Mike Cullen, Melina A. Freitag, Stefan Kindermann, Robert Scheichl (Eds.)

Large Scale Inverse Problems

Radon Series on Computational and Applied Mathematics

Volume 13

Large Scale Inverse Problems

Computational Methods and Applications in the Earth Sciences

Edited by
Mike Cullen
Melina A. Freitag
Stefan Kindermann
Robert Scheichl

DE GRUYTER

Mathematics Subject Classification 2010
Primary: 65F22, 47A52, 35R30, 47J06, 93E11, 62M20, 68U10, 94A08, 86A10, 86A05, 86A22;
Secondary: 93E10, 60G35, 62P12, 62C10, 62F15, 62M05, 90C06, 62G35, 62G08, 37N10,
65C40, 65C05, 65K10, 35A15

ISBN 978-3-11-028222-1
e-ISBN 978-3-11-028226-9
ISSN 1865-3707

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie;
detailed bibliographic data are available in the Internet at http://dnb.dnb.de.

Typesetting: le-tex publishing services GmbH, Leipzig
Printing and binding: Hubert & Co. GmbH & Co. KG, Göttingen
∞ Printed on acid-free paper
Printed in Germany

www.degruyter.com

Preface

This book contains five invited expository articles resulting from the workshop *"Large-Scale Inverse Problems and Applications in the Earth Sciences"* which took place from October 24th to October 28th, 2011, at the Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences at the Johannes Kepler University in Linz, Austria. This workshop was part of a special semester at the RICAM devoted to "*Multiscale Simulation and Analysis in Energy and the Environment*" which took place from October 3rd to December 16th, 2011. The special semester was designed around four workshops with the ambition to invoke interdisciplinary cooperation between engineers, hydrologists, meteorologists, and mathematicians.

The workshop on which this collection of articles is based was devoted more specifically to establishing ties between specialists engaged in research involving real-world applications, e.g. in meteorology, hydrology and geosciences, and experts in the theoretical background such as statisticians and mathematicians working on Bayesian inference, inverse problem and control theory.

The two central problems discussed at the workshop were the processing and handling of large scale data and models in earth sciences, and the efficient extraction of the relevant information from them. For instance, weather forecasting models involve hundreds of millions of degrees of freedom and the available data easily exceed millions of measurements per day. Since it is of no practical use to predict tomorrow's weather from today's data by a process that takes a couple of days, the need for efficient and fast methods to manage large amounts of data is obvious. The second crucial aspect is the extraction of information (in a broad sense) from these data. Since this information is often "hidden" or perhaps only accessible by indirect measurements, it takes special mathematical methods to distill and process it. A general mathematical methodology that is useful in this situation is that of inverse problems and regularization and, closely related, that of Bayesian inference. These two paths of information extraction can very roughly be distinguished by the fact that in the former, the information is usually considered a deterministic quantity, while in the latter, it is treated as a stochastic one.

A loose arrangement of the articles in this book follows this structuring of information extraction paradigms; all in view of large scale data and real-world applications:

- *Aspects of inverse problems, regularization and data assimilation.* The article by Freitag and Potthast provides a general theoretical framework for data assimilation, a special type of inverse problem and puts the theory of inverse problems in context, providing similarities and differences between general inverse problems and data assimilation problems. Lawless discusses state-of-the-art methodologies for data assimilation as a state estimation problem in current real-world applications, with particular emphasis on meteorology. In both cases, the need to treat spatial and temporal correlations effectively makes the application somewhat different from many other applications of inverse problems.

- *Aspects of inverse problems and Bayesian inference*. The survey paper by Reich and Cotter gives an introduction to mathematical tools for data assimilation coming from Bayesian inference. In particular, ensemble filter techniques and Monte Carlo methods are discussed. In this case, the need to incorporate spatial and temporal correlations makes cost-effective implementation very challenging.
- *Aspects of inverse problems and regularization in imaging applications*. The article by Burger, Dirks and Müller is an overview of the process of acquiring, processing, and interpretation of data and the associated mathematical models in *imaging sciences*. While this article highlights the benefits of the nowadays very popular nonlinear (l_1-based) regularizations, the article by van den Doel, Ascher and Haber complements the picture by contrasting these benefits with the draw-backs of l_1-based approaches and by attempting to somewhat restore the "lost honor" of the more traditional and effective, linear l_2-type regularizations.

The review-type articles in this book contain basic material as well as many interesting aspects of inverse problems, regularization and data assimilation, with the provision of excellent and extensive references to the current literature. Hence, it should be of interest to both graduate students and researchers, and a valuable reference point for both practitioners and theoretical scientists.

We would like to thank the authors of these articles for their commendable contributions to this book. Without their time and commitment, the production of this book would not have been possible. We would also like to thank Nathan Smith (University of Bath) and Peter Jan van Leeuwen (University of Reading) who helped review the articles. Additionally, we would like to express our gratitude to the speakers and participants of the workshop, who contributed to a successful workshop in Linz.

Moreover, we would like to thank Prof. Heinz Engl, founder and former director of RICAM, and Prof. Ulrich Langer, former director of RICAM for their hospitality and for giving us the opportunity to organize this workshop at the RICAM. In addition, we would like to acknowledge the work of the administrative and computer support team at RICAM, Susanne Dujardin, Annette Weihs, Wolfgang Forsthuber and Florian Tischler, as well as the local scientific organizers Jörg Willems, Johannes Kraus and Erwin Karer. The special semester, the workshops and this book would not have been possible without their efforts.

More information on the special semester and the four workshops can be found at `http://www.ricam.oeaw.ac.at/specsem/specsem2011/`.

Exeter	*Mike Cullen*
Bath	*Melina A. Freitag*
Linz	*Stefan Kindermann*
Bath	*Robert Scheichl*

Contents

Melina A. Freitag and Roland W. E. Potthast

Synergy of inverse problems and data assimilation techniques

Abstract: This review article aims to provide a theoretical framework for data assimilation, a specific type of an inverse problem arising, for example, in numerical weather prediction, hydrology and geology.

We consider the general mathematical theory for inverse problems and regularization, before we treat Tikhonov regularization, as one of the most popular methods for solving inverse problems. We show that data assimilation techniques such as three-dimensional and four-dimensional variational data assimilation (3DVar and 4DVar) as well as the Kalman filter and Bayes' data assimilation are, in the linear case, a form of cycled Tikhonov regularization. We give an introduction to key data assimilation methods as currently used in practice, link them and show their similarities. We also give an overview of ensemble methods. Furthermore, we provide an error analysis for the data assimilation process in general, show research problems and give numerical examples for simple data assimilation problems. An extensive list of references is given for further reading.

Keywords: Inverse problems, ill-posedness, regularization theory, Tikhonov regularization, error analysis, 3DVar, 4DVar, Bayesian perspective, Kalman filter, Kalman smoother, ensemble methods, advection diffusion equation, Lorenz-95 system

2010 Mathematics Subject Classification: 65F22, 47A52, 35R30, 47J06, 93E11, 62M20

Melina A. Freitag: Department of Mathematical Sciences, University of Bath, Claverton Down, Bath, BA2 7AY, United Kingdom, m.freitag@maths.bath.ac.uk
Roland W. E. Potthast: Department of Mathematics, University of Reading, Whiteknights, PO Box 220, RG6 6AX, U.K. and Research and Development, Deutscher Wetterdienst, Section FE 12, Frankfurter Strasse 135, 63067 Offenbach, Germany, r.w.e.potthast@reading.ac.uk

We would like to thank the Johann Radon Institute for Computational and Applied Mathematics (RICAM) for hosting us in Linz during the Special Semester on Multiscale Simulation & Analysis in Energy and the Environment. The second author would like to acknowledge EPSRC, UK, for funding research on the "error dynamics of data assimilation methods," and the German government's special research program on "Remote Sensing Techniques in Data Assimilation" at Deutscher Wetterdienst (DWD).

1 Introduction

Inverse problems appear in many applications and have received a great deal of attention from applied mathematicians, engineers and statisticians. They occur, for example, in geophysics, medical imaging (such as ultrasound, computerized tomography and electrical impedance tomography), computer vision, machine learning, statistical inference, geology, hydrology, atmospheric dynamics and many other important areas of physics and industrial mathematics.

This article aims to provide a theoretical framework for data assimilation, a specific inverse problem arising, for example, in numerical weather prediction (NWP) and hydrology [48, 57, 58, 70, 83]. A few introductory articles on data assimilation in the atmospheric and ocean sciences are available, mainly from the engineering and meteorological point of view, for example, [20, 44, 48, 51, 63, 66, 71]. However, a comprehensive mathematical analysis in light of the theory of the inverse problem is missing. This expository article aims to achieve this.

An inverse problem is a problem which is posed in a way that is inverse to most direct problems. The so-called direct problem we have in mind is that of determining the effect f from given causes and conditions φ when a definite physical or mathematical model H in form of a relation

$$H(\varphi) = f \tag{1.1}$$

is given. In general, the operator H is nonlinear and describes the governing equations that relate the model parameters to the observed data. Hence, in an inverse problem, we are looking for φ, that is, a special cause, state, parameter or condition of a mathematical model. The solution of an inverse problem can be described as the construction of φ from data f (see, for example, [22, 49]). We now consider the specific inverse problem arising in data assimilation which usually also contains a dynamic aspect.

Data assimilation is, loosely speaking, a method for combining observations of the state of a complex system with predictions from a computer model output of that same state where both the observations and the model output data contain errors and (in case of the observations) are often incomplete. The task in data assimilation (and hence the inverse problem) is seeking the best state estimate with the available information about the physical model and observations.

Let X be the state space. For the remainder of this article, we generally assume that X (and also Y) are Hilbert spaces unless otherwise stated. Let $\varphi \in X$, where φ is the state (of the atmosphere, for example), that is, a vector containing all state variables. Furthermore, let $\varphi_k \in X$ be the state at time t_k and $M_k : X \to X$ the (generally nonlinear) model operator at time t_k which describes the evolution of the states from time t_k to time t_{k+1}, that is, $\varphi_{k+1} = M_k(\varphi_k)$. For the moment, we consider a perfect model, that is, the true system dynamics are assumed to be known. We also use the notation

$$M_{k,\ell} = M_{k-1}M_{k-2}\cdots M_{\ell+1}M_{\ell}, \quad k > \ell \in \mathbb{N}_0, \tag{1.2}$$

to describe the evolution of the system dynamics from time t_ℓ to time t_k.

Let Y_k be the observation space at time t_k and $f_k \in Y_k$ be the observation vector, collecting all the observations at time t_k. Finally, let $H_k : X \to Y_k$ be the (generally nonlinear) observation operator at time t_k, mapping variables in the state space to variables in the observation space. The data assimilation problem can then be defined as follows.

Definition 1.1 (Data assimilation problem). Given observations $f_k \in Y_k$ at time t_k, determine the states $\varphi_k \in X$ from the operator equations

$$H_k(\varphi_k) = f_k, \quad k = 0, 1, 2, \dots \tag{1.3}$$

subject to the model dynamics $M_k : X \to X$ given by $\varphi_{k+1} = M_k(\varphi_k)$, where $k = 0, 1, 2, \dots$.

In numerical weather prediction, the operator M_k involves the solution of a time-dependent nonlinear partial differential equation. Usually, the observation operator H_k is dynamic, that is, it changes at every time step. However, for simplicity, we often let $H_k := H$. Both the operator H_k and the data f_k contain errors. Also, in practice, the dynamical model M_k involves errors, that is, M_k does not represent the true system dynamics because of model errors. For a detailed account on errors occurring in the data assimilation problem, we refer to Section 4. Moreover, the model dynamics represented by the nonlinear operators M_k are usually chaotic. In the context of data assimilation, additional information might be given through known prior information (background information) about the state variable denoted by $\varphi_k^{(b)} \in X$.

The operator equation (1.3) (see also (1.1)) is usually ill-posed, that is, at least one of the following well-posedness conditions according to Hadamard [33] is not satisfied.

Definition 1.2 (Well-Posedness [49, 82]). Let X, Y be normed spaces and $H : X \to Y$ be a nonlinear mapping. Then, the operator equation $H(\varphi) = f$ from (1.1) is called well-posed if the following holds:

- Existence: For every $f \in Y$, there exists at least one $\varphi \in X$ such that $H(\varphi) = f$, that is, the operator H is surjective.
- Uniqueness: The solution φ from $H(\varphi) = f$ is unique, that is, the operator H is injective.
- Stability: The solution φ depends continuously on the data f, that is, it is stable with respect to perturbations in f.

Equation (1.1) is ill-posed if it is not well-posed.

Note that for a general nonlinear operator H, both the existence and uniqueness of the operator equation need not be satisfied. If the existence condition in Definition 1.2 is not satisfied, then it is possible that $f \in \mathcal{R}(H)$. However, for a perturbed right-hand side f^{δ}, we have $f^{\delta} \notin \mathcal{R}(H)$, where $\mathcal{R}(H) = \{f \in Y, f = H(\varphi), \varphi \in X\}$ is the range of H. Existence of a generalized solution can sometimes (for instance, in the finite-dimensional case) be ensured by solving the minimization problem

$$\min ||f - H(\varphi)||_Y^2 \ , \tag{1.4}$$

which is equivalent to (1.1) if $f \in \mathcal{R}(H)$. The norm $\| \cdot \|_Y$ is a generic norm in Y. The second condition in Definition 1.2 implies that an inverse operator $H^{-1} : \mathcal{R}(H) \subseteq Y \to X$ with $H^{-1}(f) = \varphi$ exists. If the uniqueness condition is not satisfied, then it is possible to ensure uniqueness by looking for special solutions, for example, solutions that are closest to a reference element $\varphi^* \in X$, or, solutions with a minimum norm. Hence, at least in the linear case, uniqueness can be ensured if

$$||f - H(\varphi_{uni})||_Y = \min_{\varphi \in X} ||f - H(\varphi)||_Y \ , \tag{1.5}$$

where $\|\varphi_{uni} - \varphi^*\|_X = \min\{\|\varphi - \varphi^*\|_X, \varphi \in X, \ \varphi$ is a minimizer in (1.5)$\}$. The third condition in Definition 1.2 implies that the inverse operator $H^{-1} : \mathcal{R}(H) \subseteq Y \to X$ is continuous. Usually, this problem is the most severe one as small perturbations in the right-hand side $f \in Y$ lead to large errors in the solution $\varphi \in X$ and the problem needs to be regularized. We will look at this aspect in Section 2.

From the above discussion, it follows that the operator equation (1.3) is well-posed if the operator H_k is bijective and has a well-defined inverse operator H_k^{-1} which is continuous. A least squares solution can be found by solving the minimization problem

$$\min_{\varphi_k \in X} ||f_k - H_k(\varphi_k)||_Y^2 \ , \quad k = 0, 1, 2, \dots \ . \tag{1.6}$$

We can solve (1.6) at every time step k, which is a sequential data assimilation problem. If we include the nonlinear model dynamics constraint $M_k : X \to X$ given by $\varphi_{k+1} = M_k(\varphi_k)$, over the time steps t_k, $k = 0, \dots, K$, and take the sum of the least squares problem in every time step, the minimization problem becomes

$$\min_{\varphi_k \in X} \sum_{k=0}^{K} ||f_k - H_k(\varphi_k)||_Y^2 = \min_{\varphi_0 \in X} \sum_{k=0}^{K} ||f_k - H_k M_{k,0}(\varphi_0)||_Y^2 \ , \tag{1.7}$$

where $M_{k,0}$ denotes the evolution of the model operator from time t_0 to time t_k, that is, $M_{k,0} = M_{k-1}M_{k-2} \cdots M_0$, using the system dynamics (1.2), and $M_{k,k} = I$. Both the sequential data assimilation system (1.6) and the data assimilation system (1.7) can be written in the form

$$\min_{\varphi \in X} \left|\left|\overline{f} - \overline{H}(\varphi)\right|\right|_Y^2 \ , \tag{1.8}$$

with an appropriate operator $\overline{H}$. Problem (1.8) is equivalent to $\overline{H}(\varphi) = \overline{f}$ (cf. (1.1)) if $\overline{f} \in \mathcal{R}(\overline{H})$. For the sequential assimilation system (1.6), we have $\overline{H} := H_k$, $\overline{f} := f_k$ and $\varphi := \varphi_k$ at every step $k = 0, 1, \ldots$. For the system (1.7), we have $\varphi := \varphi_0$,

$$\overline{H} := \begin{bmatrix} H_0 \\ H_1 M_{1,0} \\ H_2 M_{2,0} \\ \vdots \\ H_K M_{K,0} \end{bmatrix} \quad \text{and} \quad \overline{f} := \begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ \vdots \\ f_K \end{bmatrix} .$$

In general, $\overline{H}$ is a *nonlinear* operator since both the model dynamics M_k and the observation operators H_k are nonlinear. If the equation $\overline{H}(\varphi) = \overline{f}$ is well-posed, then $\overline{H}$ has a well-defined continuous inverse operator $\overline{H}^{-1}$ and $\mathcal{R}(\overline{H}) = Y$.

Now, if $\overline{H}$ is a *linear* operator in Banach spaces, then well-posedness follows from the first two conditions in Definition 1.2, which are equivalent to $\mathcal{R}(\overline{H}) = Y$ and $\mathcal{N}(\overline{H}) = \{0\}$ where $\mathcal{N}(\overline{H})$ is the null space of $\overline{H}$. Moreover, if $\overline{H}$ is a *linear* operator on a finite-dimensional Hilbert space (in particular, if $\mathcal{R}(\overline{H})$ is of finite dimension), then the stability condition in Definition 1.2 holds automatically and well-posedness follows from either one of the first two conditions in 1.2. (The last condition in Definition 1.2 follows from the compactness of the unit ball in finite dimensions [49].) For linear $\overline{H}$, the uniqueness condition $\mathcal{N}(\overline{H}) = \{0\}$ is clearly satisfied if the *observability matrix* $\overline{H}$ has full row rank. In this case, the system is observable, that is, it is possible to determine the behavior of the entire system from the systems output, see [47, 73].

The remaining question is the stability of the (injective) operator equation $\overline{H}(\varphi) = \overline{f}$ (or $H\varphi = H(\varphi) = f$, a notation which we are going to use from now on) for a compact linear operator $H : X \to Y$ in infinite dimensions. As a compact linear operator is always ill-posed in an infinite-dimensional space (as $\mathcal{R}(H)$ is not closed), we need some form of regularization.

Note that the discretization of an infinite-dimensional unstable ill-posed problem naturally leads to a finite-dimensional problem which is well-posed, that is, according to Definition 1.2. However, the discrete problem will be ill-conditioned, that is, an error in the input data will still lead to large errors in the solution. Hence, some form of regularization is also needed for finite-dimensional problems arising from infinite-dimensional ill-posed operators.

In the following, we consider compact linear operators H for which a singular value decomposition exists (see, for example, [49]).

