, where $\bar{U}_{n_0} \subset U_{n_0}$ and $U_1 \subset M^2$. The number of functionally independent invariants on $(\mathbb{R}^2)^{\times(n_0+1)}$ is

$$\begin{aligned} 2(n_0 + 1) - s_{n_0+1} &= 2(n_0 + 1) - \delta, \\ &= 2(n_0 + 1) - (n_0 + n^* - 1), \text{ by Lemma 3.8,} \\ &= n_0 - n^* + 3, \end{aligned}$$

so there exists another invariant H and an open subset $U_{n_0+1} \subset \bar{U}_{n_0} \times U_1$ such that $\{J_1, \dots, J_{n_0-n^*+2}, H\}$ is a complete fundamental set of invariants on U_{n_0+1} .

To show the second part of the statement, observe that the Jacobian matrix

$$\frac{\partial (J_{n_0-n^*+2}, H)}{\partial (z_1, \dots, z_{n_0+1})}$$

must have rank two, otherwise one could write an n_0 -point joint invariant I as

$$I(z_1, \dots, z_{n_0}) = f(J_{n_0-n^*+2}(z_1, \dots, z_{n_0+1}), H(z_1, \dots, z_{n_0+1})),$$

a function of $J_{n_0-n^*+2}$ and H , which would contradict the fact that $J_{n_0-n^*+2}$ and H are functionally independent of $\{J_1, \dots, J_{n_0-n^*+1}\}$, a complete fundamental set of n_0 -point joint invariant. The conclusion follows by the implicit function theorem. \square

Again, note that we can use the moving frame normalization method to obtain both the invariants J and H . We can use $I_1 = J_{n_0-n^*+2}$ and $I_2 = H$ to parameterize a signature characterizing closed curves or curve segments modulo G . The recognition algorithms are just slightly different than those of Case 1.

Given a curve segment $C = \{\alpha(t)\}$ with n_0 landmarks $(\zeta_1, \dots, \zeta_{n_0})$ in the direction of the parameterization and the same number of landmarks $(\lambda_1, \dots, \lambda_{n_0})$ in the other direction, we define its signature as the pair $\Sigma(C) = (\Sigma_+(C), \Sigma_-(C))$, where

$$\begin{aligned} \Sigma_+(C) = & \left\{ (J(\zeta_1, \dots, \zeta_{n^*}), J(\zeta_2, \dots, \zeta_{n^*+1}), \dots, J(\zeta_{n_0-n^*+1}, \dots, \zeta_{n_0})), \right. \\ & \left. \{J(\zeta_{n_0-n^*+2}, \dots, \zeta_{n_0}, \alpha(t)), H(\zeta_1, \dots, \zeta_{n_0}, \alpha(t))\}_{0 \leq t \leq 1} \right\} \end{aligned}$$

and $\Sigma_-(C) = \Sigma_+(\{\alpha(1-t)\}_{0 \leq t \leq 1})$. So each component of the signature is a set containing the string of $n_0 - n^* + 1$ real numbers given by evaluating J on all n^* consecutive landmarks together with the curve parameterized by $J(\zeta_{n_0-n^*+2}, \dots, \zeta_{n_0}, z)$ and $H(\zeta_1, \dots, \zeta_{n_0}, z)$ with z varying along the curve. Let \mathbb{Z}_2 act on the signature $\Sigma(C)$ by permuting $\Sigma_+(C)$ and $\Sigma_-(C)$.

THEOREM 3.14 (For curve segment recognition). *Assuming that $n_0 > n^* + 1$, let J , H and U_{n_0+1} be as defined in Lemma 3.13. Consider two curve segments C and \bar{C} with landmarks $[(\zeta_1, \dots, \zeta_{n_0}), (\lambda_1, \dots, \lambda_{n_0})]$ and $[(\bar{\zeta}_1, \dots, \bar{\zeta}_{n_0}), (\bar{\lambda}_1, \dots, \bar{\lambda}_{n_0})]$ respectively. If, for all $0 \leq t \leq 1$, we have*

$$\begin{aligned} (\zeta_1, \dots, \zeta_{n_0}, \alpha(t)), (\bar{\zeta}_1, \dots, \bar{\zeta}_{n_0}, \bar{\alpha}(t)) &\in U_{n_0+1} \\ \text{and } (\lambda_1, \dots, \lambda_{n_0}, \alpha(1-t)), (\bar{\lambda}_1, \dots, \bar{\lambda}_{n_0}, \bar{\alpha}(1-t)) &\in U_{n_0+1}, \end{aligned}$$

then $C \equiv \bar{C} \pmod{G}$ if and only if $\Sigma(C) \equiv \Sigma(\bar{C}) \pmod{\mathbb{Z}_2}$.

PROOF. Invariance modulo \mathbb{Z}_2 follows from the invariance of the construction up to the choice of direction for the parameterization.

Assume that $\Sigma(C) = \Sigma(\bar{C})$ (reparameterize one of the curve segments if necessary). This implies that the vector value of

$$\begin{aligned} &(J(\zeta_1, \dots, \zeta_{n^*}), J(\zeta_2, \dots, \zeta_{n^*+1}), \dots, J(\zeta_{n_0-n^*+1}, \dots, \zeta_{n_0})) \\ &\text{and } (J(\bar{\zeta}_1, \dots, \bar{\zeta}_{n^*}), J(\bar{\zeta}_2, \dots, \bar{\zeta}_{n^*+1}), \dots, J(\bar{\zeta}_{n_0-n^*+1}, \dots, \bar{\zeta}_{n_0})) \end{aligned}$$

are equal. By Lemma 3.12, this means that there exists $g \in G$ such that $g \cdot (\zeta_1, \dots, \zeta_{n_0}) = (\bar{\zeta}_1, \dots, \bar{\zeta}_{n_0})$. By Lemma 3.13, we have $g \cdot C = \bar{C}$. \square

Given a closed curve segments $C = \{\alpha(t)\}$ with $k \geq n_0$ landmarks, we define its signature as $\Sigma(C) = (\Sigma_+(C), \Sigma_-(C))$ where $\Sigma_+(C)$ is the set of k curves

$$\Sigma_+(C) = (\Sigma_{1+}(C), \dots, \Sigma_{k+}(C))$$

parameterized by

$$\Sigma_{i+} = (J(\zeta_{i+n_0-n^*}, \dots, \zeta_{n_0+i-1}, z)H(\zeta_i, \dots, \zeta_{n_0+i-1}, z))$$

for $z \in C$ going from ζ_{n_0+i} to ζ_{n_0+i-1} . As a corollary of 3.14, we have the following theorem.

THEOREM 3.15 (For closed curve recognition). *Assuming that $n_0 > n^* + 1$, let J , H and U_{n_0+1} be as defined in Lemma 3.13. Consider two closed (simple) curves $C = \{\alpha(t)\}$ and $\bar{C} = \{\bar{\alpha}(t)\}$ with $k \geq n_0$ landmarks $(\zeta_1, \dots, \zeta_k)$ and $(\bar{\zeta}_1, \dots, \bar{\zeta}_k)$ respectively. Set $(\zeta_{k+1}, \dots, \zeta_{2k}) = (\zeta_1, \dots, \zeta_k)$ and $(\bar{\zeta}_{k+1}, \dots, \bar{\zeta}_{2k}) = (\bar{\zeta}_1, \dots, \bar{\zeta}_k)$. If, for all $0 \leq t \leq 1$ and all $i = 1, \dots, k$, we have*

$$(\zeta_i, \dots, \zeta_{n^*+i-1}, \alpha(t)), (\bar{\zeta}_i, \dots, \bar{\zeta}_{n^*+i-1}, \bar{\alpha}(t)) \in U_{n_0+1},$$

then $C \equiv \bar{C} \pmod{G}$ if and only if $\Sigma_+(C) \equiv \Sigma_+(\bar{C}) \pmod{\mathbb{Z}_k}$ or $\Sigma_+(C) \equiv \Sigma_-(\bar{C}) \pmod{\mathbb{Z}_k}$.

Observe that, in all cases presented, the main tricks used consisted in finding two invariants $I_1(z_1, \dots, z_n), I_2(z_1, \dots, z_n)$ such that given z_1, \dots, z_{n-1} , then the last point z_n is uniquely prescribed by the value of $I_1(z_1, \dots, z_n)$ and $I_2(z_1, \dots, z_n)$. The natural question to ask, of course, is whether one could build a signature in a similar fashion using invariants that depend on less than $n_0 + 1$ points. Unfortunately, there is no way to repeat our trick with such invariants.

LEMMA 3.16. *Whenever $n < n_0 + 1$, there do not exist two invariants $I_1, I_2 : U_n \subset (M^2)^{\times(n)} \rightarrow \mathbb{R}$ such that, on U_n , the last point z_n can be expressed as a function*

$$z_n = f(z_1, \dots, z_{n-1}, I_1(z_1, \dots, z_n), I_2(z_1, \dots, z_n)).$$

PROOF. This is because the rank of the Jacobian matrix

$$\mathcal{J} = \frac{\partial(I_1, I_2)}{\partial z_n}$$

is never equal to two on an open and dense subset of any open subset of $(M^2)^{\times(n)}$, since the number $\#_n$ of n -point fundamental invariants is $\#_n - \#_{n-1} = 0$ or 1, but never 2, as discussed in the proof of Lemma 3.8. \square

4. Conclusions

Based on a method described in a warm up paper [MB01] on polygon recognition, we proved the existence of two suitable joint invariants which can be used for parameterizing a signature curve for curve segments or closed curves in a two-dimensional manifold. These two invariants can be obtained by the moving frame method. The signature curve is such that two curves are equivalent modulo a Lie group G if and only if their signature is the same up to the choice of orientation of the curve. We provided a full solution for the group of area preserving planar transformations (equi-affine). Suitable invariants for other important cases, including the Euclidean group and the similarity group (scaling, rotations and translations), are given in the warm up paper. The construction of our signature requires a minimal number of landmarks (n_0 , the stabilization order of the prolonged group action on many copies of a manifold), which can themselves be used for testing the equivalence of curves, and thus ruling out the unlikely candidates.

The two main advantages of this method are: the invariants used are more robust than differential invariants and the dimension of the signature is optimal. For lack of space, we did not discuss the use of this signature for G -symmetry detection, although this is just a subproblem of curve recognition. However, a method for symmetry detection in curves can be easily deduced from the method for symmetry detection in polygons described in details in the warm up paper.

ACKNOWLEDGMENTS

I want to thank my advisor Peter Olver for encouragements, support and for many useful comments. I am grateful to the Graduate School of the University of Minnesota for providing me with financial support through a doctoral dissertation fellowship and to the mathematics department of Purdue University for its hospitality during part of the period in which I wrote this paper.

References

- [BSTG] S.I. Betelu, G. Sapiro, A. Tannenbaum and P.J. Giblin, *On the computation of Affine Skeletons of Plane Curves and the Detection of Skew Symmetries*, Pattern Recognition, **34**:5 (2001), 943–952.
- [MB00] M. Boutin, *Numerically invariant signature curves*, Int. J. Comp. Vision, **44**:3 (2000), 235–248.
- [MB01] ———, *Polygon recognition and symmetry detection*, Preprint (2001), University of Minnesota.
- [MB02] ———, *On orbit dimensions under a simultaneous Lie group action on n copies of a manifold*, J. Lie theory, **12** (2002), 191–203.
- [COSTH] E. Calabi, P.J. Olver, C. Shakiban, A. Tannenbaum and S. Haker, *Differential and Numerically Invariant Signature Curves Applied to Object Recognition*, Int. J. Comput. Vision, **26** (1998), 107–135.
- [C] É. Cartan, *La méthode du repère mobile, la théorie des groupes continus et les espaces généralisés*, Exposés de géométrie No. 5, Hermann, Paris, 1935.
- [F] G. Frobenius, *Über das Pfaff'sche Problem*, J. Reine Angew. Math., **82** (1877), 230–315.
- [G] H.W. Guggenheimer, *Differential Geometry*, Dover Books on Advanced Mathematics, Dover Publications, New York, 1977.
- [O95] P.J. Olver, *Equivalence, Invariants and Symmetry*, Cambridge University Press, Cambridge, 1995.
- [O99] ———, *Classical Invariant Theory*, London Mathematical Society Student Text, 44. Cambridge University Press, Cambridge, 1999.
- [O01] ———, *Joint Invariants Signatures*, Found. Comput. Math., **1**:1 (2001), 3–67.

P.O. Box F, PROVIDENCE, RI, USA, 02912

E-mail address: mimi@dam.brown.edu

Inpainting Based on Nonlinear Transport and Diffusion

Tony F. Chan and Jianhong Shen

ABSTRACT. Inpainting is an inverse problem in image restoration, with wide applications in image processing and lower-level vision analysis. It refers to the process of filling in the missing image information where it is lost, destroyed, or blocked, etc. Modern digital inpainting techniques can be classified into two interdependent methods: the variational approach and the PDE approach. This paper focuses on the latter and intends to explore the role of the (non-linear) transport and diffusion mechanisms for inpainting modeling. Recent empirical works on these two separate mechanisms are first reviewed, based upon which, we attempt to derive (via quasi-axiomatization) a new class of third order nonlinear PDEs for image inpainting, whose rationality still needs further criticisms from the mathematics community.

1. Introduction

The word “inpainting” is an artistic synonym for “image interpolation,” as originally circulated among museum restoration artists, who manually fill in the missing patches of cracked ancient paintings without leaving any visible trace of modification [5, 16, 34].

The concept of *digital inpainting* was first transplanted into digital image processing in the paper of Bertalmio, Sapiro, Caselles, and Ballester [5]. It now generally refers to any restoration problem that mainly involves interpolation of unavailable image information, and has found broad applications in the digital and communication technology, as well as in lower-level vision analysis [3, 5, 11, 12, 14, 15, 21, 22, 28, 33, 35].

Conventional tools for inpainting and interpolation have been very diversified, mainly including the filtering method, the Bayesian method, wavelets and spectral methods, the learning-and-growing method, etc. We are still able to witness many works in this area (see for example, [1, 6, 20, 23, 35]).

It is only quite recently that the variational and PDE methods have been introduced as new competing tools for image inpainting [3, 4, 7, 11, 12, 17, 25, 28]. These two methods on one hand are closely connected and complement each

1991 *Mathematics Subject Classification*. Primary 94A08; Secondary 68U10, 65K10.

Key words and phrases. Inpainting, PDE, transport, diffusion, curvature, morphological invariance.

Research has been supported by both NSF and ONR..

other, and on the other, indeed bear their own identities. We refer to our recent survey papers [13, 32] for a more detailed discussion.

This paper is focused on the PDE approach. More specifically, we study two infinitesimal mechanisms for extending available information onto blank domains: transport and diffusion. The nonlinear transport method first appeared in the remarkable paper by Bertalmio, et al. [5], while the nonlinear diffusion idea was developed later by the authors of the present paper [11, 12]. In section 2, we shall give a brief introduction to these two models and discuss their advantages and shortcomings as well.

The main result of the current paper is in Section 3 where we intend to *axiomatize* these two empirical works [5, 11, 12]. As inspired by our recent work in [9], we propose four axioms for establishing a new class of third order nonlinear geometric PDEs for image inpainting, which combine both transport and diffusion. Due to nonlinearity, our approach is only *quasi*-axiomatic in a sense explained in Section 3. Many of the similar ideas therein have also been developed in earlier works in the axiomatic scale-space theory (see for example [2]).

Among the four axioms, we first preview the one that involves the so-called *morphological invariance* (or as some authors prefer, the *contrast* invariance). Morphological invariance means that the way one connects a broken isophote (i.e. level line) should be independent of its gray value. Vividly speaking, given a damaged ancient painting with cracked regions to be restored, the way a restoration artist inpaints it should be independent of whether he stays inside a dim room or sits in the sun (since the pigments that the artist uses undergo the same luminance change). Morphological invariance is thus a very natural principle in image processing [2]. Its mathematical meaning can be put down rigorously. Let u_0 be the given observed data (for inpainting, u_0 is simply a portion of the complete image). An image processing tool (or *operator*) $T : u = T(u_0)$ is said to be morphologically invariant, if for any strictly increasing function (or *morphological transform*) $g : [0, 1] \rightarrow [0, 1]$, one always has $T(g(u_0)) = g(T(u_0))$.

Throughout the paper, $\Omega \subset R^2$ denotes the entire open image domain, which is bounded and rectangular for most digital applications, $D \subset \Omega$ a compact sub-domain to be inpainted, u^0 the available part of the image on $\Omega \setminus D$, and u the targeted inpainted output. The standardized symbols ∇ , $\nabla \cdot$ and Δ represent the gradient, divergence, and Laplacian operators separately. Conforming to computer science, we shall adopt the terminology *isophotes* for level sets or level lines.

2. PDE Inpainting by Transport and Diffusion

In this section, we review both the empirical work by Bertalmio et al. [5] on transport based inpainting, and that of Chan and Shen [11, 12] on diffusion based inpainting.

2.1. Transport based image inpainting. The first high order PDE inpainting model of Bertalmio, Sapiro, Caselles and Ballester [5] is based on the beautiful intuition of *smoothness transport* along isophotes. Imagine to inpaint an ideal step edge as in Fig. 1, one naturally requests the intensity jump to propagate along the edge, so that a sharp edge can be restored. Generally, Let $L(u)$ be a smoothness measure of an image u . For example, a second order smoothness measure can be

expressed in the general form of

$$L(u) = f(\nabla u, \nabla \otimes \nabla u),$$

where ∇u is the gradient vector, and $\nabla \otimes \nabla u$ the Hessian. The one experimented in [5] is the Laplacian:

$$L = \Delta u = \text{trace}(\nabla \otimes \nabla u).$$

The Bertalmio-Sapiro-Caselles-Ballester inpainting model is then defined by a third order evolutionary equation:

$$(1) \quad \frac{\partial u}{\partial t} = \nabla^\perp u \cdot \nabla L(u),$$

where, $\nabla^\perp u = (-u_y, u_x) = |\nabla u| \vec{t}$ points to the tangent \vec{t} . The model carries the transport (or propagation) nature since as the evolution approaches its equilibrium state, we have (as long as $|\nabla u| \neq 0$)

$$(2) \quad \vec{t} \cdot \nabla L(u) = 0 \text{ or equivalently } \frac{\partial L(u)}{\partial \vec{t}} = 0,$$

which means, along an isophote, the smoothness measure is conserved. Thus in terms of the available boundary data, the inpainting process evolves like transporting the boundary smoothness information along the extended isophotes into the inpainting domain (see Fig. 1).

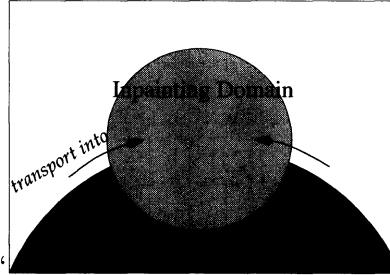


FIGURE 1. The transport model of Bertalmio, Sapiro, Caselles, and Ballester [5].

However, due to the lack of communication among the isophotes, the transport may result in kinks inside inpainting domains, just as shocks may develop in traffic models. Thus in [5], Eq. (1) is implemented with the help of intermediate steps of anisotropic diffusions. Recently, the authors have borrowed some beautiful ideas from vortex dynamics to stabilize transport with diffusion [4].

The second issue with model (1) is the smoothness measure L . The choice of the Laplacian is convenient but less ideal in two aspects:

- (a) The equilibrium equation $\partial(\Delta u)/\partial \vec{t} = 0$ is not morphologically invariant. Let

$$g(\lambda) : [0, 1] \rightarrow [0, 1]$$

be a smooth morphological transform so that $g'(\lambda) > 0$, then

$$(3) \quad \Delta g(u) = g'(u)\Delta u + g''(u)(\nabla u)^2.$$

Thus

$$\frac{\partial(\Delta g(u))}{\partial \vec{t}} = g'(u) \frac{\partial(\Delta u)}{\partial \vec{t}} + g''(u) \frac{\partial(\nabla u)^2}{\partial \vec{t}}.$$

(Notice that u is constant along \vec{t} .) Hence, generally, if u is the final equilibrium inpainting to a given image $u_0|_{\Omega \setminus D}$, then $g(u)$ is not the equilibrium inpainting to $g(u_0)|_{\Omega \setminus D}$ due to the second term (unless that $g(\lambda)$ is a *linear* scaling: $g(\lambda) = a + b\lambda$).

- (b) For the equilibrium inpainting u , according to (2), the smoothness measure $L(u)$ must be constant along the isophotes. Therefore, if p and q are two pixels along the inpainting boundary and belong to the same isophote, but with different L values computed from the available data $u_0|_{\Omega \setminus D}$, then theoretically there can be no equilibrium inpainting. Such situation often occurs in large-scale inpainting problems, which is not caused by noise, but instead, by the natural variation of L itself along an isophote. Thus asking L to be a constant along the isophotes, like the gray value u itself, is perhaps demanding too much.

The new model we derive in Section 3 shall remedy these drawbacks.

2.2. Diffusion based image inpainting. In this section, we introduce our recent works on image inpainting based on nonlinear diffusions [11, 12].

In [11], as inspired by the Bayesian approach [27] and the celebrated restoration model of Rudin, Osher, and Fatemi [31], we treat inpainting as an inverse problem, for which we minimize the *posterior* “energy” functional for the given u^0 and D :

$$(4) \quad E[u] = \int_{\Omega} |\nabla u| dx + \frac{\lambda}{2} \int_{\Omega \setminus D} (u - u^0)^2 dx,$$

where $dx = dx_1 dx_2$ is the 2-D area element, and λ a fitting weight or Lagrange multiplier, which is inversely proportional to the variation of the noise [10, 14, 30]. Mathematically speaking, we are assuming that the original complete and clean image $u \in BV(\Omega)$ (the functional space of bounded variations [8, 14]), and the formal Sobolev norm of the first term in (4) should be replaced by the TV Radon measure $\int_{\Omega} |Du|$ [19].

The steepest descent minimization scheme for $E[u|u^0, D]$ is given by

$$(5) \quad \frac{\partial u}{\partial t} = \nabla \cdot \left[\frac{\nabla u}{|\nabla u|} \right] - \lambda_e(u - u^0),$$

valid on the entire image domain Ω , and the extended Lagrange multiplier $\lambda_e(x) = \lambda \cdot 1_{\Omega \setminus D}(x)$. The boundary integral coming from integration-by-parts then leads to the natural adiabatic boundary condition $\partial u / \partial \vec{n} = 0$, with \vec{n} denoting the normal direction along $\partial\Omega$. The marching can start with any reasonable initial guess [11]. Now inside the inpainting domain, the model employs a simple anisotropic diffusion process,

$$(6) \quad \frac{\partial u}{\partial t} = \nabla \cdot \left[\frac{\nabla u}{|\nabla u|} \right].$$

The application of anisotropic diffusions in image denoising and enhancement now has become a classical topic since Perona and Malik [29] (also see [36, 26]). It is however fresh for the application to the inpainting problem.

From Eq. (5), in the absence of noise (i.e., $\lambda_e = \infty$ outside the inpainting domain), the equilibrium inpainting of the TV model is indeed morphologically invariant since the right hand side of Eq. (6) is exactly the curvature of the isophotes.

On the other hand, if one requires the entire evolution (6) to be morphologically invariant, then the factor $|\nabla u|$ can be brought in to balance the time derivative:

$$(7) \quad \frac{\partial u}{\partial t} = |\nabla u| \nabla \cdot \left[\frac{\nabla u}{|\nabla u|} \right],$$

which is exactly the *mean curvature motion* [2, 18], and is also very useful for the speeding-up of numerical convergence, as recently studied by Marquina and Osher [24].



FIGURE 2. An example of TV inpainting for text removal.

Fig. 2 shows one example of TV inpainting for text removal. After being properly digitized, the model also finds its successful applications in digital zooming and edge-based image coding schemes. More details can be found in Chan and Shen [11]. Fig. 3 shows how the TV model can reconstruct a whole image based only on the intensity values along a narrow tube (1 or 2-pixel wide digitally) surrounding the detected edges.

There are also two major drawbacks. The first one is that the TV model is only a linear interpolant, i.e., the broken isophotes are interpolated by straight lines. Thus it can generate corners along the inpainting boundary. The second one is that TV often fails to connect *widely* separated parts of a whole object, due to the high cost in TV measure of making long-distance connections [11, 9] (see Fig. 4).

Such a violation of the so-called *connectivity principle* [11] inspired the CDD (curvature driven diffusion) inpainting model of Chan and Shen [12]. CDD inpainting further refines the TV anisotropic diffusion (5). To encourage long-distance connections, CDD also employs the curvature information for the diffusion. It is based on the simple observation (such as from Fig. 4) that when the TV model gets reluctant in making connection, the edge isophotes typically contain corners (a, b, c, d in Fig. 4) which has large curvatures. The good news is that large curvatures can be incorporated into the diffusion process to “push” out the false edges (ab and cd in Fig. 4) formed in the TV diffusion:

$$(8) \quad \frac{\partial u}{\partial t} = \nabla \cdot \left(\frac{g(\kappa)}{|\nabla u|} \nabla u \right), \quad \kappa = \nabla \cdot \left[\frac{\nabla u}{|\nabla u|} \right],$$

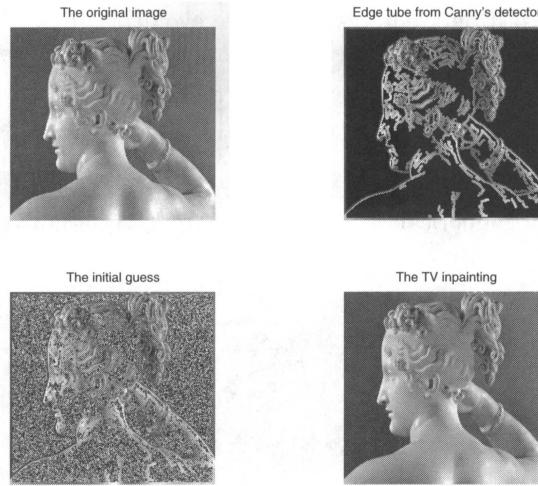


FIGURE 3. An example of TV inpainting for edge decoding. (Image source: test image of the Computational Vision Lab at California Institute of Technology.)

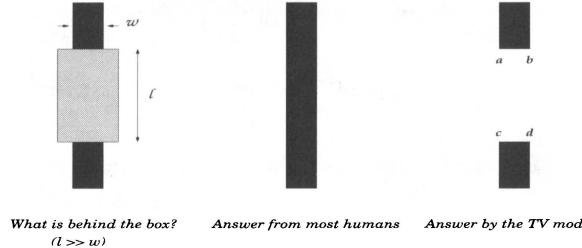


FIGURE 4. TV fails to realize the *Connectivity Principle* on inpainting images with large scale (or *aspect ratio*) missing domains.

where $g : R \rightarrow [0, +\infty)$ is a continuous function satisfying $g(0) = 0$ and $g(\pm\infty) = +\infty$. The introduction of $g(\kappa)$ is to penalize large curvatures and encourage small ones (or flatter and smoother isophotes), since $D = g(\kappa)/|\nabla u|$ denotes the diffusion strength. A simple example would be $g(s) = |s|^p$ for some positive power p . Fig. 5 shows one example of CDD inpainting, where even very weak edges are connected successfully (like the shadow of the nose).

In application, the fidelity term as in (5) is added to make the inpainting process robust to noise. In addition, since the curvature is expected to play a role only over the inpainting domain, outside, we can still employ Rudin-Osher-Fatemi's second-order TV denoising model [31], to speed up computation. In summary, as proposed in [12], a more practical and efficient model is the two-phase formulation

$$(9) \quad \frac{\partial u}{\partial t} = \nabla \cdot \left[\frac{G(x, \kappa)}{|\nabla u|} \nabla u \right] - \lambda_e(u - u^0).$$

Here the diffusivity coefficient has two phases:

$$G(x, \kappa) = 1_{\Omega \setminus D}(x) + g(\kappa)1_D(x).$$

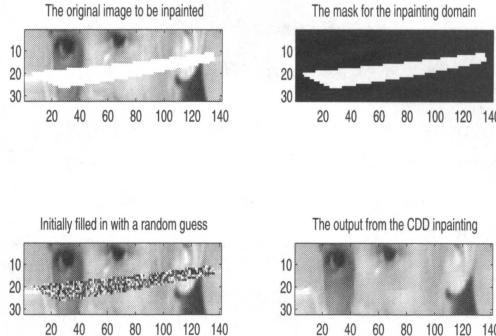


FIGURE 5. An example of CDD inpainting for scratch removal.

The associated boundary condition is still adiabatic $\partial u / \partial \vec{n} = 0$.

CDD inpainting (8) is a third order PDE model, and is indeed morphologically invariant since both the curvature κ and the normal vector \vec{n} are. The model encourages long-distance connections. But one drawback of the TV inpainting model still stays. That is, the isophotes are still approximated by straight lines.

3. Combining Transport and Diffusion via Quasi-Axiomatization

Based on the above two empirical works, in this section, we derive a new third order inpainting PDE from a set of principles or axioms. From the mathematical point of view, the axiomatic approach is an important step in the big blueprint for putting image processing on a firm mathematical foundation. Previous remarkable works can be found in [2, 7] for the axiomatic approach to the scale-space theory. In fact, most similar results in Section 3.1 also appeared in [2, 7] in the context of adaptive image denoising, enhancement, and the notion of scale-space. However, as we shall see below that the major difference between inpainting and the conventional scale-space theory is the transport mechanism in Section 3.2. Transport is prohibited in the conventional scale-space theory.

3.1. Geometric representation of differentials: curvature κ and its conjugate σ . Given an image u , its Cartesian differentials up to the second order are given by

$$(10) \quad \nabla u = \begin{pmatrix} u_x \\ u_y \end{pmatrix}, \quad \nabla \otimes \nabla u = \begin{bmatrix} u_{xx} & u_{xy} \\ u_{yx} & u_{yy} \end{bmatrix}.$$

They are easy to compute if the image u is given in the x and y coordinates, yet less ideal from the invariant (or geometric) point of view. For example, if the observer rotates by some angle, then both ∇u and $\nabla \otimes \nabla u$ change. Another simpler reason for the less idealness is that for a given image u , the x and y directions have no specific significance as far as the visual information is concerned.

However, near a regular pixel of a given image u , we do have two orthogonal directions that come naturally with the image itself: the normal \vec{n} and the tangent \vec{t} of the isophotes. Let $\mathbf{p} = \nabla u = p \vec{n}$ ($p \geq 0$) and $H = \nabla \otimes \nabla u$ denote the Cartesian differentials. We make the following transform from $R^2 \setminus \{(0,0)\} \times R^{2 \times 2}$

to $R^+ \times S^1 \times R^{2 \times 2}$:

$$(11) \quad (\mathbf{p}, H) \rightarrow \left(p, \vec{n}, \frac{1}{p} [\vec{t}, \vec{n}]^T H[\vec{t}, \vec{n}] \right) = (p, \vec{n}, G).$$

Obviously the transform is invertible and smooth (where $p \neq 0$). The transformed differentials have nicer geometric and morphological properties:

- (a) $p = |\mathbf{p}| = |\nabla u|$ is rotationally invariant while \vec{n} is morphologically invariant. and even more importantly,
- (b) The new second order differential matrix G carries much more explicit geometric information about the image. The first diagonal of G :

$$\frac{1}{p} \vec{t}^T H \vec{t} = \frac{1}{|\nabla u|} (\nabla \otimes \nabla u)(\vec{t}, \vec{t}) := \kappa$$

is exactly the scalar curvature of the oriented (by the gradient) isophotes. It is a geometric quantity characterizing each individual isophote, and thus is both rotationally and morphologically invariant. The off-diagonal of G :

$$\frac{1}{p} \vec{t}^T H \vec{n} = \frac{1}{|\nabla u|} (\nabla \otimes \nabla u)(\vec{t}, \vec{n}) := \sigma$$

is also both rotationally and morphologically invariant, and yet has played almost no role in the classical scale-space or filtering theory due to the ellipticity constraint [2, 7]. However, as shown below, for inpainting, it can be an important ingredient for the transport mechanism. In this paper, we shall call this scalar the *conjugate curvature* of the image.¹ It can easily be shown that

$$\sigma = \frac{1}{|\nabla u|} \frac{\partial |\nabla u|}{\partial \vec{t}} = \frac{\partial (\ln |\nabla u|)}{\partial \vec{t}},$$

from which the rotational and morphological invariance is immediate. (To the best knowledge of the authors, the quantity σ was first mentioned in the classical paper of Rudin and Osher [30].) The last diagonal of G :

$$\tau = \frac{1}{p} \vec{n}^T H \vec{n} = \frac{1}{|\nabla u|} (\nabla \otimes \nabla u)(\vec{n}, \vec{n}) = \frac{1}{|\nabla u|} \Delta u - \kappa$$

is only rotationally invariant and generally not morphologically invariant due to formula (3).

By considering images u of the general quadratic form

$$u = \frac{1}{2} \mathbf{x}^T H \mathbf{x} + \mathbf{p}^T \mathbf{x} + c,$$

we can easily establish the following theorem based on the new set of geometric coordinates: $p, \vec{n}, \kappa, \sigma, \tau$, as in (11).

THEOREM 1. *Let $f = f(\nabla u, \nabla \otimes \nabla u)$ be a function with up to the second order differentials. Then f is morphologically invariant if and only if it can be written in the form of*

$$f = f(\vec{n}, \kappa, \sigma).$$

If furthermore, f is also rotationally invariant, then

$$f = f(\kappa, \sigma).$$

¹The name “conjugate” has been motivated by the following consideration. Suppose that u is harmonic, and v denotes its conjugate, i.e., v solves the Cauchy-Riemann equations $v_x = u_y$, $v_y = -u_x$, then σ is exactly the curvature field of v : $\sigma = \nabla \cdot (\nabla v / |\nabla v|)$.

That is, f is both rotationally and morphologically invariant if and only if it is a function of the curvature κ and its conjugate σ .

3.2. The quasi-axiomatic approach to a class of third order inpainting PDEs. As inspired by the variational inpainting techniques studied in [9, 11], we look for a third order inpainting PDE in the divergence form

$$\frac{\partial u}{\partial t} = \nabla \cdot \vec{V}.$$

Therefore, the flux field \vec{V} shall be of second order only,

$$\vec{V} = \vec{V}(\nabla u, \nabla \otimes \nabla u).$$

It can be naturally decomposed in the normal and tangential directions,

$$\vec{V} = f \vec{n} + g \vec{t}$$

with

$$f = f(\nabla u, \nabla \otimes \nabla u), \quad g = g(\nabla u, \nabla \otimes \nabla u).$$

Our goal is to establish a new class of inpainting models given in this form that naturally combines transport and diffusion.

Axiom 1: Morphological invariance.

This first axiom requires that the equilibrium equation $0 = \nabla \cdot \vec{V}$ be morphologically invariant. Since both \vec{n} and \vec{t} are already morphologically invariant, it amounts to saying that

both f and g are morphologically invariant.

Then by Theorem 1, we must have

$$(12) \quad f = f(\vec{n}, \kappa, \sigma) \quad \text{and} \quad g = g(\vec{n}, \kappa, \sigma).$$

Axiom 2: Rotational invariance.

The second axiom requires that the equilibrium equation $0 = \nabla \cdot \vec{V}$ is rotationally invariant. Since all the following scalars and operators are rotationally invariant:

$$\nabla \cdot \vec{n}, \quad \nabla \cdot \vec{t}, \quad \vec{n} \cdot \nabla, \quad \text{and} \quad \vec{t} \cdot \nabla,$$

it requires that both f and g are rotationally invariant. Therefore by Theorem 1, we must have

$$(13) \quad f = f(\kappa, \sigma) \quad \text{and} \quad g = g(\kappa, \sigma).$$

The following two axioms or principles are imposed on the normal flux $f\vec{n}$ and tangential flux $g\vec{t}$ individually. It is more inspired by all the practical approaches discussed above, rather than by the strict *superposition principle*, which we do not have due to nonlinearity. It is for this reason that we say our approach is *quasi-axiomatic*. We speculate that making some rigorous mathematical analysis on these highly nonlinear PDEs can be very challenging.

Axiom 3: Stability principle for the pure diffusion.

As well known in the PDE theory, backward diffusion is unstable. Thus this principle asks for the stability of the pure diffusion term

$$\frac{\partial u}{\partial t} = \nabla \cdot (f \vec{n}) = \nabla \cdot \left(\frac{f}{|\nabla u|} \nabla u \right).$$

Stability requires that $f \geq 0$, or the strong stability $f \geq a > 0$, as in the elastica inpainting model of Masnou and Morel [25], and Chan, Kang, and Shen [9].

Axiom 4: Linear interpolation principle for the pure transport.

For the pure transport term

$$(14) \quad \frac{\partial u}{\partial t} = \nabla \cdot (g \vec{t}),$$

as learned from the drawback of Bertalmio et al.'s model [5], we now impose the *linear interpolation* constraint. First notice that

$$\begin{aligned} \nabla \cdot (g \vec{t}) &= \nabla \cdot ((g|\nabla u|) \nabla^\perp u) = \nabla^\perp u \cdot \nabla(g|\nabla u|) \\ &= |\nabla u| \vec{t} \cdot \nabla(g|\nabla u|) = |\nabla u| \frac{\partial}{\partial \vec{t}}(g|\nabla u|). \end{aligned}$$

By the current principle we mean that there must exist some smoothness measure L so that

$$(15) \quad g|\nabla u| = \frac{\partial L}{\partial \vec{t}}.$$

With this, it is guaranteed that the equilibrium solution to (14) satisfies

$$(16) \quad \frac{\partial^2 L}{\partial \vec{t}^2} = 0.$$

Thus along any inpainted isophote, L must be linear: $L = a + bs$, where s denotes the arc-length parameter of the isophote, and a and b are two constants that are determined by the L values at the two boundary pixels.

As discussed in Section 2.1, Bertalmio et al.'s pure transport model demands a constant value for the smoothness measure along any inpainted isophote. But for a generic image function u , the level lines of u are generally different from those of L . Therefore, it appears to us that the linearity condition (16) is more natural and feasible. The issue now is whether this axiom is compatible to all the previous constraints and three axioms, so that a consistent model can indeed be established.

Since g is a second order feature, by (15) L must only involve the first order differential ∇u , or $L = L(\nabla u)$. Furthermore, since g , $|\nabla u|$, and $\partial/\partial \vec{t}$ are all rotationally invariant, so must be L by (15), which means $L = L(|\nabla u|)$. Therefore,

$$g = \frac{1}{|\nabla u|} \frac{\partial L(|\nabla u|)}{\partial \vec{t}} = L'(|\nabla u|) \frac{1}{|\nabla u|} \frac{\partial |\nabla u|}{\partial \vec{t}} = L'(|\nabla u|) \sigma,$$

where σ is the conjugate curvature. Together with Eq. (13), it implies that

$$L'(|\nabla u|) = a, \quad \text{a constant.}$$

(Notice that g/σ is a function of κ and σ only, and from the transform (11), generically, $p = |\nabla u|$ is independent of κ and σ .) Therefore,

$$L(|\nabla u|) = a|\nabla u| + b, \quad \text{and} \quad g = a \sigma = a \frac{\partial (\ln |\nabla u|)}{\partial \vec{t}}.$$

In summary, we have established the following theorem.

THEOREM 2. *Under the previous four principles, a third order inpainting model in the divergence form must be given by*

$$(17) \quad \frac{\partial u}{\partial t} = \nabla \cdot (f(\kappa, \sigma) \vec{n} + a\sigma \vec{t}),$$

where a is a non-zero constant, $f(\kappa, \sigma)$ a positive function, and

$$\kappa = \nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right) \quad \text{and} \quad \sigma = \frac{\partial (\ln |\nabla u|)}{\partial \vec{t}},$$

are the scalar curvature and its conjugate.

Apparently, the TV and CDD inpainting models discussed in Section 2 are special examples of the current model with $a = 0$.

For instance, as inspired by Euler's elastica inpainting model [9, 25], one could choose $f = a + b\kappa^2$ for two positive constants a and b , or even $f = \exp(c\kappa^2)$ for some positive constant c . As discussed in Section 2, such choices penalize large curvatures and thus favor the *connectivity principle* [12].

In application, the available part of the image u^0 on $\Omega \setminus D$ is often noisy. Therefore, the new scheme (17) is implemented by its two-phase extension on the entire domain Ω , as in the case of CDD inpainting (Eq. (9)), with the adiabatic condition $\partial u / \partial \vec{n} = 0$ along $\partial \Omega$.

4. Conclusion and Two Notes

In this paper, we have reviewed the role of nonlinear transport and diffusion for inpainting modeling, and discussed the existing empirical works in [5, 11, 12]. To intrinsically integrate both the transport and diffusion mechanisms, we have proposed and analyzed four principle axioms that eventually lead to the discovery of a new class of third order nonlinear PDE models. In the future, we plan to further investigate this new class of models both theoretically and computationally.

Note A. The main result in Section 3 first appeared in our preprint "Morphologically invariant PDE inpainting," UCLA CAM Report 2001-15 (referred to as MI below). Due to the unfortunate mishandling by the journal we submitted MI to, by the time we submitted the current proceeding paper, MI had not been in any other reviewing process. Thus, technically speaking, we are here for the first time exposing our main new results. Moreover, by describing the techniques to our mathematics community (instead of the engineering literature as MI was initially intended), we are to inspire more discussions and criticisms on our new model from the mathematics point of view.

Note B. *Who did the digital inpainting first?* One of our referees said it should be [7, 25], instead of what goes in the second paragraph of Introduction. If we set absolutely that "inpainting = image interpolation," then the history of inpainting could have gone even much more distant than [7, 25]. In fact, image interpolation is such a standard (but interesting) topic that one can read it in almost any classical book on digital image processing. However, the notion of "inpainting" was indeed cleverly transplanted from museum conservation artists for the first time in the paper by Bertalmio et al. [5]. Ever since, the notion of digital inpainting has not

only unified numerous seemingly very different interpolation problems in image processing, but also opened the door to many important applications in the digital and telecommunication technology. It is out of this consideration that we have emphasized much the contribution of the paper by Bertalmio et al. [5], which has also cited [7, 25]. What is more important for inpainting in the future, is perhaps not “who did the inpainting first,” instead, “who has found the right way of doing it.” After all, as mathematicians, we are all trying to contribute, in a collective way, to the healthy development of this new booming field called *Mathematical Image and Vision Analysis*.

Acknowledgments

We would like to thank Professor Sapiro’s group for first introducing us to the inpainting problem.

References

- [1] A. Aldroubi and K. Gröchenig. Nonuniform sampling and reconstruction in shift-invariant spaces. *SIAM Review*, 43(4):585–620, 2001.
- [2] L. Alvarez, F. Guichard, P.-L. Lions, and J.-M. Morel. Axioms and fundamental equations of image processing. *Arch. Rational Mech. Anal.*, 123:199–257, 1993.
- [3] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera. Filling-in by joint interpolation of vector fields and grey levels. *IEEE Trans. Image Process.*, 10(8):1200–1211, 2001.
- [4] M. Bertalmio, A. L. Bertozzi, and G. Sapiro. Navier-Stokes, fluid dynamics, and image and video inpainting. IMA Preprint 1772 at: www.ima.umn.edu/preprints/jun01, June, 2001.
- [5] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. Computer Graphics, SIGGRAPH 2000, July, 2000.
- [6] T. Blu, P. Thvenaz, and M. Unser. Moms: maximal-order interpolation of minimal support. *IEEE Trans. Image Process.*, 10(7):1069–1080, 2001.
- [7] V. Caselles, J.-M. Morel, and C. Sbert. An axiomatic approach to image interpolation. *IEEE Trans. Image Processing*, 7(3):376–386, 1998.
- [8] A. Chambolle and P. L. Lions. Image recovery via Total Variational minimization and related problems. *Numer. Math.*, 76:167–188, 1997.
- [9] T. F. Chan, S.-H. Kang, and J. Shen. Euler’s elastica and curvature based inpaintings. UCLA CAM Report 2001-12 at: www.math.ucla.edu/~imagers; to appear in *SIAM J. Appl. Math.*, 2002.
- [10] T. F. Chan, S. Osher, and J. Shen. The digital TV filter and nonlinear denoising. *IEEE Trans. Image Process.*, 10(2):231–241, 2001.
- [11] T. F. Chan and J. Shen. Mathematical models for local nontexture inpaintings. *SIAM J. Appl. Math.*, 62(3):1019–1043, 2001.
- [12] T. F. Chan and J. Shen. Nontexture inpainting by curvature driven diffusions (CDD). *J. Visual Comm. Image Rep.*, 12(4):436–449, 2001.
- [13] T. F. Chan and J. Shen. Bayesian inpainting based on geometric image models. Preprint, 2002.
- [14] T. F. Chan and J. Shen. On the role of the BV image model in image restoration. UCLA’s Mathematics Department CAM Report 02-14 available at: www.math.ucla.edu/~imagers, 2002.
- [15] L. Chanas, J. P. Cocquerez, and J. Blanc-Talon. Highly degraded sequences restoration and inpainting. Preprint, 2001.
- [16] G. Emile-Male. *The Restorer’s Handbook of Easel Painting*. Van Nostrand Reinhold, New York, 1976.
- [17] S. Esedoglu and J. Shen. Digital inpainting based on the Mumford-Shah-Euler image model. *European J. Appl. Math.*, in press, 2002.
- [18] L. C. Evans and J. Spruck. Motion of level sets by mean curvature. *J. Diff. Geom.*, 33(3):635–681, 1991.
- [19] E. Giusti. *Minimal Surfaces and Functions of Bounded Variation*. Birkhäuser, Boston, 1984.

- [20] H. Igehy and L. Pereira. Image replacement through texture synthesis. *Proceedings of IEEE Int. Conf. Image Processing*, 1997.
- [21] K.-H. Jung, J.-H. Chang, and C. W. Lee. Error concealment technique using data for block-based image coding. *SPIE*, 2308:1466–1477, 1994.
- [22] W. Kwok and H. Sun. Multidirectional interpolation for spatial error concealment. *IEEE Trans. Consumer Electronics*, 39(3), 1993.
- [23] X. Li and M.T. Orchard. New edge-directed interpolation. *IEEE Trans. Image Process.*, 10(10):1521 – 1527, 2001.
- [24] A. Marquina and S. Osher. Explicit algorithms for a new time dependent model based on level set motion for nonlinear deblurring and noise removal. *Siam. J. Sci. Comput.*, 22:387–405, 2000.
- [25] S. Masnou and J.-M. Morel. Level-lines based disocclusion. *Proceedings of 5th IEEE Int'l Conf. on Image Process., Chicago*, 3:259–263, 1998.
- [26] J.-M. Morel and S. Solimini. *Variational Methods in Image Segmentation*, volume 14 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser, Boston, 1995.
- [27] D. Mumford. *Geometry Driven Diffusion in Computer Vision*, chapter “The Bayesian rationale for energy functionals”, pages 141–153. Kluwer Academic, 1994.
- [28] M. Nitzberg, D. Mumford, and T. Shiota. *Filtering, Segmentation, and Depth*. Lecture Notes in Comp. Sci., Vol. 662. Springer-Verlag, Berlin, 1993.
- [29] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Machine Intell.*, 12:629–639, 1990.
- [30] L. Rudin and S. Osher. Total variation based image restoration with free local constraints. *Proc. 1st IEEE ICIP*, 1:31–35, 1994.
- [31] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.
- [32] J. Shen. Geometric inpainting and applications. To appear in *Proc. SPIE*, vol. 4792, 2002.
- [33] A. Tsai, Jr. A. Yezzi, and A. S. Willsky. Curve evolution implementation of the Mumford-Shah functional for image segmentation, denoising, interpolation and magnification. *IEEE Trans. Image Process.*, 10(8):1169–1186, 2001.
- [34] S. Walden. *The Ravished Image*. St. Martin’s Press, New York, 1985.
- [35] L.-Y. Wei and M. Levoy. Fast texture synthesis using tree-structured vector quantization. Preprint, Computer Science, Stanford University, 2000; Also in *Proceedings of SIGGRAPH*, 2000.
- [36] J. Weickert. *Anisotropic Diffusion in Image Processing*. Teubner-Verlag, Stuttgart, Germany, 1998.

DEPARTMENT OF MATHEMATICS, AND INSTITUTE OF PURE AND APPLIED MATHEMATICS (IPAM),
UCLA, LOS ANGELES, CA 90095, USA

E-mail address: TonyC@college.ucla.edu

SCHOOL OF MATHEMATICS, (206 CHURCH STREET, SE,) UNIVERSITY OF MINNESOTA, MINNEAPOLIS, MN 55455, USA. FAX: (612) 626-2017. (CORRESPONDING AUTHOR)

E-mail address: jhshen@math.umn.edu

This page intentionally left blank

Towards fast non-rigid registration

U. Clarenz, M. Droske, M. Rumpf

ABSTRACT. A fast multiscale and multigrid method for the matching of images in 2D and 3D is presented. Especially in medical imaging this problem - denoted as the registration problem - is of fundamental importance in the handling of images from multiple image modalities or of image time series. The paper restricts to the simplest matching energy to be minimized, i.e., $E[\phi] = \frac{1}{2} \int_{\Omega} |f_1 \circ \phi - f_2|^2$, where f_1, f_2 are the intensity maps of the two images to be matched and ϕ is a deformation. The focus is on a robust and efficient solution strategy.

Matching of images, i.e., finding an optimal deformation ϕ which minimizes E is known to be an ill-posed problem. Hence, to regularize this problem a regularization of the descent path is considered in a gradient flow method. Thus the initial value problem $\partial_t \phi = -\text{grad}_g E[\phi]$ with some regular initial deformation $\phi(0) = \phi_0$ is solved on a suitable space of deformations $\Omega \rightarrow \Omega$. The gradient grad_g is measured w.r.t. a suitable regularizing metric g . Existence and uniqueness of solutions is demonstrated for different types of regularizations. For the implementation a metric based on multigrid cycles on hierarchical grids is proposed, using their superior smoothing properties. This is combined with an effective time-step control in the descent algorithm. Furthermore, to avoid convergence to local minima, multiple scales of the images to be matched are considered. Again, these image scales can be generated applying multigrid operators and we propose to resolve the pyramid of scales on a properly chosen pyramid of hierarchical grids. Examples on 2D and large 3D image matching problems prove the robustness and efficiency of the proposed approach.

1. Introduction

Image assisted diagnostics and surgery planning requires robust and valid segmentation and classification results and an analysis of the temporal change of anatomic structures. This can only be achieved properly if images recorded with different imaging machinery or at different times can suitably be correlated to each other. Various techniques have been proposed to solve this registration problem. They all ask for an “optimal” deformation which deforms one image such that there is an “optimal” correlation to another image with respect to a suitable coherence measure.

Mainly two different approaches have been discussed in the literature [5, 6, 7, 9, 15, 17, 21]. On the one hand, so called elastic registration techniques deal with a regularization of the energy, typically adding a convex energy functional based on gradients to the actual matching energy. The regularization energy is regarded as a penalty for “elastic stresses” resulting from the deformation of the images. This approach is related to the well known classical Tikhonov regularization of the originally ill-posed problem. On the other hand,

1991 *Mathematics Subject Classification.* Primary 65M30, 65M32, 65M55; Secondary 62H35, 68U05, 93A30.

© 2002 American Mathematical Society

viscous flow techniques are taken into account. They compute smooth paths from some initial deformation towards the set of minimizers of the matching energy. Thereby, a suitable regularization of the velocity, e.g., adding an artificial viscosity, ensures a certain problem dependent smoothness modulus. This class of methods can be interpreted as a gradient flow approach with respect to a metric which penalizes non-regular descent directions. Taking into account a time-step discretization this methodology is closely related to iterative Tikhonov regularization methods [12, 20]. Preparing this paper we got aware of recent results by Henn and Witsch [13]. They proposed to use multigrid smoothers in iterative Tikhonov regularization and proved its applicability for non-rigid registration in medical imaging. Concerning the impact of multigrid smoothing, this approach is closely related to our approach. In fact using the gradient flow perspective we enlighten the problem from a different point of view. In addition we embed the approach in a scale of matching problems which enable the computation of more global deformations.

Furthermore let us recall the *optical flow* method in image processing. The task is to extract motion fields from image time sequences. We ask for the time discrete motion velocity between two images of a time sequence, i.e., a short time deformation which is again a matching problem. If the motion is only piecewise smooth a simple regularization adding a Dirichlet-integral would not be able to retain the often discontinuous deformations on image edges. Thus Nagel and Enkelmann proposed an anisotropic quadratic form for the gradient of the deformation which regularizes edges of the image only in the tangential direction [8, 18].

Due to the non-convexity of the minimization problem in image registration it might be difficult to find the absolute minimum in case of larger deformations. Alternatively, one can consider a convolution of the images with a large corresponding filter width which destroys much of the detailed structure, match those images, and then successively reduce the filter-width and iterate the process [3, 19, 24]. This procedure is comparable to an annealing algorithm, where the filter width plays the role of the temperature.

In this paper we will consider one of the simplest image intensity based matching energies and apply a gradient descent approach for its minimization. The focus is on the robustness and efficiency of the proposed method and not on the generality of the approach with respect to its range of applications. In Section 8 we will give an outline on further research directions. The building blocks of the presented method are:

- a suitable choice of the regularizing metric (especially based on multigrid cycles),
- standard effective time-step control methods in the gradient descent method but now taking into account the selected new metric,
- a multiscale approach considering a series of successively smoothed images
- scale dependent grid resolution, i.e., solving coarse scale problems for sufficiently smooth images on coarse grids,
- and finally, scale dependent stopping criteria, which prevents us from resolving fine deformation details already on much too coarse scales.

Altogether these ingredients ensure a superior performance of the resulting algorithm. It allows the efficient computation of large scale matching problems with large solution deformations. Specifically, medical images of a resolution 129^3 can be matched based on deformations in the space of piecewise trilinear, continuous functions in a few minutes on a LINUX PC with a Pentium IV, 1.7 GHz processor and the resulting deformations are reasonably smooth. In what follows we will introduce the continuous model in Section 2,

show existence and uniqueness of solutions in Section 3. Section 4 contains the description of the chosen scale space method. We explain the spatial and temporal discretization in Section 5. In Sections 6 and 7 we collect the algorithmic ingredients and applications respectively and in Section 8 further extensions towards nonlinear metrics and different matching energies are sketched and we draw conclusions.

2. Problem and approach

Given two images $f_1, f_2 : \Omega \rightarrow \mathbb{R}$, where $\Omega \subset \mathbb{R}^d$ and $d = 2, 3$, we would like to determine a deformation $\phi : \Omega \rightarrow \mathbb{R}^d$ which maps Ω onto Ω and maps grey values in the first image f_1 via a deformation ϕ to grey values at the deformed position in the second image f_2 such that

$$f_1 \circ \phi \approx f_2.$$

For the ease of presentation we assume $\Omega = [0, 1]^d$ throughout this paper. We consider u as the perturbation of ϕ from the identity II which means $\text{II} + u = \phi$. To optimize the deformation with respect to a proper match of the two images we define the most basic energy E depending on the displacement u (resp. the deformation ϕ):

$$(E) \quad E[u] = \frac{1}{2} \int_{\Omega} |f_1 \circ (\text{II} + u) - f_2|^2.$$

In what follows we use either ϕ or u as the argument of the energy E . If u is an ideal deformation the above energy vanishes. Thus we ask for minimizers of the problem $E[u] \rightarrow \min$ in some Banach space \mathcal{V} . Obviously, this problem is ill-posed. Consider a deformation ϕ and for $c \in \mathbb{R}$ the level sets $\mathcal{M}_c^1 = \{x \in \Omega \mid f_1(x) = c\}$. Then for any displacement Λ which keeps \mathcal{M}_c^1 fixed for all c , the energy does not change, i.e.,

$$E[\phi] = E[\Lambda \circ \phi]$$

This especially holds true for a possible minimizer. Hence, a minimizer – if it exists – is non-unique and the set of minimizers is expected to be non-regular and not closed in a usual set of admissible displacements.

A minimizer u of (E) is characterized by the necessary condition $E'[u] = 0$, where $E'[u] \in \mathcal{V}'$ for the dual space \mathcal{V}' of \mathcal{V} . This condition can be expressed in weak form:

$$\int_{\Omega} (f_1 \circ (\text{II} + u) - f_2) \nabla f_1 \circ (\text{II} + u) \cdot \theta = 0,$$

for all $\theta \in [C_0^\infty(\Omega)]^d$. Suppose $[L^2(\Omega)]^d$ is embedded in the space \mathcal{V}' . Under obvious integrability conditions we obtain the L^2 -representation of E'

$$(2.1) \quad \text{grad}_{L^2} E[u] = (f_1 \circ (\text{II} + u) - f_2) \nabla f_1 \circ (\text{II} + u).$$

We investigate a gradient flow approach to solve this matching problem. A gradient of a functional $E : \mathcal{V} \rightarrow \mathbb{R}$ is defined as the representation of the Fréchet derivative $E' \in \mathcal{V}'$ in a metric $g(\cdot, \cdot)$ on \mathcal{V} , i.e.,

$$g(\text{grad}_g E[u], \theta) = \langle E'[u], \theta \rangle.$$

One frequently identifies $E'[u]$ with the gradient of E with respect to the L^2 -product. Here we introduce a different length measurement on the space of deformations and consider a general gradient flow

$$\begin{aligned} \partial_t u &= -\text{grad}_g E[u], \\ u(0) &= u_0. \end{aligned}$$

for a suitable metric $g(\cdot, \cdot)$ on \mathcal{V} . Thus we ask for a solution $u : \mathbb{R}_0^+ \rightarrow \mathcal{V}$, such that

$$(2.2) \quad g(\partial_t u, \theta) + \int_{\Omega} (f_1 \circ (\mathbb{I} + u) - f_2) \nabla f_1 \circ (\mathbb{I} + u) \cdot \theta = 0,$$

for all $\theta \in [C_0^\infty(\Omega)]^d$. The choice of \mathcal{V} and the metric g on \mathcal{V} is related to a regularization of the matching problem (cf. Section 3). At least for finite time we ensure \mathcal{V} -regularity of the deformation. The representation of the metric g in the duality pairing $(\mathcal{V}', \mathcal{V})$ will be denoted by $A : \mathcal{V} \rightarrow \mathcal{V}'$, i.e.,

$$g(\varphi, \psi) = \langle A\varphi, \psi \rangle$$

for all $\psi \in \mathcal{V}$. Hence, if the inverse of A is regular in a suitable sense (cf. Section 3) the gradient flow can be rewritten as an ODE in the Banach space \mathcal{V} :

$$\partial_t u = -A^{-1} E'[u].$$

In the following section we give examples for A and corresponding metrics g and show existence and uniqueness of solutions.

3. Existence and uniqueness

In what follows let us assume \mathcal{V} to be a Banach space. Furthermore suppose that there is a second Banach space $\mathcal{W} \supset \mathcal{V}$ which is embedded in the dual space \mathcal{V}' of \mathcal{V} . Hence, we can state the following

THEOREM 3.1. *Let A be a linear isomorphism $A : \mathcal{V} \rightarrow \mathcal{W}$. If $E'[\mathcal{V}] \subset \mathcal{W}$ and $E'[\cdot] : \mathcal{V} \rightarrow \mathcal{W}$ is Lipschitz continuous, then there exists a unique solution of the problem:*

For given initial data $u_0 \in \mathcal{V}$, find a solution $u : \mathbb{R}_0^+ \rightarrow \mathcal{V}$, such that

$$\begin{aligned} \partial_t u &= -A^{-1} E'[u], \\ u(0) &= u_0. \end{aligned}$$

Remark: Theorem 3.1 especially ensures that solutions of the gradient flow are \mathcal{V} -regular for finite times. Let us emphasize that in general we can neither expect the \mathcal{V} -norm to be uniformly bounded in time nor that there exists a steady state.

The proof is a straightforward application of the Picard-Lindelöf Theorem in Banach spaces. We have shown that there is a L^2 -representation $\text{grad}_{L^2} E$ of E' (cf. Section 2.1), if f_1 and f_2 are of suitable regularity. Therefore in case $\mathcal{W} = [L^2(\Omega)]^d$ the inclusion $E'(\mathcal{V}) \subset \mathcal{W}$ is valid. Let us prove Lipschitz continuity of $\text{grad}_{\mathcal{W}} E = \text{grad}_{L^2} E$.

LEMMA 3.2. *Let $\mathcal{V} = \mathcal{V}' = \mathcal{W} = [L^2(\Omega)]^d$; then the derivative of the energy E w.r.t. \mathcal{W} is Lipschitz continuous, i.e.,*

$$\|\text{grad}_{L^2} E[u_1] - \text{grad}_{L^2} E[u_2]\|_{L^2} \leq C \|u_1 - u_2\|_{L^2}$$

if $f_1 \in C^{1,1}(\mathbb{R}^d)$ and $f_2 \in L^\infty(\Omega)$.

Proof: Let $u_1, u_2 \in \mathcal{V}$. Then we have

$$\begin{aligned} &\text{grad}_{L^2} E[u_1] - \text{grad}_{L^2} E[u_2] \\ &= (f_1 \circ (\mathbb{I} + u_1) - f_2) \nabla f_1 \circ (\mathbb{I} + u_1) - \\ &\quad (f_1 \circ (\mathbb{I} + u_2) - f_2) \nabla f_1 \circ (\mathbb{I} + u_2) \\ &= [(f_1 \circ (\mathbb{I} + u_1) - f_2) - (f_1 \circ (\mathbb{I} + u_2) - f_2)] \nabla f_1 \circ (\mathbb{I} + u_1) - \\ &\quad (f_1 \circ (\mathbb{I} + u_2) - f_2) [\nabla f_1 \circ (\mathbb{I} + u_1) - \nabla f_1 \circ (\mathbb{I} + u_2)]. \end{aligned}$$

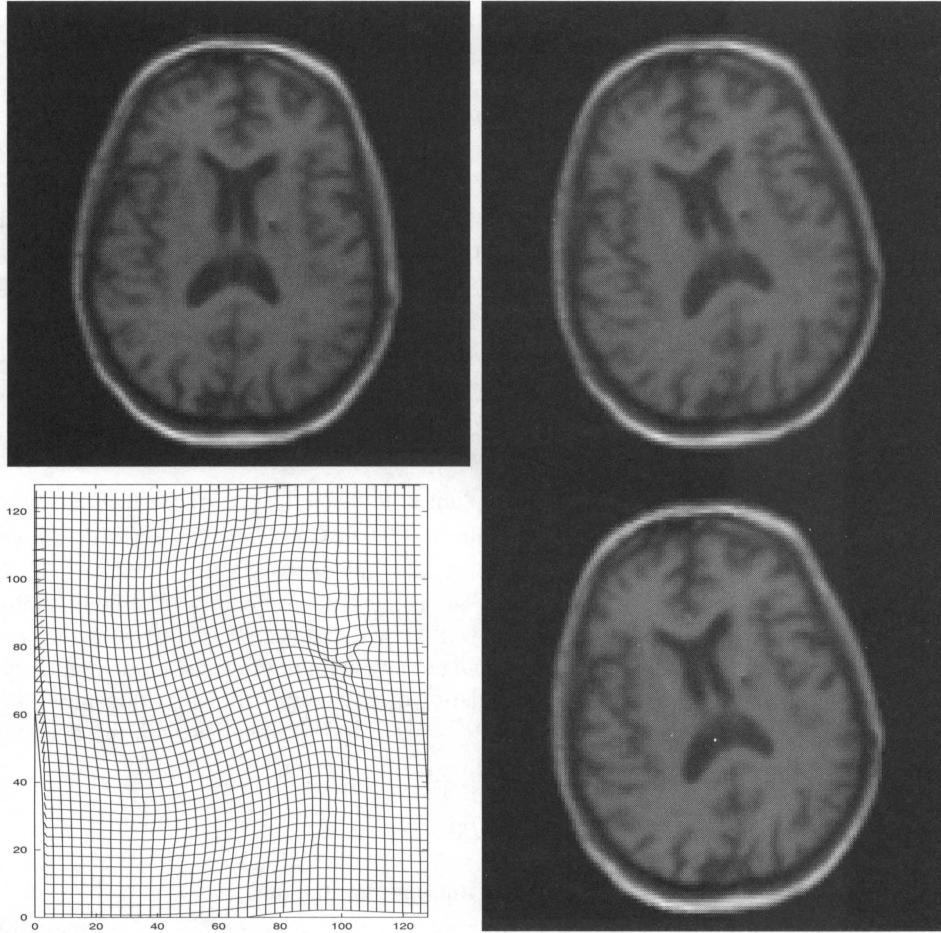


FIGURE 1. *The images show the 2D-matching result of an artificially deformed MRI-slice with 129^2 pixels. From top left to bottom right: original image, distorted image by rotational twist, application of the deformation to a uniform grid , matching result.*

On account of our regularity assumptions we can finish our proof of Lipschitz continuity:

$$\begin{aligned} & |\text{grad}_{L^2} E[u_1] - \text{grad}_{L^2} E[u_2]| \\ & \leq \|f_1\|_{C^1} \|f_1\|_{C^{0,1}} |u_1 - u_2| + (\|f_1\|_{C^0} + |f_2|) \|f_1\|_{C^{1,1}} |u_1 - u_2| \end{aligned}$$

which leads to

$$\begin{aligned} & \|\text{grad}_{L^2} E[u_1] - \text{grad}_{L^2} E[u_2]\|_{L^2(\Omega)} \\ & \leq \|f_1\|_{C^1} \|f_1\|_{C^{0,1}} \|u_1 - u_2\|_{L^2(\Omega)} + \\ & \quad \|f_1\|_{C^0} \|f_1\|_{C^{1,1}} \|u_1 - u_2\|_{L^2(\Omega)} + \\ & \quad \|f_1\|_{C^{1,1}} \|f_2\|_{L^\infty} \|u_1 - u_2\|_{L^2(\Omega)}. \end{aligned}$$

□

Let us now consider several examples:

- (i) In case $A = \mathbb{I}$ existence and uniqueness are shown by the above Lemma and Theorem 3.1.
- (ii) The proof of Lemma 3.2 clearly extends to $\mathcal{V} = [H^{s,2}(\Omega)]^d$, where $s \geq 0$ and $\mathcal{V}' = [H^{s,2}(\Omega)']^d$ because in this case $[H^{s,2}(\Omega)]^d \hookrightarrow [L^2(\Omega)]^d \hookrightarrow [H^{s,2}(\Omega)']^d$.

For our purpose of image matching the regularity induced by the L^2 -metric will not be sufficient to obtain proper approximations of energy minimizers for our ill-posed problem w.r.t. actual applications. Thus we cannot expect to obtain smooth deformations in case $A = \mathbb{I}$ and $\mathcal{V} = \mathcal{V}' = \mathcal{W} = [L^2(\Omega)]^d$, even if we start with smooth initial deformations. Therefore we deal with spaces \mathcal{V} of higher regularity and suitable operators A representing a metric:

- (iii) We might choose the Helmholtz operator $A = \mathbb{I} - \frac{\sigma^2}{2} \Delta$ for $\sigma \in \mathbb{R}^+$. The metric representing A is

$$g(v, w) = (v, w)_{L^2} + \frac{\sigma^2}{2} (\nabla v, \nabla w)_{L^2}.$$

This choice corresponds to an implicit time discretization of the heat equation with time-step $\tau = \frac{\sigma^2}{2}$ and is thus related to Gaussian filtering with a filter width σ . As corresponding spaces we take into account $\mathcal{V} = [H^{1,2}(\Omega)]^d$, $\mathcal{V}' = [H^{1,2}(\Omega)']^d$ and $\mathcal{W} = [L^2(\Omega)]^d$. The isomorphism property of A and thereby the Lipschitz continuity of A^{-1} is well known in this case. Thus we have an existence and uniqueness result at hand but now with improved solution regularity.

- (iv) We can further improve the regularity of the deformations. Choosing $\mathcal{V} = [H^{2,2}(\Omega)]^d$, $\mathcal{W} = [L^2(\Omega)]^d$ and $\mathcal{V}' = [H^{2,2}(\Omega)']^d$ together with the operator $A = (\mathbb{I} - \frac{\sigma^2}{2} \Delta)^2$. The corresponding metric is given by

$$g(v, w) = (v, w)_{L^2} + 2 \frac{\sigma^2}{2} (\nabla v, \nabla w) + \frac{\sigma^4}{4} (\Delta v, \Delta w),$$

and A^{-1} is again well defined and Lipschitz continuous.

So far we have shown that using suitable metrics g one can improve the regularity of resulting deformations. Finally let us add a remark on the use of a true Gaussian filtering instead of $A = \mathbb{I} - \frac{\sigma^2}{2} \Delta$ (cf. (iii)). Consider the ODE

$$\partial_t u = -B \operatorname{grad}_{L^2} E[u],$$

where $B = HESG\left(\frac{\sigma^2}{2}\right)$ and $HESG$ indicates the heat equation semi-group. We look at $\mathcal{V} = C_0^m(\Omega)$ for $m \geq 0$. In this case we don't have an interpretation of this ODE as a gradient flow with respect to a norm induced by a metric. Nevertheless, we obtain C^m regular deformations for any $m \geq 0$ and finite time.

4. A scale space approach

As already stated in the introduction for typical image intensity functions f_1, f_2 the energy $E[\cdot]$ is non-convex and we expect an energy landscape with many local minima. This implies that gradient descent paths mostly tend to asymptotic states which only locally minimize the energy. Following Alvarez et al. [2] we consider a continuous annealing method based on a scale of image pairs $f_{1,\epsilon}, f_{2,\epsilon}$, where $\epsilon \geq 0$ is the scale parameter. Here we consider scale spaces of images generated by a scale space operator $S(\cdot)$ which maps an initial image f onto some coarser image, i.e.,

$$f_\epsilon = S(\epsilon)f.$$

The scale parameter ϵ allows to select fine grain representations corresponding to small values of ϵ and coarse grain representations with most of the image details skipped for larger values of ϵ . A frequently used scale space is the linear one based on Gaussian filtering. In fact we can take into account the heat equation semi group $\{HESG(\frac{\epsilon^2}{2})\}_\epsilon$ on the bounded domain Ω with imposed Neumann boundary conditions, where ϵ is the filter width parameter, i.e., $S(\epsilon) = HESG(\frac{\epsilon^2}{2})$. Here, we confine with this basic filter. Alternatively, other scale space operators such as morphological ones may be incorporated. Let us emphasize, that with respect to the final implementation we actually consider an efficient and effective approximation of this operator. Finally, we formulate the arising scale of problems: For given $\epsilon \geq 0$ we consider an energy

$$E_\epsilon[u] = \frac{1}{2} \int_{\Omega} |f_{1,\epsilon} \circ (\mathbb{I} + u) - f_{2,\epsilon}|^2 .$$

and the corresponding gradient flow

$$\begin{aligned} g(\partial_t u_\epsilon, \theta) &= -\langle E'_\epsilon[u_\epsilon], \theta \rangle \\ u_\epsilon(0) &= u_{0,\epsilon} . \end{aligned}$$

We are left to choose the initial data $u_{0,\epsilon}$ for the evolution on scale ϵ . Here we expect the minimizer or a sufficiently good approximation of the same problem on a coarse scale to be a suitable starting point to approach the global minimum on the finer scale. Algorithmically, we select a sequence of scales

$$(4.1) \quad \epsilon_k = \beta_1 2^{-\beta_2 k}, \quad \beta_1, \beta_2 > 0 ,$$

where $k < k_{max}$ and $\epsilon_{k_{max}} = 0$. Thus we compute discrete counterparts of the continuous solutions $u_{\epsilon_k}(T_k)$ for end times T_k sufficiently large and set

$$u_{\epsilon_k}(0) = u_{\epsilon_{k-1}}(T_{k-1}) .$$

For fixed β_2 the parameter β_1 is chosen such that $\epsilon_{k_{max}-1} = h$. For details we refer to Section 6.

5. Discretization

Concerning the time discretization our approach can be interpreted as a gradient flow in a Banach space. The energy functional E is Fréchet differentiable if we assume certain regularity of f_1 and f_2 (see Section 3). Therefore taking into account the energy, its Gâteaux derivative and the metric g on L^2 , we are able to recall time-step controlled descent algorithms well-known in continuous optimization problems [16].

We will consider a time discretization as well as a spatial discretization in the following section. The spatial discretization is a standard finite element method. In addition we make use of multigrid techniques.

Time discretization. Aiming at an efficient implementation of a discrete gradient flow we apply a suitable time-step control. Thus, it pays off to consider the gradient flow perspective not only as a conceptually intuitive setting but also in the application of classical numerical tools. A time-step control strategy for the minimization of energy functionals on \mathbb{R}^m turns into a time-step control for our discrete generalized gradient descent algorithm. We only have to replace the Euclidian distance in \mathbb{R}^m by the norm induced by $g(\cdot, \cdot)$ on \mathcal{V} . We consider the explicit scheme:

$$\frac{u^{n+1} - u^n}{\tau_n} = -A^{-1}E'[u^n] .$$

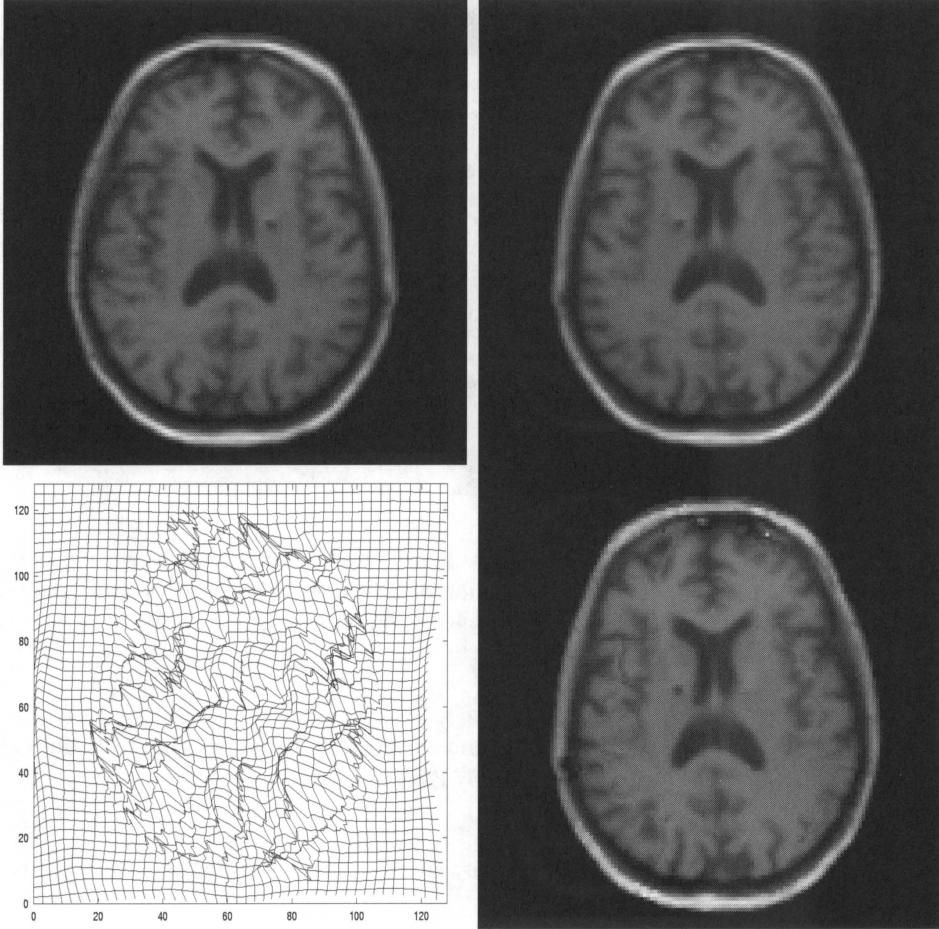


FIGURE 2. Here the 2D-matching result of a mirrored MRI-slice with 129^2 pixels is shown. From top left to bottom right: original image, mirrored image, application of the deformation to a uniform grid, matching result.

Thus we construct a sequence $(u^n)_{n=0,\dots}$, such that u^n approximates $u(t_n)$ with $t_n = \sum_{i=1}^n \tau_i$. The actual focus is not on the quality of the approximation but on a fast and robust descent. In our implementation we determine τ_n using Armijo's rule. As will become clear in our considerations, various other time-step control strategies can also be considered.

Armijo's rule determines each time-step τ_n by choosing τ_n such that for $\sigma \in (0, \frac{1}{2})$

$$-\frac{E[u^{n+1}] - E[u^n]}{\tau_n \langle E'[u^n], A^{-1}E'[u^n] \rangle} \geq \sigma.$$

Using the metric g this inequality can be expressed as

$$(5.1) \quad -\frac{E[u^{n+1}] - E[u^n]}{\tau_n g(A^{-1}E'[u^n], A^{-1}E'[u^n])} \geq \sigma.$$

We provide in each time-step a solution τ_n of the above inequality computing $k_n \in \mathbb{Z}$ as the smallest integer such that for some fixed $\beta \in (0, 1)$

$$(5.2) \quad E[u^n] - E[u^n - \beta^{k_n} A^{-1} E'[u^n]] \geq \sigma \beta^{k_n} \|A^{-1} E'[u^n]\|_g^2.$$

Here $\|\cdot\|_g$ denotes the norm induced by the metric g . It is always possible to find τ_n by that algorithm as long as u^n is not a local minimum of E . Indeed the function

$$h(t) := \frac{E[u^n] - E[u^n - t A^{-1} E'[u^n]]}{t \|A^{-1} E'[u^n]\|_g^2}$$

clearly converges to 1 if $t \rightarrow 0$ and tends (on account of the boundedness of E) to 0 if $t \rightarrow \infty$.

Spatial discretization. Now we describe the spatial discretization of equation (2.2). The set $\Omega = [0, 1]^d$ is given as the union of squares or cubes E_i for i in an index set J_h . The set of elements $\{E_i\}_{i \in J_h}$ forms the mesh \mathcal{M}_h . Here the subscript h indicates the grid size. We confine to grids which are generated by iterated subdivision into 4 squares or 8 cubes respectively.

Thus the resulting grids form a pyramid with grid sizes $h_l = 2^{-l}$ for $l = 0, \dots, l_{\max}$. The set of vertices of the mesh \mathcal{M}_h is denoted by \mathcal{N}_h . Interpreting pixel or voxel values of a 2D or 3D image as nodal values we consider discrete images (F_1, F_2) as piecewise multilinear continuous functions on \mathcal{M}_h . The corresponding multilinear finite element space is denoted by \mathcal{V}^h .

We suppose $\{\Psi^i\}_{i \in I_h}$ to be the canonical nodal basis of \mathcal{V}^h , where I_h is the index set corresponding to \mathcal{N}_h . Hence we obtain $F_i = \sum_{j \in I_h} F_i^j \Psi_j$ as the representation of the image F_i in this basis, where $F_i^j = F_i(x_j)$ for the node $x_j \in \mathcal{N}_h$ corresponding to the basis function Ψ_j . Analogously, we take into account $[\mathcal{V}^h]^d$ as the set of discrete deformations. Hence the fully discrete algorithm reads as follows:

For given initial displacement U^0 find a sequence of displacements $(U^n)_n$ in $[\mathcal{V}^h]^d$ which solve

$$g_h \left(\frac{U^{n+1} - U^n}{\tau_n}, \Theta \right) = -\langle E'[U^n], \Theta \rangle,$$

for all test functions $\Theta \in [\mathcal{V}^h]^d$. Here the metric g_h is supposed to be a suitable approximation of the original metric g . The computation of E' induces the evaluation of $f_1 \circ \phi$. The spatial discretization of ϕ is defined on all nodes x_i and we define $f_1 \circ \phi$ as the bi- or trilinear interpolation of $(f_1 \circ \phi)(x)$ for all $x \in \mathcal{N}_h$.

Let us throughout this presentation denote the nodal vector or nodal vector dependent functional by a bar on top of the corresponding function or functional respectively. Then we can rewrite our scheme and obtain

$$A_h(\bar{U}^{n+1} - \bar{U}^n) = -\tau_n \bar{E}'[\bar{U}^n],$$

or alternatively

$$(5.3) \quad \bar{U}^{n+1} = \bar{U}^n - \tau_n A_h^{-1} \bar{E}'[\bar{U}^n].$$

Here $A_h = M_h G_h$, where M_h is a mass matrix – in our case the lumped mass matrix [22] – and G_h the matrix representation of the discrete metric with respect to the product induced by M_h , i.e.,

$$g_h(X, Y) = M_h G_h \bar{X} \cdot \bar{Y}.$$

This defines G_h uniquely. In the nodal basis $\{\Psi^i\}_{i \in I_h}$ we have

$$(G_h)_{ij} = (M_h)_{ii}^{-1} g_h(\Psi^i, \Psi^j),$$

where $((M_h)_{ii})_{i \in I_h}$ is the diagonal lumped mass matrix.

Furthermore, $\bar{E}'[\bar{U}]$ is the vector of partial derivatives of $\bar{E}[\bar{U}]$ with respect to the nodal values \bar{U} of U on the grid nodes.

Hence, starting with U^0 we compute a sequence of discrete displacement functions $(U^n)_{n=0,\dots}$ approximating $u(t_n)$ for $t_n = \sum_{i=1}^n \tau_i$. Let us once more emphasize that our main intention is not a proper consistency with the time continuous problem. Indeed we intend a rapid energy descent along the piecewise linear path in the space \mathcal{V}^h . Simultaneously, we expect the timestep control to have a significant impact on the \mathcal{V} -norm of the solution. Hence an a priori upper bound of the time step seems feasible. An analysis of the problem has to be investigated in the future.

6. Improving the method's efficiency

In this section we will outline how the registration algorithm can be further improved concerning efficiency. The ingredients are a multigrid approximation of the smoothing operator A_h^{-1} , the numerical treatment of problems corresponding to coarse image scales on coarse grids and effective stopping criteria for the minimization procedure on coarse scales. In what follows these efficiency aspects are discussed in detail.

Multigrid. In Section 3 we showed existence as well as uniqueness for the flow $\partial_t u = -A^{-1}E'[u]$ on the space $\mathcal{V} = [H^{1,2}(\Omega)]^d$ and for $A = \mathbb{I} - \frac{\sigma^2}{2}\Delta$. The approach known to be the most efficient to solve such a linear system of equations is a multigrid method. It leads to an already optimal complexity of $O(n_h)$ if n_h is the cardinality of \mathcal{N}_h . But even better, already a single multigrid V-cycle is characterized by nice smoothing properties [4, 10] which we suppose to be the essential and sufficient property of A_h^{-1} . In fact we are not interested in any convergence but only in the smoothing properties of the multigrid cycle.

Henn and Witsch [13] already used this improvement in their iterative Tikhonov regularization approach.

We replace the operator A_h^{-1} in our discrete flow by the operator MGM_h representing a single multigrid V-cycle for the solution of a linear problem with the discrete operator $\mathbb{I} - \frac{\sigma^2}{2}\Delta_h$. Hence, we consider a sequence of grids $(\mathcal{M}_{h_l})_{l=0,\dots,l_{\max}}$ with successively finer grid size h_l (e.g. $h_l = 2^{-l}$). Then the building blocks of our multigrid operator are

- on each grid \mathcal{M}_{h_l} with discrete function space $\mathcal{V}^l := \mathcal{V}^{h_l}$ Jacobi iterations as smoother and
- standard prolongation and restriction operators defined on \mathcal{V}^l .

Finally, we are left to choose the number of pre-smoothing and post-smoothing steps in our V-cycle (cf. Fig. 3 for comparison of the resulting filter kernels). In our applications we confine with a single pre- and post-smoothing step. Table 1 lists the resulting computing times for the components of a single time-step in the discrete descent algorithm. Finally, let us underline that the discrete metric g_h is now induced by the multigrid operator MGM_h as follows: We consider MGM_h as an approximation of $\left(\mathbb{I} - \frac{\sigma^2}{2}\Delta_h\right)^{-1}$. For our discrete metric g_h this means $g_h(U, V) = MGM_h^{-1}\bar{U} \cdot \bar{V}$. Still this can be regarded as some approximation of the original metric $g(u, v) = (u, v)_{L^2} + \frac{\sigma^2}{2}(\nabla u, \nabla v)_{L^2}$.

Coarse scale problems on coarse grids. As the scale space operator S (cf. Section 4) actually applied in our approach we initially use an implicit step of the heat equation semigroup, i.e., we consider

$$\left(\text{II} + \frac{\epsilon^2}{2} M_h^{-1} L_h \right) \bar{F}_{i,\epsilon} = \bar{F}_i,$$

where M_h is the lumped mass matrix and L_h the usual stiffness matrix. Here ϵ plays the role of a filter width parameter. Again, with respect to an improved efficiency, we replace the exact linear solver for this system by the corresponding multigrid V-cycle, now acting on images and not on displacement descent directions (cf. Section 3 example (iii)). Let us recall that the scale parameter is chosen as in (4.1). Now we couple the sequence of meshes \mathcal{M}_{h_l} and corresponding function spaces \mathcal{V}^l on the one hand and the sequence of scales on the other hand. Thus we restrict on scale k the whole problem to grid level $l(k)$. In the case where S corresponds to the heat equation, a suitable choice for $l(k)$ would be the smallest integer such that

$$h_{l(k)} \leq \alpha \epsilon_k$$

for some constant $\alpha > 0$. In the concrete applications we have chosen the parameter $\alpha \in [\frac{1}{4}, 2]$.

Effectively coupling scales. Furthermore it is not required to reach a local minimum of the corresponding energy E_{ϵ_k} by the discrete gradient descent on level k . Expressed in formal geometric terms it is sufficient if for some integer n_k a discrete displacement $U_k^{n_k}$ of the discrete gradient descent sequence $(U_k^n)_{n=0,\dots}$ on scale k enters the attractive region of the global minimum of the energy $E_{\epsilon_{k+1}}$ on the next finer scale $k+1$. By construction in the k_{\max} -th step we would then end up in the contraction region of the actual energy E . Unfortunately, the above condition is rather implicit.

Hence, we confine with a heuristic stopping criterion for the discrete gradient descent on scale k , i.e., we turn to the next finer scale, if for some fixed constant γ we observe that for some norm $\|\cdot\|$

$$\|U_k^{n+1} - U_k^n\| \leq \gamma \epsilon_k$$

holds. Actually, in the application we consider $\gamma = \frac{1}{2}$ and evaluate the L^2 -norm of the displacement update.

If the iteration is finished on a certain scale k a regularization is applied to the deformation obtained on this scale before starting the iteration on the next scale. (We use our multigrid regularization approximating $HESG(\frac{\epsilon_k^2}{2})$.)

Based on this multilevel approach, the overall cost of the multiscale method is optimal. Here by optimal we mean that the number of considered scales does not contribute significantly to the overall cost. Indeed the overall cost reduces to the cost for the gradient descent on the finest grid and the finest scale times a constant C . For $\alpha = 1$ and an equal number of gradient descent iterations on all scales the overall cost geometrically decays with decreasing scale and grid level respectively, i.e., $C = 4/3$ in 2D and $C = 8/7$ in 3D. Due to our adaptive stopping criteria the actual factor for the offset cost for solving a multiscale problem is even smaller (cf. Table 2 which lists the required number of iterations and the computing times on every scale for an application problem in 3D).

7. Applications

Let us now collect numerical results in 2D and 3D for some test cases. In two dimensions it is still feasible to solve the resulting linear systems for $A = \mathbb{I} - \frac{\sigma^2}{2} \Delta$ by CG iterations. In three dimensions we define A^{-1} by one multigrid cycle related to the operator $A = \mathbb{I} - \frac{\sigma^2}{2} \Delta$ (cf. Section 6). We applied the presented algorithm on synthetic problems with large deformations as well as on medical MR-images which also lead to a non-local matching problem. In Figures 1, 2, and 4 2D-results for the matching of an artificially twisted resp. mirrored brain with the original image are depicted.

Figures 7 and 5 show the 3D-matching of a strongly rotated synthetic image and a reflected MR-image versus the corresponding original. All Figures corresponding to results in 3D show planar slices through the 3D volume.

The synthetically generated matching problem should demonstrate an important advantage of the multiscale approach, namely the capability to handle the registration of heavily distorted images, where the distortion itself is additionally highly non-rigid and non-local.

Naturally, the applicability to medical images is of fundamental importance for the evaluation of the method. Although we have confined to the most simple matching energy for a starting point, we wanted to get some insight on the fundamental behaviour of the gradient descent on realistic MR-images. Due to the fact, that both hemispheres of a healthy brain have – apart from minor geometrical differences – the same fundamental structure, a reflection provides a useful and solvable test example and is comparable to a matching problem from a patient to a reference image from an atlas. Thus, our aim is to find the displacement which describes both hemispheres given the corresponding other hemispheres and not to find the global minimum, which would be the reflection itself. This is naturally ruled out by the regularization and the gradient descent approach. The method is capable to compute a rather regular approximation of a local minimum, but with convincing coherence of the deformed first image and the initial second one, which can be seen by comparing the sulci of the cortex on the reference image with those of the matching result. Due to its high contrast to the surrounding tissue, the ventricles are perfectly matched as well. In contrast to the synthetic example, the resulting matching deformation varies locally in magnitude because some regions match initially quite well while others have to be deformed quite drastically. It turns out in the experiments, that the choice of σ in the metric operator A should not be too large in order not to destroy local variations in the deformation.

Process	Duration
V-cycle (single component)	3.3s
computation of $E[u]$ and grad $E[u]$	5.25s
computation of $\langle E'[u], \phi \rangle$	5.38s
computation of $E[u]$	1.23s
time-step control	1-3s

TABLE 1. *Approximate computing times for the key ingredients of each gradient descent step in our algorithm on a reference PC (Pentium IV, 1.7 Ghz, 1Gb RAM) applied on 3D images with 129^3 voxels.*

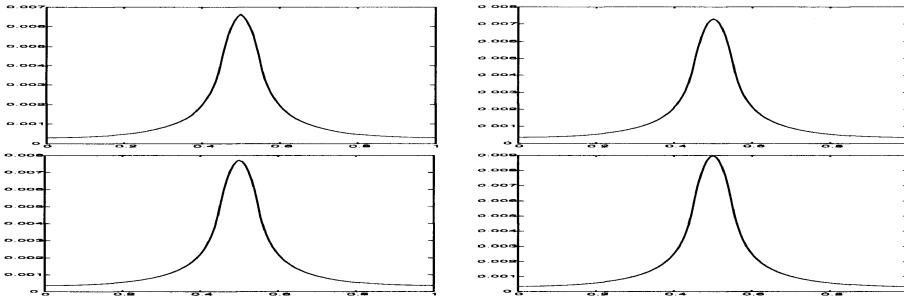


FIGURE 3. *Profiles through a three dimensional data set after the application of one multigrid cycle for the solution of the discrete heat equation with corresponding filter width $\sigma = 0.1$ and a discrete centered Dirac distribution as initial image. We applied 1 through 3 pre- and post-smoothing steps. These correspond to the actually applied kernels in a multigrid smoothing cycle. The bottom right image shows the profile corresponding to the exact discrete solution. One can clearly observe, that the overall shape varies only very slightly when the number of smoothing steps is changed.*

scale k	filter width ϵ_k	steps n_k	grid level $l(k)$	time
0	.250	5	5	<1s
1	.177	3	5	<1s
2	.125	3	5	<1s
3	.088	4	6	9s
4	.062	3	6	7s
5	.044	4	7	67s
6	.031	6	7	95s
7	.022	6	7	96s
8	.016	5	7	82s
9	.0	5	7	83s

TABLE 2. *Iteration counts on different scales due to the adaptive stopping criteria and the corresponding absolute timings for the computation on the correspondingly chosen grid levels. Here we assume $\alpha = \frac{1}{4}, \gamma = \frac{1}{2}$.*

8. Conclusion and Outlook

We have presented an efficient and robust registration method which is capable to compute large deformations on fine two and three dimensional grids. Here the focus is not on the generality of the presented method but on the acceleration potential based on the gradient flow perspective and the multigrid and multiscale approach for this class of optimization problems. Computations on medical images have pointed out the method's applicability and the quality of the achievable results. Nevertheless let us emphasize that the method is restricted to intensity matching applications. It is applicable to images of the same modality for example to register a medical MR-image of a patient to a equally weighted reference image in a medical atlas library or to describe temporal deformations

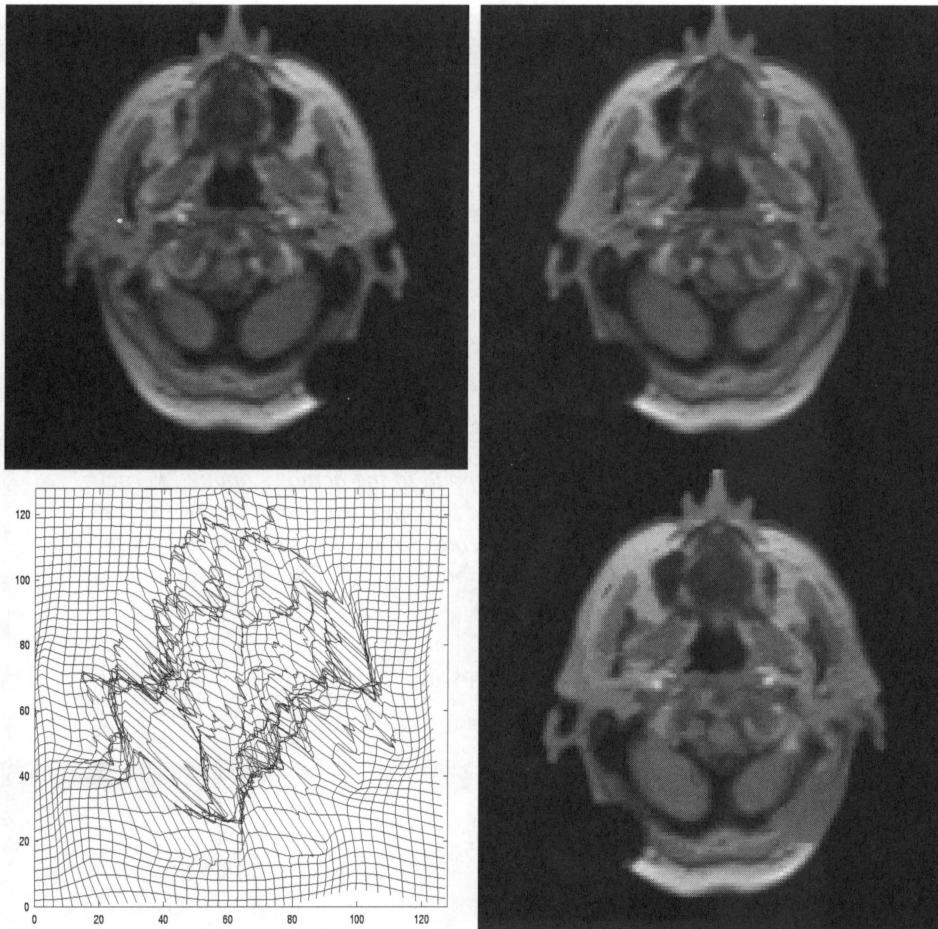


FIGURE 4. Due to the multiscale approach, the method is capable to generate relatively large displacement fields, here depicted for a 2D example. From top left to bottom right: original image, plane-mirrored image, application of the deformation to a uniform grid, matching result.

of subsequently acquired images of the same patient. It is however not capable to correlate image morphologies, i. e. it is not grey scale invariant [1], which is important for the registration of CT, MRI, PET and ultrasound datasets for example. Instead of matching image intensity one may consider image morphologies only and try to match them between images of different modality or different time steps from a sequence of images. The morphologies are characterized uniquely by the entity of level sets and their Gauss maps respectively. Hence, we will investigate a cost functional to be minimized which measures the effect of the deformation on the image Gauss maps instead of the image intensities. Furthermore in certain applications it turns out to be useful (cf. Section 1) to consider a non-homogeneous metric on the space of deformations depending on the images and image features. Furthermore constraints on the velocity fields such as a vanishing divergence – which ensures volume preservation of the resulting deformation at least in the time continuous case – can be included in the approach via a projection method.

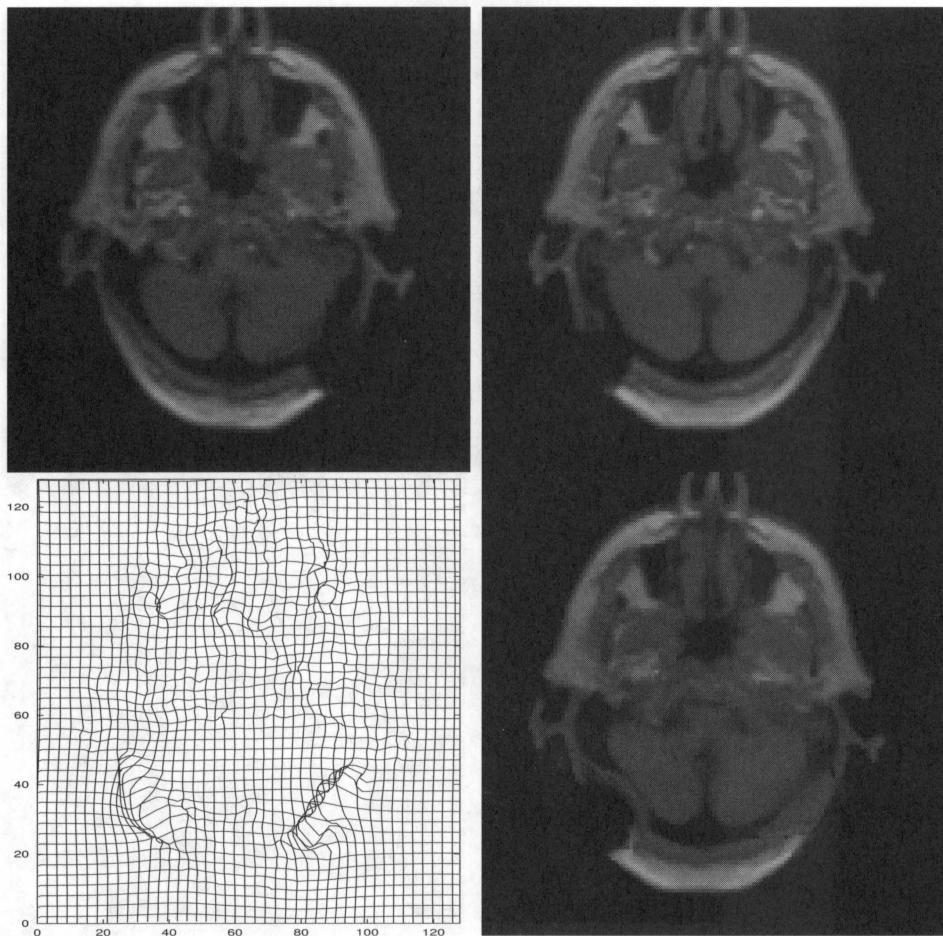


FIGURE 5. In this 3D-matching example the second image is generated from the first by reflecting the original at a central mirror plane. Thus the matching process has to cope with locally large deformations. From top left to bottom right: axial slice through the original 3D image, second image generated by reflection, deformation applied to a uniform grid, matching result.

Acknowledgements

We thank Folkmar Bornemann for giving us insight on a different topic of image smoothing and image segmentation which nevertheless mainly inspired the presented approach to the coupling of scales and the corresponding stopping criteria, Joachim Weickert for interesting discussions on optical flow problems and Rainer Lachner from the Brain-LAB AG, Munich for providing some of the test cases.

This work is partially funded by the Deutsche Forschungsgemeinschaft (DFG) – Programme SPP 1114, Mathematical methods for time series analysis and digital image processing.

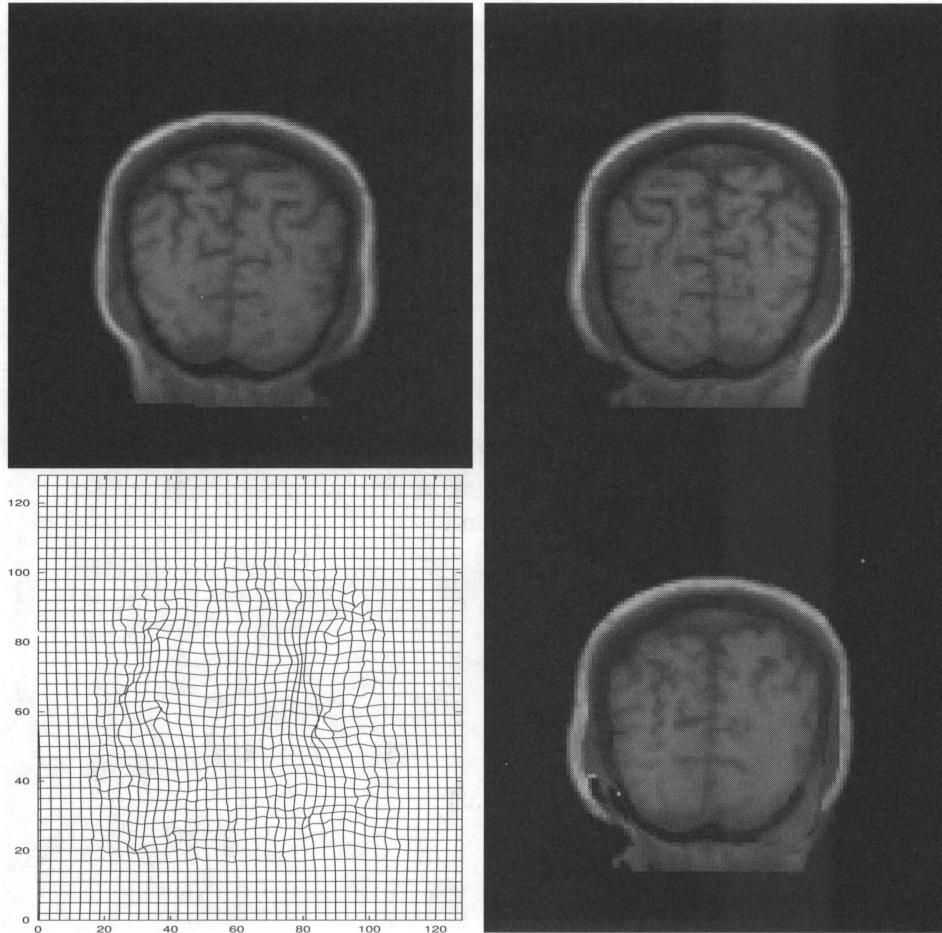


FIGURE 6. From top left to bottom right: saggital slice through the original 3D image, second image generated by reflection, deformation applied to a uniform grid, matching result.

References

- [1] L. Alvarez, F. Guichard, P. L. Lions, and J. M. Morel. Axioms and fundamental equations of image processing. *Arch. Ration. Mech. Anal.*, 123:199–257, 1993.
- [2] L. Alvarez, J. Weickert, and J. Sánchez. A scale-space approach to nonlocal optical flow calculations. In M. Nielsen, P. Johansen, O. F. Olsen, and J. Weickert, editors, *Scale-Space Theories in Computer Vision. Second International Conference, Scale-Space '99, Corfu, Greece, September 1999*, Lecture Notes in Computer Science; 1682, pages 235–246. Springer, 1999.
- [3] L. Alvarez, J. Weickert, and J. Sánchez. Reliable estimation of dense optical flow fields with large displacements. *International Journal of Computer Vision*, 39:41–56, 2000.
- [4] F. Bornemann and P. Deuflhard. The cascadic multigrid method for elliptic problems. *Num. Math.*, 75(2):135–152, 1996.
- [5] G. E. Christensen, S. C. Joshi, and M. I. Miller. Volumetric transformations of brain anatomy. *IEEE Trans. Medical Imaging*, 16, no. 6:864–877, 1997.
- [6] G. E. Christensen, R. D. Rabbitt, and M. I. Miller. Deformable templates using large deformation kinematics. *IEEE Trans. Medical Imaging*, 5, no. 10:1435–1447, 1996.

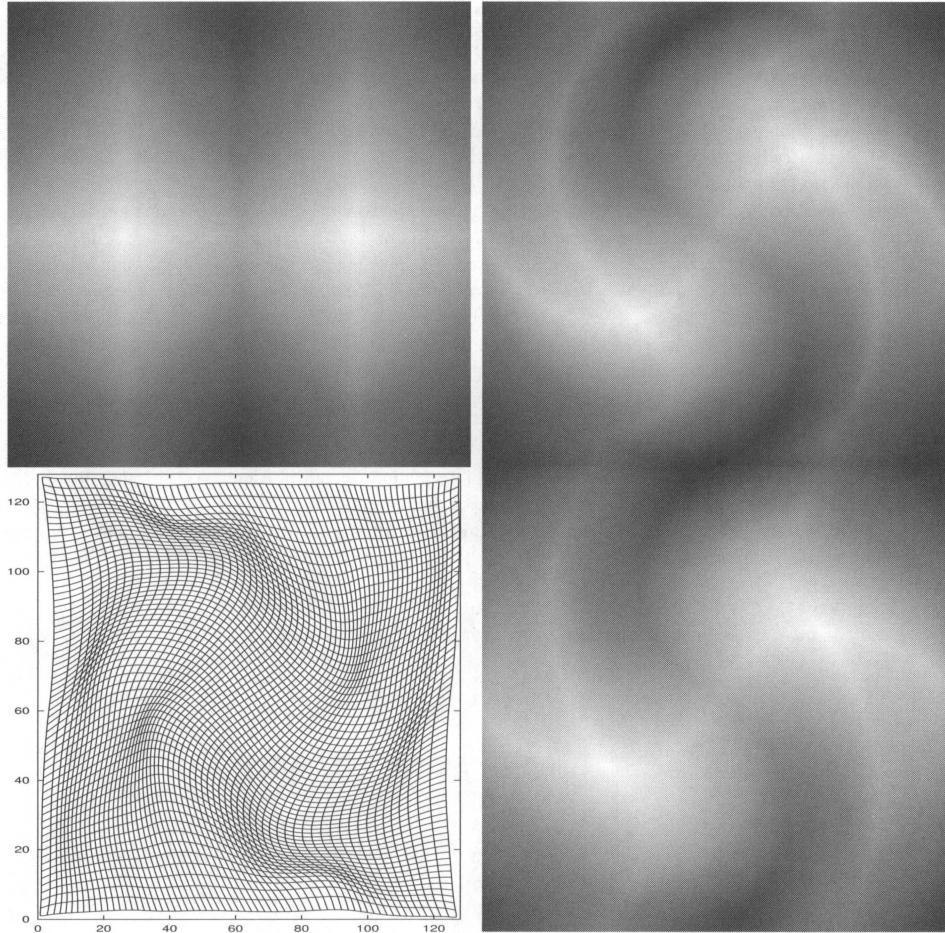


FIGURE 7. The images show the 3D-matching results of a synthetically generated problem (rotational twist by $\frac{\pi}{3}$) with resulting rather large deformations. From top left to bottom right: slice through the original volume image, second image generated via artificial deformation of original image, deformation applied to a uniform grid, matching result.

- [7] C. A. Davatzikos, R. N. Bryan, and J. L. Prince. Image registration based on boundary mapping. *IEEE Trans. Medical Imaging*, 15, no. 1:112–115, 1996.
- [8] R. Deriche, P. Kornobst, and G. Aubert. Optical-flow estimation while preserving its discontinuities: A variational approach. In *Proc. Second Asian Conf. Computer Vision (ACCV '95, Singapore, December 5–8, 1995)*, volume 2, pages 290–295, 1995.
- [9] U. Grenander and M. I. Miller. Computational anatomy: An emerging discipline. *Quarterly Appl. Math.*, LVI, no. 4:617–694, 1998.
- [10] W. Hackbusch. *Multigrid Methods and Applications*. Springer, Berlin/Heidelberg, 1985.
- [11] W. Hackbusch. *Iterative solution of large sparse systems*. Springer, Berlin, 1994.
- [12] M. Hanke and C. Groetsch. Nonstationary iterated tikhonov regularization. *J. Optim. Theory and Applications*, 98:37–53, 1998.
- [13] S. Henn and K. Witsch. Iterative multigrid regularization techniques for image matching. *SIAM J. Sci. Comput. (SISC)*, Vol. 23 no. 4:pp. 1077–1093, 2001.

- [14] W. Hinterberger, O. Scherzer, C. Schnörr, and J. Weickert. Analysis of optical flow models in the framework of calculus of variations. Technical Report 8, Computer Science Series, University Mannheim, 2001.
- [15] S. C. Joshi and M. I. Miller. Landmark matching via large deformation diffeomorphisms. *IEEE Trans. Medical Imaging*, 9, no. 8:1357–1370, 2000.
- [16] P. Kosmol. *Optimierung und Approximation*. de Gruyter Lehrbuch, 1991.
- [17] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multi-modal volume registration by maximization of mutual information. *IEEE Trans. Medical Imaging*, 16, no. 7:187–198, 1997.
- [18] H. H. Nagel and W. Enkelmann. An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8:565–593, 1986.
- [19] E. Radmoser, O. Scherzer, and J. Weickert. Scale-space properties of regularization methods. In M. Nielsen, P. Johansen, O. F. Olsen, and J. Weickert, editors, *Scale-Space Theories in Computer Vision. Second International Conference, Scale-Space '99, Corfu, Greece, September 1999*, Lecture Notes in Computer Science; 1682, pages 211–220. Springer, 1999.
- [20] O. Scherzer and J. Weickert. Relations between regularization and diffusion filtering, 1998.
- [21] J. P. Thirion. Image matching as a diffusion process: An analogy with maxwell's demon. *Medical Imag. Analysis*, 2:243–260, 1998.
- [22] V. Thomée. *Galerkin - Finite Element Methods for Parabolic Problems*. Springer, 1984.
- [23] P. A. Viola. Alignment by maximization of mutual information. Technical Report AITR-1548, 1995.
- [24] J. Weickert. *Anisotropic diffusion in image processing*. Teubner, 1998.
- [25] W. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis. Multi-modal volume registration by maximization of mutual information, 1996.

GERHARD-MERCATOR-UNIVERSITÄT DUISBURG, LOTHARSTRASSE 65, 47048 DUISBURG
E-mail address: [clarenz, droske, rumpf]@math.uni-duisburg.de

A Note on Wavelet-based Inversion Algorithms

Christine De Mol and Michel Defrise

ABSTRACT. Linear ill-posed inverse problems can be regularized by enforcing prior constraints on the solution. In this paper, we consider constraints which can be expressed in the wavelet domain and reflect typically the sparsity of the wavelet transform of the solution. Contrary to wavelet-based inversion techniques which have been formulated in a stochastic framework, we adopt here a purely deterministic setting: we consider a general inverse problem $Kf = g$ with a constraint $\|f\|_B \leq \rho$ on the norm of the solution in some Besov space, and we obtain a stability estimate for this problem.

An explicit regularized solution is defined as the minimizer of a functional $\|Kf - g\|_{L^2}^2 + \mu\|f\|_B^p$ similar to the well-known Tikhonov functional. An iterative algorithm to minimize this functional is presented, based on the Landweber iteration and on soft thresholding in the wavelet domain.

1. Introduction

We consider a linear inverse problem cast in the following form: solve the operator equation

$$(1.1) \quad Kf = g$$

where the solution f (the object) and the data g (its image) belong to some specified function spaces, typically Hilbert or Banach spaces. The choice of the data space must be appropriate for describing real-life situations, i.e. sets of noisy data, so that we take L^2 to be the data space. We assume, in a first stage, that f also belongs to L^2 (and later on, to some Banach space embedded in L^2). The operator K is assumed to be bounded in L^2 , a condition ensuring the continuous dependence of the data g on the solution f , i.e. the well-posedness of the direct problem. In many applications, K is an integral operator with a kernel representing the response of the imaging device. If this linear device is translation-invariant, K reduces to a convolution operator. For simplicity, we assume that the inverse operator exists (i.e. that the null space $N(K)$ is reduced to zero) but is unbounded (in L^2) so that the inverse problem is ill-posed. To tackle the inverse problem, we need an inverse

2000 *Mathematics Subject Classification.* Primary 47A52, 45Q05, 42C40; Secondary 65J20, 65T60, 65J22.

Key words and phrases. Inverse Problems, Regularization, Wavelets.

Significant help from Ingrid Daubechies and Albert Cohen on the topics discussed here is gratefully acknowledged.

solver or restoration algorithm, i.e. an operator mapping the observed image on a good estimate of the unknown object.

Because of ill-posedness, resulting in uncontrolled numerical instability, the inverse of the operator K cannot be used in practice for this task. Moreover, when g does not belong to the range $R(K)$ of the operator, as it is likely to happen for noisy data, there is no exact solution of eq.(1.1). Usually, one looks instead for a *pseudo*- or *least-squares solution*, i.e. for a function \tilde{f} which minimizes the *discrepancy*

$$(1.2) \quad \Delta(f) = \|Kf - g\|_{L^2}^2 .$$

Equivalently, one has to solve the Euler equation $K^*Kf = K^*g$, whose solution is unique when K is invertible (existence, however, is not guaranteed for any $g \in L^2$). Notice that in the case where $N(K) \neq \{0\}$, one can choose, among the set of pseudo-solutions, the unique element of minimal norm f^\dagger , which belongs to $N(K)^\perp$ and is called the *generalized solution* of eq.(1.1). When the inverse (or generalized inverse) of K is unbounded (or provides unstable solutions because of ill-conditioning), it has to be replaced by bounded and stable approximants, so that regularized stable solutions can be defined and used as meaningful approximations of the true solution corresponding to the exact data.

The usual paradigm for restoring stability is to enforce prior knowledge about the solution by means of an additional bound, typically a bound on the L^2 -norm of f (or a quadratic Sobolev norm involving a few derivatives). This is equivalent to adding a penalization term to the discrepancy (1.2) i.e. to looking for a penalized least-squares solution. In classical linear regularization methods (with quadratic constraints), this introduces a penalization of the highly oscillating components, which are the most sensitive to noise, and the resulting rule for penalization depends only on the properties of the operator and on the global smoothness of the solution class. This is clearly understood in the case of a convolution operator which becomes diagonal in the Fourier domain. Classical regularization methods introduce a smooth or sharp cut-off on the highest Fourier components, independently of the data and of the specific object to be restored. Such cut-off implies that the resolution with which the fine details of the solution can be stably retrieved is necessarily limited and that the achievable resolution is essentially the same at all points (see e.g. the book [Ber98] for an extensive discussion of these topics). The classical framework works fine for recovering smooth objects which have their relevant structure contained in the lower part of the spectrum and which have spectral content homogeneously distributed across the space or time domain. However, the Fourier domain is clearly not the appropriate representation for expressing smoothness properties of objects that are either spatially inhomogeneous (non-stationary), with varying local frequency content, and/or present some discontinuities (the Gibbs phenomenon being inherent to Fourier representations). Our arguments apply mutatis mutandis to penalized or truncated SVD solutions in the case of compact operators.

Wavelet bases, on the other hand, have proved very attractive for representing such non-stationary objects with possible sharp edges. Indeed, because of the good localization properties of wavelets in both “time” and “scale”, non-stationary (spatially inhomogeneous) objects with very localized fine-scale structure can be nicely represented by means of a few significant wavelet coefficients. The sparsity of such objects in the wavelet domain has been extensively exploited, mainly for two types of applications: compression and noise removal or ”denoising”. In both

cases, thresholding procedures have been proposed which discard the wavelet coefficients smaller than some prescribed threshold as either insignificant or spurious. As recalled in the next section, it turns out that the natural smoothness spaces associated with wavelets are the so-called Besov spaces, since the norm in such spaces admits a diagonal representation in the wavelet domain. Moreover, the object properties enforced by boundedness in such norms (at least for well chosen representatives in the Besov space family) can enforce both sparsity and local smoothness, yet allowing for the presence of some kinds of discontinuities (see e.g. [Mal98], [Coh00], [Cha98]). Hence, Besov norms appear as quite attractive for expressing the penalization term needed for regularization. To quote [Don95], we can say that wavelet bases allow for “a diagonalization of the prior information” in inverse problems. In section 3, we show that penalization in some appropriate Besov norm yields a regularization method and we derive a corresponding stability estimate, using a purely deterministic framework. An explicit regularized solution is then defined as the minimizer of a penalized least-squares cost functional replacing the usual quadratic Tikhonov functional. A major difficulty, however, is that the operator to be inverted does not in general admit a diagonal representation in wavelet bases, though large classes of operators get sparse representations in such bases, a property which is exploited for several applications in numerical analysis ([Bey91], [Coh00]). A special framework for diagonalizing both the prior constraint and the operator has been proposed by Donoho [Don95] under the name “Wavelet-Vaguelette Decompositions”. We summarize this approach in section 4, stressing the fact that such decompositions can be achieved in practice only for rather special types of operators. Therefore, for general linear inverse problems, enforcing a diagonal constraint in the wavelet domain leads to a difficult nonlinear optimization problem. As an alternative, several authors ([Dic96], [Lou97], [Coh02]) have investigated wavelet-based Galerkin methods.

We propose another route for bypassing this difficulty, using a technique of so-called surrogate functionals, inspired from other applications ([DeP95]). It consists in minimizing the cost functional iteratively, replacing at each iteration the (non-diagonal) discrepancy term by a surrogate diagonal one, so that the minimization completely decouples in the wavelet domain and can be done for each wavelet coefficient separately. This method is described in the last section. We show that in the absence of prior constraint, we simply recover the so-called Landweber iteration, and that with a certain type of Besov prior, soft thresholding in the wavelet domain is applied at each iteration. In other words, the minimization of the cost functional is achieved through a sequence of successive denoising problems.

2. Wavelet bases and Besov spaces

We need to expand the solution of our problem on a wavelet basis. To specify the notations, and in dimension one for simplicity, we define

$$(2.1) \quad \psi_{jk}(x) = 2^{j/2} \psi(2^j x - k)$$

and we assume that the collection $\{\psi_{jk}\}$ for all $j, k \in \mathbb{Z}$, constitutes an orthonormal basis of $L^2(\mathbb{R})$. The translation parameter k represents a discrete “time” whereas

the dilation parameter j is called the “scale”. We will use the following (inhomogeneous) wavelet expansion of $f \in L^2$

$$(2.2) \quad f = \sum_{k=-\infty}^{+\infty} (f, \phi_{0k}) \phi_{0k} + \sum_{j=0}^{+\infty} \sum_{k=-\infty}^{+\infty} (f, \psi_{jk}) \psi_{jk}$$

where $\{\phi_{0k} = \phi(x - k)\}$, $k \in \mathbb{Z}$, are the scaling functions at some coarsest scale, taken to be $j = 0$. Wavelet bases in higher dimension can be build e.g. by tensor products. We refer to [Dau92], [Lou97] or [Mal98] for the general mathematical framework of wavelet theory.

A very useful mathematical property of wavelet bases is that they also provide unconditional bases of many function spaces such as Hölder, Sobolev or Besov spaces. We just very briefly recall what is needed for our purpose and refer to [Dau92], [Cha98], [Coh00] and to the references quoted therein for the background and the precise theorems. Generally, the unconditional basis property can be ensured by requiring the mother wavelet ψ to be a little smoother than typical functions of the corresponding space. Moreover, membership in this space can be revealed by a simple test on the wavelet coefficients of the candidate function. Let us mention that Besov spaces appear to be particularly suitable to describe smoothness properties of spatially inhomogeneous functions. For the Besov space $B = B_{p,q}^s$, which contains functions having essentially s derivatives or a smoothness of order s in L^p , one has the following norm equivalence:

$$(2.3) \quad \|f\|_B \sim \left(\sum_k |(f, \phi_{0k})|^p \right)^{\frac{1}{p}} + \left(\sum_{j=0}^{+\infty} \left(2^{j\sigma p} \sum_k |(f, \psi_{jk})|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}}$$

where $\sigma = s + \frac{1}{2} - \frac{1}{p}$ in dimension 1 and $\sigma = s + d(\frac{1}{2} - \frac{1}{p})$ in \mathbb{R}^d . The norm equivalence $\|f\|_X \sim \|f\|_Y$ means that there exist strictly positive constants A and B such that

$$(2.4) \quad A\|f\|_X \leq \|f\|_Y \leq B\|f\|_X .$$

Notice that for $p = q = 2$ and $s = 0$, the equivalence (2.3) simply reduces to the Parseval equality expressing the unitarity of the wavelet transform in L^2 . Hence, as abundantly emphasized in the wavelet literature, the smoothness properties of a given class of objects (a ball in Besov space) can be easily expressed in diagonal form on the wavelet coefficients.

In the following of the paper, we will adopt simplified notations, used e.g. in [Coh00], namely

$$(2.5) \quad f = \sum_{\lambda} f_{\lambda} \psi_{\lambda} , \quad f_{\lambda} = (f, \psi_{\lambda})$$

where $\lambda = (j, k)$. We also make the convention that $|\lambda| = j$. The sum runs on $k \in \mathbb{Z}$ and on j from the coarsest scale $j = 0$ to $j = +\infty$. By a slight abuse of notations, at the coarsest scale, we include the scaling functions ϕ_{λ} among the basis functions ψ_{λ} . For wavelets in \mathbb{R}^d , the index k will run on \mathbb{Z}^d and we include in λ the necessary extra labels needed to specify the basis wavelets.

3. A stability estimate for penalization in Besov norms

We consider in this section the regularization of the inverse problem (1.1) by a prior constraint expressed in terms of the norm $\|f\|_B$ of the object in some Besov space $B = B_{p,q}^s$. We take $1 \leq p \leq 2$ and, since the parameter q plays a minor role, we also take $q = p$. As concerns the smoothness parameter s , to ensure embedding in L^2 , we require that $s \geq d(\frac{1}{p} - \frac{1}{2})$. We assume then that the unknown exact solution of the problem, \bar{f} , satisfies the constraint $\|\bar{f}\|_B \leq \rho$, where ρ is some prescribed quantity. We also assume that the error on the data g is bounded by some positive quantity ϵ , i.e. that $\|\bar{g} - g\|_{L^2} = \|K\bar{f} - g\|_{L^2} \leq \epsilon$, where \bar{g} represents the exact (noise-free) data of the problem. Hence, the information provided by the data and by the prior knowledge allows to localize the exact solution within the set

$$(3.1) \quad F(\epsilon, \rho) = \{f \in L^2 ; \|Kf - g\|_{L^2} \leq \epsilon, \|f\|_B \leq \rho\}.$$

The diameter of this set is a measure of the uncertainty of the solution for a given prior and a given noise level ϵ . The maximum diameter of F , namely $\text{diam}(F) = \sup\{\|f - f'\|_{L^2}; f, f' \in F\}$ is bounded by twice the following quantity

$$(3.2) \quad M(\epsilon, \rho) = \sup\{\|h\|_{L^2}; \|Kh\|_{L^2} \leq \epsilon, \|h\|_B \leq \rho\}$$

called the *modulus of continuity* of K^{-1} under the prior. If the set defined by the prior is compact (in L^2), then it follows from a general topological lemma that $M(\epsilon, \rho)$ tends to zero when ϵ tends to zero.

Under some assumption on the smoothing properties of the operator K , characterized by a parameter α , we can derive the following stability estimate

$$(3.3) \quad M(\epsilon, \rho) \leq C \epsilon^{\frac{\sigma}{\sigma+\alpha}} \rho^{\frac{\alpha}{\sigma+\alpha}}$$

where C is some constant depending on σ and α . The proof goes as follows. We have to find a bound for the norm $\|h\|_{L^2}$ under the constraint $\|h\|_B \leq \rho$ on the one hand, which for $q = p$ is equivalent to

$$(3.4) \quad \sum_{\lambda} 2^{|\lambda|\sigma p} |h_{\lambda}|^p \leq \rho^p,$$

and the constraint $\|Kh\|_{L^2} \leq \epsilon$ on the other hand. We assume that the operator K is a smoothing operator of order α , a property which we formulate as the following norm equivalence

$$(3.5) \quad \|Kh\|_{L^2}^2 \sim \sum_{\lambda} 2^{-2|\lambda|\alpha} |h_{\lambda}|^2$$

which is also equivalent to the norm in a Sobolev space of negative order $H^{-\alpha}$, i.e. in a Besov space $B_{2,2}^{-\alpha}$ (see e.g. [Eng96], [Lou97], [Coh02]). Therefore our second constraint is simply

$$(3.6) \quad \sum_{\lambda} 2^{-2|\lambda|\alpha} |h_{\lambda}|^2 \leq \epsilon^2.$$

We divide $\|h\|_{L^2}^2$ as follows into small- and large-scale coefficients

$$(3.7) \quad \sum_{\lambda} |h_{\lambda}|^2 = \sum_{\lambda, |\lambda| \leq J} |h_{\lambda}|^2 + \sum_{\lambda, |\lambda| > J} |h_{\lambda}|^2.$$

To get a bound for the first sum, we use the constraint (3.6):

$$(3.8) \quad \sum_{\lambda, |\lambda| \leq J} |h_\lambda|^2 \leq \sum_{\lambda, |\lambda| \leq J} 2^{2\alpha(J-|\lambda|)} |h_\lambda|^2 \leq 2^{2\alpha J} \epsilon^2.$$

To bound the second term, we use the second constraint and we distinguish between the case $p = 2$, which is the usual quadratic Sobolev case, and the case $1 \leq p < 2$. For $p = 2$, we get the bound

$$(3.9) \quad \sum_{\lambda, |\lambda| > J} |h_\lambda|^2 \leq \sum_{\lambda, |\lambda| > J} 2^{2\sigma(|\lambda|-J-1)} |h_\lambda|^2 \leq 2^{-2\sigma(J+1)} \rho^2.$$

In the case $1 \leq p < 2$, we notice that the sum of squares under the constraint (3.4) is maximized when all the "energy" is concentrated in a single coefficient whose value is the largest value compatible with (3.4), namely $2^{-\sigma(J+1)}\rho$. Hence the bound (3.9) is still valid and we get in both cases that

$$(3.10) \quad \sum_{\lambda} |h_\lambda|^2 \leq 2^{2\alpha J} \epsilon^2 + 2^{-2\sigma(J+1)} \rho^2$$

for any J . Optimizing this bound with respect to J by balancing the two terms we get the bound (3.3). Noticing that the function h , which has a unique non-zero coefficient at that optimal scale J equal to $\epsilon^{\frac{\sigma}{\sigma+\alpha}} \rho^{\frac{\alpha}{\sigma+\alpha}}$, satisfies the two required constraints, we also get the lower bound

$$(3.11) \quad \epsilon^{\frac{\sigma}{\sigma+\alpha}} \rho^{\frac{\alpha}{\sigma+\alpha}} \leq M(\epsilon, \rho).$$

For any $\sigma > 0$, the modulus of continuity (3.2) tends to zero when $\epsilon \rightarrow 0$, and the uncertainty on the solution thereby vanishes in this limit. From (3.11) we also see that the rate $\epsilon^{\frac{\sigma}{\sigma+\alpha}}$ is optimal. We recall that $\sigma = s + d(\frac{1}{2} - \frac{1}{p})$ characterizes the smoothness of the solution, whereas α is the smoothing order of the operator. For $\frac{\sigma}{\sigma+\alpha}$ close to one, the problem is mildly ill-posed, but the stability degrades for large α .

In practice, it is not enough, however, to have a stability estimate: we still need to explicitly define an approximate solution. Following a standard approach in regularization theory (see e.g. [Ber98], [Eng96], [Kir96]), we introduce the following functional on L^2 :

$$(3.12) \quad \Phi_\mu(f) = \|Kf - g\|_{L^2}^2 + \mu \|f\|_B^p$$

and define an approximate solution f_μ^* as a minimizer of $\Phi_\mu(f)$. In (3.12), μ is a positive regularization parameter, and p is the index of the Besov space $B = B_{p,p}^s$. Note that the functional reduces to the well-known Tikhonov functional when $B = L^2$ or B is a Sobolev space $B_{2,2}^s$. An upper bound on the error $\|f_\mu^* - \bar{f}\|_{L^2}$ is obtained by noticing that $\Phi_\mu(f_\mu^*) \leq \Phi_\mu(\bar{f}) \leq \epsilon^2 + \mu\rho^p$ because f_μ^* is a minimizer and $\bar{f} \in F(\epsilon, \rho)$. The two terms in the functional must satisfy this inequality separately, and therefore

$$(3.13) \quad \|Kf_\mu^* - g\|_{L^2}^2 \leq \epsilon^2 + \mu\rho^p$$

$$(3.14) \quad \mu \|f_\mu^*\|_B^p \leq \epsilon^2 + \mu\rho^p$$

or, equivalently, $f_\mu^* \in F(\epsilon', \rho')$ with

$$(3.15) \quad \epsilon' = (\epsilon^2 + \mu\rho^p)^{\frac{1}{2}}$$

$$(3.16) \quad \rho' = \left(\rho^p + \frac{\epsilon^2}{\mu} \right)^{\frac{1}{p}}.$$

The *modulus of convergence* is then defined as

$$(3.17) \quad M_\mu(\epsilon, \rho) = \sup \{ \|f_\mu^* - f\|_{L^2}; f, g \in L^2, \|Kf - g\|_{L^2} \leq \epsilon, \|f\|_B \leq \rho \}.$$

This quantity can be bounded using the triangular inequality. Indeed, for any $f \in F(\epsilon, \rho)$ and $f' \in F(\epsilon', \rho')$, we have

$$(3.18) \quad \begin{aligned} \|K(f' - f)\|_{L^2} &= \|Kf' - Kf\|_{L^2} \\ &\leq \|Kf' - g\|_{L^2} + \|Kf - g\|_{L^2} \leq \epsilon' + \epsilon \end{aligned}$$

$$(3.19) \quad \|f' - f\|_B \leq \|f'\|_B + \|f\|_B \leq \rho' + \rho.$$

From the definition of the modulus of continuity (3.2), this yields the following upper bound on the reconstruction error:

$$(3.20) \quad \|f_\mu^* - \bar{f}\|_{L^2} \leq M_\mu(\epsilon, \rho) \leq M(\epsilon + \epsilon', \rho + \rho').$$

The choice $\mu = \epsilon^2/\rho^p$ yields $\epsilon' = \sqrt{2}\epsilon$ and $\rho' = 2^{1/p}\rho$, and with this choice, $f_\mu^* \rightarrow \bar{f}$ when $\epsilon \rightarrow 0$, i.e. the problem has been *regularized*. Even though the constructed approximate solution f_μ^* does not in general belong to the smallest set $F(\epsilon, \rho)$ in which the exact solution has been localized by the data and by the prior constraint, the stability estimate (3.20) is almost optimal. Indeed, one easily checks that

$$(3.21) \quad M(\epsilon, \rho) \leq M_\mu(\epsilon, \rho) \leq M(\epsilon(1 + \sqrt{2}), \rho(1 + 2^{1/p})).$$

To prove the first inequality, we just remark that the minimizer f_μ^* of the functional (3.12) corresponding to the data $g = 0$ is $f_\mu^* = 0$. For that particular choice of g , we see that the value of $M_\mu(\epsilon, \rho)$ coincides with $M(\epsilon, \rho)$, whence the desired lower bound. For $p = 2$, the above results are standard in regularization theory.

4. Wavelet-Vaguelette and Vaguelette-Wavelet Decompositions

In a stochastic framework and for the so-called “denoising” problem, i.e. when K is merely the identity operator I , Donoho and Johnstone [Don94] have proposed to perform a nonlinear thresholding on the noisy wavelet coefficients of the data $g_\lambda = (g, \psi_\lambda)$ on some orthonormal wavelet basis $\{\psi_\lambda\}$. This amounts to replacing all the (g_λ) ’s by their thresholded values $T_\tau(g_\lambda)$ or $S_\tau(g_\lambda)$ where

$$(4.1) \quad T_\tau(t) = \begin{cases} t & |t| \geq \tau \\ 0 & |t| < \tau \end{cases}$$

for the so-called *hard thresholding* and

$$(4.2) \quad S_\tau(t) = \begin{cases} t - \tau & t \geq \tau \\ 0 & |t| < \tau \\ t + \tau & t \leq -\tau \end{cases}$$

for the so-called *soft thresholding* or *shrinkage*. The soft-thresholded estimate is thus given by

$$(4.3) \quad f_\tau = \sum_\lambda S_\tau(g_\lambda) \psi_\lambda.$$

The threshold value τ ($\tau > 0$) is chosen as a function of the variance of the noise and it may also depend on the scale $|\lambda|$. Such “shrinkage” procedure allows to take profit of the sparsity of the representation of spatially inhomogeneous objects in wavelet bases. Moreover, it is justified by some optimality properties discussed in [Don94] and [Mal98]. Connection with variational techniques has later been shown in [Cha98] so that by minimizing the functional (3.12) for $K = I$ and $p = 1$, which writes in the wavelet domain

$$(4.4) \quad \Phi_\mu(f) = \sum_{\lambda} [|f_{\lambda} - g_{\lambda}|^2 + \mu 2^{|\lambda|\sigma} |f_{\lambda}|] ,$$

one precisely gets the soft thresholding estimate (4.3) with a scale-dependent threshold $\tau = \frac{\mu}{2} 2^{|\lambda|\sigma}$. Hard thresholding can be interpreted as a keep-or-kill procedure, which consists in minimizing the functional (4.4) through estimators of the type $f_{\lambda} = T_{\tau}(g_{\lambda})$. The minimum is obtained by keeping only the terms above the threshold $\tau = \mu 2^{|\lambda|\sigma}$. For $1 < p < 2$, the minimizer has no simple expression but can easily be determined numerically, since this can be done in a separable way, coefficient by coefficient.

There have been two proposals for generalizing the previous framework to solve linear inverse problems, namely the Wavelet-Vaguelette Decomposition (WVD) introduced by Donoho [Don95] and the Vaguelette-Wavelet Decomposition (VWD) of Abramovich and Silverman [Abr98]. These denominations refer to the fact that such decompositions generalize the Singular Value Decomposition (SVD) which provides a diagonal representation of compact operators. On the other hand, as already discussed, wavelet bases are particularly suitable for expressing prior information about the object smoothness in diagonal form. As noticed by Donoho [Don95], there is “a special class of inverse problems in which we can simultaneously quasi-diagonalize both the operator and the a priori information”.

The VWD consists in expanding Kf in an orthonormal wavelet basis, $Kf = \sum_{\lambda} (Kf, \psi_{\lambda}) \psi_{\lambda}$ and in using the following representation for the solution f

$$(4.5) \quad f^{VWD} = \sum_{\lambda} (g, \psi_{\lambda}) \kappa_{\lambda} v_{\lambda}$$

where the v_{λ} can be defined by $v_{\lambda} = K^{-1}\psi_{\lambda}/\kappa_{\lambda}$ and are called ”vaguelettes” since, in general, they are no longer orthonormal wavelets. The VWD consists in denoising the data by applying soft- (or hard-) thresholding to the coefficients $(g_{\lambda}, \psi_{\lambda})$ and then in applying the inverse operator to the denoised data.

To derive the WVD, one defines instead ”vaguelettes” v_{λ} and u_{λ} by

$$(4.6) \quad v_{\lambda} = \frac{1}{\kappa_{\lambda}} K \psi_{\lambda} ; \quad u_{\lambda} = \kappa_{\lambda} (K^*)^{-1} \psi_{\lambda} ,$$

normalized to have $\|u_{\lambda}\|_{L^2} = 1$, $\|v_{\lambda}\|_{L^2} = 1$, and which satisfy the biorthogonality relation $(v_{\lambda}, u_{\lambda'}) = \delta_{\lambda\lambda'}$, so that if f is expanded in an orthonormal wavelet basis $\{\psi_{\lambda}\}$, $f = \sum_{\lambda} (f, \psi_{\lambda}) \psi_{\lambda}$, and Kf in the vaguelettes, $Kf = \sum_{\lambda} (Kf, u_{\lambda}) v_{\lambda}$, we get the following WVD expansion

$$(4.7) \quad f^{WVD} = \sum_{\lambda} \frac{g_{\lambda}}{\kappa_{\lambda}} \psi_{\lambda} , \quad g_{\lambda} = (g, u_{\lambda}) .$$

Hard or soft thresholding is then applied to the coefficients $(g_{\lambda}/\kappa_{\lambda})$. The values κ_{λ} (chosen to be positive) are reminiscent of the singular values and their decay rate reflects the degree of smoothing of the operator K . Donoho assumes that they only

depend on the scale $j = |\lambda|$ and decay as $2^{-\alpha|\lambda|}$. It can be shown that such property actually holds true under condition (3.5) on the operator K . A proof can be found in [Lou97]. In a stochastic framework, Donoho has established convergence rates for the minimax risks corresponding to these WVD thresholded estimators under the assumption that f belong to some ball in a scale of Besov spaces ([Don95]).

In our deterministic setting, using the WVD representation, we can define estimates of the solution by minimizing the functional (3.12) which writes here in diagonal form

$$(4.8) \quad \Phi_\mu(f) = \sum_{\lambda} [|\kappa_\lambda f_\lambda - g_\lambda|^2 + \mu 2^{|\lambda|\sigma p} |f_\lambda|^p]$$

with $f_\lambda = (f, \psi_\lambda)$ and $g_\lambda = (g, u_\lambda)$. For $p = 1$, the minimizer is easily obtained by applying the soft thresholding rule on the WVD coefficients

$$(4.9) \quad (f_\mu^*)_\lambda = S_\tau \left(\frac{g_\lambda}{\kappa_\lambda} \right)$$

with the scale-dependent threshold

$$(4.10) \quad \tau = \frac{\mu}{2} 2^{|\lambda|\sigma} \kappa_{|\lambda|}^{-2}.$$

For the choice $\mu = \epsilon^2/\rho$ and under the condition (3.5), i.e. for $\kappa_\lambda = 2^{-\alpha|\lambda|}$, regularization is achieved and the stability estimate derived in the previous section applies. As a special case, for $\alpha = 0$, we get results for the denoising problem. Approximate minimizers are obtained by replacing in (4.9) the soft thresholding operator S_τ by the hard thresholding one T_τ , with threshold equal to twice the threshold (4.10). For $p = 2$, we recover the classical linear regularization framework with a data-independent penalization of the largest unstable scales j (the fine details). For example, minimizing the functional (4.8) for $p = 2$ with a keep-or-kill procedure yields then truncated WVD estimators, where only the scales $|\lambda| \leq J$ are kept, with J such that

$$(4.11) \quad 2^{-(J+1)(\alpha+\sigma)} < \sqrt{\mu} \leq 2^{-J(\alpha+\sigma)}.$$

This is the analogue of the widely used truncated SVD estimator for the inversion of compact operators.

Now, for special types of operators presenting some homogeneity properties with respect to dilations, the vaguelettes are themselves wavelets, i.e. they are generated by translation and dilation of a single mother function. This happens for integration operators, fractional integration such as the Abel transform, two-dimensional Radon transform and for some convolution operators; then the WVD can be easily determined and computed (see [Don95]). For the Radon transform, through a variational approach, Lee and Lucier [Lee01] generalize the WVD framework and get explicit prescriptions for the choice of the thresholds as a function of the variance of the noise. However, except for cases where the vaguelettes can be easily constructed by hand or computed in a fast way, the difficulty of determining the vaguelettes is equivalent with that of inverting the operator K (see eq(4.6)). To bypass this problem, several authors have investigated wavelet-based Galerkin schemes and their numerical implementations ([Dic96], [Lou97], [Coh02]). As an alternative, we propose an iterative procedure, described in the next section and in which the problem reduces, at each iteration, to an easily solvable diagonal minimization problem.

5. A thresholded Landweber algorithm

Let us consider again the functional

$$(5.1) \quad \Phi_\mu(f) = \Delta(f) + \mu \Omega(f)$$

where $\Delta(f)$ is the discrepancy (1.2) and $\Omega(f)$ represents the prior. We define the following "surrogate" functional for the discrepancy

$$(5.2) \quad \Delta^S(f, \hat{f}) = \|K\hat{f} - g\|_{L^2}^2 + \|D(f - \hat{f})\|_{L^2}^2 - 2\operatorname{Re}(g - K\hat{f}, K(f - \hat{f}))$$

where D is a diagonal positive definite operator whose elements will play the role of relaxation parameters for the iteration. The surrogate functional enjoys, for any f and \hat{f} , the following properties

$$(5.3) \quad \Delta^S(\hat{f}, \hat{f}) = \Delta(\hat{f})$$

$$(5.4) \quad \Delta^S(f, \hat{f}) \geq \Delta(f).$$

The first property is evident while the second is satisfied by means of an appropriate assumption on the operator D . Indeed, it is easy to check that

$$(5.5) \quad \Delta^S(f, \hat{f}) - \Delta(f) = \|D(f - \hat{f})\|_{L^2}^2 - \|K(f - \hat{f})\|_{L^2}^2$$

which is non negative provided that the operator $D^*D - K^*K \geq 0$. For simplicity, we will renormalize the operator K in order that $\|K\| \leq 1$ and take D to be the identity operator. As easily seen by a direct computation, the surrogate residual then takes the following form

$$(5.6) \quad \Delta^S(f, \hat{f}) = \|f - \hat{f}_C\|_{L^2}^2 - \|\hat{f} - \hat{f}_C\|_{L^2}^2 + \|K\hat{f} - g\|_{L^2}^2$$

where

$$(5.7) \quad \hat{f}_C = \hat{f} + K^*(g - K\hat{f}).$$

The function f minimizing $\Delta^S(f, \hat{f})$ is clearly \hat{f}_C , since the two last terms are independent of f .

The cost functional $\Phi_\mu(f)$ will now be minimized through the iterative procedure

$$(5.8) \quad f^{n+1} = \arg \min_f \Phi_\mu^S(f, f^n)$$

where

$$(5.9) \quad \Phi_\mu^S(f, \hat{f}) = \Delta^S(f, \hat{f}) + \mu \Omega(f)$$

i.e. by successive minimization of the surrogate functional anchored at the previous iterate. The procedure has to be initialized, e.g. by setting

$$(5.10) \quad f^0 = 0.$$

It is guaranteed that the cost is decreasing at each iteration. Indeed, we have

$$(5.11) \quad \Phi_\mu(f^{n+1}) \leq \Phi_\mu^S(f^{n+1}, f^n) \leq \Phi_\mu^S(f^n, f^n) = \Phi_\mu(f^n)$$

using the property (5.4) and the fact that f^{n+1} minimizes $\Phi_\mu^S(f, f^n)$.

In the case where we minimize the residual without any prior, i.e. when $\mu = 0$, the above iteration yields

$$(5.12) \quad f^{n+1} = f^n + r^n$$

where

$$(5.13) \quad r^n = K^*(g - Kf^n)$$

is the backprojected residual of the n^{th} iterate. This is nothing else than the well-known Landweber iterative method (see e.g. [Ber98], [Eng96], [Kir96]) whose convergence to the (generalized) solution of $Kf = g$ has been established (remember that we have taken $\|K\| < 1$ and $f^0 = 0$).

With the L^2 prior $\Omega(f) = \|f\|_{L^2}^2$, we get a damped or regularized Landweber iteration

$$(5.14) \quad f^{n+1} = \frac{1}{1 + \mu} [f^n + K^*(g - Kf^n)]$$

(see e.g. [Ber98]). The convergence of (5.14) is more easily proved since it defines a contractive mapping.

Finally, for a general Besov prior with $p = 1$, we get the following thresholded Landweber iteration, expressed in the wavelet domain by

$$(5.15) \quad f_\lambda^{n+1} = S_\tau(f_\lambda^n + r_\lambda^n)$$

where S_τ is the soft thresholding operator (4.2) with threshold $\tau = \frac{\mu}{2} 2^{|\lambda|}\sigma$.

Work is in progress concerning the study of the properties of the algorithm on the one hand and its numerical implementation on the other hand. As for the convergence properties, the difficulty is that the iteration is defined by a mapping which is not contracting (although non expansive). Hence general results on such mappings cannot guarantee the strong convergence of the iterates to a fixed point of the iteration. We also have to prove that a fixed point of the iteration is a minimum of the cost functional (3.12) and to settle the uniqueness question. Concerning the numerical implementation of the algorithm, let us remark that we have to perform thresholding in the wavelet domain at each iteration. To go back and forth to the wavelet domain we can rely on fast wavelet transform algorithms, which have linear complexity. Hence the complexity of the present algorithm is not much greater than the usual Landweber algorithm. We finally remark that our thresholded Landweber algorithm could also be used to enforce the sparsity of the restored solution in its original (space or time) domain. This amounts to a penalization of the L^1 -norm of the object, or of a closely related norm. Such sparsity would be helpful e.g. for some problems encountered in medical imaging ([Li01]). For other specific problems, sparsity in the Fourier domain may be desirable instead, and our algorithm applied to such case yields for convolution operators a modified Gerchberg-Papoulis algorithm with nonlinear thresholding in the Fourier domain.

References

- [Abr98] F. Abramovich and B. W. Silverman, *Wavelet Decomposition Approaches to Statistical Inverse Problems*. Biometrika **85** (1998), 115–129.
- [Ber98] M. Bertero and P. Boccacci, *Introduction to Inverse Problems in Imaging*, Institute of Physics, Bristol, 1998.
- [Bey91] G. Beylkin, R. Coifman and V. Rokhlin, *Fast Wavelet Transforms and Numerical Algorithms I*. Comm. Pure Appl. Math. **44** (1991), 141–183.
- [Cha98] A. Chambolle, R. A. DeVore, N.-Y. Lee and B. J. Lucier, *Nonlinear Wavelet Image Processing: Variational Problems, Compression, and Noise Removal through Wavelet Shrinkage*. IEEE Trans. Image Processing **7** (1998), 319–335.
- [Coh00] A. Cohen, *Wavelet methods in numerical analysis*, Handbook of Numerical Analysis, vol. VII, P. G. Ciarlet and J. L. Lions eds., Elsevier, Amsterdam, 2000.

- [Coh02] A. Cohen, M. Hoffmann and M. Reiss, *Adaptive wavelet Galerkin methods for linear inverse problems*, in preparation and private communication.
- [Dau92] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.
- [DeP95] A. R. De Pierro, *A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography*. IEEE Trans. Med. Imag. **14** (1995), 132–137.
- [Dic96] V. Dicken and P. Maass, *Wavelet-Galerkin methods for ill-posed problems*. J. Inv. Ill-Posed Problems **4** (1996) 203–222.
- [Don94] D. Donoho and I. Johnstone, *Ideal spatial adaptation via wavelet shrinkage*. Biometrika **81** (1994), 425–455.
- [Don95] D. Donoho, *Nonlinear solution of Linear Inverse Problems by Wavelet-Vaguelette Decomposition*. Appl. Comp. Harmonic Anal. **2** (1995), 101–126.
- [Eng96] H. W. Engl, M. Hanke and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, Dordrecht, 1996.
- [Kir96] A. Kirsch, *An Introduction to the Mathematical Theory of Inverse Problems*, Springer, New-York, 1996.
- [Lee01] N.-Y. Lee and B. J. Lucier, *Wavelet Methods for Inverting the Radon Transform with Noisy Data*. IEEE Trans. Image Processing **10** (2001), 79–94.
- [Li01] M. Li, H. Yang and H. Kudo, *Accurate Iterative Reconstruction Algorithm for Sparse Objects: Application to 3-D Blood-Vessel Reconstruction from a Limited Number of Projections*, Proc. 6th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine, Asilomar (CA), 2001, pp. 245–248.
- [Lou97] A. K. Louis, P. Maass and A. Rieder, *Wavelets: Theory and Applications*, Wiley, Chichester, 1997.
- [Mal98] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, 1998.

DEPARTMENT OF MATHEMATICS, UNIVERSITE LIBRE DE BRUXELLES, CAMPUS PLAINE C.P.217,
1050 BRUSSELS, BELGIUM

E-mail address: `demol@ulb.ac.be`

DEPARTMENT OF NUCLEAR MEDICINE, VRIJE UNIVERSITEIT BRUSSEL, AZ-VUB, LAARBEEKLAAN
101, 1090 BRUSSELS, BELGIUM

E-mail address: `mdefrise@minf.vub.ac.be`

A Comparison Between the Wavelet-Galerkin and the Sinc-Galerkin Methods in Solving Nonhomogeneous Heat Equations

Mohamed El-Gamel and Ahmed I. Zayed

ABSTRACT. One of the new techniques used in solving boundary-value problems involving partial differential equations is the wavelet-Galerkin method. This method has been shown to be a powerful numerical tool for finding fast and accurate solutions. A less known technique that has been around for almost two decades is the sinc-Galerkin method.

In this paper we solve boundary-value problems involving nonhomogeneous heat equations using these two methods and then compare the results. It is shown that the sinc-Galerkin in many instances gives better results. In the wavelet-Galerkin solutions, Daubechies 6 wavelets are used because they give better results than those of lower degree wavelets. The results are then compared with those obtained using the sinc-Galerkin method. Although the sinc-Galerkin solution required slightly more computational effort than the wavelet-Galerkin solution, it resulted in more accurate results than the wavelet-Galerkin method, especially in the presence of singularities. This may be attributed to the fact that in the wavelet method in order to obtain an error of order $2^{-\alpha j N}$ at resolution j for some α , the solution and the nonhomogeneous terms have to be nice and smooth.

1. Introduction

There is vast literature on numerical solutions of boundary-value problems involving ordinary and partial differential equations. Some of the well-known techniques used in solving these problems are the finite differences, finite elements, and multi-grid methods.

The Galerkin method is another type of numerical techniques used to solve partial differential equations with boundary or initial conditions. In this method the solution is assumed to be in a Hilbert space H with inner product $\langle \cdot, \cdot \rangle$, and one seeks an approximate solution to the problem in the form $\phi(x) = \sum_{k=1}^N a_k \psi_k(x)$, where $\{\psi_k(x)\}_{k=1}^N$ is a basis for an N -dimensional subspace of functions, S . The functions $\psi_k(x)$, $k = 1, 2, \dots, N$, are called test functions and the space S is called the test space.

1991 *Mathematics Subject Classification*. Primary: 65N30, 65T60; Secondary: 35K05, 42C40.

Key words and phrases. Wavelet-Galerkin, Sinc-Galerkin, Heat Equations, Numerical Solutions.

The approximate solution is chosen in such a way that the error is minimum. This can be achieved by choosing the error to be orthogonal to the test space. In other words, the approximate solution is the projection of the actual solution onto the test space. Sometimes the test space is enlarged by admitting weak solutions to the problem, which, depending on the differential operator and the boundary conditions involved, can be obtained by integration by parts first.

To simplify the computations, the basis test functions $\{\psi_k(x)\}_{k=1}^N$ are taken to be orthogonal and in many cases they are polynomials or splines. In essence, the Galerkin method is a discretization scheme in which the expansion coefficients $\{a_k\}_{k=1}^N$ are obtained by solving a set of N algebraic equations. For example, consider the problem $Lu = f$, where L is a self-adjoint operator and f is a known function. The Galerkin method yields the system of equations

$$\sum_{k=1}^N a_k \langle L\psi_k(x), \psi_j(x) \rangle = \langle f(x), \psi_j(x) \rangle, \quad j = 1, \dots, N,$$

which is an algebraic system of equations that can be solved for the unknown coefficients a_k .

There are a number of hybrids of the Galerkin method that use different types of test functions. In the last decade or so, wavelets have been used in many applications, including numerical solutions of ordinary and partial differential equations. In the wavelet-Galerkin method, the approximate solution is obtained in a multiresolution analysis setting; hence, yielding approximate solutions at different levels of resolution. The wavelets of choice in this approach are the Daubechies' because they are orthonormal and compactly supported. But one of the drawbacks of this method is that the Daubechies wavelets are not known in closed form which makes the evaluation of their integrals, derivatives, and moments more tedious.

A less known Galerkin scheme is the Sinc-Galerkin in which the test functions are translates of the sinc function, $S(x) = \sin \pi x / \pi x$. The sinc method, which was introduced and developed by F. Stenger more than twenty years ago [29, 31], is based on the Whittaker-Shannon-Kotelnikov sampling theorem for entire functions. This method, which uses entire functions as bases, has many advantages over classical methods that use polynomials as bases. For example, in the presence of singularities, it gives a much better rate of convergence and accuracy than polynomial methods.

The aim of this paper is to compare the wavelet-Galerkin and the sinc-Galerkin methods in solving parabolic problems, in particular, in solving boundary-value problems involving perturbed heat equations. It will be shown that although the wavelet-Galerkin method is more popular, the sinc-Galerkin method gives better results, especially in the presence of singularities. To the best of our knowledge such a comparison has not been done before.

The paper is organized as follows. In Section 2, we introduce the wavelet-Galerkin method and show how wavelets are used to solve partial differential equations numerically. In Section 3 the sinc-Galerkin method is discussed. In Section 4, we apply both methods to specific problems, compare the results, and close with conclusions.

2. The Wavelet-Galerkin Method

Wavelet-based numerical solutions of partial differential equations have recently been developed [1, 7, 12, 25, 26]. To a certain extent, the wavelet technique is a strong competitor to the finite element method, at least for problems with simple geometry. Although the wavelet method provides an efficient alternative for solving partial differential equations numerically, it is not as easy to implement as the traditional finite difference method. The reason is that the use of the wavelet-Galerkin method to solve partial differential equations leads to the problem of computing integrals whose integrands involve products of compactly supported wavelets and their derivatives. These integrals are evaluated using what is known as the connection coefficient method. Algorithms for evaluating connection coefficients for bounded and unbounded domains have been developed in [4, 16]. Surprisingly, the connection coefficients for unbounded domains are easier to evaluate than those for bounded domains.

In this section, we describe how wavelets may be used to solve parabolic partial differential equations. For clarity, the method is presented for the heat equation

$$(2.1) \quad LU(x, t) = U_t(x, t) - U_{xx}(x, t) = f(x, t), \quad 0 \leq x \leq 1, \quad t > 0,$$

subject to the boundary conditions

$$(2.2) \quad U(0, t) = 0, \quad U(1, t) = 0,$$

and the initial condition

$$(2.3) \quad U(x, 0) = 0.$$

2.1. Daubechies Wavelet Bases. First, we will give a brief summary of Daubechies wavelets [9, 10, 20]. Denote by $L^2(\mathbb{R})$ the space of square integrable functions on the real line. The orthonormal basis, $\{\psi_{jk}(x)\}$, of compactly supported wavelets of $L^2(\mathbb{R})$ is formed by dilations and translations of a single function $\psi(x)$, called a “mother wavelet”

$$(2.4) \quad \psi_{jk}(x) = 2^{j/2}\psi(2^jx - k), \quad j, k \in \mathbb{Z}.$$

The function $\psi(x)$ has a companion, the scaling function $\phi(x)$. They both satisfy the following two-scale relations

$$(2.5) \quad \phi(x) = \sum_{i=0}^{N-1} P_i \phi(2x - i),$$

$$(2.6) \quad \psi(x) = \sum_{i=2-N}^1 (-1)^i P_{1-i} \phi(2x - i),$$

where the coefficients P_i ($i = 0, 1, \dots, N - 1$) appearing in the 2-scale relations (2.5) and (2.6) are called the wavelet filter coefficients. The support of the scaling function ϕ is the interval $[0, N - 1]$ while that of the corresponding wavelet ψ is the interval $[1 - N/2, N/2]$. The Daubechies wavelet filter coefficients for $N = 4, 6, 8$ and 10 are listed in [15, 16].

In this paper, we take $N = 6$; hence, the Daubechies wavelet filter coefficients satisfy the following conditions

$$(2.7) \quad \sum_{i=0}^5 P_i = 2,$$

$$(2.8) \quad \sum_{i=0}^5 P_i P_{i-m} = 2\delta_{0,m},$$

$$(2.9) \quad \sum_{k=0}^5 (-1)^k P_k k^l = 0, \quad l = 0, 1, 2,$$

and

$$(2.10) \quad \sum_{i=-4}^1 (-1)^i P_{1-i} P_{i-2m} = 0, \quad \text{for any integer } m,$$

where $\delta_{0,m}$ is the Kronecker delta. These relations follow from the following relations satisfied by the scaling function $\phi(x)$ and its mother wavelet $\psi(x)$:

$$(2.11) \quad \int_{-\infty}^{\infty} \phi(x) dx = 1,$$

$$(2.12) \quad \int_{-\infty}^{\infty} \phi(x-j) \phi(x-i) dx = \delta_{i,j},$$

$$(2.13) \quad \int_{-\infty}^{\infty} x^k \psi(x) dx = 0, \quad k = 0, 1, 2,$$

and

$$(2.14) \quad \int_{-\infty}^{\infty} \phi(x) \psi(x-k) dx = 0, \quad \text{for any integer } k.$$

We define V_j as the closure of the space generated by $\{\phi_{jk}, k \in Z\}$ and W_j , its orthogonal complement in V_{j+1} , as the closure of the space generated by $\{\psi_{jk}, k \in Z\}$. This condition implies that

$$(2.15) \quad V_{j+1} = V_j \oplus W_j,$$

where \oplus denotes the orthogonal direct sum. The sub-spaces V_j verify the following conditions

$$(2.16) \quad V_j \subset V_{j+1},$$

$$(2.17) \quad \cup_{j \in Z} V_j = L^2(R), \text{ and } \cap_{j \in Z} V_j = \{0\}.$$

The space $L^2(R)$ is represented as a direct sum

$$(2.18) \quad L^2(R) = \oplus_{j \in Z} W_j.$$

On each fixed scale j , the wavelets $\{\psi_{jk}(x)\}_{k \in Z}$ form an orthonormal basis of W_j and the scaling functions $\{\phi_{jk}(x)\}_{k \in Z}$ form an orthonormal basis of V_j . The set of spaces V_j is called a multiresolution analysis of $L^2(R)$. These spaces will be used to approximate the solutions of partial differential equations using the Galerkin method.

2.2. Connection Coefficients. Let the integral of the product of the scaling function and its n -th order derivative $\phi^{(n)}(x - k)$ be denoted by

$$(2.19) \quad \Gamma_k^n(x) = \int_0^x \phi^{(n)}(y - k)\phi(y)dy.$$

We need to evaluate $\Gamma_k^n(x)$ because they play an important role in the wavelet-Galerkin method to solve partial differential equations. Here it should be noted that these integrals are not zero because the orthogonality of the translates of ϕ is usually lost by differentiation. These integrals could be calculated by standard (Gauss) quadrature, however, the standard quadrature is not available for many wavelet bases. But due to the unusual smoothness characteristics of the scaling function and its derivatives, one can replace the role of the Gauss quadrature by the connection coefficient method in order to evaluate the integrals in equation (2.19) accurately [16]. We also need to calculate a few other connection coefficients such as

$$(2.20) \quad M_k^m(x) = \int_0^x y^m \phi(y - k)dy.$$

Algorithms for computing these coefficients are given in [22].

2.3. Wavelet-Galerkin Solution. In this section, we discuss the use of the wavelet-Galerkin method to solve the problem given by Equations (2.1)–(2.3). Let the solution $u(x, t)$ of the problem be approximated by its j -th level wavelet series on the interval $(0, 1)$, i.e.,

$$(2.21) \quad U(x, t) = \sum_{k=-4}^{2^j-1} u_{jk}(t) \phi_{jk}(x), \quad k \in Z,$$

where u_{jk} are unknown coefficients, and $\phi_{jk}(x) = 2^{j/2} \phi_{jk}(2^j x - k)$, $j > 0$. Substitution of Equation (2.21) into equation (2.1), yields

$$(2.22) \quad \sum_{k=-4}^{2^j-1} \frac{d}{dt} u_{jk}(t) \phi_{jk}(x) = \sum_{k=-4}^{2^j-1} u_{jk}(t) \frac{d^2}{dx^2} \phi_{jk}(x) + f(x, t).$$

To determine the coefficient u_{jk} , we take the inner product of both sides of equation (2.22) with ϕ_{jl} ,

$$(2.23) \quad \begin{aligned} \sum_{k=-4}^{2^j-1} u'_{jk}(t) \int_0^1 \phi_{jk}(x) \phi_{jl}(x) dx &= \sum_{k=-4}^{2^j-1} u_{jk}(t) \int_0^1 \phi''_{jk}(x) \phi_{jl}(x) dx \\ &\quad + \int_0^1 f(x, t) \phi_{jl}(x) dx, \quad l = -4, -3, \dots, 2^j - 1. \end{aligned}$$

We assume that $f(x, t) = h(t)z(x)$, where $z(x) = a_m x^m + a_{m-1} x^{m-1} + \dots + a_1 x + a_0$, is a polynomial of degree m in x , otherwise, we approximate z by such a polynomial if necessary. Then equation (2.23) can be written as

$$(2.24) \quad \sum_{k=-4}^{2^j-1} a_{kl}^j u'_{jk} = \sum_{k=-4}^{2^j-1} c_{kl}^j u_{jk} + h(t) d_{ml}^j, \quad l = -4, -3, \dots, 2^j - 1,$$

where a_{kl}^j , c_{kl}^j and d_{ml}^j are given by

$$(2.25) \quad a_{kl}^j = \int_0^1 \phi_{jk}(x)\phi_{jl}(x)dx = \Gamma_{k-l}^0(2^j - l) - \Gamma_{k-l}^0(-l),$$

$$(2.26) \quad c_{kl}^j = \int_0^1 \phi_{jk}''(x)\phi_{jl}(x)dx = \Gamma_{k-l}^2(2^j - l) - \Gamma_{k-l}^2(-l),$$

and

$$(2.27) \quad d_{ml}^j = \int_0^1 [a_m x^m + a_{m-1} x^{m-1} + \cdots + a_1 x + a_0] \phi_{jl}(x) dx \\ = a_m \frac{1}{2^{(m+\frac{1}{2})j}} M_l^m(2^j) + a_{m-1} \frac{1}{2^{((m-1)+\frac{1}{2})j}} M_l^{m-1}(2^j) \\ + \cdots + a_1 \frac{1}{2^{(\frac{3}{2})j}} M_l^1(2^j) + a_0 \frac{1}{2^{(\frac{1}{2})j}} M_l^0(2^j)$$

The algorithm for calculating Γ_{k-l}^0 , Γ_{k-l}^2 , and M_l^m has been described in [22]. The initial condition for the differential equation (2.22) is derived from the initial condition $u(x, 0)$ of the problem.

Here, we first consider the general case in which the initial condition is of the form $u(x, 0) = g(x)$, where as before $g(x)$ is assumed to be polynomial of degree m , otherwise we approximate it by such a polynomial. For this case the initial conditions $u_{jk}(0)$ satisfy

$$(2.28) \quad u(x, 0) = \sum_{k=-4}^{2^j-1} u_{jk}(0) \phi_{jk}(x) = g(x).$$

We take the inner product of both sides of equation (2.28) with ϕ_{jl} , $l \in Z$,

$$(2.29) \quad \sum_{k=-4}^{2^j-1} u_{jk}(0) \int_0^1 \phi_{jk}(x) \phi_{jl}(x) dx = \int_0^1 g(x) \phi_{jl}(x) dx, \quad l = -4, -3, \dots, 2^j - 1,$$

Hence, the initial condition $u_{jk}(0)$ can be determine by solving the following linear system of algebraic equations

$$(2.30) \quad \sum_{k=-4}^{2^j-1} a_{kl}^j u_{jk}(0) = d_{ml}^j,$$

where

$$(2.31) \quad d_{ml}^j = \int_0^1 [a_m x^m + a_{m-1} x^{m-1} + \cdots + a_1 x + a_0] \phi_{jl}(x) dx, \\ l = -4, -3, \dots, 2^j - 1,$$

To get the expansion coefficient u_{jk} in the approximate solution (2.21) firstly we have to obtain the initial conditions $u_{jk}(0)$ from the system in equation (2.30), then use them with the boundary condition to solve the first order system of differential equation (2.24), which can be written in the matrix form

$$(2.32) \quad AU' = CU + h(t)D,$$

where A is a matrix of order $(2^j + 4) \times (2^j + 4)$ with entries $[a_{kl}^j]$ and C is a matrix of order $(2^j + 4) \times (2^j + 4)$ with entries $[c_{kl}^j]$, and D is a $(2^j + 4) \times 1$ vector with

entries $[d_{ml}^j]$. Clearly, for the problem (2.1)–(2.3), $g(x) = 0$, and hence $d_{ml}^j = 0$, which implies that $u_{jk}(0) = 0$ because A is nonsingular.

We will find an approximate solution by using the finite difference method and setting

$$(2.33) \quad U'|_i = \frac{U_{i+1} - U_i}{\Delta t},$$

and the average value of U as

$$(2.34) \quad U = \frac{U_{i+1} + U_i}{2},$$

then the equation (2.32) may be written in the form

$$(2.35) \quad A_1 U_{i+1} = C_1 U_i + B_1 D,$$

where

$$(2.36) \quad A_1 = A - \frac{\Delta t}{2} C,$$

$$(2.37) \quad C_1 = A + \frac{\Delta t}{2} C,$$

and

$$(2.38) \quad B_1 = \Delta t h(t)$$

Equation (2.35) is a linear system of $(2^j + 4)$ equations in $(2^j + 4)$ unknown coefficients. This system may be easily solved by a variety of methods. In this paper we use *Q-R* method [13, 24]. By solving this system, we obtain an approximate solution at resolution level j . If we substitute $u_{jk} = u_{jk}(t_i)$ in equation (2.21) we obtain an approximate solution to the problem (2.1)–(2.3) at time $t = t_i$.

3. The Sinc-Galerkin Method

The Sinc-Galerkin procedure for solving the problem (2.1)–(2.3) begins by selecting composite sinc functions appropriate to the intervals $(0, 1)$ and $(0, \infty)$ so that their translates form a basis functions for the expansion of the approximate solution $u(x, t)$. A thorough review of properties of the sinc function and the general sinc-Galerkin method can be found in [2, 3, 6, 11, 21].

The Sinc-Galerkin procedure we shall use, which is sometimes called the approximation of derivatives method, is not as general as the one used in [30] and which can be used to solve nonlinear problems. Both procedures are based on the Sinc method, but the latter, which is called the *Sinc convolution* method, can be applied to the integral equation formulation of the partial differential equation under consideration; see Section 4.6 in [29]. Both the approximation of derivatives method and the Sinc convolution method reduce the boundary-value problem into a system of algebraic equations of the form

$$(3.1) \quad U A_t + A_x U = C$$

where U , A_t , A_x , and C are matrices in which U is the unknown. In our study the matrix A_t is non-singular whenever its order is even. The next section contains an overview of properties of the Sinc function that are used in the sequel.

3.1. Sinc Interpolation. The sinc function is defined on the whole real line by

$$(3.2) \quad \text{sinc}(x) = \frac{\sin(\pi x)}{\pi x}, \quad -\infty < x < \infty,$$

For $h > 0$, the translated sinc functions with evenly spaced nodes are given as

$$(3.3) \quad S(k, h)(x) = \text{sinc}\left(\frac{x - kh}{h}\right), \quad k = 0 \pm 1, \pm 2, \dots$$

If f is defined on the real line, then for $h > 0$ the series

$$(3.4) \quad C(f, h) = \sum_{k=-\infty}^{\infty} f(hk) \text{sinc}\left(\frac{x - hk}{h}\right).$$

is called the Whittaker cardinal expansion of f whenever this series converges. The properties of Whittaker cardinal expansions have been studied and are thoroughly surveyed in [29, 32]. These properties are derived in the infinite strip D_s of the complex plane where for $d > 0$

$$(3.5) \quad D_s = \left\{ \zeta = \xi + i\eta : |\eta| < d \leq \frac{\pi}{2} \right\}$$

Approximations can be constructed for infinite, semi-finite, and finite intervals. To construct approximations on the interval $(0, 1)$ and $(0, \infty)$ respectively, which are used in this paper, consider the conformal maps

$$(3.6) \quad \phi(z) = \ln\left(\frac{z}{1-z}\right),$$

and

$$(3.7) \quad \gamma(w) = \ln(w).$$

The map ϕ carries the eye-shaped region

$$(3.8) \quad D_E = \left\{ z = x + iy : \left| \arg\left(\frac{z}{1-z}\right) \right| < d \leq \frac{\pi}{2} \right\},$$

onto the infinite strip D_s . Similarly, the map γ carries the infinite wedge

$$(3.9) \quad D_w = \left\{ w = t + is : |\arg(w)| < d \leq \frac{\pi}{2} \right\},$$

onto the strip D_s . The compositions

$$(3.10) \quad S_i(x) = S(i, h_x) \circ \phi(x),$$

and

$$(3.11) \quad S_j(t) = S(j, h_t) \circ \gamma(t).$$

define the basis elements for equation (2.1) on the intervals $(0, 1)$ and $(0, \infty)$, respectively.

The “mesh sizes” h_x and h_t represent the mesh sizes in D_s for the uniform grids $\{kh_x\}$, and $\{kh_t\}$, $-\infty < k < \infty$. The sinc grid points $z_k \in (0, 1)$ in D_E will be denoted by x_k because they are real. Similarly, the grid points $w_k \in (0, \infty)$ in D_w will be denoted by t_k . Both are inverse images of the equi-spaced grids, that is

$$(3.12) \quad x_k = \phi^{-1}(kh_x) = \frac{e^{kh_x}}{1 + e^{kh_x}},$$

and

$$(3.13) \quad t_k = \gamma^{-1}(kh_t) = e^{kh_t}.$$

To simplify the notation throughout the remainder of this section, the pairs ϕ , D_E and γ , D_w are referred to generically as χ , and D . It is understood that the subsequent definition and theorems hold in either setting. Furthermore, the inverse of χ is denoted by ψ .

The class of functions suitable for sinc interpolation and quadrature is denoted by $B(D)$ and defined below.

DEFINITION 3.1. Let $B(D)$ be the class of functions F that are analytic in D , and satisfy

$$(3.14) \quad \int_{\psi(L+t)} |F(z)dz| \rightarrow 0, \quad \text{as } t = \pm\infty,$$

where

$$(3.15) \quad L = \{iy : |y| < d \leq \frac{\pi}{2}\},$$

and on the boundary of D (denoted ∂D) satisfy

$$(3.16) \quad N(F) = \int_{\partial D_E} |F(z)dz| < \infty.$$

The proof of the following theorem for functions in $B(D)$ may be found in [18, 28].

THEOREM 3.1. Let Γ be $(0, 1)$ or $(0, \infty)$ when $\chi = \phi$ or γ , respectively. If $F \in B(D)$ and $\tau_j = \psi(jh) = \chi^{-1}(jh)$, $j = 0, \pm 1, \pm 2, \dots$, then for $h > 0$ sufficiently small

$$(3.17) \quad \int_{\Gamma} F(\tau)d\tau - h \sum_{j=-\infty}^{\infty} \frac{F(\tau_j)}{\chi'(\tau_j)} = \frac{i}{2} \int_{\partial D} \frac{F(z)k(\chi, h)(z)}{\sin(\pi\chi(z)/h)} dz \equiv I_F,$$

where

$$(3.18) \quad |k(\chi, h)|_{z \in \partial D} = \left| \exp \left[\frac{i\pi \chi(z)}{h} \operatorname{sgn}(Im(\chi(z))) \right] \right|_{z \in \partial D} = e^{-\pi d/h}.$$

For the sinc-Galerkin method, the infinite quadrature rule must be truncated to a finite sum. The following theorem indicates the conditions under which exponential convergence results.

THEOREM 3.2. If there exist positive constants α, β and C such that

$$(3.19) \quad \left| \frac{F(\tau)}{\chi'(\tau)} \right| \leq C \begin{cases} \exp(-\alpha|\chi(\tau)|), & \tau \in \psi((-\infty, 0)), \\ \exp(-\beta|\chi(\tau)|), & \tau \in \psi((0, \infty)). \end{cases}$$

then the error bound for the quadrature rule (3.17) is

$$(3.20) \quad \left| \int_{\Gamma} F(\tau)d\tau - h \sum_{j=-M}^N \frac{F(\tau_j)}{\chi'(\tau_j)} \right| \leq C \left(\frac{e^{-\alpha M h}}{\alpha} + \frac{e^{-\beta N h}}{\beta} \right) + |I_F|.$$

The infinite sum in (3.17) is truncated with the use of (3.19) to arrive at the inequality in (3.20). Making the selections

$$(3.21) \quad h = \sqrt{\frac{\pi d}{\alpha M}},$$

and

$$(3.22) \quad N \equiv \left[\left| \frac{\alpha}{\beta} M + 1 \right| \right],$$

where $[x]$ is the integer part of x , then

$$(3.23) \quad \int_{\Gamma} F(\tau) d\tau = h \sum_{j=-M}^N \frac{F(\tau_j)}{\chi'(\tau_j)} + \mathcal{O}(e^{-(\pi \alpha d M)^{1/2}}).$$

Theorems 3.1 and 3.2 are used to approximate the integrals that arise in the formulation of the discrete systems corresponding to equation (2.1).

The sinc-Galerkin method requires the derivatives of composite sinc functions be evaluated at the nodes. These quantities are denoted by

$$(3.24) \quad \delta_{jk}^{(P)} = h^P \left[\frac{d^P}{d\chi^P} [S(j, h) \circ \chi(\tau)] \right]_{\tau=\tau_k}$$

In particular, the following convenient notation will be useful in formulating the discrete system

$$(3.25) \quad \delta_{jk}^{(0)} = [S(j, h) \circ \chi(\tau)]_{\tau=\tau_k} = \begin{cases} 1, & j = k, \\ 0, & j \neq k, \end{cases}$$

$$(3.26) \quad \delta_{jk}^{(1)} = h \left[\frac{d}{d\chi} [S(j, h) \circ \chi(\tau)] \right]_{\tau=\tau_k} = \begin{cases} 0, & j = k, \\ \frac{(-1)^{k-j}}{k-j}, & j \neq k, \end{cases}$$

$$(3.27) \quad \delta_{jk}^{(2)} = h^2 \left[\frac{d^2}{d\chi^2} [S(j, h) \circ \chi(\tau)] \right]_{\tau=\tau_k} = \begin{cases} \frac{-\pi^2}{3}, & j = k, \\ \frac{-2(-1)^{k-j}}{(k-j)^2}, & j \neq k, \end{cases}$$

In equations (3.25)–(3.27), h is the step size and τ_k is a sinc grid point as in (3.12) or (3.13).

3.2. The Sinc-Galerkin Solution. The Sinc-Galerkin method described in this paper, at one level, simply consists of the assembly of the discrete sinc-Galerkin expansion of a second order boundary-value problem in the spatial domain with a sinc-Galerkin discretization for a first order problem in the temporal domain.

This Sinc method, which was originally derived by Stenger over 20 years ago (see also [18]), is based on *integration by parts*. One of the referees of this article has pointed out that there is another Sinc method that is simpler to apply and equally accurate and which is based on *collocation*. Both the *integration by parts* and *collocation* methods are presented, along with proofs of convergence, in Section 7.2 of [29].

The approximate solution to (2.1)-(2.3) is defined by

$$(3.28) \quad u_{m_x, m_t}(x, t) = \sum_{j=-M_t}^{N_t} \sum_{i=-M_x}^{N_x} u_{ij} S_{ij}(x, t),$$

where $m_x = M_x + N_x + 1$, $m_t = M_t + N_t + 1$. The basis functions $\{S_{ij}(x, t)\}$ for $-M_x \leq i \leq N_x$, $-M_t \leq j \leq N_t$ are given as products of basis functions. In this paper we take

$$(3.29) \quad \begin{aligned} S_{ij}(x, t) &= S_i(x)S_j(t), \\ &= [S(i, h_x) \circ \phi(x)][S(j, h_t) \circ \gamma(t)], \end{aligned}$$

where the conformal map in the spatial domain is given by

$$(3.30) \quad \phi(x) = \ln \frac{x}{1-x},$$

and the conformal map in the temporal domain is given by

$$(3.31) \quad \gamma(t) = \ln(t).$$

The unknown coefficients $\{u_{ij}\}$ in equation (3.28) are determined by orthogonolizing the residual with respect to the functions $\{S_{kl}(x, t)\}$, $-M_x \leq k \leq N_x$, $-M_t \leq l \leq N_t$. This yields the discrete Galerkin system

$$(3.32) \quad \langle Lu_{m_x, m_t} - f, S_{kl} \rangle = 0, \quad -M_x \leq k \leq N_x, \quad -M_t \leq l \leq N_t$$

The inner product is defined by

$$(3.33) \quad \langle f, g \rangle = \int_0^\infty \int_0^1 f(x, t)g(x, t)W(x, t)dxdt,$$

where

$$(3.34) \quad W(x, t) = w(x)v(t) = \left(\frac{\sqrt{\gamma'(t)}}{\sqrt{\phi'(x)}} \right),$$

Here, a product weight function is chosen depending on the boundary conditions, the domain, and the differential equation. In this paper we choose $w(x) = 1/\sqrt{\phi'(x)}$ as the weight in the spatial domain and $v(t) = \sqrt{\gamma'(t)}$ as the weight in the temporal domain. A complete discussion on the choices of the weight functions can be found in [18, 19, 27].

To simplify the notation, we shall drop the subscripts and set $u(x, t) = u_{m_x, m_t}$. The derivatives can be removed from the dependent variable u by integrating by parts, twice in x and once in t , to arrive at the identity

$$(3.35) \quad \begin{aligned} - \int_0^\infty \int_0^1 u(x, t) [S_k(x)S_l(t)w(x)v(t)]_t dxdt \\ - \int_0^\infty \int_0^1 u(x, t) [S_k(x)S_l(t)w(x)v(t)]_{xx} dxdt + B \\ = \int_0^\infty \int_0^1 f(x, t) [S_k(x)S_l(t)w(x)v(t)] dxdt \end{aligned}$$

where

$$(3.36) \quad B = \int_0^1 [u(x, t) S_k(x) S_l(t) w(x) v(t)]_0^\infty dx - \int_0^\infty [u_x(x, t) S_k(x) S_l(t) w(x) v(t)]_0^1 dt + \int_0^\infty [u(x, t) (S_k(x) S_l(t) w(x) v(t))_x]_0^1 dt$$

Setting

$$(3.37) \quad \frac{d^n}{d\chi^n} S(\tau) \equiv S^n(\tau), \quad n = 0, 1, 2,$$

where χ is either ϕ or γ and τ is either x or t , and assume that $B = 0$, then equation (3.35) may be written as

$$(3.38) \quad - \int_0^\infty \int_0^1 u(x, t) S_k(x) w(x) [S'_l(t) \gamma'(t) v(t) + S_l(t) v'(t)] dx dt - \int_0^\infty \int_0^1 u(x, t) S_l(t) v(t) [S''_k(x) (\phi')^2(x) w(x) + S'_k(x) (\phi''(x) w(x) + 2\phi'(x) w'(x)) + S_k(x) w''(x)] dx dt = \int_0^\infty \int_0^1 f(x, t) S_k(x) S_l(t) w(x) v(t) dx dt.$$

Applying the quadrature rule to the iterated integrals, deleting the error terms, replacing $u(x_k, t_l)$ by u_{kl} and dividing by $h_x h_t$, yields the discrete sinc system

$$(3.39) \quad \frac{w(x_k)}{\phi'(x_k)} \sum_{i=-M_t}^{N_t} u_{ki} \left[\frac{1}{h_t} \delta_{ki}^{(1)} v(t_i) + \delta_{ki}^{(0)} \frac{v'(t_i)}{\gamma'(t_i)} \right] + \frac{v(t_l)}{\gamma'(t_l)} \sum_{j=-M_x}^{N_x} \left[\frac{1}{h_x^2} \delta_{kj}^{(2)} + \frac{1}{h_x} \delta_{kj}^{(1)} \left(\frac{\phi''(x_j)}{(\phi')^2(x_j)} + \frac{2w'(x_j)}{\phi'(x_j) w(x_j)} \right) + \delta_{kj}^{(0)} \frac{w''(x_j)}{(\phi')^2(x_j) w(x_j)} \right] w(x_j) \phi'(x_j) u_{jl} = - \frac{w(x_k)}{\phi'(x_k)} F(x_k, t_l) \frac{v(t_l)}{\gamma'(t_l)}.$$

This system is identical to that generated from orthogonalizing the residual via equation (3.32). Recall the notation of Toeplitz matrices [14]. Let $I_{m_x}^{(P)}$, $P = 0, 1, 2$ be the $m_x \times m_x$ matrices $I^{(P)}$, with jk -th entry $\delta_{jk}^{(P)}$ as given by Eqs. (3.25)-(3.27). Further, $D(g_x)$ is an $m_x \times m_x$ diagonal matrix whose diagonal entries are $[g(x_{-M_x}), g(x_{-M_x+1}), \dots, g(x_0), \dots, g(x_{N_x})]^T$, i.e,

$$(3.40) \quad D(g_x) = \begin{pmatrix} g(x_{-M_x}) & & & & \\ & g(x_{-M_x+1}) & & & \\ & & \ddots & & \\ & & & g(x_0) & \\ & & & & \ddots \\ & & & & & g(x_{N_x}) \end{pmatrix}.$$

The matrix $I_m^{(0)}$ is the identity matrix. The matrices $I_m^{(1)}$ and $I_m^{(2)}$ take the form

$$(3.41) \quad I_m^{(1)} = \begin{pmatrix} 0 & -1 & \frac{1}{2} & & \frac{(-1)^{(m-1)}}{m-1} \\ 1 & 0 & & & \\ -\frac{1}{2} & & \dots & & \\ & & \dots & & \frac{1}{2} \\ & & & 0 & -1 \\ \frac{(-1)^{(m)}}{m-1} & & & -\frac{1}{2} & 1 & 0 \end{pmatrix},$$

and

$$(3.42) \quad I_m^{(2)} = \begin{pmatrix} -\frac{\pi^2}{3} & 2 & -\frac{2}{2^2} & \dots & \frac{2(-1)^{(m)}}{(m-1)^2} \\ 2 & -\frac{\pi^2}{3} & & & \dots \\ -\frac{2}{2^2} & & \ddots & & \vdots \\ \vdots & & & \dots & \frac{1}{2} \\ \frac{2(-1)^{(m)}}{(m-1)^2} & \dots & -\frac{2}{2^2} & 2 & -\frac{\pi^2}{3} \end{pmatrix}.$$

The matrices $I_{m_t}^{(P)}$, $P = 0, 1$ and $D(g_t)$ are similarly defined though of size $m_t \times m_t$. Introducing this notation in Eq. (3.39) leads to the matrix form

$$(3.43) \quad \begin{aligned} & \left[\frac{1}{h_x^2} I_{m_x}^{(2)} + \frac{1}{h_x} I_{m_x}^{(1)} D \left(\frac{\phi''}{(\phi')^2} + \frac{2w'}{\phi' w} \right) + D \left(\frac{w''}{(\phi')^2 w} \right) \right] D(\phi') D(w) U D \left(\frac{v}{\gamma'} \right) \\ & + D \left(\frac{w}{\phi'} \right) U D \left(\frac{v}{\sqrt{\gamma'}} \right) D \left(\sqrt{\gamma'} \right) \left[\frac{1}{h_t} I_{m_t}^{(1)} + D \left(\frac{v'}{v \gamma'} \right) \right]^T \\ & = -D \left(\frac{w}{\phi'} \right) F D \left(\frac{v}{\gamma'} \right), \end{aligned}$$

where the matrix F has entries that are the point evaluations of the forcing term, i.e., the ij -th entry of F is $f(x_i, t_j) = f \left(\frac{e^{ith_x}}{e^{ith_x} + 1}, e^{jth_t} \right)$.

The weight function has been chosen so that

$$\frac{\phi''}{(\phi')^2} + \frac{2w'}{\phi' w} = 0.$$

Therefore, premultiplying by $D(\phi')$ and postmultiplying by $D(\sqrt{\gamma'})$ yields the equivalent system

$$(3.44) \quad AV + VB^T = G,$$

where A , B , V and G are matrices of size $m_x \times m_x$, $m_t \times m_t$, $m_x \times m_t$, and $m_x \times m_t$, respectively, and given by

$$(3.45) \quad A = D(\phi') \left[\frac{1}{h_x^2} I_{m_x}^{(2)} + D \left(\frac{w''}{(\phi')^2 w} \right) \right] D(\phi'),$$

$$(3.46) \quad B = D \left(\sqrt{\gamma'} \right) \left[\frac{1}{h_t} I_{m_t}^{(1)} + D \left(\frac{v'}{v \gamma'} \right) \right] D \left(\sqrt{\gamma'} \right),$$

$$(3.47) \quad V = D(w)UD\left(\frac{v}{\sqrt{\gamma'}}\right),$$

and

$$(3.48) \quad G = -D(w)FD\left(\frac{v}{\sqrt{\gamma'}}\right).$$

The matrix A is symmetric and the coefficients in Eq. (3.28) are stored in U . The diagonal matrix $D(\phi')$ is the $m_x \times m_x$ matrix with entries $[\phi'(x_i)]$. The matrix $D(\sqrt{\gamma'})$ is the $m_t \times m_t$ matrix with entries $[\sqrt{\gamma'(t_j)}]$. To obtain the approximate solution Eq. (3.28), we need to solve the system for U which requires solving (3.44) for V . In the next section we show how to solve (3.44) for V .

3.3. Solution of the Discrete System.

The special system

$$(3.49) \quad AX + XB = C,$$

which often arises in numerical partial differential equations may be solved by concatenating each side of the equation. The following definitions are necessary for this step

DEFINITION 3.2. For a matrix $B = (b_{ij})$, $1 \leq j \leq m$, $1 \leq i \leq n$, the concatenation of B is the $m n \times 1$ vector

$$(3.50) \quad \text{vec}(B) = \begin{pmatrix} b_{i1} \\ b_{i2} \\ \vdots \\ b_{in} \end{pmatrix},$$

where b_{ik} is the k th column of B .

DEFINITION 3.3. Let $A = [a_{ij}]$ be a matrix of order $(m \times n)$ and $B = [b_{ij}]$ be a matrix of order $(p \times q)$. The Kronecker or tensor product of the two matrices, denoted by $A \otimes B$, is a matrix of order $(m p \times n q)$ defined as

$$(3.51) \quad A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{pmatrix}.$$

A useful property of concatenation is given in Theorem 3.3.

THEOREM 3.3. Let $A, B, X = [x_{ij}]$ and $C = [c_{ij}]$ be $m \times m$, $n \times n$, $m \times n$ and $m \times n$ matrices, respectively. Define x and c by

$$(3.52) \quad x = \text{vec}(X) = (x_{11}, x_{21}, \dots, x_{m1}, x_{12}, x_{22}, \dots, x_{m2}, x_{13}, \dots, x_{mn})^T,$$

and

$$(3.53) \quad c = \text{vec}(C) = (c_{11}, c_{21}, \dots, c_{m1}, c_{12}, c_{22}, \dots, c_{m2}, c_{13}, \dots, c_{mn})^T,$$

Then the system of equations

$$(3.54) \quad AXB = C,$$

can also written in the equivalent matrix-vector form:

$$(3.55) \quad (B^T \otimes A)x = c.$$

Moreover, if we denote the eigenvalues of A and B respectively by α_i and β_j , then the eigenvalues of $A \otimes B$ are $\alpha_i\beta_j$. Further $I_n \otimes A + B \otimes I_m$ has eigenvalues $\alpha_i + \beta_j$.

The proof of this theorem is given in [23]. This technique can be applied to equation (3.49). First, we rewrite (3.49) in the form

$$(3.56) \quad (I_n \otimes A + B^T \otimes I_m) \text{vec}(X) = \text{vec}(C),$$

which has the unique solution

$$(3.57) \quad \text{vec}(X) = (I_n \otimes A + B^T \otimes I_m)^{(-1)} \text{vec}(C),$$

if and only if

$$(3.58) \quad (I_n \otimes A + B^T \otimes I_m).$$

is nonsingular. The solution X can be derived using definition 3.2

Therefore, it is easy to see that a necessary and sufficient condition that equation (3.49) has a unique solution for all C is that $\alpha_i + \beta_j \neq 0$ where α_i are the eigenvalues of A and β_j are the eigenvalues of B ; see [5]. In the case $C = 0$, a necessary and sufficient condition for the equation to have a nontrivial solution is $\alpha_i + \beta_j = 0$ for some i and j .

In particular, if A, B, V and G are matrices of size $m_x \times m_x, m_t \times m_t, m_x \times m_t$, and $m_x \times m_t$, respectively, then the system $AV + VB^T = G$ can be written in the form

$$(3.59) \quad (I_{m_t} \otimes A + B \otimes I_{m_x}) \text{vec}(V) = \text{vec}(G),$$

with I_{m_x} and I_{m_t} unit matrices of order m_x and m_t respectively, $\text{vec}(V)$ and $\text{vec}(G)$ are concatenations of the columns of V and G respectively. Once V has been obtained, we can get U from V via Eq. (3.47).

4. Numerical Examples

The examples reported in this section were selected from a large collection of problems to which the Sinc-Galerkin and Wavelet-Galerkin methods could be applied. For purposes of comparison, contrast and performance, examples with known solutions were chosen.

We give three examples, in two of them the non-homogeneous term, $f(x, t)$, has singularities. These examples demonstrate how the Sinc-Galerkin method outperforms the wavelet-Galerkin method, especially when singularities are present. This may be explained in view of the fact that in order to obtain a good convergence rate using the wavelet-Galerkin method, the non-homogeneous term has to be smooth, e.g., $f(x, t)$, has to be in some Sobolev space.

For the Sinc-Galerkin, d is taken to be $\pi/2$. The step sizes h_x and h_t and the summation limits M_x, N_x, M_t and N_t are selected so that the error in each coordinate direction is asymptotically balanced. Once M_x is chosen, the step sizes and remaining summation limits can be determined as follows

$$(4.1) \quad h_x = h_t = \sqrt{\frac{\pi d}{\alpha M_x}}$$

$$(4.2) \quad N_x = \left\lceil \left| \frac{\alpha M_x}{\beta} \right| \right\rceil,$$

and

$$(4.3) \quad M_t = M_x,$$

and N_t is also arbitrary. Each example was run for a sequence of M_x with $h_x = h_t = h$ and all other parameters selected as in Eqs (4.1) -(4.3). In all examples, a sequence of runs for $M_x = 4, 6, 8, 12, 14$ and 16 is reported. In all cases reported $M_t > N_t$ which yields a much smaller discrete system than the choice $M_t = N_t$ that may give results in larger matrices with no corresponding increase in accuracy.

For the Wavelet-Galerkin solution, each example was run for a sequence of $j = 2, 3, 4, 5, 6, 7$, but we found the difference between the results at level $3 \leq j \leq 7$ to be small.

Example 4.1: Consider the problem

$$(4.4) \quad u_t - u_{xx} = [(x - x^2)(1 - t) + 2t] e^{-t}, \quad 0 \leq x \leq 1, t > 0,$$

subject to the boundary conditions

$$(4.5) \quad u(0, t) = 0, \quad u(1, t) = 0,$$

and the initial condition

$$(4.6) \quad u(x, 0) = 0, \quad 0 \leq x \leq 1.$$

The solution is

$$(4.7) \quad u(x, t) = x(1 - x)te^{-t}$$

The parameter $M_x = N_x = M_t = 14$ and $N_t = 2$, and also $\alpha = \beta = \frac{1}{2}$ are used for the Sinc-Galerkin solution, while for the wavelet-Galerkin method we used $j = 3$. The three solutions are given in **Table 4.1** at time $t = 0.01$.

Table 4.1

x	Exact Solution	Sinc-Galerkin	Wavelet-Galerkin
0.0	0.0	0.0	0.0
$1/2^4$	0.00058	0.00058	0.00051
$2/2^4$	0.00108	0.00108	0.00096
$3/2^4$	0.00150	0.00151	0.00133
$4/2^4$	0.00185	0.00186	0.00166
$5/2^4$	0.00212	0.00213	0.00190
$6/2^4$	0.00232	0.00232	0.00208
$7/2^4$	0.00243	0.00244	0.00218
$8/2^4$	0.00247	0.00248	0.00222
$9/2^4$	0.00243	0.00244	0.00218
$10/2^4$	0.00232	0.00232	0.00208
$11/2^4$	0.00212	0.00213	0.00190
$12/2^4$	0.00185	0.00186	0.00166
$13/2^4$	0.00150	0.00151	0.00136
$14/2^4$	0.00108	0.00108	0.00099
$15/2^4$	0.00058	0.00058	0.00052
1.0	0.00	0.00	0.0

Example 4.2: Consider the problem

$$(4.8) \quad u_t - u_{xx} = f(x, t), \quad 0 \leq x \leq 1, t > 0,$$

where

$$(4.9) \quad f(x, t) = \frac{x(1-x)^2(1-t^2) + \frac{1}{4}(4-3x)(t+t^3)}{(1+t^2)^2(1-x)^{3/2}}$$

subject to the boundary conditions

$$(4.10) \quad u(0, t) = 0, \quad u(1, t) = 0,$$

and the initial condition

$$(4.11) \quad u(x, 0) = 0, \quad 0 \leq x \leq 1.$$

The solution is

$$(4.12) \quad u(x, t) = \frac{tx(1-x)^{1/2}}{1+t^2}$$

In this problem the function $f(x, t)$ has a singularity at $x = 1$. The parameter $M_x = M_t = N_x = 32$ and $N_t = 10$, and also $\alpha = \beta = 1/2$ are used for the Sinc-Galerkin solution, while for the wavelet-Galerkin method we used $j = 3$. Note that the wavelet-Galerkin solution requires that the function $f(x, t)$ be approximated by a polynomial in x . The solution at time $t = 0.01$ is given in **Table 4.2**.

Table 4.2

x	Exact Solution	Sinc-Galerkin	Wavelet-Galerkin
0.0	0.0	0.0	0.0
$1/2^4$	0.00060	0.00060	0.00054
$2/2^4$	0.00116	0.00117	0.00104
$3/2^4$	0.00168	0.00169	0.00149
$4/2^4$	0.00216	0.00217	0.00194
$5/2^4$	0.00259	0.00260	0.00230
$6/2^4$	0.00296	0.00297	0.00263
$7/2^4$	0.00328	0.00329	0.00285
$8/2^4$	0.00353	0.00355	0.00301
$9/2^4$	0.00372	0.00373	0.00292
$10/2^4$	0.00382	0.00384	0.00274
$11/2^4$	0.00384	0.00386	0.00220
$12/2^4$	0.00374	0.00376	0.00149
$13/2^4$	0.00351	0.00353	0.00077
$14/2^4$	0.00309	0.00311	0.00013
$15/2^4$	0.00234	0.00235	0.00007
1.0	0.00	0.00	0.0

Example 4.3: The problem

$$(4.13) \quad u_t - u_{xx} = f(x, t) \quad 0 \leq x \leq 1, t > 0,$$

where

$$f(x, t) = t^{1/2} e^{-t} (1-x)^{-1/2} \left[\left(\frac{3}{2} - t \right) x (1-x)^2 + \frac{1}{4} t (12 - 15x) \right],$$

subject to the boundary conditions

$$(4.14) \quad u(0, t) = 0, \quad u(1, t) = 0,$$

and the initial condition

$$(4.15) \quad u(x, 0) = 0, \quad 0 \leq x \leq 1.$$

has the solution

$$(4.16) \quad u(x, t) = t^{3/2} e^{-t} x (1 - x)^{3/2}$$

In this problem the function $f(x, t)$ has a singularity at $x = 1$. The parameter $M_x = M_t = N_x = 32$ and $N_t = 10$, and also $\alpha = \beta = 3/2$ are used for the Sinc-Galerkin solution, while for the wavelet-Galerkin method we used $j = 3$. Note that the wavelet-Galerkin solution requires that the function $f(x, t)$ be approximated by a polynomial in x . The solutions at time $t = 0.01$ are given in **Table 4.3**.

Table 4.3

x	Exact Solution 1.0e-03	Sinc-Galerkin 1.0e-03	Wavelet-Galerkin 1.0e-03
0.0	0.0	0.0	0.0
$1/2^4$	0.05616	0.05616	0.04373
$2/2^4$	0.10129	0.10129	0.07857
$3/2^4$	0.13595	0.13596	0.10492
$4/2^4$	0.16076	0.16077	0.12498
$5/2^4$	0.17636	0.17637	0.13776
$6/2^4$	0.18344	0.18345	0.14318
$7/2^4$	0.18273	0.18274	0.14438
$8/2^4$	0.17501	0.17502	0.13869
$9/2^4$	0.16115	0.16116	0.13072
$10/2^4$	0.14209	0.14210	0.11716
$11/2^4$	0.11890	0.11891	0.10315
$12/2^4$	0.09281	0.09282	0.08505
$13/2^4$	0.06531	0.06531	0.06742
$14/2^4$	0.03828	0.03828	0.04722
$15/2^4$	0.01450	0.01450	0.02451
1.0	0.00	0.00	0.0

5. Conclusions

Although the wavelet-Galerkin method has been shown to be a powerful numerical tool for fast and accurate solutions of partial differential equations, in many instances the sinc-Galerkin method seems to give better results, especially when condition (3.19) is satisfied as demonstrated by Examples 1 and 3. In general, the sinc-Galerkin method seems to give very good results in solving heat problems, even non-linear ones [17]; see also [8].

We obtained solutions of nonhomogeneous heat equations using Daubechies 6 wavelets and then compared them with those obtained using the sinc-Galerkin method. Although the sinc-Galerkin solution required slightly more computational effort than the wavelet-Galerkin solution, it resulted in more accurate results, especially in the presence of singularities. This may be attributed to the fact that in the wavelet method in order to obtain an error of order $2^{-\alpha j N}$ at level j for some α , the solution $u(x, t)$ and the functions $f(x, t)$ have to be smooth.