Lemma 1.3 (Singular system of compact linear operators). *Let $H : X \to Y$ be a compact linear operator. Then, there exist sets of indices $J = \{1, \ldots, m\}$ for dim$(R(H)) = m$ and $J = \mathbb{N}$ for dim$(R(H)) = \infty$, orthonormal systems $\{u_j\}_{j \in J}$ in X and $\{v_j\}_{j \in J}$*

in Y and a sequence $\{\sigma_j\}_{j\in J}$ of positive real numbers with the following properties:

$$\{\sigma_j\}_{j\in J} \text{ is non-increasing} \quad \text{and} \quad \lim_{j\to\infty} \sigma_j = 0 \text{ for } J = \mathbb{N}\,, \tag{1.9}$$

$$Hu_j = \sigma_j v_j,\ (j \in J) \quad \text{and} \quad H^* v_j = \sigma_j u_j\,, \quad (j \in J)\,. \tag{1.10}$$

For all $\varphi \in X$, there exists an element $\varphi_0 \in \mathcal{N}(H)$ with

$$\varphi = \varphi_0 + \sum_{j\in J} \left\langle \varphi, u_j \right\rangle_X u_j \quad \text{and} \quad H\varphi = \sum_{j\in J} \sigma_j \left\langle \varphi, u_j \right\rangle_X v_j\,. \tag{1.11}$$

Furthermore,

$$H^* f = \sum_{j\in J} \sigma_j \left\langle f, v_j \right\rangle_Y u_j \tag{1.12}$$

holds for all $f \in Y$. The countable set of triples $\{\sigma_j, u_j, v_j\}_{j\in J}$ is called a singular system, $\{\sigma_j\}_{j\in J}$ are called singular values, $\{u_j\}_{j\in J}$ are right singular vectors and form an orthonormal basis for $\mathcal{N}(H)^\perp$ and $\{v_j\}_{j\in J}$ are left singular vectors and form an orthonormal basis for $\overline{\mathcal{R}(H)}$.

In the following, we mostly consider compact linear operators, although the concept of ill-posedness can be extended to nonlinear operators [23, 40, 49, 82] by considering linearizations of the nonlinear problem using, for example, the Fréchet derivative of the nonlinear operator. One can show that for compact nonlinear operators, the Fréchet derivative is compact as well, leading to the concept of locally ill-posed problems for nonlinear operator equations. For solving nonlinear problems computationally, usually some form of linearization is required. Hence, most of our results for linear problems can be extended to the case of iterative solutions to nonlinear problems (where a linear problem needs to be solved at each iteration).

2 Regularization theory

Problems of the form $H\varphi = f$ with a compact operator H are ill-posed in infinite dimensions since the inverse of H is not uniformly bounded. However, in order to solve $H\varphi = f$ (or, for $f \notin \mathcal{R}(H)$, its equivalent minimization problem $\min \|H\varphi - f\|^2$), regularization is needed.

Let $H : X \to Y$ and denote its adjoint operator by $H^* : Y \to X$. Furthermore, let φ be the unique solution to the least squares minimization problem $\min \|H\varphi - f\|^2$. Then, the solution to the minimization problem is equivalent to the solution of the normal equations

$$H^* H\varphi = H^* f\,. \tag{1.13}$$

Clearly, if $H : X \to Y$ is compact, then H^*H is compact and the normal equations (1.13) remain ill-posed. However, if we replace (1.13) by

$$(\alpha I + H^* H)\,\varphi_\alpha = \alpha\varphi_\alpha + H^* H\varphi_\alpha = H^* f \tag{1.14}$$

with $\alpha > 0$, then the operator $(\alpha I + H^* H)$ has a bounded inverse. The equation (1.14) is typically referred to as Tikhonov regularization and α is a regularization parameter. We have the following theorem (see, for example, [17, 40, 62, 78, 82]).

Theorem 1.4 (Tikhonov regularization). *Let $H : X \to Y$ be a compact linear operator. Then, the operator $(\alpha I + H^* H)$ has a bounded inverse and the problem* (1.14) *is well-posed for $\alpha > 0$ and $\varphi_\alpha = (\alpha I + H^* H)^{-1} H^* f$ is the Tikhonov approximation of a minimum-norm least squares solution φ of* (1.13). *Furthermore, the solution φ_α is equivalent to the unique solution of the minimization problem*

$$\min_{\varphi \in X} T_\alpha(\varphi) := \min_{\varphi \in X} \left\{ ||f - H\varphi||_Y^2 + \alpha \, \|\varphi\|_X^2 \right\} , \tag{1.15}$$

where $T_\alpha(\varphi)$ is the so-called Tikhonov functional.

In general, Tikhonov regularization can be used with a known reference element $\varphi^{(b)}$, that is, the term $\|\varphi\|_X^2$ in (1.15) is replaced by $\|\varphi - \varphi^{(b)}\|_X^2$, and the problem is often referred to as generalized Tikhonov regularization. We consider this problem in Section 3.

We have the following definition for a general linear regularization scheme.

Definition 1.5 (Regularization scheme). A family of bounded linear operators $\{R_\alpha\}_{\alpha>0}$, $R_\alpha : Y \to X$ is a linear regularization scheme for the compact bounded linear injective operator H if

$$\lim_{\alpha \to 0} R_\alpha H \varphi = \varphi \quad \forall \varphi \in X . \tag{1.16}$$

Clearly, the family of approximate inverses $R_\alpha = (\alpha I + H^* H)^{-1} H^* : Y \to X$ is a linear regularization scheme for H. If the range of H, $\mathcal{R}(H)$, is not closed, then

$$\lim_{\alpha \to 0} \|R_\alpha\| = \infty . \tag{1.17}$$

If we apply the regularization operator R_α to noisy data f^δ with noise level δ, that is, $\|f^\delta - f\|_Y \le \delta$, we get regularized solutions

$$\varphi_\alpha^\delta = R_\alpha f^\delta .$$

Using the singular system of a compact operator from Lemma 1.3, we may also write the regularized solution arising from Tikhonov regularization via the minimization problem in (1.15) as

$$\varphi_\alpha^\delta = \sum_{j \in J} \frac{\sigma_j}{\sigma_j^2 + \alpha} \left\langle f^\delta, v_j \right\rangle_Y u_j . \tag{1.18}$$

We observe that for $\alpha = 0$, the solution φ_α^δ amplifies the noise in f^δ, since for compact operators $\lim_{j \to \infty} \sigma_j = 0$.

Furthermore, for the exact unique solution, we have $\varphi = H^\dagger f$, where $H^\dagger : \mathcal{R}(H) + \mathcal{R}(H)^\perp \to X$ denotes the Moore–Penrose pseudoinverse of H [82] and it is continuous if $\mathcal{R}(H)$ is closed. Therefore, we may estimate the total regularization error

$$\left\|\varphi_\alpha^\delta - \varphi\right\|_X \le \|R_\alpha\|\,\delta + \left\|R_\alpha f - H^\dagger f\right\|_X ,$$

or, for $\mathcal{N}(H) = \{0\}$,

$$\left\|\varphi_\alpha^\delta - \varphi\right\|_X \le \|R_\alpha\|\,\delta + \|R_\alpha H\varphi - \varphi\|_X . \tag{1.19}$$

Hence, the total regularization error consists of a stability component $\|R_\alpha\|\delta$ which represents the influence of the data error δ and a component $\|R_\alpha H\varphi - \varphi\|_X$ which represents the approximation error of the regularization scheme. For small α, the second component will be small (1.16), but the first component will be large (1.17). However, for large values of α, the first term will be small and the second one large. We will see this in the examples in Section 9. Hence, finding a good value for the regularization parameter α is important. Techniques for regularization parameter estimation aim to find a reasonably good value for α (see, for example, [37, 38, 82]). The most prominent ones are the L-curve method, generalized cross-validation and the discrepancy principle.

A regularization scheme is called convergent if from the convergence of the data error to zero, it follows that the regularized solution converges to the exact solution. One can show that a regularization scheme $R_\alpha = (\alpha I + H^*H)^{-1}H^* : Y \to X$ arising in Tikhonov regularization is a convergent regularization if $\alpha(\delta) \to 0$ and $\frac{\delta^2}{\alpha(\delta)} \to 0$ as $\delta \to 0$ [22]. For Tikhonov regularization, one may choose $\alpha = \mathcal{O}(\delta)$ such that this holds [82].

Other regularization schemes for inverse problems are also possible, some of the most famous ones being the truncated singular value decomposition (TSVD) and the Landweber iteration (see, for example, [22, 34, 35]). Moreover, it is possible to change the penalty term $\|\varphi\|_X^2$ in (1.15). Other penalty functionals can be used to incorporate *a priori* information about the solution φ. Prominent methods are total variation regularization or the use of sparsity promoting norms (like the L_1-norm, for example) in the penalty functional. There is a fast growing literature on this topic, see, for example, [1, 7, 13, 82, 86] and the articles by Burger et al. [10] and van den Doel et al. [81] in this book.

In the following, we use the results from inverse problems and regularization theory to develop a coherent mathematical framework for several data assimilation techniques used in practice.

3 Cycling, Tikhonov regularization and 3DVar

Data assimilation aims to solve a dynamic inverse problem which includes measurement data $f_1, f_2, f_3, \dots, f_k, \dots$ at various times $t_1 < t_2 < t_3 < \cdots < t_k < \cdots$. At every time t_k, the inversion problem is given by (1.3). However, usually the data f_k do not contain enough information to recover the state φ_k at time t_k completely. Thus, it is crucial to take the dynamical evolution of the states into account.

Assume that we are given some reconstruction $\varphi_k^{(a)}$ at time t_k for some $k \in \mathbb{N}$. Then, we expect that

$$\varphi_{k+1}^{(b)} := M_k\left(\varphi_k^{(a)}\right) \tag{1.20}$$

is a reasonable first guess for the system state at time t_{k+1}, where M_k describes the model dynamics and is given in Definition 1.1. In data assimilation, $\varphi^{(b)}$ is called the *background* or *first guess*. At time t_{k+1}, we would like to assimilate the data f_{k+1} to calculate a reconstruction $\varphi_{k+1}^{(a)}$, which is also called the *analysis* in data assimilation. Then, the background $\varphi_{k+2}^{(b)}$ at time t_{k+2} can be calculated using (1.20) with k replaced by $k+1$ and another reconstruction can be carried out at time t_{k+2}. This approach is called *cycling* of reconstruction and dynamics.

Definition 1.6 (Cycling for data assimilation). Start with some initial state $\varphi_0^{(a)}$ at time t_0. For $k = 0, 1, 2, \ldots$, carry out the cycling steps:

(i) *Propagation Step*. Use the system dynamics M_k to calculate a *background* $\varphi_{k+1}^{(b)}$ at time t_{k+1} using (1.20).

(ii) *Analysis Step*. With the data f_{k+1} at time t_{k+1} (and the knowledge of the background $\varphi_{k+1}^{(b)}$), calculate a reconstruction or *analysis* $\varphi_{k+1}^{(a)}$.

Increase the index k to $k+1$ and go to Step (i).

A key characteristic of a data assimilation system is its Analysis Step (ii). Here, for any step k, the task is to calculate a reconstruction $\varphi_k^{(a)}$ using the data f_k and the knowledge of the background $\varphi_k^{(b)}$. We need to choose or develop a reconstruction method which optimally combines the given information.

To carry out the analysis, we will study two basic approaches, one coming from optimization and *optimal control theory*, the other arising from *stochastics and probability theory*. In this section, we focus on the *optimization* approach and Section 5 will provide an introduction to the stochastic approach using Bayes' formula. The relationship between the two approaches will be discussed in detail in Section 5.

With a norm $\|\cdot\|_X$ in the state space X and a norm $\|\cdot\|_Y$ in the data (or observation) space Y, we can combine the given information at step k, namely, the observation data $f_k \in Y$ and the background $\varphi_k^{(b)} \in X$ by minimizing the *inhomogeneous Tikhonov functional*

$$J_k(\varphi) := \alpha \left\|\varphi - \varphi_k^{(b)}\right\|_X^2 + \|f_k - H\varphi\|_Y^2 \tag{1.21}$$

at time t_k. $H : X \to Y$ is the observation operator defined in Section 1. With $\tilde{\varphi}_k := \varphi - \varphi_k^{(b)}$, this is transformed into the Tikhonov functional (1.15) in the formula

$$\tilde{J}_k(\tilde{\varphi}_k) := \alpha \|\tilde{\varphi}_k\|_X^2 + \left\|(f_k - H\varphi_k^{(b)}) - H\tilde{\varphi}_k\right\|_Y^2 . \tag{1.22}$$

According to Theorem 1.4, it is minimized by

$$\tilde{\varphi}_k^{(a)} := \left(\alpha I + H^* H\right)^{-1} H^* \left(f_k - H\varphi_k^{(b)}\right), \tag{1.23}$$

leading to the minimizer

$$\varphi_k^{(a)} = \varphi_k^{(b)} + \left(\alpha I + H^* H\right)^{-1} H^* \left(f_k - H\varphi_k^{(b)}\right) \tag{1.24}$$

of the functional (1.21). We denote the cycling of Definition 1.6 with an analysis calculated by (1.24) as *cycled Tikhonov regularization*.

Often, data assimilation works in spaces $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$ of dimensions $n \in \mathbb{N}$ and $m \in \mathbb{N}$. The norms in the spaces X and Y are given explicitly using the standard L^2-norms and some weighting matrices $B \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{m\times m}$. In Section 5, these matrices will be chosen to coincide with the error covariance matrices of the state distributions in X and the error covariance matrices of the observation distributions in Y. For the moment, we assume the matrices to be symmetric, positive definite and invertible. Then, we define a weighted scalar product in $X = \mathbb{R}^n$ by

$$\langle \varphi, \psi \rangle_{B^{-1}} := \varphi^T B^{-1} \psi, \quad \varphi, \psi \in X = \mathbb{R}^n, \tag{1.25}$$

and a weighted scalar product in $Y = \mathbb{R}^m$ by

$$\langle f, g \rangle_{R^{-1}} := f^T R^{-1} g, \quad f, g \in Y = \mathbb{R}^m. \tag{1.26}$$

With the corresponding norms $\|\cdot\|_{B^{-1}}$ in X and $\|\cdot\|_{R^{-1}}$ in Y, we can rewrite the functional (1.21) into the form

$$J_k(\varphi) = \alpha \left(\varphi - \varphi_k^{(b)}\right)^T B^{-1} \left(\varphi - \varphi_k^{(b)}\right) + \left(f_k - H\varphi\right)^T R^{-1} \left(f_k - H\varphi\right). \tag{1.27}$$

In the framework of the cycling given by Definition 1.6, this functional is known as the *three-dimensional variational data assimilation scheme* (3DVar), see, for example, [20, 51]. Often, the notation x and $x^{(b)}$ for the state and the background, as well as y for the observations, is used in the meteorological literature of data assimilation. Here, by building a bridge to the functional analytic framework, we will use $\varphi \in X$ for the states and $f \in Y$ for the observations. Also, x, y will be points in the physical space $\mathbb{R}^3$, respectively. This is also advantageous when we employ ensemble methods and analyze localization techniques.

The functional (1.27) can easily be transformed into the general Tikhonov regularization form. By H', we denote the adjoint operator of H with respect to the standard L^2 scalar products in $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$. The notation H^* is used for the adjoint operator with respect to the weighted scalar products $\langle .,. \rangle_{B^{-1}}$ and $\langle .,. \rangle_{R^{-1}}$.

Then, we calculate

$$
\begin{aligned}
\langle \varphi , H\psi\rangle_{R^{-1}} &= \left\langle \varphi, R^{-1} H\psi \right\rangle \\
&= \left\langle H' R^{-1} \varphi, \psi \right\rangle \\
&= \left\langle H' R^{-1} \varphi, BB^{-1} \psi \right\rangle \\
&= \left\langle BH' R^{-1} \varphi, B^{-1} \psi \right\rangle \\
&= \left\langle BH' R^{-1} \varphi, \psi \right\rangle_{B^{-1}} \\
&= \langle H^* \varphi, \psi \rangle_{B^{-1}} ,
\end{aligned}
\tag{1.28}
$$

leading to

$$H^* = BH'R^{-1} .$$

This means that the minimizer (1.24) of (1.21) with the norms based on the scalar products (1.25) and (1.26) is given by

$$
\begin{aligned}
\varphi_k^{(a)} &= \varphi_k^{(b)} + (\alpha I + H^* H)^{-1} H^* (f_k - H\varphi_k^{(b)}) \\
&= \varphi_k^{(b)} + (\alpha I + BH'R^{-1}H)^{-1} BH'R^{-1} (f_k - H\varphi_k^{(b)}) .
\end{aligned}
\tag{1.29}
$$

The operator $\alpha I + H^* H$ maps the state space X into itself. In large scale data assimilation problems, the dimension n of the state space is often much larger than the dimension m of the data space Y. In this case, the inversion of $\alpha I + H^* H$ is not feasible, and it is advantageous to derive a different form of the update formula known as *measurement space inversion*. Using the invertibility of the operators $\alpha I + H^* H$ in X and $\alpha I + HH^*$ in Y, we start from

$$(\alpha I + H^* H) H^* = H^* (\alpha I + HH^*) .$$

We multiply with the inverse $(\alpha I + H^* H)^{-1}$ from the left and by $(\alpha I + HH^*)^{-1}$ from the right to obtain

$$H^* (\alpha I + HH^*)^{-1} = (\alpha I + H^* H)^{-1} H^* . \tag{1.30}$$

With the help of (1.30), we transform (1.29) into

$$
\begin{aligned}
\varphi_k^{(a)} &= \varphi_k^{(b)} + H^* (\alpha I + HH^*)^{-1} (f_k - H\varphi_k^{(b)}) \\
&= \varphi_k^{(b)} + BH'R^{-1} (\alpha I + HBH'R^{-1})^{-1} (f_k - H\varphi_k^{(b)}) \\
&= \varphi_k^{(b)} + BH' (\alpha R + HBH')^{-1} (f_k - H\varphi_k^{(b)}) .
\end{aligned}
\tag{1.31}
$$

Here, the inversion of $(\alpha I + HH^*)$ or $(\alpha R + HBH')$, respectively, takes place in the space $Y = \mathbb{R}^m$. The solution is then projected into the state space by the application of BH'. In the meteorological literature of data assimilation, the solution (1.29) is often referred to as the solution arising from Optimal Interpolation (OI) [29, 68]. It refers to a direct method being used to solve the 3DVar minimization problem (1.27) rather than an iterative optimization technique. In the linear case, Optimal Interpolation and 3DVar are equivalent. Method (1.31) is often called the PSAS (physical space statistical analysis) scheme in the literature on meteorology and oceanography [16, 18].

We summarize our results in the following theorem.

Theorem 1.7 (Equivalence of cycled Tikhonov regularization and 3DVar). *3DVar or three-dimensional variational data assimilation* (1.29) *or* (1.31) *is equivalent to cycled Tikhonov regularization* (1.24) *when the norms are arising from the weighted inner products* (1.25) *and* (1.26).

Theorem 1.7 shows that 3DVar is merely a cycled Tikhonov regularization in an appropriately chosen norm.

4 Error analysis

In this part, we investigate the error arising in data assimilation, that is, we consider the error between the true solution and the solution obtained from a data assimilation scheme. The solution obtained from solving a data assimilation problem is often referred to as *analysis* in the data assimilation literature. As a generic method, we will study cycled Tikhonov regularization, which, according to Theorem 1.7, includes three-dimensional variational assimilation. We will later see that this also carries over to cycled four-dimensional variational data assimilation, which we will discuss in Section 6.

We need to take into account errors which can arise when we cycle the update formula (1.24) according to Definition 1.6. Assume that $\varphi_k^{(\text{true})}$ is the true state at time t_k, $k = 0, 1, 2, \dots$ and $f_k^{(\text{true})}$ are the true values of the data. The errors we need to take into account include

(1) *Measurement error:* Errors in the data f_k, that is, we measure f_k^δ with a *data error* $d_k^\delta := f_k^\delta - f_k^{(\text{true})}$ of size $\|d_k^\delta\| \leq \delta$. This error was discussed in Section 2 and arises through errors in the measurements and noisy data.

(2) *Observation operator error:* Errors in the *measurement operator* H, that is, we use a measurement operator H which is different from the true mapping $H^{(\text{true})}$ of the state φ to the data f.