References

- [1] K. Amaratunga, et al., Wavelet-Galerkin solutions for one-dimensional partial differential equations, *Int. J. Numer. Meth. Engng.*, Vol. 37 (1994), 2705-2716.
- [2] C. Anne and et al., Convergence of the sinc overlapping domain decomposition method, *Appl. Math. Comput.*, Vol. 98 (1999), 209-227.
- [3] C. Anne and K. Bowers, The schwarz alternating sinc domain decomposition method, *App. Numer. Math.*, Vol. 25 (1997), 461-483.
- [4] A. Avudainayagam, Wavelet-Galerkin method for integro-differential equations, *Appl. Numer. Math.*, Vol. 32 (2000), 247-254.
- [5] N. Bellman, *Introduction to Matrix Analysis*, Mc Graw-Hill, Inc., New York, 1960.
- [6] N. Bellomo, Solution of nonlinear initial boundary value problems by sinc collocation interpolation methods, *Comput. Math. Applic.*, Vol. 29 (1995), 15-28.
- [7] G. Beylkin, On the representation of operators in bases of compactly supported wavelets, *SIAM J. Numer. Anal.*, Vol. 29 (1992), 1716-17409.
- [8] K. Bowers, T. Carlson, and J. Lund, Advection-diffusion equations:Temporal Sinc methods, *Num. Methods for Part. Diff. Eqns*, Vol. 11 (1995), 399-422.
- [9] I. Daubechies, Orthonormal bases of compactly supported wavelets, *Commun. Pure Appl. Math.*, Vol. 41 (1988), 909-996.
- [10] I. Daubechies, *Ten Lectures on Wavelets*, Captial City Press, Vermont, 1992.
- [11] K. El-Kamel, Sinc numerical solution for solitons and solitary waves, *J. Comput. Appl. Math.*, Vol. 130 (2001), 283-292.
- [12] R. Glowiniski, W. M. Lawton, M. Ravachol and E. Tenenbaum, Wavelet solutions of linear and nonlinear elliptic, parabolic and hyperbolic problems in one space dimension, *Comput. Meth. Appl. Sci. Eng.*, Chapter 4, (1990), 55-120.
- [13] G. H. Golub and C. F. Vanloan, *Matrix Computations*, Third Ed., The Johns Hopkins Press Ltd., London, 1996.
- [14] V. Grenander and G. Szego, *Toeplitz Forms and Their Applications*, Second Ed., Chelsea Publishing Co., Orlando, 1985.
- [15] F. Jin and T. Q. Ye, Instability analysis of prismatic members by wavelet-Galerkin method, *Advances in Engineering Software*, Vol. 30 (1999), 361-367.
- [16] A. Latto, H. L. Resniko and E. Tenenbaum, The evaluation of connection coefficients of compactly supported wavelets, *Proc. French-U.S.A. Workshop on Wavelets and Turbulence*, Princeton Univ., June 1991, Springer, New York, 1992.
- [17] A. Lippke, Analytic Solution and Sinc Function Approximation in Thermal Conduction with Nonlinear Heat Generation, *J. Heat Transfer (Transactions of the ASME)*, v. 113 (1991) 5-11.
- [18] J. Lund and K. Bowers, *Sinc Methods for Quadrature and Differential Equations*, SIAM, Philadelphia, PA, 1992.
- [19] J. Lund, Symmetrization of the Sinc-Galerkin method for boundary value problems ,*Math. Comp.*, Vol. 47 (1986), 571-588.
- [20] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, New York, 1999.
- [21] K. Michael, Fast iterative methods for symmetric Sinc-Galerkin system, *IMA J. Numer. Anal.*, Vol. 19 (1999), 357-373.
- [22] C. Ming-quayer and H. Chyi, The computation of wavelet-Galerkin approximation on a bounded interval, *Int. J. Numer. Meth. Eng.*, Vol. 39 (1996), 2921-2944.
- [23] H. Neudecker, A note on kronecker matrix products and matrix equation systems, *SIAM J. Appl. Math.*, Vol. 17 (1969), 603-606.
- [24] E. Part-Enander, A. Sjoberg, B. Melin and P. Isaksson, *The Matlab Handbook*, Addison Wesley Longman, 1996.
- [25] S. Qian and J. Weiss, Wavelets and the numerical solution of boundary value problems, *Appl. Math. Lett.*, Vol. 6 (1993), 47-52.
- [26] S. Qian and J. Weiss, Wavelets and the numerical solution of partial differential equations, *J. Comp. Phys.*, Vol. 106 (1993), 155-175.
- [27] C. Ralph and K. Bowers, The Sinc-Galerkin method for fourth-order differential equations, *SIAM J. Numer. Anal.*, Vol. 28 (1991), 760-788.
- [28] F. Stenger, Matrices of sinc methods, *J. Comput. Appl. Math.*, Vol. 86 (1997), 297-310.

- [29] F. Stenger, *Numerical Methods Based on Sinc and Analytic Functions*, Springer, New York, 1993.
- [30] F. Stenger, B. Barkey & R. Vakili, Sinc Convolution Method of Solution of Burgers' Equation, pp. 341–354 of *Proceedings of Computation and Control III*, edited by K. Bowers and J. Lund, Birkhäuser (1993).
- [31] F. Stenger, Numerical methods based on Whittaker cardinal or sinc functions, *SIAM Rev.*, Vol. 23 (1981), 165-223.
- [32] A.I. Zayed, *Advances in Shannon's Sampling Theory*, CRS Press, Boca Raton, 1993.

DEPARTMENT OF MATHEMATICAL SCIENCES, FACULTY OF ENGINEERING, MANSOURA UNIVERSITY, EGYPT

E-mail address: `gamel_eg@yahoo.com`

DEPARTMENT OF MATHEMATICAL SCIENCES, DEPAUL UNIVERSITY, CHICAGO, IL 60614, U.S.A
E-mail address: `azayed@math.depaul.edu`

Fast diffusion registration

Bernd Fischer and Jan Modersitzki

This paper is dedicated to the 65th birthday anniversary of Gerhard Opfer.

ABSTRACT. Image registration is one of the most challenging tasks within digital imaging, in particular in medical imaging. Typically, the underlying problems are high dimensional and demand for fast and efficient numerical schemes.

Here, we propose a novel scheme for automatic image registration by introducing a specific regularizing term. The new scheme is called *diffusion registration* since its implementation is based on the solution of a diffusion type partial differential equation (PDE). The main ingredient for a fast implementation of the diffusion registration is the so-called additive operator splitting (AOS) scheme. The AOS-scheme is known to be as accurate as a conventional semi-implicit scheme and has a linear complexity with respect to the size of the images. We present a proof of these properties based purely on matrix analysis.

The performance of the new scheme is demonstrated for a typical medical registration problem. It is worth noticing that the diffusion registration is extremely well-suited for a parallel implementation.

Finally, we also draw a connection to Thirion's demon based approach.

1. Introduction

Image registration is an important problem in computer vision and in particular in medical imaging. The problem arises when images of objects which are taken, for example, at different times, from different perspectives, or with different imaging devices, respectively, are to be studied. Typically, the geometry of the objects changes considerably and the images can not be related directly. To illustrate this problem, consider the images in Figure 2, which are pictures of two consecutive tissue sections of a human brain. The overall goal in this particular application is the reconstruction of the whole human brain out of a series of approximately 7.000 such sections. Due to the sectioning processes nonlinear distortions of the tissue are introduced, as it is apparent from Figure 2. Here, the aim of the registration is to recover the original anatomical structure of the brain.

2000 *Mathematics Subject Classification.* 68U10, 65F22, 65F05.

Key words and phrases. Image registration, matching, AOS, partial differential equation.

© 2002 American Mathematical Society

In the last decade a number of non-rigid, automatic registration algorithms have been proposed, see, for example, [Bro81, BK89, Ami94, Chr94, BN96] and references therein. Most of these schemes may be viewed as a procedure which minimizes a suitable distance measure subject to a regularization term. There are essentially two approaches to solve these optimization problems numerically. One is to deal directly with the original formulation, whereas the other is to solve a related partial differential equation (PDE). Here, we focus on the latter method. Typical members out of this class are the elastic [Chr94] and fluid [BN96] deformation models. The main bottleneck of these approaches is thought of to be the solution of the corresponding linear systems. However, Fischer and Modersitzki [FM99] recently showed that one may solve these systems in just $\mathcal{O}(n \log n)$ operations, where n is the number of pixel.

Here, we propose a novel gradient based regularization term and devise a fast and stable implementation for a finite difference approximation of the underlying partial differential equation. Since this PDE may be viewed as a generalized diffusion equation, we call our new scheme *diffusion registration*. Actually, we show that the solution of the corresponding linear system requires only $\mathcal{O}(n)$ operations, that is, its complexity is linear with respect to the number of pixel. The main tool is the so-called additive operator splitting (AOS) scheme, see Weickert [Wei98]. The idea is to split the original problem into a number of simpler problems, which allow for a fast numerical solution. Here, we give a proof for the accuracy of the AOS-scheme. The proof is based purely on matrix analysis and therefore the result applies as well to more general situations.

Beside this, we discuss Thirion's [Thi98] demon based approach. Thirion proposed a method which works well in practice but its derivation is guided by intuition and not entirely understood. In the literature one may find several attempts to shed some light on his approach (see, e.g., [PCA99, BNG96]). Because Thirion offers a variety of possible implementations, the underlying theory is widespread. However, the bottom line is, that he calculates the deformations by regularizing certain driving forces by a Gaussian convolution filter. We show that this technique may be viewed as a special (low order) approximation to the partial differential equation connected to our new scheme and thereby gaining some insight into Thirion's approach.

The paper is organized as follows. In Section 2, we introduce a mathematical formulation of the general registration problem and define our new regularization term. Section 3 is concerned with the formulation of the problem in terms of a system of parabolic nonlinear partial differential equations. The fast and efficient numerical treatment of these equations is topic of Section 4. In particular we introduce the AOS-scheme and discuss its properties. The connections to Thirion's work are drawn in Section 5. Finally, in Section 6 we present a numerical example and give some performance measurements obtained for the registration of tissue sections.

2. Diffusion based registration

We refer to the *template image* as T and the *study image* as S where $T, S : \Omega = [0, 1]^d \rightarrow \mathbb{R}$. The registration algorithm described in this paper is applicable to images with arbitrary dimension d . For a particular point $x \in \Omega$, the value $T(x)$

denotes the intensity at x . The purpose of the registration is to determine a transformation, sometimes called warping, of T onto S . Ideally, one wants to determine a displacement field $u : \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that $T(x - u(x)) = S(x)$. The question is how to find a suitable mapping u . A straightforward approach would be to minimize the following distance measure

$$(2.1) \quad \mathcal{D}[u] := \frac{1}{2} \|T(\cdot - u) - S\|_{L_2} := \frac{1}{2} \int_{\Omega} (T(x - u(x)) - S(x))^2 dx.$$

Here, it is assumed that the images T and S belong to $L_2(\Omega)$. Of course other functionals, including the so-called *mutual information* based measure (see, e.g., Viola [Vio95]) or the *normalized cross-correlation* (compare, e.g., [GW93]), respectively, might be used as well. The theory is along the same lines, as long as the resulting functional does have a Gateaux derivative. The choice of the distance measure depends on the particular application. For the application discussed in Section 6 the L_2 based measure (2.1) turns out to be appropriate.

A straightforward approach to attack the above minimization problem would be to employ a gradient based strategy. In general, however, this method might lead to discontinuous and/or suboptimal displacement fields as a result of the ill-posedness of the problem. A standard technique to overcome these problems is to introduce a *smoothing* or *regularizing* term \mathcal{S} . A proper choice of \mathcal{S} may be used as well to privilege likely solutions. It should be pointed out that in many applications the focus is more on a smooth, non-oscillatory solution rather than a perfect match between the images. A striking example in this direction is our application, the reconstruction of a human brain from a histological serial sectioning, cf. Section 6. With an additional regularization parameter α , the problem now reads, find a displacement field u which minimizes the joint criterion

$$(2.2) \quad \mathcal{J}[u] := \mathcal{D}[u] + \alpha \mathcal{S}[u].$$

For example, the well-known deformable grid methods based on elasticity or fluid mechanics may be phrased in terms of minimizing the functional (2.2) for specific choices of \mathcal{S} . We note that the parameter α determines the relative weight of the regularizing term.

In this note we investigate the smoother \mathcal{S} , defined by

$$(2.3) \quad \mathcal{S}[u] := \frac{1}{2} \sum_{j=1}^d \|\nabla u_j\|_{L_2}^2.$$

The reason for this particular choice is three-fold. It is designed to penalize oscillating deformations and consequently leads to smooth displacement fields. As we will see in Section 3, it permits a fast and efficient implementation which enables one to apply this technique even to the registration of very high dimensional image data. Finally, as a byproduct, it has a strong connection to the approach of Thirion and thereby giving a new mathematical interpretation of this popular method.

It is worth noticing that in our application the new approach does not only produce registrations in a reasonable amount of time but also the results resemble the anatomical structure of the tissue.

3. Euler-Lagrange equations and time marching

In accordance with the calculus of variations, the function u which minimizes the functional (2.2) with respect to (2.3) has to satisfy the Euler-Lagrange equations

$$(3.1) \quad f(x, u(x)) + \alpha \Delta u(x) = 0, \quad x \in \Omega,$$

subject to appropriate boundary conditions. Here, Δ denotes the Laplace operator with $\Delta u = (\Delta u_1, \dots, \Delta u_d)^T$. The so-called *force field*

$$(3.2) \quad f(x, u(x)) := (T(x - u(x)) - S(x)) \cdot \nabla(T(x - u(x)))$$

is used to drive the deformation. It is worth noticing that f is the derivative of the functional \mathcal{D} with respect to u . Changing the distance measure \mathcal{D} results in a different force field (see the comment after (2.1)).

A popular approach to solve a non-linear partial differential equation like (3.1) is to introduce an artificial time t and to compute the steady state solution $\partial_t u(x, t) = 0$ of the time dependent partial differential equation

$$(3.3) \quad \partial_t u(x, t) = f(x, u(x, t)) + \alpha \Delta u(x, t), \quad x \in \Omega, t \geq 0,$$

via a time marching algorithm. To overcome the nonlinearity in f , we employ the following semi-implicit iterative scheme

$$(3.4) \quad \partial_t u(x, t_{k+1}) - \alpha \Delta u(x, t_{k+1}) = f(x, u(x, t_k)), \quad k = 0, 1, \dots,$$

where $u(x, t_0)$ is some initial deformation, typically $u(x, t_0) = 0$. In other words, the trick is to compute the driving force f for the previous solution $u(x, t_k)$ and subsequently to solve for $u(x, t_{k+1})$.

An important property of the system of equations (3.4) is that they are essentially decoupled. The coupling is only through the right hand side. The j th equation looks like

$$(3.5) \quad \partial_t u_j(x, t_{k+1}) - \alpha \Delta u_j(x, t_{k+1}) = f_j(x, u(x, t_k)), \quad k = 0, 1, \dots$$

Note that equation (3.5) is nothing but an inhomogeneous heat-equation and well understood (see, e.g., Folland [Fol95]).

4. AOS-scheme

There exist a whole bunch of schemes to solve (3.5) numerically. Here, we are interested in schemes which are on the one side accurate and stable and on the other side fast and efficient.

As a representative example we illustrate these points by considering a finite difference discretization. To set up the notation, we start by introducing the standard semi-implicit scheme for a discretized version of (3.5). Subsequently, we discuss the AOS-scheme. Due to the simplicity of the underlying region $\Omega = [0, 1]^d$, we have chosen a standard finite difference method with a canonical $2d + 1$ stencil. For the time discretization we introduce a time-step $\tau > 0$ and for the spatial discretization a grid \vec{X} , respectively. For ease of presentation, we use a lexicographical ordering and consider $\vec{X} = (\vec{X}_1, \dots, \vec{X}_d) \in \mathbb{R}^{n \times d}$, where n is the number of gridpoints and \vec{X}_j collects the j th component of the gridpoints. For a fixed j , let $\vec{V}^k := u_j(\vec{X}, k\tau)$ and $\vec{F}_j^k := f_j(\vec{X}, u(\vec{X}, \tau k))$. The discrete version of equation (3.5) then reads

Also F^k belongs to R^n :

$$(4.1) \quad \frac{\vec{V}^{k+1} - \vec{V}^k}{\tau} - \sum_{\ell=1}^d A_\ell \vec{V}^{k+1} = \vec{F}_j^k.$$

V^k belongs to R^n : at each time step, evaluate u_j in all n grid points
--> it should be V^k

Here,

$$\frac{\vec{V}^k - \vec{V}^{k-1}}{\tau} \approx \partial_t u_j(\vec{X}, k\tau)$$

is a forward difference approximation of the time derivative $\partial_t u_j$ with time-step τ and $A_\ell \in \mathbb{R}^{n \times n}$ is an appropriate finite difference approximation of the second order derivative of u_j with respect to the ℓ th space coordinate,

$$(4.2) \quad A_\ell \vec{V}^k \approx \alpha \partial_{\ell\ell} u_j(\vec{X}, \tau k).$$

For the numerical example presented in Section 6, we have chosen a simple five-point star leading to an essentially tridiagonal matrix A_ℓ .

After rearranging Equation (4.1), we obtain

$$(4.3) \quad \vec{V}^{k+1} = \left(I - \tau \sum_{\ell=1}^d A_\ell \right)^{-1} \left(\vec{V}^k + \tau \vec{F}_j^k \right), \quad k = 0, 1, \dots$$

Equation (4.3) is the semi-implicit scheme for (3.5) and it is known, that this scheme is of order one with respect to the time-step τ and of order two with respect to the spatial meshsize.

The iteration (4.3) requires the solution of a linear system with n unknowns at each time-step. Note that the systems connected to the individual matrices A_ℓ are essentially tridiagonal and may be solved by an $\mathcal{O}(n)$ direct scheme. On the other hand, however, the sum is not tridiagonal and therefore the system in (4.3) does not permit such a fast implementation, in general.

The idea of AOS is to replace the inverse of the sum by a sum of inverses. The corresponding iterates are defined by

$$(4.4) \quad \vec{V}_A^{k+1} := \frac{1}{d} \sum_{\ell=1}^d (I - d\tau A_\ell)^{-1} \left(\vec{V}_A^k + \tau \vec{F}_j^k \right) \quad k = 0, 1, \dots$$

This clever decomposition allows an $\mathcal{O}(n)$ implementation by employing the Thomas-algorithm [Tho49]. In Weickert [Wei98] it is stated that the order of the local truncation error for both schemes is the same. However, no proof is given.

The next theorem relates the iteration matrices of the semi-implicit and the AOS-scheme to each other. It turns out that the distance between these two matrices is surprisingly small. As our result is based on matrix analysis, it is not restricted to matrices stemming from PDE-discretizations and is therefore of interest in its own. The preceding theorem is formulated for general matrices.

THEOREM 4.1. *Let $d \in \mathbb{N}$, $\tau \geq 0$, and let $A_1, \dots, A_d \in \mathbb{R}^{n \times n}$ be simultaneously diagonalizable with eigenvalues in the left half plane.*

Then there exists a constant $C \in \mathbb{R}$ with

$$\left\| \left(I - \tau \sum_{\ell=1}^d A_\ell \right)^{-1} - \frac{1}{d} \sum_{\ell=1}^d (I - d\tau A_\ell)^{-1} \right\|_2 \leq C \cdot \tau^2.$$

PROOF. The idea is to employ a basis of eigenvectors such that the corresponding matrices become diagonal. To this end consider

$$WA_\ell W^{-1} = \Lambda_\ell = \text{diag}(\lambda_{\ell,j}, 1 \leq j \leq n),$$

where W is an eigenvector matrix of any A_ℓ and the Λ_ℓ 's are the diagonal matrices based on the individual eigenvalues. Hence,

$$\begin{aligned} W(I - d\tau A_\ell)W^{-1} &= I - d\tau \Lambda_\ell, \\ W(I - \tau \sum_{\ell=1}^d A_\ell)W^{-1} &= (I - \tau \sum_{\ell=1}^d \Lambda_\ell), \end{aligned}$$

and

$$\begin{aligned} W[(I - \tau \sum_{\ell=1}^d A_\ell)^{-1} - \frac{1}{d} \sum_{\ell=1}^d (I - d\tau A_\ell)^{-1}]W^{-1} \\ = (I - \tau \sum_{\ell=1}^d \Lambda_\ell)^{-1} - \frac{1}{d} \sum_{\ell=1}^d (I - d\tau \Lambda_\ell)^{-1} \end{aligned}$$

is a diagonal matrix, where the k th diagonal entry is given by

$$q_k = \phi\left(\tau \sum_{\ell=1}^d \lambda_{\ell,k}\right) - \frac{1}{d} \sum_{\ell=1}^d \phi(d\tau \lambda_{\ell,k}), \quad \phi(x) = \frac{1}{1-x}.$$

A Taylor-expansion of the analytic function ϕ at $x_0 = 0$ reads

$$\phi(x) = 1 + x + \frac{2x^2}{(1-\xi)^3}, \quad \xi = \xi(x) \in [0, x].$$

This yields

$$\begin{aligned} q_k &= 1 + \tau \sum_{\ell=1}^d \lambda_{\ell,k} + \frac{2\tau^2 (\sum_{\ell=1}^d \lambda_{\ell,k})^2}{(1-\xi)^3} \\ &\quad - \frac{1}{d} \sum_{\ell=1}^d \left(1 + d\tau \lambda_{\ell,k} + \frac{2(d\tau \lambda_{\ell,k})^2}{(1-\xi_\ell)^3} \right) \\ &= \tau^2 \left(\frac{2(\sum_{\ell=1}^d \lambda_{\ell,k})^2}{(1-\xi)^3} - \frac{1}{d} \sum_{\ell=1}^d \frac{2(d\lambda_{\ell,k})^2}{(1-\xi_\ell)^3} \right) \\ &=: \tau^2 g(\lambda_{1,k}, \dots, \lambda_{d,k}). \end{aligned}$$

By assumption we can find compact sets Q_ℓ contained in the left complex half plane which enclose all eigenvalues of A_ℓ . Consequently, the function g is continuous on $Q := Q_1 \times \dots \times Q_d$ and attains its maximum

$$\tilde{C} := \max\{|g(z)| : z \in Q\}.$$

We thus have $|q_k| \leq \tilde{C}\tau^2$, for $k = 1, \dots, n$, and the statement follows from

$$\begin{aligned} &\|(I - \tau \sum_{\ell=1}^d A_\ell)^{-1} - \frac{1}{d} \sum_{\ell=1}^d (I - d\tau A_\ell)^{-1}\|_2 \\ &\leq \|W\|_2 \|W^{-1}\|_2 \cdot \|(I - \tau \sum_{\ell=1}^d \Lambda_\ell)^{-1} - \frac{1}{d} \sum_{\ell=1}^d (I - d\tau \Lambda_\ell)^{-1}\|_2 \leq C \cdot \tau^2. \end{aligned}$$

□

It is worth noticing that matrices are simultaneously diagonalizable if and only if they commute with each other, a proof of which can be found in Horn & Johnson [HJ90, Theorem 1.3.19]. It is this property which provides a convenient tool for checking the assumption of the above theorem.

In the statement of the theorem it is assumed that all eigenvalues are contained in the left half plane. This assumption ensures that, independent of the value of τ , the matrices

$$I - \tau \sum_{\ell=1}^d A_\ell \quad \text{and} \quad I - d\tau A_\ell$$

are nonsingular. A close inspection of the proof shows that the theorem holds for arbitrary eigenvalues as well, as long as τ is small enough.

Now lets return to the PDE-discretization (4.3) and (4.4), respectively. The next corollary relates the solutions of the two time marching processes to each other. In accordance with the theorem, the statement of the corollary is valid for general matrices.

COROLLARY 4.2. *Let $d, K \in \mathbb{N}$, $\tau \geq 0$, and let $A_1, \dots, A_d \in \mathbb{R}^{n \times n}$ be simultaneously diagonalizable with eigenvalues in the left half plane. Moreover, let \vec{V}^{k+1} and \vec{V}_A^{k+1} denote the solution of (4.3) and (4.4), respectively. Then there exists a constant $C > 0$ with*

$$\|\vec{V}^{k+1} - \vec{V}_A^{k+1}\|_2 \leq C \cdot \tau^2, \quad 0 \leq k \leq K.$$

The particular matrices introduced by (4.2) are given by

$$A_\ell = I \otimes \cdots I \otimes B_\ell \otimes I \otimes \cdots \otimes I,$$

where the ℓ th factor B_ℓ is an approximation of the second order derivative in only one spatial direction and \otimes denotes the Kronecker product of matrices. More precisely, we have

$$B_\ell = \begin{pmatrix} \alpha_{\ell,1} & \beta_{\ell,2} & & 0 \\ \gamma_{\ell,2} & \alpha_{\ell,2} & \ddots & \\ & \ddots & \ddots & \beta_{\ell,m} \\ 0 & & \gamma_{\ell,m} & \alpha_{\ell,m} \end{pmatrix}, \quad \text{Here it's where boundary conditions come into play}$$

with appropriate values of $\alpha_{\ell,j}$, $\beta_{\ell,j}$, and $\gamma_{\ell,j}$, satisfying $-\alpha_{\ell,j} \geq |\beta_{\ell,j+1}| + |\gamma_{\ell,j}|$. Obviously, the matrices A_ℓ commute and since the B_ℓ have negative eigenvalues, the above theorem and its corollary apply.

In other words, the iterates of the semi-implicit scheme and the AOS-scheme differ by $\mathcal{O}(\tau^2)$, which is what we want.

As it is apparent from the proof of Theorem 4.1 the constant C depends on the eigenvalues and eigenvectors of the matrices A_ℓ . For the particular matrices arising in the diffusion registration, it turns out that C is of moderate size and essentially independent on the number of gridpoints. To illustrate this fact we have computed the quantity

$$\left\| \left(I - \tau \sum_{\ell=1}^d A_\ell \right)^{-1} - \frac{1}{d} \sum_{\ell=1}^d (I - d\tau A_\ell)^{-1} \right\|_2$$

for various sizes of A_ℓ and various τ . Figure 1 shows a representative result for $n = 1024^2$, where $C = 25$ serves as an upper bound. The plots for other sizes of A_ℓ are visually indistinguishable from the displayed one.

The overall algorithm for the AOS-scheme is summarized in Table 1. Note, that the main steps in the algorithm are the computation of the force \vec{F}^k related to the chosen distance measure and the solution of the linear systems related to our particular smoother. We use a standard $\mathcal{O}(n)$ bilinear interpolation technique for the computation of the force. As already mentioned, the matrices $I - d\tau A_\ell$ are tridiagonal and strictly diagonal dominant. Hence, the $\mathcal{O}(n)$ Thomas-algorithm is a numerically stable solution technique. In conclusion, we end up with a fast and efficient $\mathcal{O}(n)$ registration algorithm.

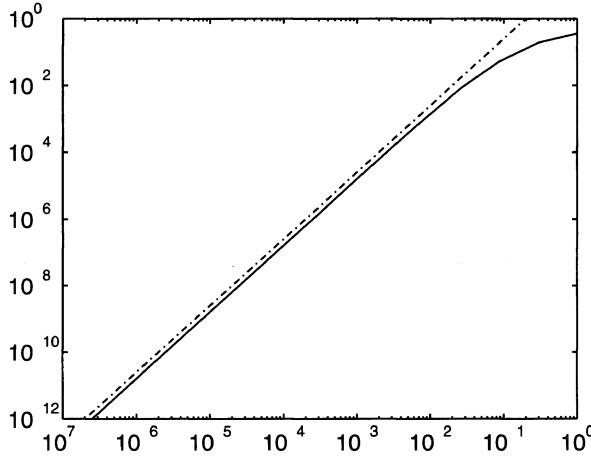


FIGURE 1. $\|(I - \tau \sum_{\ell=1}^d A_\ell)^{-1} - \frac{1}{d} \sum_{\ell=1}^d (I - d\tau A_\ell)^{-1}\|_2$ (solid line) and $C \cdot \tau^2$ (dash-dotted line) versus τ , where $C = 25$, for the matrices arising in diffusion registration.

TABLE 1. Diffusion registration for two d -dimensional images S and T .

```

Set  $k = 0$ ,  $\vec{U}^k = (\vec{U}_1^k, \dots, \vec{U}_d^k) = 0$ .
For  $k = 0, 1, 2, \dots$ 
  For  $j = 1, \dots, d$ ,
    % Compute the  $j$ th component of the force field
     $\vec{F}_j^k = (T(\vec{X} - \vec{U}^k) - R) \cdot \partial_j T(\vec{X} - \vec{U}^k)$ .
    % Compute the AOS iterate
    For  $\ell = 1, \dots, d$ ,
      Solve  $(I - d\tau A_\ell) \vec{V}_\ell = \vec{U}_j^k + \tau \vec{F}_j^k$ .
    End.
    Set  $\vec{U}_j^{k+1} = \frac{1}{d} \sum_{\ell=1}^d \vec{V}_\ell$ .
  End.
End.

```

Moreover, the implementation offers a coarse grain parallelism based on the ℓ -loop in the algorithm. Due to the special Kronecker-product structure of the matrices A_ℓ , a fine grain parallelism can be exploited. For example in two dimensions we have to the two linear systems

$$((I - 2\tau B_1) \otimes I) \vec{V}_1 = \vec{U}_j^k + \tau \vec{F}_j^k, \quad (I \otimes (I - 2\tau B_2)) \vec{V}_2 = \vec{U}_j^k + \tau \vec{F}_j^k.$$

Each of these systems decouples into a number of small systems which can be solved independently in parallel.

5. Connections to Thirion's approach

As already pointed out, there are several ways to solve (3.4). Actually, if we would have to solve (3.4) with respect to the whole space, i.e., $\Omega = \mathbb{R}^d$ then, under

mild conditions on the driving force $f^k(x) = f(x, u(x, t_k))$, it is possible to come up with an analytic solution.

A representative result in this direction reads (see [Fol95]): if $f^k \in L_1$, then the convolution

$$u(x, t)^{k+1} = K_t(x) * f^k(x) = \int_{-\infty}^t \int_{\mathbb{R}^d} K_{t-s}(x-y) f^k(y) dy ds, \quad t > 0,$$

is well defined almost everywhere and is a distributional solution of (3.4). It will be even a classical solution if $f^k \in C^p$ for $p > 1$. Here

$$K_t(x) = (4\pi t)^{-d/2} \exp(-\|x\|_2^2/(4t))$$

denotes the *Gaussian kernel*. Hence in order to solve (3.4) with respect to the bounded region $\Omega = [0, 1]^d$ one may approximate the Gaussian kernel by a Gaussian filter of suitable length, that is, to compute at each time step the force convolved with a Gaussian filter K_σ of characteristic width σ . Note, as it is well-known, the Gaussian filter based scheme is less accurate than the outlined finite difference scheme.

The Gaussian filter based approach is essentially what Thirion calls “*Demons 1: a complete grid of demons*” (see [Thi98]). However, he gives no hint on how to choose the parameter σ for a given application. It turns out in practice, that a proper choice of this free parameter is a tricky business. On the other hand, independent of the choice of σ , Thirion’s approach is also of linear complexity $\mathcal{O}(n)$.

6. Results

To illustrate the performance of the new approach based on the finite difference formulation we present the registration of two consecutive frontal sections from a series of histological tissue sections of a human brain (see [MSF01] for further details). Before we applied the diffusion registration scheme an affine linear pre-registration has been performed. The regularization parameter α (compare 2.2) is chosen such that the first displacement field \vec{U}^1 is restricted by $\max\{\vec{U}^1\} = 1$. Note, in this particular application one is not interested in removing all differences between two consecutive sections. The overall goal is to remove the differences due to the sectioning processes but to maintain the differences due to the anatomical structure of the tissue.

We are indebted to Dr. Oliver Schmitt (Institute of Anatomy, Medical University of Lübeck) for providing the medical data. Figure 2 displays the arbitrarily chosen consecutive sections 3799 and 3800 of size 1024×1024 pixel and the differences between these sections before and after registration. Note that the difference has been reduced by about 27%. We remark that in this particular application one is not interested in removing all differences between the two given images. Instead, the idea is to get rid of those differences introduced by the sectioning processes but to maintain the ones due to the anatomical structures of the tissue.

Table 2 shows the performance on a SGI OCTANE (175 MHz, MIPS R10000, 128 MB RAM under IRIX 6.5) using MATLAB 5.3.

In accordance with our theory, the CPU-times resemble nicely the linear behavior of the proposed scheme.

TABLE 2. Execution time and floating point operations per pixel (flops/pixel) for one time-step of the AOS-scheme and for different image sizes.

images size	128^2	256^2	512^2	1024^2
cpu time	0.6s	2.6s	9.7s	36.7s
flops/pixel	72.3	70.1	59.8	50.3

7. Conclusions

We have presented the novel diffusion registration technique and have devised a fast and accurate implementation. The speed is obtained by exploiting the AOS-scheme. We have shown that this scheme is as accurate as conventional implementations. The performance and potential of the new technique is demonstrated for a medical real life problem.

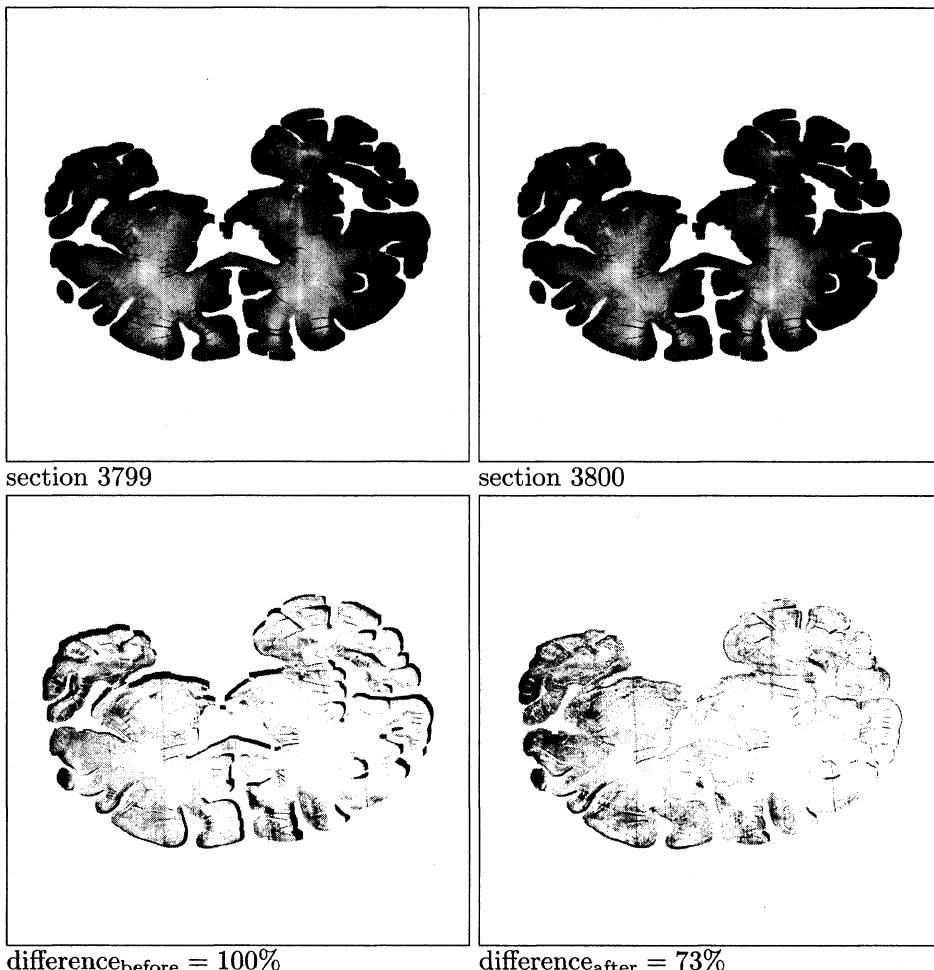


FIGURE 2. Histological frontal sections of a human brain (top row), difference between these consecutive sections before (bottom left) and after (bottom right) diffusion registration.

Beside this, we have drawn connections of the new scheme to Thirion's demon based approach.

We have shown that the new scheme is well suited for a parallel environment. More improvement in terms of execution time are expected by applying a multi-resolution or Gaussian-pyramid approach. In future work, we will also investigate the multiplicative operator scheme (MOS) instead of the AOS-scheme.

References

- [Ami94] Y. Amit, *A nonlinear variational problem for image matching*, SIAM J. Sci. Comput. **15** (1994), no. 1, 207–224.
- [BK89] R. Bajcsy and S. Kovačič, *Multiresolution elastic matching*, Computer Vision, Graphics and Image Processing **46** (1989), 1–21.
- [BN96] M. Bro-Nielsen, *Medical image registration and surgery simulation*, Ph.D. thesis, IMM, Technical University of Denmark, 1996.
- [BNG96] M. Bro-Nielsen and C. Gramkow, *Fast fluid registration of medical images*, Lecture Notes in Computer Science **1131** (1996), 267–276.
- [Bro81] C. Broidt, *Optimal registration of deformed images*, Ph.D. thesis, Computer and Information Science, Uni Pennsylvania, 1981.
- [Chr94] G. E. Christensen, *Deformable shape models for anatomy*, Ph.D. thesis, Sever Institute of Technology, Washington University, 1994.
- [FM99] B. Fischer and J. Modersitzki, *Fast inversion of matrices arising in image processing*, Num. Algo. **22** (1999), 1–11.
- [Fol95] G. B. Folland, *Introduction to partial differential equations*, 2 ed., Princeton University Press, Princeton, New Jersey, 1995.
- [GW93] R. C. Gonzales and R. E. Woods, *Digital image processing*, Addison-Wesley, 1993.
- [HJ90] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge University Press, Cambridge, 1990.
- [MSF01] J. Modersitzki, O. Schmitt, and B. Fischer, *Effiziente, nicht-lineare Registrierung eines histologischen Serienschnittes durch das menschliche Gehirn*, Bildverarbeitung für die Medizin 2001 (H. Handels et al., ed.), Springer, 2001.
- [PCA99] X. Pennec, P. Cachier, and N. Ayache, *Understanding the "demon's algorithm"* 3D non-rigid registration by gradient descent, Medical image computing and computer assisted intervention (Chris Taylor and Alan Colchester, eds.), Springer-Verlag, 1999, pp. 597–605.
- [Thi98] J. P. Thirion, *Image matching as a diffusion process: an analogy with Maxwell's demons*, Medical Image Analysis **2** (1998), no. 3, 243–260.
- [Tho49] L. H. Thomas, *Elliptic problems in linear difference equations over a network*, Tech. report, Watson Scientific Computing Laboratory, Columbia University, New York, 1949.
- [Vio95] P. A. Viola, *Alignment by maximization of mutual information*, Ph.D. thesis, Massachusetts Institute of Technology, june 1995, pp. 1–155.
- [Wei98] J. Weickert, *On discontinuity-preserving optic flow*, Proc. Computer Vision and Mobile Robotics Workshop CVMR '98 (S. Orphanoudakis, P. Trahanias, J. Crowley, and N. Katevas, eds.), 1998, pp. 115–122.

INSTITUTE OF MATHEMATICS, MEDICAL UNIVERSITY OF LÜBECK, WALLSTRASSE 40, 23560
LÜBECK, GERMANY.

E-mail address: {fischer,modersitzki}@math.mu-luebeck.de

URL: <http://math.mu-luebeck.de>

This page intentionally left blank

Iterative Stabilization and Edge Detection

C. W. Groetsch and O. Scherzer

ABSTRACT. The theory of the iterated Tikhonov-Morozov method for stable evaluation of unbounded linear operators is surveyed and a novel convergence theory, based on Dykstra's algorithm, is presented. The method is illustrated for edge detection in image processing and evaluation of the gradient of unbounded operators.

1. Introduction

Evaluation of a closed unbounded linear operator presents practical challenges that arise from two sources. First, the operator is not defined everywhere – if it were, then Banach's theorem would imply that the operator is bounded. Typically, the domain of definition of the operator is limited to certain “regular” functions that possess some degree of smoothness and perhaps are required to satisfy other ancillary conditions, such as boundary or initial conditions. On the other hand, functions representing available data may lack the attributes required for membership in the domain of the operator. Second, since the operator is unbounded arbitrarily small elements of the domain, which might manifest inaccuracies in the measurement or representation of data, may be mapped by the operator into arbitrarily large vectors. By this means small data errors may swamp the useful information one hopes to extract in the evaluation process.

A simple example from one-dimensional heat conduction illustrates some of the technicalities involved in the evaluation of unbounded operators. If in the model

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial s^2} + y(s), \quad 0 < s < \pi, \quad 0 < t < 1,$$

where $u(s, t)$ is subject to the boundary and initial conditions

$$u(0, t) = u(\pi, t) = 0, \quad u(s, 0) = 0 \quad \text{for } 0 \leq s \leq \pi,$$

one wishes to reconstruct the source distribution $y(s)$ from the spatial temperature distribution $x(s) = u(s, 1)$, one is led by formal separation of variables techniques

1991 *Mathematics Subject Classification.* 65J10, 65J20, 65M12, 47A50.

Key words and phrases. Regularization, Tikhonov-Morozov method, Dykstra's algorithm, edge detection.

The work of O.S. has been supported by the Austrian Science Foundation (FWF), grant Y-123 INF.

to the representation

$$y(s) = (Lx)(s) = \sum_{n=1}^{\infty} a_n \frac{n^2}{1 - e^{-n^2}} \sin ns$$

where

$$a_n = \frac{2}{\pi} \int_0^\pi x(s) \sin ns ds.$$

That is, $y = Lx$, where L is the linear operator on $L^2[0, \pi]$ with domain

$$\mathcal{D}(L) = \left\{ x \in L^2[0, \pi] : \sum_{m=1}^{\infty} m^4 a_m^2 < \infty, \quad a_m = \frac{2}{\pi} \int_0^\pi x(s) \sin ms ds \right\}$$

which is defined above. Note that $\mathcal{D}(L)$ is dense in $L^2[0, \pi]$ since all finite linear combinations of the eigenfunctions $\phi_n(s) = \sin ns$ are contained in $\mathcal{D}(L)$. Moreover, L is unbounded since $\|\phi_n\| = \sqrt{\pi/2}$ while

$$\|L\phi_n\| = \sqrt{\frac{\pi}{2}} \frac{n^2}{1 - e^{-n^2}} \rightarrow \infty$$

as $n \rightarrow \infty$. Finally, L is a *closed* operator, that is, its graph is a closed subspace of $L^2[0, \pi] \times L^2[0, \pi]$. Indeed, if $\{x_k\} \subseteq \mathcal{D}(L)$ and $x_k \rightarrow x$ in $L^2[0, \pi]$ as $k \rightarrow \infty$, while $Lx_k \rightarrow y$ in $L^2[0, \pi]$, then $\langle y - Lx_k, \phi_n \rangle \rightarrow 0$ as $k \rightarrow \infty$ for each n . Also,

$$\langle y - Lx_k, \phi_n \rangle = \langle y, \phi_n \rangle - \frac{n^2}{1 - e^{-n^2}} \langle x_k, \phi_n \rangle \rightarrow \langle y, \phi_n \rangle - \frac{n^2}{1 - e^{-n^2}} \langle x, \phi_n \rangle$$

Therefore,

$$\langle y, \phi_n \rangle = \frac{n^2}{1 - e^{-n^2}} \langle x, \phi_n \rangle$$

and hence

$$\sum_{m=1}^{\infty} m^4 |\langle x, \phi_m \rangle|^2 \leq \sum_{m=1}^{\infty} \left(\frac{m^2}{1 - e^{-m^2}} \right)^2 |\langle y, \phi_m \rangle|^2 = \|y\|^2.$$

Therefore, $x \in \mathcal{D}(L)$ and

$$Lx = \sum_{n=1}^{\infty} \frac{n^2}{1 - e^{-n^2}} \frac{2}{\pi} \langle x, \phi_n \rangle \phi_n = \frac{2}{\pi} \sum_{n=1}^{\infty} \langle y, \phi_n \rangle \phi_n = y.$$

In image processing filtering of noisy data is a necessary prerequisite step for *edge detection*. For edge detection one typically evaluates the gradient of images and additionally takes into account second derivative information. Edge detection typically consists of the following three steps (see e.g. [1, 8, 9]):

- Filtering: Smoothing of noise data as a basis to evaluate gradients in a stable way.
- Evaluation of the gradient and thresholding: Regions where the absolute value of a directional gradient exceeds a certain threshold are considered as candidates of edges.
- Line Thinning: After thresholding the regions of high gradients look glumpy. In order to extract one-dimensional structures, considered as edges, second derivative information is used.

We now frame the evaluation process in the context of abstract Hilbert space. Suppose $L : \mathcal{D}(L) \subseteq H_1 \rightarrow H_2$ is a closed, densely defined, unbounded linear operator from a Hilbert space H_1 into a Hilbert space H_2 . (We use $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ indiscriminately to denote the inner product and norm, respectively, in each Hilbert space.) Given $x \in \mathcal{D}(L)$ we wish to compute the value $y = Lx$. However, we assume that x itself is not available, but rather a vector $x^\delta \in H_1$ is on hand where $\|x - x^\delta\| \leq \delta$ and δ is a known bound on the approximation error. The problem of computing $y = Lx$ on the basis of the data x^δ is then ill-posed since, in general, $x^\delta \notin \mathcal{D}(L)$ and even if $x^\delta \in \mathcal{D}(L)$, generally $Lx^\delta \not\rightarrow Lx$ as $\delta \rightarrow 0$ since the operator L is unbounded. What is needed in order to stabilize the evaluation process is a method of mapping the data x^δ into $\mathcal{D}(L)$ and then using this transformed, or *mollified*, data to approximate Lx . A general approach to stabilized evaluation, based on von Neumann's theorem and spectral theory, is outlined in the next section.

2. A General Stabilization Scheme

In [6] an abstract scheme of “smoothing” the data x^δ before the application of the operator L is introduced. The basis of the method is von Neumann's theorem (see [12]) which asserts that the operators

$$\check{L} := (I + L^*L)^{-1} \quad \text{and} \quad L\check{L}$$

are defined *everywhere* and are *bounded*. The essence of the scheme is to approximate Lx by $L\check{L}T_\alpha(\check{L})x^\delta$ where T_α is a family of continuous real-valued functions on $\sigma(\check{L}) \subseteq [0, 1]$ that approximates the reciprocal function in a certain sense. One can then expect, with proper tuning of the parameter α with respect to the error level δ , that $L\check{L}T_\alpha(\check{L})x^\delta \approx Lx$. The specific requirements are that $T_\alpha(t) \rightarrow 1/t$ as $\alpha \rightarrow 0^+$ for each $t \in (0, 1]$ and that $|tT_\alpha(t)|$ is uniformly bounded. The mapping

$$x^\delta \rightarrow L\check{L}T_\alpha(\check{L})x^\delta$$

is then everywhere defined and continuous since $L\check{L}$ and $T_\alpha(\check{L})$ are bounded linear operators. In other words, the approximate evaluation process provided by this scheme is *stable* with respect to data perturbation. It is also important to notice that the process just described may be viewed as consisting of two stages: a smoothing of the data in the form $z_\alpha^\delta := \check{L}T_\alpha(\check{L})x^\delta$, followed by a stable evaluation of the operator, namely, $y_\alpha^\delta := Lz_\alpha^\delta$.

A condition for the regularity of this general scheme is developed in [6]. Specifically, if $\alpha = \alpha(\delta) \rightarrow 0$ as $\delta \rightarrow 0$ in such a way that $\delta\sqrt{r(\alpha)} \rightarrow 0$, where $r(\alpha) := \max|(1-t)T_\alpha(t)|$, then $y_\alpha^\delta = L\check{L}T_\alpha(\check{L})x^\delta \rightarrow Lx$. A convergence rate for the general method can also be purchased at the price of additional regularity on the vector x . If one assumes the stronger condition that $x \in \mathcal{D}(LL^*L)$, then it can be shown that

$$(1) \quad \|y - y_\alpha^\delta\| = \|Lx - Lz_\alpha^\delta\| = O(\omega(\alpha)) + \delta\sqrt{r(\alpha)}$$

where $\omega(\alpha) := \max|(1-tT_\alpha(t))t|$ (see, [6]).

3. The Tikhonov-Morozov Method

The Tikhonov-Morozov method is a special case of the general scheme described above which has received considerable attention (see [10], [11]). In this method $T_\alpha(t) = [\alpha + (1-\alpha)t]^{-1}$, or equivalently

$$(2) \quad y_\alpha^\delta = L(\alpha L^* L + I)^{-1} x^\delta.$$

For this method $r(\alpha) = 1/\alpha$ and $\omega(\alpha) = O(\alpha)$ and hence, by (1)

$$\|y - y_\alpha^\delta\| = O(\delta^{2/3}) \quad \text{if } \alpha \sim \delta^{2/3}$$

In [7] it is shown that the convergence rate $O(\delta^{2/3})$ mandated by the general theory under the assumption that $x \in \mathcal{D}(LL^*L)$ is in fact optimal in the sense that if the stronger rate $o(\delta^{2/3})$ were achievable for all data x^δ , then, at least for a wide class of operators L , x would lie in the nullspace of L and hence the evaluation of Lx is trivial. It is also shown that if the regularization parameter α is chosen according to the discrepancy principle, namely

$$\|(\alpha L^* L + I)^{-1} x^\delta - x^\delta\| = \delta,$$

then the suboptimal convergence rate $O(\sqrt{\delta})$ results and this suboptimal rate is best possible when the discrepancy principle is employed. However, if a modified discrepancy principle of the Gfrerer/Engl type (see, e.g. [3]) is used, then the optimal rate $O(\delta^{2/3})$ is achievable (see [7]).

The upshot of these results on the Tikhonov-Morozov method is clear: the convergence rate $O(\delta^{2/3})$ is an asymptotic brickwall which the Tikhonov-Morozov method is incapable of scaling. A similar phenomenon has long been known for the ordinary Tikhonov method for computing values of the Moore-Penrose inverse of a compact operator (see [5]). It has also long been known that there are techniques for defeating this “saturation” phenomenon – iterative methods. The remainder of this paper is therefore concerned with iterated versions of the Tikhonov-Morozov method. In the next section we review the iterated Tikhonov-Morozov method and provide what we believe to be a novel convergence analysis of the method based on Dykstra’s theorem.

4. The Iterated Method

The Tikhonov-Morozov method (2) has an important variational characterization. Indeed, it can be routinely verified that the $z_\alpha^\delta = (\alpha L^* L + I)^{-1} x^\delta$ is characterized by

$$z_\alpha^\delta = \arg \min_{z \in \mathcal{D}(L)} \|z - x^\delta\|^2 + \alpha \|Lz\|^2.$$

(this minimizer will necessarily be contained in the subspace $\mathcal{D}(L^*L)$ of smoother functions). In the standard Tikhonov-Morozov method each change in the regularization parameter α entails that the calculations must begin *ab initio*. On the other hand, the variational characterization suggests the possibility of keeping the parameter α fixed and using the result of one minimization to stabilize the next minimization in the following iterative manner:

$$z_n = \arg \min_{z \in \mathcal{D}(L)} \|z - x\|^2 + \alpha \|Lz - Lz_{n-1}\|^2 \quad n = 1, 2, \dots$$

where $z_0 = 0$. Equivalently, the method may be expressed as

$$(I + \alpha L^* L) z_n^\delta = x^\delta + \alpha L^* L z_{n-1}^\delta.$$

In this formulation one sees that the operator to be inverted, $(I + \alpha L^* L)$, is the same at each stage of the iteration, and hence the iterated method is computationally more attractive than the standard Tikhonov-Morozov method.

In the iterated Tikhonov-Morozov method the role of the regularization parameter is assumed by the iteration number n rather than the parameter α . The convergence analysis of the iterated Tikhonov-Morozov method may be couched in terms of the general spectral approach of Section 2. However, for the sake of novelty and to provide additional insight into the method, we offer a non-spectral approach based on Dykstra's theorem (see [2], p. 216). For our purposes we may formulate this result as follows:

Suppose V_1 and V_2 are closed affine subsets of a Hilbert space, and P_i is the metric projector onto V_i . Then $(P_2 P_1)^n x \rightarrow P_{V_1 \cap V_2} x$ as $n \rightarrow \infty$ for each x if and only if $V_1 \cap V_2 \neq \emptyset$.

The analysis is carried out in the product Hilbert space $\mathcal{H} = H_1 \times H_2$ with inner product $\langle \cdot, \cdot \rangle$ and norm $|\cdot|$, respectively, given by

$$\langle (u_1, v_1), (u_2, v_2) \rangle = \langle u_1, u_2 \rangle_{H_1} + \alpha \langle v_1, v_2 \rangle_{H_2},$$

and

$$|(u, v)|^2 = \|u\|_{H_1}^2 + \alpha \|v\|_{H_2}^2.$$

Suppose $x \in H_1$ and let $V_1 = \{x\} \times H_2$ and let $V_2 = \mathcal{G}(L)$, the graph of L . Then V_1 is closed (in \mathcal{H}) and affine, and V_2 is closed (since L is a closed linear operator) and affine. Also,

$$V_1 \cap V_2 \neq \emptyset \iff x \in \mathcal{D}(L).$$

Moreover, $V_1 \cap V_2 = \{(x, Lx)\}$ when $x \in \mathcal{D}(L)$.

Since

$$\begin{aligned} z_1 &= \arg \min_{z \in \mathcal{D}(L)} \|z - x\|^2 + \alpha \|Lz - L0\|^2 \\ &= \arg \min_{z \in \mathcal{D}(L)} |(z, Lz) - (x, 0)|^2, \end{aligned}$$

(z_1, Lz_1) is the \mathcal{H} -orthogonal projection of $(x, 0)$ onto $\mathcal{G}(L) = V_2$. Similarly,

$$z_2 = \arg \min_{z \in \mathcal{D}(L)} |(z, Lz) - (x, Lz_1)|^2$$

That is, (z_2, Lz_2) is the orthogonal projection of (x, Lz_1) onto V_2 . Hence,

$$(z_2, Lz_2) = P_2(x, Lz_1) = P_2 P_1(z_1, Lz_1)$$

and

$$(z_1, Lz_1) = P_2(x, 0) = (P_2 P_1)(0, 0)$$

In this way we see that

$$(z_n, Lz_n) = (P_2 P_1)^n (0, 0)$$

and hence by Dykstra's theorem

$$(z_n, Lz_n) \rightarrow (x, Lx) \in V_1 \cap V_2, \quad \text{as } n \rightarrow \infty \iff x \in \mathcal{D}(L)$$

We note that $P_{V_1 \cap V_2}(u, v) = (x, Lx)$ for all $(u, v) \in \mathcal{H}$ if $x \in \mathcal{D}(L)$ since $V_1 \cap V_2$ is in this case the singleton $\{(x, Lx)\}$. Therefore the iterative method converges in graph norm to x for *any* initial approximation $z_0 \in H_1$.

The alternating projection viewpoint is also useful for providing a stability analysis in the case of approximate data x^δ . Suppose $x^\delta \in H_1$ ($x^\delta \notin \mathcal{D}(L)$, in

general), and $\|x - x^\delta\| \leq \delta$. Let P_1^δ be the \mathcal{H} -orthogonal projector of \mathcal{H} onto $V_1^\delta := \{x^\delta\} \times H_2$. Then

$$\begin{aligned} \alpha \|Lz_n^\delta - Lz_n\|^2 &\leq |(z_n^\delta, Lz_n^\delta) - (z_n, Lz_n)|^2 \\ &= |P_2 P_1^\delta(z_{n-1}, Lz_{n-1}) - P_2 P_1(z_{n-1}, Lz_{n-1})|^2 \\ &\leq |P_1^\delta(z_{n-1}, Lz_{n-1}) - P_1(z_{n-1}, Lz_{n-1})|^2 \\ &= |(x^\delta, Lz_{n-1}^\delta) - (x, Lz_{n-1})|^2 \\ &= \|x^\delta - x\|^2 + \alpha \|Lz_{n-1}^\delta - Lz_{n-1}\|^2 \\ &\leq \delta^2 + \alpha \|Lz_{n-1}^\delta - Lz_{n-1}\|^2. \end{aligned}$$

Hence,

$$\|Lz_n^\delta - Lz_n\|^2 \leq n\delta^2/\alpha$$

i.e.,

$$Lz_n^\delta \rightarrow Lx \quad \text{as } \delta \rightarrow 0 \quad \text{if } n\delta^2 \rightarrow 0$$

5. The Nonstationary Iterated Method

A nonstationary version of the iterated Tikhonov-Morozov method is developed in [4]. The approximation z_n^δ is defined by

$$z_n^\delta = \arg \min_{z \in \mathcal{D}(L)} \|x^\delta - z\|^2 + \alpha_n \|Lz - Lz_{n-1}^\delta\|^2.$$

This is equivalent to the iteratively defined method

$$(I + \alpha_n L^* L)z_n^\delta = x^\delta + \alpha_n L^* Lz_{n-1}^\delta, \quad z_0^\delta = 0,$$

where $\{\alpha_n\}$ is a suitable sequence of positive parameters.

Before presenting convergence results for this nonstationary iterated Tikhonov - Morozov method, we provide some motivation. The basic idea of smoothing methods is to produce smooth approximate versions z_n^δ in $\mathcal{D}(L)$ of the data x^δ , which might not lie in $\mathcal{D}(L)$. One can imagine such smoothing being accomplished by the evolution problem

$$\frac{du}{dt} = -L^* Lu, \quad u(0) = x^\delta.$$

For example, if $L = \nabla$, the gradient operator, then this is the classical heat equation – a notorious smoother. Since $L^* L$ has an unbounded spectrum, $u(t)$ will be a smoothed version of $u(0)$, even for small values of t . The smoothing effect may be approximated by a single step of the implicit Euler method (a method known for its stability) with step size α_1 . Setting $z_0^\delta = u(0) = x^\delta$ and $z_1^\delta \approx u(\alpha_1)$, we have

$$\frac{z_1^\delta - z_0^\delta}{\alpha_1} = -L^* Lz_1^\delta \quad \text{or} \quad (I + \alpha_1 L^* L)z_1^\delta = z_0^\delta.$$

This, of course, is the standard Tikhonov-Morozov method. In the approximation z_1^δ , certain noise components in x^δ are damped by the factors $(1 + \alpha_1 \mu_n)^{-1}$, where μ_n are eigenvalues of $L^* L$. The possibility of subtracting out this smoothed approximation and continuing the filtration process on the remaining noise suggests

itself. That is, one takes z_2^δ to be the approximation to $u(\alpha_2)$ obtained by applying the implicit Euler method, with step size α_2 , to the problem

$$\frac{du}{dt} = -L^*Lu, \quad u(0) = x^\delta - z_1^\delta.$$

This leads to the next approximation z_2^δ , where

$$(I + \alpha_2 L^* L)(z_2^\delta - z_1^\delta) = x^\delta - z_1^\delta.$$

In this way one produces a sequence of filtered approximations $\{z_n^\delta\}$ defined by

$$(I + \alpha_n L^* L)z_n^\delta = x^\delta + \alpha_n L^* L z_{n-1}^\delta, \quad z_0^\delta = 0.$$

It can be shown that if the iteration number $n = n(\delta)$ is chosen so that $\delta \sigma_{n(\delta)} \rightarrow 0$ as $\delta \rightarrow 0$, where $\sigma_n = \sum_{j=1}^n \alpha_j^{-1}$, then $z_{n(\delta)}^\delta \rightarrow x$ in the graph norm of L , i.e., $z_{n(\delta)}^\delta \rightarrow x$ and $Lz_{n(\delta)}^\delta \rightarrow Lx$ (see [4]). Moreover, an a posteriori parameter strategy may be used which, under certain conditions, achieves an order of convergence $O(\delta^\nu)$ for any $\nu \in [0, 1]$. In fact, if the signal-to-noise ratio is strictly greater than one, i.e., $\|x^\delta\|_1 > \tau\delta$, for some $\tau > 1$, and if the stopping index of the iteration is chosen to be the first index $n = n(\delta)$ satisfying

$$\|z_{n(\delta)}^\delta - x^\delta\|_1 \leq \tau\delta$$

then

$$\|Lz_{n(\delta)}^\delta - Lx\|_2 = O(\delta^{1-\frac{1}{\nu}})$$

assuming that the true data satisfies the source condition $x \in R((I + L^* L)^{-\nu})$ for some $\nu \in (0, 1)$. This shows that the nonstationary iterated Tikhonov - Morozov method does not suffer the saturation order $O(\delta^{2/3})$ of the ordinary Tikhonov - Morozov method, but in fact can attain an essentially optimal order of convergence.

6. Numerical Illustrations

In this section we present several numerical examples in image processing and the evaluation of unbounded operators.

6.1. Edge Detection. Here we consider denoising of an ultrasound data set for edge detection. The filtered images are obtained via the Tikhonov-Morozov method. Afterwards a Canny edge detector is applied to the filtered image.

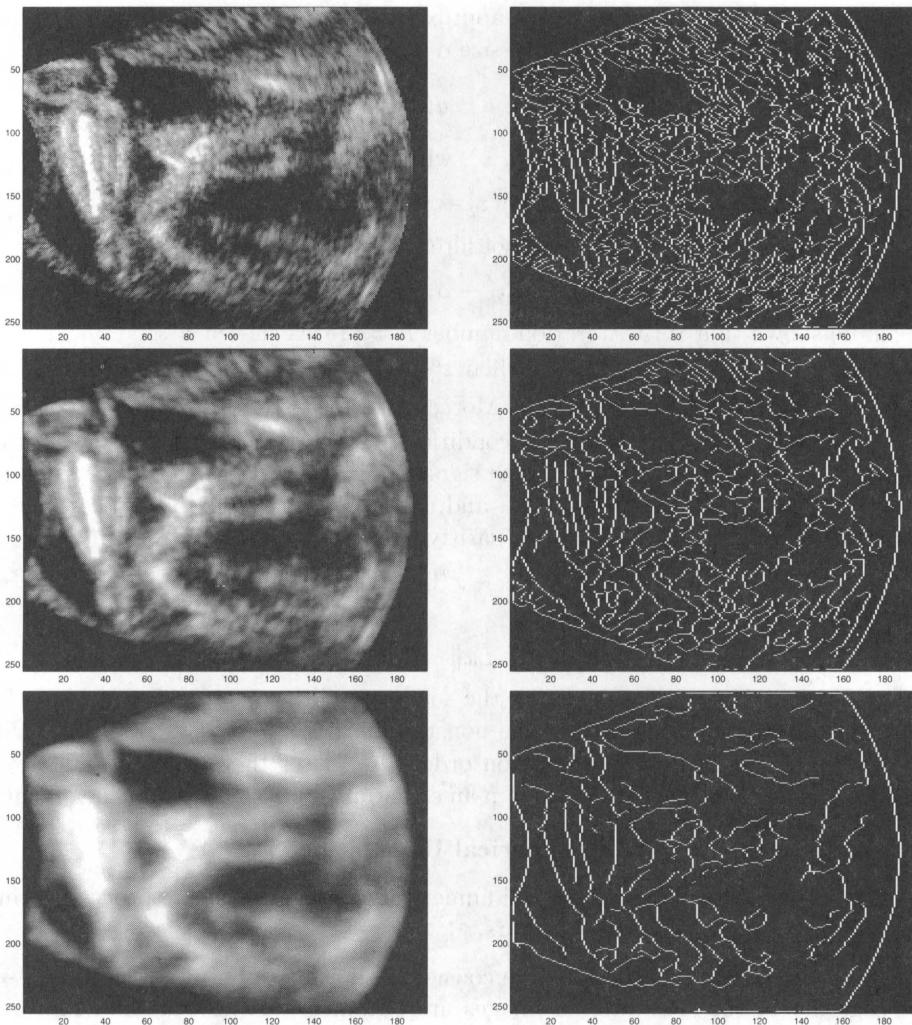


FIGURE 1. Denoising and edge-detection: Regularization with parameters $\alpha = 0, 1, 10, 100$ (left images). Canny edge detector applied to the filtered images.

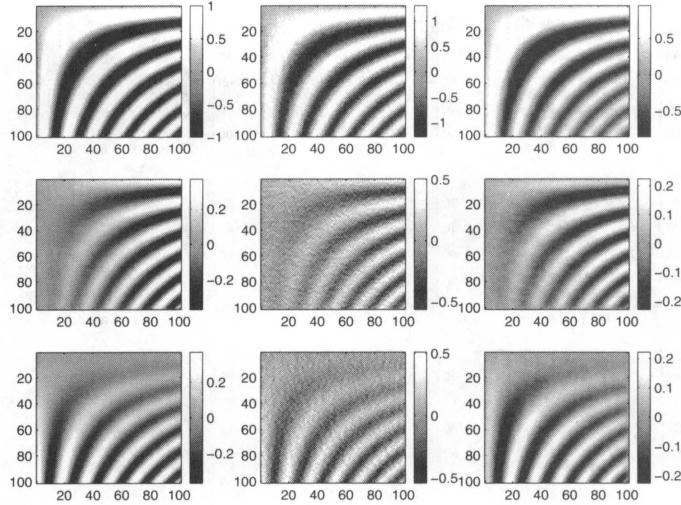


FIGURE 2. Evaluation of the gradient: Left Row: Noise free function (top) and the gradients in the two coordinate directions evaluated numerically. Middle Row: Measurement data with medium noise (top) and the gradients in the two coordinate directions evaluated numerically. Right Row: Denoised image (top) and the gradients.

6.2. Numerical Differentiation. Here we apply the Tikhonov-Morozov for evaluation of the gradient of a noisy two-dimensional data set at several noise levels.

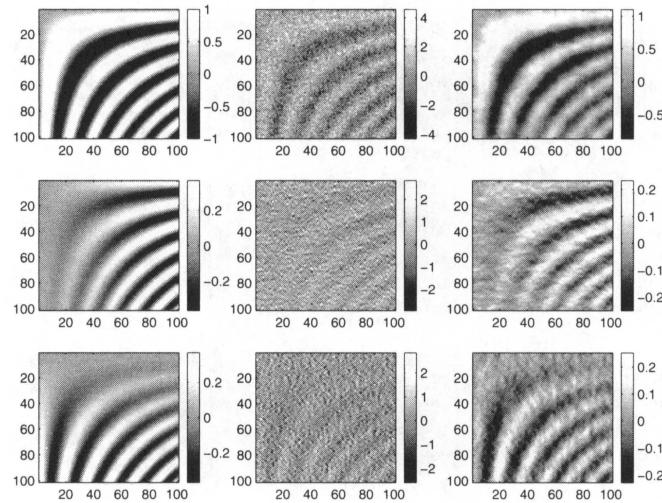


FIGURE 3. Evaluation of the gradient: Left Row: Noise free function (top) and the gradients in the two coordinate directions evaluated numerically. Middle Row: Measurement data with high noise (top) and the gradients in the two coordinate directions evaluated numerically. Right Row: Denoised image (top) and the gradients.

6.3. Tikhonov-Morozov method versus iterated Tikhonov-Morozov method. Here we compare the effect of frequent iteration. If the iterated regularization parameters $\{\hat{\alpha}_i\}_{i=1}^n$ satisfy

$$\sum_{i=1}^n \frac{1}{\hat{\alpha}_i} = \alpha,$$

then the result with the Tikhonov-Morozov method and the corresponding iterated Tikhonov-Morozov method look very similar.

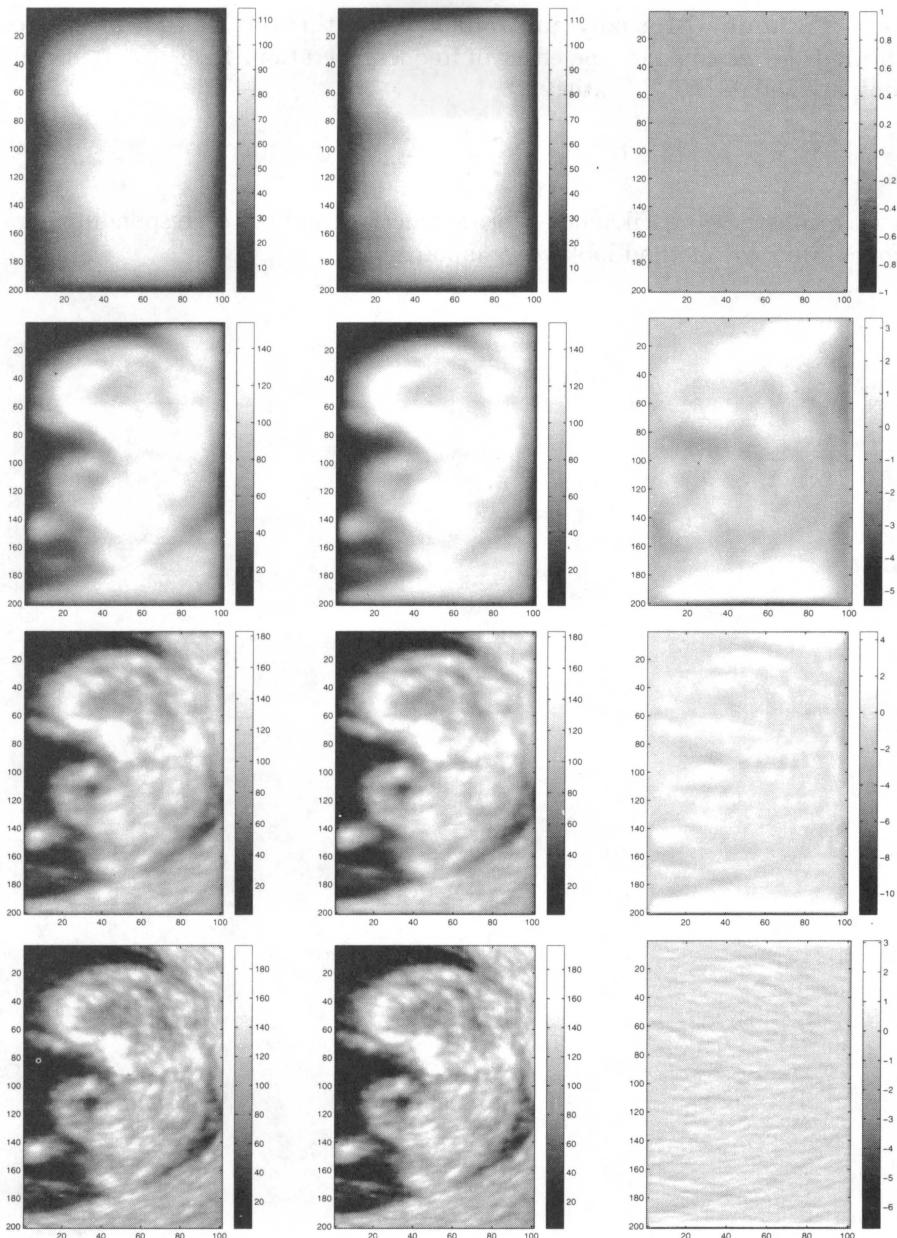


FIGURE 4. The effect of iteration: Pictures left: Tikhonov-Morozov method with regularization parameters $\alpha_1 = 200, \alpha_2 = \frac{200}{11}, \alpha_3 = \frac{200}{111}, \alpha_4 = \frac{200}{1111}$: Pictures middle: Iterated Tikhonov-Morozov: top image: 1 iteration with regularization parameter $\hat{\alpha}_1 = 200$, middle image - two iterations with parameters $\hat{\alpha}_1$ and $\hat{\alpha}_2 = 20$; middle image - three iterations with parameters $\hat{\alpha}_1, \hat{\alpha}_2$ and $\hat{\alpha}_3 = 2$; bottom images - four iterations with parameters $\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3$ and $\hat{\alpha}_4 = 0.2$. Difference between iterated and Tikhonov-Morozov method.

References

- [1] J.F. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Machine Intell.*, PAMI-8(6):679–697, 1986.
- [2] F. Deutsch. *Best Approximation in Inner Product Spaces*. Springer, New York, 2001.
- [3] H.W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, Dordrecht, 1996.
- [4] C. W. Groetsch and O. Scherzer. Nonstationary iterated tikhonov-morozov method and third order differential equations for the evaluation of unbounded operators. *Math. Meth. Appl. Sci.*, 23:1287–1300, 2000.
- [5] C.W. Groetsch. *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*. Pitman, Boston, 1984.
- [6] C.W. Groetsch. Spectral methods for linear inverse problems with unbounded operators. *J. Approx. Th.*, 70:16–28, 1992.
- [7] C.W. Groetsch and O. Scherzer. Optimal order of convergence for stable evaluation of differential operators. *Electronic Journal of Differential Equations*, 4:1–10, 1993. <http://ejde.math.unt.edu>.
- [8] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, 1998.
- [9] S. Mallat and S. Zhong. Characterization of signals from multiscale edges. *IEEE Trans. Patt. Anal. Machine Intell.*, 14(7):710–732, 1992.
- [10] V.A. Morozov. A stable method for computation of values of unbounded operators. *Soviet Math. Doklady*, 10:339–342, 1969.
- [11] V.A. Morozov. *Methods for Solving Incorrectly Posed Problems*. Springer Verlag, New York, Berlin, Heidelberg, 1984.
- [12] F. Riesz and B. Sz-Nagy. *Functional Analysis*. Ungar, New York, 1955.

DEPARTMENT OF MATHEMATICAL SCIENCES, UNIVERSITY OF CINCINNATI, CINCINNATI, OH 45221-0025 U.S.A.

E-mail address: `groetsch@email.uc.edu`

DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF INNSBRUCK, TECHNIKER STR. 25, A-6020 INNSBRUCK, AUSTRIA

E-mail address: `otmar.scherzer@uibk.ac.at`

This page intentionally left blank

BACKPROJECTIONS IN TOMOGRAPHY, SPHERICAL FUNCTIONS AND ADDITION FORMULAS: A FEW CHALLENGES

F. ALBERTO GRÜNBAUM

1. INTRODUCTION AND STATEMENT OF RESULTS

The purpose of this short paper is to present to a wider audience a pair of problems originating with very classical considerations in X-ray tomography. In a certain sense the problems are implicitly posed in my joint paper with M. Davison back in 1981 [DG]. I had a recent opportunity to present these open problems at MSRI, in Berkeley, during the semester long parallel programs in Inverse Problems and Integral Geometry, and judging from the responses there I decided to put these concerns down on paper.

There are two further purposes behind this paper: I talk briefly about some highly *nonlinear* versions of the *linear* tomographic problems that led to the issues discussed in this paper, and I touch upon the subject of *matrix valued* spherical functions that has yet to play a role even in linear tomography problems. This is a result of an ongoing joint effort with my colleagues I. Pacharoni and J.A. Tirao. To keep the paper within bounds these two sections are short but I give lots of references where the reader can find more detailed accounts of the work.

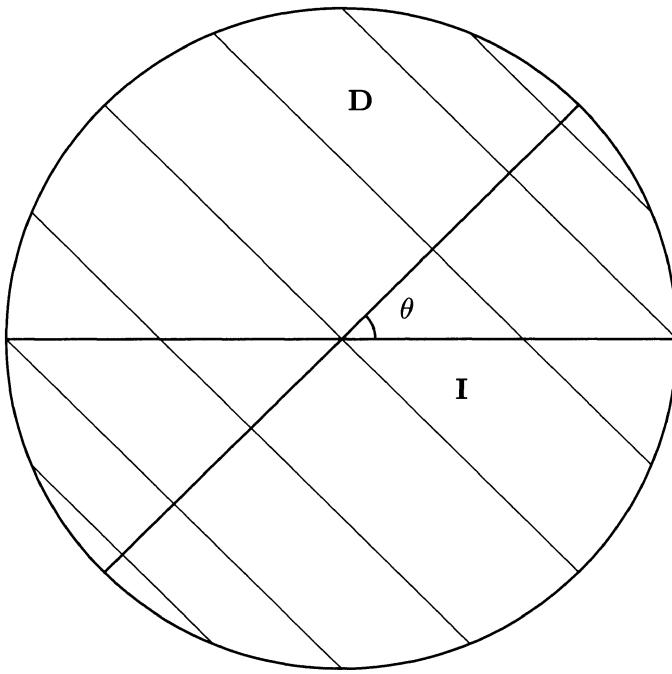
The underlying theme in the paper is that inverse problems as those raised by medical imaging are a natural playground for many of the tools of classical analysis. In turn, the consideration of new physical/geometrical set-ups in medical imaging poses new mathematical challenges that need to be looked into. The other side of the coin is that recent developments in mathematics, growing purely out of a desire to push classical theories forward, could eventually find a use in either medical or other imaging problems.

2. PARALLEL BEAM TOMOGRAPHY AND SOME SPHERICAL FUNCTIONS

In this case the geometrical setup is as indicated in the picture below.

This paper is partially supported by NSF grant FD9971151.

© 2002 American Mathematical Society



If the angles $\theta_1, \theta_2, \dots, \theta_m$ are given arbitrarily, we can consider as data the integrals (or projections) of the unknown function f supported in the disk of radius $1/2$

$$(P_{\theta_i} f)(t) = \int_{x \cos \theta_i + y \sin \theta_i = t} f.$$

Introduce as in [DG] the operation of backprojection by means of

$$(B_\theta h)(x, y) = h(x \cos \theta + y \sin \theta).$$

Our goal is now to choose *filters* α_i such that the expression

$$\frac{1}{m} \sum_{i=1}^m (P_{\theta_i} f * \alpha_i)(x \cos \theta_i + y \sin \theta_i)$$

is as close as possible to the unknown function f .

More precisely, and this is one of the key ideas in [DG], we give up *point evaluation*, pick a *desired point response function* ϕ and try to approximate

$$f * \phi.$$

This means that we set out to choose α_i such that we minimize

$$\sup_{\bar{y} \in \text{supp } f} \left| (f * \phi)(\bar{y}) - \frac{1}{m} \sum_{i=1}^m (P_{\theta_i} f * \alpha_i)(\bar{y}, \bar{\theta}_i) \right|.$$

The second term above is expressible as

$$\left(f * \frac{1}{m} \sum_{i=1}^m B_{\theta_i} \alpha_i \right) (\bar{y})$$

and thus we are looking at minimizing the supremum over y in the support of f of

$$\left| \int_D f(\bar{y} - \bar{x}) \left(\phi - \frac{1}{m} \sum_{i=1}^m B_{\theta_i} * \alpha_i \right) (\bar{x}) d\bar{x} \right|.$$

If we introduce an arbitrary weight function w_2 we can use Schwartz' inequality to bound the error by

$$\left(\sup_{\bar{y} \in \text{supp } f} \|f_{\bar{y}}\|_{w_2} \right) \left\| \phi - \frac{1}{m} \sum_{i=1}^m B_{\theta_i} \alpha_i \right\|_{1/w_2}.$$

This gives us then the problem of finding α_i so as to minimize the second factor above.

There is standard approach: introduce weight functions, consider the corresponding Hilbert spaces and solve the resulting *normal equations*.

If w_2 is the weight function on the unit disk D and w_1 is a weight function on the unit interval I we consider the map P that assigns to a function in $L^2(D, w_2)$ the collection of its one dimensional projections introduced above.

In this way we have a map P

$$P : L^2(D, w_2) \rightarrow \bigoplus L^2(I, w_1)$$

and its adjoint

$$P^t : \bigoplus L^2(I, w_1) \rightarrow L^2(D, w_2).$$

The equations we should try to solve are

$$PP^t \alpha = P\phi, \quad \alpha = (\alpha_1, \dots, \alpha_m)$$

i.e., a system of linear equations in $\bigoplus L^2(I, w_1)$.

In general these equations cannot be solved effectively. But there is a *miracle*. If the weights w_2 and w_1 are chosen with great care then in an appropriate basis we get a block diagonal system of equations with blocks of size m , the number of angles. In practice we keep as many of these blocks as the number N of “Fourier coefficients” that we want to determine for each function α_i .

For the technical details the reader should consult [DG]. Here we recall how this miracle comes into play and what the connection with spherical functions is.

Assume that w_2 is rotationally invariant. Then

$$P_{\theta_j} P_{\theta_i}^t$$

depends only on $\langle \bar{\theta}_i, \bar{\theta}_j \rangle$ and we need to consider the family of integral operators

$$P_0 P_\theta^t$$

θ arbitrary, mapping functions of one variable to functions of one variable. The presence of weights introduces extra factors and we get the following explicit expression

$$(P_0 P_\theta^t g)(s) = \int_{-\sqrt{1-s^2}}^{\sqrt{1-s^2}} \frac{g(s \cos \theta + t \sin \theta) w_1(s \cos \theta + t \sin \theta)}{w_2(s, t)} dt.$$

The main issue is the following: can the weights be chosen so that

- a) we get a commutative family of integral operators.
- b) can we compute a basis of common eigenvectors?

The relevant result from [DG] is that this is possible if

- a) $P_0 \left(\frac{1}{w_2} \right) = \frac{1}{w_1}$.
- b) $w_2(\bar{x}) \cong (1 - r^2)^{1-\lambda}$, $\lambda > 0$.

In this case

$$w_1(x) \cong (1 - x^2)^{1/2-\lambda}$$

and the common set of eigenfunctions are

$$C_n^\lambda(x)(1 - x^2)^{\lambda-1/2}$$

with eigenvalues given by

$$C_n^\lambda(\cos \theta).$$

Indeed in this case we get

$$\begin{aligned} (P_0 P_\theta^t g)(s) &= (P_0 M_{1/w_2} B_\theta M_{w_1} g)(s) \\ &= \int_{-(1-s^2)^{1/2}}^{(1-s^2)^{1/2}} \frac{g(s \cos \theta + t \sin \theta)(1 - (s \cos \theta + t \sin \theta)^2)^{1/2-\lambda}}{(1 - s^2 - t^2)^{1-\lambda}} dt \end{aligned}$$

and making the change of variables

$$\begin{aligned} s &= \cos \phi, \quad t = \sin \cos \psi \\ dt &= -\sin \phi \sin \psi d\psi \\ 1 - s^2 - t^2 &= \sin^2 \phi \sin^2 \psi \end{aligned}$$

and putting

$$g(s) \equiv f(s)(1 - s^2)^{\lambda-1/2}$$

we get

$$(P_0 P_\theta^t g)(s) \cong (\sin \phi)^{2\lambda-1} \int_0^\pi f(\cos \phi \cos \theta + \sin \phi \sin \theta \cos \psi) \sin^{2\lambda-1} \psi d\psi.$$

In the simple case of $\lambda = 1/2$ (Legendre) we recognize here that the integral identity satisfied by the Legendre polynomials, (expressing the fact that they are the spherical functions for the pair $SO(3), SO(2)$) becomes

$$(P_0 P_\theta^t P_n)(\cos \phi) = P_n(\cos \theta) P_n(\cos \phi)$$

illustrating a case of the result [DG] quoted above.

In the more general case one has to deal with Gegenbauer polynomials which for special values of the parameter λ give spherical functions for pairs

$SO(n+1), SO(n)$. In particular, if λ is 1 we get Chebychev polynomials of the second kind, if λ is 1/2 we are in the Legendre case, and for λ equal to zero we have the Chebychev polynomials of the first kind. This case is discussed in [HS] and [LS].

3. FAN AND CONE BEAM TOMOGRAPHY

This section is very short for a very good reason: it consists only of challenges. Suppose the source of X-rays which in the previous section should be thought of as being at infinite distance from the patient is brought in closer and moves either on a circle on one plane (fan beam) or on a spiral in three dimensional space (cone beam). In either case one can define anew operators of projection mapping functions of several variables to functions of one variable by integrating along the rays of a given fan or cone position. One can also try to introduce weight functions and talk about the adjoints of these operators and thus as in the case of the parallel beam one can consider a family of integral operators.

The challenge is to find weights that would make this family of integral operators a commutative family with a common set of eigenfunctions. The eigenvalues should depend on the parameter describing the location of the source. If this is possible it would be very useful to get one's hands on the eigenfunctions.

4. IS THERE ANY USE FOR *addition formulas* IN TOMOGRAPHY?

The product formula for the Gegenbauer polynomials that we used earlier on, and which serves as the definition for spherical functions in general, see [He], is a trivial consequence of something called *addition formulas* in the literature, see [A], [AAR].

These formulas are much harder to come by as the dates below indicate.

In the case of the Legendre polynomials, they were established by Laplace in 1782. In the case of the Gegenbauer polynomials, they were established by him in 1875.

Both of these are special cases of the Jacobi polynomials. The formula for these was only established (for general α and β) by T. Koornwinder in 1972.

The question that I want to raise is now clear: is there any concrete result in tomography or in integral geometry in general that requires the use of these complicated formulas? Can one (if nothing better is around) work backwards and find a problem appropriate for these tools?

5. SPHERICAL FUNCTIONS FOR GELFAND PAIRS (G, K) , ONE EXAMPLE

In this section we recall the definition of spherical functions for a Gelfand pair made up of a group G and a compact subgroup K of it. For the benefit of the reader we illustrate the definition in the case of a familiar example, related to the Legendre polynomials that appeared earlier in this paper.

One says that a scalar valued function $f : G \rightarrow \mathbb{C}$ defined on a group G is a spherical function if for every pair $g_1, g_2 \in G$ we have

$$f(g_1)f(g_2) = \int_K f(g_1kg_2)dk.$$

Here K is a compact subgroup of G , and dk denotes Haar measure on K . One also assumes $f(e) = 1$. From the definition it follows that

$$f(g_1k_0)f(g_2) = \int_K f(g_1k_0kg_2)dk = f(g_1)f(g_2)$$

as well as

$$f(g_1)f(k_0g_2) = \int_K f(g_1kk_0g_2)dk = f(g_1)f(g_2).$$

The consequence is that for all g, k_1, k_2

$$f(k_1gk_2) = f(g).$$

For the benefit of the reader we look at this in detail when $G = SO(3)$, $K = SO(2)$. We think of K as rotations around the z axis.

Using Euler angles, introduced in 1776, any element in G can be represented as

$$\begin{aligned} & \begin{pmatrix} \cos \varphi & -\sin \varphi & 0 \\ \sin \varphi & \cos \varphi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta \\ 0 & \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} \cos \psi & -\sin \psi & 0 \\ \sin \psi & \cos \psi & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} * & * & \sin \varphi \sin \theta \\ * & * & -\cos \varphi \sin \theta \\ * & * & \cos \theta \end{pmatrix} \end{aligned}$$

This map is $1 - 1$ except when $\theta = 0, \pi$. We ignore this here. This allows us to establish a $1 - 1$ correspondence between equivalence classes gK and vectors in the unit sphere of \mathbb{R}^3 , namely

$$gK \leftrightarrow g \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \sin \varphi \sin \theta \\ -\cos \varphi \sin \theta \\ \cos \theta \end{pmatrix}$$

i.e., we have

$$G/K = SO(3)/SO(2) \leftrightarrow S_2.$$

Now spherical functions, as noticed above, are defined on $K \backslash G / K$ which is clearly identified with the interval $[-1, 1]$, i.e., the segment joining the south and north poles of the unit sphere.

Observe that g and g^{-1} have the same value for the cosine of its colatitude θ . Indeed if g has Euler angles φ, θ, ψ , g^{-1} has Euler angles $-\psi, -\theta, -\varphi$.

Since a spherical function is well defined on $K \backslash G / K$, we can compute $f(g_1)$ and $f(g_2)$ by assuming

$$g_1 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \sin \alpha \\ 0 \\ \cos \alpha \end{pmatrix}; \quad g_2 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \sin \beta \\ 0 \\ \cos \beta \end{pmatrix}$$

and we have, from the definition,

$$\begin{aligned} f(\cos \alpha) f(\cos \beta) &= f(g_1) f(g_2) = f(g_1^{-1}) f(g_2) \\ &= \int_K f(g_1^{-1} k g_2) dk \\ &= \int_K f(\text{colatitude of } g_1^{-1} k g_2) dk. \end{aligned}$$

Now if $k = \begin{pmatrix} \cos \xi & -\sin \xi & 0 \\ \sin \xi & \cos \xi & 0 \\ 0 & 0 & 1 \end{pmatrix}$ we get $k g_2 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \cos \xi \sin \beta \\ \sin \xi \sin \beta \\ \cos \beta \end{pmatrix}$ and for

the colatitude of $g_1^{-1} k g_2$, we obtain

$$\left\langle g_1^{-1} k g_2 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\rangle = \left\langle k g_2 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, g_1 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\rangle = \cos \alpha \cos \beta + \sin \alpha \sin \beta \cos \xi$$

and we finally get

$$f(\cos \alpha) f(\cos \beta) = \frac{1}{2\pi} \int_0^{2\pi} f(\cos \alpha \cos \beta + \sin \alpha \sin \beta \cos \xi) d\xi.$$

6. MATRIX VALUED SPHERICAL FUNCTIONS ON THE COMPLEX PROJECTIVE PLANE $P_2(\mathbb{C}) = G/K$, WITH $G = \text{SU}(3)$ AND $K = \text{S}(\text{U}(2) \times \text{U}(1)) \simeq \text{U}(2)$ AND FOR THE THREE DIMENSIONAL HYPERBOLIC SPACE $\mathbb{H} = \mathbb{C} \times \mathbb{R}^+ = G/K$, WHERE $G = \text{SL}(2, \mathbb{C})$ AND $K = \text{SU}(2)$

In this section I limit myself to giving the general definition and then give a few comments regarding the two examples where things have been worked out in detail so far. There is an extensive list of references that the reader may want to consult.

Let G be a locally compact unimodular group and let K be a compact subgroup of G . Let \hat{K} denote the set of all equivalence classes of complex finite dimensional irreducible representations of K ; for each $\delta \in \hat{K}$, let ξ_δ denote the character of δ , $d(\delta)$ the degree of δ , i.e. the dimension of any representation in the class δ , and $\chi_\delta = d(\delta)\xi_\delta$. We shall choose once and for all the Haar measure dk on K normalized by $\int_K dk = 1$.

We shall denote by V a finite dimensional vector space over the field \mathbb{C} of complex numbers and by $\text{End}(V)$ the space of all linear transformations of V into V .

A spherical function Φ on G of type $\delta \in \hat{K}$ is a continuous function on G with values in $\text{End}(V)$ such that

- i) $\Phi(e) = I$. (I = identity transformation).
- ii) $\Phi(x)\Phi(y) = \int_K \chi_\delta(k^{-1})\Phi(xky) dk$, for all $x, y \in G$.

The reader can find a number of general results in [T] and [GV].

It is well known that there is a fruitful connection between the hypergeometric function of Euler and Gauss and the spherical functions of trivial type associated to a rank one symmetric pair (G, K) . On the other hand the spherical functions of types of dimension bigger than one have been worked out explicitly very recently in the two examples given in the title of this section.

This is accomplished by associating to a spherical function Φ on G a matrix valued function H on G/K . The entries of H are solutions of two coupled systems of ordinary differential equations.

The solution to this pair of systems can be exhibited explicitly in terms of a special class of generalized hypergeometric functions ${}_p+1F_p$.

The interested reader should consult [GPT] ,[GPT1] ,[GPT2] and [GPT3].

7. NONLINEAR TOMOGRAPHY. i.e. VERY LOW ENERGY SOURCES

Optical, or diffuse tomography, refers to the use of low energy probes to obtain images of highly scattering media.

The main motivation for this line of work is, at present, the use of an infrared laser to obtain images of diagnostic value. This is intended for use in a neonatal clinic to measure oxygen content in the brain of premature babies as well as in the case of repeated mammographies. With the development of highly specific markers that respond well in the optical or infrared region there are many potential applications of this emerging area. It is clear that the main objective is not to replace other well established imaging modalities but rather to find out ways of exploiting a different (and mathematically harder to analyze) part of the spectrum.

For a very nice and up-to-date discussion of work one can see [A1],as well as [NW]. For early attempts in this area see [G1] and [SGKZ].

The purpose of this short section is to call attention to some recent results dealing with the exact inversion, up to an appropriate manifold of arbitrary parameters, in a problem that is motivated by diffuse tomography.

The inverse problem for one of the earliest and crudest models of optical tomography amounts to reconstructing the one-step transition probability matrix for a Markov chain (with three kinds of states) from boundary measurements. This model is too simplistic and general to faithfully reflect the physics of diffuse tomography but might be of interest in other set-ups. It gives a difficult class of *nonlinear* inverse problems for certain *general class of networks* with a complex pattern of connections which are motivated by the diffuse tomography picture.

A remarkable feature of this simple model is that, at least for systems arising from very coarse tomographic discretizations, it gives an exactly solvable system of nonlinear equations, i.e. a certain number of unknowns are expressible in terms of the data and a number of free parameters. The

advantages of this rather uncommon accident are clear: for instance it is possible to go beyond iterative methods of solution, which are very common for nonlinear problems.

An interesting feature is that in dimensions two and three, and for very coarse discretizations, if one is willing to use the zeroth and first moment of time of flight, then all higher order moments depends already on these ones and the recovery of the unknown one-step transition probability matrix is possible up to an (almost) obvious gauge.

The interested reader can consult [GM], [G3] , [G4] for the most recent results as well as other references included in the bibliography for earlier results.

REFERENCES

- [A1] Arridge, S., *Optical tomography in medical imaging*, Inverse Problems **15** (1999), R41–R93.
- [G1] Grünbaum, F. A., *Tomography with diffusion*, in “Inverse Problems in Action”, P. C. Sabatier (ed.), Springer-Verlag, Berlin, pp. 16–21.
- [G2] _____, *Diffuse tomography: The isotropic case*, Inverse Problems **8** (1992), 409–419.
- [G3] _____, *Diffuse tomography: Using time-of-flight information in a two-dimensional model*, International J. of Imaging Technolgy. **11**, (2001) , 283–286.
- [G4] _____, *A nonlinear inverse problem inspired by three dimensional diffuse tomography* Inverse Problems **17** 6, (2001), 1907–1922.
- [GM] Grünbaum,F.A., and Matusevich,L. , *Explicit inversion formulas for a model in diffuse tomography*, to appear in Advances in Applied Mathematics.
- [GP1] Grünbaum, F. A., and Patch, S., *The use of Grassmann identities for inversion of a general model in diffuse tomography*, Proceedings of the Lapland Conference on Inverse Problems, Saariselka, Finland, June 1992.
- [GP2] _____, *Simplification of a general model in diffuse tomography*, in “Inverse Problems in Scattering and Imaging”, M. A. Fiddy (ed.), Proc. SPIE **176**, 744–754.
- [GP3] _____, *How many parameters can one solve for in diffuse tomography?*, Proceedings of the IMA Workshop on Inverse Problems in Waves and Scattering, March 1995.
- [GZ] Grünbaum, F. A., and Zubelli, J., *Diffuse tomography: Computational aspects of the isotropic case*, Inverse Problems **8** (1992), 421–433.
- [NW] Natterer, F., and Wubbeling, F., *Mathematical methods in image reconstruction*, SIAM (2001).
- [P1] Patch, S., *Recursive recovery of a family of Markov transition probabilities from boundary value data*, J. Math. Phys. **36**(7) (July 1995), 3395–3412.
- [P2] _____, *A recursive algorithm for diffuse planar tomography*, Chapter 20 in “Discrete Tomography: Foundations, Algorithms, and Applications”, G. Herman and A. Kuba (eds.), Birkhauser, Boston, 1999.
- [P3] _____, *Recursive recovery of Markov transition probabilities from boundary value data*, Ph.D. thesis, UC Berkeley, 1994.
- [SGKZ] Singer, J., Grünbaum, F. A., Kohn, P., and Zubelli, J., *Image reconstruction of the interior of bodies that diffuse radiation*, Science **248** (1990), 990–993.
- [A] R. Askey, Orthogonal Polynomials and Special Functions, SIAM, (1975).
- [AAR] G. Andrews, R. Askey and R. Roy, Special functions, Encyclopedia of Mathematics and its applications, Cambridge University Press, 1999.

- [C] A. Cormack, Representation of a function by its line integrals, with some radiological applications I, *J. Appl. Physics* **34** (1963), 2722-2727.
- [DG] M. E. Davison and F. A. Grünbaum, Tomographic reconstructions with arbitrary directions, *Comm. Pure and Appl. Math.* **34** (1981), 77-120.
- [GPT] F. A. Grünbaum, I. Pacharoni and J. A. Tirao, Matrix valued spherical functions associated to the complex projective plane, *J. Functional Analysis* **188** (2002) 350-441.
- [GPT1] _____, A matrix valued solution to Bochner's problem, *J. Physics A: Math. Gen.* **34** (2001), 10647-10656.
- [GPT2] _____, An invitation to matrix valued spherical functions: linearization of products in the case of the complex projective space $P_2(\mathbb{C})$, to appear.
- [GPT3] _____, Matrix valued spherical functions associated to the three dimensional hyperbolic space, submitted.
- [GV] R. Gangolli and V. S. Varadarajan, Harmonic analysis of spherical functions on real reductive groups, Springer-Verlag, Berlin, New York, 1988. Series title: *Ergebnisse der Mathematik und ihrer Grenzgebiete*, **101**.
- [HS] Ch. Hamaker and D. Solmon, The angles between the null-spaces of X-rays, *J. Math. Analysis and Its Applications* **62** (1978), 1-23.
- [He] S. Helgason, Groups and geometric analysis, Mathematical Surveys and Monographs, Vol. 83, Amer. Math. Soc., Providence, 2000
- [LS] B. Logan and L. Shepp, Optimal reconstruction of a function from its projections, *Duke Math. J.* (1975), 645-659.
- [T] J. Tirao, Spherical Functions, *Rev. de la Unión Matem. Argentina* **28** (1977), 75-98.

DEPARTAMENT OF MATHEMATICS, UNIVERSITY OF CALIFORNIA, BERKELEY CA 94705
E-mail address: grunbaum@math.berkeley.edu

Mathematical Models for 2D Positron Emission Tomography

B. A. Mair and J. A. Zahnen

ABSTRACT. The standard model for positron emission tomography is a First Kind Fredholm integral equation relating the emission means to the detection means in which the kernel is the probability that an annihilation at a point in image space is detected in a detector tube. This paper contains an overview of recent results on the precise mathematical representation of this kernel and resulting reconstruction algorithms by orthogonal series methods. These algorithms are compared with the standard filtered backprojection (FBP) and expectation maximization maximum likelihood (EMML) algorithms for reconstructing a discontinuous cardiac phantom. The simulations indicate that at least one of the new orthogonal series algorithms produces images with resolution superior to the FBP images, in about 2% of the time required to compute the EMML reconstruction.

1. Introduction

Positron emission tomography (PET) is a medical imaging technique that provides a means for assessing levels of biochemical activity in living tissue [TPRS80]. Using data obtained from PET scanners, reconstruction algorithms generate images containing levels of uptake of radiopharmaceuticals customized for diagnosing the disease of interest. This information is a valuable tool in the diagnosis of tumors, in determining their rates of growth, and in determining the effects of various drugs [LORB97, MW91, TPRS80]. Consequently, the development of accurate PET reconstruction algorithms is an important and ongoing area of research.

To obtain PET data, a radiopharmaceutical is administered to the patient, which is absorbed by sub-regions according to the local levels of metabolic activity. The decaying radioisotope results in a positron which travels a short distance (called positron range) until it is annihilated by a nearby electron, resulting in two 511 KeV photons, which propagate in uniformly random, opposite directions.

In this paper we consider two dimensional (2D) PET in which the scanner ring is viewed as the unit circle \mathbb{T} , and a detector is represented by an arc (of \mathbb{T}), of fixed length α . A detector tube is the region inside the circle, bounded by two distinct arcs (detectors), and the chords joining the ends of these arcs. The (finite) set of all such tubes is denoted by T . The detector tube registers a detection when the two detectors register nearly simultaneous events from an annihilation that

1991 *Mathematics Subject Classification.* Primary 92C55, 65R32; Secondary 31A10, 44A12.

Key words and phrases. Positron emission tomography, PET, potential theory, integral equations, image reconstruction.

© 2002 American Mathematical Society

occurred somewhere in the tube. The scanner data consists of the total number of detections in each tube. This scheme is illustrated in Figure 1(a).

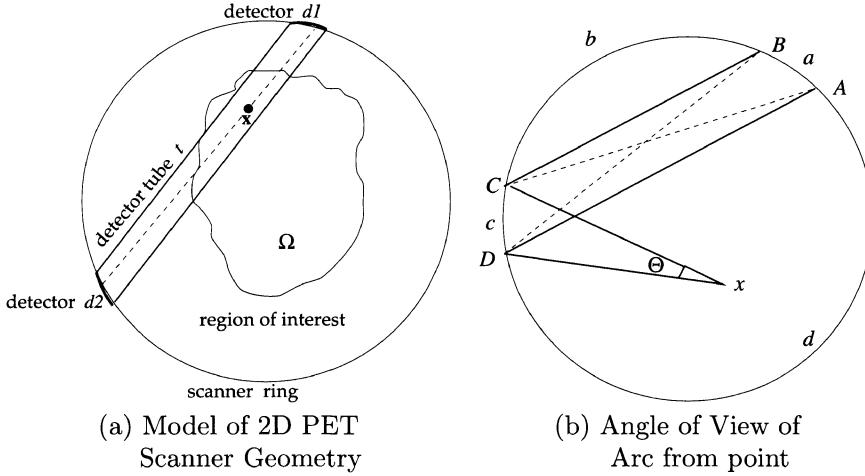


FIGURE 1. Basic Mathematical Model of 2D PET

The point marked “x” in Figure 1(a) represents an annihilation point and the dashed line through x represents the line of flight of the resulting photons. The arcs d_1, d_2 represent two detectors on the scanner ring, and the corresponding detector tube t is the region bounded by d_1, d_2 , and the unbroken lines joining their end points. Since the photons’ line of flight intersect the arcs d_1, d_2 , the scanner records that a detection occurred in the tube t . The region that is being imaged, denoted by Ω , is situated inside the scanner ring and is assumed to be a compact subset of the region interior to T , which is the unit disk \mathbb{D} . In reality, the angle between the photons is not quite 180° but this deviation is not modeled here. We also neglect inaccuracies due to positron range, attenuation, accidental coincidences, and scatter.

In this paper we adopt the basic mathematical model for the ideal emission-detection process in PET that was derived in the seminal work of L. Shepp, Y. Vardi, and L. Kaufman [**SV82**, **VS85**]. In this model, it is assumed that the emissions are governed by a spatial Poisson point process with intensity function $f : \Omega \rightarrow \mathbb{R}$ on the region of interest, Ω . So, for any Borel measurable set $E \subset \Omega$, the number of emissions (or annihilations) in E is a Poisson random variable with mean $\int_E f(\mathbf{x})d\mathbf{x}$. For each $\mathbf{x} \in \mathbb{D}$ and tube $t \in T$, the number of detections in t is a Poisson random variable U_t having mean u_t , and $p(t, \mathbf{x})$ denotes the probability that an emission at \mathbf{x} is detected in t . Then, from [**SV82**, **VS85**]

$$(1.1) \quad u_t = \int_{\Omega} p(t, \mathbf{x})f(\mathbf{x})d\mathbf{x}.$$

The data obtained from a PET scan consists of random samples \tilde{u}_t , $t \in T$, from the independent random variables U_t , $t \in T$, and the PET image reconstruction problem is to estimate the emission intensity function f from this data.

This paper develops reconstruction algorithms based on mathematical representations of p , so it is important to have a precise definition of p . We first define the simpler concept of the angle-of-view of an arc from a point.

DEFINITION 1.1. The angle-of-view of an arc Γ from \mathbf{x} is the angle subtended by Γ at \mathbf{x} , and is denoted here by $\Theta(\Gamma; \mathbf{x})$.

For example, the angle Θ in Figure 1(b) represents the angle-of-view of the arc c from \mathbf{x} . We now extend this definition to tubes. For each $\varphi \in [0, \pi]$, and $\mathbf{x} \in \mathbb{D}$, let $\mathcal{V}(\varphi, \mathbf{x})$ denote the family of all regions V , such that V consists of all points between some pair of lines which intersect at \mathbf{x} and are inclined at the angle φ to each other. Then, there exists a $V_0 \in \mathcal{V}(\varphi, \mathbf{x})$ such that each member of $\mathcal{V}(\varphi, \mathbf{x})$ is a rotation of V_0 .

DEFINITION 1.2. Let the tube t be determined by detectors a, c . For any $\mathbf{x} \in \mathbb{D}$, the angle-of-view of t from \mathbf{x} is the largest angle φ such that $V \cap \mathbb{T} \subset a \cup c$ for some $V \in \mathcal{V}(\varphi, \mathbf{x})$.

Now, $p(t, \mathbf{x})$ is the angle-of-view of t from \mathbf{x} divided by π [SV82]. Hence, if the tube t is determined by the detectors a, c ; and b, d represent the other arcs on \mathbb{T} as illustrated in Figure 1(b), then $p(t, \mathbf{x})$ can be expressed in terms of the angles-of-view of the arcs a, b, c, d as follows.

$$(1.2) \quad p(t, \mathbf{x}) = \begin{cases} \frac{1}{\pi} \min\{\Theta(c; \mathbf{x}), \pi - \Theta(b; \mathbf{x}), \Theta(a; \mathbf{x}), \pi - \Theta(d; \mathbf{x})\}, & \mathbf{x} \in t \\ 0, & \mathbf{x} \notin t \end{cases}$$

If PET detectors are so small that they can be represented by points, then detector tubes become lines, and the PET model equation (1.1) is equivalent to the usual Radon transform model used in computerized tomography (CT). Thus, PET images are often reconstructed by using the filtered back projection (FBP) algorithm for inverting the Radon transform [Dea83, Nat01]. This algorithm is very fast, but usually produces images with poor resolution [LS91]. On the other hand, the statistical, iterative EMML (expectation maximization maximum likelihood) algorithm produces high quality images, but is notoriously slow. Various methods for accelerating it have been proposed, the most notable being the ordered subsets EM (OSEM) algorithm and its variations [HL94, Byr98].

As early as 1982, Hoffman et al [HHPP82] discussed the importance of accurate geometric modeling of the spatial variation of the kernel, which is neglected by the Radon transform model. More recent numerical studies have demonstrated that a more accurate computation of p improves the resolution of PET images [BDB⁺97, QLC⁺98, REFO98, TWH⁺96]. However, these models only produce computational approximations for use in *numerical* reconstruction algorithms and do not provide any insight into the mathematical properties of the operator mapping emissions to detections.

Recently, Mair [Mai00] obtained the first precise mathematical representation of p , which is valid for detectors of arbitrary size. Also, by applying a linear transformation to equation (1.1), Carroll and Mair [CM01a, CM01b] obtained an equivalent model based on detector arcs instead of tubes. The kernels in both models were obtained in terms of the familiar Green's functions and Poisson kernels for the Laplace operator on \mathbb{D} . Section 2 contains a detailed description of these models, and the potential theoretic representations of the kernels involved. It also contains outlines of the FBP and EMML algorithms.

Despite their mathematical complexity, both models can be used to obtain reconstruction algorithms by using classical orthogonal series. The emission intensity function is represented by First Kind Bessel functions and Fourier exponentials, and

the data means are represented by Chebyshev polynomials and Fourier exponentials. Similar methods have been proposed for the image and projection functions in the Radon transform model [Dea83, JS90, Lew90, Lew92, Nat01, SJ77]. These include bases derived from First and Second Kind Bessel functions for the image and Chebyshev functions for the data. Orthogonal series of Chebyshev polynomials have also been employed in reconstructing single photon emission computed tomography images [ZGJL93]. Section 3 contains a brief description of the orthogonal series reconstruction algorithms, called TOS (tube orthogonal series) and AOS (arc orthogonal series).

Section 4 compares images reconstructed from the TOS and AOS algorithms with the EMML and FBP algorithms using noise-free and noisy data generated from a discontinuous emission image. Final comments are made in Section 5.

We express our thanks to the reviewers for their helpful comments which led to improvements in this article.

2. Mathematical Models

We now describe the standard finite dimensional and Radon transform models for PET and the more recent models developed by Mair and Carroll. First, we establish some notation. The detectors are labeled d_0, d_1, \dots, d_{M-1} , in counter-clockwise order, so the set of tubes can be identified with the (unordered) pairs $\{[d_i, d_j] : i \neq j, 0 \leq i, j \leq M-1\}$.

2.1. Finite Dimensional Model. All statistical reconstruction algorithms assume that Ω consists of finitely many disjoint subsets E_1, E_2, \dots, E_N and that on each E_j , f has the constant value f_j and $p(t, \cdot)$ has the constant value p_{tj} . Then, the mean number of emissions in E_j is, $h_j = f_j \times \text{Area of } E_j$, and equation (1.1) reduces to the linear equations

$$(2.1) \quad u_t = (Ph)_t \stackrel{\Delta}{=} \sum_{j=1}^N p_{tj} h_j$$

where P is the matrix (p_{tj}) , and \mathbf{h} is the vector with entries h_1, h_2, \dots, h_N . This model is used to obtain least squares (LS), weighted LS, penalized LS, maximum likelihood (ML), and penalized ML (or maximum a posteriori) estimators of \mathbf{h} [Fes94, GM85, Kau93].

We now describe the EMML (expectation maximization ML) algorithm which is regarded as a gold standard in this area. By using the Poisson distribution, maximizing the data likelihood is equivalent to minimizing $\sum_{t \in T} ((Ph)_t - \tilde{u}_t \log(Ph)_t)$ subject to $\mathbf{h} \geq 0$. The EMML algorithm

$$(2.2) \quad h_j^{(0)} = \frac{1}{Np_j} \sum_{t \in T} \tilde{u}_t \quad \text{and} \quad h_j^{(k+1)} = \frac{h_j^{(k)}}{p_j} \sum_{t \in T} \frac{\tilde{u}_t p_{tj}}{(Ph^{(k)})_t} \quad \text{for } k = 0, 1, \dots$$

where $p_j \stackrel{\Delta}{=} \sum_{t \in T} p_{tj}$, converges to an ML estimate [SV82, VSK85]. However, to reduce the effect of noisy data, this algorithm is either terminated prior to theoretical convergence, or is modified by inserting a smoothing step, or adding a penalty term [EL95, GM85, HL89, SJNW90].

2.2. Radon Transform Model. Throughout this paper, $L(\rho, \phi)$ denotes the chord of \mathbb{T} such that the perpendicular from the origin to $L(\rho, \phi)$ is of length $|\rho|$ and this perpendicular makes an angle of ϕ with the positive x -axis (see also Figure 2). In this model, tubes are assumed to be infinitely thin, and the image f is regarded as a function on \mathbb{R}^2 by extending it to be zero off Ω . Then, the mean number of detections along $L(\rho, \phi)$ is given by

$$(2.3) \quad u(\rho, \phi) = \int_{L_\infty(\rho, \phi)} f(\xi) d\xi, \quad \text{for all } \rho \in \mathbb{R}, 0 \leq \phi \leq \pi$$

where $L_\infty(\rho, \phi)$ is the infinite extension of $L(\rho, \phi)$. The Projection Slice Theorem then states that the Fourier transform of the projections is the two dimensional Fourier transform of the emission intensity along a slice in frequency space [DM84, Her80, Nat01]. The FBP algorithm is a method of computing the exact inversion formula $f = \frac{1}{4\pi} R^* \mathcal{H} u$, where $\mathcal{H} u(s, \phi) \triangleq \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{u(\rho, \phi)}{s - \rho} d\rho$, and $R^* g(x) \triangleq \int_0^\pi g(x \cdot (\cos \phi, \sin \phi), \phi) d\phi$ are the Hilbert and adjoint Radon transforms respectively. Thus the emission image can be obtained by first filtering the data (by an appropriate approximation of the Hilbert transform), then applying the “back-projection” operator. To discuss the recently obtained models of Mair and Carroll, we need to introduce some basic ideas from potential theory.

2.3. Basic Potential Theory. A function u is harmonic if it satisfies the Laplace equation $\Delta u = 0$ on \mathbb{D} and superharmonic if $\Delta u \leq 0$ on \mathbb{D} . The harmonic measure ω_Γ of an arc Γ is the harmonic function which has boundary values 1 on Γ and 0 off of Γ , except at the endpoints of Γ . If Γ has endpoints $e^{i\theta}$, $e^{i\phi}$, with $\theta < \phi$ then

$$(2.4) \quad \omega_\Gamma(z) = \int_\theta^\phi \mathbb{P}(z, s) ds \quad \text{where } \mathbb{P}(z, s) \triangleq \frac{1}{2\pi} \Re \left(\frac{e^{is} + z}{e^{is} - z} \right)$$

is the Poisson kernel and $\Re z$ denotes the real part of z [Gar81, Hel69]. The Green’s function for \mathbb{D} is defined by

$$(2.5) \quad G(z; w) \triangleq \begin{cases} \log |1 - \bar{z}w| / |z - w|, & z \neq w \\ \infty, & z = w \end{cases}$$

where \bar{z} denotes the complex conjugate of z .

2.4. Infinite Dimensional Tube Model. Now, recall that the PET model equation (1.1) is only valid for finitely many tubes. However, the FBP algorithm described in Section 2.2 is based on applying the Fourier transform to an idealization of equation (1.1) in which there are infinitely many tubes and each detector is a point. In order to develop algorithms which are also based on mathematical transforms *without reducing the size of the detectors to points*, we allow the *location* of the individual detectors to be arbitrary.

That is, in our idealization, we disregard the fact that the detectors are at fixed locations on the scanner ring, and assume that data consists of values of the mean detections in the tubes determined by *any pair* of distinct arcs of fixed length α . So, we consider the infinite set of tubes

$$(2.6) \quad T_\infty \triangleq \{[a, b] : a, b \text{ are arcs of length } \alpha\}.$$

Consider a typical detector tube $ABCD$ in Figure 2(a). Then, it is shown in [Mai00] that if the “main diagonal” AC is $L(\rho, \phi)$, then BD is $L(\rho, \phi + \alpha)$ and this

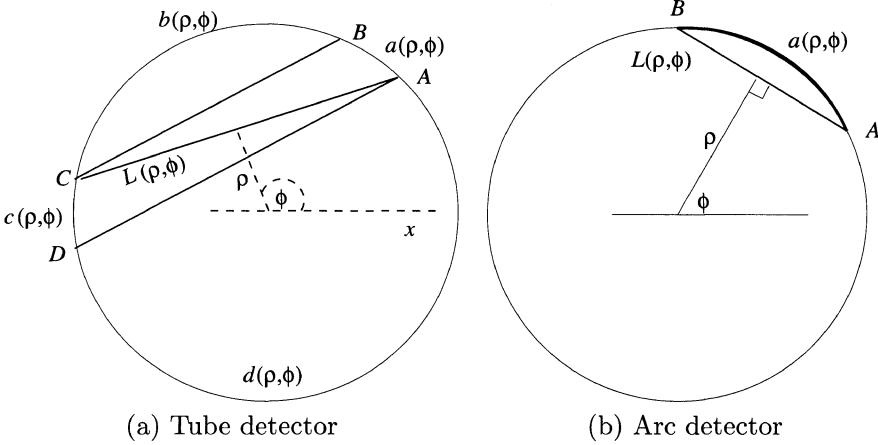


FIGURE 2. Typical tube and arc detectors

leads to a one-to-one correspondence between T_∞ and $[-\cos(\alpha/2), \cos(\alpha/2)] \times [0, \pi]$. Since $L(\rho, \phi) = L(-\rho, \phi + \pi)$, we parameterize T_∞ by

$$\mathbb{Y}_\alpha \stackrel{\Delta}{=} [-\cos(\alpha/2), \cos(\alpha/2)] \times [0, 2\pi).$$

This set will also be used in the arc-detector model to be discussed later. Let $D(\rho, \phi)$ denote the tube corresponding to (ρ, ϕ) . Then for each $(\rho, \phi) \in \mathbb{Y}_\alpha$, the mean number of emissions in the tube $D(\rho, \phi)$ is

$$(2.7) \quad u(\rho, \phi) = \int_{\mathbb{D}} p(\rho, \phi; \mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

where $p(\rho, \phi; \mathbf{x})$ is given by equation (1.2) with $t = D(\rho, \phi)$.

Figure 3 illustrates the three dimensional plot of a typical probability function $p(\rho, \phi; (x, y))$, for a fixed tube $D(\rho, \phi)$ of length 1 ($0 \leq x \leq 1$) and width 0.1 ($0 \leq y \leq 0.1$). Two dimensional cross sections are shown in Figure 4. Figure 4(a), (b) show cross-sections parallel to the sides of the tubes, in the middle (by the plane $y = 0.05$), and approximately halfway between the middle and one side (by the plane $y = 0.02$), respectively. Figure 4(c), (d) show cross-sections perpendicular to the sides of the tubes, in the middle (by the plane $x = 0.5$), and approximately halfway between the middle and one end (by the plane $x = 0.2$), respectively.

These figures clearly demonstrate that, as a function of $\mathbf{x} = (x, y)$, $p(\rho, \phi, \mathbf{x})$ is neither convex nor concave, is continuous on \mathbb{D} , and is not differentiable on the diagonals and sides of $D(\rho, \phi)$. This is proved mathematically by the following result obtained in [Mai00].

THEOREM 2.1. *For each tube $D(\rho, \phi)$, there exists a signed Borel measure $\mu_{\rho, \phi}$ supported on $L(\rho, \phi) \cup L(\rho, \phi + \alpha) \cup L(a(\rho), \phi + \alpha/2) \cup L(b(\rho), \phi + \alpha/2)$, such that*

$$p(\rho, \phi; \mathbf{x}) = \int_{\mathbb{D}} G(\mathbf{x}, \mathbf{y}) d\mu_{\rho, \phi}(\mathbf{y}) + \frac{\alpha}{2\pi} (\omega_{a(\rho, \phi)}(\mathbf{x}) + \omega_{c(\rho, \phi)}(\mathbf{x})).$$

where $\theta_\rho \stackrel{\Delta}{=} \arccos \rho$, $a(\rho) \stackrel{\Delta}{=} \cos(\theta_\rho - \alpha/2)$, $b(\rho) \stackrel{\Delta}{=} \cos(\theta_\rho + \alpha/2)$.

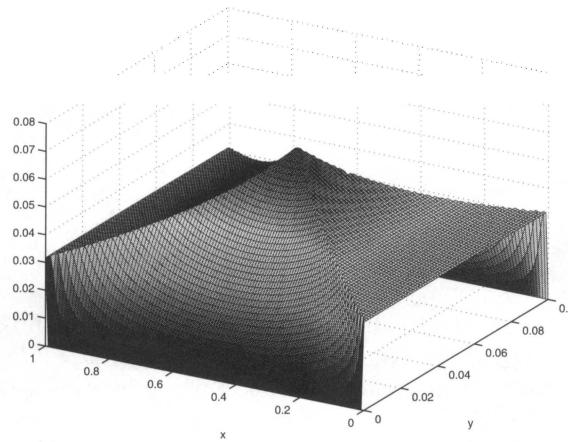


FIGURE 3. Plot of probability of detection in a typical tube

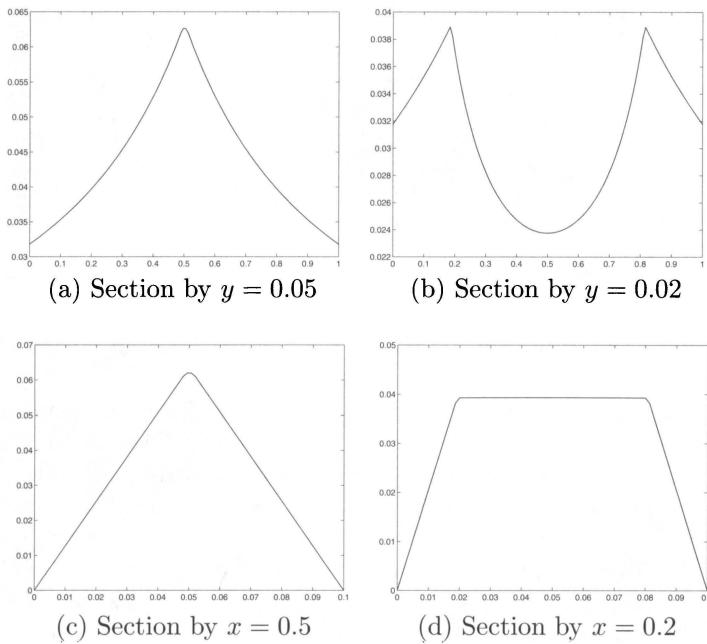


FIGURE 4. Cross sections of probability $p(t, (x, y))$. (a) Section by plane $y = 0.05$ (b) Section by plane $y = 0.02$ (c) Section by plane $x = 0.5$ (d) Section by plane $x = 0.2$

Let $a(\rho, \phi)$, $c(\rho, \phi)$ be the arcs indicated in Figure 2(a) and define

$$K(\rho, \mathbf{x}) \triangleq \frac{2}{\pi} \frac{\sqrt{1 - \rho^2}}{1 - |\mathbf{x}|^2}.$$

The following representation of p is obtained in [Mai00].

THEOREM 2.2. *For each $(\rho, \phi) \in \mathbb{Y}_\alpha$ and $\mathbf{x} \in \mathbb{D}$,*

$$\begin{aligned} p(\rho, \phi; \mathbf{x}) = & \int_{L(\rho, \phi) \cup L(\rho, \phi+\alpha)} G(\mathbf{x}, \mathbf{y}) K(\rho, \mathbf{y}) d\mathbf{y} \\ & - \int_{L(a(\rho), \phi+\alpha/2)} G(\mathbf{x}, \mathbf{y}) K(a(\rho), \mathbf{y}) d\mathbf{y} \\ & - \int_{L(b(\rho), \phi+\alpha/2)} G(\mathbf{x}, \mathbf{y}) K(b(\rho), \mathbf{y}) d\mathbf{y} \\ & + \frac{\alpha}{2\pi} (\omega_{a(\rho, \phi)}(\mathbf{x}) + \omega_{c(\rho, \phi)}(\mathbf{x})). \end{aligned}$$

This more accurate model shows that the mathematical properties of the projection operator mapping emission to detection means in PET are quite different from those indicated by the Radon transform approximation in equation (2.3). To demonstrate this, consider a simple example of an emission intensity function which is a non-zero constant on the rectangle R and 0 outside of R , illustrated in Figure 5.

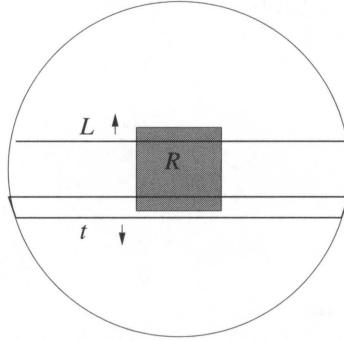


FIGURE 5. Characteristic function example

Consider the mean numbers of emissions along the line L , in the Radon transform model, and in the more realistic tube t in this model. Then, as the line L moves in the direction of the arrow, the corresponding detections drop off suddenly to zero as L crosses the boundary of R . However, for the nontrivial tubes t , the emissions vary continuously as the tube t moves in the direction of the arrow, crossing the boundary of R . From [Mai00], we know that the means for the nontrivial tubes not only vary continuously (with respect to the tubes), but are actually Lipschitz continuous. We now describe how the usual *finite* PET data is related to the ideal infinite dimensional function $u(\rho, \phi)$ in equation (2.7).

Assuming that the arc d_k has end points $e^{2\pi ik/M}$, $e^{2\pi i(k+1)/M}$, the tube $[d_j, d_k]$ corresponds to $L(\rho_{jk}, \phi_{jk})$ where

$$(2.8) \quad \rho_{jk} \triangleq \cos(\pi(k-j)/M), \quad \phi_{jk} \triangleq \pi(k+j)/M.$$

and the corresponding tube detections satisfy

$$(2.9) \quad \tilde{u}_{jk} \approx u(\rho_{jk}, \phi_{jk}), \text{ for all } 0 \leq j < k \leq M-1.$$

Hence, we have achieved an infinite dimensional extension of the model in equation (1.1) which has the same parameterization as that for the Radon transform model (2.3), without reducing the size of the individual PET detectors. It seems

that this is a more natural idealization since now the finite dimensional data can be regarded as a genuine sample of values from the infinite dimensional idealization, preserving the true character of PET detectors. We now discuss a model using arc-based detectors rather than the usual detector tubes.

2.5. Infinite Dimensional Arc Model. Consider the arcs obtained by combining all possible *contiguous* arcs from the set of detectors $\{d_0, d_1, \dots, d_{M-1}\}$. By using arithmetic modulo M in the detector indices, each such arc can be represented by an ordered k -tuple of detectors starting from an initial arc as follows

$$(2.10) \quad a_{jk} \stackrel{\Delta}{=} [d_j, d_{j+1}, \dots, d_{j+k-1}]$$

for some $0 \leq j < M, 1 \leq k \leq M$. For notational convenience, we also extend the actual PET tube data \tilde{u}_{jk} from the indices $0 \leq j < k < M$ to $0 \leq j, k < 2M$ by using arithmetic modulo M and setting $\tilde{u}_{kk} = 0$. Then, we consider the set of arc-detectors

$$\mathcal{A} \stackrel{\Delta}{=} \{a_{jk} : 0 \leq j < M, 1 \leq k \leq M\}$$

instead of the set of tubes T_∞ . To each arc-detector a_{jk} , we define the random “number of detections” in a_{jk} by $V_{jk} \stackrel{\Delta}{=} \sum_{t \in T_{jk}} U_t$, where T_{jk} denotes the set of all tubes having at least one detector in a_{jk} . That is, the number of detections in an arc-detector is the total number of detections in the tubes which have at least one end in the arc-detector. So, an emission at \mathbf{x} is “detected” in a_{jk} if at least one end of the resulting photon path intersects a_{jk} . Hence we obtain the following finite arc-detector model.

$$(2.11) \quad v_{jk} = \int_{\mathbb{D}} q_{jk}(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

where v_{jk} is the mean of V_{jk} , and $q_{jk}(\mathbf{x})$ is the probability that an emission at \mathbf{x} is detected in a_{jk} .

To obtain an infinite dimensional extension of equation (2.11), observe that the (finitely many) arc-detectors in \mathcal{A} consist of arcs whose lengths are integer multiples of α with fixed initial point $e^{2\pi i j/M}$ and final point $e^{2\pi i(j+k)/M}$. So, we consider the extension to the case in which the ideal data consists of the number of detections in all possible arcs of length at least α , with unrestricted location. As in the tube model, it is clear that this set of arc-detectors is parameterized by \mathbb{Y}_α . Figure 2(b) illustrates the arc-detector $a(\rho, \phi)$ which corresponds to (ρ, ϕ) . Observe that each line $L(\rho, \phi)$ determines two arc-detectors, $a(\rho, \phi)$ and $\mathbb{T} \setminus a(\rho, \phi)$. That explains why we need to use angles in $[0, 2\pi)$ rather than just angles in $[0, \pi]$ to represent arc-detectors. So each $(\rho, \phi) \in \mathbb{Y}_\alpha$, determines an arc-detector, denoted by $a(\rho, \phi)$, in which the mean number of emissions, $v(\rho, \phi)$, satisfies

$$(2.12) \quad v(\rho, \phi) = \int_{\mathbb{D}} q(\rho, \phi; \mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

where $q(\rho, \phi; \mathbf{x})$ is the probability that a uniformly random line through \mathbf{x} intersects $a(\rho, \phi)$. Then elementary geometry shows that

$$(2.13) \quad q(\rho, \phi; \mathbf{x}) = \min \left\{ 1, \frac{1}{\pi} \Theta(a(\rho, \phi); \mathbf{x}) \right\}.$$

As in the detector tube model, we use potential theory to obtain the following representations for q [CM01a, CM01b]. Observe the simplicity of this representation compared to the tube model in Theorem 2.2.

THEOREM 2.3. *For each $(\rho, \phi) \in \mathbb{Y}_\alpha$*

$$\begin{aligned} q(\rho, \phi; \cdot) &= \int_{\mathbb{D}} G(\cdot, \mathbf{y}) d\nu_{\rho, \phi}(\mathbf{y}) + \omega_{a(\rho, \phi)} + (1 - \omega_{a(\rho, \phi)})\theta_\rho/\pi \\ &= \int_{L(\rho, \phi)} G(\cdot, \mathbf{y}) K(\rho, \mathbf{y}) d\mathbf{y} + \omega_{a(\rho, \phi)} + (1 - \omega_{a(\rho, \phi)})\theta_\rho/\pi \end{aligned}$$

where $\nu_{\rho, \phi}$ is a positive Borel measure supported on $L(\rho, \phi)$ and $a(\rho, \phi)$ is the corresponding arc-detector.

By the definition of V_{jk} , an approximation \tilde{v}_{jk} of the true mean v_{jk} , can be computed from the actual PET tube data \tilde{u}_t , $t \in T$, by

$$(2.14) \quad \tilde{v}_{jk} \stackrel{\Delta}{=} \sum_{t \in T_{jk}} \tilde{u}_t$$

and is related to the infinite dimensional arc-detector mean function v by

$$(2.15) \quad \tilde{v}_{jk} \approx v(\cos(\pi k/M), \pi(j+k)/M).$$

3. Reconstruction Method

In this section we describe the TOS (tube orthogonal series), and AOS (arc orthogonal series) algorithms, which are based on the tube and arc models described in Sections 2.4 and 2.5 respectively. We define $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ and consider the polar representation for points in, and functions on, \mathbb{D} . To treat both models simultaneously, let Q denote either kernel p or q , and consider the model equation

$$g(\rho, \phi) = \int_0^{2\pi} \int_0^1 Q(\rho, \phi; r, \theta) f(r, \theta) r dr d\theta$$

where g denotes either u or v .

Unlike the case of the Radon transform model, taking the Fourier transform (in ρ) of $g(\rho, \phi)$ does not produce any useful relationship between the data function g and the image f . To motivate our choice of transforms, we extend the domain of the data function g from \mathbb{Y}_α to all of $[-1, 1] \times [0, 2\pi]$ by assigning it the value 0 off \mathbb{Y}_α . Note that the term $\sqrt{1 - \rho^2}$ in the function K_ρ , which occurs in the formulas for p, q in Theorems 2.2, 2.3, is exactly the weight function for the orthogonality relationships for the Chebyshev polynomials, which are orthogonal on $[-1, 1]$ [Riv90]. Of course, the dependence of $g(\rho, \phi)$ on ϕ is naturally represented by the usual trigonometric series. So it is quite reasonable to use the product of Chebyshev polynomials in ρ and trigonometric series in ϕ to represent $g(\rho, \phi)$.

For each $n \in \mathbb{N}_0$, let C_n , J_n denote the Chebyshev polynomial of degree n [Riv90], and the Bessel function of the first kind of order n [Bow58, Wat62] respectively. The positive zeros of J_n are labeled $\tau_{n,m}$, $m \in \mathbb{N}_0$ with $0 < \tau_{n,m} < \tau_{n,m+1}$. For each $k \in \mathbb{Z}$, $m \in \mathbb{N}_0$, $\theta \in [0, 2\pi]$, $\rho \in [-1, 1]$ and $r > 0$ define

$$\begin{aligned} C_{k,m}(\rho, \theta) &\stackrel{\Delta}{=} \frac{1}{\pi \varepsilon_m} C_m(\rho) e^{ik\theta} \text{ where } \varepsilon_0 \stackrel{\Delta}{=} \sqrt{2} \text{ and } \varepsilon_m \stackrel{\Delta}{=} 1 \text{ for } m \geq 1 \\ J_{k,m}(r, \theta) &\stackrel{\Delta}{=} B_m(r) e^{ik\theta} \text{ where } B_m(r) \stackrel{\Delta}{=} \frac{J_{|k|}(\tau_{|k|,m} r)}{\sqrt{\pi} J_{|k|+1}(\tau_{|k|,m})}. \end{aligned}$$

Then, by using standard orthogonality properties [Bow58, Riv90] it is easy to see that $\{J_{k,m} : k \in \mathbb{Z}, m \in \mathbb{N}\}$ and $\{C_{k,m} : k \in \mathbb{Z}, m \in \mathbb{N}\}$ are orthonormal bases of $L^2(\mathbb{D})$ and $L^2(\mathbb{Y}, w)$ respectively, where the weight function is $w(\rho) = 1/\sqrt{1-\rho^2}$ [Car98, Mai00]. So for any $f \in L^2(\mathbb{D})$ and $g \in L^2(\mathbb{Y}, w)$

$$(3.1) \quad f = \sum_{k=-\infty}^{\infty} \sum_{m=0}^{\infty} f_{k,m} J_{k,m} \text{ and } g = \sum_{k=-\infty}^{\infty} \sum_{m=0}^{\infty} g_{k,m} C_{k,m}$$

where

$$(3.2) \quad f_{k,m} \triangleq \int_0^{2\pi} \int_0^1 f(r, \theta) \bar{J}_{k,m}(r, \theta) r dr d\theta$$

$$(3.3) \quad g_{k,m} \triangleq \int_0^{2\pi} \int_{-1}^1 g(\rho, \phi) \bar{C}_{k,m}(\rho, \phi) w(\rho) d\rho d\phi$$

Equations (2.4), (2.5) imply that

$$\mathbb{P}(ze^{-i\phi}, s) = \mathbb{P}(z, se^{i\phi}) \text{ and } G(ze^{i\phi}; we^{i\phi}) = G(z; w)$$

for all $z, w \in \mathbb{T}, s \in \mathbb{T}, \phi \in [0, 2\pi]$. So, $Q(\rho, \phi; r, \theta) = Q(\rho, \phi - \theta; r, 0)$. Hence the representations in Theorems 2.2 and 2.3 show that

$$(3.4) \quad g_{k,m} = \sum_{n=0}^{\infty} Q_{k,m,n} f_{k,n} \text{ for all } k \in \mathbb{Z}, m \in \mathbb{N}_0$$

where

$$(3.5) \quad Q_{k,m,n} \triangleq 2\pi \int_0^1 \int_{-1}^1 \int_0^{2\pi} Q(\rho, \phi; r, 0) B_n(r) \frac{\bar{C}_{k,m}(\rho, \phi)}{\sqrt{1-\rho^2}} d\phi d\rho r dr$$

So, the emission intensity function f can be estimated by solving the linear systems in equation (3.4) for the coefficients $f_{k,n}$ and using these to estimate f by equation (3.1). Equation (3.4) can be written as matrix equations

$$(3.6) \quad Q^{(k)} \mathbf{f}^{(k)} = \mathbf{g}^{(k)}, \quad k \in \mathbb{Z}$$

where $Q_{mn}^{(k)} \triangleq Q_{k,m,n}$, $\mathbf{g}_m^{(k)} \triangleq g_{k,m}$, and $\mathbf{f}_n^{(k)} \triangleq f_{k,n}$. From equations (3.5), (3.2) and (3.3) it is clear that $Q^{(-k)}$, $\mathbf{f}^{(-k)}$ and $\mathbf{g}^{(-k)}$ are the complex conjugates of $Q^{(k)}$, $\mathbf{f}^{(k)}$ and $\mathbf{g}^{(k)}$ respectively, so equation (3.6) can be written as a single matrix equation

$$(3.7) \quad Q\mathbf{f} = \mathbf{g}$$

where Q is the block diagonal matrix whose k^{th} block is $Q^{(k)}$, \mathbf{f} is the concatenation of the vectors $\mathbf{f}^{(k)}$ for $k \in \mathbb{N}_0$, and \mathbf{g} is similarly defined. Now, to compute the entries in the system matrix Q , note that the Green's function and Poisson kernel can be expressed in terms of Bessel functions [Roa82], [Mai00, Lemma 6.3]. Using properties of Bessel functions and Chebyshev polynomials, it is shown in [CM01a, Mai00] that $Q_{k,m,n}$ can be rapidly and accurately computed by Simpson's rule for numerical integration. By using equations (2.9), (2.14) and (2.15), and properties of the Chebyshev polynomials, results in [CM01a], show that the data vector \mathbf{g} is easily estimated by applying the Fast Fourier Transform to the tube data.

Now, the ill-posedness of the original problem shows up as an ill-conditioning of each diagonal block $Q^{(k)}$. In our experiments, we found that the standard method of TSVD (truncated singular value decomposition) [Vog86, Wah77] produced

low quality images, and was very unstable and sensitive, relative to the noise and truncation levels. Now, it has been observed that applying the CG (conjugate gradient) method to solve an ill-conditioned system results in the formation of low frequency components of the solution in early iterations, with the high frequency components appearing in later iterates [Pla90, SV90, Vog87]. This has been made more precise in [Han92] which shows that the i^{th} iterate of the CG algorithm can be expressed as a regularized solution of a least squares problem, where the regularization constrains the solution to be in a Krylov subspace determined by the system matrix. Thus, the CG algorithm may be viewed as a regularization technique in which the degree of regularization is controlled by the number of iterations. Since each CG iterate is rapidly computed from the preceding iterate, it is a very convenient regularization method for solving our problem.

We found that the entries past the first 80×80 principal minor in each matrix $Q^{(k)}$ all computed to be zero, so our algorithms are based on solving the first 80 equations in the first 80 blocks of Q . The TOS and AOS algorithms first smooth the data by applying the standard Hann filter as used in clinical scans, with cut-off frequency 35% of the Nyquist frequency (i.e. the reciprocal of the sample grid size). Then, we solve the linear system in equation (3.7) by using the conjugate gradient (CG) method applied to the normal equation

$$Q^T Q f = Q^T g$$

terminating prior to theoretical convergence, depending on a visual inspection of the corresponding reconstructed images. This is the same stopping criterion applied in the EMML algorithm.

4. Numerical Simulations

In this section we compare the performance of the TOS and AOS reconstruction algorithms with the standard EMML and FBP algorithms outlined in Section 2. The images were determined to be the best reconstructions from a visual inspection of each iterate. The true image is illustrated in Figure 6 and consists of a computer generated model of a slice of the thorax containing sections of the heart, lungs, soft tissue, and spinal column. We refer to this as the cardiac phantom. In all the images here, the darker regions correspond to higher concentrations of the radiopharmaceutical. Thus, the heart (horse-shoe shaped region) in Figure 6 has the highest concentration, and the spine (circular section at base) has the lowest concentration. This image represents a discontinuous emission intensity function f , which is constant on pixels, hence, due to the Gibbs property for Fourier series, we expect the images reconstructed by the orthogonal series methods TOS and AOS, to suffer from ringing artifacts caused by the inability of the approximation to capture the sharp jumps at the discontinuities. However, the images reconstructed from the EMML algorithm should have no such artifacts since it is based on a pixel representation of the emission intensity. Another advantage of the EMML algorithm is due to the manner in which the data is generated.

It is important to note that the data that is referred to as “noise-free” here is generated by using the finite dimensional model equation (2.1) with 128 detectors, which approximates p by a function which is constant on pixels. Thus the noise-free data does not exactly fit the true PET model on which the orthogonal series reconstruction method is based, but exactly fits the approximate model on which the EMML algorithm is based. To reduce the model mismatch effects, the

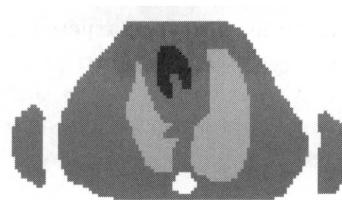


FIGURE 6. Cardiac phantom

data was generated on a finer grid (256×256) than the grid that was used for reconstructions (128×128). This provides an unbiased, conservative appraisal of the new orthogonal series methods since the simulations favor the standard EMML method. The noisy data is obtained by generating pseudo-random samples from Poisson distributions having the means in the noise-free data. Hence the noisy data not only contains Poisson noise, but also does not match our presumed continuous model. Figure 7(a),(b) contain the noise-free and noisy data respectively, arranged in the usual sinogram format [OF97].

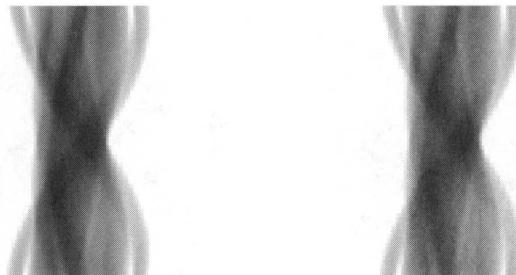


FIGURE 7. Detector tube data for cardiac phantom

Reconstructions using the EMML, FBP, TOS, and AOS algorithms are shown in Figures 8 and 9. Figure 8(a), (b), (c), and (d), contain the images obtained from the noise-free data in Fig 7(a) using the EMML, FBP, TOS, and AOS algorithms, respectively. Figure 9(a), (b), (c), and (d), contain the images obtained from the noisy data in Fig 7(b) using the EMML, FBP, TOS, and AOS algorithms, respectively. As expected, the EMML reconstructions are superior to the other methods. From Figures 8(c),(d), and 9(c),(d) we see that the TOS algorithm is able to recover the heart shape from both noise-free and noisy data, and the AOS algorithm captures the correct shape in the noise-free case, but not in the noisy case. Figures 8(c), and 9(b) indicate that the FBP algorithm is unable to recover the correct shape of the heart region, even in the noise-free data case. This is probably due to

the inability of the Radon transform approximation to model the spatially varying detector blur. The artifacts in the TOS and AOS reconstructions are clearly due to noise in the data. This could probably be remedied by using a different smoothing filter on the data, and by smoothing the reconstructed images. No such smoothing was performed on these images. Also, note that the TOS images are better than the AOS images. We suspect this is partly due to the fact that the arc-detector data has a higher variance since it is obtained by summing portions of the tube data.

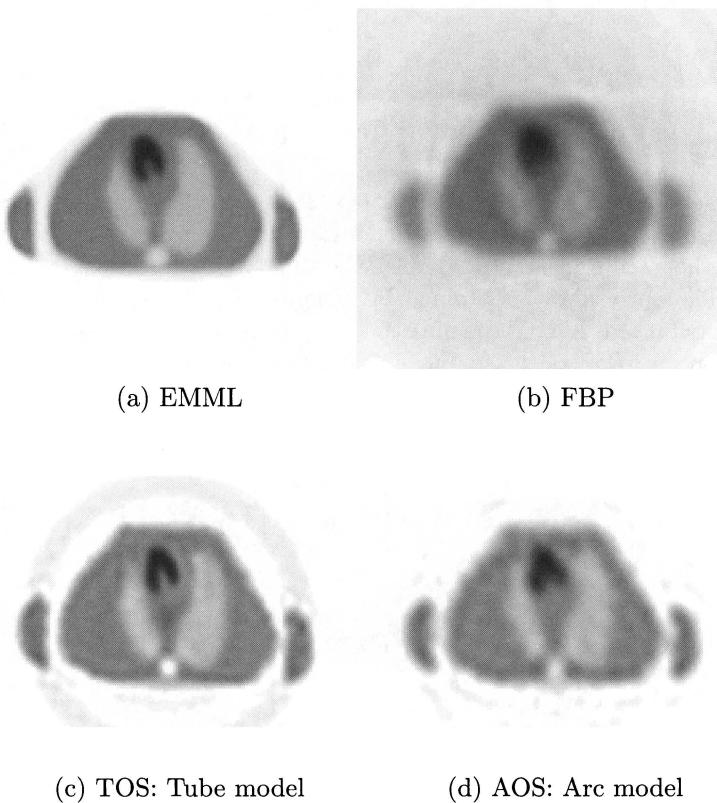


FIGURE 8. Reconstructions of cardiac phantom from noise-free data.

The images reconstructed from the TOS and AOS algorithms are formed from approximately 800 parameters, whereas the EMML algorithm requires over 16,000 parameters since it computes the value for each pixel. This contributes to the dramatic difference in speed between the two methods. The orthogonal series images here required only an average of 10 seconds of processing time, whereas the EMML images required 520 seconds. This indicates that it is feasible to obtain PET images comparable in quality to those produced by the very slow EMML algorithm in about the same time as the FBP algorithm.

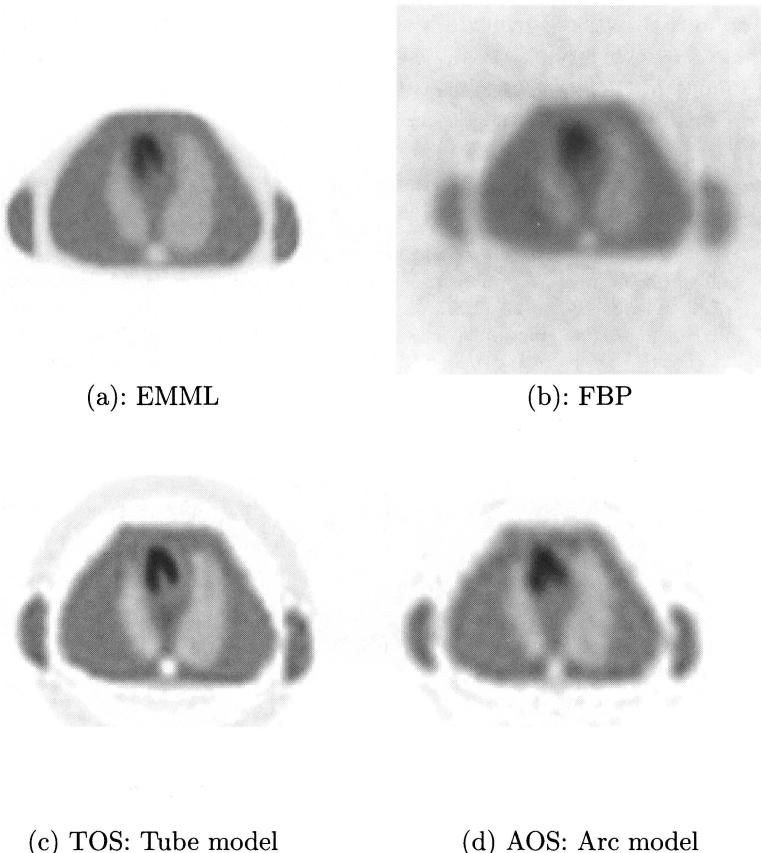


FIGURE 9. Reconstructions of cardiac phantom from noisy data.

5. Conclusion

This paper discusses two recently developed mathematical models for 2D PET, the resulting reconstruction algorithms TOS and AOS, and compares these algorithms with the usual FBP and EMML algorithms.

The simulations in this paper indicate that, on computer generated images and data, the TOS algorithm produces images which are superior in quality to those reconstructed by the FBP algorithm, and require only about 2% of the time required for the EMML reconstructions. Of the four methods, the AOS algorithm produced images with the most artifacts and a resolution comparable to the FBP. Although the TOS images contain more artifacts than the EMML images it is important to note the following.

- The smoothness properties of the data function underlying the TOS and FBP algorithms are very different (see Section 2.4).
- The TOS and FBP algorithms use different bases to represent the data and reconstructed images (see Sections 2.2,3).
- The noisy data was smoothed by the *same* filter prior to being used in the FBP and TOS algorithms.
- The TOS images were not filtered after reconstruction.

These observations indicate that it may be possible to improve the quality of the TOS images by using appropriate filters both on the noisy data and on the reconstructed image. This problem is currently being investigated.

References

- [BDB⁺97] G. Brix, J. Doll, M. E. Bellemann, H. Trojan, U. Haberkorn, P. Schmidlin, and H. Ostertag, *Use of scanner characteristics in iterative image reconstruction for high resolution positron emission tomography studies of small animals*, European Journal of Nuclear Medicine **24** (1997), no. 7, 779–786.
- [Bow58] F. Bowman, *Introduction to Bessel Functions*, Dover Publications, New York, 1958.
- [Byr98] C. Byrne, *Accelerating the EMML algorithm and related iterative algorithms by rescaled block-iterative methods*, IEEE Trans. Image Processing **7** (1998), 100–109.
- [Car98] R. B. Carroll, *An orthogonal series approach to positron emission tomography*, Ph.D. thesis, University of Florida, Gainesville, FL, October 1998.
- [CM01a] R. B. Carroll and B. A. Mair, *A new model and reconstruction method for 2D PET based on transforming detector tube data into detector arc data*, Journal of Mathematical Imaging and Vision **14** (2001), 165–185.
- [CM01b] ———, *Orthogonal series approach to reconstructing 2D PET images using data obtained from detector tubes of arbitrary width*, Mathematical Modeling, Estimation and Imaging, Proceedings of SPIE, vol. 41, SPIE, 2001, pp. 191–201.
- [Dea83] S. R. Deans, *The Radon Transform and Some of Its Applications*, John Wiley & Sons, Inc., 1983.
- [DM84] D. E. Dudgeon and R. M. Mersereau, *Multidimensional Digital Signal Processing*, Prentice Hall, New Jersey, 1984.
- [EL95] P. P. B. Eggermont and V. N. LaRiccia, *Maximum smoothed likelihood density estimation for inverse problems*, The Annals of Statistics **23** (1995), 199–220.
- [Fes94] J. A. Fessler, *Penalized weighted least-squares image reconstruction for positron emission tomography*, IEEE Trans. Med. Imag. **13** (1994), 290–300.
- [Gar81] J. B. Garnett, *Bounded Analytic Functions*, Academic Press, New York, 1981.
- [GM85] S. Geman and D. E. McClure, *Bayesian image analysis: An application to single photon emission tomography*, Proc. Statist. Comput. Sect., Amer. Statist. Assn. (1985), 12–18.
- [Han92] P. C. Hansen, *Regularization tools: A Matlab package for analysis and solution of discrete ill-posed problems*, Tech. Report UNIC-92-04, UNIC, June 1992.
- [Hel69] L. L. Helms, *Introduction to Potential Theory*, Wiley-Interscience, New York, 1969.
- [Her80] G. T. Herman, *Image Reconstruction from Projections: The Fundamentals of Computerized Tomography*, Academic Press, San Francisco, 1980.
- [HHPP82] E. J. Hoffman, S-C Huang, D. Plummer, and M. E. Phelps, *Quantitation in positron emission computed tomography: 6. effect of nonuniform resolution*, Jour. Computer Assisted Tomography **6** (1982), no. 5, 987–999.
- [HL89] T. Hebert and R. Leahy, *A generalized EM algorithm for the 3-D Bayesian reconstruction from Poisson data using Gibbs priors*, IEEE Trans. Med. Imag. **8** (1989), 194–202.
- [HL94] H. Hudson and R. Larkin, *Accelerated image reconstruction using ordered subsets of projection data*, IEEE Trans. Med. Imag. **13** (1994), no. 4, 601–609.
- [JS90] I. M. Johnstone and B. W. Silverman, *Speed of estimation in positron emission tomography and related inverse problems*, The Annals of Statistics **18** (1990), no. 1, 251–280.
- [Kau93] L. Kaufman, *Maximum likelihood, least squares, and penalized least squares for PET*, IEEE Trans. Med. Imag. **12** (1993), 200–214.
- [Lew90] R. M. Lewitt, *Multidimensional digital image representations using generalized Kaiser-Bessel window functions*, J. Opt. Soc. Am. A **7** (1990), no. 10, 1834–1846.
- [Lew92] ———, *Alternatives to voxels for image representation in iterative reconstruction algorithms*, Phys. Med. Biol. **37** (1992), no. 3, 705–716.
- [LORB97] J. D. Lane, E. C. Opara, J. E. Rose, and F. M. Behm, *Quitting smoking raises whole blood glutathione*, Physiology and Behavior (1997).

- [LS91] J. S. Liow and S. C. Strother, *Practical tradeoffs between noise, quantitation, and number of iterations for maximum likelihood-based reconstructions*, IEEE Trans. Med. Imag. **10** (1991), no. 4, 563–571.
- [Mai00] B. A. Mair, *A mathematical model incorporating the effects of detector width in 2D PET*, Inverse Problems **16** (2000), 223–246.
- [MW91] R. J. Mathew and W. H. Wilson, *Substance abuse and cerebral blood flow*, American J. Psychiatry **148** (1991), 292–305.
- [Nat01] F. Natterer, *The Mathematics of Computerized Tomography*, SIAM, Philadelphia, 2001.
- [OF97] J. M. Ollinger and J. A. Fessler, *Positron Emission Tomography*, IEEE Signal Processing Magazine (1997), 43–55.
- [Pla90] R. Plato, *Optimal algorithms for linear ill-posed problems yield regularization methods*, Numer. Funct. Anal. Optimiz. **11** (1990), 111–118.
- [QLC⁺98] J. Qi, R. M. Leahy, S. R. Cherry, A. Chatzioannou, and T. H. Farquhar, *High-resolution 3D Bayesian image reconstruction using the microPET small-animal scanner*, Phys. Med. Biol. **43** (1998), 1001–1013.
- [REFO98] A. J. Reader, K. Erlandsson, M. A. Flower, and R. J. Ott, *Fast accurate iterative reconstruction for low-statistics positron volume imaging*, Phys. Med. Biol. **43** (1998), 835–846.
- [Riv90] T. J. Rivlin, *Chebyshev Polynomials: From Approximation Theory to Algebra and Number Theory*, second ed., John Wiley & Sons, Inc., New York, 1990.
- [Roa82] G. F. Roach, *Green's functions*, second ed., Cambridge University Press, London, 1982.
- [SJ77] D. L. Snyder and Jr J.R.Cox, *An overview of reconstructive tomography and limitations imposed by a finite number of projections*, Reconstructive Tomography in Diagnostic Radiology and Nuclear Medicine (M. M. Ter-Pogossian et al, ed.), University Park Press, Baltimore, MD, 1977, pp. 3–32.
- [SJNW90] B. W. Silverman, M. C. Jones, D. W. Nychka, and J. D. Wilson, *A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography*, J. R. Statist. Soc. Series B, **52** (1990), 271–324.
- [SV82] L. A. Shepp and Y. Vardi, *Maximum likelihood reconstruction for emission tomography*, IEEE Trans. Med. Imag. **1** (1982), 113–122.
- [SV90] A. Van Der Sluis and H. A. Van Der Vorst, *SIRT and CG methods for the iterative solutions of sparse linear least squares problems*, Lin. Alg. Appl. **130** (1990), 76–83.
- [TPRS80] M. M. Ter-Pogossian, M. E. Raichle, and B. E. Sobel, *Positron emission tomography*, Scientific American **243** (1980), 170–181.
- [TWH⁺96] A. Terstegge, S. Weber, H. Herzog, H. W. Müller-Gärtner, and H. Halling, *Resolution and better quantification by tube of response modelling in 3D PET reconstruction*, 1996 IEEE Nuclear Science Symposium, 1996.
- [Vog86] C. Vogel, *Optimal choice of a truncation level for the truncated SVD solution of linear first kind integral equations when the data are noisy*, SIAM J. Numer. Anal. **23** (1986), 109–117.
- [Vog87] C. R. Vogel, *Solving ill-conditioned linear systems using the conjugate gradient method*, Tech. report, Department of Mathematical Sciences, Montana State University, 1987.
- [VSK85] Y. Vardi, L. A. Shepp, and L. Kaufman, *A statistical model for positron emission tomography*, Journal of the American Statistical Association **80** (1985), 8–37.
- [Wah77] G. Wahba, *Practical approximate solutions to linear operator equations when the data are noisy*, SIAM J. Numer. Anal. **14** (1977), 651–667.
- [Wat62] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, Cambridge University Press, London, 1962.
- [ZGJL93] G. L. Zeng, G. T. Gullberg, R. J. Jaszczak, and J. Li, *Fan-Beam reconstruction algorithm for a spatially varying focal length collimator*, IEEE Trans. Med. Imag. **12** (1993), no. 3, 575–582.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF FLORIDA, GAINESVILLE, FL 32611
E-mail address: bam@math.ufl.edu

This page intentionally left blank

Explicit versus Implicit Relative Error Regularization on the Space of Functions of Bounded Variation

O. Scherzer

ABSTRACT. The noise detected with common measurement devices frequently correlates with the data. To denoise signals and images with correlated noise we propose to use relative error regularization models. A framework is developed for which the analysis is tractable. We put these regularization concepts in connection with morphological differential equations, which can be considered as natural limits of regularization techniques. Some numerical experiments are presented.

1. Introduction

The noise in data detected with common measurement devices frequently correlates with the data. The particular situations that are relevant for this work are when the noise locally correlates with the amplitude or the variation of the data.

In order to motivate the proposed methodology for denoising signals we recall the method of Tikhonov regularization for *denoising* data u^δ . Tikhonov regularization is among others, such as diffusion filtering (see e.g. [52], where an up to date list of references on diffusion filtering can be found), an efficient method for denoising.

Denoising data u^δ with Tikhonov regularization in its classical form requires the minimization of the functional

$$\mathcal{F}_{\text{TIK}}(u) := \|u - u^\delta\|_X^2 + \alpha\|u\|_Y^2,$$

where typically $\|\cdot\|_X$ is the L^2 -norm and Y is an Hilbert space with associated norm $\|\cdot\|_Y$. A frequently used setting is $Y = H^1(\Omega)$ with norm $\|\cdot\|_Y^2 = \int_{\Omega}(|\nabla u|^2 + |u|^2) dx$ or seminorm $\|\cdot\|_Y^2 = \int_{\Omega}|\nabla u|^2 dx$. $\alpha > 0$ is a positive parameter. All along this paper $|\cdot|$ denotes the Euclidean norm.

Tikhonov regularization can as well be regarded to provide a trade off between a *fidelity term* ($\|u - u^\delta\|_X^2$) and a *regularization term* ($\|\cdot\|_Y^2$). Assuming correlation of the data and noise we are led to the fidelity terms of the form (where for the

1991 *Mathematics Subject Classification.* 65R30,65J20,35J60.

Key words and phrases. Regularization, relative error functionals, morphological differential equations, mean curvature flow.

The work of O.S. has been supported by the Austrian Science Foundation (FWF), grant Y-123 INF.

© 2002 American Mathematical Society

sake of simplicity of presentation we concentrate on just three models)

$$\begin{aligned} & \frac{1}{2} \int_{\Omega} \frac{(u - u^\delta)^2}{|u|^p} dx \quad p = 0, 1, 2, \dots \\ & \frac{1}{2} \int_{\Omega} \frac{(u - u^\delta)^2}{|\nabla u|} dx \\ & \frac{1}{4} \int_{\Omega} \frac{(u - u^\delta)^4}{|\nabla u|^3} dx ; \end{aligned}$$

here ∇u denotes the gradient of u in an appropriate sense. To clarify the appropriate setting of $|\nabla u|$ is a central (nontrivial) part of this work.

In Figure 1 and Figure 2 we have plotted noisy data revealing the difference between uncorrelated and correlated noise.

All along this paper we will concentrate on the bounded variation seminorm for regularization. This is motivated from numerical reconstructions presented in previous literature and a general convergence theory of regularization methods on the space of functions of bounded variation, $BV(\Omega)$, to reconstruct discontinuous features in the underlying data. This leads to regularization models of the form (*relative error regularization*) :

$$(1.1) \quad \frac{1}{2} \int_{\Omega} \frac{(u - u^\delta)^2}{|u|^p} dx + \alpha \|Du\|;$$

$$(1.2) \quad \frac{1}{2} \int_{\Omega} \frac{(u - u^\delta)^2}{|\nabla u|} dx + \alpha \|Du\|;$$

$$(1.3) \quad \frac{1}{4} \int_{\Omega} \frac{(u - u^\delta)^4}{|\nabla u|^3} dx + \alpha \|Du\|;$$

over $BV(\Omega)$; here $\alpha > 0$ and $\|Du\| := \|Du\|(\Omega)$ denotes the bounded variation seminorm on Ω .

In order to put this work into context with recent results on *partial differential equations* and *calculus of variations* it is convenient to consider also *iterative* relative error regularization.

In particular, we consider the models of iteratively minimizing the functionals:

$$(1.4) \quad \frac{1}{2} \int_{\Omega} \frac{(u - u^{(k-1)})^2}{|u|^p} dx + \alpha \|Du\| ,$$

$$(1.5) \quad \frac{1}{2} \int_{\Omega} \frac{(u - u^{(k-1)})^2}{|\nabla u|} dx + \alpha \|Du\| ,$$

$$(1.6) \quad \frac{1}{4} \int_{\Omega} \frac{(u - u^{(k-1)})^4}{|\nabla u|^3} dx + \alpha \|Du\|;$$

and denoting the minimizers (presuming their existence) by $u^{(k)}$; moreover, we use the convention $u^{(0)} := u^\delta$.

Since the optimization problems (1.4) – (1.6) are *nonconvex* and thus quite delicate to handle analytically and numerically (cf. Section 5), we also consider

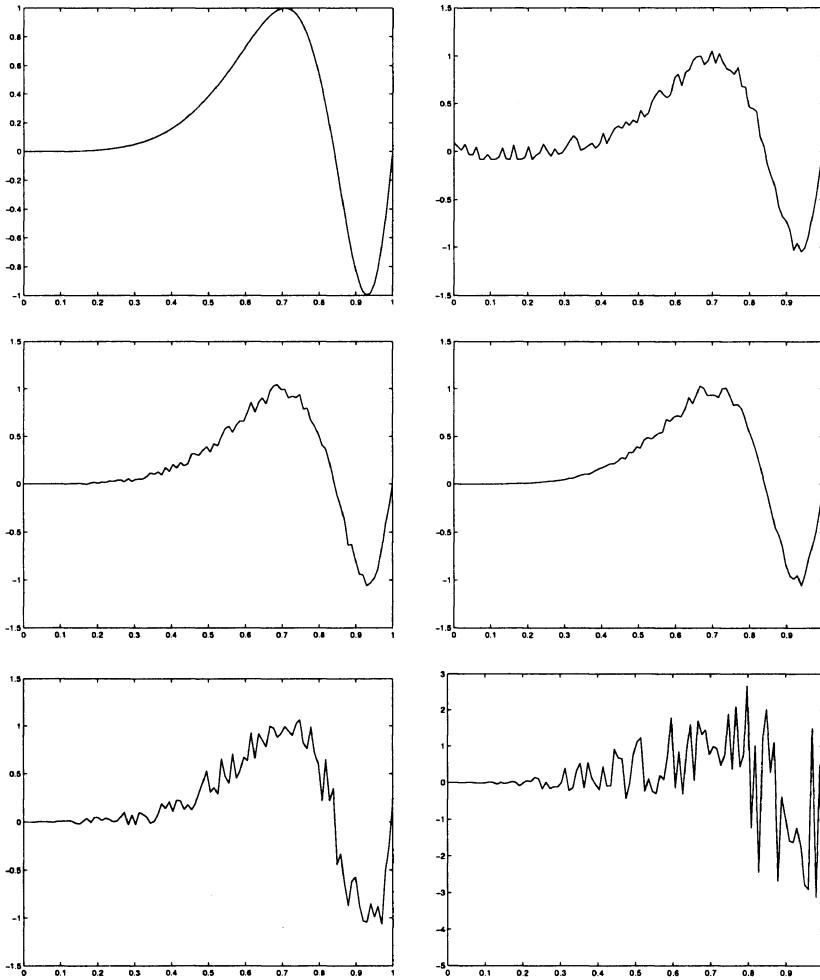


FIGURE 1. Noise in 1D signals: top left: noise free data; top right: uncorrelated noise; middle left: $\int_{\Omega} \frac{(u-u^{\delta})^2}{|u|} dx = \delta^2$; middle right: $\int_{\Omega} \frac{(u-u^{\delta})^2}{|u|^2} dx = \delta^2$; bottom left: $\int_{\Omega} \frac{(u-u^{\delta})^2}{|\nabla u|} dx = \delta^2$; bottom right: $\int_{\Omega} \frac{(u-u^{\delta})^4}{|\nabla u|^3} dx = \delta^2$.

some semi-implicit variants which consist in minimization of the functionals

$$(1.7) \quad \frac{1}{2} \int_{\Omega} \frac{(u - u^{(k-1)})^2}{|u^{(k-1)}|^p} dx + \alpha \|Du\|;$$

$$(1.8) \quad \frac{1}{2} \int_{\Omega} \frac{(u - u^{(k-1)})^2}{|\nabla u^{(k-1)}|} dx + \alpha \|Du\|;$$

$$(1.9) \quad \frac{1}{4} \int_{\Omega} \frac{(u - u^{(k-1)})^4}{|\nabla u^{(k-1)}|^3} dx + \alpha \|Du\|.$$

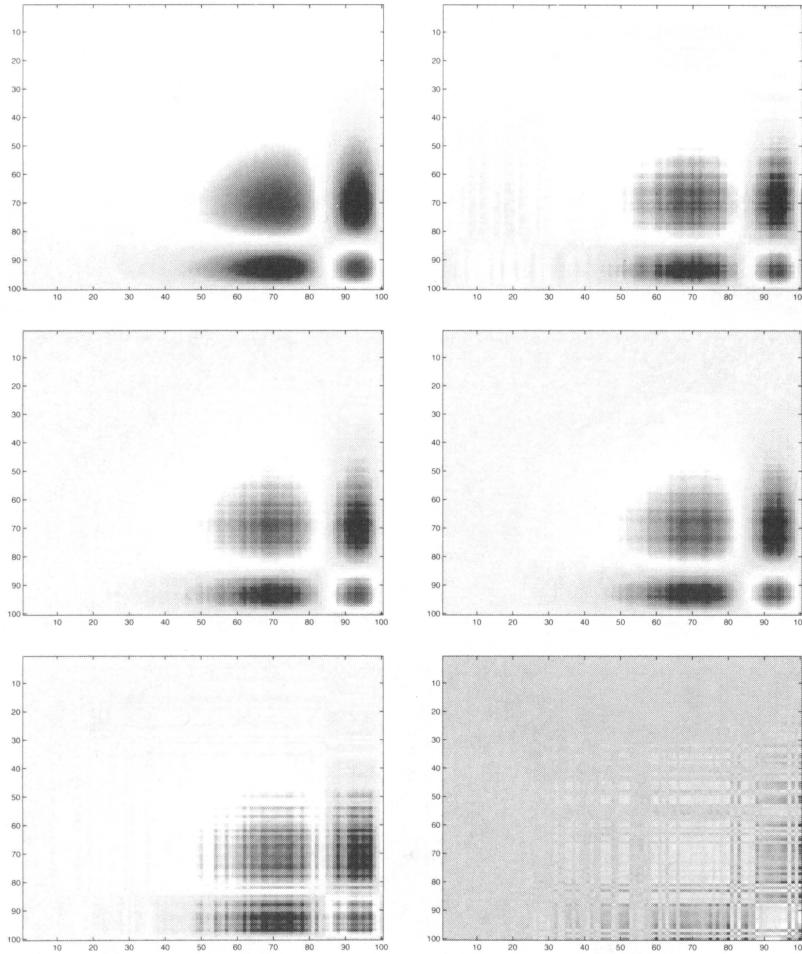


FIGURE 2. Noise in 2D signals: top left: noise free data; top right: uncorrelated noise; middle left: $\int_{\Omega} \frac{(u-u^\delta)^2}{|u|} dx = \delta^2$; middle right: $\int_{\Omega} \frac{(u-u^\delta)^2}{|u|^2} dx = \delta^2$; bottom left: $\int_{\Omega} \frac{(u-u^\delta)^2}{|\nabla u|} dx = \delta^2$; bottom right: $\int_{\Omega} \frac{(u-u^\delta)^4}{|\nabla u|^3} dx = \delta^2$.

The functionals (1.7) – (1.9) are convex with respect to u and thus much easier to analyze. However, from a modeling point of view these functionals are *not* completely satisfactory (cf. Section 2).

The outline of the paper is as follows: in Section 2 we prove well-posedness of the functionals (1.7) – (1.9). In Section 3 we study the nonconvex functionals (1.4) – (1.6) and compare the convex and nonconvex approach. In Section 5 we present some numerical results of relative error filtering methods and compare these results with *morphological* diffusion filtering methods. A comparison with morphological differential equations reveals nice properties of relative error regularization. Since morphological differential equations, like the mean curvature flow equation, are *contrast invariant*, the corresponding relative error regularization is quite stable

with respect to contrast variations of the input data. We found relative error regularization to be most useful for modeling contrast invariant methods also in other areas of applied mathematics such as “optical flow simulations”.

In Section 6 we add some remarks concerning optimality conditions for minimizers of relative error regularization methods.

2. Explicit relative error regularization

We study the variational methods of minimizing the functionals (1.1) – (1.3). There are at least two difficulties associated with rigorously defining these functionals:

- (1) $|\nabla u|$ or $|u|$ is the denominator of these functionals might vanish;
- (2) for $u \in BV(\Omega)$ the term $|\nabla u|$ just exists as a measure.

In order to overcome these difficulties there are at least two approaches: *relaxation* and *approximation of the gradient measure*.

2.1. Relaxation. We only consider the functional (1.2) with $f := u^\delta$; similar arguments apply to the other functionals, too. Let $\Omega \subseteq \mathbb{R}^n$; in relaxation methods the function values of

$$F(u) = \int_{\Omega} g(x, u(x), \nabla u(x)) dx,$$

with

$$g(x, \rho, P) = \frac{1}{2} \frac{(\rho - f(x))^2}{|P|} + \alpha |P|,$$

with $x \in \mathbb{R}^n$, $\rho \in \mathbb{R}$, and $P \in \mathbb{R}^n$, are defined via an approximation procedure.

Typically in this situation one would attempt to define (see e.g. [25, 26]) for given $u \in BV(\Omega)$ the *relaxed functional* as

$$\mathcal{F}(u) = \inf_{\{u_k\} \subseteq W^{1,1}(\Omega)} \left\{ \liminf_{k \rightarrow \infty} \frac{1}{2} \int_{\Omega} \frac{(u_k - f)^2}{|\nabla u_k|}(x) dx + \alpha \|Du_k\| : u_k \rightarrow u \text{ in } L^1(\Omega) \right\}.$$

Then our problem would be to minimize the relaxed functional $\mathcal{F}(u)$ with respect to u . This is technically quite complicated since we cannot rely on the structure of $\mathcal{F}(u)$. If g is convex with respect to P and satisfies some additional mild conditions, then the relaxed functional has an integral representation (see e.g. [26]), which consists of the sum of three parts: a part taking care of the regular parts and two parts which take care of the singular parts of the gradient of u .

Since $g(x, \rho, P)$ is not convex with respect to P , existence of an integral representation of a relaxed functional is not ensured.

We summarize some basics from measure theory. Let u be a real-valued function defined almost everywhere on an open (bounded or unbounded) subset Ω of \mathbb{R}^n , $n = 1, 2, 3, \dots$, with $\partial\Omega$ Lipschitz. If Ω is unbounded we assume that it satisfies the strong local Lipschitz property (see e.g. [2]).

According to the Lebesgue-Radon-Nikodym Theorem (see e.g. [44]) the measure $\|Du\|$ can be decomposed into an absolutely continuous part $\|Du\|_a$ and a singular part $\|Du\|_s$, i.e.,

$$\|Du\| = \|Du\|_a + \|Du\|_s.$$

Moreover,

$$\|Du\|_a(E) = \int_E |Du| dx \text{ for each measurable set } E,$$

where $|Du|$ denotes the density of the absolutely continuous part of the measure $\|Du\|_a$; note that

$$|Du| = (D\|Du\|),$$

i.e., $|Du|$ is the derivative of the measure $\|Du\|$. In particular $|Du| \in L^1(\Omega)$ for $u \in BV(\Omega)$.

In the following we use the following operators and sets: for $f \in BV(\Omega)$ we define

$$\Theta = \{x \in \Omega : |Df|(x) > 0\} \text{ and } \mathcal{G}_0(f) = \Omega \setminus \Theta.$$

For $f \in L^1(\Omega) \cap L^2(\Omega)$ and $L : L^1(\Omega) \cap L^2(\Omega) \rightarrow W^{1,1}(\Omega)$ we introduce the sets

$$\Theta^s = \{x \in \Omega : |\nabla(Lf)|(x) > 0\} \text{ and } \mathcal{G}_0^s(f) = \Omega \setminus \Theta^s.$$

Note that

$$\begin{aligned} \Theta &= \bigcup_{m=1}^{\infty} \left\{ x \in \Omega : |Df|(x) > \frac{1}{m} \right\}, \\ \Theta^s &= \bigcup_{m=1}^{\infty} \left\{ x \in \Omega : |\nabla(Lf)|(x) > \frac{1}{m} \right\} \end{aligned}$$

and $\mathcal{G}_0(f), \mathcal{G}_0^s(f)$ are measurable. For a function f defined almost everywhere in Ω we denote by $f|_{\Theta}$ the restriction onto the set Θ .

We analyze the following methods:

METHOD 2.1. *Minimization of the functional*

$$(2.1) \quad \mathcal{F}_{\text{ex}}^0(u) := \frac{1}{2} \int_{\Theta} \frac{(u-f)^2}{|Df|} dx + \alpha \|Du\|$$

over an appropriate subset of functions of bounded variation.

Note that in (2.1) the singular parts of f are neglected; this is reasonable since singular parts of f occur on regions with high gradients. In regions with singular parts the first term in the integral degenerates and does not contribute to the functional $\mathcal{F}_{\text{ex}}^0(u)$.

METHOD 2.2. *Let $L : L^1(\Omega) \cap L^2(\Omega) \rightarrow W^{1,1}(\Omega)$ be bounded. We consider minimization of the functional*

$$(2.2) \quad \mathcal{F}_{\text{ex-s}}^0(u) := \frac{1}{2} \int_{\Theta^s} \frac{(u-f)^2}{|\nabla(Lf)|}(x) dx + \alpha \|Du\|.$$

Since in comparison to (2.1) the *smoothing operator* L is involved, (2.2) represents a variant which can be handled numerically in a more easy way.

2.2. Existence of a minimizer of $\mathcal{F}_{\text{ex}}^0$ and $\mathcal{F}_{\text{ex-s}}^0$. In this Subsection we prove the existence of minimizers of the functional $\mathcal{F}_{\text{ex}}^0$ and $\mathcal{F}_{\text{ex-s}}^0$ in the class of functions of bounded variation.

We recall the definition of the space of functions of bounded variation and state some basic properties of functions of bounded variation which are relevant for this paper. These results are collected from Rudin [44], Giusti [27], Evans and Gariepy [22].

Let $C_c^1(\Omega; \mathbb{R}^n)$ be the space of n -dimensional vector valued functions which are continuously differentiable with compact support in Ω .

The BV -seminorm of a function $u \in BV(\Omega)$ is defined by

$$\|Du\| = \sup \left\{ \int_{\Omega} u(x) \nabla \cdot \varphi(x) dx : \varphi \in C_c^1(\Omega; \mathbb{R}^n) \text{ with } |\varphi(x)| \leq 1 \right\}.$$

The space $BV(\Omega)$ is defined as the space of L^1 -integrable functions for which $\|Du\|$ is finite. The space $BV(\Omega)$ associated with the norm

$$\|u\|_{BV(\Omega)} = \|u\|_{L^1(\Omega)} + \|Du\|$$

is a Banach space.

Derivatives of functions of bounded variation can be characterized by the structure theorem for BV functions (see [22]):

THEOREM 2.3. (Structure theorem for BV functions) *Let $u \in BV(\Omega)$. Then there exists a Radon measure λ on Ω and a λ measurable function $\sigma : \Omega \rightarrow \mathbb{R}^n$ such that*

- (1) $|\sigma(x)| = 1$ λ -almost everywhere, and
- (2) $\int_{\Omega} u \nabla \cdot \varphi dx = - \int_{\Omega} \varphi \cdot \sigma d\lambda$ for all $\varphi \in C_c^1(\Omega; \mathbb{R}^n)$.

We recall that for a function $u \in W^{1,1}(\Omega)$

$$(D\lambda) = |\nabla u| \text{ and } \sigma = \begin{cases} \frac{\nabla u}{|\nabla u|} & \text{if } \nabla u \neq 0 \\ 0 & \text{if } \nabla u = 0. \end{cases}$$

For a function $u \in BV(\Omega)$

$$\lambda = \|Du\|.$$

To characterize minimizers of $\mathcal{F}_{\text{ex}}^0$ and $\mathcal{F}_{\text{ex-s}}^0$ we use the following weighted L^2 -spaces:

$$\begin{aligned} L_f^2 &= \left\{ u : \Theta \rightarrow \mathbb{R} \left| \frac{u}{\sqrt{|Df|}} \in L^2(\Theta) \right. \right\}, \\ L_{f-s}^2 &= \left\{ u : \Theta^s \rightarrow \mathbb{R} \left| \frac{u}{\sqrt{|\nabla(Lf)|}} \in L^2(\Theta^s) \right. \right\}. \end{aligned}$$

We note that $u/\sqrt{|Df|}$ is measurable since $|Df| > 0$ on Θ . Moreover, it follows from the Cauchy-Schwarz inequality that

$$\begin{aligned} L_f^2 &= \{ \hat{u} \sqrt{|Df|} : \hat{u} \in L^2(\Theta) \} \subseteq L^1(\Theta), \\ L_{f-s}^2 &= \{ \hat{u} \sqrt{|\nabla(Lf)|} : \hat{u} \in L^2(\Theta^s) \} \subseteq L^1(\Theta^s). \end{aligned}$$

LEMMA 2.4. (1) Let $f \in BV(\Omega)$. Then L_f^2 with the inner product

$$\langle u, v \rangle_f = \int_{\Theta} \frac{u \cdot v}{|Df|} dx$$

is a Hilbert space.

- (2) Let $L : L^1(\Omega) \cap L^2(\Omega) \rightarrow W^{1,1}(\Omega)$ satisfying $\|L \cdot\|_{W^{1,1}(\Omega)} \leq C \|\cdot\|_{L^1(\Omega)}$, with $C > 0$. Moreover, let $f \in L^1(\Omega) \cap L^2(\Omega)$. Then L_{f-s}^2 with the inner product

$$\langle u, v \rangle_{f-s} := \int_{\Theta} \frac{u \cdot v}{|\nabla(Lf)|} dx$$

is a Hilbert space.

PROOF. In the first case we introduce a weight function

$$g(x) = \frac{1}{|Df|}(x).$$

In the second case we introduce a weight function

$$g(x) = \frac{1}{|\nabla(Lf)|}(x).$$

Introducing the measure

$$d\mu = gdx$$

shows that

$$L_f^2 = L^2(\Theta; \mu).$$

Since the later is a Hilbert space the assertion is proved. \square

Stampacchia's Lemma (see e.g. [22]) states that for any $\underline{K} < \bar{K}$ and any function u in the Sobolev-space $W^{1,p}(\Omega)$

$$\int_{\Omega} |\nabla(\max\{\min\{u, \bar{K}\}, \underline{K}\})|^p dx \leq \int_{\Omega} |\nabla u|^p dx.$$

A similar result holds for functions of bounded variation. Although the result is obviously true, we were not able to find it in the literature and thus a proof is added here for the sake of completeness.

LEMMA 2.5. (Stampacchia's Lemma for BV functions) *Let $u \in BV(\Omega)$ and let $\underline{K} < \bar{K}$ be fixed. Then $u^{\underline{K}, \bar{K}} := \max\{\min\{u, \bar{K}\}, \underline{K}\}$ satisfies*

$$\|Du^{\underline{K}, \bar{K}}\| \leq \|Du\|.$$

PROOF. For $u \in BV(\Omega)$ there exists a sequence $\{u_k\}_{k \in \mathbb{N}}$ in $C^\infty(\Omega)$ satisfying

$$u_k \rightarrow u \text{ in } L^1(\Omega) \text{ and } \int_{\Omega} |\nabla u_k| dx \rightarrow \|Du\|.$$

Since $u_k^{\bar{K}, \underline{K}} \rightarrow u^{\bar{K}, \underline{K}}$ in $L^1(\Omega)$, $\{u_k^{\bar{K}, \underline{K}}\}_{k \in \mathbb{N}} \in W^{1,1}(\Omega)$, with $\left\{ \|u_k^{\bar{K}, \underline{K}}\|_{W^{1,1}(\Omega)} \right\}_{k \in \mathbb{N}}$ uniformly bounded we see that

$$\begin{aligned} \|Du^{\bar{K}, \underline{K}}\| &\leq \liminf_{k \in \mathbb{N}} \int_{\Omega} |\nabla u_k^{\bar{K}, \underline{K}}| dx \\ &\leq \liminf_{k \in \mathbb{N}} \int_{\Omega} |\nabla u_k| dx \\ &= \|Du\|, \end{aligned}$$

and $u^{\bar{K}, \underline{K}} \in BV(\Omega)$. This proves the assertion. \square

With each function $f \in BV(\Omega)$ we associate the affine linear spaces,

$$\mathcal{L}_f^2 := f + L_f^2 \text{ and } \mathcal{L}_{f-s}^2 := f + L_{f-s}^2.$$

Now we are able to formulate and prove an existence result. The following result uses simple lower semi-continuity and compactness arguments. Although the proof is simple it is included here, since it gives a clear indication of spaces where we can expect a minimizer of the explicit relative error functionals.

THEOREM 2.6. *Let Ω be bounded with $\partial\Omega$ Lipschitz or unbounded with $\partial\Omega$ satisfying a strong local Lipschitz property.*

Let $f \in BV(\Omega) \cap L^\infty(\Omega)$.

- (1) *Then there exists a unique minimizer of $\mathcal{F}_{\text{ex}}^0$ in $\mathcal{S} = BV(\Omega) \cap L^\infty(\Omega) \cap \mathcal{L}_f^2$.*
- (2) *Let $L : L^1(\Omega) \cap L^2(\Omega) \rightarrow W^{1,1}(\Omega)$ satisfying $\|L \cdot\|_{W^{1,1}(\Omega)} \leq C \|\cdot\|_{L^1(\Omega)}$. Then there exists a unique minimizer of $\mathcal{F}_{\text{ex-s}}^0$ in $\mathcal{S}_s := BV(\Omega) \cap L^\infty(\Omega) \cap \mathcal{L}_{f-s}^2$.*

PROOF. We only prove the first item. The second item can be proven analogously.

For $\bar{K} = \|f\|_{L^\infty(\Omega)}$ we have $\|f\|_{L^2(\Omega)}^2 \leq \bar{K}\|f\|_{L^1(\Omega)}$ and thus $f \in L^2(\Omega)$. This in particular shows that $f \in \mathcal{S}$, and consequently $\mathcal{S} \neq \emptyset$. Suppose that the minimum of $\mathcal{F}_{\text{ex}}^0$ is not attained in \mathcal{S} , then there exists a sequence $\{v_k\}_{k \in \mathbb{N}} \subseteq \mathcal{S}$ such that

$$\mathcal{F}_{\text{ex}}^0(v_k) \rightarrow \inf_{u \in \mathcal{S}} \mathcal{F}_{\text{ex}}^0(u) \text{ as } k \rightarrow \infty.$$

Let $v_k^{\bar{K}} := v_k^{-\bar{K}, \bar{K}}$ be as defined above, then by Lemma 2.5, $v_k^{\bar{K}} \in BV(\Omega)$ and

$$\|Dv_k^{\bar{K}}\| \leq \|Dv_k\|.$$

Moreover, from the definition of $v_k^{\bar{K}}$ it follows that

$$\int_{\Theta} \frac{|v_k^{\bar{K}} - f|^2}{|Df|}(x) dx \leq \int_{\Theta} \frac{|v_k - f|^2}{|Df|}(x) dx.$$

Since by the Lebesgue-Radon-Nikodym Theorem $|Df| \in L^1(\Omega)$ we find that there exists a set \mathcal{C} , satisfying $\text{meas}(\mathcal{C}) < \infty$ (the existence of such a set is only relevant in the proof if $\text{meas}(\Omega) = \infty$; if $\text{meas}(\Omega) < \infty$ we can set $\mathcal{C} = \emptyset$), such that

$$|Df|(x) \leq 1 \text{ almost everywhere for } x \in \Omega \setminus \mathcal{C}.$$

Since $\mathcal{F}_{\text{ex}}^0(v_k^{\bar{K}}) \leq \mathcal{F}_{\text{ex}}^0(v_k)$ we have

$$\mathcal{F}_{\text{ex}}^0(v_k^{\bar{K}}) \rightarrow \inf_{u \in \mathcal{S}} \mathcal{F}_{\text{ex}}^0(u).$$

In particular for any $\varepsilon > 0$ there exists $N_0 \in \mathbb{N}$ such that for all $k \geq N_0$

$$\int_{\Theta} \frac{|v_k^{\bar{K}} - f|^2}{|Df|} dx + \alpha \|Dv_k^{\bar{K}}\| \leq \alpha \|Df\| + \varepsilon.$$

Consequently

$$(2.3) \quad \begin{aligned} \int_{\Theta} \frac{|v_k^{\bar{K}} - f|^2}{|Df|} dx &\leq \alpha \|Df\| + \varepsilon, \\ \int_{\Theta \setminus \mathcal{C}} |v_k^{\bar{K}} - f|^2 dx &\leq \alpha \|Df\| + \varepsilon, \end{aligned}$$

for all $k \geq N_0$.

In particular the first inequality in (2.3) and Lemma 2.4 imply that $\{(v_k^{\bar{K}} - f)|_{\Theta}\}_{k \in \mathbb{N}}$ has a subsequence that converges weakly to \hat{z} in Θ with respect to the L_f^2 inner product. Since no confusion can arise we denote this subsequence

again by $\{v_k^{\bar{K}} - f\}_{k \in \mathbb{N}}$. Since L_f^2 is a Hilbert space we conclude from the weak lower semi-continuity of the norm in a Hilbert space that

$$(2.4) \quad \|\hat{z}\|_{L_f^2} \leq \liminf \left\| \left(v_k^{\bar{K}} - f \right) \Big|_{\Theta} \right\|_{L_f^2}.$$

From (2.3) it follows that for all $k \geq N_0$

$$\begin{aligned} \int_{\Omega \setminus C} |v_k^{\bar{K}}|^2 dx &= \int_{\Theta \setminus C} |v_k^{\bar{K}}|^2 dx + \int_{G_0(f) \setminus C} |f|^2 dx \\ &\leq 2 \int_{\Theta \setminus C} |v_k^{\bar{K}} - f|^2 dx + 2 \int_{\Omega \setminus C} |f|^2 dx \\ &\leq 2\alpha \|Df\| + 2 \int_{\Omega \setminus C} |f|^2 dx + 2\varepsilon. \end{aligned}$$

In particular this shows that $\int_{\Omega \setminus C} |v_k^{\bar{K}}|^2 dx$ is uniformly bounded.

Since

$$\|v_k^{\bar{K}}\|_{L^2(C)}^2 \leq \bar{K}^2 \text{meas}(C) < \infty,$$

we see that $\{v_k^{\bar{K}}\}_{k \in \mathbb{N}}$ is uniformly bounded in $L^2(\Omega)$. Thus $\{v_k^{\bar{K}}\}_{k \in \mathbb{N}}$ has a weakly convergent subsequence in $L^2(\Omega)$ (which we denote again by $\{v_k^{\bar{K}}\}_{k \in \mathbb{N}}$ since no notational ambiguity can arise) with weak limit u . Consequently (see e.g. [27])

$$\|Du\| \leq \liminf \|Dv_k^{\bar{K}}\|.$$

Since $v_k^{\bar{K}} = f$ almost everywhere on $G_0(f)$, we also have $u = f$ in $G_0(f)$. It remains to be shown that $u = \hat{u} := \hat{z} + f$ almost everywhere in Θ . This, together with (2.4) finally shows that u is a minimizer of $\mathcal{F}_{\text{ex}}^0$. Taking

$$\Omega_\varepsilon = \{x \in \Theta : \frac{1}{\varepsilon} \geq |Df|(x) \geq \varepsilon\}.$$

Then for any $w \in L^2(\Omega)$, $\frac{w}{|Df|} \chi(\Omega_\varepsilon) \in L^2(\Omega)$. Consequently, for all $w \in L^2(\Omega)$

$$\begin{aligned} \int_{\Theta} (u - f) \frac{w}{|Df|} \chi(\Omega_\varepsilon) dx &= \lim_{k \rightarrow \infty} \int_{\Theta} (v_k^{\bar{K}} - f) \frac{w}{|Df|} \chi(\Omega_\varepsilon) dx \\ &= \lim_{k \rightarrow \infty} \left\langle (v_k^{\bar{K}} - f) \chi_{\Omega_\varepsilon}, w \right\rangle_f \\ &= \int_{\Theta} (\hat{u} - f) \frac{w}{|Df|} \chi(\Omega_\varepsilon) dx. \end{aligned}$$

Consequently $u = \hat{u}$ on Ω_ε . Since the last identity holds for all $\varepsilon > 0$ we get $u = \hat{u}$ on Θ .

The uniqueness is an easy consequence of the strict convexity of the bias term and the convexity of the bounded variation seminorm. \square

We note that if the boundary of Θ is Lipschitz, then

$$\mathcal{S} = f + L_f^2(\Theta) \cap L^\infty(\Theta) \cap BV(\Theta).$$

In particular each $s \in \mathcal{S}$ is in $BV(\Omega)$.

2.3. Properties of the explicit relative error regularization. In this subsection let Ω be a bounded domain with $\partial\Omega$ Lipschitz.

LEMMA 2.7. *Let $f \in \mathcal{S}$ with $|Df|$ strictly positive almost everywhere and satisfying $\frac{1}{|Df|} \in L^2(\Omega)$. Then the minimizer u^\dagger of $\mathcal{F}_{\text{ex}}^0$ satisfies*

$$\int_{\Omega} \frac{u^\dagger - f}{|Df|} dx = 0.$$

PROOF. Take $h = \text{constant}$. Under the assumptions of this Lemma $u^\dagger + h \in \mathcal{S}$ and thus

$$\int_{\Omega} \frac{(u^\dagger - f)^2}{|Df|} dx \leq \int_{\Omega} \frac{(u^\dagger + h - f)^2}{|Df|} dx.$$

Thus

$$2h \int_{\Omega} \frac{u^\dagger - f}{|Df|} dx + h^2 \int_{\Omega} \frac{1}{|Df|} dx \geq 0.$$

Since this inequality holds for positive and negative h , it follows that

$$2 \int_{\Omega} \frac{u^\dagger - f}{|Df|} dx + h \int_{\Omega} \frac{1}{|Df|} dx = 0.$$

Taking the limit $h \rightarrow 0$ proves the assertion. \square

In the following we study parameter dependent properties of the minimizers of $\mathcal{F}_{\text{ex}}^0$. Let u_α be the minimizer of $\mathcal{F}_{\text{ex}}^0$.

LEMMA 2.8. *Let $f \in \mathcal{S}$. Then*

- (1) $\lim_{\alpha \rightarrow 0} \|u_\alpha - f\|_{L_f^2} = 0$.
- (2) *If $|Df| > \varepsilon > 0$ on a domain K , then $u_\alpha \rightarrow f$ in $L^2(K)$.*
- (3) *If $|Df| > \varepsilon > 0$ on Ω , then $\lim_{\alpha \rightarrow \infty} \|Du_\alpha\|(\Omega) = 0$.*

PROOF. From the definition of a minimizer of the functional $\mathcal{F}_{\text{ex}}^0$ it follows that

$$(2.5) \quad \mathcal{F}_{\text{ex}}^0(u_\alpha) \leq \alpha \|Df\|.$$

In particular this shows that

$$(2.6) \quad \|Du_\alpha\| \leq \|Df\|.$$

From this it follows by taking the limit $\alpha \rightarrow 0$ in (2.5) that

$$\|u_\alpha - f\|_{L_f^2} \rightarrow 0 \text{ as } \alpha \rightarrow 0.$$

This proves the first assertion. Moreover, under the assumption that $|Df| > \varepsilon$ on K it follows that $\lim_{\alpha \rightarrow 0} u_\alpha = f$ in $L^2(K)$. Under the assumptions of the third item we have $\frac{(f)^2}{|Df|} \in L^1(\Omega)$. Consequently

$$\frac{1}{2} \int_{\Omega} \frac{(u_\alpha - f)^2}{|Df|} dx + \alpha \|Du_\alpha\| \leq \frac{1}{2} \int_{\Omega} \frac{(f)^2}{|Df|} dx < \infty.$$

Division of the inequality by α and taking the limit $\alpha \rightarrow \infty$ gives the assertion. \square

Method 2.1 exhibits standard properties (cf. Lemma 2.8) of Tikhonov regularization (see e.g. [34, 21], to name but a few).

However, several properties of a standard Tikhonov regularization technique are not satisfied:

- (1) In general, if $\mathcal{G}_0(f) \neq \emptyset$ u_α may not approach a constant function. This is simply due to the fact that we force $u_\alpha = f$ in $\mathcal{G}_0(f)$.

- (2) A stationary state of (2.1) is a piecewise constant function. This is not the case in Tikhonov regularization where the stabilization term always vanishes for $\alpha \rightarrow \infty$ (see e.g. [34, 21]).

3. Implicit relative Error Regularization

DEFINITION 3.1. Let $X = BV(\Omega) \cap L^\infty(\Omega)$, $X_r = W^{1,1}(\Omega) \cap L^\infty(\Omega)$ and let

$$\mathcal{F}_{\text{imp}}^\varepsilon : X_r \rightarrow \mathbb{R} \cup \{+\infty\}$$

$$u \rightarrow \frac{1}{2} \int_{\Omega} \frac{(u - f)^2}{\sqrt{|\nabla u|^2 + \varepsilon^2}} dx + \alpha \int_{\Omega} |\nabla u| dx.$$

For $u \in X$ we set

$$\begin{aligned} \mathcal{F}_{\text{imp}}(u) &:= \Gamma^- \liminf_{\varepsilon \rightarrow 0^+} \mathcal{F}_{\text{imp}}^\varepsilon(u) \\ &:= \inf_{\substack{\{(u_k, \varepsilon_k)\}_{k \in \mathbb{N}}, \\ u_k \in W^{1,1}(\Omega), \varepsilon_k > 0, \\ u_k \rightharpoonup u \text{ in } L^2(\Omega), \\ \varepsilon_k \rightarrow 0}} \liminf_{k \rightarrow \infty} \mathcal{F}_{\text{imp}}^{\varepsilon_k}(u_k). \end{aligned}$$

The weak Γ^- limit is understood in the sense of a weak sequential limes inferior: for any *positive* sequence $\{\varepsilon_k\}_{k \in \mathbb{N}}$, satisfying $\varepsilon_k \rightarrow 0^+$ and $u_k \rightharpoonup u$ in $L^2(\Omega)$ the limes inferior of $\mathcal{F}_{\text{imp}}^{\varepsilon_k}(u_k)$ is calculated. The infimum over all such pairs of sequences is $\mathcal{F}_{\text{imp}}(u)$.

REMARK 3.2. (1) Let u satisfy $\mathcal{F}_{\text{imp}}(u) < \infty$, then from the definition of $\mathcal{F}_{\text{imp}}(u)$ it follows that for given $m \in \mathbb{N}$ there exists a pair of sequences $\{\varepsilon_k\}_{k \in \mathbb{N}}$, $\{u_k\}_{k \in \mathbb{N}} \subseteq X_r$ satisfying $\varepsilon_k \rightarrow 0^+$, $u_k \rightharpoonup u$ in $L^2(\Omega)$ and

$$(3.1) \quad \left| \lim_{k \rightarrow \infty} \mathcal{F}_{\text{imp}}^{\varepsilon_k}(u_k) - \mathcal{F}_{\text{imp}}(u) \right| \leq \frac{1}{m}.$$

Thus there also exists a *strictly* monotonically decreasing subsequence $\{\varepsilon_k\}_{k \in \mathbb{N}}$ such that (3.1) holds.

- (2) If $\mathcal{F}_{\text{imp}}(u) = \infty$, then there exists a monotonically decreasing, positive sequence $\{\varepsilon_k\}_{k \in \mathbb{N}}$ converging to 0 such that $\{u_k\}_{k \in \mathbb{N}} \subseteq X_r$ and $u_k \rightharpoonup u$ in $L^2(\Omega)$ and $\lim_{\varepsilon_k \rightarrow 0} \mathcal{F}_{\text{imp}}^{\varepsilon_k}(u_k) = \infty$.
- (3) To see that for any $u \in X$ $\mathcal{F}_{\text{imp}}(u)$ is well-defined, we have to prove that there exists *at least* one sequence in X_r such that $u_k \rightharpoonup u$ in $L^2(\Omega)$.
- For $u \in X_r$ we can choose $u_k = u$.
 - For $u \in BV(\Omega)$ there exists a sequence $u_k \in C^\infty(\Omega)$ satisfying

$$(3.2) \quad u_k \rightarrow u \text{ in } L^1(\Omega) \text{ and } \|Du_k\| \rightarrow \|Du\|$$

(see e.g. [27]). Using that $u \in L^\infty(\Omega)$, say $u_{\min} \leq u \leq u_{\max}$, we set

$$\bar{u}_k := \max\{\min\{u_k, u_{\max}\}, u_{\min}\}, \bar{K} := \max\{|u_{\min}|, |u_{\max}|\}.$$

From Stampacchia's Lemma 2.5 it follows that $\bar{u}_k \in BV(\Omega)$; moreover, we have

$$\|D\bar{u}_k\| \leq \|Du_k\|.$$

Since \bar{u}_k is bounded in $L^\infty(\Omega)$ and $L^1(\Omega)$ we have

$$\int_{\Omega} |\bar{u}_k|^2 \leq \bar{K} \|\bar{u}_k\|_{L^1(\Omega)},$$

and thus it is bounded in $L^2(\Omega)$. Consequently there exists a weakly convergent subsequence in $L^2(\Omega)$, which for simplicity of notation is again denoted by $\{\bar{u}_k\}_{k \in \mathbb{N}}$. Since from (3.2) it follows that

$$\bar{u}_k \rightarrow u \text{ in } L^1(\Omega),$$

we find that the weak limit of $\{\bar{u}_k\}_{k \in \mathbb{N}}$ is also u , which proves the assertion.

In the following we prove that the minimum of \mathcal{F}_{imp} is attained. For this we use the following Lemma:

LEMMA 3.3. *Let $f \in L^\infty(\Omega)$ with*

$$f_{\inf} := \text{essinf}(f) \leq f \leq \text{esssup}(f) =: f_{\sup}.$$

Let $u \in X_r$, then we denote by

$$\bar{u} := \max\{\min\{u, f_{\sup}\}, f_{\inf}\}.$$

Then

$$\mathcal{F}_{\text{imp}}^\varepsilon(\bar{u}) \leq \mathcal{F}_{\text{imp}}^\varepsilon(u).$$

THEOREM 3.4. *Let $\Omega \subseteq \mathbb{R}^n$ be bounded with $\partial\Omega$ Lipschitz. Let $f \in L^\infty(\Omega)$ then the minimum of \mathcal{F}_{imp} is attained in X and $z := \inf_{u \in X} \mathcal{F}_{\text{imp}}(u) < \infty$.*

PROOF. We split the proof into two parts:

- (1) Since $u(x_1, \dots, x_n) = x_1 \in X_r$ it follows that for $\varepsilon \rightarrow 0$

$$\begin{aligned} \mathcal{F}_{\text{imp}}^\varepsilon(u) &= \int_{\Omega} \frac{(x_1 - f)^2}{\sqrt{1 + \varepsilon^2}} dx + \alpha \text{meas}(\Omega) \\ &\rightarrow \int_{\Omega} (x_1 - f)^2 dx + \alpha \text{meas}(\Omega) < \infty. \end{aligned}$$

Thus, taking the sequence $\{u_k := u\}_{k \in \mathbb{N}}$ and $\{\varepsilon_k = 1/k\}_{k \in \mathbb{N}}$ implies that $\mathcal{F}_{\text{imp}}(u) < \infty$.

- (2) Now suppose that the minimum of \mathcal{F}_{imp} is not attained in X , then there exists a sequence $\{u^k\}_{k \in \mathbb{N}}$ in X such that

$$\mathcal{F}_{\text{imp}}(u^k) \rightarrow z < \infty.$$

By the definition of $\mathcal{F}_{\text{imp}}(u^k)$ there exists a sequence $\{\varphi_m := \varepsilon_m^k\}_{m \in \mathbb{N}}$ converging to 0 for $m \rightarrow \infty$ and

$$\left\{w_m := v_{\varepsilon_m^k}^k\right\}_{m \in \mathbb{N}} \subseteq X_r$$

such that

$$w_m \rightharpoonup u^k \text{ in } L^2(\Omega) \text{ and } \mathcal{F}_{\text{imp}}^{\varphi_m}(w_m) \rightarrow \mathcal{F}_{\text{imp}}(u^k) \text{ as } m \rightarrow \infty.$$

Let $\{\tau_k = \varepsilon_{m(k)}^k\}_{k \in \mathbb{N}}$ be monotonically decreasing such that

$$|\mathcal{F}_{\text{imp}}^{\tau_k}(v_{\tau_k}^k) - \mathcal{F}_{\text{imp}}(u^k)| \leq k^{-1}.$$

The sequence $\{\bar{v}_{\tau_k}^k\}_{k \in \mathbb{N}}$ is bounded in $L^\infty(\Omega)$ and thus has a weakly convergent subsequence in $L^2(\Omega)$, which for simplicity of notation is again denoted by $\{\bar{v}_{\tau_k}^k\}_{k \in \mathbb{N}}$; the weak limit will be denoted by ρ . Since each element of the sequence $\{\bar{v}_{\tau_k}^k\}_{k \in \mathbb{N}}$ is pointwise bounded from below by f_{\inf} and from above by f_{\sup} , so is the weak limit ρ .

Since

$$|\mathcal{F}_{\text{imp}}^{\tau_k}(v_{\tau_k}^k)| \leq |\mathcal{F}_{\text{imp}}(u^k)| + k^{-1},$$

it follows in particular from Stampacchia's Lemma that

$$\begin{aligned} \alpha \|D\bar{v}_{\tau_k}^k\| &\leq \alpha \|Dv_{\tau_k}^k\| \\ &\leq |\mathcal{F}_{\text{imp}}^{\tau_k}(v_{\tau_k}^k)| \\ &\leq |\mathcal{F}_{\text{imp}}(u^k)| + k^{-1}, \end{aligned}$$

which shows that the sequence $\{\bar{v}_{\tau_k}^k\}_{k \in \mathbb{N}}$ is uniformly bounded in $BV(\Omega)$. Moreover, we also have that

$$\|D\rho\| \leq \liminf_{k \in \mathbb{N}} \|D\bar{v}_{\tau_k}^k\|.$$

This in particular shows that $\rho \in X$. Moreover, from the definition of \mathcal{F}_{imp} and Lemma 3.3 it follows that

$$\mathcal{F}_{\text{imp}}(\rho) \leq \liminf_{k \rightarrow \infty} \mathcal{F}_{\text{imp}}^{\tau_k}(\bar{v}_{\tau_k}^k) \leq \liminf_{k \rightarrow \infty} \mathcal{F}_{\text{imp}}^{\tau_k}(v_{\tau_k}^k) = z.$$

This shows that ρ is a minimizer of \mathcal{F}_{imp} . \square

4. Relation to morphological partial differential equations

In the nonlinear diffusion framework, natural relations between biased diffusion and regularization theory exist via the Euler equation for the regularization functional. This Euler equation can be regarded as the steady-state of a suitable nonlinear diffusion process with a bias term [37, 48, 12, 47, 39, 40]. The regularization parameter and the diffusion time can be identified if one regards regularization as time-discrete diffusion filtering with a single implicit time step [33, 49, 46].