(3) *Reconstruction/approximation error:* Reconstruction errors by using the inverse $R_\alpha = (\alpha I + H^*H)^{-1}H^*$ as an approximation to the inverse H^{-1} of H. This error was discussed in Section 2.

(4) *Model error:* The model operator which we defined in Section 1 is usually only an approximation M to the true system dynamics $M^{(\text{true})}$. Model error arises as the dynamical model does not usually describe the system behavior exactly. It incorporates numerical error arising from discretization of the partial differential equations that need to be solved and includes inaccuracies in the physical pa-

rameters, forcing terms and as well as in the model itself which is usually merely a simplification of the reality.

(5) *Accumulated errors:* There will be *accumulated errors* in the background in the sense that the analysis error from the previous step leads to an error in the background of the next step in contrast to the background which would be arising from the true state $\varphi^{(\mathrm{true})}$.

In every analysis step of the assimilation, we obtain an error contribution by the measurement error, by the error in the observation operator H and by the regularization operator R_α approximating the inversion of H. For the propagation step, we obtain an error caused by the model M approximating the true dynamics $M^{(\mathrm{true})}$. Moreover, the errors may accumulate over time.

Theorem 1.8. *The evolution of the analysis error* $e_k := \varphi_k^{(a)} - \varphi_k^{(true)}$ *for cycled Tikhonov regularization and three-dimensional variational assimilation is given by*

$$\begin{aligned} e_{k+1} = {} & \overbrace{(I - R_\alpha H)}^{\textit{reconstruction error}} \overbrace{\left\{M_k e_k + \left(M_k - M_k^{(true)}\right) \varphi_k^{(true)}\right\}}^{\textit{propagation of previous error and model error}} \\ & + \overbrace{R_\alpha d_{k+1}^{\delta}}^{\textit{data error influence}} \quad \overbrace{+ R_\alpha \left((H^{(true)} - H) \varphi_{k+1}^{(true)}\right)}^{\textit{observation operator error}} . \end{aligned} \tag{1.32}$$

Proof. We know from Theorem 1.7 that 3DVar and Tikhonov regularization are equivalent. We use the update formula (1.24) and the Tikhonov regularization operator $R_\alpha := (\alpha I + H^* H)^{-1} H^*$. With (1.20), as well as

$$\varphi_{k+1}^{(\mathrm{true})} = M_k^{(\mathrm{true})} \varphi_k^{(\mathrm{true})} \quad \text{and} \quad f_k^{(\mathrm{true})} = H^{(\mathrm{true})} \varphi_k^{(\mathrm{true})},$$

and subtracting $\varphi_{k+1}^{(\mathrm{true})}$ from $\varphi_{k+1}^{(a)}$, we calculate

$$\begin{align} e_{k+1} : = {} & \varphi_{k+1}^{(a)} - \varphi_{k+1}^{(\mathrm{true})} \nonumber \\ = {} & \varphi_{k+1}^{(b)} - \varphi_{k+1}^{(\mathrm{true})} + R_\alpha \left(f_{k+1} - f_{k+1}^{(\mathrm{true})}\right) \nonumber \\ & + R_\alpha \left(f_{k+1}^{(\mathrm{true})} - H \varphi_{k+1}^{(b)}\right) \tag{1.33} \\ = {} & M_k \varphi_k^{(a)} - M_k^{(\mathrm{true})} \varphi_k^{(\mathrm{true})} + R_\alpha d_{k+1}^{\delta} \nonumber \\ & + R_\alpha \left(H^{(\mathrm{true})} \varphi_{k+1}^{(\mathrm{true})} - H \varphi_{k+1}^{(b)}\right) \nonumber \\ = {} & M_k \left(\varphi_k^{(a)} - \varphi_k^{(\mathrm{true})}\right) + \left(M_k - M_k^{(\mathrm{true})}\right) \varphi_k^{(\mathrm{true})} + R_\alpha d_{k+1}^{\delta} \nonumber \\ & + R_\alpha \left((H^{(\mathrm{true})} - H) \varphi_{k+1}^{(\mathrm{true})} + H(\varphi_{k+1}^{(\mathrm{true})} - \varphi_{k+1}^{(b)})\right). \tag{1.34} \end{align}$$

We treat the last term in (1.33) similarly to the first term in (1.34). Then, collecting all parts, we derive (1.32). □

If the model error and the error in the observation operator in Theorem 1.8 is excluded, we obtain

$$e_{k+1} = R_\alpha d_{k+1}^\delta + (I - R_\alpha H)\, M_k e_k\,,$$

and, taking norms and using $\|d_k^\delta\| \leq \delta$, this is precisely the regularization error arising in Tikhonov regularization (1.19). If we select an appropriate value for α, this error can be made very small.

However, in many (practical) cases, the errors arising from the model and the observation operator are much bigger than the regularization error. *Model error*, in particular, can be very large due to insufficient resolution and inaccuracies in the physical model dynamics. This is specifically the case for a chaotic behavior of the system. The model error is a very important part of the total error and a very active area of current research (see, for example, [14, 27, 52, 80, 87]).

We also notice that even if there is no model error, no observation error and no data error, then $e_{k+1} = (I - R_\alpha H)M_k e_k$, and the errors can accumulate if α is chosen too large, in particular, if $\|(I - R_\alpha H)M_k\| > 1$ (see also [60, 67]). Note that for any regularization scheme, condition (1.16) holds and therefore α needs to be chosen small enough.

We have shown that within cycled data assimilation schemes, various forms of errors occur and influence each other which is important to consider when applying data assimilation methods in practice.

We will see in Section 6 that cycled four-dimensional variational data assimilation can be covered by the same framework of error analysis since cycled 4DVar is a form of cycled nonlinear Tikhonov regularization.

In the remainder of this article, we assume that no model error is present, that is, the model operator M_k represents the perfect model dynamics.

5 Bayesian approach to inverse problems

Probability theory provides a wide set of tools which can be used to solve inverse problems. In particular, the Bayesian theory has become quite popular as a generic approach which can be applied to inverse and ill-posed problems as well (see, for example, [5, 12, 75, 85]).

Bayesian theory has the potential to provide a stochastic background for many ideas which might appear *ad hoc* in the area of deterministic inverse problems and functional analysis. Also, Bayesian theory provides much more than just a solution to the inverse or data assimilation problem, but a full-grown theory to calculate estimates for the uncertainty as well.

However, we will see that all algorithms which can be formulated on a Bayesian background have their deterministic counterpart and, alternatively, can be studied purely within the framework of functional analysis and optimization. In this section, we apply Bayesian ideas to the observation and background errors.

Let us consider the equation

$$H(\varphi) = f, \tag{1.35}$$

as introduced in (1.1) as a starting point, where in this section we assume that $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$, $m, n \in \mathbb{N}$. The more general case with probability measures on infinite-dimensional spaces can be done formally in a similar way, but involves some nontrivial technicalities.

In the stochastic framework, the task of inverting equation (1.35) given some measurement f does not ask for *one* special solution. Since f is just one draw from some random distribution π_Y, any particular solution is of limited value and significance, but we want to know the *conditional probability distribution* of φ given some information about the error distribution of f. This conditional distribution can then be used either to calculate an *expectation value* for φ given f or to evaluate the *uncertainty* of this estimate measured, for example, by its *variance*.

We need to formulate our setup in more detail and with well-defined spaces and operators. Stochastic theory assumes that the quantity φ is a *random variable* on some probability space (Ω, Σ, P) with values in X. Here, Σ denotes some σ-algebra and P is a probability measure which maps any subset $A \subset \Omega$ for which $A \in \Sigma$ into a number $P(A) \in [0, 1]$. Also, $P(A)$ is the probability of the set A. We then obtain a probability P_X of the values of φ to be in some set $C \subset X$ by

$$P_X(\varphi \in C) := P(\{\omega : \varphi(\omega) \in C\}). \tag{1.36}$$

We also assume that the measurement f is a random variable with some probability distribution P_Y on Y. This probability distribution will depend on the true value $f^{(\text{true})}$ and is our model for measurement error during the process of measuring f. Here, we assume that the probability distribution (1.36) on X has a probability density $\pi_X : X \to [0, 1]$ such that

$$P_X(C) = \int_C \pi_X(\varphi) d\varphi\,, \tag{1.37}$$

for every open subset $C \subset X$. In the same way, we assume that P_Y has a probability density π_Y on Y such that

$$P_Y(U) = \int_U \pi_Y(f) df\,,$$

for every open subset $U \subset Y$. Usually, for simplicity, we drop the letters X and Y.

Clearly, since the conditional probability of some event $C \subset X$ given some event $\tilde{C} \subset X$ is defined by $P(C|\tilde{C}) := P(C \cap \tilde{C})/P(\tilde{C})$, we have that the conditional probability of event C given U is

$$P(C|U) = \frac{P\left(\{\omega : \varphi(\omega) \in C \text{ and } f(\omega) \in U\}\right)}{P\left(\{\omega : f(\omega) \in U\}\right)}\,,$$

where $P(\{\omega : f(\omega) \in U\}) > 0$. In terms of the *probability density functions* (PDFs), the conditional probability is formulated by

$$\pi(\varphi|f) = \frac{\pi(\varphi, f)}{\pi(f)}, \tag{1.38}$$

where $\pi(\varphi, f)$ is the joint probability density of φ and f living on the space $X \times Y$ and $\pi(f) \neq 0$ is the probability density of f in X. Equation (1.38) also holds with the role of φ and f exchanged, i.e. we have

$$\pi(f|\varphi) = \frac{\pi(\varphi, f)}{\pi(\varphi)}, \tag{1.39}$$

assuming that $\pi(\varphi) \neq 0$. Now, from equations (1.38) and (1.39), we get the famous *Bayes' formula* for conditional probability densities, that is,

$$\pi(\varphi|f) = \frac{\pi(\varphi)\pi(f|\varphi)}{\pi(f)}. \tag{1.40}$$

Note that the value of $\pi(f)$ can be obtained by the knowledge that the integral of $\pi(\varphi|f)$ over the whole space X should be equal to one, i.e. it is not necessary to know $\pi(f)$ (it is merely a normalizing constant).

Bayes' formula now provides a "simple" solution to the stochastic inverse problem of inverting equation (1.35). Given a probability density $\pi(\varphi)$ on X and some error density π on Y which can be used to calculate the density of the data distribution (often called the "measurement model" in statistics),

$$\pi(f|\varphi) = \pi(f - H(\varphi)). \tag{1.41}$$

We employ (1.40) to calculate the conditional probability density function $\pi(\varphi|f)$. This probability density is also known as *posterior density* or *analysis density function*. It is the density of the unobservable $\varphi \in X$ given the data $f \in Y$, that is, the probability of observing the data f as a function of φ. The density function $\pi(\varphi)$ on X is denoted as *prior* density. The posterior density is considered as the solution to the inverse problem.

Remark 1.9. Note that Bayes' formula seems to provide a very easy and stable approach to solving the inverse problem. The calculation of the posterior density $\pi(\varphi|f)$ is obtained by a *multiplication* of two given distributions $\pi(\varphi)$ and $\pi(f - H(\varphi))$. However, the calculation of the mean of the posterior distribution involves the solution of an ill-posed equation. In general, the full ill-posedness of the task is implicitly involved in Bayes' data assimilation as it is in all other schemes as well.

We can now formulate a general approach to data assimilation based on Bayes' formula.

Definition 1.10 (Bayes' data assimilation). Bayes' data assimilation determines probability density functions $\pi_k^{(a)}$ at time t_k for the states $\varphi \in X$ given data $f_k \in Y$ at time t_k by cycling the following propagation and analysis steps:

(i) *Propagation Step.* Calculate the prior density $\pi_k^{(b)}(\varphi)$ at time t_k by propagating the analysis density $\pi_{k-1}^{(a)}$ from time t_{k-1} to t_k based on the (linear or nonlinear) model dynamics M_{k-1}.

(ii) *Analysis Step.* Calculate the posterior or *analysis density* $\pi_k^{(a)}(\varphi|f_k)$ at time t_k by Bayes' formula (1.40) using the measurement model (1.41).

An important special case of Bayes' formula is the setup where all densities are *normal* or *Gaussian* distributions. For the prior distribution, we assume that it is a multivariate Gaussian distribution, that is, the probability density function is given by

$$\pi(\varphi) = \frac{1}{\sqrt{(2\pi)^n \det(B)}} e^{-\frac{1}{2}(\varphi-\mu)^T B^{-1}(\varphi-\mu)}, \quad \varphi \in \mathbb{R}^n \tag{1.42}$$

around some state $\mu := \varphi^{(b)} \in X = \mathbb{R}^n$ with some symmetric positive define matrix B. Gaussian densities are completely determined by their mean value $\mu = \mathbb{E}(\varphi) \in \mathbb{R}^n$ and the matrix B, which is well known to be the covariance matrix, that is,

$$B = \mathbb{E}\left((\varphi - \mu)(\varphi - \mu)^T\right), \tag{1.43}$$

of the Gaussian distribution (1.42). We write $\varphi \sim \mathcal{N}(\mu, B)$. The normalization is based on the integral formula

$$\int_{\mathbb{R}^n} e^{-\frac{1}{2}\varphi^T B^{-1}\varphi} d\varphi = \sqrt{\frac{(2\pi)^n}{\det(B^{-1})}} = \sqrt{(2\pi)^n \det(B)}\,.$$

Let us study the case where the probability density $\pi(f|\varphi)$ of the measurements f is also given by a Gaussian distribution with probability density function

$$\pi(f|\varphi) = \frac{1}{\sqrt{(2\pi)^m \det(R)}} e^{-\frac{1}{2}(f-H(\varphi))^T R^{-1}(f-H(\varphi))}, \quad f \in \mathbb{R}^m, \tag{1.44}$$

around the values $H(\varphi) \in Y = \mathbb{R}^m$ with the symmetric positive definite covariance matrix $R \in \mathbb{R}^{m\times m}$ of the observation error. Then, according to Bayes' formula (1.40), we obtain

$$\pi(\varphi|f) \propto \exp\left\{-\frac{1}{2}((\varphi-\mu)^T B^{-1}(\varphi-\mu) + (f - H(\varphi))^T R^{-1}(f - H(\varphi)))\right\}$$

for the probability density function of the posterior distribution. If H is *linear*, this is again a normal distribution with probability density

$$\pi(\varphi|f) \propto \exp\left\{-\frac{1}{2}\left(\varphi - \tilde{\mu}\right)^T \tilde{B}^{-1}\left(\varphi - \tilde{\mu}\right)\right\}.$$

Using $\mu = \varphi^{(b)}$, its mean $\tilde{\mu}$ is given by

$$\tilde{\mu} = \varphi^{(b)} + BH^*(R + HBH^*)^{-1}(f - H\varphi^{(b)}) = \varphi^{(b)} + K(f - H\varphi^{(b)}), \tag{1.45}$$

and its covariance matrix $\tilde{B}$ is given by

$$\tilde{B} = (B^{-1} + H^*R^{-1}H)^{-1} = (I - KH)B, \tag{1.46}$$

where $K = BH^*(R + HBH^*)^{-1}$ is called the (Kalman) gain. The proof of (1.45) and (1.46) will be worked out in detail in Section 7 on the Kalman filter, see equations (1.77) and (1.79). The equivalence of the two different expressions in (1.46) can also be obtained via the Sherman–Morrison–Woodbury formula (see, for example, [31]), though here it is worked out elementarily in Lemma 1.16. We summarize the above arguments in the following theorem.

Theorem 1.11 (Bayes' data assimilation for Gaussian probability densities). *In the case of a linear observation operator H, assume that the prior distribution is Gaussian with probability density function $\pi(\varphi)$ and the same is true for the distribution of the measurements with probability density function $\pi(f|\varphi)$ as given in* (1.44). *Then, the posterior distribution with density function $\pi(\varphi|f)$ is Gaussian as well. Its mean is calculated by the update formula* (1.45) *and its covariance matrix is given by* (1.46).

Note that the update formula (1.45) for the mean of the posterior Gaussian distribution is the same as for the update vector (or reconstruction) $\varphi_k^{(a)}$ obtained from (cycled) Tikhonov regularization (1.31), which is equivalent to 3DVar. In this respect, we see that Bayes' data assimilation gives more information by calculating a whole probability distribution of a state estimate, whereas Tikhonov regularization/3DVar only provides the mean of the estimate.

Further, when the dynamics M of a dynamical system is *linear*, then it maps a Gaussian distribution into a Gaussian distribution. The covariance matrix B in (1.45) and (1.46) needs to be replaced by its transported version $B^{(b)}$ calculated from the matrix B at the previous assimilation step by $B^{(b)} := MBM^*$. The propagation $B^{(b)}$ arises from the definition of the covariance matrix (1.43) and the linearity of the expected value. In this case, we can formulate the full cycling of the Bayesian approach explicitly.

Definition 1.12 (Gaussian Bayes' data assimilation for linear systems). For linear dynamical systems M_k and linear observation operators H_k, we start with some prior distribution with probability density function $\pi_0^{(a)}(\varphi)$ given by its mean $\varphi_0^{(a)}$ and its covariance matrix $B_0^{(a)}$. Then, for $k = 1, 2, 3, \ldots$, we carry out Bayes' data assimilation by cycling the following propagation and analysis steps.

(i) *Propagation Step.* Calculate the mean state $\varphi_k^{(b)}$ and the covariance matrix $B_k^{(b)}$ of the prior density $\pi_k^{(b)}(\varphi)$ at time t_k by

$$\varphi_k^{(b)} = M_{k-1}\varphi_{k-1}^{(a)}, \quad B_k^{(b)} := M_{k-1}B_{k-1}^{(a)}M_{k-1}^*. \tag{1.47}$$

(ii) *Analysis Step*. Calculate the Gaussian posterior or *analysis density* $\pi_k^{(a)}(\varphi|f_k)$ at time t_k by its mean and covariance

$$\varphi_k^{(a)} := \varphi_k^{(b)} + B_k^{(b)} H_k^* \left(R + H_k B_k^{(b)} H_k^*\right)^{-1} \left(f_k - H_k \varphi_k^{(b)}\right), \tag{1.48}$$

$$\left(B_k^{(a)}\right)^{-1} := \left(B_k^{(b)}\right)^{-1} + H_k^* R^{-1} H_k. \tag{1.49}$$

The above calculations treat the case of linear systems. Of course, Bayes' formula also works for nonlinear dynamics and nonlinear observation operators for which the numerics is much more difficult to carry out efficiently. A numerical method to approximately calculate the densities by *ensemble approaches* will be introduced in Section 8.

6 4DVar

A natural approach to the solution of a time-dependent state estimation problem is to put all available measurements into one big minimization problem. Given measurements $f_{k+1}, \dots, f_{k+K} \in Y$, this leads to

$$J_k(\varphi) := \left\|\varphi - \varphi_k^{(b)}\right\|_X^2 + \sum_{j=1}^{K} \left\|f_{k+j} - HM_{k+j,k}(\varphi)\right\|_Y^2, \tag{1.50}$$

where $M_{k+j,k}$ is defined in (1.2). For simplicity, we use a fixed (possibly nonlinear) observation operator H. Similar to the approach in Section 1, we can rewrite the problem (1.50) in a 3DVar type form like (1.21) by putting all the measurements $f_{k+1}, \dots, f_{k+K}$ into one long vector and removing the sum and defining a new (possibly nonlinear) operator $\overline{H}_k$, that is,

$$J_k(\varphi) := \left\|\varphi - \varphi_k^{(b)}\right\|_X^2 + \left\|\overline{f}_k - \overline{H}_k(\varphi)\right\|_Y^2,$$

where

$$\overline{f}_k = \begin{bmatrix} f_{k+1} \\ f_{k+2} \\ \vdots \\ f_{k+K} \end{bmatrix} \quad \text{and} \quad \overline{H}_k = \begin{bmatrix} HM_{k+1,k} \\ HM_{k+2,k} \\ \vdots \\ HM_{k+K,k} \end{bmatrix}.$$

The minimization of (1.50) corresponds to the fit of the full dynamic trajectory of the states to the given measurements f_{k+j}, $j = 1, \dots, K$ over the time window between t_k and t_{k+K}. As in Section 3, we can transform the functional (1.50) into a (generally nonlinear) Tikhonov functional of the form (1.15), for example, [28, 45]. Note that sometimes the observation f_k at time step t_k is included in the sum (here, in the functional (1.50) it is not included).

Denote the minimum of (1.50) by $\varphi_k^{(a)}$. A cycling of the assimilation is then obtained by using a new background at time t_{k+K} defined by

$$\varphi_{k+K}^{(b)} := M_{k+K,k}\left(\varphi_k^{(a)}\right), \tag{1.51}$$

for $k = 0, K, 2K, 3K, \ldots$. The process of minimizing the functional (1.50) and using the minimizing ϕ as the initial condition for the forecast is known as four-dimensional variational data assimilation (4DVar) [6, 19, 50, 51, 72]. The repeated minimization of (1.50) combined with (1.51) is then a cycled 4DVar scheme. As we can write 4DVar in the form of 3DVar, this is merely a form of (nonlinear) cycled Tikhonov regularization as shown in Section 3.

Usually, the minimization of (1.50) is carried out by a *gradient method*, that is, we calculate the gradient $\nabla_\varphi J_k(\varphi)|_{\varphi^{(\ell)}}$ at points $\varphi^{(\ell)}$ in the state space and update

$$\varphi^{(\ell+1)} := \varphi^{(\ell)} - h\nabla_\varphi J_k(\varphi)|_{\varphi^{(\ell)}} \tag{1.52}$$

with some appropriately chosen step-size $h > 0$ and starting guess $\varphi^{(0)}$ (often $\varphi^{(0)} := \varphi_k^{(b)}$ is used).

For simplicity, we consider the case where $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$, and the scalar products are the l^2 scalar products. Let us study terms of the form

$$g(\varphi) := ||f - HM\varphi||_Y^2 , \tag{1.53}$$

with $f \in Y$ and some linear operator $M : X \to X$. The gradient of $g(\varphi)$ with respect to φ is given by

$$\nabla_\varphi g(\varphi) = -2\Big(M^* H^* (f - HM\varphi)\Big). \tag{1.54}$$

If M is a nonlinear operator, then we obtain the nonlinear version

$$\nabla_\varphi g(\varphi) = -2\left(\Big(\frac{dM(\varphi)}{d\varphi}\Big)^* H^* \Big(f - HM(\varphi)\Big)\right) \tag{1.55}$$

of (1.54), where $dM(\varphi)/d\varphi$ denotes the Fréchet derivative of $M(\varphi)$ with respect to φ. The derivative

$$\mathbf{M}(\varphi) := \frac{dM(\varphi)}{d\varphi} \tag{1.56}$$

is also known as the *tangent linear* model [26, 50].

For many applications, the dynamical model is given as a system of ordinary differential equations in the form

$$\dot{\varphi} = F(\varphi), \quad \varphi(0) = \varphi_0. \tag{1.57}$$

Since the model dynamics is given by $\varphi(t) = M_{t,0}(\varphi(0)) = M_{t,0}(\varphi_0)$, this means that

$$F(\varphi) = \frac{d}{dt} M_{t,0}(\varphi_0). \tag{1.58}$$

We denote the derivative with respect to the initial state φ_0 by

$$\varphi'(t) := \frac{d\varphi}{d\varphi_0}. \tag{1.59}$$

Note that φ' is a linear mapping from X into X; when $X = \mathbb{R}^n$, it is the $n \times n$-matrix with elements $\partial\varphi_j/\partial\varphi_{0,i}$ for $i, j = 1, \dots, n$.

We assume that the solution $\varphi = \varphi(t)$ is continuously differentiable with respect to the initial state φ_0 as well as with respect to the time t. In this case, we can exchange the differentiation with respect to time t and the initial state φ_0 and, differentiating (1.57) with respect to φ_0, we obtain

$$\frac{d\,\dot{\varphi}}{d\varphi_0} = \frac{d}{d\varphi_0}\frac{d}{dt}M_{t,0}(\varphi_0) = \frac{d}{dt}\frac{d}{d\varphi_0}M_{t,0}(\varphi_0) = \frac{d}{dt}\varphi'(t). \tag{1.60}$$

Therefore, the time evolution of the derivative φ' is given by

$$\frac{d}{dt}\varphi'(t) = \frac{d}{d\varphi_0}F(\varphi(t)) = F'(\varphi(t))\frac{d\varphi(t)}{d\varphi_0}. \tag{1.61}$$

At time $t = 0$, this is equal to $F'(\varphi_0) = dF(\varphi)/d\varphi_0|_{\varphi=\varphi_0}$, that is,

$$\frac{d}{dt}\varphi'(t)|_{t=0} = F'(\varphi_0). \tag{1.62}$$

This means that the tangent linear model φ' can be calculated by solving the system

$$\frac{d}{dt}\varphi'(t) = F'(\varphi(t))\varphi'(t), \quad t \geq 0 \tag{1.63}$$

of ordinary differential equations with initial condition $\varphi'(0) = I$ and with the solution φ of the original system of equations (1.57). Using $\varphi(t) = M_{t,0}(\varphi_0)$ and $\varphi'(t) = dM_{t,0}(\varphi_0)/d\varphi_0$ as well as (1.56), we obtain

$$\varphi'(t) = \frac{dM_{t,0}(\varphi_0)}{d\varphi_0} =: \mathbf{M}_{t,0}(\varphi_0)$$

for the tangent linear model.

We remark that the tangent linear adjoint is an $n \times n$ matrix which might be huge when n is large. Thus, efficient methods for its evaluation need to be setup. To evaluate the adjoint in (1.54), we define a function $\psi(t) \in X$ on the interval $[t_{k+1}, t_k]$ by

$$\dot{\psi} = -F'(\varphi(t))^* \psi(t), \tag{1.64}$$

with *final condition*

$$\psi(t_{k+1}) = H^*(f_{k+1} - HM(\varphi_k)). \tag{1.65}$$

Lemma 1.13. *For $t \in [t_k, t_{k+1}]$, the inner product*

$$h(t) := \left\langle \varphi'(t)(\delta\varphi_0), \psi(t) \right\rangle$$

is constant over time for any $\delta\varphi_0 \in X$.

Proof. We differentiate $h(t)$ with respect to t and calculate

$$\begin{aligned}\frac{dh(t)}{dt} &= \frac{d}{dt}\Big\langle \varphi'(t)(\delta\varphi_0), \psi(t)\Big\rangle \\ &= \Big\langle \frac{d}{dt}\varphi'(t)(\delta\varphi_0), \psi(t)\Big\rangle + \Big\langle \varphi'(t)(\delta\varphi_0), \frac{d}{dt}\psi(t)\Big\rangle \\ &= \Big\langle F'(\varphi(t))\varphi'(t)(\delta\varphi_0), \psi(t)\Big\rangle + \Big\langle \varphi'(t)(\delta\varphi_0), -F'(\varphi(t))^*\psi(t)\Big\rangle \\ &= \Big\langle \varphi'(t)(\delta\varphi_0), F'(\varphi(t))^*\psi(t)\Big\rangle - \Big\langle \varphi'(t)(\delta\varphi_0), F'(\varphi(t))^*\psi(t)\Big\rangle \\ &= 0 \end{aligned} \tag{1.66}$$

where we have used (1.63) and (1.64). Since the derivative of $h(t)$ is zero by (1.66), we obtain the statement of the lemma. □

Let e_j, $j = 1, \dots, n$ be the canonical basis of $\mathbb{R}^n$. We can now calculate the gradient ∇g of (1.54) by

$$\begin{aligned}\nabla g_j(\varphi_k) &= -2\Big\langle \varphi'(t_{k+1})e_j, H^*(f_{k+1} - HM(\varphi_k))\Big\rangle \\ &= -2\Big\langle \varphi'(t_{k+1})e_j, \psi(t_{k+1})\Big\rangle \\ &= -2\Big\langle \varphi'(t_k)e_j, \psi(t_k)\Big\rangle \\ &= -2\Big\langle e_j, \psi(t_k)\Big\rangle = -2\psi(t_k)_j \end{aligned} \tag{1.67}$$

for $j = 1, \dots, n$. Thus, the gradient is calculated by propagating the field forward in time by (1.57), then propagating the observation error back by (1.64), (1.65) and calculating the gradient using (1.67).

In general, we consider the time step t_k as the initial time step or, subsequently, the intermediate time step, and thus (1.57) becomes

$$\dot{\varphi} = F(\varphi), \quad \varphi(0) = \varphi_k, \quad \text{where} \quad \varphi_k := \varphi(t_k), \tag{1.68}$$

and the derivative $'$ with respect to the initial state φ_k is given by $\varphi'(t) := \frac{d\varphi}{d\varphi_k}$. Hence, discretizing (1.68) using, for example, a simple finite difference between time steps t_k and t_{k+1} leads to

$$\frac{\varphi_{k+1} - \varphi_k}{\Delta t} = F(\varphi_k)\,, \tag{1.69}$$

and therefore the discretized model operator M_k from time step t_k to time step t_{k+1} is given by

$$\varphi_{k+1} = \varphi_k + \Delta t F(\varphi_k) = M_k(\varphi_k) = M_{k+1,k}(\varphi_k)\,.$$

Moreover, discretizing (1.63) leads to

$$\frac{\varphi'_{k+1} - \varphi'_k}{\Delta t} = F'(\varphi_k). \tag{1.70}$$

Hence, using $\varphi'_k = d\varphi'_k/d\varphi'_k = I$, the (discretized) tangent linear model is given by

$$\varphi'_{k+1} = I + \Delta t F'(\varphi_k) = \mathbf{M}_k(\varphi_k) = \mathbf{M}_{k+1,k}(\varphi_k) := \frac{dM_k}{d\varphi}|_{\varphi_k} = \frac{dM_{k+1,k}}{d\varphi}|_{\varphi_k}\,,$$

which can also be obtained by differentiating (1.69) with respect to φ_k. Note that we can similarly find the (nonlinear) operators $M_{k+j,k}$ and their tangent linear models $\mathbf{M}_{k+j,k}(\varphi_k) := \frac{dM_{k+j,k}}{d\varphi}|_{\varphi_k}$ for any $j = 1,\dots,K$, and, by the chain rule applied to (1.2), it follows that

$$\mathbf{M}_{k+j,k}(\varphi_k) = \mathbf{M}_{k+j,k+j-1}\cdots\mathbf{M}_{k+2,k+1}\mathbf{M}_{k+1,k}(\varphi_k)\,.$$

Studying the case $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$, and using the weighted scalar product (1.25) and (1.26), we may compute the gradient $\nabla_\varphi J_k(\varphi)$ of the full functional $J_k(\varphi)$ given in (1.50) by

$$\nabla_\varphi J_k(\varphi) := 2B^{-1}\left(\varphi - \varphi_k^{(b)}\right) - 2\sum_{j=1}^{K}\mathbf{M}_{k+j,k}(\varphi)^* H^* R^{-1}\left(f_{k+j} - HM_{k+j,k}(\varphi)\right)\,. \tag{1.71}$$

A gradient method like (1.52) can then be used to obtain a local minimizer for the functional $J_k(\varphi)$ in (1.50). Another method which may be used to find a local minimum of $J_k(\varphi)$ in (1.50) is the Gauss–Newton method [21]. We solve $\nabla_\varphi J_k(\varphi) = 0$ in order to find the minimum of (1.50) using Newton's method, that is,

$$\varphi^{(\ell+1)} := \varphi^{(\ell)} - \left(\nabla\nabla_\varphi J_k(\varphi)|_{\varphi^{(\ell)}}\right)^{-1}\nabla_\varphi J_k(\varphi)|_{\varphi^{(\ell)}}\,,$$

with some starting guess $\varphi^{(0)}$ where $\nabla\nabla_\varphi J_k(\varphi)|_{\varphi^{(\ell)}}$ is the Jacobian of $\nabla_\varphi J_k(\varphi)$ at $\varphi^{(\ell)}$, that is, the Hessian. Usually, the starting guess $\varphi^{(0)} = \varphi_k^{(b)}$ is taken. Often, instead of the correct Hessian $\nabla\nabla_\varphi J_k(\varphi)|_{\varphi^{(\ell)}}$, an approximate version is used, neglecting terms involving the gradient of the tangent linear model, thereby leading to a quasi-Newton method. The gradient method usually only gives linear convergence. The Gauss–Newton method with approximate Hessian converges superlinearly for well-posed problems and a sufficiently close starting guess. For linear observation operators H and linear model dynamics M_k, the Newton and Gauss–Newton method are the same and any local minimizer of (1.50) is clearly also a global minimizer (see, for example, [32]) and the convergence speed to the global minimum is quadratic.

7 Kalman filter and Kalman smoother

The Kalman filter is a method to solve the data assimilation problem (1.3) similarly to the cycled Tikhonov regularization, 3DVar or 4DVar. But in addition to calculating an analysis in every step, it also iteratively updates the norm of the state space to include the knowledge from previous assimilation cycles.

We can introduce the Kalman filter using deterministic and stochastic arguments. Here, we will start with a deterministic approach, which also proves equivalence of the Kalman filter and Kalman smoother to the four-dimensional variational data assimilation for linear model dynamics $M_k : X \to X$ and linear observation operators $H : X \to Y$. Then, we discuss a stochastic approach to the Kalman filter.

Let us study assimilation for a *linear* model dynamics M_k, a linear observation operator H and measurements f_1 and f_2 at times t_1 and t_2. Then, four-dimensional variational data assimilation with weighted norms as in Section 3 minimizes the functional (1.50)

$$J_{4\mathrm{DVar}}(\varphi) := \left\|\varphi - \varphi_0^{(b)}\right\|_{B^{-1}}^2 + \|f_1 - HM_0\varphi\|_{R^{-1}}^2 + \|f_2 - HM_1M_0\varphi\|_{R^{-1}}^2, \quad (1.72)$$

with $B \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{m\times m}$. Alternatively, we study the assimilation of the data f_1 in a first step by minimization of

$$J_1(\varphi) := \left\|\varphi - \varphi_0^{(b)}\right\|_{B^{-1}}^2 + \|f_1 - HM_0\varphi\|_{R^{-1}}^2, \quad (1.73)$$

with minimizer $\tilde{\varphi}^{(a)}$ and the assimilation of f_2 in a *second step* by minimizing

$$J_2(\varphi) := \left\|\varphi - \tilde{\varphi}^{(a)}\right\|_{\tilde{B}^{-1}}^2 + \|f_2 - HM_1M_0\varphi\|_{R^{-1}}^2, \quad (1.74)$$

with a weight matrix $\tilde{B}$. The key question here is to determine the new weight $\tilde{B}$ such that the minimizer of J_2 is equal to the minimizer of the full functional J_{4DVar} in (1.72). This is the case if we can choose $\tilde{B}$ such that $J_2(\varphi) = J_{4DVar}(\varphi) + c$ with some constant c, where J_1 is implicitly used via $\tilde{\varphi}^{(a)}$ in (1.74). The problem is solved if we can determine $\tilde{\varphi}^{(a)}$ and $\tilde{B}$ such that J_1 and the first term of J_2 are identical. Starting with J_1, we obtain

$$\begin{aligned} J_1(\varphi) &= \left\langle \varphi - \varphi_0^{(b)}, B^{-1}(\varphi - \varphi_0^{(b)}\right\rangle \\ &\quad + \left\langle f_1 - HM_0\varphi, R^{-1}(f_1 - HM_0\varphi)\right\rangle \\ &= \left\langle \varphi, (B^{-1} + M_0^* H^* R^{-1} HM_0)\varphi\right\rangle \\ &\quad - 2\left\langle \varphi, B^{-1}\varphi_0^{(b)} + M_0^* H^* R^{-1} f_1\right\rangle + c, \end{aligned} \quad (1.75)$$

with some constant c independent of φ. The first term of J_2 is given by

$$\left\|\varphi - \tilde{\varphi}^{(a)}\right\|_{\tilde{B}^{-1}}^2 = \left\langle \varphi, \tilde{B}^{-1}\varphi\right\rangle - 2\left\langle \varphi, \tilde{B}^{-1}\tilde{\varphi}^{(a)}\right\rangle + \tilde{c}, \quad (1.76)$$

with some constant $\tilde{c}$ not depending on φ. A comparison of the coefficients of the quadratic and linear terms in (1.75) and (1.76) immediately shows that with

$$\tilde{B}^{-1} := B^{-1} + M_0^* H^* R^{-1} HM_0 \quad (1.77)$$

and

$$\tilde{B}^{-1}\tilde{\varphi}^{(a)} := B^{-1}\varphi_0^{(b)} + M_0^* H^*, R^{-1} f_1 \quad (1.78)$$

the functional J_1 given by (1.75) and the first term of the functional J_2 given by (1.76) are the same up to some constant not depending on φ. Finally, from (1.78) using (1.77), we derive

$$\begin{aligned}\tilde{\varphi}^{(a)} &= \tilde{B}\left(B^{-1}\varphi_0^{(b)} + M_0^* H^* R^{-1} f_1\right) \\ &= (I + BM_0^* H^* R^{-1} HM_0)^{-1}\left(\varphi_0^{(b)} + BM_0^* H^* R^{-1} f_1\right).\end{aligned} \tag{1.79}$$

After some algebraic manipulations inserting

$$I = \left(I + BM_0^* H^* R^{-1} HM_0\right) - BM_0^* H^* R^{-1} HM_0,$$

we obtain

$$\begin{aligned}\tilde{\varphi}^{(a)} &= \varphi_0^{(b)} + \left(I + BM_0^* H^* R^{-1} HM_0\right)^{-1} BM_0^* H^* R^{-1}\left(f_1 - HM_0\varphi_0^{(b)}\right) \\ &= \varphi_0^{(b)} + BM_0^* H^* \left(R + HM_0 BM_0^* H^*\right)^{-1}\left(f_1 - HM_0\varphi_0^{(b)}\right),\end{aligned}$$

which is the minimizer of J_1 as in (1.29) or (1.31) when the propagation M_0 from φ_0 at time t_0 to φ_1 at time t_1, that is, $\varphi_1 = M_0\varphi_0$ is used. The above approach can be carried out successively for the measurements f_1, f_2, f_3 etc. This sequential approach leads to the Kalman smoother (see, for example, [27, 53, 59]). We will see later in Theorem 1.18 that the Kalman smoother is equivalent to the Kalman filter at the final time.

Definition 1.14 (Kalman smoother (KS)). Let $H_k : X \to Y$ and $M_k : X \to X$, $k = 0, 1, 2, \dots$ given in Definition 1.1 be linear and assume that measurements $f_1, f_2, \dots$ at times $t_1, t_2, \dots$ are given. Then, we calculate weight matrices

$$\tilde{B}_k^{-1} := \tilde{B}_{k-1}^{-1} + M_{k,0}^* H_k^* R^{-1} H_k M_{k,0}\,, \quad k = 1, 2, \dots, \tag{1.80}$$

with $\tilde{B}_0 := B$, where $M_{k,0}$ is defined in (1.2), and analysis states $\tilde{\varphi}_k^{(a)}$ at time t_k defined by

$$\begin{aligned}\tilde{\varphi}_k^{(a)} :&= \tilde{\varphi}_{k-1}^{(a)} \\ &+ \tilde{B}_{k-1} M_{k,0}^* H_k^* \left(R + H_k M_{k,0} \tilde{B}_{k-1} M_{k,0}^* H_k^*\right)^{-1}\left(f_k - H_k M_{k,0}\tilde{\varphi}_{k-1}^{(a)}\right)\end{aligned} \tag{1.81}$$

for $k = 1, 2, \dots$ with $\tilde{\varphi}_0^{(a)} := \varphi_0^{(b)}$.