The approach relating mean curvature flow to Method 2.1 is conceptually rather similar to relating total variation flow

$$u_t = \nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right)$$

to bounded variation regularization. For some reference concerning bounded variation regularization we refer to [43, 1, 9, 35, 31, 10, 11, 14, 30, 13, 19, 17, 18, 50, 21]. Since the following considerations serve for motivation purposes we are not precise on the meaning of ∇u ; it could denote a measure or a function. In the literature of bounded variation regularization $-\nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right)$ is considered as the functional derivative of the functional $\|Du\|$ with respect to u . In Section 6 we give a rigorous interpretation of the term $-\nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right)$.

For the numerical solution of the *mean curvature flow* equation

$$(4.1) \quad u_t = |\nabla u| \nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right),$$

(1.8) can be regarded as the *2/3-semi-time-implicit* numerical scheme

$$\frac{u^{(k)} - u^{(k-1)}}{|\nabla u^{(k-1)}|} = \nabla \cdot \left(\frac{\nabla u^{(k)}}{|\nabla u^{(k)}|} \right).$$

This can be seen by identifying the regularization parameter with the time step

$$\alpha = t_k - t_{k-1},$$

and identifying $u^{(k)} \approx u(t_k)$ and $u^{(k-1)} \approx u(t_{k-1})$.

The terminology *2/3-semi-implicit* refers to the fact that 2 occurrences of ∇u in $|\nabla u| \nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right)$ are implemented implicitly and one gradient is implemented explicitly. In the literature the following numerical methods are suggested for the solution of the mean curvature equation :

(1) **explicit time discretization:**

$$\frac{u^{(k+1)} - u^{(k)}}{t_{k+1} - t_k} = |\nabla u^{(k)}| \nabla \cdot \left(\frac{\nabla u^{(k)}}{|\nabla u^{(k)}|} \right),$$

(2) **1/3 - semi-implicit time discretization:**

$$\frac{u^{(k+1)} - u^{(k)}}{t_{k+1} - t_k} = |\nabla u^{(k)}| \nabla \cdot \left(\frac{\nabla u^{(k+1)}}{|\nabla u^{(k)}|} \right).$$

Formally, the first order optimality condition for a minimizer of (2.1) (with $f = u^{(0)}$) is

$$(4.2) \quad \frac{u - u^{(0)}}{|\nabla u^{(0)}|} \chi(\Theta) - \alpha \nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right) = 0 \text{ in } \Omega.$$

If Ω is a bounded domain, then this differential equation should be considered together with homogeneous Neumann boundary data on $\partial\Omega$.

To point out the connection between (2.1) and mean curvature flow, we note that in a connected domain of positive measure where $|\nabla u^{(0)}|$ vanishes, we force $u - u^{(0)} = 0$. Then, (4.2) reveals that in this particular domain the curvature $\nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right)$ stays bounded.

Analogously we find that (1.9) is a *semi-implicit* method (in time) for solving the *affine invariant mean curvature flow* equation

$$(4.3) \quad u_t = |\nabla u| \nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right)^{1/3}.$$

(1.7) is a *semi-implicit* method for solving

$$(4.4) \quad u_t = |u|^p \nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right).$$

The mean curvature flow equation attains a *viscosity solution* if the initial data $u^{(0)}$ is continuously differentiable on $\overline{\Omega}$ (see e.g. [15]). So far we were not successful in proving that the “elliptic” differential equations

$$\begin{aligned} u - u^{(k-1)} &= \alpha |\nabla u^{(k-1)}| \nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right), \\ u - u^{(k-1)} &= \alpha |\nabla u^{(k-1)}| \left(\nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right) \right)^{1/3}, \\ u - u^{(k-1)} &= \alpha |u^{(k-1)}|^p \nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right), \end{aligned}$$

which constitute the formal Euler equations for the minimizers of the functionals (4.1), (4.3), and (4.4), respectively, satisfy the general assumptions for the existence of viscosity solutions. Through the approach presented in this paper it is possible to consider non-continuous functions as initial data $u^{(0)}$.

The formal Euler equation for the minimizer of (1.5) is

$$(4.5) \quad \frac{u - u^{(k-1)}}{|\nabla u|} = \nabla \cdot \left(\left(\alpha - \frac{1}{2} \frac{(u - u^{(k-1)})^2}{|\nabla u|^2} \right) \frac{\nabla u}{|\nabla u|} \right).$$

Division by α gives

$$\frac{u - u^{(k-1)}}{\alpha} = |\nabla u| \nabla \cdot \left(\left(1 - \frac{1}{2} \frac{(u - u^{(k-1)})^2}{\alpha^2} \right) \frac{\nabla u}{|\nabla u|} \right).$$

Taking the formal limit $\alpha \rightarrow 0$ and considering again $u^{(k-1)} \approx u(t_{k-1})$, $u^{(k)} \approx u(t_k)$ and $\alpha = t_k - t_{k-1}$ gives

$$\frac{\partial u}{\partial t} = |\nabla u| \nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right),$$

the mean curvature flow equation.

(4.5) can be considered as a Perona-Malik model with positive and negative diffusion. Thus in general the solution of this differential equation is ill-posed. However, in practical applications for denoising one could use a parameter setting α such that the diffusion coefficient

$$\alpha - \frac{1}{2} \frac{(u - u^{(k-1)})^2}{|\nabla u|^2}$$

is positive. This can be motivated from classical regularization theory. Set for simplicity of notation $k = 1$ and $u^\delta = u^{(0)}$; suppose that the measurement error satisfies the relative error criterion

$$\frac{(u - u^\delta)^2}{|\nabla u|^2} \leq \delta^2 \text{ uniformly in } \Omega,$$

then for α satisfying $\frac{\delta^2}{\alpha} < 2$ the diffusion terms stays positive. If $\frac{\delta^2}{\alpha} \rightarrow 0$ then the diffusion term tends to the diffusion term of the mean curvature flow equation. Conditions such as $\frac{\delta^2}{\alpha} < \infty$ and $\frac{\delta^2}{\alpha} \rightarrow 0$ are typically used in regularization of ill-posed problems and guarantee stability and convergence of the regularized solutions (see e.g. [28, 21]).

In summary both the iterative semi-implicit and the iterative implicit method can be considered as time discrete approximations of the mean curvature flow equation.

5. Numerical results

In this section we compare numerically *morphological diffusion filtering* and *implicit relative error regularization*. A more detailed discussion on the numerical minimization of relative error functionals is given in [29].

We present three numerical examples of denoising:

- (1) We used the artificially generated data set at the top of Figure 2. The several reconstructions in Figure 3 have been performed with bounded variation regularization (1.1) with $p = 0$; mean curvature filtering, affine mean curvature filtering, and implicit error regularization (1.2). The stopping time in the diffusion filtering method and the regularization parameters were selected such that all reconstructions have about the same amplitudes.
- (2) The second example concerns denoising of an MR data set which has been artificially covered with Gaussian noise.

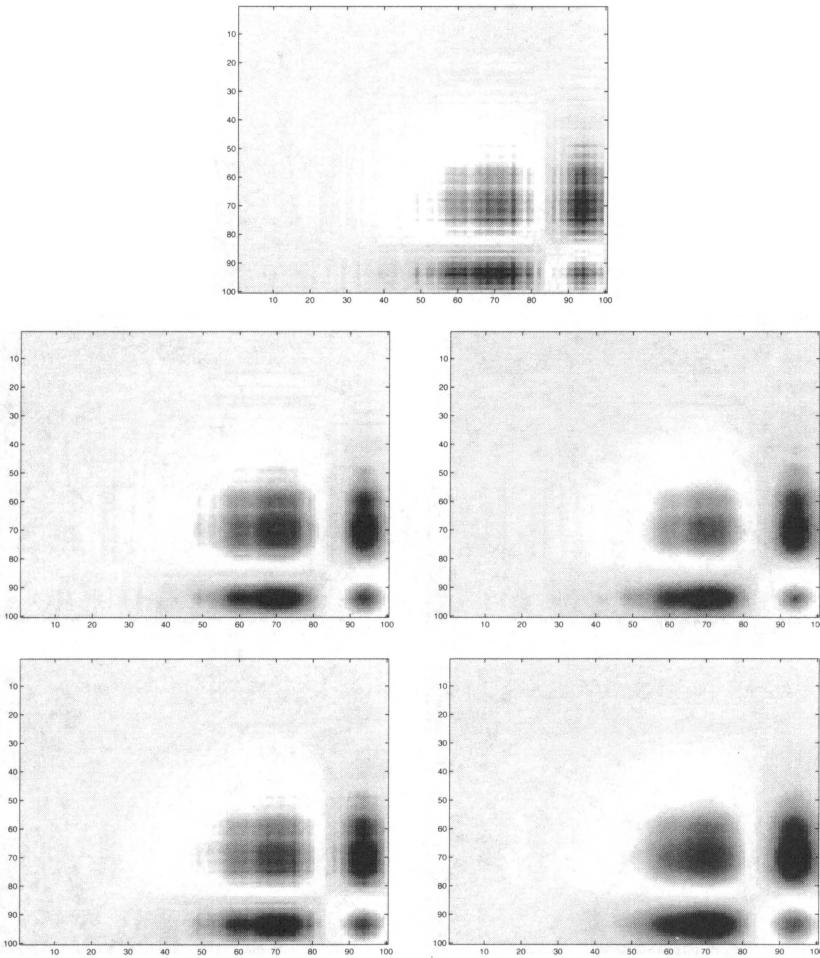


FIGURE 3. Original image (top) and filter images: middle left: mean curvature flow; middle right: affine mean curvature flow; bottom left: implicit regularization; bottom right: *BV* regularization

- (3) The third example is concerned with the denoising of an ultrasound data set.

From the numerical reconstructions one finds that mean curvature flow and implicit error regularization produce very similar results if the regularization parameter and the diffusion time are identified. When comparing standard regularization techniques and diffusion filtering methods (non morphological methods) this observation has been made before in [47, 39, 40].

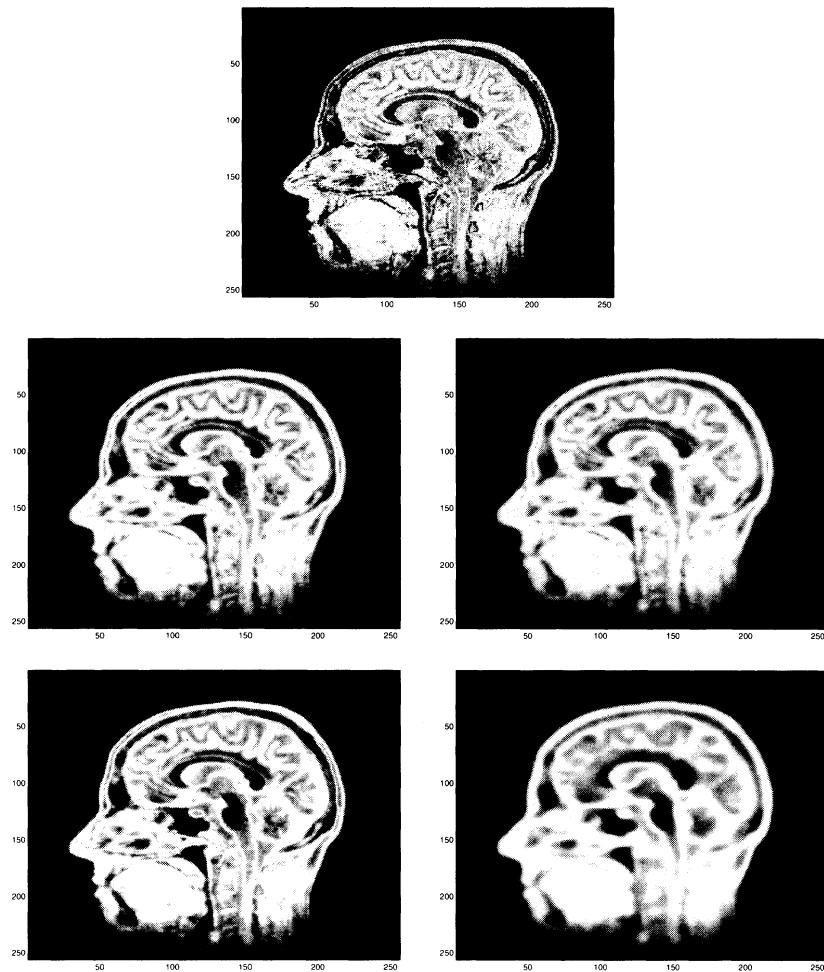


FIGURE 4. Original image (top) and filter images: middle left: mean curvature flow; middle right: affine mean curvature flow; bottom left: implicit regularization; bottom right: *BV* regularization

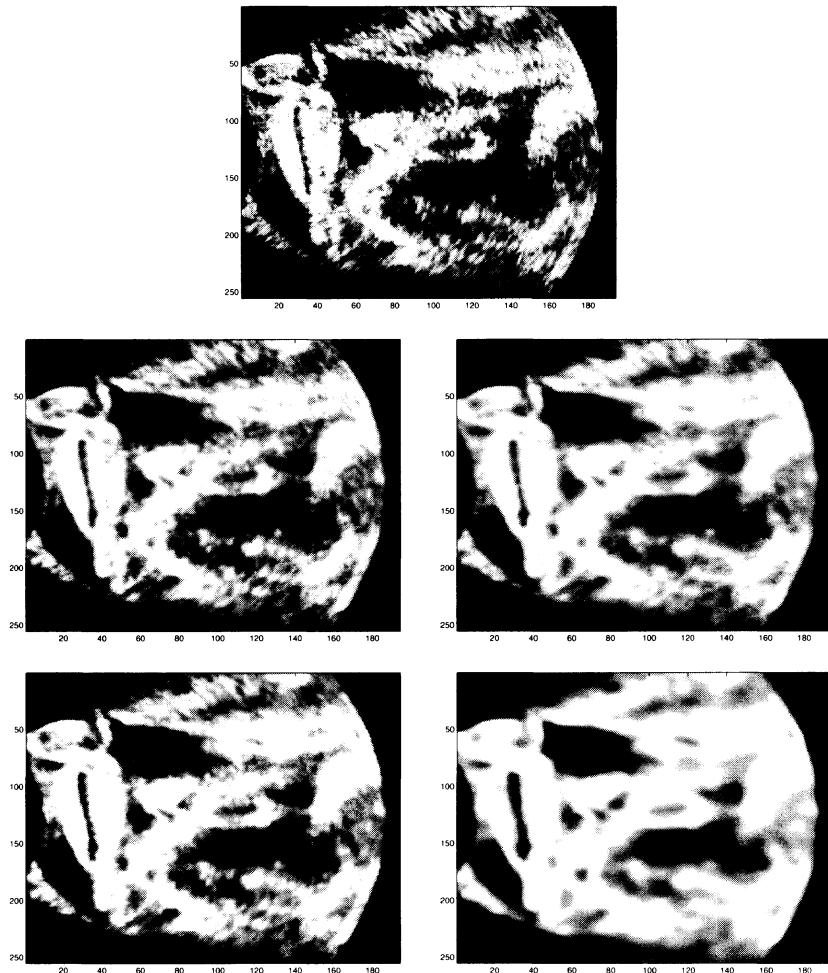


FIGURE 5. Original image (top) and filter images: middle left: mean curvature flow; middle right: affine mean curvature flow; bottom left: implicit regularization; bottom right: BV regularization

6. Some Remarks on the Curvature of a BV -function

In this section we recall some elementary results from nonlinear functional analysis on subdifferentials.

Many numerical experiments with bounded variation regularization utilize the curvature

$$(6.1) \quad \nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right)$$

of the function u . Also the numerical results presented in this paper were obtained by solving the first order optimality condition, which involves the curvature term. The goal of this section is to provide a rigorous definition of curvature of a BV -function and hopefully provide some insight to this term. We deduce below that the curvature of a BV -function can be interpreted as the subdifferential of its BV -seminorm. This has some important consequences, since it allows us to *analytically* calculate regularized solutions on the space of functions of bounded variation – therefore providing test examples for numerical algorithms. Moreover, we can put this concept in context with recent work by Strong & Chan [49] on exact solutions with BV -regularization.

Two properties of the curvature are striking:

- (1) The subdifferential is *set valued*. Below we give analytical examples where the subdifferential consists of more than one element.
- (2) The subdifferential is *non local*. That is the curvature of a non-smooth function *cannot* be evaluated pointwise, which would be intuitive from its definition (6.1).

Numerical handling of set-valued and nonlocal operators seems extremely difficult as one can see for instance when solving the nonlinear differential equation, representing the first order optimality condition for BV -regularization. That is

$$u - u^\delta \in \alpha \nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right).$$

Note that this is an inclusion equation, since the curvature is set valued.

6.1. The terminology curvature. The terminology curvature originates from differential geometric arguments (see e.g. [38, 45, 3]) and is outlined here for the readers convenience. Let

$$\begin{aligned} c : [0, 2\pi] &\rightarrow \mathbb{R}^2 \\ p &\mapsto \begin{pmatrix} x(p) \\ y(p) \end{pmatrix} \end{aligned}$$

be a closed counter clockwise parametrized curve in \mathbb{R}^2 , then the standard definition (see e.g. [8]) of curvature is

$$\mathcal{K} = \frac{x_p y_{pp} - x_{pp} y_p}{(x_p^2 + y_p^2)^{3/2}},$$

where $._p$ denotes the derivative with respect to the curve parameter p . Let

$$\begin{aligned} C : [0, 2\pi] \times [0, \infty[&\rightarrow \mathbb{R}^2 \\ (p, t) &\mapsto \begin{pmatrix} x(p, t) \\ y(p, t) \end{pmatrix} \end{aligned}$$

be a timely varying counter clockwise oriented curve. We consider the *curvature based evolution process*

$$(6.2) \quad C_t(p, t) = \beta(\mathcal{K})(p, t)\nu(p, t),$$

where ν denotes the vector pointing outside in normal direction to the curve C , and β is an appropriate scalar valued function.

Let $f \in C^2(\overline{\Omega})$ be a locally invertible function, i.e., $|\nabla f| \neq 0$ in $\overline{\Omega}$. We assume that the zero level set

$$\mathcal{L}(t) := \{x : f(x, t) = 0\}$$

can be parametrized by a curve $C(\cdot, t)$ which evolves according to (6.2), then $f(C(p, t), t) = 0$ for all $p \in [0, 2\pi[$ and $t \in [0, \infty[$. Consequently, by differentiation with respect to t and p we get

$$\begin{aligned} \nabla f(C(p, t), t) \cdot C_t(p, t) + f_t(C(p, t), t) &= 0, \\ \nabla f(C(p, t), t) \cdot C_p(p, t) &= 0, \end{aligned}$$

for all $p \in [0, 2\pi[, t \in [0, \infty[$. The later equation shows that ∇f and $(y_p, -x_p)$ are proportional, i.e., $\nabla f(C(p, t), t) = \psi(p)(y_p, -x_p)^t$. This in turn implies that

$$\begin{aligned} (\psi^2 y_{pp})(p, t) &= (-f_{xx}f_y + f_{xy}f_x)(C(p, t), t) - \psi_p(p, t)f_x(C(p, t), t), \\ (\psi^2 x_{pp})(p, t) &= (-f_{yy}f_x + f_{xy}f_y)(C(p, t), t) + \psi_p(p, t)f_y(C(p, t), t). \end{aligned}$$

Consequently, we set

$$(6.3) \quad \mathcal{K} = \frac{f_y^2 f_{xx} - 2f_{xy}f_x f_y + f_{yy}f_x^2}{|\nabla f|^3}.$$

Moreover,

$$\nu(p, t) = \frac{1}{\sqrt{x_p^2 + y_p^2}} \begin{pmatrix} y_p(p, t) \\ -x_p(p, t) \end{pmatrix} = \frac{\nabla f}{|\nabla f|}.$$

Using the abbreviations H_f for the Hessian of f and \cdot^t for the transpose it follows that

$$\begin{aligned} (6.4) \quad \text{curv}(f) &= \nabla \cdot \left(\frac{\nabla f}{|\nabla f|} \right) \\ &= \frac{|\nabla f|^2 \Delta f - \nabla f^t H_f \nabla f}{|\nabla f|^3} \\ &= \mathcal{K}. \end{aligned}$$

This shows that the corresponding level set formulation in Eulerian coordinates of (6.2) is

$$(6.5) \quad f_t(x, t) = -\beta((\text{curv}f)(x, t))|\nabla f(x, t)|.$$

6.2. The subdifferential of non-smooth functionals. To introduce the curvature of non-smooth functions we use the concept of *Fenchel-duality* (see e.g. [20, 16, 54]) on a real Banach space X .

DEFINITION 6.1. Let X be a real Banach space with dual X^* and $\Psi : X \rightarrow \mathbb{R} \cup \{+\infty\}$. Moreover, let

$$D_\Psi := \{x \in X : \Psi(x) < +\infty\} \neq \emptyset.$$

Then $\Psi^* : X^* \rightarrow \mathbb{R} \cup \{+\infty\}$, defined by

$$\Psi^*(x^*) = \sup_{x \in D_\Psi} \{x^*[x] - \Psi(x)\},$$

is called the **conjugate** or **Fenchel transform** of Ψ .

In the following we summarize some basic properties of the Fenchel-Transform (see [23, 24]). The particular stated results are collected from Deimling [16, Definition 23.4]

THEOREM 6.2. [16, Proposition 23.2] *Let X be a real Banach space and $\Psi : X \rightarrow \mathbb{R} \cup \{+\infty\}$ a lower semi-continuous convex function¹, then Ψ^* is convex and weak*-lower semi-continuous*

DEFINITION 6.3. The *subdifferential* of Ψ (as introduced by Moreau [32] and Rockafellar [41] (for $X = \mathbb{R}^N$), see also [6]) is defined by

$$\partial\Psi(x) := \{w \in X^* : w[x] = \Psi(x) + \Psi^*(w)\}.$$

Note that the subdifferential is in general *set valued*.

There are other definitions of subdifferentials: The classical definition of a subdifferential is as follows: $w \in X^*$ is in $\partial^1\Psi(x)$ if

$$\Psi(\zeta) \geq \Psi(x) + w[\zeta - x].$$

In \mathbb{R}^n the two definitions of subdifferentials correspond:

THEOREM 6.4. (see e.g. [42]) *For any lower semi continuous, proper, convex function f one has*

$$f(x) + f^*(w) = w \cdot x \text{ if and only if } w \in \partial^1 f(x).$$

I.e.

$$w \in \partial f(x) \text{ if and only if } w \in \partial^1 f(x).$$

In fact the definitions of subdifferentials corresponds if $\Psi : X \rightarrow \mathbb{R}$ is convex.

6.3. Definition of curvature of BV-functions. In the following we apply the general results of Subsection 6.2 to the mapping

$$\hat{\Psi}(f) = \|Df\|$$

defined on an appropriate subspace of *functions of bounded variation*.

We will show that the subdifferential of $\hat{\Psi}$ generalizes the concept of curvature of smooth functions.

To prove this we introduce some notation and recall some basic properties of the functional $\hat{\Psi}$.

DEFINITION AND LEMMA 6.5. *Let $\Omega \subseteq \mathbb{R}^n$ be bounded with $\partial\Omega$ Lipschitz.*

$$BV_{\text{mean}} := \{f \in BV(\Omega) : \int_{\Omega} f(x) dx = 0\}.$$

BV_{mean} with norm $\hat{\Psi}$ is a Banach space.

¹A function $\Psi : X \rightarrow \mathbb{R} \cup \{+\infty\}$ is lower semi-continuous if for all $t \in \mathbb{R}$ the sets $X_t := \{x \in X : \Psi(x) \leq t\}$ are closed.

The proof of the Banach space property follows from the Sobolev embedding theorem.

Since the classical curvature is *shift invariant* with respect to constant functions, i.e., $\text{curv}(f + c) = \text{curv}(f)$ for any constant function c , we require an analogous property for the curvature of non-smooth functions and we define the curvature of an arbitrary function $f \in BV(\Omega)$ by the curvature of $\hat{f} = f - \frac{1}{\text{meas}(\Omega)} \int_{\Omega} f dx$.

In the following we state some elementary properties of the functional $\hat{\Psi}$.

LEMMA 6.6. *The functional $\hat{\Psi}(f)$ is lower semi-continuous and convex on BV_{mean} .*

PROOF. Since $\hat{\Psi}$ is a norm on BV_{mean} it is per definition continuous. Any norm is convex and every continuous functional is lower semi-continuous (see e.g. [55, p760]) which yields the assertion. \square

From Lemma 6.6 and the results in Section 6.2 we immediately get:

THEOREM 6.7. *Let $X = BV_{\text{mean}}$ and let $\hat{\Psi}(f) = \|Df\|$ the bounded variation seminorm. Then $\hat{\Psi}^*$ is weak*-lower semi-continuous.*

Applying Theorem 6.2 in the BV_{mean} -setting we immediately deduce the following well-known result from nonlinear functional analysis; we include a proof for the readers convenience.

PROPOSITION 6.8. (1) *For any $f^* \in BV_{\text{mean}}^*$ satisfying $\|f^*\|_{BV_{\text{mean}}^*} \leq 1$ we have $\hat{\Psi}^*(f^*) = 0$.*
(2) *For any $f^* \in BV_{\text{mean}}^*$ satisfying $\|f^*\|_{BV_{\text{mean}}^*} > 1$ we have $\hat{\Psi}^*(f^*) = +\infty$.*

PROOF. For $f^* \in BV_{\text{mean}}^*$ we have by definition

$$\|f^*\|_{BV_{\text{mean}}^*} = \sup_{\{f \in BV_{\text{mean}} : \hat{\Psi}(f) = 1\}} f^*[f].$$

For any $\varepsilon > 0$ let $f^\varepsilon \in BV_{\text{mean}}$ with $\hat{\Psi}(f^\varepsilon) = 1$ satisfying

$$f^*[f^\varepsilon] - \varepsilon \leq \|f^*\|_{BV_{\text{mean}}^*} \leq f^*[f^\varepsilon] + \varepsilon.$$

(1) Let $\|f^*\|_{BV_{\text{mean}}^*} > 1$ and set $0 < \varepsilon < \|f^*\|_{BV_{\text{mean}}^*} - 1$. For any positive constant C and test function Cf^ε we have

$$\begin{aligned} \hat{\Psi}^*(f^*) &= \sup_{f \in BV_{\text{mean}}} \{f^*[f] - \hat{\Psi}(f)\} \\ &\geq C \{f^*[f^\varepsilon] - \hat{\Psi}(f^\varepsilon)\} \\ &\geq C(\|f^*\|_{BV_{\text{mean}}^*} - \varepsilon - 1). \end{aligned}$$

The right hand side converges to ∞ as C converges to ∞ and thus $\hat{\Psi}^*(f^*) = \infty$ if $\|f^*\|_{BV_{\text{mean}}^*} > 1$.

(2) If $\|f^*\|_{BV_{\text{mean}}^*} \leq 1$, then

$$\hat{\Psi}^*(f^*) = \sup_{f \in BV_{\text{mean}}} \{f^*[f] - \hat{\Psi}(f)\} \leq (\|f^*\|_{BV_{\text{mean}}^*} - 1) \inf_{f \in BV_{\text{mean}}} \hat{\Psi}(f).$$

Thus $\hat{\Psi}^*(f^*) \leq 0$. Taking $f = 0$ shows that $\hat{\Psi}^*(f^*) = 0$ for $\|f^*\|_{BV_{\text{mean}}^*} \leq 1$. \square

From Proposition 6.8 it follows that $\hat{\Psi}^*$ is an indicator function for the closed unit-ball in BV_{mean}^* . Thus the **subdifferential** of $\hat{\Psi}(f)$ (as introduced in Definition 6.3) is

$$(6.6) \quad \partial\hat{\Psi}(f) = \{f^* \in BV_{\text{mean}}^*, \|f^*\|_{BV_{\text{mean}}^*} \leq 1 : f^*[f] = \hat{\Psi}(f)\}.$$

Thus formally we have

$$\partial\hat{\Psi}(f) = \left\{ f^* = -\mathcal{K} : \mathcal{K} \text{ formally equals } \nabla \cdot \left(\frac{Df}{|Df|} \right) \right\}$$

where \mathcal{K} is an element of the *curvature* of f .

DEFINITION 6.9. Let $f \in BV(\Omega)$, then $\partial\hat{\Psi}(f)$ is the negative curvature of f .

In the following we analytically calculate the curvature of several one dimensional functions to get a feeling for the curvature term.

6.4. Analytical examples.

EXAMPLE 6.10. An element of the curvature of

$$f(t) := \begin{cases} -1 & -\frac{1}{2} < t < 0 \\ 1 & 0 < t < \frac{1}{2} \end{cases}$$

is

$$\mathcal{K}(t) := \begin{cases} 2 & -\frac{1}{2} < t < 0 \\ -2 & 0 < t < \frac{1}{2} \end{cases}$$

To see this we note that the curvature is the negative subdifferential $f^* = -\mathcal{K}$ and that

$$-\mathcal{K}[f] - \|Df\|(-1/2, 1/2) = - \left(\int_0^{1/2} -2 \, dx + \int_{-1/2}^0 -2 \, dx \right) - 2 = 0.$$

It remains to be proven that $\|\mathcal{K}\|_{BV_{\text{mean}}^*} \leq 1$. From the definition of $\|\cdot\|_{BV_{\text{mean}}^*}$ it follows that

$$(6.7) \quad \begin{aligned} \|\mathcal{K}\|_{BV_{\text{mean}}^*} &= \sup_{\{g \in BV_{\text{mean}} : \|Dg\|(-1/2, 1/2) = 1\}} \mathcal{K}[g] \\ &= \sup_{\{g \in BV_{\text{mean}} : \|Dg\|(-1/2, 1/2) = 1\}} \int_{-1/2}^{1/2} \mathcal{S}Dg \end{aligned}$$

where $\mathcal{S}(t) = 1 - 2|t|$. To get (6.7) rigorous, we note that for any function $g \in BV_{\text{mean}}$ there exists $g_k \in C_c^\infty(-1/2, 1/2)$ satisfying $\int_{-1/2}^{1/2} g_k \, dx = 0$ such that $g_k \rightarrow g \in L^1(-1/2, 1/2)$ and $\|Dg_k\|(-1/2, 1/2) \rightarrow \|Dg\|(-1/2, 1/2)$. Since $\mathcal{K} \in L^\infty(-1/2, 1/2)$ we see that $\mathcal{K}[g_k] \rightarrow \mathcal{K}[g]$. By integration by parts it follows

$$\sup_{\{g \in BV_{\text{mean}} \cap C_c^\infty : \|Dg\|(-1/2, 1/2) = 1\}} g[f] \leq 1.$$

This shows (6.7). Thus

$$\|\mathcal{K}\|_{BV_{\text{mean}}^*} \leq \max_{t \in (-1/2, 1/2)} |\mathcal{S}(t)| \leq 1,$$

which proves the assertion.

Via trace operations elements of the curvature of a BV -function can easily be found: Take for instance

$$\begin{aligned} f^* : BV_{\text{mean}} &\rightarrow \mathbb{R}, \\ v &\rightarrow T_{1/2}[v] - T_{-1/2}[v] \end{aligned}$$

where T_x denotes the trace operator, defined on BV , at the point x . Then for $\mathcal{K} = -f^*$

$$-\mathcal{K}[f] - \|Df\|(-1/2, 1/2) = 0.$$

Since $\|f^*\|_{BV_{\text{mean}}} \leq 1$ we have another representant of the subgradient.

EXAMPLE 6.11. Let $f(t) = \sin(t)$ in $[-\pi, 0]$. Let

$$f^* : BV_{\text{mean}} \rightarrow \mathbb{R}.$$

$$u \rightarrow T_0[u] - 2T_{-\pi/2}[u] + T_{-\pi}[u]$$

This operator is well-defined (see e.g. [22]). Since

$$f^*[f] = 2 \text{ and } \|Df\|(-\pi, 0) = 2$$

we get

$$-\mathcal{K}[f] = f^*[f] = \hat{\Psi}(f) = 2.$$

A function f in BV_{mean} and total variation 1 satisfies

$$|T_0[f] - 2T_{-\pi/2}[f] + T_{-\pi}[f]| \leq |T_0[f] - T_{-\pi/2}[f]| + |T_{-\pi/2}[f] - T_{-\pi}[f]| \leq \|f\|_{BV_{\text{mean}}}.$$

This shows that $\|\mathcal{K}\|_{BV_{\text{mean}}^*} \leq 1$.

In general the curvature of a BV_{mean} -function can be defined via considering all subregions where the function is monotone. Let $[a_i, b_i]$, $i \in \mathcal{I}$ be a subdivision of Ω into regions where f is monotone, then an element of the subgradient is

$$(6.8) \quad f^*[v] = \sum_{i \in \mathcal{I}} c_i(T_{b_i}[v] - T_{a_i}[v]).$$

with appropriate weighting constants $\{c_i\}$.

6.5. Applications. A widely inspected model in image processing is *bounded variation regularization* (see e.g. [1, 17, 9, 13, 19, 51, 30, 35, 47] to name but a few) which consists in minimizing the functional

$$(6.9) \quad \min_{u \in BV(\Omega)} \frac{1}{2} \|u - u^\delta\|_{L^2(\Omega)}^2 + \alpha \|Du\|.$$

For $u^\delta \in L^2(\Omega)$ and $\alpha > 0$ there exists a unique minimizer in $BV(\Omega)$, which can as well be characterized as the solution of the elliptic set valued differential equation

$$\frac{u - u^\delta}{\alpha} \in \partial \hat{\Psi}(u).$$

Some analytical results in [49] bring additional light to the definition of curvature of a function. Let u^δ be the test function from Example 6.10. The considerations in Strong & Chan [49] show that for sufficiently small $\alpha > 0$ the minimizer u_α of (6.9) is

$$u_\alpha(t) := \begin{cases} -1 + 2\alpha & -\frac{1}{2} < t < 0 \\ 1 - 2\alpha & 0 < t < \frac{1}{2} \end{cases}$$

Therefore,

$$\lim_{\alpha \rightarrow 0} \frac{u_\alpha - f}{\alpha}(t) = \begin{cases} 2 & -\frac{1}{2} < t < 0 \\ -2 & 0 < t < \frac{1}{2} \end{cases}$$

If the following two equations hold (which seems plausible by assuming that we can carry over standard argument of convex analysis in Hilbert spaces, but so far

we have no rigorous analysis), then by the usual identification $u_\alpha(x) = u(x, \alpha)$ and $u_0(x) = u(x, 0) = u^\delta(x)$ we get

$$\lim_{\alpha \rightarrow 0} \frac{u_\alpha - u^\delta}{\alpha}(t) = \frac{\partial u^\delta}{\partial t}(t)$$

and $\lim_{\alpha \rightarrow 0} \partial \hat{\Psi}(u_\alpha) = \partial \hat{\Psi}(u^\delta)$

we find $\partial \hat{\Psi}(u^\delta) = \mathcal{K}$, where \mathcal{K} is as in Example 6.10. In particular regularization selects an element of the subdifferential.

In a recent paper Andreu, Ballester, Caselles, Mazón [4, 5] proved the existence of the parabolic process

$$(6.10) \quad \frac{\partial u}{\partial t} \in \nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right)$$

via *nonlinear semigroup theory* and the theory of *accretive operators*. They used set-valued arguments to prove existence and uniqueness of (6.10).

References

- [1] R. Acar and C.R. Vogel. Analysis of bounded variation penalty methods for ill-posed problems. *Inverse Probl.*, 10:1217–1229, 1994.
- [2] R.A. Adams. *Sobolev Spaces*. Academic Press, New York, 1975.
- [3] L. Alvarez and J.-M. Morel. Formalization and computational aspects of image analysis. *Acta Numerica*, pages 1–59, 1994.
- [4] F. Andreu, C. Ballester, V. Caselles, and J. M. Mazón. Minimizing total variation flow. *C. R. Acad. Sci. Paris Sér. I Math.*, 331:867–872, 2000.
- [5] F. Andreu, C. Ballester, V. Caselles, and J. M. Mazón. Minimizing total variation flow. *Differential Integral Equations*, 14:321–360, 2001.
- [6] J.-P. Aubin. *Mutational and Morphological Analysis*. Birkhäuser, Boston, 1999.
- [7] M. Bardi, M.G. Crandall, L.C. Evans, H.M Soner, and P.E. Souganidis. *Viscosity Solutions and Applications*. Springer, Berlin, Heidelberg, 1997. Lecture Notes in Mathematics.
- [8] I.N. Bronstein, K.A. Semendjaew, G. Musiol, and H. Muehlig. *Taschenbuch der Mathematik, (Handbook of mathematics)*. Harri Deutsch, Frankfurt am Main, 1997. 3rd edition.
- [9] A. Chambolle and P.L. Lions. Image recovery via total variation minimization and related problems. *Numer. Math.*, 76:167–188, 1997.
- [10] T. F. Chan, G. H. Golub, and P. Mulet. A nonlinear primal-dual method for total variation-based image restoration. *SIAM J. Sci. Comput.*, 20:1964–1977 (electronic), 1999.
- [11] T. F. Chan and P. Mulet. On the convergence of the lagged diffusivity fixed point method in total variation image restoration. *SIAM J. Numer. Anal.*, 36:354–367 (electronic), 1999.
- [12] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proc. IEEE Int. Conf. Image Processing*, volume 2, pages 168–172, Los Alamitos, 1994. IEEE Computer Society Press. ICIP–94, Austin, Nov. 13–16, 1994.
- [13] G. Chavent and K. Kunisch. Regularization of linear least squares problems by total bounded variation. *ESAIM Control Optim. Calc. Var.*, 2:359–376 (electronic), 1997.
- [14] A. Cohen, R. DeVore, P. Petrushev, and H. Xu. Nonlinear approximation and the space $BV(\mathbb{R}^2)$. *Amer. J. Math.*, 121:587–628, 1999.
- [15] M.G. Crandall. Viscosity solutions: a primer. In [7], pages 1–43, 1997.
- [16] K. Deimling. *Nonlinear Functional Analysis*. Springer-Verlag, New York, 1985.
- [17] D.C. Dobson and F. Santosa. An image enhancement technique for electrical impedance tomography. *Inverse Probl.*, 10:317–334, 1994.
- [18] D.C. Dobson and O. Scherzer. Analysis of regularized total variation penalty methods for denoising. *Inverse Probl.*, 12:601–617, 1996.
- [19] D.C. Dobson and C.R. Vogel. Convergence of an iterative method for total variation denoising. *SIAM J. Num. Anal.*, 34:1779–1791, 1997.

- [20] I. Ekeland and R. Temam. *Convex Analysis and Variational Problems*. North Holland, Amsterdam, 1976.
- [21] H.W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, Dordrecht, 1996.
- [22] L.C. Evans and R.F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, Boca Raton, 1992.
- [23] W. Fenchel. On conjugate convex functions. *Canad. J. Math.*, 1:73–77, 1949.
- [24] W. Fenchel. *Convex Cones, Sets and Functions*. Princeton University Press, Princeton, NJ, 1951.
- [25] I. Fonseca and St. Müller. Quasi-convex integrands and lower semicontinuity in L^1 . *SIAM J. Math. Anal.*, 23:1081–1098, 1992.
- [26] I. Fonseca and St. Müller. Relaxation of quasiconvex functionals in $BV(\Omega, \mathbb{R}^p)$ for integrands $f(x, u, \nabla u)$. *Arch. Ration. Mech. Anal.*, 123:1–49, 1993.
- [27] E. Giusti. *Minimal Surfaces and Functions of Bounded Variation*. Birkhäuser, Boston, 1984.
- [28] C.W. Groetsch. *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*. Pitman, Boston, 1984.
- [29] W. Hinterberger and O. Scherzer. Well-posedness of a class of non-convex minimization problems in image processing. 2002. work in progress.
- [30] K. Ito and K. Kunisch. An active set strategy based on the augmented Lagrangian formulation for image restoration. *M2AN Math. Model. Numer. Anal.*, 33:1–21, 1999.
- [31] A. S. Leonov. Application of functions of several variables with limited variations for piecewise uniform regularization of ill-posed problems. *J. Inverse Ill-Posed Probl.*, 6:67–93, 1998.
- [32] J. Moreau. Weak and strong solutions of dual problems. In [53], 1971.
- [33] J.M. Morel and S. Solimini. *Variational Methods in Image Segmentation*. Birkhäuser, Boston, 1995.
- [34] V.A. Morozov. *Methods for Solving Incorrectly Posed Problems*. Springer Verlag, New York, Berlin, Heidelberg, 1984.
- [35] M.Z. Nashed and O. Scherzer. Least squares and bounded variation regularization with non-differentiable functional. *Num. Funct. Anal. and Optimiz.*, 19:873–901, 1998.
- [36] M. Nielsen, P. Johansen, O.F. Olsen, and J. Weickert, editors. *Scale-Space Theories in Computer Vision*. Springer, Berlin, 1999. Lecture Notes in Computer Science.
- [37] N. Nordström. Biased anisotropic diffusion – a unified regularization and diffusion approach to edge detection. *Image and Vision Computing*, 8:318–327, 1990.
- [38] S. Osher and J. A. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations. *J. Comput. Phys.*, 79:12–49, 1988.
- [39] E. Radmoser, O. Scherzer, and J. Weickert. Scale-space properties of regularization methods. In [36], 1999.
- [40] E. Radmoser, O. Scherzer, and J. Weickert. Scale-space properties of nonstationary iterative regularization methods. *Journal of Visual Communication and Image Representation*, 11:96–114, 2000.
- [41] R. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.
- [42] R. T. Rockafellar. *The theory of subgradients and its applications to problems of optimization*. Heldermann Verlag, Berlin, 1981. Convex and nonconvex functions.
- [43] L.I. Rudin, St. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.
- [44] W. Rudin. *Real and Complex Analysis*. McGraw-Hill, New York, 1974. 2nd edition, 1st edition 1966.
- [45] G. Sapiro and A. Tannenbaum. Affine invariant scale-space. *Int. J. Comput. Vision*, 11:25–44, 1994.
- [46] O. Scherzer. Stable evaluation of differential operators and linear and nonlinear multi-scale filtering. *Electronic Journal of Differential Equations*, 15:1–12, 1997. <http://ejde.math.unt.edu>.
- [47] O. Scherzer and J. Weickert. Relations between regularization and diffusion filtering. *J. Math. Imag. Vision*, 12:43–63, 2000.
- [48] C. Schnörr. Unique reconstruction of piecewise smooth images by minimizing strictly convex non-quadratic functionals. *J. Math. Imag. Vision*, 4:189–198, 1994.
- [49] D. Strong and T. F. Chan. Exact solutions to the total variation regularization problem. Technical report, University of California, Los Angeles, 1996. CAM 96 - 41.

- [50] C.R. Vogel and M.E. Oman. Iterative methods for total variation denoising. *SIAM J. Sci. Comput.*, 17:227–238, 1996.
- [51] J. Weickert. *Anisotropic Diffusion in Image Processing*. Teubner, Stuttgart, 1998.
- [52] J. Weickert. On discontinuity-preserving optic flow. *Diku, Copenhagen*, 1998. preprint,submitted.
- [53] E. Zarantello, editor. *Contributions to nonlinear Functional Analysis*. Academic Press, New York, 1971.
- [54] E. Zeidler. *Nonlinear Functional Analysis and its Applications III*. Springer–Verlag, New York, 1985.
- [55] E. Zeidler. *Nonlinear Functional Analysis and its Applications I*. Springer–Verlag, New York, 1993. corrected printing.

DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF INNSBRUCK, TECHNIKER STR. 25,
A-6020 INNSBRUCK, AUSTRIA

E-mail address: otmar.scherzer@uibk.ac.at

SAMPLING METHODS FOR APPROXIMATE SOLUTION OF PDE

Frank Stenger[†], Ahmad Reza Naghsh–Nilchi, Jenny Niebsch, and Ronny Ramlau

ABSTRACT. This paper describes a novel procedure that combines indefinite convolution and *Sinc* approximation, for solving PDE (partial differential equations). The PDE is first transformed to an equivalent integral equation, this “Sinc convolution” procedure then enables a remarkable method of separation of variables to solve the problem. Whereas different numerical methods are used in practice for solving elliptic, parabolic, and hyperbolic PDE, the present paper uses essentially the same procedure for all three of these equations, over bounded, unbounded, and even curvilinear regions. The time complexity of computation to solve a d -dimensional problem on a sequential machine (i.e., the amount of time required to obtain a solution to within a uniform error of ϵ) under suitable assumptions of analyticity, allowing for possible singularities in the coefficients on the boundaries of the regions is of the order of $(\log(\epsilon))^{2d+2}$. The method also lends itself readily to parallel computation, although we have not illustrated this feature in this paper. We do not need to store the large matrices that current methods require, enabling us to achieve high accuracy, whereas this is not possible via current algorithms. Examples of solutions are presented for every type of equation, and time comparisons are made with efficient existing methods.

1. Introduction and Summary

Sinc methods offer a variety of approaches for solving PDE [15, 16, 1]. In the present article we present a unified approach to solve the integral equation formulation of solutions to each of the three classes of PDE, via use of *indefinite convolutions* combined with *Sinc approximation*, a procedure described in [16] and [15], §4.6, and henceforth in this paper to be referred to as *Sinc convolution*. We expect that the algorithms which we present

2000 *Mathematics Subject Classification.* 35A35, 35A40, 65M70, 65N35, 65P05, 65R20.

[†]Supported by USAF Contract # F33615–00–C–5004.

here via use of *Sinc methods* can also be carried out via use of wavelets [4], although we have not yet attempted this latter approach. The main tool, i.e., the *Sinc convolution* procedure is derived in [16] for approximating one-dimensional indefinite convolution integrals. It is readily extended to the approximation of definite convolution integrals, and to the approximation of multidimensional definite and indefinite integrals.

The method of separation of variables is popular, and taught regularly in undergraduate engineering classes. Its popularity possibly stems from Chapter 5 of [10], where it is stated:

“... once the Green’s function is found, as shown in Chapter 7, any desired solution of the homogeneous or inhomogeneous equation may be found, in principle. However, as the phrase “in principle” indicates, the integral solution is not always the most satisfactory solution, for in many cases, the integral cannot be integrated in closed form and numerical values are then extremely difficult to obtain.”

Furthermore, the authors of Chapter 5 of [10] then derive *all* of about a dozen of the transformations which enable separation of variables for the Helmholtz equation, $\nabla^2 u + k^2 u = 0$ in \mathbb{R}^3 . In effect, the transformations introduced in [10] transform the region over which a solution is sought onto a rectangular region, over which the solution can then be expressed as a sum of products of functions in each variable, in the form

$$u(x, y, z) = \sum_{ijk} a_{ijk} f_i(x) g_j(y) h_k(z)$$

In this paper we do, in fact, advocate the solution of PDE by solution of the integral equation representation of the PDE. Our method is based on *Sinc convolution*, which requires the multidimensional “Laplace transforms” of Green’s functions. Integral equation difficulties referred in the above quote are non-existent by our method. Furthermore, our method involves solution of such equations via separation of variables, enabling the solution of the PDE via performance of a few one-dimensional matrix multiplications.

The *Sinc convolution* procedure was first derived in [16]. This novel method of approximation has to date belonged solely to the family of Sinc methods. These methods enable the replacement of a vector of function values at Sinc points by another vector, depending on the approximation. For example, to get an accurate approximation of an indefinite integral such as

$$r(x) = \int_a^x g(t) dt, \quad x \in (a, b)$$

we would evaluate g at a set of *Sinc points* (see Example 3.1) $\{x_i\}_1^m$ on (a, b) , to form a vector $\mathbf{g} = (g_1, g_2, \dots, g_m)^T$, with $g_i = g(x_i)$. We

then pre-multiply \mathbf{g} by an explicitly defined matrix A (see Eq. (3.3)) to get another vector, $\mathbf{r} = A\mathbf{g} = (r_1, r_2, \dots, r_m)^T$, with $r_i \approx r(x_i)$, and where the accuracy increases with increasing m . Moreover, using *Sinc bases* (see Eq. (3.6)), $\{w_i(x)\}_1^m$, we would then be able to get an approximation for any $x \in (a, b)$, in the form

$$r(x) \approx \sum_1^m r_i w_i(x),$$

for which the (uniform) error is of the order of $e^{-cm^{1/2}}$, for some positive constant c . We could thus get a highly accurate approximation of integrals of functions with integrable singularities, such as (see Theorem 3.1)

$$\int_0^x \frac{\sqrt{1 + e^{-t}} \log(t) dt}{t^{1/4}(1 + t^{7/3})}, \quad x \in (0, \infty),$$

which we are able to approximate to e.g., 5 places of accuracy via use of a matrix of size $m = 20$ (see Theorem 3.5).

The Sinc method of indefinite integration is basic to Sinc convolution, which in one dimension, enables us to accurately approximate integrals of the form

$$p(x) = \int_0^x f(x - t) g(t) dt, \quad x \in (a, b).$$

To do this, we need a vector \mathbf{g} of values of $g(x_i)$ defined as above, as well as the “Laplace transform”,

$$F(s) = \int_0^c f(t) \exp(-t/s) dt$$

where $c \geq b - a$, but is otherwise arbitrary. Then, as is shown in [16], the $m \times m$ matrix $F(A)$ with A defined as above is well defined, and the vector $\mathbf{p} = (p_1 \ p_2, \dots \ p_m)^T$ given by $\mathbf{p} = F(A)\mathbf{g}$ yields a uniformly accurate approximation (see Theorem 3.6)

$$p(x) \approx \sum_1^m p_i w_i(x) \quad x \in (a, b).$$

A definite convolution integral can be similarly dealt with, since a definite convolution integral can be written as a sum of two indefinite ones.

Of course, we can evaluate $F(A)\mathbf{g}$ by first diagonalizing $A = X S X^{-1}$, where X is the matrix of eigenvectors of A , and $S = \text{diag}(s_1, s_2, \dots, s_m)$ denotes the diagonal matrix of eigenvalues, and while this procedure requires a non-trivial amount of computation for a one dimensional problem, there is considerable “pay-off” when solving a d -dimensional PDE via *Sinc convolution*, since we then (usually) need to diagonalize only d such one-dimensional matrices.

Consider, for example, obtaining a particular solution to the Poisson problem

$$\nabla^2 U = -f \quad \text{in } B = (0, a) \times (0, b).$$

Such a particular solution can be expressed in the form of the two dimensional convolution integral,

$$u(\bar{r}) = \int \int_B \mathcal{G}(\bar{r} - \bar{\rho}) f(\bar{\rho}) d\bar{\rho}$$

where $\bar{r} = (x, y)$, $\bar{\rho} = (\xi, \eta)$, and where $\mathcal{G}(\bar{r} - \bar{\rho}) = -(2\pi)^{-1} \log |\bar{r} - \bar{\rho}|$ is the Green's function. After expressing this two dimensional definite integral as a sum of four indefinite integrals, the method of this paper enables us to compute the entries of an $m \times m$ matrix $U = [U_{ij}]$ where $u(x_i, y_j) \approx U_{ij}$, yielding an approximate solution of the form

$$u(x, y) \approx \sum_{i=1}^m \sum_{j=1}^m U_{ij} w_i^{(1)}(x) w_j^{(2)}(y),$$

where the $w_k^{(\ell)}$ are one-dimensional *Sinc basis functions*. This approximate solution is uniformly accurate in B , even though f may have integrable singularities that “blow up” on the boundary of B . Upon defining an $m \times m$ matrix F by $F = [f(x_i, y_j)]$, the U can, in fact, be evaluated via the following simple four–statement *Matlab* program, (the second statement being equivalent to Algorithm 3.1),

```

U = X*(G.*((Xi.*F.*Xi.'))*X.' ;
U = U + Y*(G.*((Yi.*F.*Xi.'))*X.' ;
U = U + X*(G.*((Xi.*F.*Yi.'))*Y.' ;
U = U + Y*(G.*((Yi.*F.*Yi.'))*Y.' ;

```

in which X , $Xi = X^{-1}$, Y , $Yi = Y^{-1}$ and G are explicitly defined $m \times m$ matrices. In fact, $Y = PX$ where P is an explicitly defined permutation matrix of order m . The $(i, j)^{th}$ element of the $m \times m$ matrix G in this program example is just $\hat{G}(s_i, t_j)$, where

$$\begin{aligned} \hat{G}(s, t) &= - \int_0^\infty \int_0^\infty \left(\frac{1}{2\pi} \log \sqrt{x^2 + y^2} \exp(-x/s - y/t) \right) dx dy, \\ &= \left\{ \frac{1}{s^2} + \frac{1}{t^2} \right\}^{-1} \left\{ -\frac{1}{4} + \frac{1}{2\pi} \left[\frac{t}{s} (\gamma - \ln(t)) + \frac{s}{t} (\gamma - \ln(s)) \right] \right\}, \end{aligned}$$

where $\gamma \approx .5772\dots$ is Euler's constant, and where $S = \text{diag}(s_1, s_2, \dots, s_m)$ and $T = \text{diag}(t_1, t_2, \dots, t_m)$ are diagonal eigenvalue matrices of the indefinite integration matrices of order m over $(0, a)$ and $(0, b)$ respectively.

Eight such *Matlab* program statements are required to solve three space–dimensional, or three space and one time–dimensioanl problems. Furthermore, this separation of variables method also enables us to solve problems

over a union of curvilinear regions (see Eqs. (3.28)–(3.46), and or the Example preceding §2.2 of this paper), such as, e.g.,

$$\mathcal{B} = \{(x, y) : a_1 < x < b_1, a_2(x) < y < b_2(x)\}.$$

or over a rotation of such regions, over infinite regions, over a union of such regions, and similarly in three or more dimensions (see §3.4 of this paper). Moreover, essentially the same procedure as the above one for Poisson problems works also for hyperbolic and parabolic PDE. We furthermore have one additional advantage over classical methods of solution, in that we only need to store a relatively small number of $m \times m$ matrices to produce a solution at $\mathcal{O}(m^d)$ points in a d -dimensional region, whereas classical methods require the storage of a much larger matrix.

We have been fortunate to be able to obtain explicit expressions for all of the multidimensional “Laplace transforms” of the standard free-space Greens functions, in all dimensions, for Poisson problems, for heat problems and for wave problems. These results are essential for our method of solving PDE, and we thus present them in this paper.

Let us summarize some features of the *Sinc convolution* procedure.

- (1) Only the Laplace transform (or its accurate approximation) of f is required to get an accurate approximation of convolution integrals of the form

$$\int_a^x f(x-t) g(t) dt, \quad \int_x^b f(x-t) g(t) dt. \quad (1.1)$$

The accurate approximation of such integrals, including e.g., *Abel-type* integrals, was hitherto difficult, especially in important cases of when $f(t)$ has an integrable singularity at $t = 0$. Moreover, being able to accurately approximate each of the integrals in (1.1) enables us to accurately approximate definite integral convolutions of the form

$$\int_a^b f(x-t) g(t) dt. \quad (1.2)$$

- (2) Whereas Fourier transforms (FFT) can also be used to approximate the integrals (1.1), FFT converges very slowly in cases of when $f(t)$ has an integrable singularity at $t = 0$, or when (a,b) is finite or semi-infinite, and/or when g has isolated singularities on (a,b) . For example, the *Sinc convolution* procedure offers a remarkably simple method of solving *Wiener-Hopf* integral equations.
- (3) As already mentioned above, the *Sinc convolution* procedure enables a surprising “separation of variables” procedure for solving

PDE. This procedure is analogous to that used to solve multidimensional problems in e.g., PDE. This makes it possible to accurately approximate the multidimensional convolution integrals via use of one-dimensional matrix multiplications. It is applicable in all dimensions.

- (4) This separation of variables feature of the method enables solution of the PDE via parallel computation, although we have not used parallel computation in our illustrative examples in this paper.
- (5) A bi-product of using one-dimensional matrix multiplications is that we need not set up the analogous “big matrices” that are required for the solution of PDE via finite difference and finite element methods. This is one of the reasons why it is not possible to achieve high accuracy in solving PDE via these classical methods. Another reason is that the rate of convergence of the *Sinc convolution* technique is much more rapid. These points are illustrated in our comparison tests in this paper.
- (6) Finally, we add that the procedure can also be applied to convolution integrals over curvilinear regions in two or more dimensions, such as, e.g., in two dimensions, regions of the form

$$B = \{(x, y) : a_1 < x < b_1, a_2(x) < y < b_2(x)\} \quad (1.3)$$

and, of course, to translations and rotations, and unions of such regions. That is, it is applicable to the solution of most PDE problems over curvilinear regions arising in science and engineering.

The present paper is *not* a complete illustrations of solution of PDE via *Sinc methods*, i.e., we have not illustrated handling initial and/or boundary conditions. However, what we have omitted can, in fact be easily dealt with via *Sinc methods*, as has already been illustrated. That initial value ordinary differential equation problems can be easily dealt with via *Sinc methods* has been illustrated in the program package [13]. The present paper illustrates the construction of an approximate solution to a non-homogeneous PDE via use of Green’s functions. Given a PDE along with initial and or boundary conditions, this same Green’s function can be used to set up a boundary integral equation for determination of a solution of the homogeneous PDE, yielding a solution to the original PDE with the correct boundary conditions. That such boundary integral equations can in fact be efficiently and accurately solved via *Sinc methods* has already been illustrated in [15], §6.5, and [14, 3, 7, 5, 2, 6, 17, 18, 12]. Furthermore, the *Sinc convolution* procedure can also be used to solve integral equations, as we illustrate in this paper via a simply example, and as was illustrated in [11] on the solution of a five-dimensional convolution type integral equation.

We shall use *Sinc terminology* in the present paper. An excellent presentation of this terminology is given in [8]. For sake of completeness, we shall

include without proofs the essence of this terminology in §3 of the present paper. We also present the *Sinc convolution* algorithm in this section.

In what follows, in §2 we illustrate the *Sinc convolution* procedure to obtain approximate solutions of several PDE, and we make time comparisons with other existing methods for solving such problems. In order to achieve accuracies that are possible via *Sinc convolution*, the matrices required by classical methods to achieve such accuracies in more than three and 4 dimensions very quickly reached the capacity of our computer, making it difficult to make complexity comparisons.

The *Sinc terminology*, including the *Sinc convolution* technique is given in §3, while §4 contains explicit results of multidimensional Laplace transforms of standard Green's functions, based on a technique developed, essentially, in [11]. We have omitted proofs, in order to limit the length of this paper. Such proofs can be obtained from F. Stenger.

2. Applications

In this section we illustrate the application of the *Sinc convolution* procedures to obtain approximate solutions of elliptic, parabolic, and hyperbolic differential equations. We also illustrate the solution of an integral equation problem, as well as the solution of a PDE problem over a curvilinear region.

For all our numerical computations a two processor PC with Intel Pentium II (400 MHz), Linux operating system and 512MB main memory was used. The code for our algorithms was written in *Matlab*. To compare the results of the *Sinc convolution* computations with results obtained by *FEM*-methods, we used for the Poisson equation and Heat equation the program package KASKADE, developed at the Konrad-Zuse-Zentrum Berlin (ZIB), Germany (for a description of the used algorithm, program code and manual see <ftp://elib.zib.de/pub/kaskade>). For the two dimensional Wave equation the PDE–Toolbox of *Matlab* was used.

2.1. Sinc Convolution Solution of Poisson Problems.

(1) Our first illustration is that for the Poisson equation

$$\Delta \Psi(\bar{r}) = -g(\bar{r}), \quad \bar{r} \in V = \mathbb{R}^3, \quad (2.1)$$

Our *Sinc convolution* computations were based on the formula

$$\Psi(x, y, z) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} \int_{a_3}^{b_3} \frac{g(\xi, \eta, \zeta)}{4\pi\sqrt{(x-\xi)^2 + (y-\eta)^2 + (z-\zeta)^2}} d\xi d\eta d\zeta. \quad (2.2)$$

for $(x, y, z) \in V$. This multidimensional convolution integral can be readily split into 8 indefinite convolution integrals, such as, e.g.,

$$\Psi^{(1)}(x, y, z)$$

$$= \int_{a_1}^x \int_y^{b_2} \int_{a_3}^z \frac{g(\xi, \eta, \zeta)}{4\pi\sqrt{(x-\xi)^2 + (y-\eta)^2 + (z-\zeta)^2}} d\zeta d\eta d\xi. \quad (2.3)$$

To evaluate the 8 integrals of the form $\Psi^{(1)}$ at all of the *Sinc points* $\{(ih, jh, kh) : i = -N, \dots, N, j = -N, \dots, N, k = -N, \dots, N\}$, we require 3 transformations $\varphi_i : (a_i, b_i) = \mathbb{R} \rightarrow \mathbb{R}, i = 1, 2, 3$, but this is trivial, since each of the transformations is the same identity map. We thus determine $h_i = h$ and we form the matrices

$$\begin{aligned} A_i = A &= h I^{(-1)} = X_i S_i X_i^{-1}, \quad i = 1, 3 \\ A_2 = A^T &= h (I^{(-1)})^T = X_2 S_2 X_2^{-1}, \end{aligned} \quad (2.4)$$

where each $S_i = \text{diag}[s_{-M_i}^{(i)}, \dots, s_{N_i}^{(i)}]$ is a diagonal matrix of eigenvalues of the matrix A_i , and X_i is the corresponding matrix of eigenvectors. We then evaluate the array $[g_{ijk}] = [g(ih, jh, kh)]$, and we use the *Sinc convolution* algorithm (an explicit 3-dimensional version is given in [16]) to transform this array into an array $[\Psi_{ijk}^{(1)}]$, by means of the “Laplace transform” of the Greens function $1/(4\pi r)$ given in Lemma 4.3 of this paper. We then repeat this computation to get accurate approximations at the *Sinc points* for all of the remaining 7 *Sinc convolutions* $\Psi^{(\ell)}$, $\ell = 2, 3, \dots, 8$. We then have

$$[\Psi(ih, jh, kh)] = \left[\sum_{\ell=1}^8 \Psi^{(\ell)}(ih, jh, kh) \right]. \quad (2.5)$$

Using *Sinc interpolation*, we can then get an almost equally accurate approximation to the function Ψ at all points of V . It may moreover be shown, assuming that the function $g(\cdot, y, z)$ (and, additionally, making similar assumptions about the functions $g(x, \cdot, z)$ and $g(x, y, \cdot)$) is analytic on (a_1, b_1) , for all $(y, z) \in [a_2, b_2] \times [a_3, b_3]$ (which is, in fact so, for the case of this example) then the uniform error of approximation is of the order of $\exp(-c N^{1/2})$, with c a constant that is independent of N . In particular, for the case of the present problem we can take $c = \pi$.

Let us also give a more explicit picture of the (unstored) matrix that is involved in the above computation. Let us form a vector \mathbf{g} from the array $[g_{ijk}]$, in which the subscripts appear in the order (call it lexicographic) dictated by the order of appearance of the subscripts in the FORTRAN do loop, “DO k = $-M_3, N_3$ ”, followed by “DO j = $-M_2, N_2$ ”, followed by “DO i = $-M_1, N_1$ ”. We then

also form the diagonal matrix $\hat{\mathbf{G}}$ in which the entries are the values $\hat{G}_{ijk} = \hat{G}(s_i^{(1)}, s_j^{(2)}, s_k^{(3)})$, with the function \hat{G} and the eigenvalues $s_j^{(i)}$ defined as above, and where we also list the values \hat{G}_{ijk} in the same lexicographic order as for g_{ijk} . Then, similarly from the array $\Psi_{ijk}^{(1)}$, we can define a vector Ψ_1 by listing the elements $\Psi_{ijk}^{(1)}$ in lexicographic order. It can then be shown that Ψ_1 is defined by the matrix (Kronecker) product

$$\begin{aligned}\Psi_1 &= \Phi^{(1)} \mathbf{g} \\ \Phi^{(1)} &= X_3 \otimes X_2 \otimes X_1 \hat{\mathbf{G}} X_3^{-1} \otimes X_2^{-1} \otimes X_1^{-1},\end{aligned}\tag{2.6}$$

and moreover, the analogous vector Ψ approximating the function Ψ defined in (2.2) above at the *Sinc points* is then given by

$$\Psi = \left(\sum_{\ell=1}^8 \Psi^{(\ell)} \right) \mathbf{g}.\tag{2.7}$$

We emphasize that the matrices $\Psi^{(\ell)}$ and $\sum_{\ell} \Psi^{(\ell)}$ need never be computed, since our algorithm involves performing a sequence of one-dimensional matrix multiplications. For example, with $N = 20$ we get at least 6 places of accuracy, and the size of the corresponding matrix Ψ is $41^3 \times 41^3$, or $68,921 \times 68,921$. Such a matrix, which is full, contains more than 4.75×10^9 elements. If such a matrix were to be obtained by a Galerkin scheme, with each entry requiring the evaluation of a three dimensional integral, and with each integral requiring 41^3 evaluation points, then more than 3.27×10^{14} function evaluations would be required, an ominous task indeed! On the other hand, our method accurately gives us *all* of these values for relatively little work.

For a test computation, we used as right hand side $g(\bar{r}) = \exp(-r^2)(6 - 4r^2)$. The exact solution is then given by $\Psi(\bar{r}) = \exp(-r^2)$, $r = |\bar{r}|$.

The computational domain for FEM (finite element method) with which we compared our results was $[-6, 6]^3$ with zero boundary conditions. This restriction caused no problems because u and f are rapidly decreasing functions.

We computed the *Sinc*-based solution with $m = 2N+1$ *Sinc points*. The corresponding values of h were computed by $h = \pi/\sqrt{N}$. The CPU time is listed in Table 1. The FEM solution was computed afterwards using more and more knots until the accuracy of the *Sinc* solutions was achieved. We compared the accuracy of the

FEM			SINC		
accuracy ($\ \cdot\ _\infty$)	# of knots	CPU (sec)	accuracy ($\ \cdot\ _\infty$)	CPU (sec)	N
0.2829	158	0.85	0.171	5.3	10
0.1437	247	1.79	0.171		
0.0652	1003	6.16			
0.0244	4365	26.25	0.0166	21.54	15
0.0125	15373	98.71			
0.0125	15373	98.71			
0.0072	28140	229.47	0.0040	57	20
0.0062	53706	482.68			
0.0026	143497	1164.63			
***	***	***	1.1672e-05	2977	60
***	***	***	1.0734e-06	11611	80

TABLE 1. FEM and SINC costs for solution of Laplace equation. *** indicates that FEM was not able to achieve a similar accuracy.

Sinc with the FEM solution based on the maximum difference of exact and approximate solution at the FEM knots. The table shows that to achieve one point of accuracy (i.e. 10^{-1}) FEM is faster than our *Sinc method*. But even for two places of accuracy our *Sinc* procedure is three times as fast whereas three points cause problems with FEM methods concerning time and storage of the matrix. For *Sinc*, we were even able to compute an example with 161 *Sinc points* and an accuracy of 10^{-6} on the region $[-6, 6]^3$.

- (2) Next, we illustrate the solution via *Sinc convolution of an elliptic problem over a two dimensional curvilinear region*. Our region \mathcal{B} is the union of two separate regions, \mathcal{B}_1 and \mathcal{B}_2 . Thus, for evaluation of the Green's function integrals of the form

$$\int_{y \in \mathcal{B}_j} G(x - y) g_j(y) dy, \quad x \in \mathcal{B}_i,$$

we need to perform *Sinc convolution* only when $i = j$; whenever $i \neq j$ we can use ordinary *Sinc quadrature* to evaluate the integrals.

As mentioned above, the region is $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$, with

$$\begin{aligned} \mathcal{B}_1 &= \left\{ (x, y) : -\frac{3}{2} < x < 0, 0 < y < x^2 + \frac{\sqrt{3}}{2} \right\} \\ \mathcal{B}_2 &= \left\{ (x, y) : 0 < x < \frac{3}{2}, 0 < y < \sqrt{1 - \left(x - \frac{1}{2} \right)^2} \right\}. \end{aligned} \quad (2.8)$$

We are given functions g_1 and g_2 defined by the equations

$$\begin{aligned} g_1(x, y) &= c_1 (-x)^{1/7} \left(\frac{3}{2} - x \right)^{-6/7} y^{-1/4} \left(\frac{\sqrt{3}}{2} + x^2 - y \right)^{-3/4} \\ g_2(x, y) &= c_2 x^{\frac{1}{\sqrt{3}} - 1} y^{-1/3} \left(1 - \left(x - \frac{1}{2} \right)^2 - y^2 \right)^{-1/3}, \end{aligned} \quad (2.9)$$

where the constants c_1 and c_2 are selected so that

$$\int \int_{B_1} g_1(x, y) dx dy = - \int \int_{B_2} g_2(x, y) dx dy = 1,$$

$$\text{i.e., } c_1 = -\sin(\pi/7)/(\sqrt{2}\pi^2), \quad c_2 = 1/\left((3/2)^{1/\sqrt{3}}\pi\right).$$

Let us use the notation $\bar{\rho} = (x, y)$, $\rho = \sqrt{x^2 + y^2}$. The partial differential equation which we propose to solve is

$$\begin{aligned} \nabla^2 U(\bar{\rho}) &= g(\bar{\rho}) \quad \bar{\rho} \in \mathbb{R}^2 \\ \lim_{\rho \rightarrow \infty} U(\bar{\rho}) &= 0, \end{aligned} \quad (2.10)$$

with $g = g_j$ in \mathcal{B}_j ($j = 1, 2$), and with $g = 0$ on $\mathbb{R}^2 \setminus \{\mathcal{B}_1 \cup \mathcal{B}_2\}$, although we shall be interested in values of the solution only on $\mathcal{B}_1 \cup \mathcal{B}_2$. (Notice that g is unbounded on the boundary of $\mathcal{B}_1 \cup \mathcal{B}_2$.) Evidently, the solution to this problem is given by

$$\begin{aligned} U(\bar{\rho}) &= \int \int_{B_1} \frac{1}{2\pi} \log \left\{ \frac{1}{|\bar{\rho} - \bar{\rho}'|} \right\} g_1(\bar{\rho}') d\bar{\rho}' \\ &\quad + \int \int_{B_2} \frac{1}{2\pi} \log \left\{ \frac{1}{|\bar{\rho} - \bar{\rho}'|} \right\} g_2(\bar{\rho}') d\bar{\rho}', \end{aligned} \tag{2.11}$$

with $\bar{\rho} \in \mathcal{B}$.

To solve this problem, we split each integral over \mathcal{B}_j into four indefinite convolution integrals, i.e.,

$$\int \int_{\mathcal{B}_j} = \int_{a_{j,1}}^{b_{j,1}(x)} \int_{a_{j,2}(x)}^{b_{j,2}(x)} G(\bar{\rho}, \bar{\rho}') g_j(\bar{\rho}') dy' dx' = \sum_{i=1}^4 Q_j^{(i)},$$

with

$$Q_j^{(1)}(\bar{\rho}) = \int_{a_{j,1}}^x \int_{a_{j,2}(x)}^y \cdots dy' dx'$$

$$Q_j^{(2)}(\bar{\rho}) = \int_{a_{j,1}}^x \int_y^{b_{j,2}(x)} \cdots dy' dx'$$

$$Q_j^{(3)}(\bar{\rho}) = \int_x^{b_{j,1}} \int_{a_{j,2}(x)}^y \cdots dy' dx'$$

$$Q_j^{(4)}(\bar{\rho}) = \int_x^{b_{j,1}} \int_y^{b_{j,2}(x)} \cdots dy' dx'.$$

Inspection of the functions g_j shows that:

- (a) $Q_1^{(i)}(x, y) \in \mathbf{Lip}_\alpha$ with respect to x , with $\alpha = \alpha_x^{(1)} = 6/7$ near $x = -3/2$, and $\alpha = \beta_x^{(1)} = 1/7$ near $x = 0$;
- (b) $Q_1^{(i)}(x, y) \in \mathbf{Lip}_\alpha$ with respect to y , with $\alpha = \alpha_y^{(1)} = 3/4$ near $y = 0$ and with $\alpha = \beta_y^{(1)} = 1/4$ near $y = x^2 + \sqrt{3}/2$;
- (c) $Q_2^{(i)}(x, y) \in \mathbf{Lip}_\alpha$ with respect to x , with $\alpha = \alpha_x^{(2)} = 1/\sqrt{3}$ near $x = 0$ and with $\alpha = \beta_y^{(2)} = 1$ near $x = 1$; and
- (d) $Q_2^{(i)}(x, y) \in \mathbf{Lip}_\alpha$ with respect to y , with $\alpha = \alpha_y^{(2)} = 2/3$ near $y = 0$ and with $\alpha = \beta_y^{(2)} = 1/3$ near $y = \sqrt{1 - (x - 1/2)^2}$.

Let us (at this point, somewhat arbitrarily) select

$$h = \frac{1}{\sqrt{N_1}}. \tag{2.12}$$

Given some $\varepsilon > 0$, we select an integer N_1 so that

$$\exp\left(-\beta_x^{(1)} N_1 h\right) = \exp\left(-\beta_x^{(1)} N_1^{1/2}\right) = \varepsilon.$$

We can then expect to achieve the same accuracy in all the variables by fixing M_j by means of the equations (see e.g., [15] §3.1)

$$\begin{aligned} \beta_x^{(1)} N_1^{(1)} &= \alpha_x^{(1)} M_1^{(1)} = \alpha_y^{(1)} M_2^{(1)} = \beta_y^{(1)} N_2^{(1)} \\ &= \beta_x^{(2)} N_1^{(2)} = \alpha_x^{(2)} M_1^{(2)} = \alpha_y^{(2)} M_2^{(2)} = \beta_y^{(2)} N_2^{(2)}. \end{aligned}$$

We then need the matrices

$$A_j^{(i)} = h I_{m_j^{(i)}}^{(-1)} D_{m_j^{(i)}}$$

$$B_j^{(i)} = h \left(I_{m_j^{(i)}}^{(-1)} \right)^T D_{m_j^{(i)}}$$

with $i, j = 1, 2$, with h defined as above, with $m_j^{(i)} = M_j^{(i)} + N_j^{(i)} + 1$, and with

$$D_{m_j^{(i)}} = D \left(\frac{e^w}{1 + e^w} \right), \quad w = k h_j, \quad k = -M_j^{(i)}, \dots, N_j^{(i)}.$$

We next approximate each of the integrals $Q_j^{(i)}$ via the above described *Sinc convolution algorithm*. To this end, we first opt to simplify the somewhat cumbersome notation that we have adopted above. In order to approximate $Q_1^{(1)}$, let us first set

$$p(x, y) = Q_1^{(1)}(x, y) = \int_{-3/2}^x \int_0^y \frac{1}{2\pi} \ln\left(\frac{1}{|\bar{\rho} - \bar{\rho}'|}\right) g_1(\bar{\rho}') d\bar{\rho}'.$$

We have to diagonalize the matrices $A_1^{(1)}$ and $B_1^{(1)}$, i.e., we set

$$A_1^{(1)} = X S_1 X^{-1}; \quad X^{-1} = [x^{ij}]$$

$$\begin{aligned} B_1^{(1)} &= Y S_2 Y^{-1}; \\ S_j &= \text{diag} \left[s_{M_j}^{(j)}, \dots, s_{M_j}^{(j)} \right]. \end{aligned}$$

With reference to *Algorithm 3.2*, we have

$$a_1^{(1)} = -\frac{3}{2}, \quad b_1^{(1)} = 0,$$

$$a_2^{(1)}(x) = 0, \quad b_2^{(1)}(x) = x^2 + \frac{\sqrt{3}}{2}.$$

The *Sinc points* which we shall require for \mathcal{B}_1 are

$$x_i^{(1)} = \frac{a_1^{(1)}}{1 + e^{ih}}, \quad y_{i,j}^{(1)} = \frac{b_2(x_i^{(1)}) e^{jh}}{1 + e^{jh}}$$

The algorithm for approximating the first integral on the right hand side of (2.11), with $(x, y) \in \mathcal{B}_1$ then is the following:

- (a) Set up $[g_{i,j}] = [g_1(x_i^{(1)}, y_{i,j}^{(1)})]$;
- (b) Form $h_{i,\cdot} = Y^{-1} g_{i,\cdot}$;
- (c) Use the “Laplace transform” $\hat{G}(u, v)$ given in (4.10);
- (d) Form

$r_{i,j}$

$$= \sum_{k=-M_1^{(1)}}^{N_1^{(1)}} x^{ik} \hat{G} \left([b_1 - a_1] s_i^{(1)}, [b_2(x_k^{(1)}) - b_2(x_i^{(1)})] s_j^{(2)} \right) h_{k,j};$$

- (e) Form

$$q_{\cdot,j} = X r_{\cdot,j}; \quad p_{i,\cdot}^{(1)} = Y q_{i,\cdot}.$$

At this point we need to:

- (a) Repeat the above steps to evaluate the 3 other indefinite convolutions over \mathcal{B}_1 , to get the total *convolution contribution* $p_{i,j}$ to U from \mathcal{B}_1 ;
- (b) Repeat the above steps for the second integral in (2.11);
- (c) We then need to do a *Sinc quadrature* over \mathcal{B}_2 , to determine the contribution $P_{ij}^{(1)}$ of the integral over \mathcal{B}_2 to the *Sinc points* in \mathcal{B}_1 , and similarly, we also need to do a *Sinc quadrature* over \mathcal{B}_1 to determine the contribution of this convolution integral to the *Sinc points* in \mathcal{B}_2 . These *Sinc quadratures* are possible since the Green’s function $G(\bar{\rho}, \bar{\rho}')$ does not have any singularity on the region of integration. The contribution P_{ij} is then determined as follows:

$$\begin{aligned}
P_{i,j} &= \int \int_{B_2} G(\bar{\rho}_{i,j} - \bar{\rho}') g_2(\bar{\rho}') d\bar{\rho}' \\
&= \int_0^{3/2} \int_0^{\sqrt{1-(x^{(2)}-1/2)^2}} G\left(x_i^{(1)}, y_{i,j}^{(1)}; x^{(2)}, y^{(2)}\right) \cdot \\
&\quad \cdot g\left(x^{(2)}, y^{(2)}\right) dy^{(2)} dx^{(2)} \\
&= \int_0^{3/2} \int_0^1 G\left(x_i^{(1)}, y_{i,j}^{(1)}; x^{(2)}, y^{(2)} \sqrt{1 - (x^{(2)} - 1/2)^2}\right) \cdot \\
&\quad \cdot g\left(x^{(2)}, y^{(2)} \sqrt{1 - (x^{(2)} - 1/2)^2}\right) \cdot \\
&\quad \cdot \sqrt{1 - (x^{(2)} - 1/2)^2} dy^{(2)} dx^{(2)} \\
&\approx h^2 \sum_{k=-M_1^{(2)}}^{N_1^{(2)}} \sum_{\ell=-M_2^{(2)}}^{N_2^{(2)}} \frac{e^{(k+\ell)h} g_{k,\ell}}{(1 + e^{kh})^2 (1 + e^{kh})^2} \cdot \\
&\quad \cdot G\left(x_i^{(1)}, y_{i,j}^{(1)}; x_k^{(2)}, y_\ell^{(2)} \sqrt{1 - (x_k^{(2)} - 1/2)^2}\right) \cdot \\
&\quad \cdot g\left(x_k^{(2)}, y_\ell^{(2)} \sqrt{1 - (x_k^{(2)} - 1/2)^2}\right) \cdot \\
&\quad \cdot \sqrt{1 - (x_k^{(2)} - 1/2)^2}
\end{aligned}$$

- (d) This sum has to be done for all integers $(i, j) \in [-M_1^{(1)}, N_1^{(1)}] \times [-M_2^{(1)}, N_2^{(1)}]$, and each of these contributions $P_{i,j}$ then needs to be added to $p_{i,j}$. Then repeat, for approximating the integral over B_2 .

2.2. Sinc Convolution Solution of a Heat Problem. Next, we consider the heat equation,

$$\frac{\partial u(\bar{r}, t)}{\partial t} - \mu \Delta u(\bar{r}, t) = f(\bar{r}, t),$$

with $\mu = 1$, $\bar{r} = (x, y, z) \in \mathbb{R}^3$ and $r = |\bar{r}|$. For the numerical tests we chose as right sides the functions

TABLE 2. Results for example 1. *** indicates that FEM was not able to achieve this accuracy.

FEM			SINC		
accuracy ($\ \cdot\ _\infty$)	CPU time (sec)	# of time steps	accuracy ($\ \cdot\ _\infty$)	CPU time (sec)	N
0.0619	10.28	10	0.0621	101	8
0.0293	121.2	10	0.0297	238	10
0.0123	3517	100	0.0132	680	12
0.0055	12994	100	0.0068	1614	14
***	***	***	0.0041	2306	15

$$f_1(x, y, z, t) = e^{-\bar{r}^2 - 0.5 \cdot t} (1 + 5.5 \cdot t - 4 \cdot t \cdot r^2)$$

(Table 2) and

$$f_2(x, y, z, t) = e^{-\bar{r}^2 - 0.5 \cdot t} \left(\frac{1}{2\sqrt{t}} + 5.5\sqrt{t} - 4\sqrt{t} \cdot r^2 \right)$$

(Table 3); the corresponding solutions are

$$u_1(x, y, z, t) = t \cdot e^{-r^2 - 0.5 \cdot t}$$

and

$$u_2(x, y, z, t) = \sqrt{t} \cdot e^{-r^2 - 0.5 \cdot t}.$$

The FEM solution was computed on a cubic area with center the origin, side length 12 and zero boundary conditions, the time interval was chosen as $t = [0, 1]$. In the time variable t , a constant step size was used.

For both problems, the accuracy of the computations was compared on the mesh generated by the FEM method. The total number of *Sinc points* used in each direction as well as in the time t was $2N + 1$. In order to get a higher accuracy for the FEM method, we had to choose smaller time steps in the second half of the computations. As indicated in the table, we failed in our attempt to get higher accuracy for FEM due to memory problems. On the other hand, we were able to achieve a higher accuracy via our *Sinc* procedure. We may note that FEM failed for example 2 even earlier than for example 1 due to a singularity of f_2 at $t = 0$, which resulted in our requirement of a finer mesh for FEM.

TABLE 3. Results for example 2. *** indicates that FEM was not able to achieve this accuracy.

FEM			SINC		
accuracy ($\ \cdot\ _\infty$)	CPU time (sec)	# of time steps	accuracy ($\ \cdot\ _\infty$)	CPU (sec)	N
0.1139	14	10	0.1458	80	6
0.0435	125	10	0.0394	438	9
0.0198	5309	200	0.0186	1228	11
***	***	***	0.0012	8238	20
***	***	***	0.0006	13592 sec	30

2.3. A Wave Equation Problem. For a numerical example for solving the 2d wave equation,

$$\frac{1}{c^2} \frac{\partial^2 u(\bar{r}, t)}{\partial t^2} - \nabla^2 u(\bar{r}, t) = f(\bar{r}, t),$$

we took as right hand side the function

$$f(x, y, t) = e^{(-|r|^2 - 0.5 \cdot t)} \left(\frac{3}{4 \cdot \sqrt{t}} - \frac{3}{2} \sqrt{t} + \sqrt{t}^3 \cdot \left(\frac{17}{4} - 4 \cdot |r|^2 \right) \right).$$

The corresponding solution is

$$u(x, y, t) = t^{3/2} \cdot e^{(-|r|^2 - 0.5 \cdot t)}.$$

As usual, $r = \sqrt{x^2 + y^2}$, and c was set to 1. The results of both methods were compared with the exact solution (with maximum-norm) only in the knots of the *FEM* mesh. For the previous two examples, the program KSAKADE was used to produce the *FEM* solution. This time, we used the *Matlab* PDE-Toolbox. *Matlab* does not have an adaptive refinement of the mesh. To get different degrees of accuracy, the mesh for *FEM* was refined by hand; for *SINC* a larger number of *Sinc points* was used (number of *Sinc points* is $2N + 1$). The time was measured by using the *Matlab* command *cputime*.

It is obvious from Table 4 that the computing time for *FEM* increases rapidly with a finer mesh without a substantial improvement of the accuracy. Again, trying to use a finer mesh caused some memory problems for *FEM*. This is not the problem with *SINC*, a substantial improvement of the accuracy is still possible as is shown in the table.

TABLE 4. Wave equation Results. *** indicates that FEM was not able to achieve this accuracy.

FEM				SINC		
accuracy ($\ \cdot\ _\infty$)	CPU (sec)	time steps	knots	accuracy ($\ \cdot\ _\infty$)	CPU (sec)	N
0.0614	7.52	10	181	0.066	1.3	8
0.0073	25.32	10	681	0.008	9.8	15
0.0021	100	20	2641	0.0023	25.83	20
0.0016	556.19	31	10401	0.0014	37.44	22
0.0015	3964	51	41218	0.0014	37.44	22
***	***	***	***	2e-06	717	50
***	***	***	***	2e-07	1421	60

2.4. Solving Burgers' Equation. This section summarizes a presentation given in [1]. An excellent *Sinc* solution of the n -dimensional generalization of this equation, i.e., of conservation law problems was given in [19], but before the development of *Sinc convolution*.

We shall illustrate an integral equation procedure for solving the Burgers' equation problem

$$\begin{aligned} \frac{\partial}{\partial t} u(x, t) - \varepsilon \frac{\partial^2}{\partial x^2} u(x, t) &= -\frac{1}{2} \frac{\partial}{\partial x} u^2(x, t), \quad x \in \mathbb{R}, \quad t > 0, \\ u(x, 0) &= u_0(x). \end{aligned} \tag{2.13}$$

We accomplish this by first transforming the problem (2.13) into the equivalent integral equation problem

$$\begin{aligned} u(x, t) &= \frac{1}{(4\pi\varepsilon t)^{1/2}} \int_{\mathbb{R}} \exp \left\{ -\frac{(x-\xi)^2}{4\varepsilon t} \right\} u_0(\xi) d\xi \\ &\quad + \pi \int_0^t \int_{\mathbb{R}} \frac{x-\xi}{\{4\pi\varepsilon(t-\tau)\}^{3/2}} \exp \left\{ -\frac{(x-\xi)^2}{4\varepsilon(t-\tau)} \right\} u^2(\xi, \tau) d\xi d\tau, \end{aligned} \tag{2.14}$$

which we discretize via the *Sinc collocation* procedure of the previous example, and then we solve the resulting discretized system via Neumann iteration.

We take

$$u_0(x) = a \exp \left\{ -b(x - c)^2 \right\}. \quad (2.15)$$

This choice of u_0 enables an explicit expression for the first term on the right-hand side of (2.16), so that we can now rewrite (2.16) in the form

$$\begin{aligned} u(x, t) &= v(x, t) \\ &+ \pi \int_0^t \left[\int_{-\infty}^x \frac{x - \xi}{\{4\pi\varepsilon(t - \tau)\}^{3/2}} \exp \left\{ -\frac{(x - \xi)^2}{4\varepsilon(t - \tau)} \right\} u^2(\xi, \tau) d\xi \right. \\ &\quad \left. - \int_x^\infty \frac{\xi - x}{\{4\pi\varepsilon(t - \tau)\}^{3/2}} \exp \left\{ -\frac{(x - \xi)^2}{4\varepsilon(t - \tau)} \right\} u^2(\xi, \tau) d\xi \right] d\tau, \end{aligned} \quad (2.16)$$

where

$$v(x, t) = \frac{a}{\{1 + 4b\varepsilon t\}^{1/2}} \exp \left\{ -\frac{b(x - c)^2}{1 + 4b\varepsilon t} \right\}. \quad (2.17)$$

Due to this explicit form of the function $v(x, t)$, the form (2.13) for u_0 makes it possible to approximate an arbitrary continuous function u_0 defined on \mathbb{R} by use of the function $F_3(\beta, h)$ defined in [15], §5.8.

We now proceed to discretize Equation (2.16) as outlined in Example 2.2. To this end we may note that it is possible to explicitly evaluate the “Laplace transform” of the convolution kernel in (2.16), i.e.,

$$\begin{aligned} F(s, \sigma) &= \int_0^\infty \int_0^\infty \exp \left\{ -\frac{x}{s} - \frac{t}{\sigma} \right\} \frac{x}{\{4\pi\varepsilon t\}^{3/2}} \exp \left\{ -\frac{x^2}{4\varepsilon t} \right\} dx dt \\ &= \frac{1}{4\varepsilon^{1/2}} \frac{s \sigma^{1/2}}{s + \varepsilon^{1/2} \sigma^{1/2}}. \end{aligned} \quad (2.18)$$

We now select $\varepsilon = 1/2$, $b = 1$, $c = 0$, $\phi_t(t) = \log(t)$, $\phi_x(x) = x$, $d_t = \pi/2$, $\alpha_t = \beta_t = 1/2$, $d_x = \pi/4$, $\alpha_x = \beta_x = 1$, and in this case it is convenient to take $M_t = N_t = M_x = N_x = N$ and $h = 2/\sqrt{N}$. We thus form matrices

$$\begin{aligned} A_x &= h_x I^{(-1)} = X_x S_x X_x^{-1}, & A'_x &= h_x (I^{(-1)})^T = (X_x^{-1})^T S_x X_x^T, \\ B_t &= h_t I^{(-1)} D(1/\phi'_t) = X_t S_t X_t^{-1}, \end{aligned} \quad (2.19)$$

where the superscript “ T ” denotes the transpose, and where S_x and S_t are diagonal matrices, and then proceed as in Example 2.2 above, and the notation of Equation (4.6) to reduce the integral equation problem (2.16) to the nonlinear matrix problem

$$[u_{ij}] = F(A_x, B_t, [u_{ij}^2]) - F(A'_x, B_t, [u_{ij}^2]) + [v_{ij}]. \quad (2.20)$$

Here, upon listing as a single vector the columns of a rectangular matrix $[c_{i,j}]$, (denote it by $\text{col}\{[c_{i,j}]\}$) then similarly listing the columns of $F(A_x, B_t, [v_{ij}])$, and then forming a diagonal matrix \mathbf{F} by listing the numbers $F(s_i, \sigma_j)$ in the same order, then

$$\text{col}\{F(A_x, B_t, [c_{i,j}])\} = X_t \otimes X_x \mathbf{F} X_t^{-1} \otimes X_x^{-1} \text{col}\{[c_{i,j}]\}, \quad (2.21)$$

where the numbers v_{ij} may be evaluated *a priori*, via the formula $v_{ij} = v(ih_x, z_j)$, with $v(x, t)$ defined as in (2.17), and with $z_j = e^{j h_t}$.

The system (2.20) may be solved by Neumann iteration, for a (defined as in (2.16) sufficiently small. Neumann iteration takes the form

$$[u_{ij}^{(k+1)}] = F(A_x, B_t, [(u_{ij}^{(k)})^2]) - F(A'_x, B_t, [(u_{ij}^{(k)})^2]) + [v_{ij}], \quad (2.22)$$

for $k = 0, 1, 2, \dots$, starting with $[u_{ij}^{(0)}] = [v_{ij}]$. For example, with $a = 1/2$, and using the map $\phi_t(t) = \log[\sinh(t)]$ we achieved convergence in 4 iterations, for all values of N (between 10 and 30) that we attempted. We can also solve the above equation via Neumann iteration for larger values of a , if we restrict the time t to a finite interval, $(0, T)$, via the map $\phi_t(t) = \log\{t/(T-t)\}$.

Let us now also consider the convergence of the iteration procedure (2.22). To this end, let us assume that we have determined the integers $M_x = N_x = M_t = N_t = N$, as well as $h = 2/\sqrt{N}$ and the time interval T to enable achievement of a certain accuracy in the approximate solution to the problem (2.20). We wish to illustrate the existence of $T = T_0$, such that if the parameters are fixed in this manner, and the “time map” is selected by $w = \phi_t(t) = \log(t/(T-t))$, then (2.20) is a contraction map for all $T < T_0$. To this end, we note from (2.19) above, that A_x and A'_x are unchanged, whereas

$$B_t = h T I^{(-1)} D \left(\frac{e^w}{(1+e^w)^2} \right), \quad w = k h_t, \quad k = -N, \dots, N,$$

that is, the eigenvalues of the diagonal matrix S_t in (2.19) are proportional to T , whereas the eigenvector matrix X_t is independent of T . By (2.18) it thus follows that e.g.,

$$\begin{aligned} & \|\text{col}\{F(A_x, B_t, [c_{i,j}])\}\| \\ & \leq \|X_t\| \|X_t^{-1}\| \|X_x\| \|X_x^{-1}\| \|\mathbf{F}\| \|\text{col}\{[c_{i,j}]\}\|, \end{aligned}$$

where, by (2.20), and the above expression for B_t ,

$$\|\mathbf{F}\| = \mathcal{O}(T^{1/2}), \quad T \rightarrow 0.$$

That is, the right hand side of (2.20) is a contraction map for all sufficiently small T .

Similar results obtain for the above cases of the wave and heat equations, as well as for the case of the electric field integral equation which is considered below, in the cases when the Green's function approach is used to reduce a PDE to an equivalent integral equation formulation.

2.5. Solving the Electric Field Integral Equation, [11]. Our final example involves the electric field integral equation, which takes the form

$$\begin{aligned} \mathbf{e}^{in}(\mathbf{r}, t) &= \mathbf{e}(\mathbf{r}, t) \\ &= - \int_V \int_0^t \left(\int_0^{t'} \gamma(\mathbf{r}', t' - \xi) \mathbf{e}(\mathbf{r}', \xi) d\xi \right) g(|\mathbf{r} - \mathbf{r}'|, t - t') dt' d^3 \mathbf{r}', \end{aligned} \quad (2.23)$$

where the time-domain Green's function is given in terms of $r = |\mathbf{r}|$, i.e.,

$$\begin{aligned} g(\mathbf{r}, t) &= \frac{1}{4\pi r} e^{-a\frac{r}{c}} \delta \left(t - \frac{r}{c} \right) \\ &\quad + \frac{a\frac{r}{c} e^{-at}}{4\pi r \left(t^2 - \frac{r^2}{c^2} \right)^{\frac{1}{2}}} I_1 \left[a \left(t^2 - \frac{r^2}{c^2} \right)^{\frac{1}{2}} \right] u \left(t - \frac{r}{c} \right), \end{aligned} \quad (2.24)$$

where u is the *Heaviside function* and I_1 is the *modified Bessel function of the first kind* of order one, with $a = \frac{\sigma_0}{2\epsilon_0}$. The constant $c = 1/\sqrt{\mu_0\epsilon_0}$ is the velocity of the wave in V . Furthermore, $\gamma(\mathbf{r}, t)$ is the time-domain scattering potential, which is given as the product of two functions for $t > 0$, one with space variables and the other with time variable, i.e.

$$\gamma(\mathbf{r}, t) = \gamma_1(\mathbf{r}) \gamma_2(\mathbf{r}, t), \quad (2.25)$$

where

$$\gamma_1(\mathbf{r}) = \mathbf{z}^{\frac{1}{2}} e^{-r^2} + j \mathbf{z}^{\frac{3}{2}} e^{-|\mathbf{r}-(0,0,3)|^2}, \quad (2.26)$$

and where the “Laplace transform” of $\gamma_2(\mathbf{r}, t)$ taken with respect to t is

$$\Gamma_2(\mathbf{r}, \tau) = \int_0^\infty \exp(-t/\tau) \gamma_2(\mathbf{r}, t) dt = \frac{(\sigma(\mathbf{r}) - \sigma_0)\tau + (\epsilon(\mathbf{r}) - \epsilon_0)}{\sigma_0\tau + \epsilon_0}, \quad (2.27)$$

We shall obtain an approximate solution to this integral equation for $\{(\mathbf{r}, t) \in V \times (0, T)\}$, where

TABLE 5. Yee's FD and Sinc Convolution Comparisons

Precision	FD Run-Time	Sinc Run-Time
10^{-1}	< 1 second	< 1 second
10^{-2}	000:00:00:27	000:00:00:06
10^{-3}	003:00:41:40*	000:00:02:26
10^{-4}	> 82 years*	000:00:43:12
10^{-5}	> 800,000 years*	000:06:42:20
10^{-6}	> 8.2 billion years*	001:17:31:11

$$V = \left\{ \mathbf{r} = (x, y, z) \in \mathbb{R}^3 : (x, y) \in \mathbb{R}^2, z > 0 \right\}. \quad (2.28)$$

The *Sinc convolution* method of solution requires the “Laplace transform” $\mathcal{G}(u, v, w, \tau)$ of the kernel of the integral equation (2.23), which is the product not only the “Laplace transform” of $\gamma_2(\mathbf{r}, t)$ with respect to t , and the four dimensional “Laplace transform” of the time domain free space Green’s function $g(x, y, z, t)$ with respect to all variables

$$\mathcal{G}(u, v, w, \tau) = \Gamma_2(\mathbf{r}, \tau) G(u, v, w, \tau). \quad (2.29)$$

These “Laplace transforms” may in fact be explicitly expressed in terms of the results given in §4 of this paper. Moreover, the resulting *Sinc convolution* layout can be solved via Neumann iteration, analogous to that of Burgers’ equation above. Furthermore, this iteration scheme may be shown to converge provided the time interval $(0, T)$ is sufficiently small, via an argument similar to that used at the end of the Burgers equation example. However, we omit the lengthy details, which will be published elsewhere. It is, nevertheless interesting to compare the performance the *Sinc convolution* method with that of Yee’s [20] FD (finite difference) solution method. For these comparisons, see Table 5. In this table, all entries except those with a “*” are actual computation times. The entries marked with a “*” are computed times based on the convergence rates obtained by Monk & Süli in [9]. In the table, Computer run-time is shown as Days: Hours: Minutes: Seconds.

3. Sinc Terminology

This appendix is a summary of the *Sinc* notation which we require for the presentation of the results of the paper. Most of the results are proved elsewhere, i.e., in [8, 15, 16]. The new results, such as the extension of *Sinc convolution* to curvilinear regions are presented with proofs. Our manner of description of the methods is in symbolic form. We include methods for collocation, function interpolation and approximation, for approximate definite and indefinite integration, for the approximation of definite and indefinite convolutions, including multidimensional extensions of these for the approximate solution of partial differential and integral equations.

3.1. One Dimensional Sinc Spaces. Let \mathcal{D} be a simply connected domain in the complex plane \mathbb{C} , let $1 \leq p \leq \infty$, and let $\mathbf{H}^p(\mathcal{D})$ denote the family of all functions f that are analytic in \mathcal{D} , such that

$$N_p(f, \mathcal{D}) \equiv \begin{cases} \left(\int_{\partial\mathcal{D}} |f(z)|^p |dz| \right)^{1/p} < \infty & \text{if } 1 \leq p < \infty, \\ \sup_{z \in \mathcal{D}} |f(z)| < \infty & \text{if } p = \infty. \end{cases} \quad (3.1)$$

In essence, we consider two spaces of functions $\mathbf{M}_{\alpha,\beta}(\Gamma)$ and $\mathbf{L}_{\alpha,\beta}(\Gamma)$ for purposes of *Sinc approximation* on an interval or contour. Consider first the case of a finite interval, (a, b) . Perhaps the simplest concept of the space of functions $\mathbf{M}_{\alpha,\beta}(a, b)$, with $0 < \alpha \leq 1$, $0 < \beta \leq 1$, is that consisting of all functions that are analytic on the open interval (a, b) , of class \mathbf{Lip}_α in a neighborhood of a , and of class \mathbf{Lip}_β in a neighborhood of b . The corresponding space $\mathbf{L}_{\alpha,\beta}(a, b)$ consists of the set of all functions $f \in \mathbf{M}_{\alpha,\beta}(a, b)$ for which $f(a) = f(b) = 0$.