From our derivation, it is clear that the following theorem holds.

Theorem 1.15 (Equivalence of 4DVar and Kalman smoother). *Let H_k and M_k for $k = 0, 1, 2, \dots$ be linear operators and data $f_1, f_2, \dots$ be given. Then, 4DVar carried out with data $f_1, \dots, f_k$ is equivalent to the Kalman smoother given in Definition* 1.14 *in the sense that the minimum of the 4DVar functional taking $k = 0$ and $k = K$ in* (1.50) *is given by the analysis $\tilde{\varphi}_k^{(a)}$ for $k = 1, 2, \dots, K$ according to* (1.81).

Proof. The proof for $k = 1$ is given in equations (1.72) to (1.79). The general case is directly obtained by iterating the arguments. □

In Definition 1.14, we worked with states at time t_0. Usually, the states of the Kalman filter are calculated at times t_1, t_2 etc. We need to propagate the states $\tilde{\varphi}_k^{(a)}$ from time t_0 to t_k by

$$\varphi_k^{(b)} = M_{k,0}\tilde{\varphi}_{k-1}^{(a)}, \quad \text{and} \quad \varphi_k^{(a)} = M_{k,0}\tilde{\varphi}_k^{(a)}, \tag{1.82}$$

for $k = 1, 2, 3, \dots$, which means that

$$\varphi_k^{(b)} = M_{k-1}\left(\varphi_{k-1}^{(a)}\right) \tag{1.83}$$

propagates the state from t_{k-1} to t_k (see also (1.20)). The matrices $\tilde{B}$ are propagated from t_0 to t_k by

$$B_k^{(b)} = M_{k,0}\tilde{B}_{k-1}M_{k,0}^*, \quad \text{and} \quad B_k^{(a)} = M_{k,0}\tilde{B}_k M_{k,0}^*, \tag{1.84}$$

for $k = 1, 2, 3, \dots$, where the *background matrix* at time t_k is obtained by propagating the *analysis matrix* from time t_{k-1} to t_k by

$$B_k^{(b)} = M_{k-1}B_{k-1}^{(a)}M_{k-1}^*. \tag{1.85}$$

Note that the propagation of the state (1.83) and the propagation of the weight matrix (1.85) are equivalent to the propagation step in Bayes' data assimilation for Gaussian probability densities and linear systems, see (1.47).

Using (1.82) and (1.84), the iterative version of (1.81) is then given by

$$\varphi_k^{(a)} = \varphi_k^{(b)} + B_k^{(b)}H_k^*\left(R + H_k B_k^{(b)}H_k^*\right)^{-1}\left(f_k - H_k\varphi_k^{(b)}\right), \tag{1.86}$$

for $k \in \mathbb{N}$, often written in the form

$$\varphi_k^{(a)} = \varphi_k^{(b)} + K_k\left(f_k - H_k\varphi_k^{(b)}\right) \tag{1.87}$$

with the *Kalman gain matrix*

$$K_k := B_k^{(b)}H_k^*\left(R + H_k B_k^{(b)}H_k^*\right)^{-1}. \tag{1.88}$$

Note that the Kalman gain matrix is identical to the Tikhonov regularization matrix (1.31). Using (1.85) and (1.80), we readily verify that the analysis matrix $B_k^{(a)}$ at time t_k is obtained from the *background matrix* $B_k^{(b)}$ at time t_k by

$$\left(B_k^{(a)}\right)^{-1} = \left(B_k^{(b)}\right)^{-1} + H_k^* R^{-1} H_k, \tag{1.89}$$

for $k \in \mathbb{N}$. Note that the analysis matrix $B_k^{(a)}$ in (1.89) and the analysis state $\varphi_k^{(a)}$ in (1.86) is equivalent to the updated covariance matrix and the updated state in the analysis step in Bayes' data assimilation for Gaussian probability densities and linear systems, see (1.48 and (1.49)).

Often, another version of (1.89) is used, where the matrices appear without their inverse (see also (1.46)).

Lemma 1.16. *For $k \in N$ and $B_k^{(a)}$ in* (1.89), *we have*

$$B_k^{(a)} = (I - K_k H_k) B_k^{(b)}, \tag{1.90}$$

where K_k is given by (1.88).

Proof. We start from (1.89) in the form

$$B_k^{(a)} = \left(I + B_k^{(b)} H_k^* R^{-1} H_k\right)^{-1} B_k^{(b)}. \tag{1.91}$$

We expand

$$\begin{aligned} T := {} & \left(I + B_k^{(b)} H_k^* R^{-1} H_k\right)(I - K_k H_k) \\ = {} & \left(I + B_k^{(b)} H_k^* R^{-1} H_k\right)\left(I - B_k^{(b)} H_k^* (R + H_k B_k^{(b)} H_k^*)^{-1} H_k\right) \\ = {} & I + \underbrace{B_k^{(b)} H_k^* R^{-1} H_k}_{=:S} - \underbrace{B_k^{(b)} H_k^* \left(R + H_k B_k^{(b)} H_k^*\right)^{-1} H_k}_{=:S_1} \\ & - \underbrace{B_k^{(b)} H_k^* R^{-1} H_k B_k^{(b)} H_k^* \left(R + H_k B_k^{(b)} H_k^*\right)^{-1} H_k}_{:=S_2} \end{aligned} \tag{1.92}$$

and remark that

$$S = B_k^{(b)} H_k^* R^{-1} \left(R + H_k B_k^{(b)} H_k^*\right)\left(R + H_k B_k^{(b)} H_k^*\right)^{-1} H_k = S_1 + S_2,$$

yielding $T = I$. Thus,

$$\left(I + B_k^{(b)} H_k^* R^{-1} H_k\right)^{-1} = (I - K_k H_k)$$

and the proof is complete. □

We are now ready to define the Kalman filter (see, for example, [2, 39, 53]).

Definition 1.17 (Kalman filter). Starting with an initial state $\varphi_0^{(b)}$ and an initial weight matrix $B_0^{(a)} := B$, for $k \in \mathbb{N}$, the Kalman filter iteratively calculates an analysis $\varphi_k^{(a)}$ at time t_k for $k = 1, 2, \ldots$ by

(i) propagating the state $\varphi_{k-1}^{(a)}$ from t_{k-1} to t_k via (1.83):

$$\varphi_k^{(b)} = M_{k-1}\left(\varphi_{k-1}^{(a)}\right),$$

(ii) propagating $B_{k-1}^{(a)}$ from t_{k-1} to t_k following (1.85):

$$B_k^{(b)} = M_{k-1} B_{k-1}^{(a)} M_{k-1}^*,$$

(iii) calculating the Kalman gain by (1.88):

$$K_k = B_k^{(b)} H_k^* \left(R + H_k B_k^{(b)} H_k^*\right)^{-1},$$

(iv) calculating an analysis state by (1.86):

$$\varphi_k^{(a)} = \varphi_k^{(b)} + K_k \left(f_k - H_k \varphi_k^{(b)} \right),$$

(v) calculating an analysis weight by (1.90):

$$B_k^{(a)} = (I - K_k H_k)\, B_k^{(b)}.$$

The first two steps of the Kalman filter are often referred to as the *predictor steps* as they predict a state and a covariance estimate by propagating them forward via the model dynamics. The last two steps are called analysis steps to update the state and covariance estimate.

The relationship between the Kalman filter, the Kalman smoother and 4DVar is summarized in the following theorem.

Theorem 1.18 (Equivalence of 4DVar, Kalman filter and Kalman smoother). *Let the operators $H_k : X \to Y$ for $k \in \mathbb{N}$ and $M_k : X \to X$ for $k \in \mathbb{N}_0$ be linear. Let $\varphi_k^{(a)}$ be the analysis of the Kalman filter at time t_k, $\tilde{\varphi}_k^{(a)}$ the analysis of the Kalman smoother with data $f_1, \dots, f_k$ at time t_0, $\tilde{\varphi}_{\text{4DVar},k}^{(a)}$ the minimizer of the 4DVar functional* (1.50) *at time t_0 and define*

$$\varphi_{\text{4DVar},k}^{(a)} := M_{k,0}\tilde{\varphi}_{\text{4DVar},k}^{(a)}, \quad k = 1, 2, 3, \dots \tag{1.93}$$

Then, 4DVar is equivalent to the Kalman filter and to the Kalman smoother in the sense that

$$\varphi_{\text{4DVar},k}^{(a)} = \varphi_k^{(a)} = M_{k,0}\tilde{\varphi}_k^{(a)}, \tag{1.94}$$

if we start the iterations with the same initial background state $\varphi_0^{(b)}$ and the same initial error covariance matrix $B_0^{(a)} := B$.

Proof. The equivalence of the Kalman smoother with the Kalman filter is obtained by our reformulation based on (1.82) worked out in equations (1.85) to (1.90). The equivalence to 4DVar is then a consequence of Theorem 1.15. □

Theorem 1.18 states that the Kalman smoother is equivalent to the Kalman filter (and 4DVar) at the end of some time window for linear operators.

We finally consider the stochastic approach to the Kalman filter, which we formulate as a basic theorem. Observing that the formulas for Bayes' data assimilation with Gaussian densities as given in Definition 1.12 are identical to the update formulas for the Kalman filter according to Definition 1.17, the proof of this result is straightforward.

Theorem 1.19 (Equivalence of Kalman filter and Bayes' data assimilation). *For linear systems $M_k : X \to X$, linear observation operators $H_k : X \to Y$, and Gaussian probability densities, the* Kalman filter *as given in Definition* 1.17 *is identical to* Bayes' data assimilation *given by Definition* 1.12.

For nonlinear system dynamics $M_k : X \to X$ and nonlinear observation operators $H_k : X \to Y$, the above equivalences do not hold any more. However, we may still apply the Kalman filter if we linearize both the model M_k and the observation operator H_k about the considered state. This leads to the Extended Kalman Filter (EKF) [2, 46]. The linearizations of the model operator M_k and the observation operator H_k, which are used within the Kalman filter (1.17), are given by

$$\mathbf{M}_k(\varphi_k) := \frac{dM_k}{d\varphi}|_{\varphi_k} \quad \text{and} \quad \mathbf{M}_k(\varphi_k) := \frac{dH_k}{d\varphi}|_{\varphi_k},$$

where $\mathbf{M}_k$ is the tangent linear model (1.56).

We have introduced several data assimilation methods and shown that for linear systems, they are all essentially equivalent to cycled Tikhonov regularization with a weighted norm. In the next section, we consider ensemble methods which provide a way of (approximately) updating probability distributions and covariance matrices within the assimilation schemes.

8 Ensemble methods

We have introduced several methods for data assimilation in the previous sections, including Tikhonov data assimilation, 3DVar, 4DVar, Bayes' data assimilation and the Kalman filter.

Evaluating the different approaches, we note that 3DVar or Tikhonov data assimilation works with fixed norms at every time step and do not fully include all the dynamic information which is available from previous assimilations. Since 4DVar uses full trajectories over some time window, it implicitly includes such information and we can expect it to be superior to the simple 3DVar. However, Bayes' data assimilation or the Kalman filter are equivalent to 4DVar for linear systems and include all available information by updating the weight matrices and propagating them through time. This is essentially done implicitly in 4DVar. In general, we can expect them to yield results comparable to those of 4DVar.

The need to propagate some probability distribution is a characteristic feature of the Bayes' data assimilation and the Kalman filter. It is also their main challenge since the matrices $B_k^{(a)}$ or $B_k^{(b)}$ have dimension $n \times n$, which for large n is usually not feasible in terms of computation time or storage, even when supercomputers are employed for the calculation as in most operational centers for atmospheric data assimilation. Thus, a key need for these methods is to formulate algorithms which give a reasonable approximation to the weight matrices $B_k^{(b)}$ with less computational costs than by the use of (1.85) and (1.89) or (1.90).

Often, the approach to ensemble methods is carried out via stochastic estimators. Here, we want to stay within the framework of the previous sections and study the ensemble approach from the viewpoint of applied mathematics. The stochastic view will be discussed in a second step. One of the most popular ensemble filter techniques is the Ensemble Kalman filter [3, 11, 24, 25, 41–43, 65, 70, 77, 84].

Definition 1.20 (Ensemble). An *ensemble* with N members is any finite set of vectors $\varphi^{(\ell)} \in X$ for $\ell = 1, \dots, N$. We can propagate the ensemble through time by applying the model dynamics $M : X \to X$ or $M_k : X \to X$, respectively. Starting with an *initial ensemble* $\varphi_0^{(\ell)}$, $\ell = 1, \dots, N$, this leads to ensemble members

$$\varphi_k^{(\ell)} = M_{k-1}\varphi_{k-1}^{(\ell)}, \quad k = 1, 2, 3, \dots \tag{1.95}$$

for $\ell = 1, \dots, N$.

We start with the construction of a particular family of ensembles generated by the eigenvalue decomposition of the weight matrix $B := B^{(b)}$ defined in Section 7 with $X = \mathbb{R}^n$. B is a self-adjoint and a positive definite matrix, hence, there is a complete set of eigenvectors of B, i.e. we have vectors $\psi^{(1)}, \dots, \psi^{(n)} \in X$ and eigenvalues $\lambda^{(1)}, \dots, \lambda^{(n)}$ such that

$$B\psi^{(\ell)} = \lambda^{(\ell)}\psi^{(\ell)}, \quad \ell = 1, \dots, n. \tag{1.96}$$

The eigenvalues are real valued and positive and we will always assume that they are ordered according to their size $\lambda^{(1)} \geq \lambda^{(2)} \geq \cdots \geq \lambda^{(n)}$. With the matrix $\Lambda := \mathrm{diag}[\sqrt{\lambda^{(1)}}, \dots, \sqrt{\lambda^{(n)}}]$ and the orthogonal matrix $U := [\psi^{(1)}, \dots, \psi^{(n)}]$, we obtain

$$B = U\Lambda^2 U^* = (U\Lambda)(U\Lambda)^*, \tag{1.97}$$

where we note that $U^* = U^{-1}$. This representation corresponds to the well-known *principle component analysis* of the *quadratic form* defined by

$$E(\varphi, \psi) := \varphi^T B\psi, \quad \varphi, \psi \in X. \tag{1.98}$$

Geometrically, B defines a hypersurface of second-order with positive eigenvalues, whose level curves form a family of $n - 1$-dimensional ellipses in X. The principal axis of this ellipse are given by the eigenvectors $\psi^{(l)}$, $\ell = 1, \dots, n$.

The application of B to some vector $\varphi \in X$ according to (1.97) is carried out by a projection of φ onto the principle axis $\psi^{(\ell)}$ of B, followed by the multiplication with $\lambda^{(\ell)}$. This setup can be a basis for further insight to construct a low-dimensional approximation of B.

Before we continue the ensemble construction, we first need to discuss the *metric* in which we want an approximation of the B-matrix. We remark that the role of B in the Kalman filter is mainly in the update formulas (1.85), (1.86) and (1.90). Here, to obtain a good approximation of the vector updates in L^2, we need B to be approximated in the operator norm based on L^2 on $X = \mathbb{R}^n$. That is what we will use as the basis for the following arguments.

Lemma 1.21. *We construct an ensemble of vectors by choosing the $N-1$ maximal eigenvalues of B and its corresponding eigenvectors $\psi^{(1)}, \dots, \psi^{(N-1)}$. We define*

$$Q := \left[\sqrt{\lambda^{(1)}}\psi^{(1)}, \dots, \sqrt{\lambda^{(N-1)}}\psi^{(N-1)}\right]. \tag{1.99}$$

Then, we have the error estimate

$$||B - QQ^*|| = \sup_{j=N,\dots,n} \left|\lambda^{(j)}\right| = \left|\lambda^{(N)}\right| = \lambda^{(N)}. \tag{1.100}$$

Proof. The proof is obtained from

$$B - QQ^* = U\tilde{\Lambda}^2 U^*, \tag{1.101}$$

with $\tilde{\Lambda}^2 = \mathrm{diag}[0, \dots, 0, \lambda^{(N)}, \lambda^{(N+1)}, \dots, \lambda^{(n)}]$, where there are $N-1$ zeros on the diagonal of $\tilde{\Lambda}$. Since U is an orthogonal matrix, the norm estimate (1.100) is straightforward. □

We are now going to use arbitrary ensembles $\varphi^{(1)}, \dots, \varphi^{(N)}$ and construct approximate weight matrices. From the Courant minimum-maximum principle, we know that

$$\lambda^{(\ell)} = \min_{\dim U = \ell-1} \max_{\varphi \in U^\perp, \|\varphi\|=1} \langle \varphi, B\varphi \rangle . \tag{1.102}$$

For an arbitrary ensemble $\varphi^{(1)}, \dots, \varphi^{(N)}$, we use the mean

$$\mu = \frac{1}{N} \sum_{\ell=1}^{N} \varphi^{(\ell)} \tag{1.103}$$

to define the *ensemble matrix*

$$Q := \left[\varphi^{(1)} - \mu, \dots, \varphi^{(N)} - \mu\right], \tag{1.104}$$

and we define the *ensemble subspace* U_Q by

$$U_Q = \mathrm{span}\left\{\varphi^{(1)} - \mu, \dots, \varphi^{(N)} - \mu\right\} . \tag{1.105}$$

We call the vectors $\varphi^{(\ell)} - \mu$, $\ell = 1, \dots, N$ the *centered ensemble*. We remark that $\dim U_Q = N - 1$. Then, we have

$$\begin{aligned} ||B - QQ^*|| &\geq \sup_{B\varphi \perp U_Q, \|\varphi\|=1} ||(B - QQ^*)\varphi|| \\ &\geq \sup_{B\varphi \perp U_Q, \|\varphi\|=1} \|B\varphi\| \\ &\geq \sup_{B\varphi \perp U_Q, \|\varphi\|=1} \langle \varphi, B\varphi \rangle \\ &\geq \min_{\dim U = N-1} \sup_{\varphi \perp U, \|\varphi\|=1} \langle \varphi, B\varphi \rangle \\ &= \lambda^{(N)}. \end{aligned} \tag{1.106}$$

The above results are summarized in the following theorem.

Theorem 1.22. *Let the eigenvalues $\lambda^{(1)} \geq \lambda^{(2)} \geq \cdots \geq \lambda^{(n)}$ of the self-adjoint weight matrix B be ordered according to its size and let $\varphi^{(1)}, \ldots, \varphi^{(N)}$ with $N \in \mathbb{N}$ be an arbitrary ensemble of states in X. Then, the error for the approximation of the weight matrix B by QQ^* with Q defined in* (1.104) *is estimated by*

$$\|B - QQ^*\|_2 \geq \lambda^{(N)}. \tag{1.107}$$

Remark 1.23. The optimal error $\lambda^{(N)}$ can be achieved if the centered ensemble spans the space of the $N-1$ eigenvectors $\psi^{(1)}, \ldots, \psi^{(N-1)}$ of B corresponding to the largest eigenvalues $\lambda^{(1)}, \ldots, \lambda^{(N-1)}$ with appropriate coefficients as in (1.99).

Ensembles can be used to approximate the weight matrix $B_{k+1}^{(b)}$ when the weight matrix $B_k^{(a)}$ is given, see (1.85). If $B_k^{(a)}$ is approximated by the ensemble $\varphi_k^{(1)}, \ldots, \varphi_k^{(N)}$ in the form

$$B_k^{(a)} \approx Q_k^{(a)} \left(Q_k^{(a)}\right)^*, \tag{1.108}$$

with $Q_k^{(a)} := [(\varphi^{(1)})^{(a)} - \mu^{(a)}, \ldots, (\varphi^{(N)})^{(a)} - \mu^{(a)}]$, then by (1.85), we derive an approximation for $B_{k+1}^{(b)}$ by

$$\begin{aligned} B_{k+1}^{(b)} &= M_k B_k^{(a)} M_k^* \\ &\approx M_k Q_k^{(a)} \left(Q_k^{(a)}\right)^* M_k^* \\ &= M_k Q_k^{(a)} \left(M_k Q_k^{(a)}\right)^* \\ &= Q_{k+1}^{(b)} \left(Q_{k+1}^{(b)}\right)^*, \end{aligned} \tag{1.109}$$

where $Q_{k+1}^{(b)} = M_k Q_k^{(a)}$.

Lemma 1.24. *Consider the approximation of $B_k^{(a)}$ by an ensemble $\varphi_k^{(1)}, \ldots, \varphi_k^{(N)}$ with ensemble matrix $Q_k^{(a)}$. If the error satisfies*

$$\left\| B_k^{(a)} - Q_k^{(a)} \left(Q_k^{(a)}\right)^* \right\| \leq \epsilon, \tag{1.110}$$

for some $\epsilon > 0$, then the error estimate for the propagated ensemble at time t_{k+1} is given by

$$\left\| B_{k+1}^{(b)} - Q_{k+1}^{(b)} \left(Q_{k+1}^{(b)}\right)^* \right\| \leq \left\|M_k\right\| \left\|M_k^*\right\| \epsilon. \tag{1.111}$$

Proof. Based on (1.109), the proof is straightforward. □

A key question of ensemble methods is how to update the ensemble in the data assimilation step. Given the data f_k at time t_k, how do we get an ensemble which approximates the analysis covariance matrix $B_k^{(a)}$ given an ensemble which approximates the background error covariance matrix $B_k^{(b)}$? We know that for the Kalman filter, the analysis covariance matrix $B_k^{(a)}$ is calculated from $B_k^{(b)}$ by (1.90). In terms of the ensemble approximations, this means

$$Q_k^{(a)} \left(Q_k^{(a)}\right)^* = (I - K_k H_k)\, Q_k^{(b)} \left(Q_k^{(b)}\right)^* \tag{1.112}$$

with the *ensemble Kalman matrix*

$$K_k := Q_k^{(b)}\left(Q_k^{(b)}\right)^* H_k^*\left(R + H_k Q_k^{(b)}\left(Q_k^{(b)}\right)^* H_k^*\right)^{-1}, \tag{1.113}$$

leading to

$$\begin{aligned} & Q_k^{(a)}\left(Q_k^{(a)}\right)^* \\ & = Q_k^{(b)}\underbrace{\left\{I - \left(Q_k^{(b)}\right)^* H_k^*\left(R + H_k Q_k^{(b)}\left(Q_k^{(b)}\right)^* H_k^*\right)^{-1} H_k Q_k^{(b)}\right\}}_{=:T}(Q_k^{(b)})^*. \end{aligned} \tag{1.114}$$

The matrix T in the curly brackets is self-adjoint and positive semidefinite, and hence there exists a matrix L such that $T = LL^*$. This finally leads to

$$Q_k^{(a)} = Q_k^{(b)} L, \tag{1.115}$$

which we denote as *square root filter* [4, 8, 65, 79].

Lemma 1.25. *Assume that $\varphi_k^{(1)}, \dots, \varphi_k^{(N)}$ is an ensemble which satisfies*

$$\left\|B_k^{(b)} - Q_k^{(b)}\left(Q_k^{(b)}\right)^*\right\| \le \epsilon, \tag{1.116}$$

with some $\epsilon < \|B_k^{(b)}\|$. Then, for the analysis ensemble defined by (1.115)*, we have*

$$\left\|B_k^{(a)} - Q_k^{(a)}\left(Q_k^{(a)}\right)^*\right\| \le C\epsilon, \tag{1.117}$$

with some constant C not depending on $Q_k^{(a)}$.

Proof. Using the notation $K_k^{(\text{true})}$ for the Kalman gain matrix in the general case ((1.88) and (1.90)), and $Q_k^{(a)}(Q_k^{(a)})^*$ from (1.112), we write

$$\begin{aligned} B_k^{(a)} - Q_k^{(a)}\left(Q_k^{(a)}\right)^* &= \left(I - K_k^{(\text{true})} H_k\right)\left(B_k^{(b)} - Q_k^{(b)}\left(Q_k^{(b)}\right)^*\right) \\ &\quad + \left(K_k - K_k^{(\text{true})}\right) H_k Q_k^{(b)}\left(Q_k^{(b)}\right)^*, \end{aligned} \tag{1.118}$$

with K_k defined by (1.113). We remark that due to its special structure, the norm of the inverse $(R + H_k Q_k^{(b)}(Q_k^{(b)})^* H_k^*)^{-1}$ in (1.113) is bounded uniformly independent of $Q_k^{(b)}$. Furthermore, using $\epsilon < \|B_k^{(b)}\|$, the norm

$$\begin{aligned} \left\|Q_k^{(b)}\left(Q_k^{(b)}\right)^*\right\| &= \left\|B_k^{(b)} + \left(Q_k^{(b)}\left(Q_k^{(b)}\right)^* - B_k^{(b)}\right)\right\| \\ &\le \left\|B_k^{(b)}\right\| + \epsilon \\ &\le 2\left\|B_k^{(b)}\right\| \end{aligned} \tag{1.119}$$

is bounded uniformly, leading to

$$\left\|K_k^{(\mathrm{true})} - K_k\right\| \le c\epsilon\,, \tag{1.120}$$

with a constant c not depending on $Q_k^{(b)}$. Finally, a similar estimate applied to (1.118) yields the desired result (1.117) and the proof is complete. □

For further insight into ensemble methods, we refer to the article [69] in this book.

9 Numerical examples

We examine data assimilation techniques discussed in this article and their relation to inverse problem theory for simple model problems. First, we consider an advection–diffusion equation in Section 9.1 and then the Lorenz-95 system in Section 9.2.

9.1 Data assimilation for an advection-diffusion system

Consider the following linear (one-dimensional) advection-diffusion problem (see, for example, [15]). The system dynamics are described by

$$\frac{\partial}{\partial t}\varphi(x,t) = \nu\frac{\partial^2}{\partial x^2}\varphi(x,t) - a\frac{\partial}{\partial x}\varphi(x,t) \tag{1.121}$$

for $x \in (0,1)$ and $t \in (0,T)$. As boundary and initial conditions, we have

$$\begin{aligned}
&\varphi(0,t) = 0, \quad t \in (0,T),\\
&\varphi(1,t) = 0, \quad t \in (0,T),\\
&\varphi(x,0) = \varphi_0(x), \quad x \in (0,1)\,.
\end{aligned}$$

Here, $\nu > 0$ is the diffusion coefficient and a is the advection parameter. We want to determine the initial condition φ_0 from the measurements of the solution $\varphi(x,t)$ at certain points in space and time. Let $0 = x_0 < x_1 \cdots < x_n = 1$ and $x_i = ih$, $i = 0,\dots,n+1$ and $h = \frac{1}{n+1}$. With the discretizations of the spatial derivatives

$$\frac{\partial^2}{\partial x^2}\varphi \approx \frac{\varphi^{i+1} - 2\varphi^i + \varphi^{i-1}}{h^2}, \quad \text{and} \quad \frac{\partial}{\partial x}\varphi \approx \frac{\varphi^i - \varphi^{i-1}}{h}\,,$$

for $i = 0,\dots,n$, we obtain a system of ordinary differential equations of the form

$$\dot{\varphi}(t) = F(\varphi), \quad t \in (0,T], \quad \varphi(0) = \varphi_0\,, \tag{1.122}$$

where, in this case, $F(\varphi) = K\varphi(t)$, that is, F is linear, with

$$K = \begin{bmatrix} -2\frac{\nu}{h^2} - \frac{a}{h} & \frac{\nu}{h^2} & & & & \\ \frac{\nu}{h^2} + \frac{a}{h} & -2\frac{\nu}{h^2} - \frac{a}{h} & \frac{\nu}{h^2} & & & \\ & \frac{\nu}{h^2} + \frac{a}{h} & -2\frac{\nu}{h^2} - \frac{a}{h} & \frac{\nu}{h^2} & & \\ & \ddots & \ddots & \ddots & & \\ & & \frac{\nu}{h^2} + \frac{a}{h} & -2\frac{\nu}{h^2} - \frac{a}{h} & \frac{\nu}{h^2} & \\ & & & \frac{\nu}{h^2} + \frac{a}{h} & -2\frac{\nu}{h^2} - \frac{a}{h} & \end{bmatrix} \in \mathbb{R}^{n+2\times n+2}$$

and $\varphi(t) = [\varphi^0(t), \dots, \varphi^{n+1}(t)]^T \in \mathbb{R}^{n+2}$. To satisfy the boundary conditions, we set $\varphi^0(t) = \varphi^{n+1}(t) = 0$ throughout. As an initial condition, we choose $\varphi^i(0) = \varphi_0(x_i)$, $i = 0 \dots, n$. The solution to the linear system of ordinary differential equations with constant coefficients (1.122) is given by

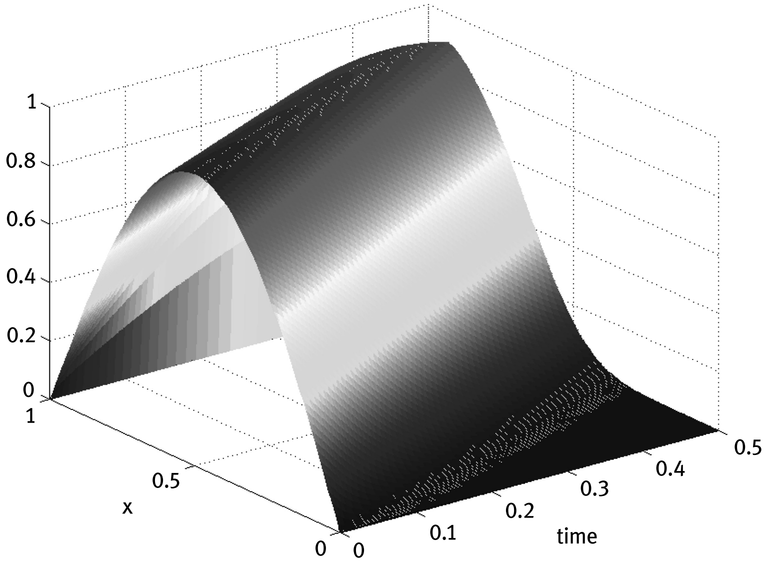


Figure 1.1: Solution of $\varphi(t) = (\exp Kt)\varphi_0$, $t \in [0, 0.5]$ (discretized advection–diffusion equation (1.121)) for initial condition $\varphi_0(x) = \sin(\pi x)$.

$$\varphi(t) = (\exp Kt)\varphi_0, \quad t \in [0, T], \tag{1.123}$$

where $\exp Kt \in \mathbb{R}^{n+2\times n+2}$, or, using an explicit first-order Euler scheme, we obtain the discrete linear model

$$\varphi_{k+1} = \varphi_k + \Delta t K \varphi_k, \quad k = 0, \dots, \frac{T}{\Delta t}, \tag{1.124}$$

where $\varphi_k = [\varphi_k^0, \dots, \varphi_k^{n+1}]^T \in \mathbb{R}^{n+2}$ and $\varphi_k^0 = \varphi_k^{n+1} = 0$ throughout. Note that we use a lower index to describe the time steps and an upper index to describe points in space/components of φ_k. The approach (1.124) is a more practical implementation as the analytical solution (1.123) would only be available for certain problems. We solve the advection-diffusion problem (1.121) (using the Forward Euler method) with $a = 1$, $\nu = 0.01$, $n = 100$, final time $T = 0.5$, time step $\Delta t = 0.001$ and initial condition $\varphi_0(x_i) = \sin(\pi x_i)$. The solution is shown in Figure 1.1.

For the inverse problem (data assimilation problem), we suppose we do not know the initial condition $\varphi_0(x)$. We want to estimate $\varphi_0(x)$ from measurements of r components $\varphi^{\frac{n}{r}}(t), \varphi^{2\frac{n}{r}}(t), \dots, \varphi^{n}(t)$ of the solution $\varphi(t)$ at times $t_1 = 0.002$, $t_2 = 0.004, \dots, t_m = 0.5$. For our experiment, we use $r = 5$, and hence we observe 5 out of $n = 100$ components. Take noisy measurements of $H\varphi(t_1), H\varphi(t_2), \dots,$ $H\varphi(t_m)$, where $H \in \mathbb{R}^{r\times n+2}$ is the observation operator matrix (which is linear in this case) given by $H_{ij} = 1$ if $j = \frac{n}{r}i$ and $H_{ij} = 0$ otherwise. We obtain the (linear) least squares problem

$$\min_{\varphi_0\in\mathbb{R}^{n+2}} \left\|\overline{H}\varphi_0 - f\right\|_2^2 , \tag{1.125}$$

with $\overline{H}$ and f for the forward Euler method and observations every second time step given by

$$\overline{H} = \begin{bmatrix} H(I + 2\Delta t K) \\ H(I + 2\Delta t K)^2 \\ \vdots \\ H(I + 2\Delta t K)^m \end{bmatrix} \in \mathbb{R}^{rm\times n+2} \quad \text{and} \quad f = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix} \in \mathbb{R}^{rm} .$$

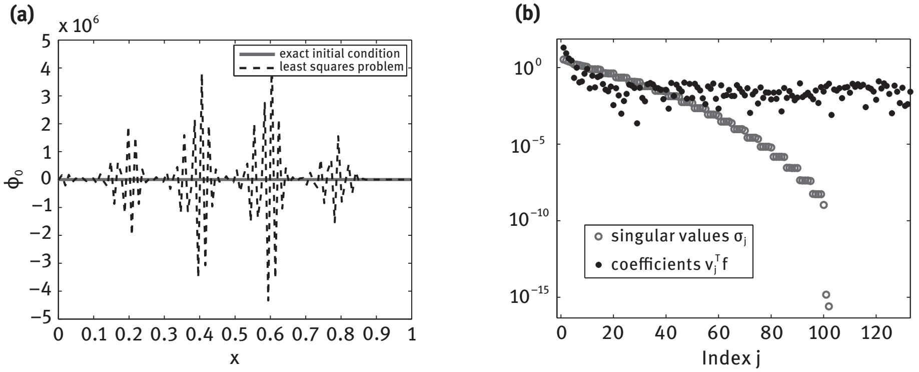
The observations are obtained using the output from the exact initial condition and the measurements usually contain noise (see Section 4 for a detailed description of the errors), that is, $f = f^{\delta} = f^{(\text{true})} + d^{\delta}$, where the noise is usually normally distributed, that is, $d^{\delta} \sim \mathcal{N}(0, \rho^2 I)$, where ρ is the standard deviation. If we solve the problem using a naive approach with a standard least squares implementation [74], we obtain the result in Figure 1.2 (a).

Using the singular value decomposition given in Lemma 1.3, we have $\overline{H} = V\Sigma U^*$ and, with $f = f^{(\text{true})} + d^{\delta}$, we obtain

$$\varphi_0^{\delta} = \sum_{j=1}^{n+2} \frac{\langle f^{\delta}, v_j\rangle_Y}{\sigma_j} u_j = \sum_{j=1}^{n+2} \left(\frac{v_j^T f^{(\text{true})}}{\sigma_j} + \frac{v_j^T d^{\delta}}{\sigma_j} \right) u_j ,$$

and clearly for small singular values σ_j, the noise is magnified, hence the naive solution in Figure 1.2 (a). Figure 1.2 (b) shows what happens for this particular example. The singular values σ_j decay rapidly and only the coefficients $|v_j^T f| = |v_j^T f^{\delta}|$ above the noise level (here we chose $d^{\delta} \sim \mathcal{N}(0, \rho^2 I)$ with $\rho = 0.1$) are useful and carry clear information about the data.

In order to compute a better solution φ_0 for the initial condition than the one given in Figure 1.2 (a), we apply Tikhonov regularization. From (1.31), the Tikhonov



Exact initial condition φ_0 and naive solution to the least squares problem.

Plots of the singular values σ_j and the coefficients $|v_j^T f|$, for $j = 1, \dots, n+2$.

Figure 1.2: Naive solution to the least squares problem (1.125) and singular values of $\overline{H}$.

regularized solution is given by

$$\varphi_0^{(a)} = \varphi_0^{(b)} + B\overline{H}^* \left(\alpha R + \overline{H}B\overline{H}^*\right)^{-1} \left(f - \overline{H}^* \varphi_0^{(b)}\right) .$$

For our problem, we use the observation error covariance matrix $R = 0.01I$ (in line with the noise on the observations). For this particular problem, we chose $\varphi_0^{(b)} = 1 - 0.5\pi^2(x - 0.5)^2$ for the background estimate, which is the truncated Taylor series expansion of the true initial condition φ_0. For the background error covariance matrix, we take B with entries $B_{ij} = 0.01 \times \exp(\frac{-|i-j|}{50})$ and for α, we choose the value $\alpha = 0.00359$ which minimizes both the total error consisting of perturbation error $\|R_\alpha d^\delta\|$ where $R_\alpha = B\overline{H}^*(\alpha R + \overline{H}B\overline{H}^*)^{-1}$ and regularization error $\|R_\alpha \overline{H}\varphi_0 - \varphi_0\|$,

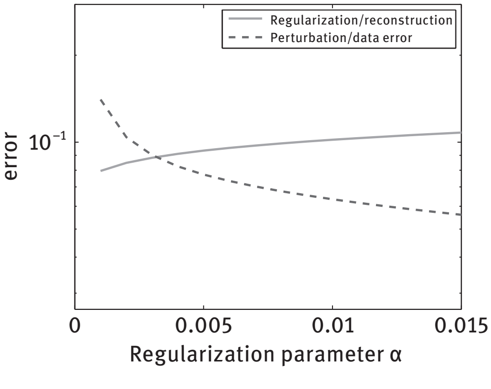
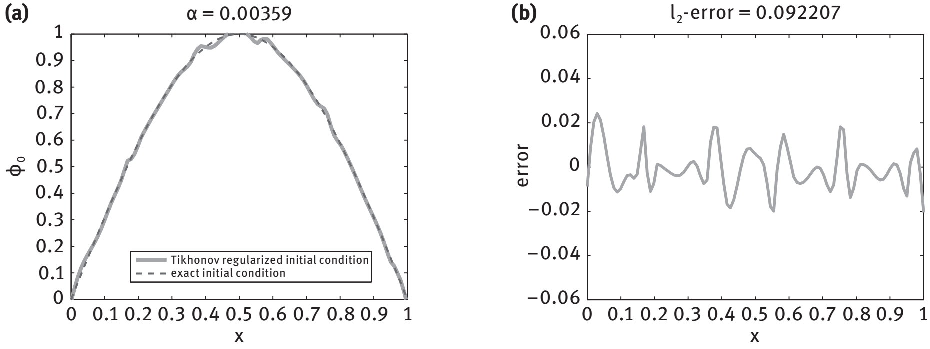


Figure 1.3: Regularization/reconstruction and data/measurement error for different values of α between 0 and 0.015. The optimal α in this case is found to be $\alpha = 0.00359$.