More generally, if (a, b) is a contour Γ , such as, e.g., the interval $(0, \infty)$, or the real line \mathbb{R} , (or even an analytic arc in the complex plane), the mapping ϕ is selected to be a conformal mapping of a domain \mathcal{D} onto \mathcal{D}_d , with \mathcal{D}_d defined as above, such that ϕ is also a one-to-one map of Γ onto \mathbb{R} . We define ρ by $\rho = e^\phi$. Note that $\rho(z)$ increases from 0 to ∞ as z traverses Γ from a to b .

Let α , β and d denote arbitrary, fixed positive numbers. We denote by $\mathbf{L}_{\alpha,\beta}(\Gamma)$ the family of all functions that are analytic and uniformly bounded in \mathcal{D} , such that

$$f(z) = \begin{cases} \mathcal{O}(|\rho(z)|^\alpha), & \text{uniformly as } z \rightarrow a \text{ from within } \overline{\mathcal{D}}, \\ \mathcal{O}(|\rho(z)|^{-\beta}), & \text{uniformly as } z \rightarrow b \text{ from within } \overline{\mathcal{D}}. \end{cases} \quad (3.2)$$

We next define the class of functions $\mathbf{M}_{\alpha,\beta}(\Gamma)$, but this time restricting α , β and d such that $\alpha \in (0, 1]$, $\beta \in (0, 1]$ and $d \in (0, \pi)$. This class consists

of all those functions $g \in \mathbf{Hol}(\mathcal{D})$, that have finite limits at a and b , so that the function $\mathcal{L}f$ is well defined, where

$$\mathcal{L}g(z) = \frac{f(a) + \rho(z) f(b)}{1 + \rho(z)}, \quad \rho = e^\phi, \quad (3.3)$$

and such that if f is defined by

$$f = g - \mathcal{L}g \quad (3.4)$$

then $f \in \mathbf{L}_{\alpha,\beta}(\Gamma)$.

Note that if $0 < d < \pi$, then $\mathcal{L}(g)$ is uniformly bounded in $\overline{\mathcal{D}}$, the closure of \mathcal{D} , and moreover, $\mathcal{L}(g)(z) - f(a) = \mathcal{O}(|\rho(z)|)$ as $z \rightarrow a$, and $\mathcal{L}(g)(z) - f(b) = \mathcal{O}(1/|\rho(z)|)$ as $z \rightarrow b$, i.e., $\mathcal{L}(g) \in \mathbf{M}_{1,1}(\Gamma)$. Furthermore, $\mathbf{M}_{1,1}(\Gamma) \subseteq \mathbf{M}_{\alpha,\beta}(\Gamma)$ for any $\alpha \in (0, 1]$, $\beta \in (0, 1]$, and $d \in (0, \pi)$, and moreover for these restrictions on α , β , and d , the class $\mathbf{L}_{\alpha,\beta}(\Gamma)$ is contained in the class $\mathbf{M}_{\alpha,\beta}(\Gamma)$.

The spaces $\mathbf{L}_{\alpha,\beta}(\Gamma)$ and $\mathbf{M}_{\alpha,\beta}(\Gamma)$ are motivated by the premise that most scientists and engineers use calculus to model differential and integral equation problems, and under this premise the solution to these problems are (at least piece-wise) analytic. The spaces $\mathbf{L}_{\alpha,\beta}(\Gamma)$ and $\mathbf{M}_{\alpha,\beta}(\Gamma)$ house nearly all solutions to such problems, including solutions with singularities at end points of (finite or infinite) intervals (or at boundaries of finite or infinite domains in more than one dimension). Although these spaces also house singularities, they are not as large as Sobolev spaces which assume the existence of only a finite number of derivatives in a solution, and consequently (see below) when *Sinc* methods are used to approximate solutions of differential or integral equations, they are usually more efficient than finite difference or finite element methods. In addition, *Sinc* methods are replete with interconnecting simple identities, including DFT (which is one of the *Sinc* methods, enabling the use of FFT), making it possible to use a *Sinc approximation* for nearly every type of operation arising in the solution of differential and integral equations.

Let us describe some specific spaces for one dimensional *Sinc approximation*.

Example 3.1: If $\Gamma = (0, 1)$, and if \mathcal{D} is the “eye-shaped” region, $\mathcal{D} = \{z \in \mathbb{C} : |\arg[z/(1-z)]| < d\}$, then $\phi(z) = \log[z/(1-z)]$, the relation (3.3) reduces to $f = g - (1-x)g(0) - xg(1)$, and $\mathbf{L}_{\alpha,\beta}(\Gamma)$ is the class of all functions $f \in \mathbf{Hol}(\mathcal{D})$, such that for all $z \in \mathcal{D}$, $|f(z)| < c|z|^\alpha|1-z|^\beta$. In this case, if e.g. $\delta = \max\{\alpha, \beta\}$, and a function w is such that $w \in \mathbf{Hol}(\mathcal{D})$, and $w \in \mathbf{Lip}_\delta(\mathcal{D})$, then $w \in \mathbf{M}_{\alpha,\beta}(\Gamma)$. The *Sinc points* z_j are $z_j = e^{jh}/(1 + e^{jh})$, and $1/\phi'(z_j) = e^{jh}/(1 + e^{jh})^2$.

Example 3.2: If $\Gamma = (0, \infty)$, and if \mathcal{D} is the “sector” $\mathcal{D} = \{z \in \mathbb{C} : |\arg(z)| < d\}$, then $\phi(z) = \log(z)$, the relation (3.3) reduces to $f(z) = g(z) - [g(0) + z g(\infty)]/(1+z)$, and the class $\mathbf{L}_{\alpha,\beta}(\Gamma)$ is the class of all functions $f \in \mathbf{Hol}(\mathcal{D})$ such that if $z \in \mathcal{D}$ and $|z| \leq 1$ then $|f(z)| \leq c|z|^\alpha$, while if $z \in \mathcal{D}$ and $|z| \geq 1$, then $|f(z)| \leq c|z|^{-\beta}$. This map thus allows for algebraic decay at both $x = 0$ and $x = \infty$. The *Sinc points* z_j are defined by $z_j = e^{jh}$, and $1/\phi'(z_j) = e^{jh}$.

Example 3.3: If $\Gamma = (0, \infty)$, and if \mathcal{D} is the “bullet-shaped” region $\mathcal{D} = \{z \in \mathbb{C} : |\arg(\sinh(z))| < d\}$, then $\phi(z) = \log(\sinh(z))$. The relation (3.3) then reduces to $f(z) = g(z) - [g(0) + \sinh(z) g(\infty)]/(1+\sinh(z))$, and $\mathbf{L}_{\alpha,\beta}(\Gamma)$ is the class of all functions $f \in \mathbf{Hol}(\mathcal{D})$ such that if $z \in \mathcal{D}$ and $|z| \leq 1$ then $|f(z)| \leq c|z|^\alpha$, while if $z \in \mathcal{D}$ and $|z| \geq 1$, then $|f(z)| \leq c \exp\{-\beta|z|\}$. This map thus allows for algebraic decay at $x = 0$ and exponential decay at $x = \infty$. The *Sinc points* z_j are defined by $z_j = \log[e^{jh} + (1 + e^{2jh})^{1/2}]$, and $1/\phi'(z_j) = (1 + e^{-2jh})^{-1/2}$.

Example 3.4: If $\Gamma = \mathbb{R}$, and if \mathcal{D} is the above defined “strip”, $\mathcal{D} = \mathcal{D}_d$, take $\phi(z) = z$. The relation (3.3) then reduces to $f(z) = g(z) - [g(-\infty) + e^z g(\infty)]/(1+e^z)$. The class $\mathbf{L}_{\alpha,\beta}(\mathcal{D})$ is the class of all functions $f \in \mathbf{Hol}(\mathcal{D})$ such that if $z \in \mathcal{D}$ and $\Re z \leq 0$ then $|f(z)| \leq ce^{-\alpha|z|}$, while if $z \in \mathcal{D}$ and $\Re z \geq 0$, then $|f(z)| \leq ce^{-\beta|z|}$. Thus this map allows for exponential decay at both $x = -\infty$ and $x = \infty$. The *Sinc points* z_j are defined by $z_j = jh$, and $1/\phi'(z_j) = 1$.

Example 3.5: If $\Gamma = \mathbb{R}$, and if \mathcal{D} is the “hour glass-shaped” region, $\mathcal{D} = \{z \in \mathbb{C} : |\arg[z + (1 + z^2)^{1/2}]| < d\}$, take $\phi(z) = \log[z + (1 + z^2)^{1/2}]$. The relation (3.3) reduces to $f(z) = g(z) - [g(-\infty) + (z + (1 + z^2)^{1/2}) g(\infty)]/[1 + z + (1 + z^2)^{1/2}]$, and the class $\mathbf{L}_{\alpha,\beta}(\Gamma)$ is the class of all functions $f \in \mathbf{Hol}(\mathcal{D})$ such that if $z \in \mathcal{D}$ and $\Re z \leq 0$, then $|f(z)| \leq c(1 + |z|)^{-\alpha}$, while if $z \in \mathcal{D}$ and $\Re z \geq 0$, then $|f(z)| \leq c(1 + |z|)^{-\beta}$. This map thus allows for algebraic decay at both $x = -\infty$ and $x = \infty$. The *Sinc points* z_j are defined by $z_j = \sinh(jh)$, and $1/\phi'(z_j) = \cosh(jh)$.

Example 3.6: If $\Gamma = \mathbb{R}$, and if \mathcal{D} is the “funnel-shaped” region, $\mathcal{D} = \{z \in \mathbb{C} : |\arg\{\sinh[z + (1 + z^2)^{1/2}]\}| < d\}$, take $\phi(z) = \log\{\sinh[z + (1 + z^2)^{1/2}]\}$. The relation (3.3) then reduces to

$f(z) = g(z) - [g(-\infty) + \sinh(z + (1 + z^2)^{1/2}) g(\infty)]/[1 + \sinh(z + (1 + z^2)^{1/2})]$, and $\mathbf{L}_{\alpha,\beta}(\Gamma)$ is the class of all functions $f \in \mathbf{Hol}(\mathcal{D})$ such that if $z \in \mathcal{D}$ and $\Re z \leq 0$, then $|f(z)| \leq c(1 + |z|)^{-\alpha}$, while if $z \in \mathcal{D}$ and $\Re z \geq 0$, then $|f(z)| \leq ce^{-\beta|z|}$. This map thus allows for algebraic decay at $x = -\infty$ and exponential decay at $x = \infty$. The *Sinc points* z_j are defined by $z_j = (1/2)[t_j - 1/t_j]$, where $t_j = \log[e^{jh} + (1 + e^{2jh})^{1/2}]$, and $1/\phi'(z_j) = (1/2)(1 + 1/t_j^2)(1 + e^{-2jh})^{-1/2}$.

3.2. Multidimensional Sinc Spaces. These spaces are defined precisely in [15] §6.53, §6.6.2 & §7.3.2. In essence, a function f defined on e.g., a bounded region $V \in \mathbb{R}^d$ belongs to a proper *Sinc space* if given a point $\mathbf{x} = (x_1, \dots, x_d)$ in the closure of V , then corresponding to each $j = 1, \dots, d$, let us fix all of the coordinates x_i , $i \neq j$, and call the resulting function $f_j(x_j)$. Denoting by $\Gamma_j = V$ the interval of longest length in the x_j direction that contains the point \mathbf{x} , we want f_j to be analytic in the interior of Γ_j , and to be of class Lip_α on the closure of Γ_j . If all of these conditions are satisfied, then by taking $h = c_1/\sqrt{N}$, and performing *Sinc approximation* with N point evaluations in each dimension, i.e., for a total of N^d points, we are able to achieve an error of the order of $\exp(-c'\sqrt{N})$.

3.3. One Variable Sinc Approximation.

(1) Notation.

Sinc approximation in $\mathbf{M}_{\alpha,\beta}(\Gamma)$ is defined as follows. Let N denote a positive integer, and let integers M , and m , a diagonal matrix $D(u)$ and an operator V_m be defined as follows

$$\begin{aligned} N &= \text{positive integer} \\ M &= [\beta N/\alpha] \\ m &= M + N + 1 \\ D(u) &= \text{diag}[u(z_{-M}), \dots, u(z_N)] \\ V_m(u) &= (u(z_{-M}), \dots, u(z_N))^T, \end{aligned} \tag{3.5}$$

where $[\cdot]$ denotes the greatest integer function, where u is an arbitrary function defined on Γ , and where “ T ” denotes the transpose.

We shall also define a norm by

$$\|f\| = \sup_{x \in \Gamma} |f(x)|,$$

and throughout this section C will denote a generic constant, independent of N .

(2) Sinc Basis.

Letting \mathbb{Z} denote the set of all integers, set

$$\begin{aligned}
\text{sinc}(z) &= \frac{\sin(\pi z)}{\pi z}, \\
h &= \left(\frac{\pi d}{\beta N} \right)^{1/2}, \\
z_j &= \phi^{-1}(jh), \quad j \in \mathbb{Z} \\
\gamma_j &= \text{sinc}\{[\phi - jh]/h\}, \quad j = -M, \dots, N, \\
w_j &= \gamma_j, \quad j = -M + 1, \dots, N - 1, \\
w_{-M} &= \frac{1}{1 + \rho} - \sum_{j=-M+1}^N \frac{1}{1 + e^{jh}} \gamma_j, \\
w_N &= \frac{\rho}{1 + \rho} - \sum_{j=-M}^{N-1} \frac{e^{jh}}{1 + e^{jh}} \gamma_j, \\
\varepsilon_N &= N^{1/2} e^{-(\pi d \beta N)^{1/2}}.
\end{aligned} \tag{3.6}$$

We may thus define a row vector \mathbf{w} of basis functions by

$$\mathbf{w} = (w_{-M}, \dots, w_N)$$

with w_j defined as in (3.6) and for given vector $\mathbf{c} = (c_{-M}, \dots, c_N)^T$, we have

$$\mathbf{w}_m \mathbf{c} = \sum_{j=-M}^N c_j w_j. \tag{3.7}$$

(3) Sinc Interpolation and Approximation.

A proof of the following result may be found in [2, 3] (see e.g., [2, pp. 126–132]).

THEOREM 3.1. *If $f \in \mathbf{M}_{\alpha,\beta}(\Gamma)$, then*

$$\|f - \mathbf{w}_m V_m f\| \leq C \varepsilon_N. \tag{3.8}$$

The constants in the exponent in the definition of ε_N are the best constants for approximation in $\mathbf{M}_{\alpha,\beta}(\Gamma)$. Hence accurate *Sinc approximation* of f is based on our being able to make good estimates on α , β , and d . If these constants cannot be accurately estimated, e.g., if instead of as in (3.6) above, we define h by $h = \gamma/N^{1/2}$, with γ a constant independent of N , then the right-hand side of (3.8) is replaced by $C e^{-\delta N^{1/2}}$, where C and δ are some positive constants independent of N . Henceforth we shall take h as defined in (3.6).

Remark: We remark, that if $f \in \mathbf{L}_{\alpha,\beta}(\mathcal{D})$, then it is convenient to take $w_j = \text{sinc}\{[\phi - jh]/h\}$, $j = -M, \dots, N$, instead of as defined

in (3.6), since the corresponding approximation of f as defined in (3.2) then also vanishes at the end points of Γ , just as f then vanishes at the end points of Γ .

Remark: The above basis $\{w_j\}_{-M}^N$ interpolates at function values in the interior of the interval or contour, Γ , allowing for interpolation of functions that may “blow up” at the end-points of Γ , in which case, such an interpolation is accurate to within a relative error. Also, e.g., if $\beta = \alpha$, and $M = N$, then for $f \in \mathbf{M}_{\alpha,d}(\varphi)$, the differences $f(a) - f(z_{-N})$ and $f(b) - f(z_N)$ are of the order of $e^{-\alpha N h}$, and being negligible, allows us to substitute $f(z_{-N}) = f(a)$, and $f(z_N) = f(b)$. This feature is a very convenient; it eliminates having to evaluate functions whose boundary values are defined only in terms of limits, as is often the case in applications. In addition the computer evaluation of the *Sinc* function $S(j, h) \circ \varphi(x)$ poses problems at the end-points a and b of Γ , since φ is not finite at these points.

(4) *Sinc Collocation.*

The following result, guarantees an accurate final approximation of f on Γ , provided that we know a good approximation to f at the *Sinc points* (for a proof, see [8], p. 132).

THEOREM 3.2. *Let $f \in \mathbf{M}_{\alpha,\beta}(\Gamma)$, and let the conditions of Theorem 3.1 be satisfied. Let $\mathbf{c} = (c_{-M}, \dots, c_N)^T$ be a complex vector of order m , such that*

$$\left(\sum_{j=-M}^N |f(z_j) - c_j|^2 \right)^{1/2} < \delta, \quad (3.9)$$

where δ is a positive number. If C and ε_N are defined as in (3.10), and if w_j is defined as in (3.6), then

$$\|f - \mathbf{w}_m \mathbf{c}\| < C\varepsilon_N + \delta. \quad (3.10)$$

(5) *Sinc Quadrature.*

We also record the standard *Sinc quadrature* formula, which belongs to the family of tools for solving differential and integral equations (see [15], §4.2).

THEOREM 3.3. *If $f/\phi' \in \mathbf{L}_{\alpha,\beta}(\Gamma)$, then*

$$\left| \int_a^b f(x) dx - h \{V_m(1/\phi')\}^T (V_m f) \right| \leq C\varepsilon_N. \quad (3.11)$$

(6) *Sinc Indefinite Integration.*

A detailed derivation and proof of *Sinc indefinite* integration is given in [15], §3.6 and 4.5.

Let us next describe *Sinc indefinite integration* or *Sinc convolution* over an interval or a contour. At the outset, we define numbers σ_k and e_k , by

$$\begin{aligned}\sigma_k &= \int_0^k \operatorname{sinc}(x) dx, \quad k \in \mathbf{Z}, \\ e_k &= 1/2 + \sigma_k.\end{aligned}\tag{3.12}$$

We use the notation of (3.6), and we define a Toeplitz matrix $I^{(-1)}$ of order m by $I^{(-1)} = [e_{i-j}]$, with e_{i-j} denoting the (i, j) th element of $I^{(-1)}$. We then define operators \mathcal{J} and \mathcal{J}' , and matrices A_m and B_m by

$$\begin{aligned}(\mathcal{J}f)(x) &= \int_a^x f(t) dt, & (\mathcal{J}^*f)(x) &= \int_x^b f(t) dt, \\ A_m &= h I^{(-1)} D(1/\phi'), & B_m &= h (I^{(-1)})^T D(1/\phi'), \\ \mathcal{J}_m &= \mathbf{w}_m A_m V_m, & \mathcal{J}_m^* &= \mathbf{w}_m B_m V_m,\end{aligned}\tag{3.13}$$

with $(I^{(-1)})^T$ denoting the transpose of $I^{(-1)}$. We thus obtain the following theorem [8],

THEOREM 3.4. *If $f/\phi' \in \mathbf{L}_{\alpha,\beta}(\Gamma)$, then*

$$\begin{aligned}\|\mathcal{J}f - \mathcal{J}_m f\| &\leq C\varepsilon_N, \\ \|\mathcal{J}^*f - \mathcal{J}_m^* f\| &\leq C\varepsilon_N.\end{aligned}\tag{3.14}$$

(7) *Sinc Indefinite Convolution.*

Indefinite convolution integrals can also be effectively collocated via *Sinc methods*. To this end, we begin with the *model integrals*,

$$\begin{aligned}p(x) &= \int_a^x f(x-t) g(t) dt, \\ q(x) &= \int_x^b f(t-x) g(t) dt,\end{aligned}\tag{3.15}$$

where $x \in \Gamma$. In presenting these convolution results, we shall assume that $\Gamma = (a, b) \subseteq \mathbb{R}$, unless otherwise indicated. Note also, that being able to collocate p and q enables us to collocate both definite convolutions

$$\begin{aligned}&\int_a^b f(x-t) g(t) dt \\ &\int_a^b f(|x-t|) g(t) dt.\end{aligned}\tag{3.16}$$

Sinc collocation of p and q is possible under the following

ASSUMPTION 3.5. *We assume that the “Laplace transform”,*

$$F(s) = \int_E f(t) e^{-t/s} dt \quad (3.17)$$

with E any subset of $\mathbb{R} = (-\infty, \infty)$ such that $E \supseteq (0, b-a)$, exists for all $s \in \Omega^+ \equiv \{s \in \mathbb{C} : \Re s > 0\}$.

In this notation, one gets the rather “esoteric results” [16],

$$p = F(\mathcal{J}) g, \quad q = F(\mathcal{J}^*) g. \quad (3.18)$$

*However, the previous theorem suggests that the approximations $\mathcal{J}g \approx \mathcal{J}_m g$ and $\mathcal{J}^*g \approx \mathcal{J}_m^*g$ are accurate, at least for g in certain spaces of functions, and this is indeed the case. In fact, with the above definition of $\mathcal{J}_m = \mathbf{w}_m A_m V_m$, upon diagonalization of A_m in the form $A_m = X_1 \Lambda X_1^{-1}$, with $\Lambda = \text{diag}[\lambda_{-M_1}, \dots, \lambda_{N_1}]$, and the esoteric forms (3.18) become computationally feasible, as follows:*

$$\begin{aligned} F(\mathcal{J}) g &\approx F(\mathcal{J}_m) g \\ &= \mathbf{w}_m F(A_m) V_m g \\ &= \mathbf{w}_m X_1 F(\Lambda) X_1^{-1} V_m g. \end{aligned} \quad (3.19)$$

As to convergence, let $P(r, x)$ be defined by

$$P(r, x) = \int_a^x f(r+x-t) g(t) dt. \quad (3.20)$$

We assume that

- (i) $P(r, \cdot) \in \mathbf{M}_{\alpha, \beta}(\Gamma)$, uniformly for $r \in [0, b-a]$; and that
- (ii) $P(\cdot, x)$ is of bounded variation on $(0, [b-a])$, uniformly for $x \in [a, b]$.

Under these assumptions, we have (see [15], §4.6, or [8] for a proof)

THEOREM 3.6. *If the above assumptions are satisfied, and if A_m and B_m are defined as in (3.13), then [16]*

$$\begin{aligned} \|p - \mathbf{w}_m F(A_m) V_m g\| &\leq C\varepsilon_N, \\ \|q - \mathbf{w}_m F(B_m) V_m g\| &\leq C\varepsilon_N. \end{aligned} \quad (3.21)$$

Remark: We remark here that it may be shown [15] that every eigenvalue of the matrices $I^{(-1)}$ lies in $\overline{\Omega^+}$, where Ω^+ denotes the right half plane, and also, we have shown by direct computation, that every eigenvalue of the matrices $I^{(-1)}$ lies in Ω^+ for $m = 1, 2, \dots, 513$. It thus follows, thus, at least for the case when (a, b) is a subinterval of \mathbb{R} , that the matrices $F(A_m)$ and $F(B_m)$ are well defined, and may be evaluated in the usual way, via diagonalization of A_m and B_m . (We have also tacitly assumed here that A_m

and B_m can be diagonalized, which has so far always been the case for the problems that we have attempted.)

3.4. Sinc Approximation of Multidimensional Convolutions. In this section we illustrate the extension of one dimensional convolution to the approximation of multidimensional convolution integrals. The reader should note that this algorithm actually yields a separation of variables, enabling the approximation of multidimensional convolution integrals via a sequence of one-dimensional matrix multiplications. Thus, the “big matrix” that one requires for the solution of partial differential equations via classical finite difference and finite element techniques need never be stored, so that problems that require matrix sizes of e.g. $10^7 \times 10^7$ can readily be dealt with. At first we summarize the known results over rectangular regions, leaving out some details which may be found in [16], [15], §4.6. We then give a detailed derivation of indefinite convolution over curvilinear regions. The combination of these two algorithms enables us to solve most PDE problems stemming from applications whose solution can be expressed as convolution integrals over curvilinear regions.

(1) *Convolutions over Rectangular Regions*

We briefly illustrate in what follows, an algorithm for evaluating a two dimensional convolution integral based on the *Sinc convolution theorem* Theorem 3.8 above.

We illustrate here, the approximation of a convolution integral of the form

$$p(x, y) = \int_{a_2}^y \int_x^{b_1} f(x - \xi, \eta - y) g(\xi, \eta) d\xi d\eta, \quad (3.22)$$

where we seek to approximate p on $\mathcal{B} = \prod_{i=1}^2 \otimes(a_i, b_i)$, and with $(a_i, b_i) \subseteq \mathbb{R}$. We assume that the mappings $\phi_j : \mathcal{D}'_j \rightarrow \mathcal{D}_{d'}$ have been determined. We furthermore assume that positive integers N_j and M_j as well as positive numbers h_j ($j = 1, 2$) have been selected, we set $m_j = M_j + N_j + 1$, and we define the *Sinc points* by $z_\ell^{(j)} = \phi^{-1}(\ell h_j)$, for $\ell = -M_j, \dots, N_j$; $j = 1, 2$. Next, we determine matrices A_j , X_j , S_j and X_j^{-1} , such that

$$\begin{aligned} A_1 &= h_1 \left(I_{m_1}^{(-1)} \right)^T D(1/\phi'_1) &= X_1 S_1 X_1^{-1}, \\ A_2 &= h_2 I_{m_2}^{(-1)} D(1/\phi'_2) &= X_2 S_2 X_2^{-1}. \end{aligned} \quad (3.23)$$

In (3.23), $I_{m_j}^{(-1)}$ is defined as in (3.13) above, and the S_j are diagonal matrices,

$$S_j = \text{diag}[s_{-M_j}^{(j)}, \dots, s_{N_j}^{(j)}]. \quad (3.24)$$

We require the two dimensional “Laplace transform”

$$F(s^{(1)}, s^{(2)}) = \int_0^\infty \int_0^\infty f(x, y) \exp\left(-\frac{x}{s^{(1)}} - \frac{y}{s^{(2)}}\right) dx dy, \quad (3.25)$$

which we assume to exist for all $s^{(j)} \in \Omega^+$, with Ω^+ denoting the right half of the complex plane. It can then be shown (see [8], or [15], §4.6) that the values $p_{i,j}$ which approximate $p(z_i^{(1)}, z_j^{(2)})$ can be computed via the following succinct algorithm. In this algorithm the we use the notations, e.g., $\mathbf{k}_{i,\cdot} = (k_{i,-M_2}, \dots, k_{i,N_2})^T$ and $\mathbf{h}_{\cdot,j} = (h_{-M_1,j}, \dots, h_{N_1,j})^T$. We again emphasize the obvious ease of adaptation of this algorithm to parallel computation.

Algorithm 3.1

- (a) Form the arrays $z_i^{(j)}$, and $\frac{d}{dx}\phi^{(j)}(x)$ at $x = z_i^{(j)}$ for $j = 1, 2$, and $i = -M_j, \dots, N_j$, and then form the block of numbers $[g_{i,j}] = [g(z_i^{(1)}, z_j^{(2)})]$.
- (b) Determine A_j , S_j , X_j , and X_j^{-1} for $j = 1, 2$, as defined in (3.23).
- (c) Form $\mathbf{h}_{\cdot,j} = X_1^{-1}\mathbf{g}_{\cdot,j}$, $j = -M_2, \dots, N_2$;
- (d) Form $\mathbf{k}_{i,\cdot} = X_2^{-1}\mathbf{h}_{i,\cdot}$, $i = -M_1, \dots, N_1$;
- (e) Form

$$r_{i,j} = F(s_i^{(1)}, s_j^{(2)}) k_{i,j}, \quad i = -M_1, \dots, N_1, \quad j = -M_2, \dots, N_2;$$

- (f) Form $\mathbf{q}_{i,\cdot} = X_2 r_{i,\cdot}$, $i = -M_1, \dots, N_1$;
- (g) Form $\mathbf{p}_{\cdot,j} = X_1 \mathbf{q}_{\cdot,j}$, $j = -M_2, \dots, N_2$.

Remark: It is unnecessary to compute the matrices X_1^{-1} and X_2^{-1} in steps c and d of this algorithm, since the vectors $\mathbf{h}_{\cdot,j}$ and $\mathbf{k}_{i,\cdot}$ can be found via the $L U$ factorization of the matrices X_1 and X_2 .

Thus starting with the rectangular array $[g_{i,j}]$, Algorithm 3.1 transforms this into the rectangular array $[p_{i,j}]$.

Suppose, for example, that we form a vector \mathbf{g} in which the subscripts appear in the order (call it lexicographic) dictated by the order of appearance of the subscripts in the FORTRAN do loop, “DO $j = -M_2, N_2$ ”, followed by “DO $i = -M_1, N_1$ ”. We then also form the diagonal matrix \mathbf{F} in which the entries are the values

$F_{ij} = F(s_i^{(1)}, s_j^{(2)})$, with the function F and the eigenvalues $s_j^{(i)}$ defined as above, and where we also list the values F_{ij} in the same lexicographic order as for g_{ij} . Then, similarly for the array $p_{i,j}$, we can define a vector \mathbf{p} by listing the elements p_{ij} in lexicographic order. It can then be shown that if \mathbf{p}_1 is defined by the matrix (Kronecker) product

$$\begin{aligned}\mathbf{p} &= \mathbf{C}\mathbf{g} \\ \mathbf{C} &= X_2 \otimes X_1 \mathbf{F} \otimes X_2^{-1} \otimes X_1^{-1}\end{aligned}$$

then the corresponding numbers p_{ij} are accurate approximations of the values $p(z_i^{(1)}, z_j^{(2)})$.

We emphasize here, due to the Kronecker product representation of the matrix \mathbf{C} the numerical determination of the vector $\mathbf{p} = \mathbf{C}\mathbf{g}$ can be carried out in parallel, without storage of the huge matrix \mathbf{C} in this equation, which may be an asset, especially for problems in 3 or more dimensions.

Once the numbers $p_{i,j}$ have been computed, we can then use these numbers to approximate p on the region \mathcal{B} via the use of a *Sinc basis*; upon setting $\rho^{(\ell)} = e^{\phi^{(\ell)}}$, we can define the functions

$$\begin{aligned}\gamma_i^{(\ell)} &= \text{sinc}\{[\phi^{(\ell)} - ih]/h\}, \quad \ell = 1, 2; \quad i = -M_\ell, \dots, N_\ell, \\ w_i^{(\ell)} &= \gamma_i^{(\ell)}, \quad \ell = 1, 2; \quad i = -M_\ell + 1, \dots, N_\ell - 1, \\ w_{-M_\ell}^{(\ell)} &= \frac{1}{1 + \rho^{(\ell)}} - \sum_{j=-M_\ell+1}^{N_\ell} \frac{1}{1 + e^{jih_\ell}} \gamma_j^{(\ell)}, \\ w_{N_\ell}^{(\ell)} &= \frac{\rho^{(\ell)}}{1 + \rho^{(\ell)}} - \sum_{j=-M_\ell}^{N_\ell-1} \frac{e^{jih_\ell}}{1 + e^{jih_\ell}} \gamma_j^{(\ell)}.\end{aligned}\tag{3.26}$$

We then get the approximation

$$p(x, y) \approx \sum_{i=-M_1}^{N_1} \sum_{j=-M_2}^{N_2} p_{i,j} w_i^{(1)}(x) w_j^{(2)}(y).\tag{3.27}$$

To get an idea of the complexity of the above procedure, we make the simplifying assumption that $M_j = N_j = N$, for $j = 1, 2$. We may readily deduce that if the above two dimensional “Laplace transform” F is either known explicitly, or if the evaluation of this transform can be reduced to the evaluation of a one-dimensional integral, then the complexity, i.e., the total amount of work required to achieve an error ε when carrying out the computations of the above algorithm (to approximate $p(x, y)$ at $(2N + 1)^2$ points) on a sequential machine, is $\mathcal{O}([\log(\varepsilon)]^6)$.

The above algorithm extends readily to ν dimensions, in which case the complexity for evaluating a ν -dimensional convolution integral (at $(2N + 1)^\nu$ points) by the above algorithm to within an error of ε is of the order of $[\log(\varepsilon)]^{2\nu+2}$.

The above results extend readily to “product regions” over more than one dimension.

(2) *Convolutions over Curvilinear Regions*

Curvilinear regions can also be dealt with relatively easily via *Sinc* methods. We now illustrate this, by deriving in detail a *Sinc convolution* algorithm over a two dimensional region, and then stating the algorithmic form of the three dimensional version (which is obvious at this point, once the derivation of the two dimensional algorithm has been carried out).

Sinc Convolution over a Two Dimensional Curvilinear Region.

Suppose that we are given a convolution integral over the region

$$\mathcal{B} = \{(t, \tau) : a_1 < t < b_1, a_2(t) < \tau < b_2(t)\} \quad (3.28)$$

where for purposes of this illustration a_1 and b_1 are finite and a_2 and b_2 are finite valued functions belonging to $\mathbf{M}_{\alpha, \beta}(\Gamma)$ can be readily dealt with by the following initial transformation of the differential equation:

$$\begin{aligned} t &= a_1 + (b_1 - a_1) \xi \\ \tau &= a_2(t) + (b_2(t) - a_2(t)) \eta, \end{aligned} \quad (3.29)$$

which transforms the square $\{(x, y) : 0 \leq \xi \leq 1, 0 \leq \eta \leq 1\}$ onto the region \mathcal{B} . Here we have not excluded the possibility that $a_2(a_1) = b_2(a_1)$ or that $a_2(b_1) = b_2(b_1)$. The resulting differential equation problem over the square can now be easily dealt with, starting with a product-type approximation, using a double sum, based on (3.3) (see (4.8) below).

We next extend the above two dimensional *Sinc convolution* algorithm to such a region \mathcal{B} defined in (3.28) above.

Consider the integral

$$r(x, y) = \int \int_{\mathcal{B}} f(x - t, y - \tau) g(t, \tau) dt d\tau, \quad (3.30)$$

where \mathcal{B} is given in (3.28) above.

We decompose r into a sum of 4 integrals:

$$r = r_1 + r_2 + r_3 + r_4 \quad (3.31)$$

with

$$\begin{aligned}
 r_1 &= \int_{a_1}^x \int_{a_2(t)}^y f(x-t, y-\tau) g(t, \tau) d\tau dt, \\
 r_2 &= \int_x^{b_1} \int_{a_2(t)}^y f(x-t, y-\tau) g(t, \tau) d\tau dt, \\
 r_3 &= \int_{a_1}^x \int_y^{b_2(t)} f(x-t, y-\tau) g(t, \tau) d\tau dt, \\
 r_4 &= \int_x^{b_1} \int_y^{b_2(t)} f(x-t, y-\tau) g(t, \tau) d\tau dt.
 \end{aligned} \tag{3.32}$$

Each of these integrals can be handled in exactly the same way. We thus illustrate here, an explicit procedure for approximating the first of the above integrals, i.e.,

$$p(x, y) = \int_{a_1}^x \int_{a_2(t)}^y f(x-t, y-\tau) g(t, \tau) d\tau dt \tag{3.33}$$

over the region B defined above.

It is again essential to illustrate in detail the steps of the reduction. To this end, we shall require the use of the “Laplace transforms”,

$$\begin{aligned}
 F(x, \sigma) &= \int_{E_2} \exp\left(-\frac{y}{\sigma}\right) f(x, y) dy, \\
 E_2 &\supset (0, \max_x(b_2(x) - a_2(x))), \\
 G(s, \sigma) &= \int_{E_1} \exp\left(-\frac{x}{s}\right) F(x, \sigma) dx, \\
 E_1 &\supset (0, \max_x(b_2 - a_2)),
 \end{aligned} \tag{3.34}$$

where we assume that both integrals exist whenever both variables are on their respective open right half planes.

We first apply the *Sinc convolution* procedure to the inner integral. To this end, we set

$$\begin{aligned} \tau &= \varphi^{-1}(w) \equiv \frac{a_2(t) + b_2(t) e^w}{1 + e^w} \\ \Leftrightarrow w &= \varphi(\tau) = \log\left(\frac{\tau - a_2(t)}{b_2(t) - \tau}\right) \\ \frac{1}{\varphi'(\tau)} &= [b_2(t) - a_2(t)] \frac{e^w}{(1 + e^w)^2}, \\ \tau_j &= \frac{a_2(t) + b_2(t) e^{jh_2}}{1 + e^{jh_2}}. \end{aligned} \tag{3.35}$$

In order to carry out indefinite convolution with respect to τ , we shall require the *indefinite integration matrix* $[b_2(t) - a_2(t)] \mathbf{A}_2$, with $\mathbf{A}_2 = h_2 I^{(-1)} D (e^w (1 + e^w)^{-2})$, and in which the variable w is evaluated at the points $j h_2$, $j = -M_2, \dots, N_2$. We then set $m_2 = M - 2 + N_2 + 1$, $\mathbf{A}_2 = \mathbf{X}_2 \mathbf{S}_2 \mathbf{X}_2^{-1}$, with \mathbf{S} a diagonal matrix with entries $s_{-M_2}^{(2)}, \dots, s_{N_2}^{(2)}$. We thus obtain

$$p(x, y) \approx \int_{a_1}^x \mathbf{w}_2(y) \mathbf{X}_2 F(x - t, [b_2(t) - a_2(t)] \mathbf{S}_2) \mathbf{X}_2^{-1} V_\tau g(t, \cdot) dt. \tag{3.36}$$

At this point $\mathbf{X}_2 F(x - t, [b_2(t) - a_2(t)] \mathbf{S}_2) \mathbf{S}_2^{-1} V_y g(t, \cdot)$ is a vector of order m_2 which accurately approximates the vector

$$V_y \int_{a_2(t)}^y f(x - t, y - \tau) d\tau$$

at the points τ_j defined in (3.35) above.

We emphasize here that the operator V_y transforms $g(t, \tau)$ into a vector with entries obtained by replacing τ with the numbers τ_j defined in (3.35) above. That is, the points τ_j also depend upon t . The vector $\mathbf{w}_2(y)$ of basis functions interpolates at the points τ_j . Removal of the basis \mathbf{w}_2 in (3.36) therefore defines this vector $\mathbf{p}(x)$; setting $\mathbf{q}(x) = \mathbf{X}_2^{-1} \mathbf{p}(x)$, $\mathbf{h}(t) = \mathbf{X}_2^{-1} V_y g(t, \cdot)$, and then taking the j^{th} component, $q_j(x)$ of $\mathbf{q}(x)$ and $h_j(t)$ of $\mathbf{h}(t)$, we get

$$q_j(x) = \int_{a_1}^x F(x - t, [b_2(t) - a_2(t)] s_j^{(2)}) h_j(t) dt. \tag{3.37}$$

We now again apply *Sinc convolution*, in the variable t to this integral, to get in the notation of (5.13), and with $m_1 = M_1 + N_1 + 1$,

$$q_j \approx G\left(\mathcal{J}_{m_1}, [b_2(\cdot) - a_2(\cdot)] s_j^{(2)}\right) h_j, \tag{3.38}$$

where, corresponding to any function u ,

$$(\mathcal{J}_{m_1} u)(x) = \mathbf{w}_1(x) [b_1 - a_1] \mathbf{A}_1 V_x u$$

$$\begin{aligned} \mathbf{A}_1 &= h_1 I^{(-1)} D \left(\frac{e^{jh_1}}{(1 + e^{jh_1})^2} \right), \\ \mathbf{A}_1 &= \mathbf{X}_1 \mathbf{S}_1 \mathbf{X}_1^{-1}, \end{aligned} \quad (3.39)$$

$$\mathbf{X}_1 = \begin{bmatrix} x_{i,j}^{(1)} \end{bmatrix}, \quad \mathbf{X}^{-1} = \begin{bmatrix} x_{(1)}^{i,j} \end{bmatrix}.$$

If one of the variables $s^{(i)}$ in the function $G(s^{(1)}, s^{(2)})$ is fixed, then $G(s, \sigma)$ is assumed to be analytic with respect to the other variable, on the whole right half plane. We can thus express G as a limit of finite sums of functions, in the form

$$G(s, \sigma) = \lim_{\nu \rightarrow \infty} \sum_{\ell=0}^{\kappa_\nu} \gamma_\ell^{(\nu)}(s) \delta_\ell^{(\nu)}(\sigma), \quad (3.40)$$

and where we may assume without loss of generality that each of the functions γ_ℓ and δ_ℓ are analytic on the right half plane. Denoting the right hand side of (3.38) by q_j^* , it thus follows from (3.40) that

$$\begin{aligned} q_j^* &= G \left(\mathcal{J}_{m_1}, [b_2(\cdot) - a_2(\cdot)] s_j^{(2)} \right) h_j(\cdot) \\ &= \lim_{\nu} \sum_{\ell} \gamma_\ell^{(\nu)} (\mathcal{J}_m) \delta_\ell^{(\nu)} \left([b_2(\cdot) - a_2(\cdot)] s_j^{(2)} \right) h_j(\cdot) \\ &= \mathbf{w}_1 \mathbf{X}_1 \lim_{\nu} \sum_{\ell} \gamma_\ell^{(\nu)} \left([b_1 - a_1] S^{(1)} \right) \mathbf{X}_1^{-1} \cdot \\ &\quad \cdot V_x \delta_\ell^{(\nu)} \left([b_2(\cdot) - a_2(\cdot)] s_j^{(2)} \right) h_j(\cdot). \end{aligned} \quad (3.41)$$

Now, by applying the operator V_x to each side (or equivalently, dropping the vector \mathbf{w}_1), then multiplying each side of this equation by \mathbf{X}_1^{-1} , setting $\mathbf{X}_1^{-1} V_x q_j^* = \mathbf{r}_j$, and denoting

$$t_k = \left(a_1 + b_1 e^{kh_1} \right) / \left(1 + e^{kh_1} \right),$$

we find that the i^{th} component of \mathbf{r}_j is just

$$\begin{aligned}
r_{i,j} &= \lim_{\nu} \sum_{\ell} \gamma_{\ell}^{(\nu)} \left([b_1 - a_1] s_i^{(1)} \right) \cdot \\
&\quad \cdot \sum_{k=-M_1}^{N_1} x_{(1)}^{i,k} \delta_{\ell}^{(\nu)} \left([b_2(t_k) - a_2(t_k)] s_j^{(2)} \right) h_j(t_k) \\
&= \sum_{k=-M_1}^{N_1} x_{(1)}^{i,k} \lim_{\nu} \sum_{\ell} \gamma_{\ell}^{(\nu)} \left([b_1 - a_1] s_i^{(1)} \right) \cdot \\
&\quad \cdot \delta_{\ell}^{(\nu)} \left([b_2(t_k) - a_2(t_k)] s_j^{(2)} \right) h_j(t_k) \\
&= \sum_{k=-M_1}^{N_1} x_{(1)}^{i,k} G \left([b_1 - a_1] s_i^{(1)}, [b_2(t_k) - a_2(t_k)] s_j^{(2)} \right) h_j(t_k).
\end{aligned} \tag{3.42}$$

Note that the interchange of order of summation in (3.42) was made possible by taking the i^{th} component of the vector \mathbf{r}_j , and that we able to take the final limit with respect to ν since $\gamma_{\ell}^{(\nu)}(s)$ and $\delta_{\ell}^{(\nu)}(\sigma)$ are evaluated at a finite number of points on the right half plane, where they are well defined, by assumption.

Now, having gotten $r_{i,j}$, we get $q_{i,j}$ and $p_{i,j}$ from the equations

$$q_{\cdot,j} = \mathbf{X}_1 r_{\cdot,j} \quad p_{i,\cdot} = \mathbf{X}_2 q_{i,\cdot}. \tag{3.43}$$

The final algorithm takes the form of Algorithm 3.2, which follows.

Algorithm 3.2

- (a) Determine parameters M_{ℓ} and N_{ℓ} , set $m_{\ell} = M_{\ell} + N_{\ell} + 1$, $\ell = 1, 2$, and form the matrices

$$\begin{aligned}
\mathbf{A}_{\ell} &= h_{\ell} I^{(-1)} D \left(\frac{e^w}{(1 + e^w)^2} \right); \quad w = k h_{\ell} \\
&= \mathbf{X}_{\ell} \mathbf{S}_{\ell} \mathbf{X}_{\ell}^{-1} \\
\mathbf{X}_1^{-1} &= \left[x_{(1)}^{i,j} \right], \quad \mathbf{S}_{\ell} = \text{diag} \left[s_{-M_{\ell}}^{(\ell)}, \dots, s_{N_{\ell}}^{(\ell)} \right], \quad \ell = 1, 2.
\end{aligned} \tag{3.44}$$

- (b) Form the values u_i and v_j , with $u_i = e^w / (1 + e^w)$, and $w = i h_1$, $i = -M_1, \dots, N_1$ and $v_j = e^w / (1 + e^w)$, with $w = j h_2$, $j = -M_2, \dots, N_2$. Then use these to get the *Sinc points* $t_i = (a_1 + (b_1 - a_1) u_i)$, and $\tau_{i,j} = a_2(t_i) + (b_2(t_i) - a_2(t_i)) v_j$.
- (c) Form the $m_1 \times m_2$ array $[g_{i,j}]$ with

$$g_{i,j} = g(t_i, \tau_{i,j}) .$$

(d) Replace the array $[g_{i,j}]$ with $[h_{i,j}]$, where

$$h_{i,\cdot} = \mathbf{X}_2^{-1} g_{i,\cdot} .$$

(e) Obtain the “Laplace transform”

$$G(s^{(1)}, s^{(2)}) = \int_{E_1} \int_{E_2} \exp \left\{ -\frac{x}{s^{(1)}} - \frac{y}{s^{(2)}} \right\} f(x, y) dy dx,$$

with the sets E_i defined as above.

(f) Replace the array $[h_{i,j}]$ with the array $[r_{i,j}]$, with

$$r_{i,j} = \sum_{k=-M_1}^{N_1} x_{(1)}^{i,k} G \left([b_1 - a_1] s_i^{(1)}, [b_2(t_k) - a_2(t_k)] s_j^{(2)} \right) h_{k,j} .$$

(g) Set

$$q_{\cdot,j} = \mathbf{X}_1 r_{\cdot,j} .$$

(h) Set

$$p_{i,\cdot} = \mathbf{X}_2 q_{i,\cdot} .$$

Remark: Having computed the array $\{p_{i,j}\}$, we can now approximate $p(x, y)$ at any point in \mathcal{B} by means of the formula

$$p(x, y) \approx \mathbf{w}_1(x) [p_{i,j}] (\mathbf{w}_2(x, y))^T , \quad (3.45)$$

where $\mathbf{w}_i = (w_{M_i}^{(i)}, \dots, w_{N_i}^{(i)})$, are defined as in (3.6), but with

$$\begin{aligned} \varphi_1(x) &= \log \left(\frac{x - a_1}{b_1 - x} \right) \\ \varphi_2(x, y) &= \log \left(\frac{y - a_2(x)}{b_2(x) - y} \right) \end{aligned} \quad (3.46)$$

Sinc Convolution over a Three Dimensional Curvilinear Region.

We now state an analogous algorithm for approximating indefinite convolutions over the region

$$\mathcal{B} = \{(x, y, z) :$$

$$a_1 < x < b_1, a_2(x) < y < b_2(x), a_3(x, y) < z < b_3(x, y)\} . \quad (3.47)$$

We wish to approximate a convolution integral of the form

$$R(x, y, z) = \int \int \int_{\mathcal{B}} f(x - \xi, y - \eta, z - \zeta) g(\xi, \eta, \zeta) d\xi d\eta d\zeta. \quad (3.48)$$

This definite convolution integral can be split up into 8 indefinite convolution integrals, the approximation of each of which is similar. We now give an explicit algorithm for approximating one of these, namely,

$$p(x, y, z) = \int_{a_1}^x \int_{a_2(x)}^y \int_{a_3(x,y)}^z f(x - \xi, y - \eta, z - \zeta) g(\xi, \eta, \zeta) d\zeta d\eta d\xi. \quad (3.49)$$

Algorithm 3.3

- (a) Determine parameters M_ℓ and N_ℓ , set $m_\ell = M_\ell + N_\ell + 1$, $\ell = 1, 2, 3$, and form the matrices

$$\begin{aligned} \mathbf{A}_\ell &= h_\ell I^{(-1)} D \left(\frac{e^w}{(1 + e^w)^2} \right); \quad w = k h_\ell \\ &= \mathbf{X}_\ell \mathbf{S}_\ell \mathbf{X}_\ell^{-1} \\ \mathbf{X}_\ell^{-1} &= [x_{(\ell)}^{i,j}], \quad \mathbf{S}_\ell = \text{diag} [s_{-M_\ell}^{(\ell)}, \dots, s_{N_\ell}^{(\ell)}], \quad \ell = 1, 2, 3. \end{aligned} \quad (3.50)$$

- (b) Form the values u_i , v_j , and w_k , where $u_i = e^\omega / (1 + e^\omega)$, with $\omega = i h_1$, $i = -M_1, \dots, N_1$, $v_j = e^\omega / (1 + e^\omega)$, with $\omega = j h_2$, $j = -M_2, \dots, N_2$, and $w_k = e^\omega / (1 + e^\omega)$, with $\omega = k h_3$, $k = -M_3, \dots, N_3$. Then use these to get the *Sinc points* $\xi_i = (a_1 + (b_1 - a_1)u_i)$, $\eta_{i,j} = a_2(\xi_i) + [b_2(\xi_i) - a_2(\xi_i)] v_j$, and

$$\zeta_{i,j,k} = a_2(\xi_i, \eta_j) + [b_3(\xi_i, \eta_j) - a_3(\xi_i, \eta_j)] \zeta_k.$$

- (c) Form the $m_1 \times m_2 \times m_3$ array $[g_{i,j,k}]$ with

$$g_{i,j,k} = g(\xi_i, \eta_{i,j}, \zeta_{i,j,k}).$$

- (d) Replace the array $[g_{i,j,k}]$ with $[h_{i,j,k}]$, where

$$h_{i,j,\cdot} = \mathbf{X}_3^{-1} g_{i,j,\cdot}.$$

- (e) It is convenient to obtain the “Laplace transform” by first taking

$$\begin{aligned} E_1 &\supset (0, [b_1 - a_1]) , \\ E_2 &\supset \left(0, \sup_{x \in (a_1, b_1)} [b_2(x) - a_2(x)] \right) , \\ E_3 &\supset \left(0, \sup_{(x,y) \in (a_1, b_1) \times (a_2(x), b_2(x))} [b_3(x, y) - a_3(x, y)] \right) , \end{aligned}$$

and then setting

$$\begin{aligned} G(s^{(1)}, s^{(2)}, s^{(3)}) \\ = \int_{E_1} \int_{E_2} \int_{E_3} \exp \left\{ -\frac{x}{s^{(1)}} - \frac{y}{s^{(2)}} - \frac{z}{s^{(3)}} \right\} f(x, y, z) dz dy dx . \end{aligned}$$

(f) Replace the array $[h_{i,j,k}]$ with the array $[r_{i,j,k}]$, with

$$r_{i,j,k} = \sum_{\ell=-M_1}^{N_1} x_{(1)}^{i,\ell} \sum_{m=-M_2}^{N_2} x_{(2)}^{j,m} G(\mu_i, \nu_{\ell,j}, \lambda_{\ell,m,k}) ,$$

and with

$$\begin{aligned} \mu_i &= [b_1 - a_1] s_i^{(1)} \\ \nu_{\ell,j} &= [b_2(\xi_\ell) - a_2(\xi_\ell)] s_j^{(2)} \\ \lambda_{\ell,m,k} &= [b_3(\xi_\ell, \eta_{l,m}) - a_3(\xi_\ell, \eta_{l,m})] s_k^{(3)} . \end{aligned}$$

(g) Form

$$r_{\cdot,j,k} = X_1 \sigma_{\cdot,j,k}$$

(h) Form

$$q_{i,\cdot,k} = \mathbf{X}_2 r_{i,\cdot,k} .$$

(i) Form

$$p_{i,j,\cdot} = \mathbf{X}_3 q_{i,j,\cdot} .$$

Remark: Having computed the array $\{p_{i,j,k}\}$, we can now approximate $p(x, y, z)$ at any point B by means of the formula

$$p(x, y, z) \approx \sum_{i=-M_1}^{N_1} \sum_{j=-M_2}^{N_2} \sum_{k=-M_3}^{N_3} p_{i,j,k} w_i^{(1)}(x) w_j^{(2)}(x, y) w_k^{(3)}(x, y, z) , \quad (3.51)$$

where $\mathbf{w}_i = (w_{M_i}^{(i)}, \dots, w_{N_i}^{(i)})$, are defined as in (3.6), but with

$$\begin{aligned}\varphi_1(x) &= \log\left(\frac{x - a_1}{b_1 - x}\right) \\ \varphi_2(x, y) &= \log\left(\frac{y - a_2(x)}{b_2(x) - y}\right) \\ \varphi_3(x, y, z) &= \log\left(\frac{z - a_3(x, y)}{b_3(x, y) - z}\right)\end{aligned}\tag{3.52}$$

4. Laplace Transforms of Green's Functions

In this section we present the “Laplace transforms” of standard Green’s functions for solving Poisson problems in 2 and 3 space dimensions, wave equation problems in one, two and three space and one time dimension, Helmholtz equation problems in two and three space dimensions, and heat equation problems in one, two and three space and one time dimension. The procedure used to carry out these derivations is essentially that developed first in [11].

The following lemma is applicable for evaluation of the “Laplace transforms” of all three types of equations, elliptic, hyperbolic, and parabolic. The lemma involves the correct evaluation of the integral

$$Q(a) = \int_C \frac{dz}{z - a} \tag{4.1}$$

for given $a \in \mathbb{C}$, with $C = \{z \in \mathbb{C} : z = e^{i\theta}, 0 < \theta < \pi/2\}$. As a convention, we let $\ln(a)$ denote the principal value of the logarithm, i.e., if $a = \xi + i\eta \in \mathbb{C}$, then $\ln(a) = \ln|a| + i \arg(a)$, with $\arg(a)$ taking its principal value in the range $-\pi < \arg(a) \leq \pi$.

LEMMA 4.1. *Let $a = \xi + i\eta \in \mathbb{C}$, set $L(a) = \xi + \eta - 1$ and let A denote the region*

$$A = \{a = \xi + i\eta \in \mathbb{C} : |a| < 1, L(a) > 0\}. \tag{4.2}$$

- If $a \in \mathbb{C} \setminus \overline{A}$, then

$$Q(a) = \ln\left(\frac{i - a}{1 - a}\right); \tag{4.3}$$

- If $a = 1$ or $a = i$, then $Q(a)$ is infinite.
- If $a \in A$, then

$$Q(a) = \ln\left|\frac{i - a}{1 - a}\right| + i\left(2\pi - \arg\left(\frac{i - a}{1 - a}\right)\right); \tag{4.4}$$

- If $L(a) = 0$ with $|a| < 1$, then

$$Q(a) = \ln \left| \frac{i-a}{1-a} \right| + i\pi; \quad (4.5)$$

- If $L(a) > 0$, and $|a| = 1$, then

$$Q(a) = \ln \left| \frac{i-a}{1-a} \right| + i\frac{\pi}{2}. \quad (4.6)$$

4.1. Transforms of Green's Functions of Poisson Problems.

- (1) *The Case for Planar Regions.* In this case, the Green's function $G(x, y)$ for which the expression

$$\Psi(x, y) = \int \int_B G(x - \xi, y - \eta) g(\xi, \eta) d\xi d\eta \quad (4.7)$$

defines a function u that solves the problem

$$\frac{\partial^2 \Psi(x, y)}{(\partial x)^2} + \frac{\partial^2 \Psi(x, y)}{(\partial y)^2} = -g(x, y), \quad (x, y) \in B \quad (4.8)$$

is given by the expression

$$G(x, y) = \frac{1}{2\pi} \log \left(\frac{1}{\sqrt{x^2 + y^2}} \right). \quad (4.9)$$

LEMMA 4.2. *If u and v are both on the open right half complex plane, then*

$$\begin{aligned} \hat{G}(u, v) &= \int_0^\infty \int_0^\infty \exp \left(-\frac{x}{u} - \frac{y}{v} \right) G(x, y) dx dy \\ &= \left\{ \frac{1}{u^2} + \frac{1}{v^2} \right\}^{-1} \left\{ -\frac{1}{4} + \frac{1}{2\pi} \left[\frac{v}{u} (\gamma - \ln(v)) + \frac{u}{v} (\gamma - \ln(u)) \right] \right\} \end{aligned} \quad (4.10)$$

- (2) *The "Laplace Transform" of $G(x, y) = (x^2 + y^2)^{-1/2}$.*

It is sometimes convenient to know this result for boundary integral equations, as well as for two obtaining the transform of the Green's function in three dimensions. We want to evaluate the integral

$$\hat{G}(u, v) = \int_0^\infty \int_0^\infty \exp \left(-\frac{x}{u} - \frac{y}{v} \right) G(x, y) dx dy. \quad (4.11)$$

Setting

$$\begin{aligned} \rho &= \sqrt{x^2 + y^2}, \quad x + i y = \rho z, \quad (z = e^{i\theta}) \\ \lambda &= \sqrt{\frac{1}{u^2} + \frac{1}{v^2}}, \quad \zeta = \sqrt{\frac{\frac{1}{u} + \frac{i}{v}}{\frac{1}{u} - \frac{i}{v}}}, \end{aligned} \tag{4.12}$$

we have

$$\begin{aligned} \frac{x}{u} + \frac{y}{v} &= \frac{\rho \lambda}{2} \left(\frac{z}{\zeta} + \frac{\zeta}{z} \right) \\ u + i v &= \lambda \zeta. \end{aligned} \tag{4.13}$$

Substituting these results into (4.11) above, we get, after integrating with respect to ρ ,

$$\hat{G}(u, v) = \frac{2}{\pi i} \int_C \frac{dz}{z \lambda \left(\frac{z}{\zeta} + \frac{\zeta}{z} \right)}, \tag{4.14}$$

where C is defined as in (4.1) above. Hence, after rational simplification, we get

$$\hat{G}(u, v) = \frac{-1}{\lambda} \{Q(i\zeta) - Q(-i\zeta)\}. \tag{4.15}$$

where Q is defined as in (4.1) above.

- (3) *The 3-d Green's function $(4\pi r)^{-1}$, with $r = \sqrt{x^2 + y^2 + z^2}$.*

We shall here give an explicit expression for

$$\hat{G}(u, v, w) = \int_0^\infty \int_0^\infty \int_0^\infty \frac{\exp \left\{ -\frac{x}{u} - \frac{y}{v} - \frac{z}{w} \right\}}{4\pi \sqrt{x^2 + y^2 + z^2}} dx dy dz. \tag{4.16}$$

This result enables us to obtain an accurate approximation for Ψ , with

$$\begin{aligned} \Psi(x, y, z) &= \int_{a_1}^{b_1} \int_{a_2}^{b_2} \int_{a_3}^{b_3} \frac{g(\xi, \eta, \zeta)}{4\pi \sqrt{(x - \xi)^2 + (y - \eta)^2 + (z - \zeta)^2}} d\xi d\eta d\zeta. \end{aligned} \tag{4.17}$$

in $V = (a_1, b_1) \times (a_2, b_2) \times (a_3, b_3)$. The function Ψ defined in (4.17) satisfies the equation $\Psi_{xx} + \Psi_{yy} + \Psi_{zz} = -g$ in V .

LEMMA 4.3. *Let $\hat{G}(u, v, w)$ be defined as in (4.16) for arbitrary complex u , v , and w located on the right half complex plane. Then*

$$\begin{aligned}\hat{G}(u, v, w) &= \left(\frac{1}{u^2} + \frac{1}{v^2} + \frac{1}{w^2} \right)^{-1} \cdot \\ &\quad \cdot \left\{ -\frac{1}{8} + H(u, v, w) + H(v, w, u) + H(w, u, v) \right\},\end{aligned}\tag{4.18}$$

where, setting

$$\begin{aligned}\lambda &= \sqrt{\frac{1}{v^2} + \frac{1}{w^2}}, \\ \zeta &= \sqrt{\frac{\frac{1}{v} + \frac{i}{w}}{\frac{1}{v} - \frac{i}{w}}}\end{aligned}\tag{4.19}$$

we have,

$$H(u, v, w) = -\frac{1}{8\pi u \lambda} \{Q(i\zeta) - Q(-i\zeta)\}\tag{4.20}$$

where $Q(a)$ is defined as in Lemma 4.1.

4.2. Transforms of Green's Functions of Wave Problems. We present here, the “Laplace transforms” of Green’s functions for problems in d space and one time dimension, for $d = 1, 2, 3$. The *Sinc convolution* technique has a considerable advantage in the solution of such problems, since for small time intervals *Sinc convolution* enables uniformly (in space and time) accurate solution of the corresponding integral equations via use of successive approximations. This feature should be useful for accurate and efficient solution of inverse problems stemming from the novel integral equations in the time domain.

(1) *The $d = 1$ Case.* We want to evaluate the integral

$$\hat{G}(u, \tau) = \int_0^\infty \int_0^\infty G(x, t) \exp\left(-\frac{x}{u} - \frac{t}{\tau}\right) d\tau dx,\tag{4.21}$$

where the Green’s function $G(x, t)$ is defined by the equations

$$\begin{aligned}\frac{1}{c^2} \frac{\partial^2 G(x, t)}{(\partial t)^2} - \frac{\partial^2 G(x, t)}{(\partial x)^2} &= \delta(t) \delta(x) \quad x \in \mathbb{R}, \quad t \in (0, T), \\ G(x, 0^+) &= \left. \frac{\partial G(x, t)}{\partial t} \right|_{t=0^+} = 0, \quad x \in \mathbb{R}.\end{aligned}\tag{4.22}$$

It is readily seen that the “Laplace transform” $\tilde{G}(x, \tau)$ of $G(x, t)$ then satisfies the differential equation

$$\tilde{G}_{xx}(x, \tau) - \frac{1}{c^2 \tau^2} \tilde{G}(x, \tau) = -\delta(x) \quad (4.23)$$

and solving this, we find that

$$\tilde{G}(x, \tau) = \frac{s \tau}{2} \exp\left(-\frac{|x|}{s \tau}\right). \quad (4.24)$$

Taking the “Laplace transform” of this equation with respect to the variable x , we now get

$$\hat{G}(u, \tau) = \frac{c \tau u}{c \tau + u}. \quad (4.25)$$

(2) *The $d = 2$ Case.* We shall derive the “Laplace transform”

$$\hat{G}(u, v, \tau) = \int_0^\infty \int_0^\infty \int_0^\infty G(x, y, t) \exp\left(-\frac{x}{u} - \frac{y}{v} - \frac{t}{\tau}\right) d\tau dx dy, \quad (4.26)$$

where G is defined for $(x, y) \in \mathbb{R}^2$ and $t \in (0, T)$ by the equations

$$\frac{1}{c^2} \frac{\partial^2 G(x, y, t)}{(\partial t)^2} - \frac{\partial^2 G(x, y, t)}{(\partial x)^2} - \frac{\partial^2 G(x, y, t)}{(\partial y)^2} = \delta(t) \delta(x) \delta(y),$$

$$G(x, y, 0^+) = \left. \frac{\partial G(x, y, t)}{\partial t} \right|_{t=0^+} = 0, \quad (x, y) \in \mathbb{R}^2. \quad (4.27)$$

LEMMA 4.4. *Let $\hat{G}(u, v, \tau)$ be defined as in (4.26) for all $\Re u > 0$, $\Re v > 0$, and $\Re \tau > 0$. Then*

$$\hat{G}(u, v, \tau) = \left(\frac{1}{c^2 \tau^2} - \frac{1}{u^2} - \frac{1}{v^2} \right)^{-1} \left(\frac{1}{4} - \hat{H}(u, v, \tau) - \hat{H}(v, u, \tau) \right), \quad (4.28)$$

where

$$\hat{H}(u, v, \tau) = \frac{1}{\pi u} \frac{1}{\sqrt{\frac{1}{c^2 \tau^2} - \frac{1}{v^2}}} \arctan \sqrt{\frac{\frac{1}{c \tau} - \frac{1}{v}}{\frac{1}{c \tau} + \frac{1}{v}}} \quad (4.29)$$

(3) *Sinc Convolution Solution of a Wave Equation Problem* For this case, Green’s function $G(\bar{r}, t) = G(x, y, z, t)$ for the wave equation satisfies the equations

$$\frac{1}{c^2} \frac{\partial^2 G(\bar{r}, t)}{\partial t^2} - \nabla^2 G(\bar{r}, t) = \delta(t) \delta^3(\bar{r}), \quad \bar{r} \in \mathbb{R}^3, \quad t \in (0, T)$$

$$G(\bar{r}, 0^+) = \frac{\partial G}{\partial t}(\bar{r}, 0^+) = 0, \quad (4.30)$$

The four dimensional “Laplace transform” of the function $G(\bar{r}, t)$ is defined by

$$\begin{aligned} & \hat{G}(u, v, w, \tau) \\ = & \int_0^\infty \int_0^\infty \int_0^\infty \int_0^\infty G(x, y, z, t) \exp \left\{ -\frac{x}{u} - \frac{y}{v} - \frac{z}{w} - \frac{t}{\tau} \right\} dx dy dz dt. \end{aligned} \quad (4.31)$$

LEMMA 4.5. Let \hat{G} be defined for all $\Re u > 0$, $\Re v > 0$, $\Re w > 0$ and $\Re \tau > 0$ by (4.31). Then

$$\begin{aligned} & \hat{G}(u, v, w, \tau) \\ = & \left(\frac{1}{c^2 \tau^2} - \frac{1}{u^2} - \frac{1}{v^2} - \frac{1}{w^2} \right)^{-1} \cdot \\ & \cdot \left\{ \frac{1}{8} - \hat{H}(u, v, w, \tau) - \hat{H}(v, w, u, \tau) - \hat{H}(w, u, v, \tau) \right\} \end{aligned} \quad (4.32)$$

where

$$\hat{H}(u, v, w, \tau) = \frac{i}{4\pi u \left(\frac{1}{c^2 \tau^2} - \frac{1}{v^2} - \frac{1}{w^2} \right)} \{Q(z_1) - Q(z_2)\}, \quad (4.33)$$

where, with

$$\zeta = \sqrt{\frac{\frac{1}{v} + \frac{i}{w}}{\frac{1}{v} - \frac{i}{w}}}, \quad (4.34)$$

$$\lambda = \sqrt{\frac{1}{v^2} + \frac{1}{w^2}},$$

we have

$$z_{1,2} = -\frac{\zeta}{\lambda} \left\{ \frac{1}{c\tau} \pm \sqrt{\frac{1}{c^2 \tau^2} - \lambda^2} \right\}. \quad (4.35)$$

and where $Q(\zeta)$ is defined as in Lemma 4.1.

- (4) *Helmholtz Equations.* The Green's functions for these equations are simple replacements of the ones above, i.e., we have

$$\begin{aligned} G(\bar{\rho}) &= \frac{i}{4} H_0^{(1)}(k \rho) \\ \nabla^2 G(\bar{\rho}) + k^2 G(\bar{\rho}) &= -\delta^2(\bar{\rho}) \end{aligned} \quad (4.36)$$

in two dimensions, with $H_0^{(1)}$ denoting the *Hankel* function, and

$$\begin{aligned} G(\bar{r}) &= \frac{e^{ikr}}{4\pi r} \\ \nabla^2 G(\bar{r}) + k^2 G(\bar{r}) &= -\delta^3(\bar{r}) \end{aligned} \quad (4.37)$$

in three dimensions. The “Laplace transforms” of these can be readily obtained from the above by replacing $1/(c\tau)$ by $-ik$ in Lemmas 4.4 and 4.5 above.

4.3. Transforms of Green's Functions of Heat Problems. We consider here, obtaining the d -dimensional “Laplace transforms” of the Free space Green's functions in $\mathbb{R}^d \times (0, \infty)$, where, with

$$r = |\bar{r}| = \sqrt{(x^1)^2 + (x^2)^2 + \dots + (x^d)^2},$$

we have

$$G(r, t) = \frac{1}{(4\pi\varepsilon t)^{d/2}} \exp\left(-\frac{r^2}{4\varepsilon t}\right). \quad (4.38)$$

For $(\bar{r}, t) \in \mathbb{R}^d \times (0, \infty)$, this Green's function satisfies the equations

$$\frac{\partial G(\bar{r}, t)}{\partial t} - \varepsilon \nabla_{\bar{r}}^2 G(\bar{r}, t) = \delta(t) \delta^d(\bar{r}) \quad (4.39)$$

$$G(\bar{r}, 0) = 0.$$

Hence, setting

$$\tilde{G}(\bar{r}, \tau) = \int_0^\infty \exp\left(-\frac{t}{\tau}\right) G(\bar{r}, t) dt, \quad (4.40)$$

we arrive at the differential equation

$$\nabla_{\bar{r}}^2 \tilde{G}(\bar{r}, \tau) - \frac{1}{\varepsilon \tau} \tilde{G}(\bar{r}, \tau) = -\frac{1}{\varepsilon} \delta^3(\bar{r}). \quad (4.41)$$

The results of §4.2 then immediately yield the Laplace transforms $\hat{G}(\bar{\lambda}, \tau)$, with $\bar{\lambda} = (u^1, \dots, u^d)$, and

$$\begin{aligned} & \hat{G}(\bar{\lambda}, \tau) \\ &= \int \cdots \int_{\mathbb{R}^d} \int_0^\infty \exp \left(-\frac{x^1}{u^1} - \cdots - \frac{x^d}{u^d} - \frac{t}{\tau} \right) G(x^1, \dots, x^d, t) d\bar{r} dt. \end{aligned} \quad (4.42)$$

We summarize these results for the cases of $d = 1, 2, 3$ in the following Lemma.

LEMMA 4.6. *Let the d -dimensional Laplace transform \hat{G} of the Green's function $G(r, t)$ be defined as in (4.42), for all $\Re u^k > 0$. Then, with $u^1 = u$, $u^2 = v$, and $u^3 = w$, we have:*

(i) *For the case of $d = 1$, from Equation (4.25),*

$$\hat{G}(u, \tau) = \frac{1}{\varepsilon} \frac{(\varepsilon \tau)^{1/2} u}{(\varepsilon \tau)^{1/2} + u}. \quad (4.43)$$

(ii) *For the case of $d = 2$, from (4.28)–(4.29),*

$$\begin{aligned} \hat{G}(u, v, \tau) &= \frac{1}{\varepsilon} \left(\frac{1}{\varepsilon \tau} - \frac{1}{u^2} - \frac{1}{v^2} \right)^{-1} \left(\frac{1}{4} - \hat{H}(u, v, \tau) - \hat{H}(v, u, \tau) \right) \\ \hat{H}(u, v, \tau) &= \frac{1}{\pi u} \frac{1}{\sqrt{\frac{1}{\varepsilon \tau} - \frac{1}{v^2}}} \arctan \sqrt{\frac{\frac{1}{\sqrt{\varepsilon \tau}} - \frac{1}{v}}{\frac{1}{\sqrt{\varepsilon \tau}} + \frac{1}{v}}}. \end{aligned} \quad (4.44)$$

(iii) *For the case of $d = 3$, from Lemma 4.5,*

$$\begin{aligned} \hat{G}(u, v, w, \tau) &= \frac{1}{\varepsilon} \left(\frac{1}{\varepsilon \tau} - \frac{1}{u^2} - \frac{1}{v^2} - \frac{1}{w^2} \right)^{-1} \cdot \\ &\quad \cdot \left(\frac{1}{8} - \hat{H}(u, v, w, \tau) - \hat{H}(v, w, u, \tau) - \hat{H}(w, u, v, \tau) \right), \end{aligned} \quad (4.45)$$

where, with $\lambda = (1/v^2 + 1/w^2)^{1/2}$ and $Q(z)$ as in Lemma 4.1,

$$\begin{aligned} \hat{H}(u, v, w, \tau) &= \frac{i}{4\pi u \left(\frac{1}{\varepsilon \tau} - \lambda^2 \right)} \{Q(z_1) - Q(z_2)\} \\ \zeta &= \sqrt{\frac{\frac{1}{u} + \frac{i}{v}}{\frac{1}{u} - \frac{i}{v}}} \\ z_{1,2} &= -\frac{\zeta}{\lambda} \left\{ \sqrt{\frac{1}{\varepsilon \tau}} \pm \sqrt{\frac{1}{\varepsilon \tau} - \lambda^2} \right\}. \end{aligned} \quad (4.46)$$

References

- [1] F. Stenger B. Barkey and R. Vakili. Sinc Convolution Method of Solution of Burgers' Equation. In K. Bowers and J. Lund, editors, *Proceedings of Computation and Control III*, pages 341–354. Birkhäuser, 1993.
- [2] B. Bialecki and F. Stenger. Sinc -Nyström Method for Numerical Solution of One Dimensional Cauchy Singular Integral Equations Given on a Smooth Open Arc in the Complex Plane. *Math. Comp.*, (51):133–165, 1988.
- [3] Y.H. Chiu. Integral Equation Solution of $\Delta u + k^2 u = 0$ in the Plane. Ph.D. Thesis, University of Utah, 1977.
- [4] I. Daubechies. *Ten Lectures on Wavelets*. SIAM, Philadelphia, 1992.
- [5] D. Elliott and F. Stenger. Sinc Method of Solution of Singular Integral Equations. In *IMACS Conference on CSIE, Philadelphia, P.A.*, pages 155–166. 1984.
- [6] B. Barkey F. Stenger and R. Vakili. Sinc Convolution Method of Solution of Burgers' Equation. In K. Bowers and J. Lund, editors, *Proceedings of Computation and Control III*, pages 341–354. Birkhäuser, 1993.
- [7] M. Hagmann F. Stenger and J. Schwing. An Algorithm for Computing the Electromagnetic Scattered Field from an Axially Symmetric Body with an Impedance Boundary Condition. *J. Math. Anal. Appl.*, (78):531–573, 1980.
- [8] K. Sikorski M. Kowalski and F. Stenger. *Selected Topics in Approximation and Computation*. Oxford University Press, 1993.
- [9] P. Monk and E. Süli. Convergence Analysis of Yee's Scheme on Nonuniform Grids. *SIAM J. Numer. Anal.*, (31):393–412, 1994.
- [10] P.M. Morse and H. Feshbach. *Method of Theoretical Physics, Part I*. McGraw–Hill, 1953.
- [11] A. Naghsh-Nilchi. Iterative Sinc–Convolution Method for Solving Three–Dimensional Electromagnetic Models. Ph.D. thesis, University of Utah (1997).
- [12] S. Narasimhan. Sinc Solution of Two Dimensional Problems in Mechanical Engineering. Ph. D. Thesis, Univesity of Utah, 1999.
- [13] F. Stenger S.-Å. Gustafson B. Keyes M. O'Reilly and K. Parker. ODE–IVP—PACK, via Sinc Indefinite Integration and Newton Iteration. *Numerical Algorithms*, (20):241–268, 1999.
- [14] J. Schwing. Numerical Solution of Integral Equations in Potential Theory Problems. Ph.D. Thesis, University of Utah, 1976.
- [15] F. Stenger. *Numerical Methods Based on Sinc and Analytic Functions*. Springer–Verlag, 1993.
- [16] F. Stenger. Collocating Convolutions. *Math. Comp.*, (64):211–235, 1995.
- [17] F. Stenger. Sinc Approximation for Cauchy–Type Singular Integrals over Arcs, paper in Honor of David Elliott. *The Anzian Journal*, (42):87–97, 2000.
- [18] F. Stenger and R. Schmidlein. Conformal Maps via Sinc Methods. In St. Ruscheweyh N. Papamichael and E.B. Saff, editors, *Computational Methods in Function Theory (CMFT '97)*, pages 505–549. World Scientific Publishing Co. Pte. Ltd., 1999.
- [19] M. Stromberg. Solution of Shock Problems by Methods Using Sinc Functions. 1986.
- [20] K. Yee. Numerical Solution of Boundary Value Problems Involving Maxwell's Equations in Isotropic Media. *IEEE Trans., Antennas and Propagation*, (AP–16):302–307, 1966.

FRANK STENGER, SCHOOL OF COMPUTING, UNIVERSITY OF UTAH,
SALT LAKE CITY, UTAH 84112
E-mail address: stenger@cs.utah.edu

AHMAD REZA NAGHSH-NILCHI, COMPUTER ENGINEERING DEPARTMENT,
THE UNIVERSITY OF ISFAHAN, ISFAHAN, POSTAL CODE: 81744, IRAN

JENNY NIEBSCH, ZENTRUM FÜR TECHNOMATHEMATIK, UNIVERSITÄT BREMEN,
GERMANY
E-mail address: niebsch@math.uni-bremen.de

RONNY RAMLAU, ZENTRUM FÜR TECHNOMATHEMATIK, UNIVERSITÄT BREMEN,
GERMANY
E-mail address: ramlau@math.uni-bremen.de

This page intentionally left blank

Diffusion and Regularization of Vector- and Matrix-Valued Images

Joachim Weickert and Thomas Brox

ABSTRACT. The goal of this paper is to present a unified description of diffusion and regularization techniques for vector-valued as well as matrix-valued data fields. In the vector-valued setting, we first review a number of existing methods and classify them into linear and nonlinear as well as isotropic and anisotropic methods. For these approaches we present corresponding regularization methods. This taxonomy is applied to the design of regularization methods for variational motion analysis in image sequences. Our vector-valued framework is then extended to the smoothing of positive semidefinite matrix fields. In this context a novel class of anisotropic diffusion and regularization methods is derived and it is shown that suitable algorithmic realizations preserve the positive semidefiniteness of the matrix field without any additional constraints. As an application, we present an anisotropic nonlinear structure tensor and illustrate its advantages over the linear structure tensor.

1. Introduction

In digital image processing, vector- and matrix-valued data sets are becoming increasingly important. This is caused by rapidly dropping prices for color imaging devices as well as by novel imaging techniques such as Diffusion Tensor MRI. Often these data suffer from noise creating the need for image restoration methods that allow to remove the noise without severely affecting important structures such as image discontinuities (edges).

In the present paper we will review some recent techniques that achieve this goal by using nonlinear diffusion or regularization approaches. A unifying description is presented that includes diffusion and regularization techniques, linear and nonlinear approaches as well as isotropic and anisotropic methods for vector- or matrix-valued data sets. However, we will not only confine ourselves to the review of existing techniques, we will also present several novel approaches that have not been considered before. Since this paper is mainly intended as a means to communicate the essential ideas and structural similarities, we do not go very deeply into

1991 *Mathematics Subject Classification*. Primary 68T10, 68T45; Secondary 35J60, 35K55.

Key words and phrases. Image processing, diffusion filtering, regularization methods.

Our research on matrix-valued smoothing methods is partly funded by the projects WE 2602/1-1 and SO 363/9-1 of the *Deutsche Forschungsgemeinschaft (DFG)*.

mathematical details. Such details and full proofs can be found in more specialized publications.

Our paper is organized as follows. In Section 2 we first review diffusion techniques for vector-valued images before we present energy functionals for corresponding regularization methods. This taxonomy is then used for classifying variational approaches for motion analysis in image sequences. Section 3 is devoted to matrix-valued image processing. In analogy to our discussions in the vector-valued case, we present diffusion and regularization methods in the isotropic and anisotropic setting. The latter methods are studied here for the first time. We argue that these methods are capable of preserving the positive semidefiniteness of an initial matrix field without the need to impose additional constraints. Finally we apply our ideas to the generalization of the linear structure tensor, a very successful tool for analysing corners, textures and flow-like structures, to the nonlinear setting. Our paper is concluded with a summary in Section 4.

2. Vector-Valued Filtering

2.1. Diffusion of Vector-Valued Images. Vector-valued images arise for example as color images, multi-spectral satellite images and multi-spin echo MR images. Diffusion filtering of some multichannel image $f = (f_1(x, y), \dots, f_m(x, y))^\top$ may be based on one of the following evolutions:

- (a) *Homogeneous diffusion* ([Iij59] in the scalar case):

$$(2.1) \quad \partial_t u_i = \Delta u_i \quad (i = 1, \dots, m)$$

- (b) *Linear isotropic diffusion* ([Fri92] in the scalar case):

$$(2.2) \quad \partial_t u_i = \operatorname{div} \left(g \left(\sum_j |\nabla f_j|^2 \right) \nabla u_i \right) \quad (i = 1, \dots, m)$$

- (c) *Linear anisotropic diffusion* ([Iij62] in the scalar case):

$$(2.3) \quad \partial_t u_i = \operatorname{div} \left(D \left(\sum_j \nabla f_j \nabla f_j^\top \right) \nabla u_i \right) \quad (i = 1, \dots, m)$$

- (d) *Nonlinear isotropic diffusion* [GKKJ92]:

$$(2.4) \quad \partial_t u_i = \operatorname{div} \left(g \left(\sum_j |\nabla u_j|^2 \right) \nabla u_i \right) \quad (i = 1, \dots, m)$$

- (e) *Nonlinear anisotropic diffusion* [Wei94]:

$$(2.5) \quad \partial_t u_i = \operatorname{div} \left(D \left(\sum_j \nabla u_j \nabla u_j^\top \right) \nabla u_i \right) \quad (i = 1, \dots, m)$$

with f as initial condition:

$$(2.6) \quad u_i(x, y, 0) = f_i(x, y) \quad (i = 1, \dots, m).$$

Here, g denotes a scalar-valued diffusivity, and D is a positive definite diffusion matrix. The diffusivity $g(s^2)$ is a decreasing function in its argument. Moreover,

we assume that the flux function $g(s^2)s$ is nondecreasing in s . One may e.g. use [NS98]

$$(2.7) \quad g(s^2) = \alpha + \frac{1}{\sqrt{\beta^2 + s^2}}.$$

with some small positive numbers α and β . In the linear case this ensures that at edges of the initial image f , where $\sum_j |\nabla f_j|^2$ is large, the diffusivity $g(\sum_j |\nabla f_j|^2)$ is close to zero. Consequently, diffusion at edges is inhibited. In the nonlinear case one introduces a feedback by adapting the diffusivity g to the evolving image u . In physics, a diffusion process with a scalar-valued diffusivity is called *isotropic*, since its diffusive behavior does not depend on the direction.

Anisotropic diffusion with a direction depending behavior may be realized by replacing the scalar-valued diffusivity g by some positive definite diffusion matrix D . One may design the diffusion matrix D such that diffusion along edges of f or u is preferred and diffusion across edges is inhibited. This may be very useful in cases when noisy edges are present.

How can edge directions in some vector-valued image f be measured? Di Zenzo [Di 86] has proposed to consider the matrix $\sum_j \nabla f_j \nabla f_j^\top$. It serves as a structure tensor for vector-valued images since its eigenvectors v_1, v_2 describe the directions of highest and lowest contrast. This contrast is given by the corresponding eigenvalues μ_1 and μ_2 .

A natural choice for the design of some diffusion matrix D as a function of a vector-valued image f would thus be to specify its eigenvectors as the eigenvectors v_1, v_2 of $\sum_j \nabla f_j \nabla f_j^\top$, and its eigenvalues λ_1, λ_2 via

$$(2.8) \quad \lambda_1 = g(\mu_1),$$

$$(2.9) \quad \lambda_2 = g(\mu_2),$$

with a diffusivity function g as e.g. in (2.7).

REMARK 2.1. The fact that in the preceding models the same diffusivity or diffusion matrix is used for all channels ensures that the evolutions between the channels are synchronized. This prevents e.g. that discontinuities are created at different locations in each channel.

REMARK 2.2. Let $J \in \mathbb{R}^{2 \times 2}$ be symmetric with eigenvectors v_1, v_2 and eigenvalues μ_1, μ_2 :

$$(2.10) \quad J = \mu_1 v_1 v_1^\top + \mu_2 v_2 v_2^\top.$$

A formal way to extend some scalar-valued function $g(s^2)$ to a matrix-valued function $g(J)$ is to define

$$(2.11) \quad g(J) := g(\mu_1) v_1 v_1^\top + g(\mu_2) v_2 v_2^\top.$$

With this notation we may characterize the linear and nonlinear isotropic models by their diffusivities $g(\sum_j \nabla f_j^\top \nabla f_j)$ and $g(\sum_j \nabla u_j^\top \nabla u_j)$, while their anisotropic counterparts are given by $g(\sum_j \nabla f_j \nabla f_j^\top)$ and $g(\sum_j \nabla u_j \nabla u_j^\top)$. Hence, isotropic and anisotropic models only differ by the location of the transposition.

REMARK 2.3. It should be noted that the preceding models are not the only diffusion methods that have been proposed for processing vector-valued images. For alternative approaches the reader is referred to [BC98, KMS00, Sap01, TD01a, Wei99].

REMARK 2.4. The requirement of having a nondecreasing flux $g(s^2)s$ has been introduced in order to ensure well-posedness in the nonlinear setting using classical frameworks such as maximal monotone operators [Bre73]. It is also possible to use more sophisticated models that allow contrast enhancement. In this case one can establish well-posedness results if some Gaussian presmoothing is introduced in the diffusivity or the diffusion matrix [CLMC92, Wei98].

Experiments. Figure 1 illustrates the effect of the different smoothing strategies for a noisy color image with three channels corresponding to the red, green and blue components. We observe that homogeneous diffusion performs well with respect to denoising, but does not respect image edges. Space-variant linear isotropic diffusion, however, may suffer from noise sensitivity as strong noise may be misinterpreted as an important edge structure where the diffusivity is reduced. Anisotropic linear diffusion allows smoothing along edges, but reduces smoothing across them. This leads to a better performance than isotropic linear diffusion if images are noisy. We can also observe that nonlinear models give better results than their linear counterparts. This is not surprising, since the nonlinear models adapt the diffusion process to the evolving image instead of the initial one.

2.2. Regularization Methods for Vector-Valued Images. Let us now explain some connections between the preceding vector-valued diffusion filters and regularization methods for vector-valued data. To this end we consider minimizers of the following energy functionals over some rectangular image domain Ω :

(a) *homogeneous regularization:*

$$(2.12) \quad E_{HV}(u) = \frac{1}{2} \int_{\Omega} \left(|f - u|^2 + \alpha \sum_k |\nabla u_k|^2 \right) dx dy$$

(b) *linear isotropic regularization:*

$$(2.13) \quad E_{LIV}(u) = \frac{1}{2} \int_{\Omega} \left(|f - u|^2 + \alpha g \left(\sum_j |\nabla f_j|^2 \right) \sum_k |\nabla u_k|^2 \right) dx dy$$

(c) *linear anisotropic regularization:*

$$(2.14) \quad E_{LAV}(u) = \frac{1}{2} \int_{\Omega} \left(|f - u|^2 + \alpha \sum_k \nabla u_k^\top g \left(\sum_j \nabla f_j \nabla f_j^\top \right) \nabla u_k \right) dx dy$$

(d) *nonlinear isotropic regularization:*

$$(2.15) \quad E_{NIV}(u) = \frac{1}{2} \int_{\Omega} \left(|f - u|^2 + \alpha \Psi \left(\sum_k |\nabla u_k|^2 \right) \right) dx dy$$

(e) *nonlinear anisotropic regularization:*

$$(2.16) \quad E_{NAV}(u) = \frac{1}{2} \int_{\Omega} \left(|f - u|^2 + \alpha \operatorname{tr} \Psi \left(\sum_k \nabla u_k \nabla u_k^\top \right) \right) dx dy$$

with some penalizing function $\Psi(s^2)$ that is differentiable in its argument and convex in s . Moreover, we assume that there exist constants $c_1, c_2 > 0$ such that $c_1 s^2 \leq \Psi(s^2) \leq c_2 s^2$ for all s . In the experiments for this paper we use the Nashed–Scherzer regularizer [NS98]

$$(2.17) \quad \Psi(s^2) := \alpha s^2 + \sqrt{\beta^2 + s^2}$$

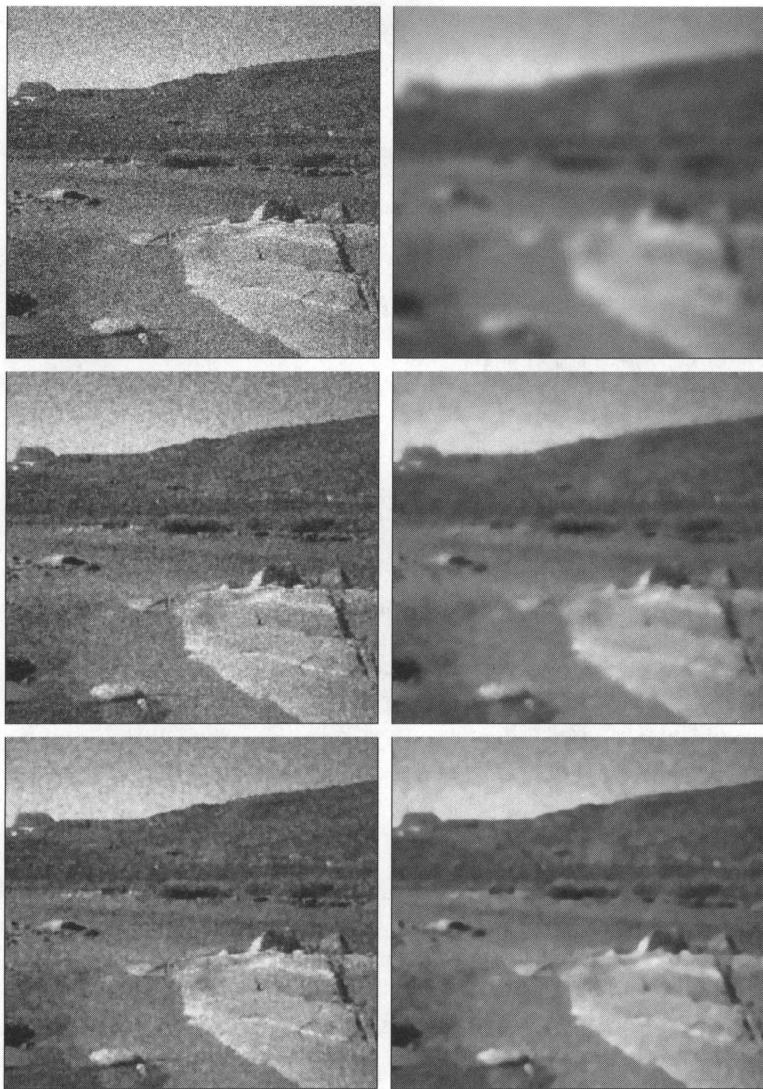


FIGURE 1. (a) TOP LEFT: Noisy color image. (b) TOP RIGHT: Homogeneous diffusion. (c) MIDDLE LEFT: Linear isotropic diffusion. (d) MIDDLE RIGHT: Linear anisotropic diffusion. (e) BOTTOM LEFT: Nonlinear isotropic diffusion. (f) BOTTOM RIGHT: Nonlinear anisotropic diffusion.

with some small parameters $\alpha, \beta > 0$. Its derivative is given by the diffusivity (2.7). Under the preceding assumptions one can show that the convex minimization problems (2.12)–(2.16) are well-posed in the Sobolev space $H^1(\Omega) \times \dots \times H^1(\Omega)$. Their unique solution satisfies the following Euler-Lagrange equations:

(a) *homogeneous regularization:*

$$(2.18) \quad \frac{u_i - f_i}{\alpha} = \Delta u_i \quad (i = 1, \dots, m)$$

(b) *isotropic linear regularization:*

$$(2.19) \quad \frac{u_i - f_i}{\alpha} = \operatorname{div} \left(g \left(\sum_k |\nabla f_k|^2 \right) \nabla u_i \right) \quad (i = 1, \dots, m)$$

(c) *anisotropic linear regularization:*

$$(2.20) \quad \frac{u_i - f_i}{\alpha} = \operatorname{div} \left(g \left(\sum_k \nabla f_k \nabla f_k^\top \right) \nabla u_i \right) \quad (i = 1, \dots, m)$$

(d) *isotropic nonlinear regularization:*

$$(2.21) \quad \frac{u_i - f_i}{\alpha} = \operatorname{div} \left(\Psi' \left(\sum_k |\nabla u_k|^2 \right) \nabla u_i \right) \quad (i = 1, \dots, m)$$

(e) *anisotropic nonlinear regularization:*

$$(2.22) \quad \frac{u_i - f_i}{\alpha} = \operatorname{div} \left(\Psi' \left(\sum_k \nabla u_k \nabla u_k^\top \right) \nabla u_i \right) \quad (i = 1, \dots, m)$$

with homogeneous Neumann boundary conditions.

While this is very easy to verify for the cases (a)–(d), the proof for the case (e) is more involved. More details can be found in a recent paper [WS01] where these anisotropic nonlinear regularizers have been analyzed first.

We may regard the elliptic equations (2.18)–(2.22) as fully implicit time discretizations of the parabolic diffusion filters (2.1)–(2.4) with initial value f and time step size α . This connection has been used in [SW00, RSW00] to establish a scale-space theory for noniterated and iterated scalar-valued regularization methods. The results include well-posedness, maximum-minimum principles, a large family of Lyapunov functionals and convergence to a flat image as $\alpha \rightarrow \infty$. This reasoning can also be extended to the vector-valued case. Experiments in [SW00, RSW00] showed that even for large regularization parameters α the regularization methods and their diffusion counterparts are visually fairly similar. This is also the case for the vector-valued setting, so we refrain from showing experimental results, since they can hardly be distinguished from those for diffusion filtering.

2.3. Application: Variational Image Sequence Analysis. Let us now apply the preceding concepts to the analysis of image sequences [WS01].

One of the main goals of image sequence analysis is the recovery of the so-called *optic flow field*. Optic flow describes the apparent motion of structures in the image plane. It can be used in a large variety of applications ranging from the recovery of motion parameters in robotics to the design of efficient algorithms for second generation video compression.

In the following we consider an image sequence $f(x, y, z)$ where $(x, y) \in \Omega$ denotes the location and $z \in [0, Z]$ is the time. We are looking for the optic flow field $\begin{pmatrix} u_1(x, y, z) \\ u_2(x, y, z) \end{pmatrix}$ which describes the correspondence of image structures at different times.

Very frequently it is assumed that image structures do not change their grey value over time. Therefore, along their path $(x(z), y(z))$ one obtains

$$(2.23) \quad 0 = \frac{df(x(z), y(z), z)}{dz} = f_x u_1 + f_y u_2 + f_z.$$

This brightness constancy assumption is called *optic flow constraint (OFC)*. It is not sufficient to determine $u := (u_1, u_2)$ uniquely. As a remedy, a regularizing smoothness constraint may be introduced such that the optic flow problem can be solved within a variational framework. We may recover the optic flow as minimizer of some convex functional of type

$$(2.24) \quad E(u) := \int_{\Omega} \left(\frac{1}{2} \underbrace{(f_x u_1 + f_y u_2 + f_z)^2}_{\text{data term}} + \alpha \underbrace{V(\nabla f, \nabla u)}_{\text{regularizer}} \right) dx dy$$

where $V(\nabla f, \nabla u)$ penalizes deviations from (piecewise) smoothness, and $\nabla u := (\nabla u_1, \nabla u_2)$. The corresponding gradient descent equations are given by

$$(2.25) \quad \partial_t u_1 = \partial_x V_{u_{1x}} + \partial_y V_{u_{1y}} - \frac{1}{\alpha} f_x (f_x u_1 + f_y u_2 + f_z),$$

$$(2.26) \quad \partial_t u_2 = \partial_x V_{u_{2x}} + \partial_y V_{u_{2y}} - \frac{1}{\alpha} f_y (f_x u_1 + f_y u_2 + f_z).$$

This diffusion–reaction system allows to recover the optic flow u as solution for $t \rightarrow \infty$. We observe that the regularizer $V(\nabla f, \nabla u)$ creates the vector-valued diffusion processes

$$(2.27) \quad \partial_t u_i = \partial_x V_{u_{ix}} + \partial_y V_{u_{iy}} \quad (i = 1, 2).$$

Specific choices of V allow to design regularizers that smooth the flow field, but respect semantically important image discontinuities or flow discontinuities. In the first case, we call the method *image-driven*, in the second case it is a *flow-driven* method. If the regularizer corresponds to an isotropic diffusion process, it is named *isotropic*, otherwise it is an *anisotropic* regularizer.

Table 1 gives an overview of the different vector-valued diffusion processes that we have just discussed, and their corresponding optic flow regularizers. Firstly we observe that diffusion filters have been discovered several years ahead of their corresponding optic flow regularizers. Secondly, it becomes clear that image-driven regularizers always correspond to linear diffusion processes, while flow-driven ones can be related to nonlinear diffusion filters.

It is possible to treat all these optic flow methods within a unifying theoretical framework. In [WS01] the following well-posedness results have been established.

THEOREM 2.5. *Let $V(\nabla f, \nabla u)$ be one of the optic flow regularizers from Table 1. Moreover, let us assume that*

- Ψ is differentiable, and $\Psi(s^2)$ is strictly convex in $s \in \mathbb{R}$.
- There exist $c_1, c_2 > 0$ such that $c_1 s^2 \leq \Psi(s^2) \leq c_2 s^2$ for all s .
- $f \in H^1(\Omega \times (0, T))$
- f_x, f_y are linearly independent in $L^2(\Omega)$, and $f_x, f_y \in L^\infty(\Omega)$.

Then the optic flow functional (2.24) has a unique minimizer $u(z) \in H^1(\Omega) \times H^1(\Omega)$ that depends in a continuous way on the image sequence f .

In order to illustrate the influence of the different regularization methods, we used the marbled block sequence of Otte and Nagel (KOGS/IAKS, University of Karlsruhe, Germany) [ON95]. These images can be downloaded from the web site <http://i21www.ira.uka.de/image-sequences>. The sequence consists of 31 frames of size 512×512 . In our case we only used frame 16 and 17. Figure 2 depicts

TABLE 1. Vector-valued diffusion processes and their corresponding optic flow regularizers. In the diffusion context, f denotes the vector-valued initial image and u its evolution. In the optic flow setting, f is the scalar-valued image sequence and u describes the optic flow field.

vector-valued diffusion process $\partial_t u_i = \partial_x V_{u_{ix}} + \partial_y V_{u_{iy}}$	optic flow regularizer $V(\nabla f, \nabla u)$
homogeneous $\partial_t u_i = \Delta u_i$ (scalar case: Iijima 1959 [Iij59])	homogeneous $\sum_{i=1}^2 \nabla u_i ^2$ (Horn/Schunck 1981 [HS81])
linear isotropic $\partial_t u_i = \operatorname{div} \left(g(\sum_j \nabla f_j ^2) \nabla u_i \right)$ (scalar case: Fritsch 1992 [Fri92])	image-driven, isotropic $g(\nabla f ^2) \sum_{i=1}^2 \nabla u_i ^2$ (Alvarez et al. 1999 [AELS99])
linear anisotropic $\partial_t u_i = \operatorname{div} \left(g(\sum_j \nabla f_j \nabla f_j^\top) \nabla u_i \right)$ (scalar case: Iijima 1962 [Iij62])	image-driven, anisotropic $\sum_{i=1}^2 \nabla u_i^\top D(\nabla f) \nabla u_i$ (Nagel 1983 [Nag83])
nonlinear isotropic $\partial_t u_i = \operatorname{div} \left(\Psi'(\sum_j \nabla u_j ^2) \nabla u_i \right)$ (Gerig et al. 1992 [GKKJ92])	flow-driven, isotropic $\Psi \left(\sum_{i=1}^2 \nabla u_i ^2 \right)$ (Schnörr 1994 [Sch94])
nonlinear anisotropic $\partial_t u_i = \operatorname{div} \left(\Psi'(\sum_j \nabla u_j \nabla u_j^\top) \nabla u_i \right)$ (Weickert 1994 [Wei94])	flow-driven, anisotropic $\operatorname{tr} \Psi \left(\sum_{i=1}^2 \nabla u_i \nabla u_i^\top \right)$ (Weickert/Schnörr 2001 [WS01])

the results for the optic flow magnitude. For better visibility, we also show a detail of the flow magnitude images in Figure 3.