Exact initial condition and Tikhonov regularized solution.

Error between exact initial condition and regularized solution.

Figure 1.4: Exact initial condition and regularized solution for the regularization parameter $\alpha = 0.00359$ and the l_2-norm error between the exact and regularized solution for the linear advection equation (1.121).

see (1.19). The plots in Figure 1.3 show both the regularization and perturbation error for this problem. For the value $\alpha = 0.00359$, the reconstruction of the initial condition is plotted in Figure 1.4 (a) and the initial condition error is displayed in Figure 1.4 (b). Note that similar computations can be done using no background $\varphi_0^{(b)}$, the standard situation in Tikhonov regularization, different background estimates, as well as different choices for the background error covariance matrices B. For the choice of α, which corresponds to the choice of the Tikhonov regularization parameter, several heuristics are available, for example, the L-curve criterion [36], generalized cross-validation [30] and the discrepancy principle [61], where the latter is most appropriate for large scale computations.

We have essentially solved a 4DVar data assimilation problem, since we have shown in Section 6 that 4DVar can be written in the form of 3DVar which is merely a Tikhonov regularization, discussed in Section 3.

The situation described above was an ideal situation. In reality, models are nonlinear and imperfect, that is, they include model error. We give examples for these situations. First, consider a nonlinear problem. Instead of (1.121), consider

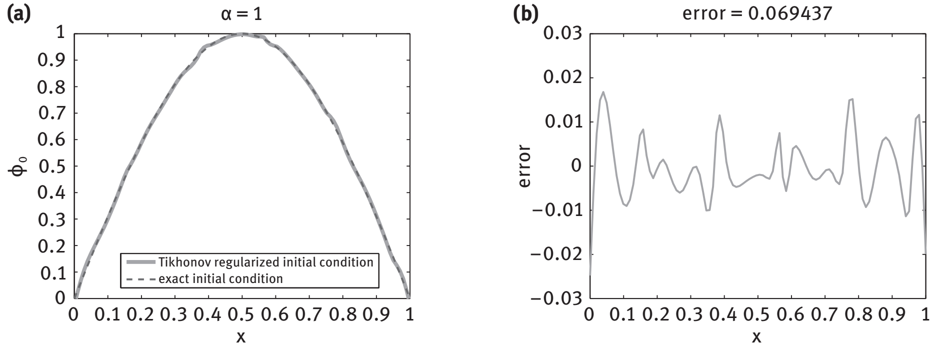
$$\frac{\partial}{\partial t}\varphi(x,t) = \nu\frac{\partial^2}{\partial x^2}\varphi(x,t) - a\frac{\partial}{\partial x}\varphi(x,t) + \varphi(x,t)^3\,,$$

and the discrete nonlinear problem becomes

$$\varphi_{k+1} = \varphi_k + \Delta t K\varphi_k + \varphi_k^3 = M_k(\varphi_k), \quad k = 0,\ldots,\frac{T}{\Delta t}\,. \tag{1.126}$$

We set up the nonlinear least squares problem

$$\min_{\varphi_0\in\mathbb{R}^{n+2}} \left\|\overline{H}(\varphi_0) - f\right\|_2^2\,,$$



Exact initial condition and Tikhonov regularized solution.

Error between exact initial condition and regularized solution.

Figure 1.5: Exact initial condition and regularized solution for the regularization parameter $\alpha = 1$ and the l_2-norm error between exact and regularized solution for the nonlinear advection equation.

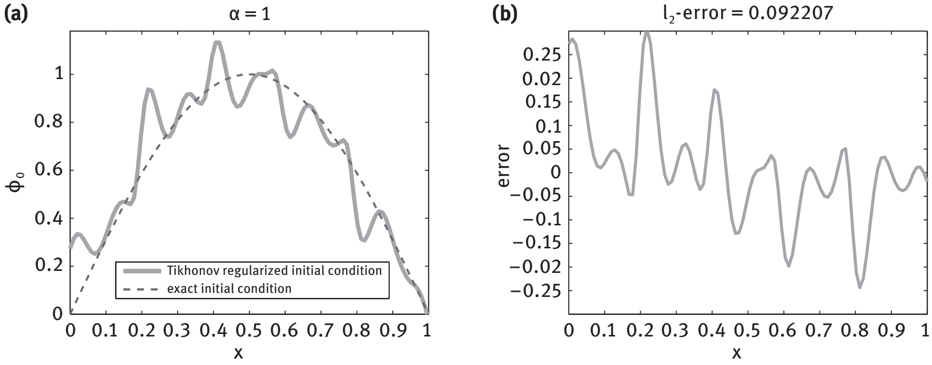
where here $\overline{H}$ is a nonlinear operator. The minimization problem can be solved using the Gauss–Newton method [21, 64]. The results for the reconstructed initial condition for the same data as for the linear problem are displayed in Figure 1.5 (a) and the initial condition error is displayed in Figure 1.5 (b).

Finally, consider the case where some model error is present. To this end, we assume that the observations are created by the true model for the nonlinear advection-diffusion equation (1.121) with $a = 1$, $\nu = 0.01$. The model used in the data assimilation process uses perturbed parameters $a^{\text{pert}} = 1.1$, $\nu^{\text{pert}} = 0.009$. The results for the reconstructed initial condition are shown in Figure 1.6 (a) and the initial condition error is displayed in Figure 1.6 (b). As the model contains an error, we are trying to fit an initial condition for the wrong model and hence the error for this problem is rather large, as seen in Figures 1.6 (a) and 1.6 (b).

However, in Figures 1.7 (a) and 1.7 (b), we see that this relatively large error in the initial condition does not lead to large errors in the solution. Figure 1.7 (a) shows the solution to the nonlinear advection equation with exact initial condition and Figure 1.7 (b) shows the solution with the perturbed initial condition obtained after solving the inverse (data assimilation) problem. We see that as the solution is propagated forward in time, the error in the initial condition is smoothed. The reason is the smoothing property of the forward operator. We have $\varphi_{k+1} = M_k(\varphi_k)$ where M_k is a linear (that is, $I + \Delta t K$) or a nonlinear (1.126) operator. If the initial condition is perturbed by ζ_k, then we have $\varphi_{k+1} + \zeta_{k+1} = M_k(\varphi_k + \zeta_k)$, and to leading order

$$\zeta_{k+1} = \mathbf{M}_k(\varphi_k)\zeta_k\,,$$

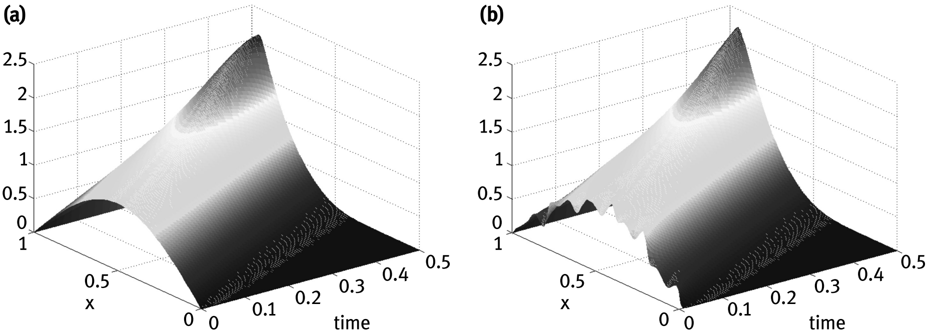
where $\mathbf{M}_k$ is the discretized tangent linear model. Assuming that $\mathbf{M}_k(\varphi_k) = \mathbf{M}$ (which holds for our linear example), then in the limit, we have $\zeta_k = \mathbf{M}^k\zeta_0$. From basic linear



Exact initial condition and Tikhonov regularized solution.

Error between exact initial condition and regularized solution.

Figure 1.6: Exact initial condition and regularized solution for the regularization parameter $\alpha = 1$ and the l_2-norm error between exact and regularized solution for the nonlinear advection equation when a model error is present.



Solution to nonlinear discretized advection-diffusion equation for initial condition $\varphi_0(x) = \sin(\pi x)$.

Solution to nonlinear discretized advection-diffusion equation for perturbed initial condition computed from data assimilation problem.

Figure 1.7: Solution to nonlinear advection-diffusion problem with exact and perturbed initial condition.

algebra [31], we have that $\zeta_k \to 0$ if $\rho(\mathbf{M}) < 1$, where $\rho(\mathbf{M}) = \max\{|\lambda|, \lambda \in \Lambda(\mathbf{M})\}$ is the spectral radius. In our example, both for the linear and linearized nonlinear model dynamics, the eigenvalues of $\mathbf{M}_k(\varphi_k)$ are within the unit circle, explaining the smoothing of the error in the initial condition as the solution propagates in time.

In the next example, we consider problems which are more sensitive to the initial conditions, that is, systems that exhibit chaotic dynamics (and hence more accurately represent the effects in, say, weather forecasting). One such system is the Lorenz-

95 model. In reality, we would expect a mix of situations arising from chaotic and smoothing systems.

9.2 Data assimilation for the Lorenz-95 system

As a second example, consider the Lorenz-95 system [55, 56], that is, a generalization of the well-known three-dimensional Lorenz-63 system [54]. The model is given by a system of N coupled nonlinear ordinary differential equations whose solution φ with components $\varphi = [\varphi^1, \dots, \varphi^N]$ satisfies

$$\frac{d\varphi^i}{dt} = -\varphi^{i-2}\varphi^{i-1} + \varphi^{i-1}\varphi^{i+1} - \varphi^i + f, \quad t \in (0,T],\ \varphi^i(0) = \varphi_0^i, \tag{1.127}$$

where $i = 0, \dots, N$, with cyclic boundary conditions $\varphi^0 = \varphi^N$, $\varphi^{-1} = \varphi^{N-1}$, $\varphi^{N+1} = \varphi^1$ and f is a forcing term. For a forcing term $f = 8$, the system is chaotic (i.e. it has positive Lyapunov exponents, see [76]). For $N = 40$, the system has 13 positive Lyapunov exponents. Lorenz [55] observed that this system has a similar error growth characteristic as an operational numerical weather prediction system if a time $T = 1$ is associated with 5 days.

We solve (1.127) using the classical 4th order explicit Runge–Kutta scheme, which gives

$$\varphi_{k+1} = M_k(\varphi_k), \quad \text{where} \quad \varphi_k = \left[\varphi_k^1, \dots, \varphi_k^N\right]^T, \tag{1.128}$$

and M_k is the nonlinear model operator which propagates φ_k to φ_{k+1}. The solution trajectory of two components of φ computed with the Runge–Kutta method, and $\Delta t = 0.01$ and $T = 21$ is displayed in Figure 1.8. In order to illustrate the chaotic dynamics of the Lorenz-95 model, we run it with slightly perturbed initial conditions. Perturbing the initial condition randomly with an error of about $10\,\%$ gives the ensemble of forecasts in Figure 1.9 (a) and using a perturbation of about $0.1\,\%$ gives the forecast ensemble in Figure 1.9 (b). We only show the trajectory of site 20.

The figures show an unperturbed solution trajectory and an ensemble where the initial conditions have been slightly perturbed. It is easy to see that the larger the perturbation in the initial condition, the more the error in the forecast grows. For this problem, the eigenvalues of the matrix $\mathbf{M}_k(\varphi_k)$ from the linearization of (1.128) are not necessarily within the unit disk.

We carry out some data assimilation experiments with this problem. First, consider the 4DVar minimization problem (1.50). We need to minimize

$$J(\varphi_0) := \left(\varphi_0 - \varphi_0^{(b)}\right)^T B^{-1} \left(\varphi_0 - \varphi_0^{(b)}\right) + \sum_{j=1}^{K} \left(f_j - H(\varphi_j)\right)^T R^{-1} \left(f_j - H(\varphi_j)\right), \tag{1.129}$$

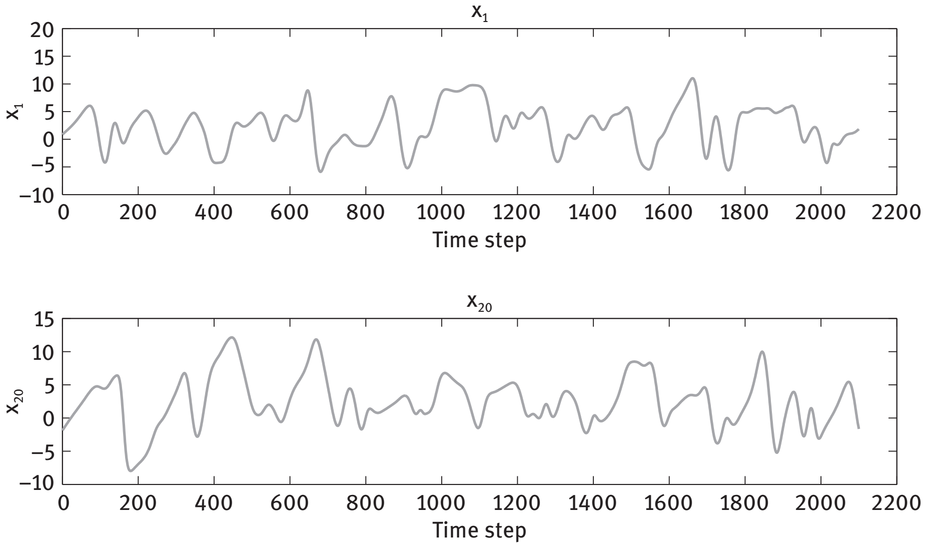
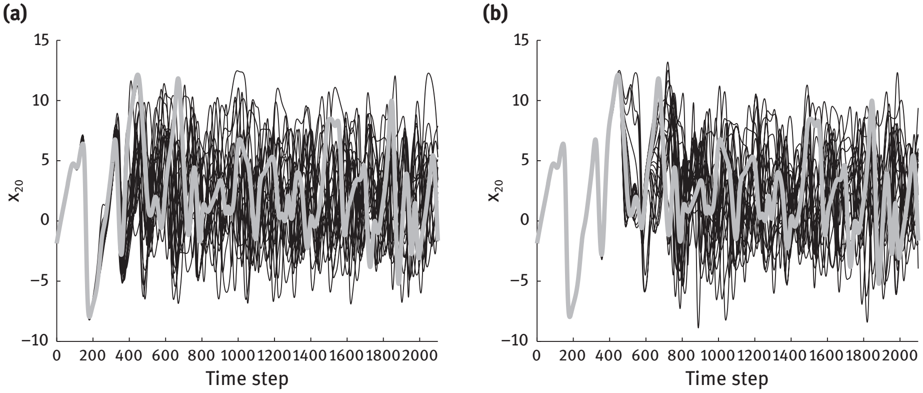


Figure 1.8: Components 1 and 20 of the solution to (1.127).



Forecast ensemble for an initial condition error of $10\,\%$.

Forecast ensemble for an initial condition error of $0.1\,\%$.

Figure 1.9: Trajectory of site 20 of Lorenz-95 system of size 40. Green thick line: unperturbed forecast. Black lines: Ensemble of 20 perturbed forecasts.

where $\varphi_j = M_{j-1}(\varphi_{j-1})$ is given by (1.128). We have

$$\nabla_{\varphi_0} J(\varphi_0) = 2B^{-1}(\varphi_0 - \varphi_0^{(b)}) - 2\sum_{j=1}^{K}(\mathbf{M}_{j,0}(\varphi_0)^T H^T R^{-1}(f_j - HM_{j,0}(\varphi_0)),$$

where $M_{j,0}$ is given by (1.2) and $\mathbf{M}_{j,0}$ is the tangent linear model. In order to minimize the cost function, we need $\nabla_{\varphi_0} J(\varphi_0)$ and in order to solve this problem, we apply Newton's method. The Hessian (or the Jacobian for Newton's method) is given by

$$\nabla\nabla_{\varphi_0} J(\varphi_0) = 2B^{-1} + 2\sum_{j=1}^{K} (\mathbf{M}_{j,0}(\varphi_0)^T H^T R^{-1} H \mathbf{M}_{j,0}(\varphi_0)) + Q(\varphi_0)\,,$$

where $Q(\varphi_0)$ involves terms including second derivatives of the system dynamics. These are usually neglected since for large problems, they are inefficient, impracticable and often infeasible to calculate. Hence, we solve

$$\begin{aligned} \nabla\nabla_{\varphi_0} J(\varphi_0)\Delta\varphi_0^{(\ell)} &= -\nabla_{\varphi_0} J(\varphi_0^{(\ell)}),\\ \varphi_0^{(\ell+1)} &= \varphi_0^{(\ell)} + \Delta\varphi_0^{(\ell)}\,, \end{aligned}$$

for $\ell = 0, 1, \dots$, where $\varphi_0^{(\ell)}$ is the ℓth iterate of Newton's method. For the initial condition, the background state is usually chosen, that is, $\varphi_0^{(0)} = \varphi_0^{(b)}$. We perform data assimilation for a single assimilation window of length 100 time steps, followed by a forecast of 2000 time steps. First, we carry out an experiment with perfect observations. For the background estimate, we choose a perturbed initial condition and $B = 0.01I$. Checking the singular values of the observability matrix for this problem, we obtain that the singular values lie between 4 and 30, and the problem is not ill-conditioned. This is in contrast to the problem in Section 9.1, where the forward operator has very small singular values, which, however, led to a smoothing property of the forecast. The problem here lies in the fact that the forecast error grows severely. The inverse problem is not actually ill-conditioned as such, but the forward problem exhibits severe error growth for small perturbations! Figure 1.10 shows the 1st and 20th component of φ before and after the data assimilation process. The error between the true solution and the trajectory before and after the 4DVar data assimilation process is shown in Figure 1.11. We observe that the error in the analysis (thick line) is reduced significantly (compared to the background) in the first 600 time steps (where the assimilation window is of length 100 time steps). After that, we see that the effect of the chaotic dynamics emerges and the error grows since the initial condition of the analysis vector is perturbed from the true initial condition. The initial condition error is of order $\mathcal{O}(10^{-3})$ at each of the sites. From Figure 1.9 (b), we cannot anticipate a better performance of the forecast. We expect the results to be best for perfect and full observations. Next, we carry out an experiment with noisy observations. The observations are generated from the truth with an error of mean zero and covariance $R = 0.01I$. Moreover, we only take observations every five time steps and we only observe 8 of the 40 variables (precisely, we observe every 5th component). For the background state, we use a perturbed initial condition, though this time with background error covariance matrix B with entries $B_{ij} = 0.01\exp(\frac{-|i-j|}{50})$. We observe that the singular values of the observability matrix for this problem lie between

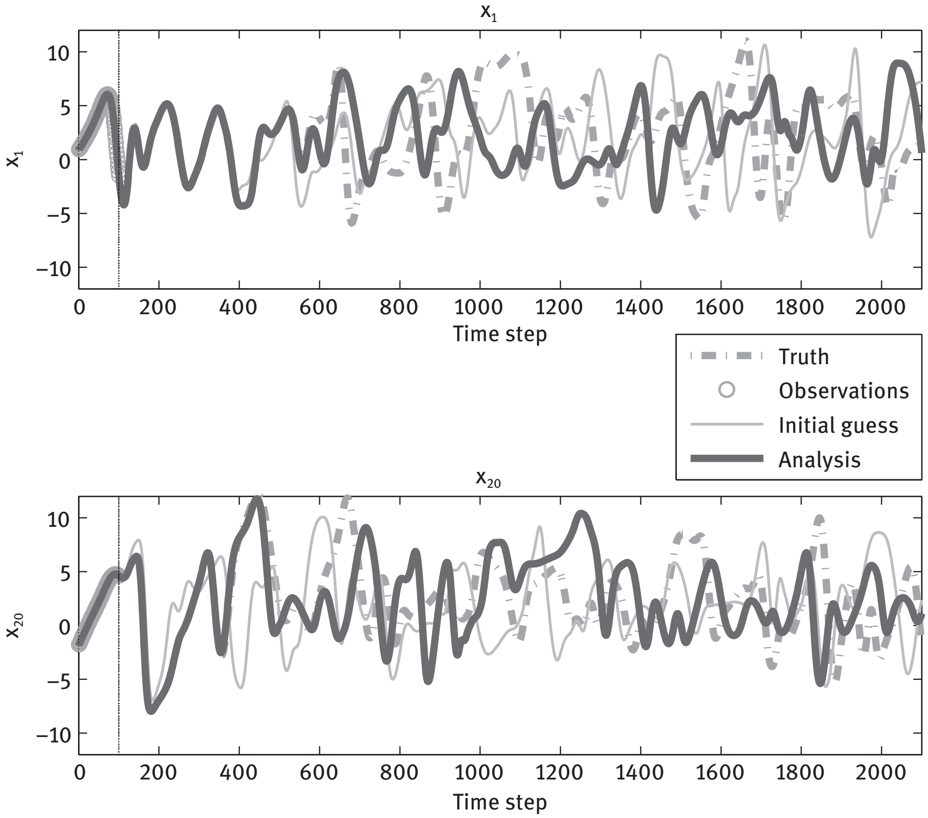


Figure 1.10: Components 1 and 20 of the solution to (1.127) for full and perfect observations. The plot shows the observations, the assimilation window, the exact trajectory, the background trajectory and the final solution (analysis) after 4DVar.