As expected, one can observe that the homogeneous regularization of Horn and Schunck creates very smooth flow fields. It is, however, unsuited to respect any flow discontinuities.

Isotropic image-driven reduces smoothing at all image edges. This may create an oversegmentation of the flow fields, as can be seen from the flow artifacts resulting from the texture of the marbled floor. This oversegmentation influences in particular the flow magnitude, while the flow direction appears to be more stable.

Anisotropic image-driven regularization permits smoothing along image edges. This leads to a more homogeneous flow field than the one from isotropic image-driven smoothing. Larger structures of the marble texture, however, are still visible in this case as well.

Flow-driven models are performing better here. The marble texture, which corresponds to image discontinuities but not to flow discontinuities, does hardly perturb the flow field. Figure 3 shows that, similar to the image-driven case, anisotropic regularization is less affected by these texture artifacts than isotropic smoothing, although the differences are a bit smaller. This shows that anisotropic flow-driven regularization is an interesting technique for optic flow problems where flow discontinuities are important and highly textured image structures are present.

3. Matrix-Valued Filtering

In this section we extend diffusion and regularization methods to fields of matrix-valued data. After giving a motivation of the practical importance of such data sets, we present our models. They also comprise novel anisotropic techniques where a joint diffusion tensor is used instead of a scalar-valued diffusivity. Afterwards we argue that these models are well-suited for the practically important smoothing of positive semidefinite matrix fields, since they maintain the property of positive semidefiniteness without additional projection steps. Finally we study an example where a novel anisotropic nonlinear variant of the structure tensor is constructed and its superiority over linear and nonlinear isotropic structure tensors is illustrated.

3.1. Motivation. The need for smoothing methods for positive semidefinite matrix fields is rapidly growing in the image processing and computer vision community. Let us illustrate this by two examples.

- (a) Matrix-valued data fields with positive semidefinite matrices arise for example in all imaging applications where the so-called *structure tensor* is used. The structure tensor of some scalar-valued image v is given by $K_\rho * (\nabla v \nabla v^\top)$ where K_ρ denotes a Gaussian with standard deviation ρ . This Gaussian convolution averages orientation over same scale of order ρ . A principal component analysis of the structure tensor gives information that is highly useful for corner detection [FG87], texture analysis [RS91], optic flow computation [BGW91], and even for designing better adaptive numerical algorithms [GMS98, Tho99].

The structure tensor, however, uses Gaussian convolution of each matrix channel. This is equivalent to linear diffusion filtering with a constant diffusivity. Thus, the question arises whether one can obtain better results by replacing the matrix-valued linear diffusion process by matrix-valued nonlinear diffusion or regularization methods. In this case one would expect to have a better preservation of discontinuities.

- (b) Another application consists of *diffusion tensor magnetic resonance imaging (DT-MRI)*, a recent medical image acquisition technique that measures the diffusion characteristics of water molecules in tissues. The resulting diffusion tensor field is a positive semidefinite matrix field that provides valuable information for brain connectivity studies as well as for multiple sclerosis or stroke diagnosis [PJB⁺96].



FIGURE 2. (a) TOP LEFT: Frame 16 of the marbled block sequence (512×512 pixels). (b) TOP RIGHT: Optic flow magnitude between Frame 16 and 17 for homogeneous regularization. (c) MIDDLE LEFT: Result for image-driven isotropic regularization (d) MIDDLE RIGHT: Image-driven anisotropic regularization. (e) BOTTOM LEFT: Flow-driven isotropic regularization (f) BOTTOM RIGHT: Flow-driven anisotropic regularization. From [WS01].

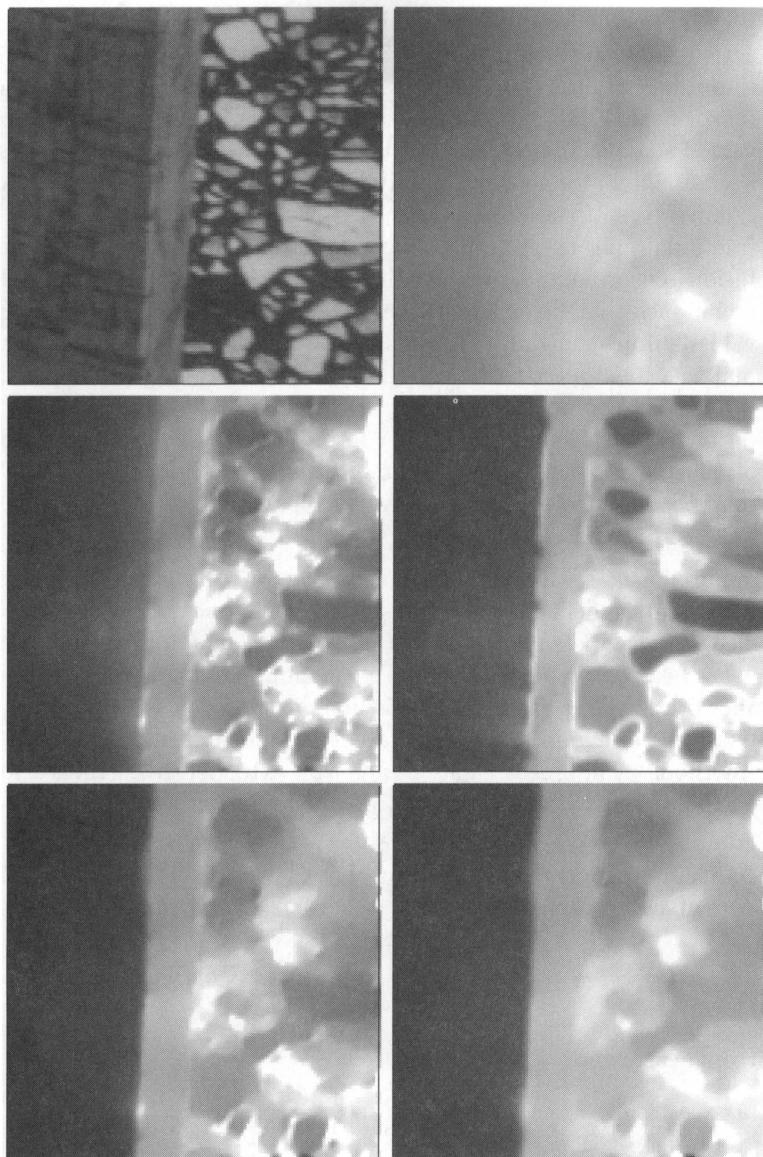


FIGURE 3. (a) TOP LEFT: Detail from the lower right part of Frame 16 (128×128 pixels). (b) TOP RIGHT: Optic flow magnitude for homogeneous regularization. (c) MIDDLE LEFT: Image-driven isotropic regularization (d) MIDDLE RIGHT: Image-driven anisotropic regularization. (e) BOTTOM LEFT: Flow-driven isotropic regularization (f) BOTTOM RIGHT: Flow-driven anisotropic regularization. For better visibility, the grey values of the optic flow results been transformed by a gamma correction with $\gamma = 0.4$. From [WS01].

The search for good smoothing techniques for DT-MRI became a very active research field in the last 3 years. Some authors suggest to perform smoothing of directional images that are used for computing the diffusion tensor field [KBFS00, PSS⁺00, VCR⁺01]. This comes down to scalar-valued smoothing processes. Other scalar- or vector-valued processes have been applied by smoothing derived expressions such as the eigenvalues and eigenvectors of the diffusion tensor [PMF⁺98, CAA01, TD01b] or its fractional anisotropy [PSS⁺00]. Methods that work directly on the diffusion tensor components use linear [WMK⁺99] or nonlinear [HPP01] techniques that filter all channels *independently*, thus performing scalar-valued filtering again. A nonlinear regularization method giving true matrix-valued filtering by coupling the channels via a common diffusivity is due to Tschumperlé and Deriche [TD01b]. They have also included additional projection steps in order to preserve the positive semidefiniteness of the matrix field. Since their diffusivity is scalar-valued, the method may be classified as isotropic. Anisotropic matrix-valued techniques have not been considered so far.

These examples illustrate that there is a clear need for a diffusion and regularization framework for matrix fields. Ideally, it should be compatible with the preceding vector-valued framework and it should take into account matrix-specific requirements such as the preservation of positive semidefiniteness. Below we shall describe such a framework for filtering matrix fields in an isotropic or anisotropic way. Since the linear isotropic and anisotropic cases are less important in practice, we focus on nonlinear methods. The linear strategies can be extended from the vector-valued situation in the same way as is described in the nonlinear setting.

3.2. Matrix-Valued Filter Design. Let us consider some matrix field $F(x) = (f_{kl}(x))$. A regularized version $U(x, \alpha) = (u_{kl}(x, \alpha))$ can be obtained by minimizing

$$(3.1) \quad E_{IM}(U) = \frac{1}{2} \int_{\Omega} \left(\|F - U\|^2 + \alpha \Psi \left(\operatorname{tr} \sum_{k,l} \nabla u_{kl} \nabla u_{kl}^\top \right) \right) dx dy$$

in the isotropic case, and

$$(3.2) \quad E_{AM}(U) = \frac{1}{2} \int_{\Omega} \left(\|F - U\|^2 + \alpha \operatorname{tr} \Psi \left(\sum_{k,l} \nabla u_{kl} \nabla u_{kl}^\top \right) \right) dx dy$$

in the anisotropic case. We assume that the penalizer Ψ satisfies the same conditions that we imposed in the vector-valued context. As matrix norm, we use the rotationally invariant Frobenius norm

$$(3.3) \quad \|F\| := \left(\sum_{k,l} f_{kl}^2 \right)^{1/2}.$$

It guarantees that both energy functionals are rotationally invariant.

Note the large structural similarities between the isotropic and the anisotropic functional: only the order of the penalizer and the trace operator is exchanged. One may also write the isotropic regularizer in a slightly simpler form as $\Psi(\sum_{k,l} |\nabla u_{k,l}|^2)$. The Euler–Lagrange equations to the isotropic functional (3.1) and its anisotropic

counterpart (3.2) are given by

$$(3.4) \quad \frac{u_{ij} - f_{ij}}{\alpha} = \operatorname{div} \left(\Psi' \left(\sum_{k,l} \nabla u_{kl}^\top \nabla u_{kl} \right) \nabla u_{ij} \right) \quad \forall i,j,$$

$$(3.5) \quad \frac{u_{ij} - f_{ij}}{\alpha} = \operatorname{div} \left(\Psi' \left(\sum_{k,l} \nabla u_{kl} \nabla u_{kl}^\top \right) \nabla u_{ij} \right) \quad \forall i,j.$$

These systems of elliptic PDEs may be regarded as implicit time discretizations of the isotropic resp. anisotropic matrix-valued diffusion processes

$$(3.6) \quad \partial_t u_{ij} = \operatorname{div} \left(\Psi' \left(\sum_{k,l} \nabla u_{kl}^\top \nabla u_{kl} \right) \nabla u_{ij} \right) \quad \forall i,j,$$

$$(3.7) \quad \partial_t u_{ij} = \operatorname{div} \left(\Psi' \left(\sum_{k,l} \nabla u_{kl} \nabla u_{kl}^\top \right) \nabla u_{ij} \right) \quad \forall i,j$$

with initial condition

$$(3.8) \quad u_{ij}(x, 0) = f_{ij}(x) \quad \forall i,j$$

and time step size α . The isotropic diffusivity may be simplified to $\Psi'(\sum_{k,l} |\nabla u_{kl}|^2)$.

As in the vector-valued case, one may also use diffusivities $\Psi'(s^2)$ with non-monotone flux functions $\Psi'(s^2)s$, if in their argument u_{kl} is replaced by a Gaussian-smoothed variant $K_\sigma * u_{kl}$.

While the preceding isotropic regularization or diffusion methods are also part of the models of Tschumperlé and Deriche [TD01b], their anisotropic counterparts are studied for the first time in the present paper. In Subsection 3.4 we shall see that anisotropy may lead to significantly improved results.

Our initial motivation for considering matrix-valued smoothing processes stems from the quadratic case with symmetric matrices that are positive semidefinite. However, it should be noted that our matrix-valued models are not restricted to quadratic matrices: They may be applied to the smoothing of arbitrary $n \times m$ matrix fields. In particular, we may regard an n -dimensional vector as an $n \times 1$ matrix. In this case it follows directly that the matrix-valued diffusion and regularization models comprise our vector-valued ones that we discussed before.

Another point that is worth mentioning is that the preceding matrix-valued smoothing processes use diffusivities or diffusion tensors that are identical for all matrix channels. As in the vector-valued case this ensures that the filtering behavior at edges remains synchronized. Moreover, it has an additional interesting consequence that shall be discussed next.

3.3. Preservation of Positive Semidefiniteness. Let us go back to quadratic matrix fields that are positive semidefinite. In this case a natural requirement for a practically useful smoothing method is that it should not destroy the positive semidefiniteness of the initial matrix field. By construction of our continuous filters it is obvious that symmetric matrix fields remain symmetric under filtering. In order to understand why the nonnegativity of the eigenvalues is preserved as well, it is helpful to consider a finite difference setting.

The sketch of the proof for the discrete diffusion case is as follows. With a slight abuse of notation, let f_{ij} be a discretization of the (i,j) component of the vector field $F(x)$. We may regard f_{ij} as a vector whose components describe the grey values of the (i,j) component at all pixel locations. Let us consider some

suitably small time step size τ and let $u^k = (u_{ij})^k$ represent in a similar way some discretization of the matrix field $U(x, t)$ at time level $t = k\tau$. In [Wei98] it is shown that there exist finite difference schemes for diffusion filtering such that $u_{i,j}^{k+1}$ may be obtained from $u_{i,j}^k$ by a matrix–vector multiplication:

$$(3.9) \quad u_{ij}^0 = f_{ij} \quad \forall i, j$$

$$(3.10) \quad u_{ij}^{k+1} = A(u_{1,1}^k, \dots, u_{n,n}^k) u_{ij}^k \quad \forall i, j, \quad \forall k \geq 0$$

where the matrix A has unit row sums and all entries are nonnegative. Since we have a common diffusivity or diffusion tensor for all channels, it follows that A is identical for all channels:

$$(3.11) \quad A = A(u_{1,1}^k, \dots, u_{n,n}^k).$$

Thus, the discrete iteration scheme performs convex combinations of the matrices from the iteration level k in order to obtain the result at level $k + 1$. Since convex combinations of positive semidefinite matrices are positive semidefinite again (see e.g. the proof of Proposition 2 in [WS01]), it follows that the positive semidefiniteness of the initial matrix field is preserved for all iteration levels.

It should be noted that this reasoning depends strongly on the use of a joint diffusivity or diffusion tensor for all matrix channels. Thus, models that apply different diffusivities in each channel may not preserve positive semidefinite matrix fields unless additional projection steps are introduced (cf. [TD01b]).

3.4. Example: Nonlinear Structure Tensors. Let us now illustrate the usefulness of nonlinear smoothing strategies for matrix-valued data sets by investigating nonlinear versions of the structure tensor that we mentioned in Subsection 3.1.

Given some image v , we consider the tensor product

$$(3.12) \quad F := (f_{ij}) := \nabla v \nabla v^\top$$

The linear structure tensor computes the convolution $K_\rho * (\nabla v \nabla v^\top)$, where K_ρ denotes a Gaussian with standard deviation ρ . This is equivalent to the linear matrix valued diffusion process

$$(3.13) \quad \partial_t u_{ij} = \Delta u_{ij} \quad \forall i, j$$

$$(3.14) \quad u_{ij}(x, 0) = f_{ij}(x) \quad \forall i, j$$

with stopping time $T = \frac{1}{2} \rho^2$.

Figure 4(a) shows a synthetic test image where we have two regions with homogeneous orientation transitions. These two regions are separated by an orientation discontinuity. An ideal orientation measure would average the orientation information within each region without affecting the orientation discontinuity.

In Figure 4(b) we can see all four components of the structure tensor field when the preceding linear diffusion process is applied. This process blurs each of the tensor components. While this is desirable within the same region, it also blurs the matrix components at orientation discontinuities. If one wants to avoid this shortcoming, one has to use adaptive filters.

A first attempt along these lines is shown in Figure 4(c). It depicts the structure tensor field when the linear diffusion process (3.13) has been replaced by the isotropic nonlinear evolution (3.6) with a diffusivity of type (2.7). We observe that this process does respect orientation discontinuities. On the other hand, it may

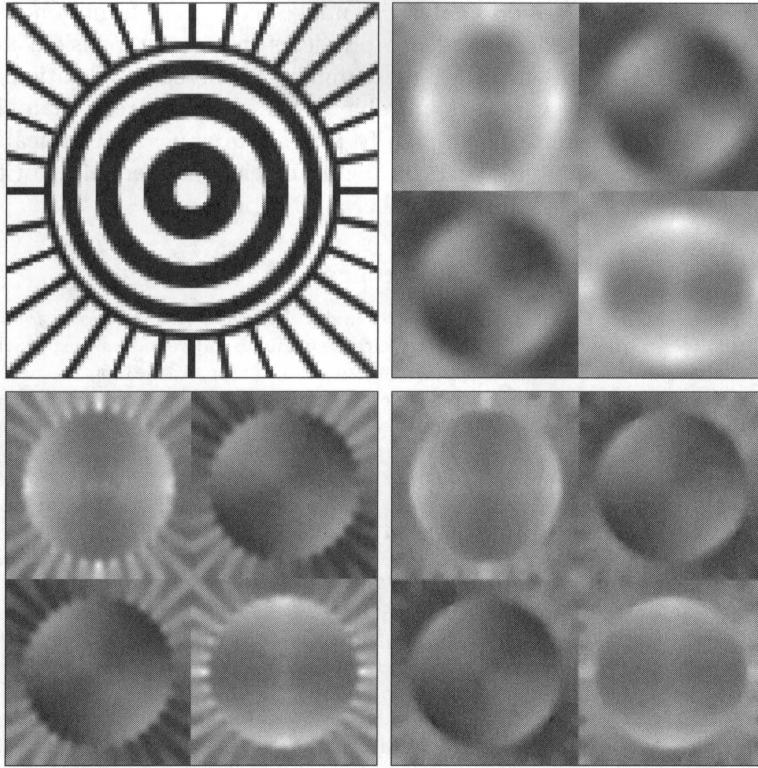


FIGURE 4. (a) TOP LEFT: Synthetic test image. (b) TOP RIGHT: Structure tensor components with matrix-valued linear diffusion filtering. (c) BOTTOM LEFT: Ditto with isotropic nonlinear diffusion filtering. (d) BOTTOM RIGHT: Ditto with anisotropic nonlinear diffusion filtering.

even be too conservative: at discontinuities, diffusion is stopped in all directions. This may lead to problems in noisy or textured images, where spatial smoothing of the tensor components does not take place.

Anisotropic matrix-valued diffusion filtering on the basis of equations (3.7) and (2.7) is illustrated in Figure 4(d). At discontinuities, only diffusion across the discontinuity is inhibited, while diffusion along the discontinuity is still maintained. As one would expect, this leads to the desired matrix averaging without blurring across orientation discontinuities. This makes the nonlinear anisotropic structure tensor an interesting candidate for a number of applications where the linear structure tensor has limited performance. We are currently trying to identify such situations in order to quantify the benefits of nonlinear anisotropic structure tensors.

4. Summary

In this paper we have given a unified description of diffusion and regularization methods for vector- and matrix-valued data sets. These ideas have been illustrated by applying them to variational motion analysis and by deriving novel nonlinear

structure tensors. Since motion analysis in image sequences is only one representative of a large class of correspondence problems in computer vision, and since the structure tensor is present in a large number of different applications, we are optimistic that this framework is applicable to many more areas than those described here. In our future work we plan to present a detailed theoretical analysis of our models, to carry out research on highly efficient numerical algorithms for these approaches, and to investigate further application areas.

References

- [AELS99] L. Alvarez, J. Esclarín, M. Lefébure, and J. Sánchez, *A PDE model for computing the optical flow*, Proc. XVI Congreso de Ecuaciones Diferenciales y Aplicaciones (Las Palmas de Gran Canaria, Spain), September 1999, pp. 1349–1356.
- [BC98] P. Blomgren and T. F. Chan, *Color TV: total variation methods for restoration of vector valued images*, IEEE Transactions on Image Processing **7** (1998), no. 3, 304–309.
- [BGW91] J. Bigün, G. H. Granlund, and J. Wiklund, *Multidimensional orientation estimation with applications to texture analysis and optical flow*, IEEE Transactions on Pattern Analysis and Machine Intelligence **13** (1991), no. 8, 775–790.
- [Bre73] H. Brezis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North Holland, Amsterdam, 1973.
- [CAA01] O. Coulon, D. C. Alexander, and S. A. Arridge, *A regularization scheme for diffusion tensor magnetic resonance images*, Information Processing in Medical Imaging – IPMI 2001 (M. F. Insana and R. M. Leahy, eds.), Lecture Notes in Computer Science, vol. 2082, Springer, Berlin, 2001, pp. 92–105.
- [CLMC92] F. Catté, P.-L. Lions, J.-M. Morel, and T. Coll, *Image selective smoothing and edge detection by nonlinear diffusion*, SIAM Journal on Numerical Analysis **32** (1992), 1895–1909.
- [Di 86] S. Di Zenzo, *A note on the gradient of a multi-image*, Computer Vision, Graphics and Image Processing **33** (1986), 116–125.
- [FG87] W. Förstner and E. Gülich, *A fast operator for detection and precise location of distinct points, corners and centres of circular features*, Proc. ISPRS Intercommission Conference on Fast Processing of Photogrammetric Data (Interlaken, Switzerland), June 1987, pp. 281–305.
- [Fri92] D. S. Fritsch, *A medial description of greyscale image structure by gradient-limited diffusion*, Visualization in Biomedical Computing '92 (R. A. Robb, ed.), Proceedings of SPIE, vol. 1808, SPIE Press, Bellingham, 1992, pp. 105–117.
- [GKKJ92] G. Gerig, O. Kübler, R. Kikinis, and F. A. Jolesz, *Nonlinear anisotropic filtering of MRI data*, IEEE Transactions on Medical Imaging **11** (1992), 221–232.
- [GMS98] T. Grahs, A. Meister, and T. Sonar, *Image processing for numerical approximations of conservation laws: nonlinear anisotropic artificial dissipation*, Tech. Report F8, Institute for Applied Mathematics, University of Hamburg, Germany, December 1998.
- [HPP01] K. Hahn, S. Pigarin, and B. Pütz, *Edge preserving regularization and tracking for diffusion tensor imaging*, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2001 (W. J. Niessen and M. A. Viergever, eds.), Lecture Notes in Computer Science, vol. 2208, Springer, Berlin, 2001, pp. 195–203.
- [HS81] B. Horn and B. Schunck, *Determining optical flow*, Artificial Intelligence **17** (1981), 185–203.
- [Iij59] T. Iijima, *Basic theory of pattern observation*, Papers of Technical Group on Automata and Automatic Control, IECE, Japan, December 1959, In Japanese.
- [Iij62] T. Iijima, *Observation theory of two-dimensional visual patterns*, Papers of Technical Group on Automata and Automatic Control, IECE, Japan, October 1962, In Japanese.
- [KBFS00] S. Keeling, R. Bammer, F. Fazekas, and R. Stollberger, *Total variation denoising for improved diffusion tensor calculation*, Proc. Eighth Scientific Meeting and Exhibition of the International Society for Magnetic Resonance in Medicine (Denver, CO), vol. 8, April 2000, p. 783.

- [KMS00] R. Kimmel, R. Malladi, and N. Sochen, *Images as embedded maps and minimal surfaces: movies, color, texture, and volumetric medical images*, International Journal of Computer Vision **39** (2000), no. 2, 111–129.
- [Nag83] H.-H. Nagel, *Constraints for the estimation of displacement vector fields from image sequences*, Proc. Eighth International Joint Conference on Artificial Intelligence (Karlsruhe, West Germany), vol. 2, August 1983, pp. 945–951.
- [NS98] M. Z. Nashed and O. Scherzer, *Least squares and bounded variation regularization with nondifferentiable functionals*, Numerical Functional Analysis and Optimization **19** (1998), 873–901.
- [ON95] M. Otte and H.-H. Nagel, *Estimation of optical flow based on higher-order spatiotemporal derivatives in interlaced and non-interlaced image sequences*, Artificial Intelligence **78** (1995), 5–43.
- [PJB⁺96] C. Pierpaoli, P. Jezzard, P. J. Bassett, A. Barnett, and G. Di Chiro, *Diffusion tensor MR imaging of the human brain*, Radiology **201** (1996), no. 3, 637–648.
- [PMF⁺98] C. Poupon, J.-F. Mangin, V. Frouin, J. Régis, F. Poupon, M. Pachot-Clouard, D. Le Bihan, and I. Bloch, *Regularization of MR diffusion tensor maps for tracking brain white matter bundles*, Medical Image Computing and Computer-Assisted Intervention – MICCAI 1998 (W. M. Wells, A. Colchester, and S. Delp, eds.), Lecture Notes in Computer Science, vol. 1496, Springer, Berlin, 1998, pp. 489–498.
- [PSS⁺00] G. J. M. Parker, J. A. Schnabel, M. R. Symms, D. J. Werring, and G. J. Barker, *Non-linear smoothing for reduction of systematic and random errors in diffusion tensor imaging*, Journal of Magnetic Resonance Imaging **11** (2000), 702–710.
- [RS91] A. R. Rao and B. G. Schunck, *Computing oriented texture fields*, CVGIP: Graphical Models and Image Processing **53** (1991), 157–185.
- [RSW00] E. Radmoser, O. Scherzer, and J. Weickert, *Scale-space properties of nonstationary iterative regularization methods*, Journal of Visual Communication and Image Representation **11** (2000), no. 2, 96–114.
- [Sap01] G. Sapiro, *Geometric partial differential equations and image analysis*, Cambridge University Press, Cambridge, UK, 2001.
- [Sch94] C. Schnörr, *Segmentation of visual motion by minimizing convex non-quadratic functionals*, Proc. Twelfth International Conference on Pattern Recognition (Jerusalem, Israel), vol. A, IEEE Computer Society Press, October 1994, pp. 661–663.
- [SW00] O. Scherzer and J. Weickert, *Relations between regularization and diffusion filtering*, Journal of Mathematical Imaging and Vision **12** (2000), no. 1, 43–63.
- [TD01a] D. Tschumperlé and R. Deriche, *Constrained and unconstrained PDE's for vector image restoration*, Proc. Twelfth Scandinavian Conference on Image Analysis (Bergen, Norway), June 2001.
- [TD01b] D. Tschumperlé and R. Deriche, *Diffusion tensor regularization with constraints preservation*, Proc. 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Kauai, HI), vol. 1, IEEE Computer Society Press, December 2001, pp. 948–953.
- [Tho99] I. Thomas, *Anisotropic adaptation and structure detection*, Tech. Report F11, Institute for Applied Mathematics, University of Hamburg, Germany, August 1999.
- [VCR⁺01] B. Vemuri, Y. Chen, M. Rao, T. McGraw, Z. Wang, and T. Marcic, *Fiber tract mapping from diffusion tensor MRI*, Proc. First IEEE Workshop on Variational and Level Set Methods in Computer Vision (Vancouver, Canada), IEEE Computer Society Press, July 2001, pp. 73–80.
- [Wei94] J. Weickert, *Scale-space properties of nonlinear diffusion filtering with a diffusion tensor*, Tech. Report 110, Laboratory of Technomathematics, University of Kaiserslautern, Germany, October 1994.
- [Wei98] J. Weickert, *Anisotropic diffusion in image processing*, Teubner, Stuttgart, 1998.
- [Wei99] J. Weickert, *Coherence-enhancing diffusion of colour images*, Image and Vision Computing **17** (1999), no. 3–4, 199–210.
- [WMK⁺99] C.-F. Westin, S. E. Maier, B. Khidhir, P. Everett, F. A. Jolesz, and R. Kikinis, *Image processing for diffusion tensor magnetic resonance imaging*, Medical Image Computing and Computer-Assisted Intervention – MICCAI 1999 (C. Taylor and A. Colchester, eds.), Lecture Notes in Computer Science, vol. 1679, Springer, Berlin, 1999, pp. 441–452.

- [WS01] J. Weickert and C. Schnörr, *A theoretical framework for convex regularizers in PDE-based computation of image motion*, International Journal of Computer Vision **45** (2001), no. 3, 245–264.

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE, SAARLAND UNIVERSITY, BUILDING 27.1,
P. O. Box 15 11 50, 66041 SAARBRÜCKEN, GERMANY

E-mail address: `weickert@mia.uni-saarland.de`, `brox@mia.uni-saarland.de`

A Numerically Robust Hybrid Steepest Descent Method for The Convexly Constrained Generalized Inverse Problems

Isao Yamada, Nobuhiko Ogura, and Nobuyasu Shirakawa

ABSTRACT. The convexly constrained generalized pseudoinverse problem is formulated as a special example of the variational inequality problem over the fixed point sets of nonexpansive mappings in a real Hilbert space [Yamada (1999,2001)], of which the solution can be approximated successively by the use of the hybrid steepest descent method [Yamada, Ogura, Yamashita and Sakaniwa (1996,1998), Deutsch and Yamada (1998), Yamada (1999,2000,2001), Ogura and Yamada (2002a,2002b)].

In the first part of this paper, we demonstrate that a variety of algorithmic solutions, to the convexly constrained generalized inverse problems, are derived in a unified manner based on the hybrid steepest descent method. In this process, it is revealed that several recent algorithms, specialized for the linear inverse problems, can also be characterized as simple realizations of the hybrid steepest descent method. In the second part of this paper, we show that, by the use of a slowly changing sequence of nonexpansive mappings having same fixed point sets, a variation of the hybrid steepest descent method is gifted with notable robustness to the numerical errors possibly unavoidable in the iterative computations. The present analysis fully extends the recent convergence analyses [Ogura and Yamada (2002a, 2002b)] of the hybrid steepest descent method as well as shows its sound applicability to the possibly ill-posed problems like the convexly constrained inverse problems.

1. Introduction

The *convexly constrained inverse problems* have attracted considerable attention not only because they cover core problems in the *signal/image processing* [4, 13, 40, 70, 75, 76, 82, 99, 104, 112, 113, 115–119, 135, 136], but also because their roots have been growing over interdisciplinary mathematical sciences including *nonlinear functional analysis, fixed point theory, optimization theory, approximation theory and numerical analysis* (See e.g., [7, 18, 26, 34, 38, 39, 53, 61, 64, 65, 69, 72, 84, 88, 92–94, 109, 121, 123, 124, 134, 138–142] and references therein).

2000 *Mathematics Subject Classification.* Primary 47H10,90C25; Secondary 47H09,90C30.

Key words and phrases. Convexly constrained inverse problem, Hybrid steepest descent method, Fixed point, Nonexpansive mapping, Variational inequality problem, Inverse problem.

This work is supported in part by the International Communications Foundations (ICF) Japan.

© 2002 American Mathematical Society

In particular, the following *convexly constrained generalized pseudoinverse problem* unifies the broad range of signal processing problems for example in (i) time or band-limited extrapolation with subspace constraints [63, 82, 101, 104], (ii) image reconstruction with positivity constraint [75, 115–118, 135, 136], and (iii) signal recovery (or synthesis) [4, 76, 99, 104, 112, 113] accompanied by constraints on amplitude, support, energy etc.

PROBLEM 1.1. (*K-constrained generalized pseudoinverse problem*) Let K be a nonempty closed convex set in a real Hilbert space \mathcal{H} equipped with an inner product $\langle \cdot, \cdot \rangle$ and its induced norm $\|\cdot\|$. Let $A : \mathcal{H} \rightarrow \mathcal{H}_o$ ($:= \mathbb{R}^m$) be a given bounded linear (experiment) operator and $\mathbf{b} = (b_1, \dots, b_m)^T \in \mathcal{H}_o$ a possibly perturbed measurement (we denote the norm defined in the real Hilbert space \mathcal{H}_o by $\|\cdot\|_o$). Suppose that

$$(1.1) \quad \mathcal{S} := \arg \inf_{x \in K} \|Ax - \mathbf{b}\|_o \neq \emptyset.$$

Then the problem is

$$(1.2) \quad \text{Find a point } x^* \in \arg \inf_{x \in \mathcal{S}} \Theta(x),$$

where $\Theta : \mathcal{H} \rightarrow \mathbb{R} \cup \{\infty\}$ is a convex function.

A generalized version of Problem 1.1 can be stated as:

PROBLEM 1.2. (*Generalized nonlinear convex inverse problem*) Let K be a nonempty closed convex set in a real Hilbert space \mathcal{H} . Let $\Psi : \mathcal{H} \rightarrow \mathbb{R} \cup \{\infty\}$ be a convex function satisfying

$$(1.3) \quad K_\Psi := \arg \inf_{x \in K} \Psi(x) \neq \emptyset.$$

Then for an additionally given convex function $\Theta : \mathcal{H} \rightarrow \mathbb{R} \cup \{\infty\}$, the problem is

$$(1.4) \quad \text{Find a point } x^* \in \arg \inf_{x \in K_\Psi} \Theta(x).$$

REMARK 1.3.

- (a) Problem 1.1 is an example of Problem 1.2. Note that Problem 1.2 covers a more general version of Problem 1.1, where (i) \mathcal{H}_o is a possibly infinite dimensional real Hilbert space, and (ii) $A : \mathcal{H} \rightarrow \mathcal{H}_o$ is a bounded linear operator. We will demonstrate in Proposition 2.16 and Remark 2.17 that the hybrid steepest descent method (Theorem 2.15) [133] can be applied soundly to a broad class of Problem 1.2.
- (b) The nonemptiness in (1.1) holds if and only if $\mathcal{P}_{\overline{A(K)}}(\mathbf{b}) \in A(K)$, where $\mathcal{P}_C : \mathcal{H}_o \rightarrow C$ assigns every point in \mathcal{H}_o to its unique nearest point in a nonempty closed convex set $C \subset \mathcal{H}_o$ [111].
- (c) The set K reflects *a priori* knowledge on the unknown estimandum or the data.
- (d) The cost function Θ in Problems 1.1 (or more generally in Problem 1.2) is reasonably defined for a proper point estimation. However the convexly constrained inverse problem in some applications has not necessarily been demanding for the cost function. Indeed, in such applications, the cost function is utilized just to realize any point estimation, and has been selected mainly based on its computational tractability [44]. Moreover, even

if the function Θ has a certain objective meaning, the load for the minimization of Θ over the set S is often highly expensive. The set-theoretic formulation [41] may lead to an alternative, rather simpler, formulation :

$$(1.5) \quad \text{Find a point } \hat{x} \in S = \arg \inf_{x \in K} \|Ax - b\|_o.$$

As will be seen later in Section 3.B (and C), the characterization of the closed convex set K , as the fixed point set of a nonexpansive mapping, realizes a strategy to reflect soundly *a priori* information into the set K .

In [60, 111], Problem 1.1 was considered for a special cost function $\Theta(x) = \frac{1}{2}\|x\|^2$. In this case, since the uniqueness of the solution of Problem 1.1 is automatically guaranteed under condition (1.1), by defining $\mathcal{D}(A_K^\dagger) := \{b \in \mathbb{R}^m \mid \overline{\mathcal{P}_{A(K)}(b)} \in A(K)\}$, a mapping $A_K^\dagger : \mathcal{D}(A_K^\dagger) \rightarrow K$ named *K-constrained pseudoinverse* is well-defined as

$$(1.6) \quad A_K^\dagger(b) := \arg \inf_{x \in S} \|x\|.$$

$\mathcal{D}(A_K^\dagger) = \mathbb{R}^m$ holds if K is bounded. In case $K = \mathcal{H}$, A_K^\dagger is reduced to the well-known *Moore-Penrose pseudoinverse* [88, 94].

Although, the mapping $A_K^\dagger : \mathcal{D}(A_K^\dagger) \rightarrow K$ is not necessarily continuous on $b \in \mathbb{R}^m$ for some closed convex set K (a simple example in \mathbb{R}^3 was given [110]) and thus ill-posed, in general, in the sense of Hadamard [72], several important cases as well were shown to guarantee the well-posedness of the mapping A_K^\dagger (See Example 1.4 below) [110, 111], which are surely maintaining the status of Problem 1.1 as one of the central inverse problems in wide range of the mathematical sciences and engineerings.

EXAMPLE 1.4.

- (a) A_K^\dagger is continuous on the relative interior of $A(K)$ for any closed convex set $K \subset \mathcal{H}$.
- (b) A_K^\dagger is uniformly continuous if K is a closed subspace or more generally a closed linear variety.
- (c) A_K^\dagger is uniformly continuous if every point in the boundary of K is a regular point.

Several algorithms, developed in the last decade, generate the sequences converging to $A_K^\dagger(b)$, provided that every necessary computation is ideally realized during the whole processes of the formulae. In [60, 105, 111], it is shown that a variation of the *projected Landweber iteration*:

$$(1.7) \quad u_{n+1} := P_K(\lambda A^*b + \beta_n(I - \lambda A^*A)u_n)$$

(for $\lambda \in (0, 2\|A\|^{-2})$ and $\beta_n := (1 + n^{-c})^{-1}$ with $\beta_0 = 0$ and $0 < c < 1$) generates a sequence $(u_n)_{n \geq 0}$ that converges strongly to $A_K^\dagger(b)$ for $b \in \mathcal{D}(A_K^\dagger)$, where P_K is the metric projection onto K . By applying a special nonexpansive mapping $T : \mathcal{H} \rightarrow \mathcal{H}$, whose fixed point set is K_Ψ (As will be discussed in Proposition 2.5, such a nonexpansive mapping can be constructed, in many situations, by the use of P_K and the derivative of Ψ), to the hybrid steepest descent method [54, 96, 97, 129–133]:

$$(1.8) \quad u_{n+1} := T(u_n) - \lambda_{n+1}\Theta'(T(u_n)),$$

where $(\lambda_n)_{n \geq 1}$ is a slowly decreasing nonnegative sequence satisfying certain conditions, it is possible to approximate successively the solution of Problem 1.2 (This fact was observed in [131, 133] for Problem 1.1). In Remark 2.17, we will reproduce the formula (1.7) as a special example of an alternative expression (2.8) of the hybrid steepest descent method (1.8).

Moreover, even if a closed form expression of P_K is not known, it is also possible to resolve (1.5) essentially by applying the convex function $\phi(x) := \frac{1}{2} \|Ax - b\|_o^2$ and any nonexpansive mapping T , whose fixed point set is K , to the hybrid steepest descent method as

$$u_{n+1} := T(u_n) - \lambda_{n+1} \phi'(T(u_n)).$$

These facts suggest the remarkable applicability of the hybrid steepest descent method to the convexly constrained inverse problems.

A still remaining problem of the above approaches must be that the hybrid steepest descent method requires a closed form expression of the nonexpansive mapping T . However, as remarked clearly in [44, 49], certain nonexpansive mappings, necessary for some applications, must be constructed in iterative ways, and thus we have to utilize reasonable approximations of such mappings somehow. This situation and the sensitive nature of the convexly constrained inverse problem (for example the ill-posedness as observed in Problem 1.1) induce a natural question: Do the algorithms using such approximations still have mathematical guarantees of convergences to the solution of the original inverse problem?

In the later part of this paper, we study the numerical robustness of a variation of the hybrid steepest descent method [54, 96, 97, 129–133] for the variational inequality problem over the fixed point set of a nonexpansive mapping. The present analysis is motivated by the above question and the recent error analysis [49], given for the parallel projection method [42], which works as the error analysis for the *Krasnoselski/Mann type iteration* [9, 43, 56, 90, 91, 114]:

$$(1.9) \quad x_{n+1} := (1 - \alpha_n)x_n + \alpha_n T(x_n)$$

aiming at the weak convergence of $(x_n)_{n \geq 0}$ to a fixed point of a nonexpansive mapping $T : \mathcal{H} \rightarrow \mathcal{H}$.

The main results of the present paper show that, by the use of a slowly changing sequence of mappings $T_{(n)} := (1 - t_{n+1})I + t_{n+1}T$ (I is identity, hence $\text{Fix}(T_{(n)}) = \text{Fix}(T)$ for $t_{n+1} \neq 0$), a variation of the hybrid steepest descent method:

$$(1.10) \quad u_{n+1} := T_{(n)}(u_n) - \lambda_{n+1} \Theta'(T_{(n)}(u_n))$$

is gifted with notable numerical robustness.

The rest of this paper is divided into four sections. For the readers' convenience, the next section contains preliminaries on the fixed point, the nonexpansive mapping, the convex projection, as well as the asymptotically shrinking mapping which was recently introduced by the first two authors [96]. This section also includes a brief introduction to the variational inequality problem and a summary of the recent progress on the generalizations of the hybrid steepest descent method for the variational inequality problem over the fixed point set of a nonexpansive mapping in a real Hilbert space.

The third section discusses several useful formulations of the convexly inverse problems as the variational inequality problems over the fixed point set of a nonexpansive mapping. Thanks to these formulations, the hybrid steepest descent

method realizes various approaches, in a unified manner, to the convexly constrained inverse problems.

In the fourth section, we present an error analysis of the variation (1.10) of the hybrid steepest descent method. The main theorem fully extends the recent convergence analyses [96, 97] for the prototype of the method as well as proves that the variation (1.10) approximates successively the solution of the original variational inequality problem even in the case where the numerical error sequence, caused at each iterative computation, is not necessarily *absolutely summable*. Lastly in the final section, we conclude the paper with some remarks.

2. Preliminaries

A. Fixed points, Nonexpansive mappings, Convex projections and Asymptotically shrinking mappings. A *fixed point* of a mapping $T : \mathcal{H} \rightarrow \mathcal{H}$ is a point $x \in \mathcal{H}$ such that $T(x) = x$. $\text{Fix}(T) := \{x \in \mathcal{H} \mid T(x) = x\}$ denotes the fixed point set of T . A mapping $T : \mathcal{H} \rightarrow \mathcal{H}$ is called κ -Lipschitzian (or κ -Lipschitz continuous) over $S \subset \mathcal{H}$ if there exists $\kappa > 0$ such that

$$(2.1) \quad \|T(x) - T(y)\| \leq \kappa \|x - y\| \text{ for all } x, y \in S.$$

In particular, a mapping $T : \mathcal{H} \rightarrow \mathcal{H}$ is called (i) *strictly contractive* if $\|T(x) - T(y)\| \leq \kappa \|x - y\|$ for some $\kappa \in (0, 1)$ and all $x, y \in \mathcal{H}$ [The *Banach-Picard fixed point theorem* guarantees the unique existence of the fixed point, say $x_* \in \text{Fix}(T)$, of T and the strong convergence of $(T^n(x_0))_{n \geq 0}$ to x_* for any $x_0 \in \mathcal{H}$.]; (ii) *nonexpansive* if $\|T(x) - T(y)\| \leq \|x - y\|$ for all $x, y \in \mathcal{H}$; (iii) *firmly nonexpansive* if $\|T(x) - T(y)\|^2 \leq \langle x - y, T(x) - T(y) \rangle$ for all $x, y \in \mathcal{H}$; (iv) *averaged* (or α -averaged) if there exists $\alpha \in [0, 1]$ and a nonexpansive mapping $\mathcal{N} : \mathcal{H} \rightarrow \mathcal{H}$ such that $T := (1 - \alpha)I + \alpha\mathcal{N}$. (Note: This definition is slightly different from the one in the sense of [5] where $\alpha = 0$ is excluded to ensure $\text{Fix}(T) = \text{Fix}(\mathcal{N})$. A firmly nonexpansive mapping is alternatively characterized as $\frac{1}{2}$ -averaged mapping.); and (v) *attracting nonexpansive* if T is nonexpansive with $\text{Fix}(T) \neq \emptyset$ and $\|T(x) - f\| < \|x - f\|$ for all $f \in \text{Fix}(T)$ and all $x \notin \text{Fix}(T)$ (We regard the identity I as a special attracting nonexpansive mapping due to $\text{Fix}(I) = \mathcal{H}$). If $T : \mathcal{H} \rightarrow \mathcal{H}$ is averaged with $\text{Fix}(T) \neq \emptyset$, then T is attracting (See for example [9, Lemma 2.4]).

Recall that a nonempty set $C \subset \mathcal{H}$ is called *convex* if $x, y \in C$ and $t \in [0, 1]$ imply $tx + (1 - t)y \in C$. Given a nonempty closed convex set $C \subset \mathcal{H}$, the mapping that assigns every point in \mathcal{H} to its unique nearest point in C is called the *metric projection* or *convex projection* onto C and is denoted by P_C ; i.e., $\|x - P_C(x)\| = d(x, C)$, where $d(x, C) := \inf_{y \in C} \|x - y\|$. The metric projection P_C is characterized by the relation:

$$(2.2) \quad x^* = P_C(x) \Leftrightarrow x^* \in C \text{ satisfies } \langle x - x^*, y - x^* \rangle \leq 0 \text{ for all } y \in C,$$

and therefore P_C is firmly nonexpansive with $\text{Fix}(P_C) = C$. It is also well-known that the functional $\frac{1}{2}d(\cdot, C)^2 : \mathcal{H} \rightarrow \mathbb{R}$ is Gâteaux differentiable over \mathcal{H} with derivative $I - P_C$ (See for example [3, Theorem 5.2]. Indeed in this case, by the continuity of $I - P_C$, it is the Fréchet derivative of $\frac{1}{2}d(\cdot, C)^2$). In particular, if \mathcal{M} is a closed subspace in \mathcal{H} , $P_{\mathcal{M}}$ is linear and $x - P_{\mathcal{M}}(x) \in \mathcal{M}^\perp$ for all $x \in \mathcal{H}$.

There are many nonexpansive mappings simple enough in the sense that their closed form expressions are known, which implies that such a mapping can be computed within a finite number of arithmetic operations. Typical example of such nonexpansive mappings is the metric projection P_C onto a simple closed convex set

$C \subset \mathcal{H}$ such as linear variety, closed ball, closed cone or closed polytope [128] etc. By using several properties shown, in Fact 2.1 collecting particularly useful results [9, 11, 22, 43, 64, 65, 96, 107, 119, 130, 137], it is possible to construct a novel nonexpansive mapping T , of which the fixed point set $\text{Fix}(T)$ becomes a feasible set of many convexly constrained inverse problems.

FACT 2.1. (Selected properties of nonexpansive mappings)

- (a) If a nonexpansive mapping $T : \mathcal{H} \rightarrow \mathcal{H}$ has at least one fixed point, $\text{Fix}(T) \subset \mathcal{H}$ is closed convex and expressed as

$$\text{Fix}(T) = \bigcap_{y \in \mathcal{H}} \{x \in \mathcal{H} \mid 2\langle y - T(y), x \rangle \leq \|y\|^2 - \|T(y)\|^2\}.$$

- (b) For nonexpansive mappings $T_i : \mathcal{H} \rightarrow \mathcal{H}$ ($i = 1, 2, \dots, m$), both $\sum_{i=1}^m w_i T_i$ and $T_m T_{m-1} \cdots T_1$ are also nonexpansive, where $(w_i)_{i=1}^m \subset [0, 1]$ and $\sum_{i=1}^m w_i = 1$.
- (c) For nonexpansive mappings $T_i : \mathcal{H} \rightarrow \mathcal{H}$ ($i = 1, 2, \dots, m$) satisfying $\bigcap_{i=1}^m \text{Fix}(T_i) \neq \emptyset$, it follows $\text{Fix}(\sum_{i=1}^m w_i T_i) = \bigcap_{i=1}^m \text{Fix}(T_i)$ where $w_i > 0$ and $\sum_{i=1}^m w_i = 1$.
- (d) $T : \mathcal{H} \rightarrow \mathcal{H}$ is firmly nonexpansive iff $2T - I$ is nonexpansive. This fact implies that (i) $I - T : \mathcal{H} \rightarrow \mathcal{H}$ is firmly nonexpansive iff $T : \mathcal{H} \rightarrow \mathcal{H}$ is firmly nonexpansive [because $2(I - T) - I = I - 2T$ is nonexpansive], and (ii) $(1 - \alpha)I + \alpha T$ is $\frac{\alpha}{2}$ -averaged and thus attracting nonexpansive for all $\alpha \in [0, 2]$ if $T : \mathcal{H} \rightarrow \mathcal{H}$ is firmly nonexpansive with $\text{Fix}(T) \neq \emptyset$. Moreover, for given firmly nonexpansive mappings $T_i : \mathcal{H} \rightarrow \mathcal{H}$ and $w_i \geq 0$ ($i = 1, 2, \dots, m$) satisfying $\sum_{k=1}^m w_k = 1$, $\sum_{i=1}^m w_i T_i$ is firmly nonexpansive because $2(\sum_{i=1}^m w_i T_i) - I = \sum_{i=1}^m w_i(2T_i - I)$ is nonexpansive.
- (e) Suppose that $T_i : \mathcal{H} \rightarrow \mathcal{H}$ ($i \in J \subset \mathbb{Z}$) is a countable family of firmly nonexpansive mappings and $F_\infty := \bigcap_{i \in J} \text{Fix}(T_i) \neq \emptyset$. Let $T := \sum_{i \in J} w_i T_i$, where $(w_i)_{i \in J} \subset (0, 1]$ and $\sum_{i \in J} w_i = 1$. Then T is firmly nonexpansive and $\text{Fix}(T) = F_\infty$ (see [43]).
- (f) If $T_i : \mathcal{H} \rightarrow \mathcal{H}$ ($i = 1, 2, \dots, m$) are attracting nonexpansive mappings with $\bigcap_{i=1}^m \text{Fix}(T_i) \neq \emptyset$, $T_m T_{m-1} \cdots T_1$ is attracting nonexpansive and $\text{Fix}(T_m T_{m-1} \cdots T_1) = \bigcap_{i=1}^m \text{Fix}(T_i)$.
- (g) Let $T_1 : \mathcal{H} \rightarrow \mathcal{H}$ be α_1 -averaged and $T_2 : \mathcal{H} \rightarrow \mathcal{H}$ α_2 -averaged for some $\alpha_1 \in [0, 1]$ and $\alpha_2 \in [0, 1]$. Then $(1-t)T_1 + tT_2$ is $\{(t-1)\alpha_1 + t\alpha_2\}$ -averaged for any $t \in [0, 1]$ (see [96]).
- (h) Let $T_1 : \mathcal{H} \rightarrow \mathcal{H}$ be α_1 -averaged and $T_2 : \mathcal{H} \rightarrow \mathcal{H}$ α_2 -averaged for some $\alpha_1 \in [0, 1]$ and $\alpha_2 \in [0, 1]$. Then it follows that $\frac{\alpha_1 + \alpha_2 - 2\alpha_1\alpha_2}{1 - \alpha_1\alpha_2} \in [0, 1]$ and $T_1 T_2$ is $\frac{\alpha_1 + \alpha_2 - 2\alpha_1\alpha_2}{1 - \alpha_1\alpha_2}$ -averaged (see [96]). In particular, for any pair of firmly nonexpansive mappings T_1 and T_2 that do not necessarily have a common fixed point, $(1 - \alpha)I + \alpha T_1 T_2$ is nonexpansive for all $\alpha \in [0, 3/2]$. Conversely, for every $\alpha \notin [0, 3/2]$, there exists a pair of firmly nonexpansive mappings T_1 and T_2 such that $(1 - \alpha)I + \alpha T_1 T_2$ is not nonexpansive (see [130]).

B. Variational Inequality Problem. A mapping $\mathcal{F} : \mathcal{H} \rightarrow \mathcal{H}$ is called (i) *monotone* over $S \subset \mathcal{H}$ if $\langle \mathcal{F}(u) - \mathcal{F}(v), u - v \rangle \geq 0$ for all $u, v \in S$. In particular, a mapping \mathcal{F} which is monotone over $S \subset \mathcal{H}$ is called (ii) *paramonotone* over S if $\langle \mathcal{F}(u) - \mathcal{F}(v), u - v \rangle = 0 \Leftrightarrow \mathcal{F}(u) = \mathcal{F}(v)$ for all $u, v \in S$; (iii) *η -inverse*

strongly monotone (or *firmlly monotone*) over S if there exists $\eta > 0$ such that $\langle \mathcal{F}(u) - \mathcal{F}(v), u - v \rangle \geq \eta \|\mathcal{F}(u) - \mathcal{F}(v)\|^2$ for all $u, v \in S$ [87]; (iv) *strictly monotone* over S if $\langle \mathcal{F}(u) - \mathcal{F}(v), u - v \rangle = 0 \Leftrightarrow u = v$ for all $u, v \in S$; (v) η -*strongly monotone* over S if there exists $\eta > 0$ such that $\langle \mathcal{F}(u) - \mathcal{F}(v), u - v \rangle \geq \eta \|u - v\|^2$ for all $u, v \in S$ [140].

The *variational inequality problem* $VIP(\mathcal{F}, C)$ is defined as follows: given $\mathcal{F} : \mathcal{H} \rightarrow \mathcal{H}$ which is monotone over a nonempty closed convex set $C \subset \mathcal{H}$, find $u^* \in C$ such that $\langle u - u^*, \mathcal{F}(u^*) \rangle \geq 0$ for all $u \in C$. If a function $\Theta : \mathcal{H} \rightarrow \mathbb{R} \cup \{\infty\}$ is *convex* over a closed convex set C ; i.e., $\Theta((1-t)u + tv) \leq (1-t)\Theta(u) + t\Theta(v)$ for all $t \in [0, 1]$ and $u, v \in C$, and Gâteaux differentiable with derivative Θ' over an open set $U \supset C$, then Θ' is paramonotone over C [36]. For such a Θ , the set $\Gamma := \{u \in C \mid \Theta(u) = \inf \Theta(C)\}$ is nothing but the solution set of $VIP(\Theta', C)$ (see for example Prop.II.2.1 of [61] and its proof).

FACT 2.2. [36, 61] Let \mathcal{F} be monotone and continuous over a nonempty closed convex set $C \subset \mathcal{H}$. Then

- (a) u^* is a solution of $VIP(\mathcal{F}, C)$ iff, for all $u \in C$, $\langle \mathcal{F}(u), u - u^* \rangle \geq 0$.
- (b) Suppose that (i) \mathcal{F} is paramonotone over C , (ii) $u^* \in C$ is a solution of $VIP(\mathcal{F}, C)$ and (iii) $u \in C$ satisfies $\langle \mathcal{F}(u), u - u^* \rangle = 0$. Then u is also a solution of $VIP(\mathcal{F}, C)$.

The characterization in (2.2) of the convex projection P_C yields at once an alternative interpretation of the VIP as a fixed point problem.

FACT 2.3. (VIP as a fixed point problem) Given $\mathcal{F} : \mathcal{H} \rightarrow \mathcal{H}$ which is monotone over a nonempty closed convex set C , the following three statements are equivalent.

- (a) $u^* \in C$ is a solution of $VIP(\mathcal{F}, C)$; i.e.,

$$\langle v - u^*, \mathcal{F}(u^*) \rangle \geq 0 \text{ for all } v \in C.$$

- (b) For an arbitrarily fixed $\mu > 0$, $u^* \in C$ satisfies

$$\langle v - u^*, (u^* - \mu \mathcal{F}(u^*)) - u^* \rangle \leq 0 \text{ for all } v \in C.$$

- (c) For an arbitrarily fixed $\mu > 0$,

$$(2.3) \quad u^* \in \text{Fix}(P_C(I - \mu \mathcal{F})).$$

Obviously, by Schwarz's inequality, if $\mathcal{F} : \mathcal{H} \rightarrow \mathcal{H}$ is η -inverse strongly monotone over \mathcal{H} , \mathcal{F} is $1/\eta$ -Lipschitzian over \mathcal{H} . But now we shed light on the following fact ensuring the converse direction when \mathcal{F} is given as the derivative of a convex function, which will be driving force for the broad applications of the hybrid steepest descent method to Problem 1.2(See Propositions 2.5 and 2.16).

FACT 2.4. (Lipschitz continuity implies inverse strong monotonicity [6, 58, 68, 87]) Let $\Psi : \mathcal{H} \rightarrow \mathbb{R} \cup \{\infty\}$ be a Gâteaux differentiable convex function with derivative $\Psi' : \mathcal{H} \rightarrow \mathcal{H}$. Then the following two statements are equivalent.

- (a) Ψ' is γ -Lipschitzian over \mathcal{H} .
- (b) Ψ' is $1/\gamma$ -inverse strongly monotone over \mathcal{H} .

These facts lead to the following invaluable fixed point characterization of the solution set of the convex optimization problem.

PROPOSITION 2.5. Suppose that $\Psi : \mathcal{H} \rightarrow \mathbb{R} \cup \{\infty\}$ is a Gâteaux differentiable convex function of which derivative $\Psi' : \mathcal{H} \rightarrow \mathcal{H}$ is γ -Lipschitzian over \mathcal{H} . Suppose also that $K \subset \mathcal{H}$ is a nonempty closed convex set and

$$K_\Psi := \{u \in K \mid \Psi(u) = \inf \Psi(K)\} \neq \emptyset.$$

Then for any $\nu \in (0, 2/\gamma]$, $I - \nu\Psi'$ is nonexpansive, and $K_\Psi = \text{Fix}(P_K(I - \nu\Psi'))$.

PROOF: By Fact 2.4, $\Psi' : \mathcal{H} \rightarrow \mathcal{H}$ is $1/\gamma$ -inverse strongly monotone over \mathcal{H} . Therefore, for an arbitrarily fixed $\nu \in [0, 2/\gamma]$, we deduce, for all $u, v \in \mathcal{H}$,

$$\begin{aligned} & \| (I - \nu\Psi')(u) - (I - \nu\Psi')(v) \|^2 \\ & \leq \|u - v\|^2 + \nu^2 \|\Psi'(u) - \Psi'(v)\|^2 - \frac{2\nu}{\gamma} \|\Psi'(u) - \Psi'(v)\|^2 \\ & = \|u - v\|^2 + \nu \left(\nu - \frac{2}{\gamma} \right) \|\Psi'(u) - \Psi'(v)\|^2 \leq \|u - v\|^2, \end{aligned}$$

which implies the nonexpansivity of $I - \nu\Psi'$. Now the remaining statement is obvious from Fact 2.3. (Q.E.D.)

EXAMPLE 2.6. Suppose that (i) \mathcal{H}_o is a possibly infinite dimensional real Hilbert space equipped with norm $\|\cdot\|_o$, and (ii) $A : \mathcal{H} \rightarrow \mathcal{H}_o$ is a bounded linear operator and $b \in \mathcal{H}_o$. Define a convex function $\Psi : \mathcal{H} \rightarrow \mathbb{R} \cup \{\infty\}$ by $\Psi(x) := \frac{1}{2}\|A(x) - b\|_o^2$, $x \in \mathcal{H}$. Then $\Psi'(x) = A^*A(x) - A^*(b)$ is $\|A\|^2$ -Lipschitzian over \mathcal{H} , where $A^* : \mathcal{H}_o \rightarrow \mathcal{H}$ denotes the adjoint operator of A [57, 84]. Proposition 2.5 shows that for $\nu \in (0, 2\|A\|^{-2}]$ (i) $P_K(I - \nu\Psi')$ is nonexpansive, and (ii) $K_\Psi = \text{Fix}(P_K(I - \nu\Psi'))$.

The notion of the generalized convex feasible set is particularly useful in the convexly constrained generalized inverse problems.

DEFINITION 2.7. (Generalized convex feasible set [129, 130]) For nonempty closed convex sets $C_i (\subset \mathcal{H})$ ($i = 1, 2, \dots, m$) and $K \subset \mathcal{H}$, define a proximity function $\Phi : \mathcal{H} \rightarrow \mathbb{R}$ by

$$(2.4) \quad \Phi(x) := \frac{1}{2} \sum_{i=1}^m w_i d(x, C_i)^2,$$

where $(w_i)_{i=1}^m \subset (0, 1]$ and $\sum_{i=1}^m w_i = 1$. Then the generalized convex feasible set $K_\Phi \subset K$ is defined by

$$(2.5) \quad K_\Phi := \{u \in K \mid \Phi(u) = \inf \Phi(K)\}.$$

We call the set K the absolute constraint because every point in K_Φ belongs to K .

REMARK 2.8.

- (a) Obviously, $K_\Phi = K \cap (\bigcap_{i=1}^m C_i)$ if $K \cap (\bigcap_{i=1}^m C_i) \neq \emptyset$. Even if $K \cap (\bigcap_{i=1}^m C_i) = \emptyset$, the set K_Φ is well-defined as the set of all minimizers of Φ over K . The set K_Φ of (2.5) is a nonempty closed convex set if one of the C_i 's or K is bounded (Note: This is an immediate consequence of Fact 2.9 and Propositions 2.13 and 2.14 below [or Proposition 7 of [42], Propositions 38.12 and 38.15 of [141]]).
- (b) The set $\mathcal{S} \subset \mathcal{H}$ in (1.1) can also be characterized as an example of the above generalized convex feasible set, which is verified as follows. For given $A : \mathcal{H} \rightarrow \mathbb{R}^m$ in Problem 1.1, define $a_i \in \mathcal{H}$ ($i = 1, \dots, m$) to get the

expression: $Ax = (\langle a_1, x \rangle, \dots, \langle a_m, x \rangle)^T$ (This is possible by F. Riesz's representation theorem[84, 134]). Then it is easy to see $\mathcal{S} = K_\Phi$ for $C_i := \Pi_i := \{x \in \mathcal{H} \mid \langle a_i, x \rangle = b_i\}$ and special weights $w_i := \frac{\|a_i\|^2}{\sum_{j=1}^m \|a_j\|^2}$ ($i = 1, \dots, m$). The fixed point characterization of \mathcal{S} based on Fact 2.9 and P_{Π_i} ($i = 1, \dots, m$) was used in [131, 133].

Note that $\Phi' = \sum_{i=1}^m w_i(I - P_{C_i}) = I - \sum_{i=1}^m w_i P_{C_i}$ is firmly nonexpansive (due to Fact 2.1(d)) and thus 1-inverse strongly monotone over \mathcal{H} . By applying Proposition 2.5 to Φ' , we deduce with Fact 2.1(h) the following useful characterizations of K_Φ .

FACT 2.9. (Fixed point characterization of K_Φ in Definition 2.7)

(a) For any $\alpha \neq 0$, it follows

$$K_\Phi = \text{Fix} \left((1 - \alpha)I + \alpha P_K \sum_{i=1}^m w_i P_{C_i} \right).$$

In particular, $T := (1 - \alpha)I + \alpha P_K \sum_{i=1}^m w_i P_{C_i}$ becomes nonexpansive for any $0 < \alpha \leq 3/2$ [130].

(b) For any $\beta > 0$, it follows

$$K_\Phi = \text{Fix} \left(P_K \left[(1 - \beta)I + \beta \sum_{i=1}^m w_i P_{C_i} \right] \right).$$

In particular, $T := P_K [(1 - \beta)I + \beta \sum_{i=1}^m w_i P_{C_i}]$ becomes nonexpansive for any $0 < \beta \leq 2$ [46].

In most digital signal/image processing applications, we can assume at least one of the closed convex sets K or C_i 's, for defining the set K_Φ in (2.5), is bounded. Motivated mainly by this fact, the notion of the *asymptotically shrinking mapping* was firstly introduced in [96].

DEFINITION 2.10. (Asymptotically shrinking mapping [96]) A mapping $T : \mathcal{H} \rightarrow \mathcal{H}$ is said to be *asymptotically shrinking* if there exists some $R > 0$ satisfying

$$(2.6) \quad \sup_{\|u\| \geq R} \frac{\|T(u)\|}{\|u\|} < 1.$$

The following propositions shown in [96] are useful to interpret the feasible sets of the inverse problems as the fixed point sets of asymptotically shrinking mappings.

PROPOSITION 2.11. [96] The following statements are equivalent for $T : \mathcal{H} \rightarrow \mathcal{H}$.

(a) T is asymptotically shrinking.

(b) There exist some $(u_1, u_2, u_3) \in \mathcal{H} \times \mathcal{H} \times \mathcal{H}$ and some $R_1 > \|u_1 - u_2\|$ satisfying

$$\sup_{\|u-u_1\| \geq R_1} \frac{\|T(u) - u_3\|}{\|u - u_2\|} < 1.$$

(c) For every $(v_1, v_2, v_3) \in \mathcal{H} \times \mathcal{H} \times \mathcal{H}$, there exists $R_2 > \|v_1 - v_2\|$ satisfying

$$\sup_{\|u-v_1\| \geq R_2} \frac{\|T(u) - v_3\|}{\|u - v_2\|} < 1.$$

PROPOSITION 2.12. [96]

- (a) $T : \mathcal{H} \rightarrow \mathcal{H}$ is asymptotically shrinking if $T(\mathcal{H})$ is bounded.
- (b) Let $T_1 : \mathcal{H} \rightarrow \mathcal{H}$ be nonexpansive and $T_2 : \mathcal{H} \rightarrow \mathcal{H}$ asymptotically shrinking. Then:
 - (i) αT_1 is asymptotically shrinking for $-1 < \alpha < 1$.
 - (ii) αT_2 is asymptotically shrinking for $-1 \leq \alpha \leq 1$.
 - (iii) $\alpha T_1 + (1 - \alpha)T_2$ is asymptotically shrinking for $0 \leq \alpha < 1$.
 - (iv) $T_1 T_2$ is asymptotically shrinking.
 - (v) $T_2 T_1$ is asymptotically shrinking if T_2 is nonexpansive.

PROPOSITION 2.13. [96] Let K and C_i ($i = 1, \dots, m$) be nonempty closed convex sets, and $(w_i)_{i=1}^m \subset (0, 1]$ satisfy $\sum_{i=1}^m w_i = 1$. Define $T : \mathcal{H} \rightarrow \mathcal{H}$ by $T := P_K(\sum_{i=1}^m w_i P_{C_i})$. Then we have :

- (a) Suppose that C_k is bounded for some $k \in \{1, 2, \dots, m\}$. Then, for any $u_1 \in C_k$ and any $\varepsilon > 0$, it follows that

$$\sup_{\|u-u_1\|\geq R} \frac{\|T(u) - T(u_1)\|}{\|u - u_1\|} < 1,$$

where $R := \sup_{u \in C_k} \|u - u_1\| + \varepsilon$, hence T is asymptotically shrinking.

- (b) Suppose that K is bounded. Then, for any $\varepsilon > 0$, it follows

$$\sup_{\|u\|\geq R} \frac{\|T(u)\|}{\|u\|} < 1,$$

where $R := \sup_{u \in K} \|u\| + \varepsilon$, hence T is asymptotically shrinking.

PROPOSITION 2.14. [96] If $T : \mathcal{H} \rightarrow \mathcal{H}$ is nonexpansive as well as asymptotically shrinking, then $\text{Fix}(T) \neq \emptyset$ and $\text{Fix}(T)$ is bounded.

C. Hybrid steepest descent method for the variational inequality problem over the fixed point set of the nonexpansive mapping. Motivated strongly by the tremendous progresses in the fixed point theory of nonexpansive mappings for the last four decades (see for example [5, 9, 10, 19–21, 37, 43, 56, 64, 65, 73, 86, 100, 106, 107, 120, 121, 125, 127] and references therein), the *hybrid steepest descent method* [54, 96, 97, 129–133] has been developed as a steepest descent type algorithm for the minimization of convex functions over the fixed point set of a nonexpansive mapping, or more generally, over the intersection of a family of the fixed point sets of nonexpansive mappings. The method is essentially an algorithmic solution to the variational inequality problem defined over the fixed point sets of nonexpansive mappings in a real Hilbert space \mathcal{H} [96, 97, 133].

The central results so far on the convergence of the *hybrid steepest descent method* for $VIP(\mathcal{F}, \text{Fix}(T))$ are summarized as follows.

THEOREM 2.15. [133] Let $T : \mathcal{H} \rightarrow \mathcal{H}$ be a nonexpansive mapping with $\text{Fix}(T) \neq \emptyset$. Suppose that a mapping $\mathcal{F} : \mathcal{H} \rightarrow \mathcal{H}$ is κ -Lipschitzian and η -strongly monotone over $T(\mathcal{H})$. Then, by using any sequence $(\lambda_n)_{n \geq 1} \subset [0, \infty)$ satisfying (W1) $\lim_{n \rightarrow +\infty} \lambda_n = 0$, (W2) $\sum_{n \geq 1} \lambda_n = +\infty$, (W3) $\sum_{n \geq 1} |\lambda_n - \lambda_{n+1}| < +\infty$ [or $(\lambda_n)_{n \geq 1} \subset (0, \infty)$ satisfying (L1) $\lim_{n \rightarrow +\infty} \lambda_n = 0$, (L2) $\sum_{n \geq 1} \lambda_n = +\infty$, (L3) $\lim_{n \rightarrow \infty} (\lambda_n - \lambda_{n+1})\lambda_{n+1}^{-2} = 0$], the sequence $(u_n)_{n \geq 0}$ generated, with arbitrary $u_0 \in \mathcal{H}$, by

$$(2.7) \quad u_{n+1} := T(u_n) - \lambda_{n+1} \mathcal{F}(T(u_n))$$

converges strongly to the uniquely existing solution of the VIP: find $u^* \in \text{Fix}(T)$ such that $\langle u - u^*, \mathcal{F}(u^*) \rangle \geq 0$ for all $u \in \text{Fix}(T)$.

Thanks to Proposition 2.5, we obtain the following proposition that presents an algorithmic solution to Problem 1.2.

PROPOSITION 2.16. *Let $K \subset \mathcal{H}$ be a closed convex set. Suppose that (i) $\Psi : \mathcal{H} \rightarrow \mathbb{R} \cup \{\infty\}$ is a Gâteaux differentiable convex function of which derivative $\Psi' : \mathcal{H} \rightarrow \mathcal{H}$ is γ -Lipschitzian over \mathcal{H} , and (ii) $\Theta : \mathcal{H} \rightarrow \mathbb{R} \cup \{\infty\}$ is a Gâteaux differentiable convex function of which derivative $\mathcal{F} := \Theta' : \mathcal{H} \rightarrow \mathcal{H}$ is κ -Lipschitzian and η -strongly monotone over $T(\mathcal{H})$, where $T := P_K(I - \nu\Psi')$ for an arbitrarily fixed $\nu \in (0, 2/\gamma]$. Then the sequence $(u_n)_{n \geq 0}$ generated, with arbitrary $u_0 \in \mathcal{H}$, by (2.7) converges strongly to the uniquely existing solution of Problem 1.2 if $K_\Psi \neq \emptyset$.*

REMARK 2.17.

- (a) Under the same conditions assumed in Theorem 2.15, the sequence $v_n := T(u_n)$ ($n = 0, 1, 2, \dots$) is generated, with any $v_0 := T(u_0) \in T(\mathcal{H})$, by

$$(2.8) \quad v_{n+1} := T(I - \lambda_{n+1}\mathcal{F})(v_n),$$

and obviously satisfies

$$0 \leq \|v_n - u^*\| = \|T(u_n) - u^*\| \leq \|u_n - u^*\| \rightarrow 0 \quad (n \rightarrow \infty).$$

The formula (2.8) can be regarded as a generalization of the projected gradient method:

$$(2.9) \quad v_{n+1} := P_C(I - \mu_{n+1}\mathcal{F})(v_n) \quad (n = 0, 1, 2, \dots),$$

which is a well-known algorithmic solution to $\text{VIP}(\mathcal{F}, C)$. Several convergence analyses on (2.9) are found for example in [55, 67, 68, 85, 140, 141]. These include the case where \mathcal{F} is an inverse strongly monotone mapping [Proposition 2.16 presents an algorithmic solution to $\text{VIP}(\widehat{\mathcal{F}}, K)$ for the case where $\widehat{\mathcal{F}}$ is an inverse strongly monotone mapping (See Fact 2.4, Proposition 2.5 and Theorem 2.15). The algorithm generates a sequence to converge strongly to the minimizer of Θ over the solution set of $\text{VIP}(\widehat{\mathcal{F}}, K)$.] The formula (2.9) leads to so called the projected Landweber method [13, 60] that has been widely used for the convexly constrained least-squares problems.

It is noteworthy that the extremely simple formula (2.8) reproduces (1.7) that resolves a special case of Problem 1.1 for $\Theta(x) := \frac{1}{2}\|x\|^2$. This fact is verified, based on Example 2.6, by applying, to (2.8), $\mathcal{F} := \Theta'$ and $T := P_K(I - \nu\Psi')$, where $\nu \in (0, 2\|A\|^{-2}]$, $\Theta(x) := \frac{1}{2}\|x\|^2$ and $\Psi'(x) = A^*A(x) - A^*(b)$. However, of course, Theorem 2.15 and Proposition 2.16 imply that the hybrid steepest descent method (2.7)[or (2.8)] can be applied to much broader class of Problem 1.2.

It must also be noted that the use of formula (2.9) is essentially based on the assumption that the closed form expression of $P_C : \mathcal{H} \rightarrow C$ is known, whereas in many situations the assumption does not hold. Theorem 2.15 or the iteration (2.8) requires only a closed form expression of any nonexpansive mapping $T : \mathcal{H} \rightarrow \mathcal{H}$ with $\text{Fix}(T) = C$. The facts in Section 2.A and Theorem 2.15 provide us with computational methods,

based on sound mathematical foundations, that can be used to solve a significantly wider class of variational inequality problems. (For other developments on the algorithmic solutions to $VIP(\mathcal{F}, C)$, see for example [36, 80, 83] and references therein, where the conditions imposed on \mathcal{F} are relaxed by splitting the formula (2.9) into two similar stages.)

- (b) Suppose that a nonexpansive mapping $T : \mathcal{H} \rightarrow \mathcal{H}$ and the derivative $\mathcal{F} = \Theta'$, of a differentiable convex function $\Theta : \mathcal{H} \rightarrow \mathbb{R} \cup \{\infty\}$, satisfy the conditions in Theorem 2.15. In such a case, the sequence $(u_n)_{n \geq 0}$ generated by (2.7) converges strongly to the unique minimizer of Θ over $Fix(T)$. In the formula, a new point u_{n+1} is generated by combining the operation of the mapping T and a small change in the steepest descent direction of Θ . Intuitively, the conditions (W1) and (W2) [or (L1) and (L2)] on $(\lambda_n)_{n \geq 1}$ ensure that the effect of the steepest descent direction in the formula declines but influences forever. The third condition (W3) [or (L3)] may be replaced by some other condition (See Theorem 2.18).
- (c) Suppose that $r \in \mathcal{H}$ and a self-adjoint bounded linear operator [84] $Q : \mathcal{H} \rightarrow \mathcal{H}$ is strongly positive; i.e., $(\exists \alpha > 0, \forall x \in \mathcal{H}) \langle Qx, x \rangle \geq \alpha \|x\|^2$ (Q is nothing but a positive definite matrix when \mathcal{H} is finite dimensional). Define a quadratic function $\Theta : \mathcal{H} \rightarrow \mathbb{R}$ by

$$\Theta(x) := \frac{1}{2} \langle Qx, x \rangle - \langle r, x \rangle, \quad x \in \mathcal{H}.$$

Then $\mathcal{F}(x) := \Theta'(x) = Qx - r$ is $\|Q\|$ -Lipschitzian and α -strongly monotone over \mathcal{H} [130] (see [54, 133] for other example of \mathcal{F} satisfying the conditions in Theorem 2.15). Obviously, Theorem 2.15 for this special example is also a generalization of the celebrated fixed point iteration [10, 73, 86, 127] so called the anchor method:

$$(2.10) \quad u_{n+1} := \lambda_{n+1}a + (1 - \lambda_{n+1})T(u_n),$$

which converges strongly to $P_{Fix(T)}(a)$. For the formula (2.10), the set of conditions (L1)–(L3) for $(\lambda_n)_{n \geq 1} \subset (0, 1]$ was introduced in [86] while the set of conditions (W1)–(W3) for $(\lambda_n)_{n \geq 1} \subset [0, 1]$ was introduced in [127]. [Note: $\lambda_n := 1/n^\varrho$ for $0 < \varrho < 1$ is a simple example of the sequence $(\lambda_n)_{n \geq 1}$ satisfying (L1)–(L3). The set of conditions (W1)–(W3) allows the case $\lambda_n = \frac{1}{n}$.]

By the use of similar $(\lambda_n)_{n \geq 1}$, an extended version, of (2.10),

$$u_{n+1} := \lambda_{n+1}a + (1 - \lambda_{n+1})T_{[n+1]}(u_n),$$

was also given in [10, 86] and shown its strong convergence to $P_F(a)$, where $[m] := m \bmod N$ and $T_i : \mathcal{H} \rightarrow \mathcal{H}$ ($i = 1, 2, \dots, N$) are nonexpansive mappings with $F := \bigcap_{i=1}^N Fix(T_i) \neq \emptyset$. This formula was also generalized to a formula for the variational inequality problem over F [133].

- (d) The ranges for $(\lambda_n)_{n \geq 1}$ in Theorem 2.15 are enlarged from $[0, 1]$ to $[0, \infty)$ [or from $(0, 1]$ to $(0, \infty)$]. These relaxations are justified by (W1)[or (L1)] (see [133, Remark 3.4(b)]) and surely broaden the possible ways how to use \mathcal{F} (= Θ' : the steepest descent direction of a convex function Θ) effectively in the early stage of the iterations.

To demonstrate simply the applicability of the formula (2.7) to more general paramonotone mapping \mathcal{F} , we focused our discussion to the case where $\dim(\mathcal{H}) < \infty$, and $T : \mathcal{H} \rightarrow \mathcal{H}$ is *attracting nonexpansive mapping* with bounded $\text{Fix}(T) \neq \emptyset$ (Note: Practically, by taking into account some physical conditions for example so called the bounded energy condition, these extra conditions are not restrictive in the applications of the method to the wide range of digital signal/image analysis).

THEOREM 2.18. [96, 97] Assume $\dim(\mathcal{H}) < \infty$. Suppose that (i) $T : \mathcal{H} \rightarrow \mathcal{H}$ is an *attracting nonexpansive mapping* with bounded $\text{Fix}(T) \neq \emptyset$, (ii) $\mathcal{F} : \mathcal{H} \rightarrow \mathcal{H}$ is κ -Lipschitzian over $T(\mathcal{H})$ as well as paramonotone over $\text{Fix}(T)$. If the following condition (a) or (b) is fulfilled, then the sequence $(u_n)_{n \geq 0}$ generated by (2.7), for arbitrary $u_0 \in \mathcal{H}$, satisfies

$$\lim_{n \rightarrow \infty} d(u_n, \Gamma) = 0,$$

where $\Gamma := \{u^* \in \text{Fix}(T) \mid \langle u - u^*, \mathcal{F}(u^*) \rangle \geq 0 \text{ for all } u \in \text{Fix}(T)\} \neq \emptyset$.

(a)

- (i) \mathcal{F} is monotone over $T(\mathcal{H}) \supset \text{Fix}(T)$,
- (ii) The nonnegative sequence $(\lambda_n)_{n \geq 1}$ in (2.7) satisfies (W1), (W2) and $(\lambda_n)_{n \geq 1} \in l^2$.

(b)

- (i) T is asymptotically shrinking (In this case, the nonemptiness and boundedness of $\text{Fix}(T)$ automatically holds [see Proposition 2.14]),
- (ii) The nonnegative sequence $(\lambda_n)_{n \geq 1}$ in (2.7) satisfies (W1) and (W2).

[Note: $\Gamma \neq \emptyset$ automatically holds because the requirements in [141, Theorem 54.A](or [61, Prop. II.3.1], [7, Theorem II.2.6]) are immediately verified. Moreover, by a simple inspection with Fact 2.2, Γ is bounded closed convex.]

3. Applications of the Hybrid steepest descent method to the convexly constrained inverse problems

A. Convexly constrained generalized pseudoinverse problem. Let's consider first Problem 1.2 in the case where the closed form expression of the metric projection $P_K : \mathcal{H} \rightarrow K$ is known and computationally available. In this case, Proposition 2.16 presents an algorithmic solution to the Problem 1.2 (See Remark 2.17).

B. Fixed point set theoretically constrained least-squares problem. As noted in the Remark 1.3(d), in many situations, the cost function Θ is not necessarily demanded. In such a case, the quality of the estimate for the convexly constrained problem (1.5) must be dominated strongly by how variety of *a priori* knowledge on the unknown estimandum is soundly incorporated into the closed convex set $K \subset \mathcal{H}$. The renunciation of the use of the cost function Θ and the fixed point set-theoretic characterization of K gift us with a notable way for the flexible incorporation of the *a priori* knowledge into the set K . We consider here the following convexly constrained inverse problem.

PROBLEM 3.1. Suppose that $K \subset \mathcal{H}$, of the Problem 1.1, is characterized as $K = \text{Fix}(T)$, where $T : \mathcal{H} \rightarrow \mathcal{H}$ is a nonexpansive mapping having a computationally available closed form expression. Then the problem is stated as (1.5).

We can incorporate various pieces of *a priori* information on the estimandum into the set K by applying the facts in Section 2.A.

Note that for given bounded linear operator $A : \mathcal{H} \rightarrow \mathbb{R}^m$ and $\mathbf{b} \in \mathbb{R}^m$ in the Problem 1.1, the derivative of the convex function

$$(3.1) \quad \phi(x) := \frac{1}{2} \|Ax - \mathbf{b}\|_o^2 = \frac{1}{2} \langle A^* A(x), x \rangle - \langle A^*(\mathbf{b}), x \rangle + \frac{1}{2} \|\mathbf{b}\|_o^2, \quad x \in \mathcal{H}$$

is given by $\mathcal{F}(x) := \phi'(x) = A^* A(x) - A^*(\mathbf{b})$, where $A^* : \mathbb{R}^m \rightarrow \mathcal{H}$ denotes the adjoint operator of A [57, 84], and thus paramonotone as well as $\|A\|^2$ -Lipschitzian over \mathcal{H} . Then if \mathcal{F} and T fulfill one of the sets of conditions in Theorems 2.15 and 2.18, the hybrid steepest descent method (2.7) generates a sequence approximating successively the solution of the Problem 3.1 (For other recent approaches to incorporate the *a priori* information into the closed convex set $K \subset \mathcal{H}$ in the problem (1.5), see for example [47], where two approaches are shown. The one is based on a fixed point theorem [127] with a special inner product $\langle \cdot, \cdot \rangle_Q : (x, y) \mapsto \langle Q(x), y \rangle$ for $Q := A^* A$ (hence Q is assumed to be invertible). The other is based on *outer approximation method* [48]).

C. Multi-layered hard-constrained convex feasibility problem. The problem of finding a point in the intersection of a given family of closed convex sets has been receiving great attention as one of the most sound foundations in the *set theoretic estimation scheme* [41] including the applications to a wide range of the inverse problems. This general problem is referred to as the *convex feasibility problem* [9, 34, 41, 53, 119] and many algorithmic solutions have been developed based on the use of convex projections P_{C_i} or on the use of certain other key operations, for example *Bregman projections* or *subgradient projections* (see for example [9, 11, 12, 15, 25–27, 34, 35, 41, 45, 71, 119, 135, 136] and references therein).

In estimation or design problems, however, because each closed convex set, say $C_i \subset \mathcal{H}$ ($i = 1, 2, \dots, m$), represents (possibly, not highly reliable) *a priori* information due to noisy measurements, or a (possibly too tight) design specification, some collection of convex sets C_i ($i = 1, 2, \dots, m$) may often become inconsistent, (i.e., $\bigcap_{i=1}^m C_i = \emptyset$) [41, 42, 46].

In the last decade, the use of the set \mathcal{H}_Φ (see its generalized version K_Φ in Definition 2.7), instead of the use of $C := \bigcap_{i=1}^m C_i$, has become one of the most promising strategies because \mathcal{H}_Φ reflects the priorities of all closed convex sets C_i 's through the weights w_i 's and $\mathcal{H}_\Phi = C$ when $C \neq \emptyset$. Indeed, many parallel algorithmic solutions to the convex feasibility problems have attracted attention not only because of their fast convergence but also because of their convergence to a point in such sets, even in the inconsistent case $C = \emptyset$ (For the convex feasibility problems or the best approximation problems over the intersection of the closed convex sets (originated from [95]), and their algorithmic solutions, consult the extensive surveys [9, 11, 26, 34, 41, 53, 119] and other papers for example [1, 8, 14, 16, 17, 23, 24, 28–33, 42, 45, 50–52, 59, 62, 77–79, 81, 102, 103, 108, 122, 126] and references therein).

We can adjust the weights w_i 's for \mathcal{H}_Φ , which results in imposing the priorities on each closed convex set C_i . However some estimate may have to belong to certain closed convex set when the set represents vital *a priori* information. Unfortunately, in general, the use of weights does not necessarily suite such a requirement. To resolve the difficulty, a more generalized problem of finding a point in the generalized convex feasible set K_Φ was firstly posed in [129, 130]. Indeed, this problem can

be solved soundly by applying any fixed point approximation formula, for example the *Krasnoselski/Mann type iteration* (1.9) [9, 43, 46, 90, 91, 114] or the *anchor method* (2.10) as well as the method (2.7), to the nonexpansive mapping T given in Fact 2.9.

Recently, we extended the notion of the set K_Φ to the multi-layered hard-constrained convex feasible set [98]. This generalization leads to more flexible ways to impose hard constraints in different levels.

PROBLEM 3.2. (m -layered hard constrained convex feasibility problem $HCF(m)$) Let $C_{i,j} \subset \mathcal{H}$ ($i \in \{1, \dots, m\}$, $j \in \{1, \dots, M_i\}$) be nonempty closed convex sets. For each $i = 1, 2, \dots, m$, define the proximity functions $\Phi_i(u) := \frac{1}{2} \sum_{j=1}^{M_i} w_{i,j} d^2(u, C_{i,j})$ with weights $w_{i,j} > 0$ satisfying $\sum_{j=1}^{M_i} w_{i,j} = 1$. Then the problem $HCF(m)$ is

Find a point $\hat{u} \in \Gamma_m$,

where

$$(3.2) \quad \Gamma_i := \begin{cases} \mathcal{H} & (i = 0) \\ \arg \min \Phi_i(\Gamma_{i-1}) & (i \in \{1, \dots, m\}). \end{cases}$$

For each i , we call $C_{i,j}$'s i -th layer target sets, and call Γ_i i -th layer hard constraint set respectively.

In the $HCF(m)$, the hard constraint set Γ_l ($l \leq m$) satisfies $\Gamma_l = C_{<l>} := \bigcap_{\substack{i \in \{1, \dots, l\} \\ j \in \{1, \dots, M_i\}}} C_{i,j}$ if $C_{<l>} \neq \emptyset$. Even in the inconsistent case (i.e. $C_{<l>} = \emptyset$), the set Γ_l can be a useful substitute for $C_{<l>}$ by reflecting the priorities of the target sets into the weights $w_{i,j}$'s.

REMARK 3.3.

- (a) In [41, 42], an algorithmic solution, named the *parallel projection method*, to the $HCF(1)$ and its signal processing applications were presented.
- (b) In [46, 54, 129, 130, 133], algorithmic solutions to $HCF(2)$ were developed for $M_1 = 1$. These have broad applications [46, 66, 131, 133].

Fortunately, as will seen below, under certain conditions, by the use of the hybrid steepest descent method (2.7), we can approximate successively the solution of the $HCF(3)$ (for $M_1 = 1$).

Let $\mathcal{F} := I - \sum_{j=1}^{M_3} w_{3,j} P_{C_{3,j}}$ and $T := P_{C_{1,1}} \sum_{i=1}^{M_2} w_{2,i} P_{C_{2,i}}$, where $C_{1,1}$, $C_{2,i}$, $C_{3,j} \subset \mathcal{H}$ are nonempty closed convex sets and $w_{i,j} > 0$ satisfies $\sum_{j=1}^{M_i} w_{i,j} = 1$ ($i = 2, 3$). In a way similar to the discussion for Fact 2.9, it is not hard to verify that (i) the Gâteaux derivative of Φ_3 is \mathcal{F} , (ii) Φ_3 is convex and \mathcal{F} is paramonotone over \mathcal{H} , (iii) \mathcal{F} is nonexpansive, i.e. 1-Lipschitzian over \mathcal{H} , (iv) T is attracting nonexpansive if $Fix(T) \neq \emptyset$ (see Fact 2.1(d),(h)), (v) T is asymptotically shrinking hence $Fix(T) \neq \emptyset$ if at least one of $C_{1,1}$ or $C_{2,i}$'s is bounded (see Prop.2.13 and 2.14) and (vi) $Fix(T) = \arg \inf \Phi_2(C_{1,1}) = \Gamma_2$ (see Fact 2.9). Then if these \mathcal{F} and T fulfill one of the sets of conditions in Theorems 2.15 and 2.18, the hybrid steepest descent method (2.7) approximates successively the solution of the $HCF(3)$ (for $M_1 = 1$).

4. A Variation of the Hybrid steepest descent method and its Numerical Robustness

It is of great interest to validate the sound applicability of the general ideas, of the hybrid steepest descent method, to the possibly ill-posed problems like the convexly constrained inverse problems. In this section, we present an error analysis of the formula (1.10), which fully extends the recent convergence analyses [96, 97] of the hybrid steepest descent method as well as proves its notable robustness to the error possibly caused at each computation of $T(u_n)$. Throughout the section, we assume that \mathcal{H} is a finite dimensional real Hilbert space, i.e., Euclidean space.

We proceed the convergence analysis on the iterative scheme:

$$(4.1) \quad u_{n+1} := T_{(n)}(u_n) + t_{n+1}e_n - \lambda_{n+1}\mathcal{F}(T_{(n)}(u_n) + t_{n+1}e_n),$$

where (i) $T : \mathcal{H} \rightarrow \mathcal{H}$ is a nonexpansive mapping with $\text{Fix}(T) \neq \emptyset$, (ii) $\mathcal{F} : \mathcal{H} \rightarrow \mathcal{H}$ is monotone over $\text{Fix}(T)$, (iii) $e_n \in \mathcal{H}$ ($n \geq 0$) is interpreted as the numerical error caused at the computation of $T(u_n)$, and (iv) $(\lambda_n, t_n)_{n \geq 1} \subset [0, \infty) \times [0, \infty)$ and $(e_n)_{n \geq 0} \subset \mathcal{H}$ will be assumed to satisfy certain subset of the following conditions (C1)–(C6):

- (C1) $(\exists N \geq 1, \forall n \geq N)$ (i) $0 < t_n \leq 1$, (ii) $T_{\tilde{t}} := (1 - \tilde{t})I + \tilde{t}T$ is attracting nonexpansive for $\tilde{t} := \sup_{n \geq N} t_n$,
- (C2) $(\forall \varepsilon > 0, \exists N \geq 1, \forall n \geq N)$ $\lambda_n \leq \varepsilon t_n$,
- (C3) $(\forall \varepsilon > 0, \exists N \geq 1, \forall n \geq N)$ $t_{n+1} \|e_n\| \leq \varepsilon \lambda_{n+1}$,
- (C4) $\sum_{n \geq 1} \lambda_n = \infty$,
- (C5) $(\exists N \geq 1) \inf_{n \geq N} \frac{t_{n+1}}{t_n} > 0$,
- (C6) $(\lambda_n)_{n \geq 1} \in l^2$.

REMARK 4.1.

- (a) $\lim_{n \rightarrow \infty} \lambda_n = 0$ by (C1) and (C2), or by (C6). $\sum_{n \geq 1} t_n = \infty$ by (C2) and (C4). $\lim_{n \rightarrow \infty} \|e_n\| = 0$ by (C2) and (C3). $\lim_{n \rightarrow \infty} t_{n+1} \|e_n\| = 0$ by (C1)–(C3).
- (b) By (C1), $T_{(n)}$ is attracting nonexpansive as well as $\text{Fix}(T_{(n)}) = \text{Fix}(T)$ for all $n \geq N$.
- (c) Suppose that (i) T is attracting nonexpansive, and (ii) there exists $N \geq 1$ such that $0 < t_n \leq 1$ for all $n \geq N$. Then (C1) automatically holds.
- (d) Suppose that (i) $\text{Fix}(T) \neq \emptyset$ and (ii) there exists N such that $t_n > 0$ for all $n \geq N$ as well as $\tilde{t} = \sup_{n \geq N} t_n < 1$. Then (C1) automatically holds because $T_{\tilde{t}}$ is \tilde{t} -averaged, hence attracting.

EXAMPLE 4.2.

- (a) T and $(\lambda_n)_{n \geq 1}$, used in Theorem 2.18 for (2.7), suffice (C1)–(C5) [or (C1)–(C6)] as a special case where $T_{(n)} = T$ and $(t_{n+1}, e_n) := (1, 0)$ for all $n \geq 0$.
- (b) Define $(\lambda_n, t_n)_{n \geq 1} \subset [0, \infty) \times [0, \infty)$ by $\lambda_n := n^{-\rho_1}$, $t_n := n^{-\rho_2}$ for $0 < \rho_2 < \rho_1 \leq 1$. Then it is easy to see that for any nonexpansive mapping $T : \mathcal{H} \rightarrow \mathcal{H}$ with $\text{Fix}(T) \neq \emptyset$, $(\lambda_n, t_n)_{n \geq 1}$ satisfies (C1), (C2), (C4) and (C5). Moreover, for arbitrarily fixed $\alpha > 0$ and any $\rho_3 > \rho_1 - \rho_2$, any $(e_n)_{n \geq 1} \subset \mathcal{H}$ satisfying

$$(\forall n \geq 0) \quad \|e_n\| \leq \alpha(n+1)^{-\rho_3}$$

fulfills (C3). In this case, we can choose $\rho_3 \leq 1$, which admits $(e_n)_{n \geq 0}$ that is not absolutely summable. Note in addition that $(\lambda_n)_{n \geq 1}$ suffices (C6) if $\rho_1 > \frac{1}{2}$.

Hereafter, for a nonempty bounded closed convex set $C \subset \mathcal{H}$ and $r \geq 0$, we use the notations: $\diamond(C, r) := \{u \in \mathcal{H} \mid d(u, C) = r\}$, $\triangleleft(C, r) := \{u \in \mathcal{H} \mid d(u, C) \leq r\}$ and $\triangleright(C, r) := \{u \in \mathcal{H} \mid d(u, C) \geq r\}$. For simplicity, we also use $\diamond(v, r) := \diamond(\{v\}, r)$, $\triangleleft(v, r) := \triangleleft(\{v\}, r)$ and $\triangleright(v, r) := \triangleright(\{v\}, r)$ for $v \in \mathcal{H}$. It is easy to verify that (i) $\diamond(C, r)$, $\triangleleft(C, r)$ and $\triangleright(C, r)$ are closed; (ii) $\diamond(C, r)$ and $\triangleleft(C, r)$ are bounded; (iii) $\triangleleft(C, r)$ is convex.

The next lemma is a key in the present analysis.

LEMMA 4.3. Let $T : \mathcal{H} \rightarrow \mathcal{H}$ be a nonexpansive mapping with bounded $\text{Fix}(T) \neq \emptyset$. Suppose that T and $(t_n)_{n \geq N}$ satisfy (C1). Then we have

(a) For any $f \in \text{Fix}(T)$, there exists $R > 0$ satisfying

$$\inf_{\substack{u \in \diamond(f, R) \\ n \geq N}} \frac{1}{t_{n+1}} (\|u - f\| - \|T_{(n)}(u) - T_{(n)}(f)\|) > 0.$$

(b) Define $D_n : [0, \infty) \rightarrow \mathbb{R}$ for all $n \geq N$ by

$$D_n(r) := \inf_{u \in \diamond(\text{Fix}(T), r)} \{r - d(T_{(n)}(u), \text{Fix}(T))\},$$

and $D : [0, \infty) \rightarrow \mathbb{R}$ by

$$D(r) := \inf_{u \in \diamond(\text{Fix}(T), r)} \{r - d(T_{\tilde{t}}(u), \text{Fix}(T))\}.$$

Then it follows that for all $n \geq N$

- (i) $0 \leq D_n(r) \leq r$ and $0 \leq D(r) \leq r$ for all $r \geq 0$,
- (ii) $D_n(r) = 0 \Leftrightarrow r = 0$, $D(r) = 0 \Leftrightarrow r = 0$,
- (iii) D_n and D are monotone increasing, i.e.,

$D_n(r_1) \geq D_n(r_2)$ and $D(r_1) \geq D(r_2)$ for all $r_1 > r_2 \geq 0$,

- (iv) $\frac{t_{n+1}}{\tilde{t}} D(r) \leq D_n(r)$ for all $r \geq 0$.

PROOF of (a): Fix $n \geq N$ and $f \in \text{Fix}(T)$ arbitrarily. Due to the boundedness of $\text{Fix}(T)$, there exists $R > 0$ satisfying $\sup_{u \in \text{Fix}(T)} \|u - f\| < R$. By the definition of $T_{(n)}$ and $T_{\tilde{t}}$, we have

$$T_{(n)}(u) - T_{(n)}(f) = \left(1 - \frac{t_{n+1}}{\tilde{t}}\right)(u - f) + \frac{t_{n+1}}{\tilde{t}}(T_{\tilde{t}}(u) - T_{\tilde{t}}(f))$$

and

$$\begin{aligned} & \|u - f\| - \|T_{(n)}(u) - T_{(n)}(f)\| \\ & \geq \|u - f\| - \left(1 - \frac{t_{n+1}}{\tilde{t}}\right) \|u - f\| - \frac{t_{n+1}}{\tilde{t}} \|T_{\tilde{t}}(u) - T_{\tilde{t}}(f)\| \\ & = \frac{t_{n+1}}{\tilde{t}} (\|u - f\| - \|T_{\tilde{t}}(u) - T_{\tilde{t}}(f)\|). \end{aligned}$$

Since $T_{\tilde{t}}$ is attracting nonexpansive and $\diamond(f, R) \cap \text{Fix}(T) = \emptyset$, the compactness of $\diamond(f, R)$ (see for example [2] for finite dimensional case) implies

$$\delta_1 := \min_{u \in \diamond(f, R)} (\|u - f\| - \|T_{\tilde{t}}(u) - T_{\tilde{t}}(f)\|) > 0.$$

These simple observations yield

$$\inf_{u \in \diamond(f, R)} (\|u - f\| - \|T_{(n)}(u) - T_{(n)}(f)\|) \geq \frac{t_{n+1}}{\tilde{t}} \delta_1 > 0.$$

$\frac{t_{n+1}}{\tilde{t}} \delta_1$ can serve as a more general lower bound as shown below.

Fix $u \in \triangleright(f, R)$ arbitrarily. Since $v := P_{\triangleleft(f, R)}(u) \in \diamond(f, R)$ suffices

$$\|u - f\| = \|u - v\| + \|v - f\|,$$

we deduce by the nonexpansivity of $T_{(n)}$

$$\begin{aligned} & \|u - f\| - \|T_{(n)}(u) - T_{(n)}(f)\| \\ & \geq \|u - v\| - \|T_{(n)}(u) - T_{(n)}(v)\| + \|v - f\| - \|T_{(n)}(v) - T_{(n)}(f)\| \\ & \geq \|u - v\| - \|u - v\| + \|v - f\| - \|T_{(n)}(v) - T_{(n)}(f)\| \\ & = \|v - f\| - \|T_{(n)}(v) - T_{(n)}(f)\| \geq t_{n+1} \frac{\delta_1}{\tilde{t}}. \end{aligned}$$

This completes the proof of (a). (Q.E.D.)

PROOF of (b):

- (i) We show only $0 \leq D_n(r) \leq r$ because the same proof works for $0 \leq D(r) \leq r$. For any $u \in \diamond(Fix(T), r)$, it follows that

$$\begin{aligned} 0 & \leq d(T_{(n)}(u), Fix(T)) \leq \|T_{(n)}(u) - T_{(n)}(P_{Fix(T)}(u))\| \\ & \leq \|u - P_{Fix(T)}(u)\| = r, \end{aligned}$$

which implies $0 \leq D_n(r) \leq r$.

- (ii) We show only $D_n(r) = 0 \Leftrightarrow r = 0$ because the same proof works for $D(r) = 0 \Leftrightarrow r = 0$.

$r = 0 \Rightarrow D_n(r) = 0$ holds obviously. We prove the converse part. Assume that

$$(4.2) \quad D_n(r) = 0 \text{ for some } r > 0.$$

Since $\diamond(Fix(T), r)$ is closed bounded and thus compact, there exists $\tilde{u} \in \diamond(Fix(T), r)$ such that

$$\begin{aligned} r - d(T_{(n)}(\tilde{u}), Fix(T)) &= \inf_{u \in \diamond(Fix(T), r)} \{r - d(T_{(n)}(u), Fix(T))\} \\ &= D_n(r) = 0. \end{aligned}$$

Then, by the nonexpansivity of $T_{(n)}$ and $Fix(T_{(n)}) = Fix(T)$, we obtain

$$\begin{aligned} r &= d(\tilde{u}, Fix(T)) = \|\tilde{u} - P_{Fix(T)}(\tilde{u})\| \\ &\geq \|T_{(n)}(\tilde{u}) - T_{(n)}(P_{Fix(T)}(\tilde{u}))\| \geq d(T_{(n)}(\tilde{u}), Fix(T)) = r. \end{aligned}$$

This implies

$$(4.3) \quad \begin{aligned} \|\tilde{u} - P_{Fix(T)}(\tilde{u})\| &= \|T_{(n)}(\tilde{u}) - T_{(n)}(P_{Fix(T)}(\tilde{u}))\| \\ &= \|T_{(n)}(\tilde{u}) - P_{Fix(T)}(\tilde{u})\|. \end{aligned}$$

However, since $T_{(n)}$ is attracting nonexpansive (see Remark 4.1(b)) and $\tilde{u} \notin Fix(T)$, we have

$$\|\tilde{u} - P_{Fix(T)}(\tilde{u})\| > \|T_{(n)}(\tilde{u}) - P_{Fix(T)}(\tilde{u})\|,$$

which violates (4.3). This shows that the assumption (4.2) does not hold.

- (iii) We show only the monotonicity of D_n because the same proof works for the monotonicity of D . For any $r_1 > r_2 \geq 0$ and any $u \in \diamond(Fix(T), r_1)$, let $v := P_{Fix(T)}(u) + \frac{r_2}{r_1}(u - P_{Fix(T)}(u))$. Then, by $v \in \diamond(Fix(T), r_2)$, it follows

$$D_n(r_2) = \inf_{x \in \diamond(Fix(T), r_2)} \{r_2 - d(T_{(n)}(x), Fix(T))\} \leq r_2 - d(T_{(n)}(v), Fix(T)),$$

and thus

$$\begin{aligned} d(T_{(n)}(u), Fix(T)) &\leq \|T_{(n)}(u) - P_{Fix(T)}(T_{(n)}(v))\| \\ &\leq \|T_{(n)}(u) - T_{(n)}(v)\| + \|T_{(n)}(v) - P_{Fix(T)}(T_{(n)}(v))\| \\ &\leq \|u - v\| + (r_2 - D_n(r_2)) \\ &= (r_1 - r_2) + (r_2 - D_n(r_2)) \\ &= r_1 - D_n(r_2). \end{aligned}$$

This proves

$$D_n(r_1) = \inf_{u \in \diamond(Fix(T), r_1)} \{r_1 - d(T_{(n)}(u), Fix(T))\} \geq D_n(r_2).$$

- (iv) Since $\diamond(Fix(T), r)$ is compact, there exists some $\tilde{u} \in \diamond(Fix(T), r)$ satisfying

$$\begin{aligned} D_n(r) &= \inf_{u \in \diamond(Fix(T), r)} \{r - d(T_{(n)}(u), Fix(T))\} \\ &= r - d(T_{(n)}(\tilde{u}), Fix(T)). \end{aligned}$$

By using

$$\left(1 - \frac{t_{n+1}}{\tilde{t}}\right) P_{Fix(T)}(\tilde{u}) + \frac{t_{n+1}}{\tilde{t}} P_{Fix(T)}(T_{\tilde{t}}(\tilde{u})) \in Fix(T),$$

we deduce

$$\begin{aligned} &d(T_{(n)}(\tilde{u}), Fix(T)) \\ &\leq \left\| T_{(n)}(\tilde{u}) - \left\{ \left(1 - \frac{t_{n+1}}{\tilde{t}}\right) P_{Fix(T)}(\tilde{u}) + \frac{t_{n+1}}{\tilde{t}} P_{Fix(T)}(T_{\tilde{t}}(\tilde{u})) \right\} \right\| \\ &= \left\| \left(1 - \frac{t_{n+1}}{\tilde{t}}\right) (\tilde{u} - P_{Fix(T)}(\tilde{u})) + \frac{t_{n+1}}{\tilde{t}} \{T_{\tilde{t}}(\tilde{u}) - P_{Fix(T)}(T_{\tilde{t}}(\tilde{u}))\} \right\| \\ &\leq \left(1 - \frac{t_{n+1}}{\tilde{t}}\right) \|\tilde{u} - P_{Fix(T)}(\tilde{u})\| + \frac{t_{n+1}}{\tilde{t}} \|T_{\tilde{t}}(\tilde{u}) - P_{Fix(T)}(T_{\tilde{t}}(\tilde{u}))\| \\ &= \left(1 - \frac{t_{n+1}}{\tilde{t}}\right) r + \frac{t_{n+1}}{\tilde{t}} d(T_{\tilde{t}}(\tilde{u}), Fix(T)), \end{aligned}$$

which implies

$$\begin{aligned} D_n(r) &= r - d(T_{(n)}(\tilde{u}), Fix(T)) \\ &\geq r - \left\{ \left(1 - \frac{t_{n+1}}{\tilde{t}}\right) r + \frac{t_{n+1}}{\tilde{t}} d(T_{\tilde{t}}(\tilde{u}), Fix(T)) \right\} \\ &= \frac{t_{n+1}}{\tilde{t}} \{r - d(T_{\tilde{t}}(\tilde{u}), Fix(T))\} \geq \frac{t_{n+1}}{\tilde{t}} D(r). \end{aligned}$$

(Q.E.D.)

In the sequel, we use

$$(4.4) \quad \Omega_N := \bigcup_{n \geq N} (T_{(n)}(\mathcal{H}) + \triangle(\mathbf{0}, \|e_n\|)).$$

REMARK 4.4. $\Omega_N = T(\mathcal{H})$ holds if $T_{(n)} = T$ and $(t_{n+1}, e_n) := (1, \mathbf{0})$ for all $n \geq 0$.

THEOREM 4.5. (Boundedness of the sequence) *Let $T : \mathcal{H} \rightarrow \mathcal{H}$ be a nonexpansive mapping with bounded $\text{Fix}(T) \neq \emptyset$. For arbitrary $u_0 \in \mathcal{H}$, the sequence $(u_n)_{n \geq 0}$ generated by (4.1) is bounded if the following condition (a) and (b) is fulfilled.*

- (a) (i) $T, (\lambda_n, t_n)_{n \geq 1} \subset [0, \infty) \times [0, \infty)$ and $(e_n)_{n \geq 0} \subset \mathcal{H}$ satisfy (C1)–(C3) and (C6).
(ii) $\mathcal{F} : \mathcal{H} \rightarrow \mathcal{H}$ is monotone as well as κ -Lipschitzian over Ω_N , where N suffices (C1).
- (b) (i) $(\lambda_n, t_n)_{n \geq 1} \subset [0, \infty) \times [0, \infty)$ satisfies (C2).
(ii) $(e_n)_{n \geq 0} \subset \mathcal{H}$ is bounded (This is fulfilled if (C2) and (C3) hold).
(iii) $T : \mathcal{H} \rightarrow \mathcal{H}$ is an asymptotically shrinking nonexpansive mapping (and thus the nonemptiness and the boundedness of $\text{Fix}(T)$ automatically hold).
(iv) $\mathcal{F} : \mathcal{H} \rightarrow \mathcal{H}$ is κ -Lipschitzian over Ω_N , where $N \geq 1$ suffices ($\forall n \geq N$) $0 < t_n \leq 1$.

PROOF in case (a): Fix $f \in \text{Fix}(T)$ arbitrarily. From Lemma 4.3(a), there exist some $R_1 > 0$ and $\delta > 0$ satisfying

$$(4.5) \quad (\forall n \geq N) \quad \inf_{u \in \Delta(f, R_1)} (\|u - f\| - \|T_{(n)}(u) - T_{(n)}(f)\|) \geq t_{n+1}\delta > 0.$$

From Remark 4.1(b), (C2), (C3) and (C6), there exist $N_1 \geq N$ and $R_2 > 0$ satisfying

$$(4.6) \quad \begin{cases} (\forall n \geq N_1) \quad t_{n+1}\delta \geq \lambda_{n+1} \|\mathcal{F}(T_{(n)}(f))\| + (1 + \kappa\lambda_{n+1})t_{n+1}\|e_n\| \\ \kappa^2 \sum_{i=N_1+1}^{\infty} \lambda_i^2 \leq \frac{1}{2} \end{cases}$$

and

$$(4.7) \quad \begin{cases} R_2 \geq \max_{n \leq N_1} \|u_n - f\| \\ R_2 \geq (1 + \kappa^2 \sup_{n \geq N_1} \lambda_{n+1}^2) R_1 \\ R_2 \geq \sup_{n \geq N_1} \lambda_{n+1} \|\mathcal{F}(T_{(n)}(f))\| \\ R_2 \geq \sup_{n \geq N_1} (1 + \kappa\lambda_{n+1})t_{n+1}\|e_n\|. \end{cases}$$

To prove the theorem in case (a), it suffices to show, by induction

$$(4.8) \quad (\forall n \geq N_1) \quad \|u_n - f\| \leq 3 \left(1 + 2\kappa^2 \sum_{i=N_1+1}^n \lambda_i^2 \right) R_2.$$

Remind that \mathcal{F} is monotone as well as κ -Lipschitzian over Ω_N . Then, we deduce for all $n \geq N_1$

$$\begin{aligned}
& \|u_{n+1} - f\| \\
= & \|T_{(n)}(u_n) - T_{(n)}(f) - \lambda_{n+1}\{\mathcal{F}(T_{(n)}(u_n)) - \mathcal{F}(T_{(n)}(f))\} \\
& - \lambda_{n+1}\{\mathcal{F}(T_{(n)}(u_n) + t_{n+1}e_n) - \mathcal{F}(T_{(n)}(u_n))\} - \lambda_{n+1}\mathcal{F}(T_{(n)}(f)) + t_{n+1}e_n\| \\
\leq & \|T_{(n)}(u_n) - T_{(n)}(f) - \lambda_{n+1}\{\mathcal{F}(T_{(n)}(u_n)) - \mathcal{F}(T_{(n)}(f))\}\| + t_{n+1}\|e_n\| \\
& + \lambda_{n+1}\|\mathcal{F}(T_{(n)}(f))\| + \lambda_{n+1}\|\mathcal{F}(T_{(n)}(u_n) + t_{n+1}e_n) - \mathcal{F}(T_{(n)}(u_n))\| \\
\leq & \left\{ \|T_{(n)}(u_n) - T_{(n)}(f)\|^2 + \lambda_{n+1}^2\|\mathcal{F}(T_{(n)}(u_n)) - \mathcal{F}(T_{(n)}(f))\|^2 \right. \\
& \left. - 2\lambda_{n+1}\langle T_{(n)}(u_n) - T_{(n)}(f), \mathcal{F}(T_{(n)}(u_n)) - \mathcal{F}(T_{(n)}(f)) \rangle \right\}^{\frac{1}{2}} \\
& + \lambda_{n+1}\|\mathcal{F}(T_{(n)}(f))\| + (1 + \kappa\lambda_{n+1})t_{n+1}\|e_n\| \\
\leq & \left\{ \|T_{(n)}(u_n) - T_{(n)}(f)\|^2 + \lambda_{n+1}^2\|\mathcal{F}(T_{(n)}(u_n)) - \mathcal{F}(T_{(n)}(f))\|^2 \right\}^{\frac{1}{2}} \\
& + \lambda_{n+1}\|\mathcal{F}(T_{(n)}(f))\| + (1 + \kappa\lambda_{n+1})t_{n+1}\|e_n\| \\
\leq & (1 + \kappa^2\lambda_{n+1}^2)^{\frac{1}{2}}\|T_{(n)}(u_n) - T_{(n)}(f)\| \\
& + \lambda_{n+1}\|\mathcal{F}(T_{(n)}(f))\| + (1 + \kappa\lambda_{n+1})t_{n+1}\|e_n\| \\
\leq & (1 + \kappa^2\lambda_{n+1}^2)\|T_{(n)}(u_n) - T_{(n)}(f)\| \\
& + \lambda_{n+1}\|\mathcal{F}(T_{(n)}(f))\| + (1 + \kappa\lambda_{n+1})t_{n+1}\|e_n\|.
\end{aligned}$$

This proves that

$$(4.9) \quad \|u_{n+1} - f\| \leq (1 + \kappa^2\lambda_{n+1}^2)\|T_{(n)}(u_n) - T_{(n)}(f)\| + \lambda_{n+1}\|\mathcal{F}(T_{(n)}(f))\| + (1 + \kappa\lambda_{n+1})t_{n+1}\|e_n\|.$$

Assume that (4.8) holds for some $n \geq N_1$ [See (4.7) for the case $n = N_1$].

(I) For $u_n \in \triangle(f, R_1)$: Applying (4.7) to (4.9), we obtain

$$(4.10) \quad \|u_{n+1} - f\| \leq 3R_2.$$

(II) For $u_n \in \triangleright(f, R_1)$: Applying (4.5) and (4.6) to (4.9), we deduce

$$\begin{aligned}
& \|u_{n+1} - f\| \\
&\leq (1 + \kappa^2 \lambda_{n+1}^2)(\|u_n - f\| - t_{n+1}\delta) \\
&\quad + \lambda_{n+1} \|\mathcal{F}(T_{(n)}(f))\| + (1 + \kappa \lambda_{n+1})t_{n+1} \|e_n\| \\
&= (1 + \kappa^2 \lambda_{n+1}^2) \|u_n - f\| - \kappa^2 \lambda_{n+1}^2 t_{n+1} \delta \\
&\quad + \{\lambda_{n+1} \|\mathcal{F}(T_{(n)}(f))\| + (1 + \kappa \lambda_{n+1})t_{n+1} \|e_n\| - t_{n+1}\delta\} \\
&\leq (1 + \kappa^2 \lambda_{n+1}^2) \|u_n - f\| \\
&\leq 3(1 + \kappa^2 \lambda_{n+1}^2) \left(1 + 2\kappa^2 \sum_{i=N_1+1}^n \lambda_i^2\right) R_2 \\
&= 3 \left(1 + \kappa^2 \lambda_{n+1}^2 + 2\kappa^2 \sum_{i=N_1+1}^n \lambda_i^2 + 2\kappa^2 \lambda_{n+1}^2 \kappa^2 \sum_{i=N_1+1}^n \lambda_i^2\right) R_2 \\
&\leq 3 \left(1 + \kappa^2 \lambda_{n+1}^2 + 2\kappa^2 \sum_{i=N_1+1}^n \lambda_i^2 + 2\frac{1}{2}\kappa^2 \lambda_{n+1}^2\right) R_2 \\
&= 3 \left(1 + 2\kappa^2 \sum_{i=N_1+1}^{n+1} \lambda_i^2\right) R_2.
\end{aligned}$$

This proves

$$(4.11) \quad \|u_{n+1} - f\| \leq 3 \left(1 + 2\kappa^2 \sum_{i=N_1+1}^{n+1} \lambda_i^2\right) R_2.$$

Therefore, the first inequality in (4.7), (4.10) and (4.11) ensure (4.8) for all $n \geq N_1$.
(Q.E.D.)

PROOF in case (b): Fix $f \in Fix(T)$ arbitrarily. Since T is asymptotically shrinking, by Proposition 2.11, there exist $\alpha \in (0, 1]$ and $R_1 > 0$ such that

$$(\forall u \in \triangleright(f, R_1)) \quad \|T(u) - T(f)\| \leq (1 - \alpha) \|u - f\|.$$

Therefore, by (b-iv), it follows that for all $u \in \triangleright(f, R_1)$ and all $n \geq N$,

$$\begin{aligned}
(4.12) \quad \|T_{(n)}(u) - T_{(n)}(f)\| &\leq (1 - t_{n+1}) \|u - f\| + t_{n+1} \|T(u) - T(f)\| \\
&\leq (1 - t_{n+1}) \|u - f\| + t_{n+1}(1 - \alpha) \|u - f\| \\
&= (1 - \alpha t_{n+1}) \|u - f\|.
\end{aligned}$$

By (C2), there exists some $N_1 \geq N$ such that

$$(4.13) \quad (\forall n \geq N_1) \quad t_n \geq \frac{2(\kappa + 1)}{\alpha} \lambda_n.$$

Take any $R_2 \geq 0$ satisfying

$$(4.14) \quad \|u_n - f\| \leq R_2 \text{ for all } n \leq N_1.$$

By the Lipschitz continuity of \mathcal{F} and the boundednesses of $Fix(T)$ and $(e_n)_{n \geq 0}$, the sequence $(\|\mathcal{F}(T_{(n)}(f) + t_{n+1}e_n)\|)_{n \geq N_1}$ is bounded.

Now, let's define a bounded closed ball C_0 by

$$(4.15) \quad C_0 := \left\{ u \in \mathcal{H} \mid \|u - f\| \leq \left(3 + \kappa \sup_{n \geq N_1} \{\lambda_{n+1}\}\right) R_3 \right\},$$

where

$$R_3 := \max \left\{ R_1, R_2, \frac{2}{\alpha} \sup_{n \geq N_1} \|e_n\|, \sup_{n \geq N_1} \|\mathcal{F}(T_{(n)}(f) + t_{n+1}e_n)\| \right\} < \infty.$$

Note that (4.14) and (4.15) ensure $u_n \in C_0$ for all $n \leq N_1$. Then, to show the boundedness of $(u_n)_{n \geq 0}$, it suffices to prove

$$(4.16) \quad (\forall n \geq N_1) \quad u_n \in C_0 \quad \Rightarrow \quad u_{n+1} \in C_0.$$

By the Lipschitz continuity of \mathcal{F} over Ω_N , we deduce for all $n \geq N_1$

$$\begin{aligned} & \|u_{n+1} - f\| \\ = & \|T_{(n)}(u_n) - T_{(n)}(f) - \lambda_{n+1}\{\mathcal{F}(T_{(n)}(u_n) + t_{n+1}e_n) - \mathcal{F}(T_{(n)}(f) + t_{n+1}e_n)\} \\ & - \lambda_{n+1}\mathcal{F}(T_{(n)}(f) + t_{n+1}e_n) + t_{n+1}e_n\| \\ \leq & \|T_{(n)}(u_n) - T_{(n)}(f)\| + \kappa\lambda_{n+1} \|T_{(n)}(u_n) + t_{n+1}e_n - \{T_{(n)}(f) + t_{n+1}e_n\}\| \\ & + \lambda_{n+1} \|\mathcal{F}(T_{(n)}(f) + t_{n+1}e_n)\| + t_{n+1} \|e_n\| \\ = & (1 + \kappa\lambda_{n+1}) \|T_{(n)}(u_n) - T_{(n)}(f)\| + \lambda_{n+1} \|\mathcal{F}(T_{(n)}(f) + t_{n+1}e_n)\| + t_{n+1} \|e_n\|. \end{aligned}$$

This fact is used to show (4.16) in the following two cases.

(I) For $u_n \in \triangle(f, R_3) \subset C_0$: By (4.13) and (4.15), we obtain

$$\|u_{n+1} - f\| \leq (1 + \kappa\lambda_{n+1})R_3 + \frac{\alpha}{2(\kappa + 1)}t_{n+1} \|\mathcal{F}(T_{(n)}(f) + t_{n+1}e_n)\| + t_{n+1} \|e_n\|,$$

which implies $u_{n+1} \in C_0$.

(II) For $u_n \in \triangleright(f, R_3) \cap C_0$: By (4.12), (4.15) and (4.13), we obtain

$$\begin{aligned} & \|u_{n+1} - f\| \\ \leq & (1 + \kappa\lambda_{n+1})(1 - \alpha t_{n+1}) \|u_n - f\| + \lambda_{n+1} \|\mathcal{F}(T_{(n)}(f) + t_{n+1}e_n)\| + t_{n+1} \|e_n\| \\ \leq & \{1 - \alpha\kappa t_{n+1}\lambda_{n+1} - (\alpha t_{n+1} - \kappa\lambda_{n+1})\} \|u_n - f\| + \lambda_{n+1}R_3 + \frac{\alpha}{2}t_{n+1}R_3 \\ \leq & \left[1 - \alpha\kappa t_{n+1}\lambda_{n+1} - \left\{ \frac{\alpha}{2}t_{n+1} - (1 + \kappa)\lambda_{n+1} \right\} \right] \|u_n - f\| \\ \leq & \|u_n - f\|, \end{aligned}$$

which implies $u_{n+1} \in C_0$.

(I) and (II) prove (4.16) hence the theorem in case (b) [Note: This proof is valid for infinite dimensional \mathcal{H} as well]. (Q.E.D.)

THEOREM 4.6. (Distance to the fixed point set) Let $T : \mathcal{H} \rightarrow \mathcal{H}$ be a non-expansive mapping with bounded $\text{Fix}(T) \neq \emptyset$. Suppose that $T, (\lambda_n, t_n)_{n \geq 1} \subset [0, \infty) \times [0, \infty)$ and $(e_n)_{n \geq 0} \subset \mathcal{H}$ satisfy the conditions (C1)–(C5). Let $\mathcal{F} : \mathcal{H} \rightarrow \mathcal{H}$ be κ -Lipschitzian over Ω_N , in (4.4), where N suffices (C1). Suppose that the sequence $(u_n)_{n \geq 0}$, generated by (4.1), with some $u_0 \in \mathcal{H}$, is bounded [see Theorem 4.5 for typical cases that guarantee the boundedness of $(u_n)_{n \geq 0}$]. Then, it follows that

$$(4.17) \quad \lim_{n \rightarrow \infty} d(u_n, \text{Fix}(T)) = 0,$$

$$(4.18) \quad \lim_{n \rightarrow \infty} \|T_{(n)}(u_n) - u_n\| = 0,$$

$$(4.19) \quad \lim_{n \rightarrow \infty} \|u_{n+1} - u_n\| = 0.$$

PROOF: Since \mathcal{F} is κ -Lipschitzian over Ω_N , the boundedness of $(u_n)_{n \geq N}$ ensures the existence of some $R_1 > 0$ satisfying

$$(\forall n \geq N) \quad \|\mathcal{F}(T_{(n)}(u_n) + t_{n+1}e_n)\| \leq R_1.$$

By a simple inspection and the use of the above $R_1 > 0$, we deduce

$$\begin{aligned} & d(u_{n+1}, \text{Fix}(T)) \\ & \leq \|u_{n+1} - P_{\text{Fix}(T)}(T_{(n)}(u_n))\| \\ & = \|T_{(n)}(u_n) - P_{\text{Fix}(T)}(T_{(n)}(u_n)) + t_{n+1}e_n - \lambda_{n+1}\mathcal{F}(T_{(n)}(u_n) + t_{n+1}e_n)\| \\ & \leq \|T_{(n)}(u_n) - P_{\text{Fix}(T)}(T_{(n)}(u_n))\| + t_{n+1}\|e_n\| + \lambda_{n+1}\|\mathcal{F}(T_{(n)}(u_n) + t_{n+1}e_n)\| \\ & \leq d(T_{(n)}(u_n), \text{Fix}(T)) + \lambda_{n+1}R_1 + t_{n+1}\|e_n\|. \end{aligned}$$

Moreover, by applying Lemma 4.3(b) to $r = a_n := d(u_n, \text{Fix}(T))$ and by the above inequality, we obtain, for all $n \geq N$

$$(4.20) \quad \begin{aligned} D_n(a_n) & \leq a_n - d(T_{(n)}(u_n), \text{Fix}(T)) \\ & \leq a_n - a_{n+1} + \lambda_{n+1}R_1 + t_{n+1}\|e_n\|. \end{aligned}$$

Now, fix $\eta > 0$ arbitrarily. Then, by (C1)–(C3) and (C5), there exists $N_1 \geq N$ such that

$$(4.21) \quad \begin{aligned} & t_{n+1} \leq 1, \\ & (\forall n \geq N_1) \quad t_{n+1}D(\eta) \geq 2\tilde{t}\lambda_{n+1}R_1 + 2\tilde{t}t_{n+1}\|e_n\|, \\ & t_{n+1}D(\eta) \geq 2\tilde{t}\lambda_nR_1 + 2\tilde{t}t_n\|e_{n-1}\|. \end{aligned}$$

Apparently, to prove (4.17), it suffices to show, by induction,

$$(4.22) \quad (\exists N_2 \geq N_1, \forall n \geq N_2) \quad a_n \leq \eta + \lambda_nR_1 + t_n\|e_{n-1}\|.$$

(I) Firstly we will show

$$(4.23) \quad (\exists N_2 \geq N_1) \quad a_{N_2} \leq \eta.$$

Assume the contrary, i.e.,

$$(4.24) \quad (\forall n \geq N_1) \quad a_n > \eta.$$

By (4.20), (4.21), (4.24) and Lemma 4.3(b), it follows that

$$\begin{aligned} a_{n+1} & \leq a_n + \lambda_{n+1}R_1 + t_{n+1}\|e_n\| - D_n(a_n) \\ & \leq a_n + \frac{1}{2\tilde{t}}t_{n+1}D(\eta) - D_n(\eta) \\ & \leq a_n + \frac{1}{2\tilde{t}}t_{n+1}D(\eta) - \frac{t_{n+1}}{\tilde{t}}D(\eta) \\ & = a_n - \frac{1}{2\tilde{t}}t_{n+1}D(\eta) \\ & \quad \vdots \\ & \leq a_{N_1} - \frac{1}{2\tilde{t}}D(\eta) \sum_{i=N_1+1}^{n+1} t_i. \end{aligned}$$

Moreover, by Remark 4.1(a), there exists some $N_2 > N_1$ such that

$$(\forall n \geq N_2) \quad a_{N_1} < \frac{1}{2\tilde{t}}D(\eta) \sum_{i=N_1+1}^{n+1} t_i \quad \text{for all } n \geq N_2.$$

This shows the contradiction:

$$(\forall n \geq N_2) \quad a_{n+1} = d(u_{n+1}, Fix(T)) < 0,$$

hence (4.23) is proved.

(II) Next, by using N_2 in (4.23), we will show

$$(4.25) \quad (\forall n \geq N_2) \quad \begin{aligned} a_n &\leq \eta + \lambda_n R_1 + t_n \|e_{n-1}\| \\ &\Rightarrow a_{n+1} \leq \eta + \lambda_{n+1} R_1 + t_{n+1} \|e_n\|. \end{aligned}$$

(i) Suppose that $a_n < \eta$ for some $n \geq N_2$. Then, by (4.20), it follows

$$\begin{aligned} a_{n+1} &\leq a_n + \lambda_{n+1} R_1 + t_{n+1} \|e_n\| - D_n(a_n) \\ &\leq a_n + \lambda_{n+1} R_1 + t_{n+1} \|e_n\| < \eta + \lambda_{n+1} R_1 + t_{n+1} \|e_n\|. \end{aligned}$$

(ii) Suppose that $\eta \leq a_n \leq \eta + \lambda_n R_1 + t_n \|e_{n-1}\|$ for some $n \geq N_2$. In this case, by (4.21) and Lemma 4.3(b), it follows

$$(4.26) \quad \begin{aligned} D_n(a_n) &\geq D_n(\eta) \geq \frac{t_{n+1}}{\tilde{t}} D(\eta) = \frac{1}{2\tilde{t}} t_{n+1} D(\eta) + \frac{1}{2\tilde{t}} t_{n+1} D(\eta) \\ &\geq \lambda_{n+1} R_1 + t_{n+1} \|e_n\| + \lambda_n R_1 + t_n \|e_{n-1}\|. \end{aligned}$$

Applying this relation to (4.20), we obtain

$$\begin{aligned} a_{n+1} &\leq a_n + \lambda_{n+1} R_1 + t_{n+1} \|e_n\| - D_n(a_n) \\ &\leq a_n - \lambda_n R_1 - t_n \|e_{n-1}\| \leq \eta. \end{aligned}$$

These two cases show (4.25).

Finally (4.23) and (4.25) imply (4.22) and thus (4.17).

(4.18) and (4.19) are obvious because, by Remark 4.1(b),

$$\begin{aligned} \|T_{(n)}(u_n) - u_n\| &= \|T_{(n)}(u_n) - P_{Fix(T)}(u_n) + P_{Fix(T)}(u_n) - u_n\| \\ &\leq \|T_{(n)}(u_n) - P_{Fix(T)}(u_n)\| + \|P_{Fix(T)}(u_n) - u_n\| \\ &\leq \|u_n - P_{Fix(T)}(u_n)\| + \|P_{Fix(T)}(u_n) - u_n\| \\ &= 2d(u_n, Fix(T)) \rightarrow 0 \ (n \rightarrow \infty) \end{aligned}$$

and

$$\begin{aligned} \|u_{n+1} - u_n\| &= \|T_{(n)}(u_n) + t_{n+1} e_n - \lambda_{n+1} \mathcal{F}(T_{(n)}(u_n) + t_{n+1} e_n) - u_n\| \\ &\leq \|T_{(n)}(u_n) - u_n\| + t_{n+1} \|e_n\| + \lambda_{n+1} \|\mathcal{F}(T_{(n)}(u_n) + t_{n+1} e_n)\| \\ &\leq \|T_{(n)}(u_n) - u_n\| + t_{n+1} \|e_n\| + \lambda_{n+1} R_1 \\ &\rightarrow 0 \ (n \rightarrow \infty). \end{aligned}$$

(Q.E.D.)

THEOREM 4.7. (Convergence of the formula (4.1)) Let $T : \mathcal{H} \rightarrow \mathcal{H}$ be a non-expansive mapping with bounded $Fix(T) \neq \emptyset$. Suppose that $T, (\lambda_n, t_n)_{n \geq 1} \subset [0, \infty) \times [0, \infty)$ and $(e_n)_{n \geq 0} \subset \mathcal{H}$ satisfy the conditions (C1)–(C5). Suppose also that $\mathcal{F} : \mathcal{H} \rightarrow \mathcal{H}$ is paramonotone over $Fix(T)$ and κ -Lipschitzian over Ω_N , where N suffices (C1). Define a sequence $(u_n)_{n \geq 0} \subset \mathcal{H}$ by (4.1) with arbitrarily given $u_0 \in \mathcal{H}$. Then, if $(u_n)_{n \geq 0}$ is bounded [See Theorem 4.5 for typical cases that guarantee the boundedness of $(u_n)_{n \geq 0}$], it follows that

$$(4.27) \quad \lim_{n \rightarrow \infty} d(u_n, \Gamma) = 0,$$

where Γ is the solution set of $VIP(\mathcal{F}, Fix(T))$; i.e., $\Gamma := \{u^* \in Fix(T) \mid \langle u - u^*, \mathcal{F}(u^*) \rangle \geq 0 \text{ for all } u \in Fix(T)\} \neq \emptyset$ (Note: $\Gamma \neq \emptyset$ automatically holds. See the comment in Theorem 2.18).

PROOF: If $\Gamma = Fix(T)$, (4.27) is immediate from Theorem 4.6. Therefore, in the following, we assume $\Gamma \subsetneq Fix(T)$. In this case, note that there exists $\delta_0 > 0$ such that

$$(4.28) \quad (\forall \delta \in (0, \delta_0]) \quad Fix(T) \cap \triangleright(\Gamma, \delta) \neq \emptyset.$$

Moreover, by the Lipschitz continuity of \mathcal{F} over Ω_N , and the boundednesses of $(u_n)_{n \geq 0}$ and $Fix(T)$, it is easy to verify the existence of $R_1 > 0$ satisfying

$$(4.29) \quad \begin{cases} (\forall u, v \in Fix(T) \subset T(\mathcal{H})) & R_1 \geq \|\mathcal{F}(u)\|, \\ & R_1 \geq \|u - v\|, \\ (\forall n \geq N) & R_1 \geq \|\mathcal{F}(T_{(n)}(u_n))\|, \\ & R_1 \geq \|u_n - P_\Gamma(u_n)\|, \\ & R_1 \geq t_{n+1} \|e_n\|, \\ & R_1 \geq \lambda_{n+1}, \end{cases}$$

and

$$(4.30) \quad (\forall n \geq N, \forall y \in \Gamma) \quad \begin{cases} R_1 \geq \|T_{(n)}(u_n) - y - \lambda_{n+1} \mathcal{F}(T_{(n)}(u_n))\|, \\ R_1 \geq \|T_{(n)}(u_n) - y - \lambda_{n+1} \mathcal{F}(T_{(n)}(u_n) + t_{n+1} e_n)\|. \end{cases}$$

(I) Our first goal is to prove

$$(4.31) \quad (\forall u^* \in \Gamma) \quad \liminf_{n \rightarrow \infty} \langle \mathcal{F}(T_{(n)}(u_n)), T_{(n)}(u_n) - u^* \rangle = 0.$$

The inner product on the left hand side of (4.31) is expressed as

$$(4.32) \quad \begin{aligned} & \langle \mathcal{F}(T_{(n)}(u_n)), T_{(n)}(u_n) - u^* \rangle \\ &= \langle \mathcal{F}(P_{Fix(T)}(T_{(n)}(u_n))), P_{Fix(T)}(T_{(n)}(u_n)) - u^* \rangle \\ &+ \langle \mathcal{F}(P_{Fix(T)}(T_{(n)}(u_n))), T_{(n)}(u_n) - P_{Fix(T)}(T_{(n)}(u_n)) \rangle \\ &+ \langle \mathcal{F}(T_{(n)}(u_n)) - \mathcal{F}(P_{Fix(T)}(T_{(n)}(u_n))), T_{(n)}(u_n) - u^* \rangle. \end{aligned}$$

By Fact 2.2(a),

$$(4.33) \quad \liminf_{n \rightarrow \infty} \langle \mathcal{F}(P_{Fix(T)}(T_{(n)}(u_n))), P_{Fix(T)}(T_{(n)}(u_n)) - u^* \rangle \geq 0.$$

For $n \geq N$, by (4.29), the nonexpansivity of $T_{(n)}$ and $Fix(T) = Fix(T_{(n)})$, it follows

$$\begin{aligned} & |\langle \mathcal{F}(P_{Fix(T)}(T_{(n)}(u_n))), T_{(n)}(u_n) - P_{Fix(T)}(T_{(n)}(u_n)) \rangle| \\ &\leq R_1 \|T_{(n)}(u_n) - P_{Fix(T)}(T_{(n)}(u_n))\| \leq R_1 \|T_{(n)}(u_n) - P_{Fix(T)}(u_n)\| \\ &\leq R_1 \|u_n - P_{Fix(T)}(u_n)\| = R_1 d(u_n, Fix(T)). \end{aligned}$$

Then, by Theorem 4.6, we have

$$(4.34) \quad \lim_{n \rightarrow \infty} \langle \mathcal{F}(P_{Fix(T)}(T_{(n)}(u_n))), T_{(n)}(u_n) - P_{Fix(T)}(T_{(n)}(u_n)) \rangle = 0.$$

Moreover, by the boundedness of $(\|T_{(n)}(u_n) - u^*\|)_{n \geq 0}$ and Schwarz's inequality, it follows

$$(4.35) \quad \lim_{n \rightarrow \infty} \langle \mathcal{F}(T_{(n)}(u_n)) - \mathcal{F}(P_{Fix(T)}(T_{(n)}(u_n))), T_{(n)}(u_n) - u^* \rangle = 0.$$

Now by applying (4.33), (4.34) and (4.35) to (4.32), we obtain

$$(4.36) \quad \begin{aligned} & \liminf_{n \rightarrow \infty} \langle \mathcal{F}(T_{(n)}(u_n)), T_{(n)}(u_n) - u^* \rangle \\ = & \liminf_{n \rightarrow \infty} \langle \mathcal{F}(P_{\text{Fix}(T)}(T_{(n)}(u_n))), P_{\text{Fix}(T)}(T_{(n)}(u_n)) - u^* \rangle \\ & + \lim_{n \rightarrow \infty} \langle \mathcal{F}(P_{\text{Fix}(T)}(T_{(n)}(u_n))), T_{(n)}(u_n) - P_{\text{Fix}(T)}(T_{(n)}(u_n)) \rangle \\ & + \lim_{n \rightarrow \infty} \langle \mathcal{F}(T_{(n)}(u_n)) - \mathcal{F}(P_{\text{Fix}(T)}(T_{(n)}(u_n))), T_{(n)}(u_n) - u^* \rangle \geq 0. \end{aligned}$$

We will show that the left hand side of (4.36) indeed achieves 0. Assume

$$(4.37) \quad c_1 := \liminf_{n \rightarrow \infty} \langle \mathcal{F}(T_{(n)}(u_n)), T_{(n)}(u_n) - u^* \rangle > 0$$

and thus there exists $N_1 \geq N$ satisfying

$$(4.38) \quad (\forall n \geq N_1) \quad \langle \mathcal{F}(T_{(n)}(u_n)), T_{(n)}(u_n) - u^* \rangle \geq \frac{1}{2}c_1.$$

Then, by $u^* \in \Gamma \subset \text{Fix}(T)$ and by the nonexpansivity of $T_{(n)}$, we deduce, for all $n \geq N_1$:

$$(4.39) \quad \begin{aligned} & \|u_{n+1} - u^*\|^2 \\ = & \|T_{(n)}(u_n) + t_{n+1}e_n - \lambda_{n+1}\mathcal{F}(T_{(n)}(u_n) + t_{n+1}e_n) - u^*\|^2 \\ \leq & \{\|T_{(n)}(u_n) - u^* - \lambda_{n+1}\mathcal{F}(T_{(n)}(u_n))\| \\ & + \|-\lambda_{n+1}\mathcal{F}(T_{(n)}(u_n) + t_{n+1}e_n) + \lambda_{n+1}\mathcal{F}(T_{(n)}(u_n))\| + \|t_{n+1}e_n\|\}^2 \\ \leq & \{\|T_{(n)}(u_n) - u^* - \lambda_{n+1}\mathcal{F}(T_{(n)}(u_n))\| + (\lambda_{n+1}\kappa\|t_{n+1}e_n\| + \|t_{n+1}e_n\|)\}^2 \\ = & \|T_{(n)}(u_n) - u^* - \lambda_{n+1}\mathcal{F}(T_{(n)}(u_n))\|^2 + (\lambda_{n+1}\kappa\|t_{n+1}e_n\| + \|t_{n+1}e_n\|)^2 \\ & + 2\|T_{(n)}(u_n) - u^* - \lambda_{n+1}\mathcal{F}(T_{(n)}(u_n))\|(\lambda_{n+1}\kappa\|t_{n+1}e_n\| + \|t_{n+1}e_n\|) \\ \leq & \|(T_{(n)}(u_n) - u^*) - \lambda_{n+1}\mathcal{F}(T_{(n)}(u_n))\|^2 + (\lambda_{n+1}\kappa + 1)^2 t_{n+1}^2 \|e_n\|^2 \\ & + 2R_1(\lambda_{n+1}\kappa + 1)t_{n+1}\|e_n\| \\ \leq & \{\|T_{(n)}(u_n) - u^*\|^2 - 2\lambda_{n+1} \langle \mathcal{F}(T_{(n)}(u_n)), T_{(n)}(u_n) - u^* \rangle \\ & + \lambda_{n+1}^2 \|\mathcal{F}(T_{(n)}(u_n))\|^2\} + (R_1\kappa + 1)^2 t_{n+1}^2 \|e_n\|^2 + 2R_1(R_1\kappa + 1)t_{n+1}\|e_n\| \\ \leq & \|u_n - u^*\|^2 - 2\lambda_{n+1} \langle \mathcal{F}(T_{(n)}(u_n)), T_{(n)}(u_n) - u^* \rangle \\ & + \lambda_{n+1}^2 R_1^2 + (R_1\kappa + 1)^2 t_{n+1}^2 \|e_n\|^2 + 2R_1(R_1\kappa + 1)t_{n+1}\|e_n\| \\ \leq & \|u_n - u^*\|^2 - \lambda_{n+1}c_1 + \lambda_{n+1}^2 R_1^2 + (R_1\kappa + 1)^2 t_{n+1}^2 \|e_n\|^2 \\ & + 2R_1(R_1\kappa + 1)t_{n+1}\|e_n\|. \end{aligned}$$

Note that (C1)–(C4) ensure the existence of $N_2 \geq N_1$ such that

$$(\forall n \geq N_2) \quad \lambda_{n+1}^2 R_1^2 + (R_1\kappa + 1)^2 t_{n+1}^2 \|e_n\|^2 + 2R_1(R_1\kappa + 1)t_{n+1}\|e_n\| \leq \frac{1}{2}\lambda_{n+1}c_1.$$

Then it follows that

$$(\forall n \geq N_2) \quad \|u_{n+1} - u^*\|^2 - \|u_n - u^*\|^2 \leq -\frac{1}{2}\lambda_{n+1}c_1.$$

Therefore we have the contradiction: for sufficiently large $n \geq N_2$

$$\begin{aligned} & \|u_{n+1} - u^*\|^2 \\ = & \|u_{N_2} - u^*\|^2 + \sum_{i=N_2}^n (\|u_{i+1} - u^*\|^2 - \|u_i - u^*\|^2) \\ \leq & \|u_{N_2} - u^*\|^2 - \frac{1}{2}c_1 \sum_{i=N_2}^n \lambda_{i+1} < 0. \end{aligned}$$

This implies that (4.37) does not hold, and thus (4.31) is proved.

- (II) The second goal is to prove that: for any $\delta \in (0, \delta_0]$, there exists some $N_0 \geq 0$ satisfying

$$(4.40) \quad d(u_{N_0}, \Gamma) \leq 2\delta$$

and

$$(4.41) \quad (\forall n \geq N_0) \quad d(u_n, \Gamma) \leq 2\delta \Rightarrow d(u_{n+1}, \Gamma) \leq 2\delta,$$

which lead to (4.27) by induction.

(i) PROOF of (4.40): Fix $\delta \in (0, \delta_0]$ arbitrarily. Then the set $Fix(T) \cap \triangleright(\Gamma, \frac{1}{2}\delta) \neq \emptyset$ (see (4.28)) is compact, and thus the function $g(u, y) := \langle \mathcal{F}(u), u - y \rangle$ has its minimizer over $(Fix(T) \cap \triangleright(\Gamma, \frac{1}{2}\delta)) \times \Gamma \subset \mathcal{H} \times \mathcal{H}$ at some $(\tilde{u}, \tilde{y}) \in (Fix(T) \cap \triangleright(\Gamma, \frac{1}{2}\delta)) \times \Gamma$. Note that, by Fact 2.2,

$$c_2 := \inf_{\substack{u \in Fix(T) \cap \triangleright(\Gamma, \frac{1}{2}\delta) \\ y \in \Gamma}} g(u, y) = g(\tilde{u}, \tilde{y}) > 0.$$

Now, by Schwarz's inequality, (4.29) and the nonexpansivity of $T_{(n)}$ ($n \geq N$), we deduce for $\forall \eta \in (0, \frac{1}{2}\delta)$, $\forall (u, y) \in (\triangleleft(Fix(T), \eta) \cap \triangleright(\Gamma, \delta)) \times \Gamma$,

$$\begin{aligned} & \langle \mathcal{F}(T_{(n)}(u)), T_{(n)}(u) - y \rangle \\ = & \langle \mathcal{F}(P_{Fix(T)}(u)), P_{Fix(T)}(u) - y \rangle + \langle \mathcal{F}(P_{Fix(T)}(u)), T_{(n)}(u) - P_{Fix(T)}(u) \rangle \\ & + \langle \mathcal{F}(T_{(n)}(u)) - \mathcal{F}(P_{Fix(T)}(u)), T_{(n)}(u) - y \rangle \\ \geq & \langle \mathcal{F}(P_{Fix(T)}(u)), P_{Fix(T)}(u) - y \rangle - R_1 \|T_{(n)}(u) - P_{Fix(T)}(u)\| \\ & - \|\mathcal{F}(T_{(n)}(u)) - \mathcal{F}(P_{Fix(T)}(u))\| \|T_{(n)}(u) - y\| \\ \geq & \langle \mathcal{F}(P_{Fix(T)}(u)), P_{Fix(T)}(u) - y \rangle - R_1 \|T_{(n)}(u) - P_{Fix(T)}(u)\| \\ & - \kappa \|T_{(n)}(u) - P_{Fix(T)}(u)\| (\|T_{(n)}(u) - P_{Fix(T)}(u)\| + \|P_{Fix(T)}(u) - y\|) \\ \geq & \langle \mathcal{F}(P_{Fix(T)}(u)), P_{Fix(T)}(u) - y \rangle - R_1 \|u - P_{Fix(T)}(u)\| \\ & - \kappa \|u - P_{Fix(T)}(u)\| (\|u - P_{Fix(T)}(u)\| + R_1) \\ \geq & \langle \mathcal{F}(P_{Fix(T)}(u)), P_{Fix(T)}(u) - y \rangle - R_1\eta - \kappa\eta(\eta + R_1). \end{aligned}$$

Applying the simple facts

$$(\forall u \in \triangleleft(Fix(T), \eta) \cap \triangleright(\Gamma, \delta)) \quad P_{Fix(T)}(u) \in Fix(T) \cap \triangleright(\Gamma, \delta - \eta)$$

and

$$(\triangleleft(Fix(T), \eta) \cap \triangleright(\Gamma, \delta)) \supset (Fix(T) \cap \triangleright(\Gamma, \delta)) \neq \emptyset,$$

to the above inequality, we deduce for all $\eta \in (0, \frac{1}{2}\delta)$

$$\begin{aligned} (4.42) \quad c_3(\eta) &:= \inf_{\substack{u \in \triangleleft(Fix(T), \eta) \cap \triangleright(\Gamma, \delta) \\ y \in \Gamma}} \langle \mathcal{F}(T_{(n)}(u)), T_{(n)}(u) - y \rangle \\ &\geq -R_1\eta - \kappa\eta(\eta + R_1) \\ &\quad + \inf_{\substack{u \in \triangleleft(Fix(T), \eta) \cap \triangleright(\Gamma, \delta) \\ y \in \Gamma}} \langle \mathcal{F}(P_{Fix(T)}(u)), P_{Fix(T)}(u) - y \rangle \\ &\geq -R_1\eta - \kappa\eta(\eta + R_1) + \inf_{\substack{u \in Fix(T) \cap \triangleright(\Gamma, \delta - \eta) \\ y \in \Gamma}} \langle \mathcal{F}(u), u - y \rangle \\ &\geq -R_1\eta - \kappa\eta(\eta + R_1) + \inf_{\substack{u \in Fix(T) \cap \triangleright(\Gamma, \frac{1}{2}\delta) \\ y \in \Gamma}} g(u, y) \\ &= c_2 - R_1\eta - \kappa\eta(\eta + R_1), \end{aligned}$$

which implies that we can choose sufficiently small $\eta_1 \in (0, \frac{1}{2}\delta)$ such that

$$c_3(\eta_1) \geq c_2 - R_1\eta_1 - \kappa\eta_1(\eta_1 + R_1) > 0.$$

Furthermore, by (C1)–(C3) and Theorem 4.6, there exists some $N_3 > N_2$ such that

$$(4.43) \quad (\forall n \geq N_3) \quad \begin{aligned} \frac{\|u_{n+1} - u_n\|}{d(u_n, Fix(T))} &\leq \delta \\ \frac{2\lambda_{n+1}c_3(\eta_1)}{2\lambda_{n+1}c_3(\eta_1)} &\geq \frac{\eta_1}{\lambda_{n+1}^2 R_1^2 + (R_1\kappa + 1)^2 t_{n+1}^2 \|e_n\|^2} \\ &\quad + 2R_1(R_1\kappa + 1)t_{n+1}\|e_n\| \end{aligned}$$

On the other hand, by noting (4.31), for any $u^* \in \Gamma$, there exists some subsequence $(T_{(n_i)}(u_{n_i}))_{i \geq 0}$, of $(T_{(n)}(u_n))_{n \geq 0}$, such that

$$(4.44) \quad \lim_{i \rightarrow \infty} \langle \mathcal{F}(T_{(n_i)}(u_{n_i})), T_{(n_i)}(u_{n_i}) - u^* \rangle = 0.$$

In addition, the boundedness of $(T_{(n_i)}(u_{n_i}))_{i \geq 0}$ implies the existence of the further subsequence $(T_{(n_{i_j})}(u_{n_{i_j}}))_{j \geq 0}$ that converges to some $\hat{u} \in \mathcal{H}$. Moreover, by the nonexpansivity of T and $T_{(n)}$, it follows that

$$\begin{aligned} &\left\| T_{(n_{i_j})}(u_{n_{i_j}}) - T(T_{(n_{i_j})}(u_{n_{i_j}})) \right\| \\ \leq &\left\| T_{(n_{i_j})}(u_{n_{i_j}}) - P_{Fix(T)}(u_{n_{i_j}}) \right\| + \left\| P_{Fix(T)}(u_{n_{i_j}}) - T(T_{(n_{i_j})}(u_{n_{i_j}})) \right\| \\ \leq &\left\| u_{n_{i_j}} - P_{Fix(T)}(u_{n_{i_j}}) \right\| + \left\| P_{Fix(T)}(u_{n_{i_j}}) - u_{n_{i_j}} \right\| = 2d(u_{n_{i_j}}, Fix(T)), \end{aligned}$$

and thus, by Theorem 4.6,

$$\lim_{j \rightarrow \infty} \left\| T_{(n_{i_j})}(u_{n_{i_j}}) - T(T_{(n_{i_j})}(u_{n_{i_j}})) \right\| = 0.$$

Application of the Opial's Demiclosedness Principle [100] to the above subsequence $(T_{(n_{i_j})}(u_{n_{i_j}}))_{j \geq 0}$ yields $\hat{u} = \lim_{j \rightarrow \infty} T_{(n_{i_j})}(u_{n_{i_j}}) \in Fix(T)$.

Furthermore, by the continuities of the inner product and \mathcal{F} , we obtain

$$\langle \mathcal{F}(\hat{u}), \hat{u} - u^* \rangle = \lim_{j \rightarrow \infty} \langle \mathcal{F}(T_{(n_{i_j})}(u_{n_{i_j}})), T_{(n_{i_j})}(u_{n_{i_j}}) - u^* \rangle = 0,$$

which implies $\hat{u} \in \Gamma$ by Fact 2.2(b). Finally, Theorem 4.6 yields

$$\lim_{j \rightarrow \infty} u_{n_{i_j}} = \lim_{j \rightarrow \infty} \left\{ T_{(n_{i_j})}(u_{n_{i_j}}) + (u_{n_{i_j}} - T_{(n_{i_j})}(u_{n_{i_j}})) \right\} = \hat{u} \in \Gamma,$$

hence (4.40) holds for some $N_0 \geq N_3$.

(ii) PROOF of (4.41): In the following, as the final step, we will prove that (4.41) holds for such N_0 .

Fix $n \geq N_0$ arbitrarily. By the definition of $d(\cdot, \Gamma)$ and (4.43), it follows

$$\begin{aligned} d(u_{n+1}, \Gamma) - d(u_n, \Gamma) &\leq \|u_{n+1} - P_\Gamma(u_n)\| - \|u_n - P_\Gamma(u_n)\| \\ &\leq \|u_{n+1} - u_n\| \leq \delta \end{aligned}$$

hence

$$(4.45) \quad d(u_n, \Gamma) \leq \delta \Rightarrow d(u_{n+1}, \Gamma) \leq 2\delta.$$

Moreover, by (4.42) and (4.43), if $d(u_n, \Gamma) \geq \delta$, it follows

$$(4.46) \quad \langle \mathcal{F}(T_{(n)}(u_n)), T_{(n)}(u_n) - P_\Gamma(u_n) \rangle \geq c_3(\eta_1).$$

Furthermore, by a deduction similar to (4.39), by using (4.43) and (4.46), we obtain, for $d(u_n, \Gamma) \geq \delta$,

$$\begin{aligned}
& d^2(u_{n+1}, \Gamma) - d^2(u_n, \Gamma) \\
&= \|u_{n+1} - P_\Gamma(u_{n+1})\|^2 - \|u_n - P_\Gamma(u_n)\|^2 \\
&\leq \|u_{n+1} - P_\Gamma(u_n)\|^2 - \|u_n - P_\Gamma(u_n)\|^2 \\
&\quad \vdots \\
&\leq -2\lambda_{n+1} \langle \mathcal{F}(T_{(n)}(u_n)), T_{(n)}(u_n) - P_\Gamma(u_n) \rangle \\
&\quad + \lambda_{n+1}^2 R_1^2 + (R_1 \kappa + 1)^2 t_{n+1}^2 \|e_n\|^2 + 2R_1(R_1 \kappa + 1)t_{n+1}\|e_n\| \\
&\leq -2\lambda_{n+1} c_3(\eta_1) \\
&\quad + \lambda_{n+1}^2 R_1^2 + (R_1 \kappa + 1)^2 t_{n+1}^2 \|e_n\|^2 + 2R_1(R_1 \kappa + 1)t_{n+1}\|e_n\| \leq 0,
\end{aligned}$$

which implies

$$(4.47) \quad \delta \leq d(u_n, \Gamma) \leq 2\delta \Rightarrow d(u_{n+1}, \Gamma) \leq 2\delta.$$

The statements (4.45) and (4.47) yield (4.41), which completes the proof.
(Q.E.D.)

Combining Theorems 4.5 and 4.7 immediately yields the following main results showing the numerical robustness of the iterative formula (1.10) and thus its sound applicability to the possibly ill-posed problems like convexly constrained inverse problems.

THEOREM 4.8. *Let $T : \mathcal{H} \rightarrow \mathcal{H}$ be a nonexpansive mapping with bounded $\text{Fix}(T) \neq \emptyset$. Suppose that $T, (\lambda_n, t_n)_{n \geq 1} \subset [0, \infty) \times [0, \infty)$ and $(e_n)_{n \geq 0} \subset \mathcal{H}$ satisfy the conditions (C1)–(C5). Suppose also that $\mathcal{F} : \mathcal{H} \rightarrow \mathcal{H}$ is paramonotone over $\text{Fix}(T)$ and κ -Lipschitzian over $\Omega_N \subset \mathcal{H}$, where N suffices (C1). Let $\Gamma := \{u^* \in \text{Fix}(T) \mid \langle u - u^*, \mathcal{F}(u^*) \rangle \geq 0 \text{ for all } u \in \text{Fix}(T)\} \neq \emptyset$. [Note: $\Gamma \neq \emptyset$ automatically holds. See the comment in Theorem 2.18.] Then the sequence $(u_n)_{n \geq 0} \subset \mathcal{H}$ generated by (4.1), with arbitrary $u_0 \in \mathcal{H}$, satisfies*

$$\lim_{n \rightarrow \infty} d(u_n, \Gamma) = 0$$

if the following condition (a) or (b) is fulfilled.

- (a) (i) \mathcal{F} is monotone over Ω_N , (ii) $(\lambda_n)_{n \geq 1}$ satisfies (C6),
- (b) T is asymptotically shrinking (In this case, the nonemptiness and boundedness of $\text{Fix}(T)$ automatically holds [see Proposition 2.14]).

REMARK 4.9.

- (a) Remind that $(e_n)_{n \geq 0}$ satisfying (C1)–(C5) is not necessarily absolute summable; i.e., possibly $(\|e_n\|)_{n \geq 0} \notin l^1$ (See Example 4.2(b)).
- (b) Let $T_{(n)} = T$, $(t_{n+1}, e_n) := (1, \mathbf{0})$ for all $n \geq 0$, and $(\lambda_n)_{n \geq 1}$ satisfy $\{(W1), (W2)\}$ or $\{(W1), (W2), (C6)\}$. Then Theorem 2.18 is reproduced from Theorem 4.8 because $\Omega_N = T(\mathcal{H})$ holds.

The following corollaries, on the minimizations of the quadratic function $\Theta : \mathcal{H} \rightarrow \mathbb{R}$ and the proximity function $\Phi : \mathcal{H} \rightarrow \mathbb{R}$ in (2.4), over $\text{Fix}(T) \neq \emptyset$, are particularly useful for the convexly constrained inverse problems [see for example (1.1)–(1.7), Example 2.6, (2.4) and the discussions in Section 3]. In both cases, the conditions on $\mathcal{F} := \Theta'$ in Theorem 4.8 are automatically fulfilled.

COROLLARY 4.10. Let $T : \mathcal{H} \rightarrow \mathcal{H}$ be a nonexpansive mapping with bounded $\text{Fix}(T) \neq \emptyset$. Define, for any $r \in \mathcal{H}$, a function $\Theta : \mathcal{H} \rightarrow \mathbb{R}$ by

$$\Theta(x) := \frac{1}{2}\langle Q(x), x \rangle - \langle r, x \rangle, \quad x \in \mathcal{H},$$

where $Q : \mathcal{H} \rightarrow \mathcal{H}$ is a self-adjoint bounded linear operator satisfying

$$\langle Q(x), x \rangle \geq 0, \quad x \in \mathcal{H}.$$

Suppose that $T, (\lambda_n, t_n)_{n \geq 1} \subset [0, \infty) \times [0, \infty)$ and $(e_n)_{n \geq 0} \subset \mathcal{H}$ satisfy the conditions (C1)–(C5). Assume that (i) $(\lambda_n)_{n \geq 1}$ satisfies (C6), or (ii) T is asymptotically shrinking. Then the sequence $(u_n)_{n \geq 0} \subset \mathcal{H}$ generated, with arbitrary $u_0 \in \mathcal{H}$, by

$$\begin{aligned} u_{n+1} &:= (1 - t_{n+1})u_n + t_{n+1}(T(u_n) + e_n) \\ &\quad - \lambda_{n+1}\Theta'((1 - t_{n+1})u_n + t_{n+1}(T(u_n) + e_n)) \end{aligned}$$

satisfies

$$\lim_{n \rightarrow \infty} d(u_n, \Gamma) = 0,$$

where $\Theta'(x) = Qx - r$ and $\Gamma := \{u^* \in \text{Fix}(T) \mid \Theta(u^*) = \inf \Theta(\text{Fix}(T))\} \neq \emptyset$.

COROLLARY 4.11. Let $T : \mathcal{H} \rightarrow \mathcal{H}$ be a nonexpansive mapping with bounded $\text{Fix}(T) \neq \emptyset$. Let $\Phi : \mathcal{H} \rightarrow \mathbb{R}$ be a proximity function defined in (2.4). Suppose that $T, (\lambda_n, t_n)_{n \geq 1} \subset [0, \infty) \times [0, \infty)$ and $(e_n)_{n \geq 0} \subset \mathcal{H}$ satisfy the conditions (C1)–(C5). Assume that (i) $(\lambda_n)_{n \geq 1}$ satisfies (C6), or (ii) T is asymptotically shrinking. Then the sequence $(u_n)_{n \geq 0} \subset \mathcal{H}$ generated, with arbitrary $u_0 \in \mathcal{H}$, by

$$\begin{aligned} u_{n+1} &:= (1 - t_{n+1})u_n + t_{n+1}(T(u_n) + e_n) \\ &\quad - \lambda_{n+1}\Phi'((1 - t_{n+1})u_n + t_{n+1}(T(u_n) + e_n)) \end{aligned}$$

satisfies

$$\lim_{n \rightarrow \infty} d(u_n, \Gamma) = 0,$$

where $\Phi' = I - \sum_{i=1}^m w_i P_{C_i}$ (see Section 2.A.) and $\Gamma := \{u^* \in \text{Fix}(T) \mid \Phi(u^*) = \inf \Phi(\text{Fix}(T))\} \neq \emptyset$.

5. Concluding Remarks

In this paper, we show that broad range of convexly constrained generalized inverse problems can be regarded as the convex optimization problems defined over the fixed point set of nonexpansive mappings, and many algorithmic solutions to the inverse problems are derived in a unified way based on the hybrid steepest descent method. Next we present a variation of the hybrid steepest descent method for the variational inequality problem over the fixed point set of a nonexpansive mapping. The variation uses a slowly changing sequence of nonexpansive mappings having same fixed point sets, and is gifted with certain notable robustness to the numerical errors possibly unavoidable in the iterative computations. The proposed method realizes sound mathematical algorithms for wide range of the convexly constrained inverse problems. The application of the hybrid steepest descent method is quite broad and of course not restricted in the standard inverse problems. One of its important applications would be in the MPEC (mathematical program with equilibrium constraints)[89], which will be discussed elsewhere.

References

- [1] R. Aharoni and Y. Censor, Block-iterative projection methods for parallel computation of solutions to convex feasibility problems, *Linear Algebra and Its Applications* **120** (1989) 165-175.
- [2] T.M. Apostol, *Mathematical Analysis* (2nd ed.), (Addison-Wesley, 1974).
- [3] J.-P. Aubin, *Optima and Equilibria — An Introduction to Nonlinear Analysis*, (Springer-Verlag, 1993).
- [4] C. Sánchez-Avila, An adaptive regularized method for deconvolution of signal with edges by convex projections, *IEEE Trans. Signal Processing* **42** (1994) 1849-1851.
- [5] J.-B. Baillon, R. E. Bruck and S. Reich, On the asymptotic behavior of nonexpansive mappings and semigroups in Banach spaces, *Houston J. Math.* **4** (1978) 1-9.
- [6] J.-B. Baillon and G. Haddad, Quelques propriétés des opérateurs angle-bornés et n -cycliquement monotones, *Israel J. Math.* **26** (1977) 137-150.
- [7] V. Barbu and Th. Precupanu, *Convexity and Optimization in Banach Spaces*, 3rd ed., (D. Reidel Publishing Company, 1986).
- [8] H.H. Bauschke and J.M. Borwein, Dykstra's alternating projection algorithm for two sets, *J. Approx. Theory* **79** (1994) 418-443.
- [9] H.H. Bauschke and J.M. Borwein, On projection algorithms for solving convex feasibility problems, *SIAM Review* **38** (1996) 367-426.
- [10] H.H. Bauschke, The approximation of fixed points of compositions of nonexpansive mappings in Hilbert space, *J. Math. Anal. Appl.* **202** (1996) 150-159.
- [11] H.H. Bauschke, J.M. Borwein and A.S. Lewis, The method of cyclic projections for closed convex sets in Hilbert space, *Contemp. Math.* **204** (1997) 1-38.
- [12] H.H. Bauschke and A.S. Lewis, Dykstra's algorithm with Bregman projections: a convergence proof, *Optimization* **48** (2000) 409-427.
- [13] M. Bertero and P. Boccacci, *Introduction to Inverse Problems in Imaging* (IOP, 1998).
- [14] J.P. Boyle and R.L. Dykstra, A method for finding projections onto the intersection of convex sets in Hilbert spaces, *Advances in Order Restricted Statistical Inference*, Lecture Notes in Statistics (Springer-Verlag, 1985) 28-47.
- [15] L.M. Bregman, The method of successive projection for finding a common point of convex sets, *Soviet Math. Dokl.* **6** (1965) 688-692.
- [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** (1967) 200-217.
- [17] L. M. Bregman, Y. Censor and S. Reich, Dykstra's algorithm as the nonlinear extension of Bregman's optimization method, *J. Convex Analysis* **6** (1999) 319-333.
- [18] H. Brezis, *Analyse Fonctionnelle*, (Masson, Editeur, Paris, 1983).
- [19] F.E. Browder, Nonexpansive nonlinear operators in Banach space, *Proc. Nat. Acad. Sci. USA* **54** (1965) 1041-1044.
- [20] F.E. Browder, Convergence of approximants to fixed points of nonexpansive nonlinear mappings in Banach spaces, *Arch. Rat. Mech. Anal.* **24** (1967) 82-90.
- [21] F.E. Browder and W.V. Petryshyn, Construction of fixed points of nonlinear mappings in Hilbert space, *J. Math. Anal. Appl.* **20** (1967) 197-228.
- [22] R. E. Bruck and S. Reich, Nonexpansive projections and resolvents of accretive operators in Banach spaces, *Houston J. Math.* **3** (1977) 459-470.
- [23] D. Butnariu and Y. Censor, On the behavior of a block-iterative projection method for solving convex feasibility problems, *International Journal of Computer Mathematics* **34** (1990) 79-94.
- [24] D. Butnariu and Y. Censor, Strong convergence of almost simultaneous block-iterative projection methods in Hilbert spaces, *Journal of Computational and Applied Mathematics* **53** (1994) 33-42.
- [25] D. Butnariu, Y. Censor and S. Reich, Iterative averaging of entropic projections for solving stochastic convex feasibility, *Computational Optimization and Applications* **8** (1997) 21-39.
- [26] D. Butnariu and A.N. Iusem, *Totally Convex Functions for fixed point computation and infinite dimensional optimization* (Kluwer Academic Publishers, 2000).
- [27] C.L. Byrne, Iterative projection onto convex sets using multiple Bregman distances, *Inverse Problems* **15** (1999) 1295-1313.

- [28] C.L. Byrne and Y. Censor, Proximity function minimization for separable, jointly convex Bregman distances, with applications, Technical Report (1998)
- [29] C.L. Byrne and Y. Censor, Proximity function minimization using multiple Bregman projections, with applications to split feasibility and Kullback-Leibler distance minimization, Technical Report (2000)
- [30] Y. Censor, Row-action methods for huge and sparse systems and their applications, *SIAM Review* **23** (1981) 444-464.
- [31] Y. Censor and A. Lent, An iterative row-action method for interval convex programming, *Journal of Optimization Theory and Applications* **34** (1981) 321-353.
- [32] Y. Censor, Parallel application of block-iterative methods in medical imaging and radiation therapy, *Math. Programming* **42** (1988) 307-325.
- [33] Y. Censor and T. Elfving, A multiprojection algorithm using Bregman projections in a product space, *Numerical Algorithms* **8** (1994) 221-239.
- [34] Y. Censor and S.A. Zenios, *Parallel Optimization: Theory, Algorithm, and Optimization* (Oxford University Press, 1997).
- [35] Y. Censor and S. Reich, The Dykstra algorithm with Bregman projections, *Communications in Applied Analysis* **2** (1998) 407-419.
- [36] Y. Censor, A.N. Iusem and S.A. Zenios, An interior point method with Bregman functions for the variational inequality problem with paramonotone operators, *Math. Programming* **81** (1998) 373-400.
- [37] W. Cheney and A.A. Goldstein, Proximity maps for convex sets, *Proc. Amer. Math. Soc.* **10** (1959) 448-450.
- [38] C.K. Chui, F. Deutsch and J.W. Ward, Constrained best approximation in Hilbert space, *Constr. Approx.* **6** (1990) 35-64.
- [39] C.K. Chui, F. Deutsch, and J.W. Ward, Constrained best approximation in Hilbert space II, *J. Approx. Theory*, **71** (1992) 213-238.
- [40] D. Colton, H.W. Engl, A.K. Louis, J.R. McLaughlin and W. Rundell (eds.), *Surveys on Solution Methods for Inverse Problems* (Springer, 2000).
- [41] P.L. Combettes, Foundation of set theoretic estimation, *Proc. IEEE* **81** (1993) 182-208.
- [42] P.L. Combettes, Inconsistent signal feasibility problems: least squares solutions in a product space, *IEEE Trans. on Signal Processing* **42** (1994) 2955-2966.
- [43] P.L. Combettes, Construction d'un point fixe commun à une famille de contractions fermes, *C.R. Acad. Sci. Paris Sér. I Math.* **320** (1995) 1385-1390.
- [44] P.L. Combettes, The convex feasibility problem in image recovery, in *Advances in Imaging and Electron Physics*, P. Hawkes, Ed., New York: Academic, **95** (1996) 155-270.
- [45] P.L. Combettes, Convex set theoretic image recovery by extrapolated iterations of parallel subgradient projections, *IEEE Trans. Image Processing*, **6** (1997) 493-506.
- [46] P.L. Combettes and P. Bondon, Hard-constrained inconsistent signal feasibility problems, *IEEE Trans. Signal Processing* **47** (1999) 2460-2468.
- [47] P.L. Combettes, A parallel constraint disintegration and approximation scheme for quadratic signal recovery, *Proc. of the 2000 IEEE International Conference on Acoustics, Speech and Signal Processing* (2000) 165-168.
- [48] P.L. Combettes, Strong convergence of block-iterative outer approximation methods for convex optimization, *SIAM J. Control Optim.* **38** (2000) 538-565.
- [49] P.L. Combettes, On the numerical robustness of the parallel projection method in signal synthesis, *IEEE Signal Processing Letters* **8** (2001) 45-47.
- [50] G. Crombez, Finding projections onto the intersection of convex sets in Hilbert spaces, *Numer. Funct. Anal. Optim.* **15** (1996) 637-652.
- [51] F. Deutsch and H. Hundal, The rate of convergence of Dykstra's cyclic projections algorithms: the polyhedral case, *Numer. Funct. Anal. Optim.* **15** (1994) 537-565.
- [52] F. Deutsch and H. Hundal, The rate of convergence for the method of alternating projections II, *J. Math. Anal. Appl.* **205** (1997) 381-405.
- [53] F. Deutsch, *Best Approximation in Inner Product Spaces*, (Springer, 2001).
- [54] F. Deutsch and I. Yamada, Minimizing certain convex functions over the intersection of the fixed point sets of nonexpansive mappings, *Numer. Funct. Anal. Optim.* **19** (1998) 33-56.
- [55] Z.O. Dolidze, Solutions of variational inequalities associated with a class of monotone maps, *Ekonomika i Matem. Metody* **18** (1982) 925-927.
- [56] W.G. Dotson, On the Mann iterative process, *Trans. Amer. Math. Soc.* **149** (1970) 65-73.

- [57] N. Dunford and J.T Schwartz, *Linear Operators Part I: General Theory, Wiley Classics Library Edition*, (John Wiley & Sons, 1988).
- [58] J.C. Dunn, Convexity, Monotonicity, and Gradient Processes, *J. Math. Anal. Appl.* **53** (1976) 145-158.
- [59] R.L. Dykstra, An algorithm for restricted least squares regression, *J. Amer. Statist. Assoc.* **78** (1983) 837-842.
- [60] B. Eicke, Iteration methods for convexly constrained ill-posed problems in Hilbert space, *Numer. Funct. Anal. Optim.* **13** (1992) 413-429.
- [61] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems, Classics in Applied Mathematics 28* (SIAM, 1999).
- [62] U.M. García-Palomares, Parallel projected aggregation methods for solving the convex feasibility problem, *SIAM J. Optim.* **3** (1993) 882-900.
- [63] R.W Gerchberg, Super-restoration through error energy reduction, *Optica Acta* **21** (1974) 709-720.
- [64] K. Goebel and S. Reich, *Uniform Convexity, Hyperbolic Geometry, and Nonexpansive Mappings* (Dekker, New York and Basel, 1984).
- [65] K. Goebel and W.A. Kirk, *Topics in Metric Fixed Point Theory* (Cambridge Univ. Press, 1990).
- [66] M. Goldburg and R. J. Marks II, Signal Synthesis in the Presence of an Inconsistent Set of Constraints, *IEEE Trans. on Circuits and Systems* **32** (1985), 647-663.
- [67] A.A. Goldstein, Convex programming in Hilbert space, *Bull. Amer. Math. Soc.* **70** (1964) 709-710.
- [68] E.G. Golshtain and N.V. Tretyakov, *Modified Lagrangians & Monotone Maps in Optimization* (Wiley and Sons, 1996).
- [69] G.H. Golub and C.F. Van Loan, *Matrix computations 3rd ed.* (The Johns Hopkins University Press, 1996).
- [70] C.W. Groetsch, *Inverse Problems in Mathematical Sciences* (Wiesbaden-Vieweg, 1993).
- [71] L.G. Gubin, B.T. Polyak, and E.V. Raik, The method of projections for finding the common point of convex sets, *USSR Computational Mathematics and Mathematical Physics* **7** (1967) 1-24.
- [72] J. Hadamard, *Lectures on Cauchy's Problem in Linear Partial Differential Equations* (Yale Univ. Press, 1923).
- [73] B. Halpern, Fixed points of nonexpanding maps, *Bull. Amer. Math. Soc.* **73** (1967) 957-961.
- [74] S.P. Han, A successive projection method, *Math. Programming* **40** (1988) 1-14.
- [75] J.C. Harsanyu and C.I. Chang, Hyperspectral image classification and dimensionality reduction: an orthogonal subspace projection approach, *IEEE Trans. Geosci. Remote Sensing* **32** (1994) 779-785.
- [76] M.H. Hayes, M.H. Lim, and A.V. Oppenheim, Signal reconstruction from phase or magnitude, *IEEE Trans. Acoust., Speech, Signal Processing* **28** (1980) 672-680.
- [77] H. Hundal and F. Deutsch, Two generalizations of Dykstra's cyclic projections algorithm, *Math. Programming* **77** (1997) 335-355.
- [78] A.N. Iusem and A.R. De Pierro, Convergence results for an accelerated nonlinear Cimmino algorithm, *Numerische Mathematik* **49** (1986) 367-378.
- [79] A.N. Iusem and A.R. De Pierro, On the convergence of Han's method for convex programming with quadratic objective, *Math. Programming* **52** (1991) 265-284.
- [80] A.N. Iusem, An iterative algorithm for the variational inequality problem, *Comp. Appl. Math.*, **13** (1994) 103-114.
- [81] K.C. Kiwiel, Free-steering relaxation methods for problems with strictly convex costs and linear constraints, *Mathematics of Operations Research* **22** (1983) 326-349.
- [82] D.P. Kolba and T.W. Parks, Optimal estimation for band-limited signals including time domain consideration, *IEEE Trans. Acoust., Speech, Signal Processing* **31** (1983) 113-122.
- [83] G.M. Korolevich, The extragradient method for finding saddle points and other problems, *Ekonomika i matematicheskie metody* **12** (1976) 747-756.
- [84] E. Kreyszig *Introductory Functional Analysis with Applications, Wiley Classics Library Edition*, (John Wiley & Sons, 1989).
- [85] E.S. Levitin and B.T. Polyak, Constrained Minimization Method, *USSR Computational Mathematics and Mathematical Physics* **6** (1966) 1-50.

- [86] P.L. Lions, Approximation de points fixes de contractions, *C. R. Acad. Sci. Paris Série A-B* **284** (1977) 1357-1359.
- [87] F. Liu and M.Z. Nashed, Regularization of Nonlinear Ill-Posed Variational Inequalities and Convergence Rates, *Set-Valued Analysis* **6** (1998) 313-344.
- [88] D.G. Luenberger, *Optimization by Vector Space Methods* (Wiley, 1968).
- [89] Z.-Q. Luo, J.-S. Pang and D. Ralph, *Mathematical Programs with Equilibrium Constraints* (Cambridge University Press, 1996).
- [90] W.R. Mann, Mean value methods in iteration, *Proc. Amer. Math. Soc.* **4** (1953) 506-510.
- [91] S. Maruster, The solution by iteration of nonlinear equations in Hilbert spaces, *Proc. Amer. Math. Soc.* **63** (1977) 69-73.
- [92] C.A. Micchelli, P.W. Smith, J. Swetits, and J.W. Ward, Constrained L_p Approximation, *Constr. Approx.* **1** (1985) 93-102.
- [93] C.A. Micchelli and F. Utreras, Smoothing and interpolation in a convex set in a Hilbert space, *SIAM J. Sci. Statist. Comput.* **9** (1988) 728-746.
- [94] M.Z. Nashed, *Generalized Inverse and Applications* (Academic Press, 1976).
- [95] J.von Neumann, Functional operators Vol.II. The Geometry of Orthogonal Spaces, *Annals of Math. Studies* **22** (Princeton Univ. Press, 1950) [Reprint of mimeographed lecture notes first distributed in 1933].
- [96] N. Ogura and I. Yamada, Non-strictly convex minimization over the fixed point set of the asymptotically shrinking nonexpansive mapping, *Numer. Funct. Anal. Optim.* **23** (2002) 113-137.
- [97] N. Ogura and I. Yamada, Nonstrictly convex minimization over the bounded fixed point set of nonexpansive mapping, *submitted for publication* (2001).
- [98] N. Ogura and I. Yamada, The multi-layered hard constrained convex feasibility problem, *The 2nd International Conference on Nonlinear Analysis and Convex Analysis*, Hirosaki, (2001).
- [99] S. Oh, R.J. Marks II, and L.E. Atlas, Kernel synthesis for generalized time frequency distribution using the method of alternating projections onto convex sets, *IEEE Trans. Signal Processing* **42** (1994) 1653-1661.
- [100] Z. Opial, Weak convergence of the sequence of successive approximations for nonexpansive mapping, *Bull. Amer. Math. Soc.*, **73** (1967) 591-597.
- [101] A Papoulis, A new algorithm in spectral analysis and band limited extrapolation, *IEEE Trans. Circuits and Syst.* **22** (1975) 735-742.
- [102] G. Pierra, Eclatement de contraintes en parallèle pour la minimisation d'une forme quadratique, *Lecture Notes Computer Sci.* **40** (1976) 200-218.
- [103] G. Pierra, Decomposition through formalization in a product space, *Math. Programming* **28** (1984) 96-115.
- [104] L.C. Potter and K.S. Arun, Energy concentration in band-limited extrapolation, *IEEE Trans. Acoust., Speech, Signal Processing*, **37** (1989) 1027-1041.
- [105] L.C. Potter and K.S. Arun, A dual approach to linear problems with convex constraints, *SIAM J. Control and Optimization*, **31** (1993) 1080-1092.
- [106] S. Reich, Weak convergence theorems for nonexpansive mappings in Banach spaces, *J. Math. Anal. Appl.* **67** (1979) 274-292.
- [107] S. Reich, Some problems and results in fixed point theory, *Contemp. Math.* **21** (1983) 179-187.
- [108] S. Reich, A limit theorem for projections, *Linear and Multilinear Algebra* **13** (1983) 281-290.
- [109] T.R. Rockafellar, Convex Analysis, (Princeton University Press, 1970).
- [110] A. Sabharwal and L.C. Potter, Some results on least-squares and regularization for convexly constrained linear inverse problems; *IPS Laboratory Technical Report TR-96-02*, The Ohio State University, (1996).
- [111] A. Sabharwal and L.C. Potter, Convexly constrained linear inverse problems: Iterative least-squares and regularization, *IEEE Trans. on Signal Processing*, **46** (1998) 2345-2352.
- [112] J.L.C. Sanz and T.S. Huang, Continuation techniques for a certain class of analytic functions, *SIAM J. Appl. Math.*, **44** (1984) 819-838.
- [113] J.L.C. Sanz and T.S. Huang, A unified approach to noniterative linear signal restoration, *IEEE Trans. Acoust., Speech, Signal Processing* **32** (1984) 403-409.
- [114] H.F. Senter and W.G. Dotson,Jr., Approximating fixed points of nonexpansive mappings, *Proc. Amer. Math. Soc.* **44** (1974) 55-67.

- [115] J.J. Settle and N.A. Drake, Linear mixing and the estimation of ground cover proportions, *Int. J. Remote Sensing* **14** (1993) 1159-1177.
- [116] J.J. Settle and N. Campbell, On the error of two estimators of sub-pixel fractional cover when mixing is linear, *IEEE Trans. Geosci. Remote Sensing* **36** (1998) 163-170.
- [117] Y.S. Shin and Z.H. Cho, SVD pseudoinversion image reconstruction, *IEEE Trans. Acoust., Speech, Signal Processing* **29** (1989) 904-909.
- [118] B. Sirkci, D. Brady and J. Burman, Restricted total least squares solutions for hyperspectral imagery, *Proceedings of 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, **1** (2000) 624-627.
- [119] H. Stark and Y. Yang, *Vector Space Projections - A Numerical Approach to Signal and Image Processing, Neural Nets, and Optics* (John Wiley & Sons, 1998).
- [120] W. Takahashi, Fixed point theorems and nonlinear ergodic theorem for a nonexpansive semigroups and their applications, *Nonlinear Analysis* **30** (1997) 1283-1293.
- [121] W. Takahashi, *Nonlinear Functional Analysis— Fixed Point Theory and its Applications* (Yokohama Publishers, 2000).
- [122] K. Tanabe, Projection method for solving a singular system of linear equations and its applications, *Numerische Mathematik* **17** (1971) 203-214.
- [123] A.N. Tikhonov, Solution of incorrectly formulated problems and the regularization method, *Sov. Doklady* **4** (1963) 1035-1038.
- [124] A.N. Tikhonov and V.Y. Arsenin, *Solutions of Ill-posed problems* (V.H. Winston & Sons, 1977).
- [125] P. Tseng, On the convergence of the products of firmly nonexpansive mappings, *SIAM J. Optim.* **2** (1992) 425-434.
- [126] P. Tseng, Dual coordinate ascent methods for non-strictly convex minimization, *Math. Programming* **59** (1993) 231-247.
- [127] R. Wittmann, Approximation of fixed points of nonexpansive mappings, *Arch. Math.* **58** (1992) 486-491.
- [128] P. Wolfe, Finding the nearest point in a polytope, *Math. Programming* **11** (1976) 128-149.
- [129] I. Yamada, N. Ogura, Y. Yamashita and K. Sakaniwa, An extension of optimal fixed point theorem for nonexpansive operator and its application to set theoretic signal estimation, *Technical Report of IEICE, DSP96-106* (1996) 63-70.
- [130] I. Yamada, N. Ogura, Y. Yamashita and K. Sakaniwa, Quadratic optimization of fixed points of nonexpansive mappings in Hilbert space, *Numer. Funct. Anal. Optim.* **19** (1998) 165-190.
- [131] I. Yamada, Approximation of convexly constrained pseudoinverse by Hybrid Steepest Descent Method, *Proceedings of 1999 IEEE International Symposium on Circuits and Systems*, **5** (1999) 37-40.
- [132] I. Yamada, Convex projection algorithm from POCS to Hybrid steepest descent method (in Japanese), *Journal of the IEICE*, **83** (2000).
- [133] I. Yamada, The hybrid steepest descent method for the variational inequality problem over the intersection of fixed point sets of nonexpansive mappings, in *Inherently Parallel Algorithm for Feasibility and Optimization*, (D. Butnariu, Y. Censor, and S. Reich, Eds.) Elsevier, 2001.
- [134] K. Yosida, *Functional analysis*, 4th ed. (Springer Verlag, 1974).
- [135] D.C. Youla, Generalized image restoration by the method of alternating orthogonal projections, *IEEE Trans. Circuits and Syst.* **25** (1978) 694-702.
- [136] D.C. Youla and H. Webb, Image restoration by the method of convex projections: Part 1- Theory, *IEEE Trans. Medical Imaging* **1** (1982) 81-94.
- [137] E.H Zarantonello, ed., *Contributions to Nonlinear Functional Analysis* (Academic Press, 1971).
- [138] E. Zeidler, *Nonlinear Functional Analysis and its Applications, I - Fixed Point Theorems* (Springer-Verlag, 1986).
- [139] E. Zeidler, *Nonlinear Functional Analysis and its Applications, II/A - Linear Monotone Operators* (Springer-Verlag, 1990).
- [140] E. Zeidler, *Nonlinear Functional Analysis and its Applications, II/B - Nonlinear Monotone Operators* (Springer-Verlag, 1990).
- [141] E. Zeidler, *Nonlinear Functional Analysis and its Applications, III - Variational Methods and Optimization* (Springer-Verlag, 1985).

- [142] E. Zeidler, *Nonlinear Functional Analysis and its Applications, IV - Applications to Mathematical Physics* (Springer-Verlag, 1988).

DEPARTMENT OF COMMUNICATIONS AND INTEGRATED SYSTEMS, TOKYO INSTITUTE OF TECHNOLOGY, OOKAYAMA, MEGURO-KU, TOKYO 152-8552 JAPAN

E-mail address: isao@comm.ss.titech.ac.jp

PRECISION AND INTELLIGENCE LABORATORY, TOKYO INSTITUTE OF TECHNOLOGY, 4259 NAGATSUTA-CHO, MIDORI-KU, YOKOHAMA 226-8503 JAPAN

E-mail address: ogura@pi.titech.ac.jp

DEPARTMENT OF COMMUNICATIONS AND INTEGRATED SYSTEMS, TOKYO INSTITUTE OF TECHNOLOGY, OOKAYAMA, MEGURO-KU, TOKYO 152-8552 JAPAN

E-mail address: sirakawa@comm.ss.titech.ac.jp

This page intentionally left blank

Titles in This Series

- 313 **M. Zuhair Nashed and Otmar Scherzer, Editors**, Inverse problems, image analysis, and medical imaging, 2002
- 312 **Aaron Bertram, James A. Carlson, and Holger Kley, Editors**, Symposium in honor of C. H. Clemens, 2002
- 311 **Clifford J. Earle, William J. Harvey, and Sevín Recillas-Pishmish, Editors**, Complex manifolds and hyperbolic geometry, 2002
- 310 **Alejandro Adem, Jack Morava, and Yongbin Ruan, Editors**, Orbifolds in mathematics and physics, 2002
- 309 **Martin Guest, Reiko Miyaoka, and Yoshihiro Ohnita, Editors**, Integrable systems, topology, and physics, 2002
- 308 **Martin Guest, Reiko Miyaoka, and Yoshihiro Ohnita, Editors**, Differentiable geometry and integrable systems, 2002
- 307 **Ricardo Weder, Pavel Exner, and Benoit Grébert, Editors**, Mathematical results in quantum mechanics, 2002
- 306 **Xiaobing Feng and Tim P. Schulze, Editors**, Recent advances in numerical methods for partial differential equations and applications, 2002
- 305 **Samuel J. Lomonaco, Jr. and Howard E. Brandt, Editors**, Quantum computation and information, 2002
- 304 **Jorge Alberto Calvo, Kenneth C. Millett, and Eric J. Rawdon, Editors**, Physical knots: Knotting, linking, and folding geometric objects in \mathbb{R}^3 , 2002
- 303 **William Cherry and Chung-Chun Yang, Editors**, Value distribution theory and complex dynamics, 2002
- 302 **Yi Zhang, Editor**, Logic and algebra, 2002
- 301 **Jerry Bona, Roy Choudhury, and David Kaup, Editors**, The legacy of the inverse scattering transform in applied mathematics, 2002
- 300 **Sergei Vostokov and Yuri Zarhin, Editors**, Algebraic number theory and algebraic geometry: Papers dedicated to A. N. Parshin on the occasion of his sixtieth birthday, 2002
- 299 **George Kamberov, Peter Norman, Franz Pedit, and Ulrich Pinkall**, Quaternions, spinors, and surfaces, 2002
- 298 **Robert Gilman, Alexei G. Myasnikov, and Vladimir Shpilrain, Editors**, Computational and statistical group theory, 2002
- 297 **Stephen Berman, Paul Fendley, Yi-Zhi Huang, Kailash Misra, and Brian Parshall, Editors**, Recent developments in infinite-dimensional Lie algebras and conformal field theory, 2002
- 296 **Sean Cleary, Robert Gilman, Alexei G. Myasnikov, and Vladimir Shpilrain, Editors**, Combinatorial and geometric group theory, 2002
- 295 **Zhangxin Chen and Richard E. Ewing, Editors**, Fluid flow and transport in porous media: Mathematical and numerical treatment, 2002
- 294 **Robert Coquereaux, Ariel García, and Roberto Trinchero, Editors**, Quantum symmetries in theoretical physics and mathematics, 2002
- 293 **Donald M. Davis, Jack Morava, Goro Nishida, W. Stephen Wilson, and Nobuaki Yagita, Editors**, Recent progress in homotopy theory, 2002
- 292 **A. Chenciner, R. Cushman, C. Robinson, and Z. Xia, Editors**, Celestial Mechanics, 2002
- 291 **Bruce C. Berndt and Ken Ono, Editors**, q -series with applications to combinatorics, number theory, and physics, 2001
- 290 **Michel L. Lapidus and Machiel van Frankenhuyzen, Editors**, Dynamical, spectral, and arithmetic zeta functions, 2001

TITLES IN THIS SERIES

- 289 **Salvador Pérez-Esteva and Carlos Villegas-Blas, Editors**, Second summer school in analysis and mathematical physics: Topics in analysis: Harmonic, complex, nonlinear and quantization, 2001
- 288 **Marisa Fernández and Joseph A. Wolf, Editors**, Global differential geometry: The mathematical legacy of Alfred Gray, 2001
- 287 **Marlos A. G. Viana and Donald St. P. Richards, Editors**, Algebraic methods in statistics and probability, 2001
- 286 **Edward L. Green, Serkan Hosten, Reinhard C. Laubenbacher, and Victoria Ann Powers, Editors**, Symbolic computation: Solving equations in algebra, geometry, and engineering, 2001
- 285 **Joshua A. Leslie and Thierry P. Robart, Editors**, The geometrical study of differential equations, 2001
- 284 **Gaston M. N'Guérékata and Asamoah Nkwanta, Editors**, Council for African American researchers in the mathematical sciences: Volume IV, 2001
- 283 **Paul A. Milewski, Leslie M. Smith, Fabian Waleffe, and Esteban G. Tabak, Editors**, Advances in wave interaction and turbulence, 2001
- 282 **Arlan Ramsay and Jean Renault, Editors**, Groupoids in analysis, geometry, and physics, 2001
- 281 **Vadim Olshevsky, Editor**, Structured matrices in mathematics, computer science, and engineering II, 2001
- 280 **Vadim Olshevsky, Editor**, Structured matrices in mathematics, computer science, and engineering I, 2001
- 279 **Alejandro Adem, Gunnar Carlsson, and Ralph Cohen, Editors**, Topology, geometry, and algebra: Interactions and new directions, 2001
- 278 **Eric Todd Quinto, Leon Ehrenpreis, Adel Faridani, Fulton Gonzalez, and Eric Grinberg, Editors**, Radon transforms and tomography, 2001
- 277 **Luca Capogna and Loredana Lanzani, Editors**, Harmonic analysis and boundary value problems, 2001
- 276 **Emma Previato, Editor**, Advances in algebraic geometry motivated by physics, 2001
- 275 **Alfred G. Noël, Earl Barnes, and Sonya A. F. Stephens, Editors**, Council for African American researchers in the mathematical sciences: Volume III, 2001
- 274 **Ken-ichi Maruyama and John W. Rutter, Editors**, Groups of homotopy self-equivalences and related topics, 2001
- 273 **A. V. Kelarev, R. Göbel, K. M. Rangaswamy, P. Schultz, and C. Vinsonhaler, Editors**, Abelian groups, rings and modules, 2001
- 272 **Eva Bayer-Fluckiger, David Lewis, and Andrew Ranicki, Editors**, Quadratic forms and their applications, 2000
- 271 **J. P. C. Greenlees, Robert R. Bruner, and Nicholas Kuhn, Editors**, Homotopy methods in algebraic topology, 2001
- 270 **Jan Denef, Leonard Lipschitz, Thanases Pheidas, and Jan Van Geel, Editors**, Hilbert's tenth problem: Relations with arithmetic and algebraic geometry, 2000
- 269 **Mikhail Lyubich, John W. Milnor, and Yair N. Minsky, Editors**, Laminations and foliations in dynamics, geometry and topology, 2001
- 268 **Robert Gulliver, Walter Littman, and Roberto Triggiani, Editors**, Differential geometric methods in the control of partial differential equations, 2000
- 267 **Nicolás Andruskiewitsch, Walter Ricardo Ferrer Santos, and Hans-Jürgen Schneider, Editors**, New trends in Hopf algebra theory, 2000
- 266 **Caroline Grant Melles and Ruth I. Michler, Editors**, Singularities in algebraic and analytic geometry, 2000

TITLES IN THIS SERIES

- 265 **Dominique Arlettaz and Kathryn Hess, Editors**, Une dégustation topologique:
Homotopy theory in the Swiss Alps, 2000
- 264 **Kai Yuen Chan, Alexander A. Mikhalev, Man-Keung Siu, Jie-Tai Yu, and Efim
I. Zelmanov, Editors**, Combinatorial and computational algebra, 2000
- 263 **Yan Guo, Editor**, Nonlinear wave equations, 2000
- 262 **Paul Igodt, Herbert Abels, Yves Félix, and Fritz Grunewald, Editors**,
Crystallographic groups and their generalizations, 2000
- 261 **Gregory Budzban, Philip Feinsilver, and Arun Mukherjea, Editors**, Probability
on algebraic structures, 2000
- 260 **Salvador Pérez-Esteva and Carlos Villegas-Blas, Editors**, First summer school in
analysis and mathematical physics: Quantization, the Segal-Bargmann transform and
semiclassical analysis, 2000
- 259 **D. V. Huynh, S. K. Jain, and S. R. López-Permouth, Editors**, Algebra and its
applications, 2000
- 258 **Karsten Grove, Ib Henning Madsen, and Erik Kjær Pedersen, Editors**, Geometry
and topology: Aarhus, 2000
- 257 **Peter A. Cholak, Steffen Lempp, Manuel Lerman, and Richard A. Shore,
Editors**, Computability theory and its applications: Current trends and open problems,
2000
- 256 **Irwin Kra and Bernard Maskit, Editors**, In the tradition of Ahlfors and Bers:
Proceedings of the first Ahlfors-Bers colloquium, 2000
- 255 **Jerry Bona, Katarzyna Saxton, and Ralph Saxton, Editors**, Nonlinear PDE's,
dynamics and continuum physics, 2000
- 254 **Mourad E. H. Ismail and Dennis W. Stanton, Editors**, q -series from a contemporary
perspective, 2000
- 253 **Charles N. Delzell and James J. Madden, Editors**, Real algebraic geometry and
ordered structures, 2000
- 252 **Nathaniel Dean, Cassandra M. McZeal, and Pamela J. Williams, Editors**,
African Americans in Mathematics II, 1999
- 251 **Eric L. Grinberg, Shiferaw Berhanu, Marvin I. Knopp, Gerardo A. Mendoza,
and Eric Todd Quinto, Editors**, Analysis, geometry, number theory: The Mathematics
of Leon Ehrenpreis, 2000
- 250 **Robert H. Gilman, Editor**, Groups, languages and geometry, 1999
- 249 **Myung-Hwan Kim, John S. Hsia, Yoshiyuki Kitaoka, and Rainer Schulze-Pillot,
Editors**, Integral quadratic forms and lattices, 1999
- 248 **Naihuan Jing and Kailash C. Misra, Editors**, Recent developments in quantum affine
algebras and related topics, 1999
- 247 **Lawrence Wasson Baggett and David Royal Larson, Editors**, The functional and
harmonic analysis of wavelets and frames, 1999
- 246 **Marcy Barge and Krystyna Kuperberg, Editors**, Geometry and topology in
dynamics, 1999
- 245 **Michael D. Fried, Editor**, Applications of curves over finite fields, 1999
- 244 **Leovigildo Alonso Tarrío, Ana Jeremías López, and Joseph Lipman, Editors**, Studies in
duality on noetherian formal schemes and non-noetherian ordinary schemes, 1999
- 243 **Tsit Yuan Lam and Andy R. Magid, Editors**, Algebra, K -theory, groups, and
education, 1999
- 242 **Bernhelm Booss-Bavnbek and Krzysztof Wojciechowski, Editors**, Geometric
aspects of partial differential equations, 1999
- 241 **Piotr Pragacz, Michał Szurek, and Jarosław Wiśniewski, Editors**, Algebraic
geometry: Hirzebruch 70, 1999

TITLES IN THIS SERIES

- 240 **Angel Carocca, Víctor González-Aguilera, and Rubí E. Rodríguez, Editors,** Complex geometry of groups, 1999
- 239 **Jean-Pierre Meyer, Jack Morava, and W. Stephen Wilson, Editors,** Homotopy invariant algebraic structures, 1999
- 238 **Gui-Qiang Chen and Emmanuele DiBenedetto, Editors,** Nonlinear partial differential equations, 1999
- 237 **Thomas Branson, Editor,** Spectral problems in geometry and arithmetic, 1999
- 236 **Bruce C. Berndt and Fritz Gesztesy, Editors,** Continued fractions: From analytic number theory to constructive approximation, 1999
- 235 **Walter A. Carnielli and Itala M. L. D'Ottaviano, Editors,** Advances in contemporary logic and computer science, 1999
- 234 **Theodore P. Hill and Christian Houdré, Editors,** Advances in stochastic inequalities, 1999
- 233 **Hanna Nencka, Editor,** Low dimensional topology, 1999
- 232 **Krzysztof Jarosz, Editor,** Function spaces, 1999
- 231 **Michael Farber, Wolfgang Lück, and Shmuel Weinberger, Editors,** Tel Aviv topology conference: Rothenberg Festschrift, 1999
- 230 **Ezra Getzler and Mikhail Kapranov, Editors,** Higher category theory, 1998
- 229 **Edward L. Green and Birge Huisgen-Zimmermann, Editors,** Trends in the representation theory of finite dimensional algebras, 1998
- 228 **Liming Ge, Huixin Lin, Zhong-Jin Ruan, Dianzhou Zhang, and Shuang Zhang, Editors,** Operator algebras and operator theory, 1999
- 227 **John McCleary, Editor,** Higher homotopy structures in topology and mathematical physics, 1999
- 226 **Luis A. Caffarelli and Mario Milman, Editors,** Monge Ampère equation: Applications to geometry and optimization, 1999
- 225 **Ronald C. Mullin and Gary L. Mullen, Editors,** Finite fields: Theory, applications, and algorithms, 1999
- 224 **Sang Geun Hahn, Hyo Chul Myung, and Efim Zelmanov, Editors,** Recent progress in algebra, 1999
- 223 **Bernard Chazelle, Jacob E. Goodman, and Richard Pollack, Editors,** Advances in discrete and computational geometry, 1999
- 222 **Kang-Tae Kim and Steven G. Krantz, Editors,** Complex geometric analysis in Pohang, 1999
- 221 **J. Robert Dorroh, Gisèle Ruiz Goldstein, Jerome A. Goldstein, and Michael Mudi Tom, Editors,** Applied analysis, 1999
- 220 **Mark Mahowald and Stewart Priddy, Editors,** Homotopy theory via algebraic geometry and group representations, 1998
- 219 **Marc Henneaux, Joseph Krasil'shchik, and Alexandre Vinogradov, Editors,** Secondary calculus and cohomological physics, 1998
- 218 **Jan Mandel, Charbel Farhat, and Xiao-Chuan Cai, Editors,** Domain decomposition methods 10, 1998
- 217 **Eric Carlen, Evans M. Harrell, and Michael Loss, Editors,** Advances in differential equations and mathematical physics, 1998
- 216 **Akram Aldroubi and EnBing Lin, Editors,** Wavelets, multiwavelets, and their applications, 1998

For a complete list of titles in this series, visit the
AMS Bookstore at www.ams.org/bookstore/.

This book contains the proceedings of the Special Session, Interaction of Inverse Problems and Image Analysis, held at the January 2001 meeting of the AMS in New Orleans, LA.

The common thread among inverse problems, signal analysis, and image analysis is a canonical problem: recovering an object (function, signal, picture) from partial or indirect information about the object. Both inverse problems and imaging science have emerged in recent years as interdisciplinary research fields with profound applications in many areas of science, engineering, technology, and medicine. Research in inverse problems and image processing shows rich interaction with several areas of mathematics and strong links to signal processing, variational problems, applied harmonic analysis, and computational mathematics.

This volume contains carefully refereed and edited original research papers and high-level survey papers that provide overview and perspective on the interaction of inverse problems, image analysis, and medical imaging.

The book is suitable for graduate students and researchers interested in signal and image processing and medical imaging.

ISBN 0-8218-2979-3



9 780821 829790

CONN/313

AMS on the Web
www.ams.org