0.02 and 7, and, not surprisingly, the problem is slightly worse conditioned than the one for full observations.

Figure 1.12 shows the error between the true solution and the trajectory before and after the 4DVar data assimilation process. We observe that the error in both components is not reduced as much as the error in Figure 1.11 (for perfect and full observations), which is to be expected as we observe fewer components and moreover, the observations are noisy. Note that with our setup, the 1st component is an “observed site,” where the 20th component is unobserved. We can therefore explain the slightly worse assimilation results of the trajectory of the 20th component compared to the trajectory of the 1st component in Figure 1.12.

To explore this relation further, Figure 1.13 shows the absolute value of the error in the initial condition for this problem, including the sites of the observations. Clearly, at the observation sites, the analysis error is generally smaller than at the unobserved

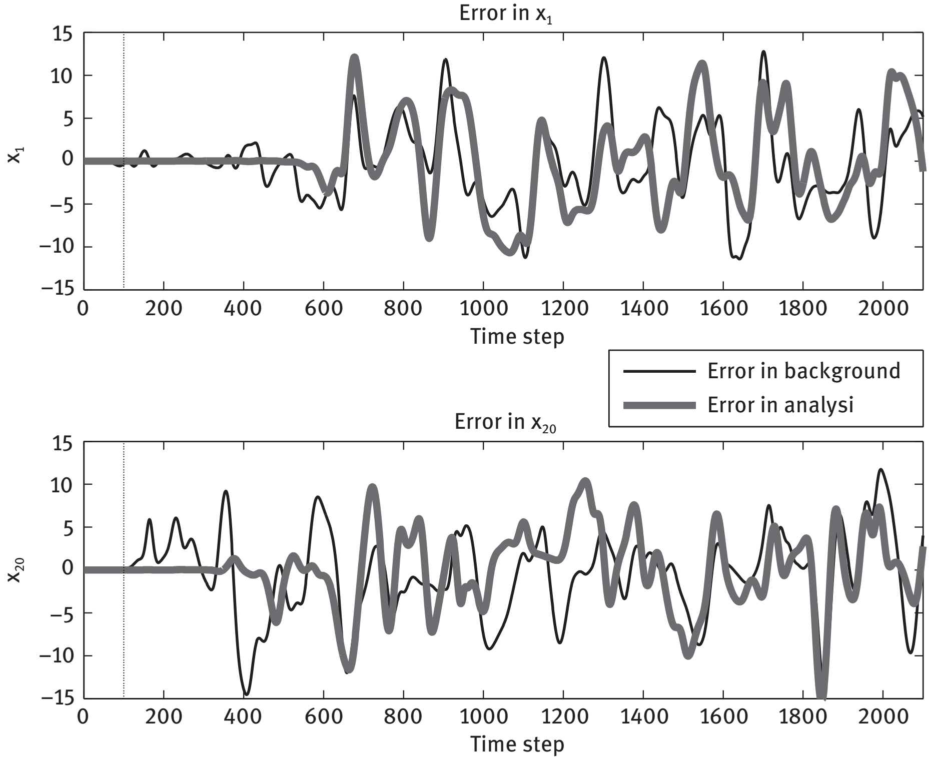


Figure 1.11: Error of components 1 and 20 of the solution to (1.127) for full and perfect observations. The plot shows the error in the background trajectory and the error in the final solution (analysis) after 4DVar.

sites. However, this is not always true as information about the true state from the observations is spread to the unobserved sites through the coupling of the problem and via the background error covariance matrix B.

We carried out tests with other data assimilation algorithms such as 3DVar and the Extended Kalman Filter (EKF). We do not report the results for 3DVar here, but mention that for full perfect observations, 3DVar produces very small errors at the end of the assimilation window as we have perfect observations which are sequentially assimilated into the trajectory. Then, the forecast is run from a very small error at the end of the assimilation window. With fewer and noisy observations, 3DVar gives worse results than 4DVar (as in 4DVar, the missing information is assimilated via the system dynamics). Also, if a model error is included in the system dynamics (that is, the observations are created from the true trajectory, whereas in the data assimilation process, we use a different, perturbed model, replicating the practical situation), we obtain worse results than for the perfect model, as would be expected (Section 9.1).

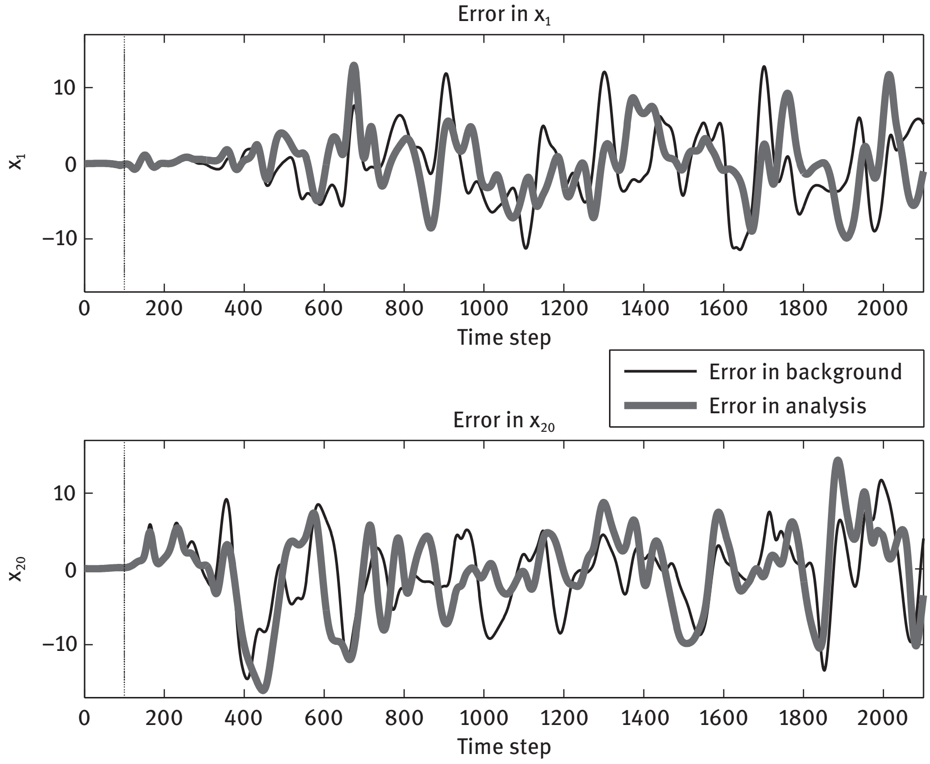


Figure 1.12: Error of components 1 and 20 of the solution to (1.127) for partial and noisy observations. The plot shows the error in the background trajectory and the error in the final solution (analysis) after 4DVar.

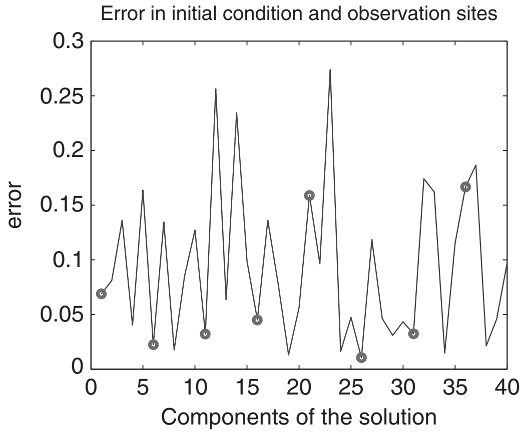


Figure 1.13: Error in the initial condition and observed sites for the solution to (1.127) after 4DVar for partial and noisy observations.

Finally, we apply the EKF to the Lorenz-95 problem. If we use the same background error covariance matrix and the same initial condition as for 4DVar, we obtain essentially the same results as for 4DVar (as would be expected from Theorem 1.18). The results here are only approximately equivalent as Theorem 1.18 only holds for the Kalman filter applied to linear system dynamics. However, when plotting the error, we hardly observe any difference.

A better result as for 4DVar is obtained for the EKF if a better background error covariance matrix is chosen. To this end, we use the covariance matrix produced by the EKF (after one data assimilation cycle at time step 100) as the initial background error covariance matrix for a new EKF experiment applied to the data assimilation problem we consider. This should give a better (flow-dependent) background error covariance matrix. This is indeed true as seen in Figure 1.14 compared to Figure 1.12. The new (flow-dependent) background covariance matrix can also be used for 4DVar, resulting in a hybrid method [9].

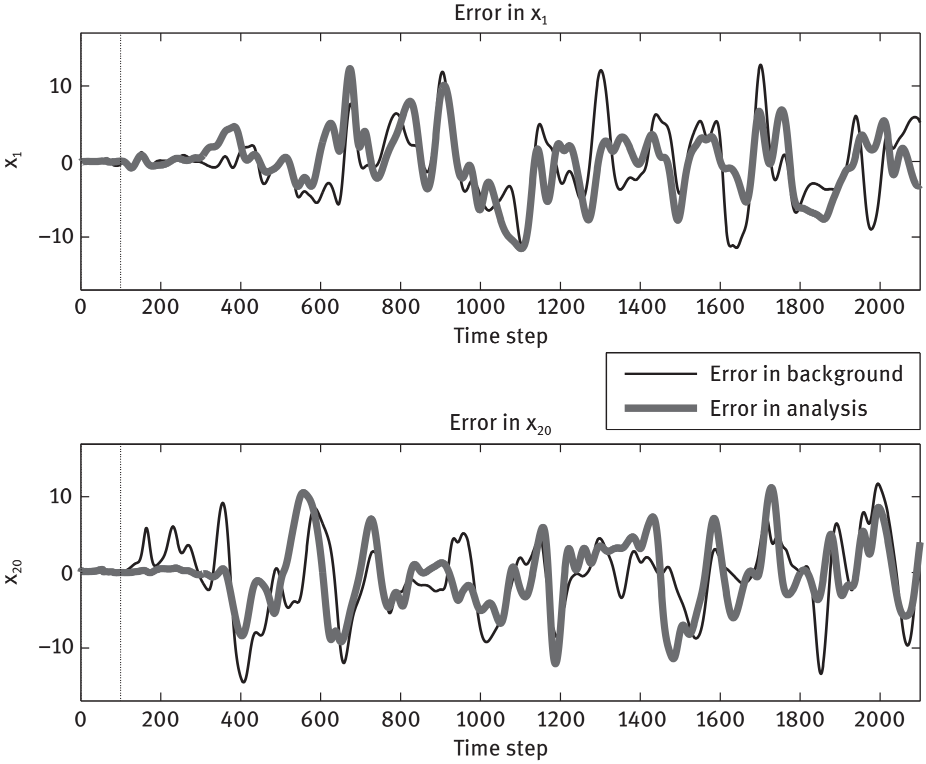


Figure 1.14: Error of components 1 and 20 of the solution to (1.127) for partial and noisy observations. The plot shows the error in the background trajectory and the error in the final solution (analysis) after applying the EKF.

10 Concluding remarks

Inverse problems are an area of research dealing with the reconstruction of functions or parameter distributions from measurements. It has evolved over nearly 100 years in many applications, for example, in electromagnetics and acoustics, in medical imaging and elastography. Today, a growing community of researchers employs both a large set of well-established methods for linear and nonlinear inverse problems as well as a large variety of specific new methods for reconstructions and imaging.

Data assimilation has evolved as a very important and popular research area from specific applications such as weather prediction or hydrology. Using measurement data to control the evolution of dynamical systems shares many of the features which are integral parts of inverse problems. Since World War II data assimilation has focused on the state estimation problem, that is, the reconstruction of the state $\varphi \in X$ of the dynamical system under consideration, where X denotes the particular state space. Often, parameter functions are also involved and lead to an extended state space which includes unknown parameter functions as well. The algorithms which have been introduced here can easily be applied to this most general situation.

Historically, the communities of *inverse problems* and *data assimilation* have evolved independently, with particular notation and approaches which are similar in content, but have been expressed in a different type of notation or terminology. One main goal of this article has been to describe key approaches to data assimilation in an *inverse problems terminology*, such that the *dynamic inverse problems* can be easily identified by the inverse problems community. At the same time, we provide an introduction into a functional analytic view for the data assimilation community which is often second priority by those working on important applications.

Today, the *convergence* of *inverse problems* and *data assimilation* is driven by the evolution of modern remote sensing measurement technologies. For example, there is an increasing set of satellite infrared and microwave sounders, such that their assimilation into atmospheric models involves the use of ill-posed measurement operators. New radar machines not only measure Doppler shift and reflectivity of atmospheric meteors, but also polarization. Ground-based LIDaRs involve further highly ill-posed measurement operators. Further techniques, for example, GPS/GNSS slant delay measurements, lead to ill-posed tomographic problems which become integral parts of operational data assimilation. We believe that the framework which we presented provides an adequate approach to the further development of these systems.

There is also a *need for convergence* on the level of *assimilation algorithms*. Clearly, methods like 3DVar or 4DVar are basically a version of Tikhonov regularization. Additionally, modern ensemble or particle methods increase the need for mathematical analysis with tools from functional analysis and approximation theory since for typical applications, only a very limited number of ensembles or particles can be used and we are in the range of low-dimensional approximation theory rather than in the stochastic limit of an infinite ensemble.

Our article has aimed to contribute to the convergence by presenting a concise introduction into key algorithms and results in a functional analytic language which has the potential to be understood by a large range of mathematicians, thus building a basis for further research and developments. We have included both the viewpoint of deterministic mathematics, numerical analysis and functional analysis as well as stochastics and Bayesian reasoning. Understanding important state-of-the-art algorithms within a uniform framework is a key step today to further develop the tools which are known to have the highest impact on society with respect to such crucial areas as high-impact weather, logistics, travel and energy supply by renewable energy resources.

References

[1] R. Acar and C. R. Vogel, Analysis of bounded variation penalty methods for ill-posed problems, *Inverse Problems* 10 (1999), 1217.

[2] B. D. O. Anderson and J. B. Moore, *Optimal filtering*, Prentice-Hall Englewood Cliffs, NJ, 1979.

[3] J. L. Anderson, An ensemble adjustment Kalman filter for data assimilation, *Monthly weather review* 129 (2001), 2884–2903.

[4] A. Andrews, A square root formulation of the Kalman covariance equations, *AIAA Journal* 6 (1968), 1165–1166.

[5] A. Apte, C. K. R. T. Jones, A. M. Stuart, and J. Voss, Data assimilation: Mathematical and statistical perspectives, *International Journal for Numerical Methods in Fluids* 56 (2008), 1033–1046.

[6] R. N. Bannister, *Elementary 4D-Var*, University of Reading, DARC Technical Report no. 2, 2001.

[7] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM Journal on Imaging Sciences* 2 (2009), 183–202.

[8] G. J. Bierman, Factorization methods for discrete sequential estimation, *Mathematics in science and engineering* 128 (1977).

[9] M. Buehner, Ensemble-derived stationary and flow-dependent background-error covariances: Evaluation in a quasi-operational NWP setting, *Quarterly Journal of the Royal Meteorological Society* 131 (2005), 1013–1043.

[10] M. Burger, H. Dirks, and J. Müller, *Inverse problems in imaging*, Large Scale Inverse Problems. Computational Methods and Applications in the Earth Sciences (M. Cullen, M. A. Freitag, S. Kindermann, and R. Scheichl, eds.), Radon Ser. Comput. Appl. Math. 13, Walter de Gruyter, Berlin, 2013, pp. 135–180.

[11] G. Burgers, P. J. Van Leeuwen, and G. Evensen, Analysis scheme in the ensemble Kalman filter, *Monthly weather review* 126 (1998), 1719–1724.

[12] D. Calvetti and E. Somersalo, *Introduction to Bayesian scientific computing*, Surveys and Tutorials in the Applied Mathematical Sciences 2, Springer, New York, 2007.

[13] E. J. Candes, M. B. Wakin, and S. P. Boyd, Enhancing sparsity by reweighted l_1 minimization, *Journal of Fourier Analysis and Applications* 14 (2008), 877–905.

[14] A. Carrassi and S. Vannitsem, Accounting for model error in variational data assimilation: A deterministic formulation, *Monthly Weather Review* 138 (2010), 3369–3386.

[15] T. F. Chan, Stability analysis of finite difference schemes for the advection-diffusion equation, *SIAM J. Numer. Anal.* 21 (1984), 272–284.

[16] S. E. Cohn, A. da Silva, J. Guo, M. Sienkiewicz, and D. Lamich, Assessing the Effects of Data Selection with the DAO Physical-Space Statistical Analysis System, *Monthly Weather Review* 126 (1998), 2913–2926.

[17] D. L. Colton and R. Kress, *Inverse acoustic and electromagnetic scattering theory*, Springer Verlag, 1998.

[18] P. Courtier, Dual formulation of four-dimensional variational assimilation, *Quarterly Journal of the Royal Meteorological Society* 123 (1997), 2449–2461.

[19] P. Courtier, J.-N. Thépaut, and A. Hollingsworth, A strategy for operational implementation of 4D-Var, using an incremental approach, *Quarterly Journal of the Royal Meteorological Society* 120 (1994), 1367–1387.

[20] R. Daley, *Atmospheric data analysis*, Cambridge: Cambridge University Press, 1991.

[21] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Classics in Applied Mathematics 16, SIAM, Philadelphia, PA, 1996.

[22] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of inverse problems*, Kluwer, Dordrecht, 1996.

[23] H. W. Engl, K. Kunisch, and A. Neubauer, Convergence rates for Tikhonov regularisation of non-linear ill-posed problems, *Inverse Problems* 5 (1989), 523.

[24] G. Evensen, Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *Journal of Geophysical Research* 99 (1994), 10143–10162.

[25] G. Evensen, The ensemble Kalman filter: Theoretical formulation and practical implementation, *Ocean dynamics* 53 (2003), 343–367.

[26] G. Evensen, *Data assimilation: the ensemble Kalman filter*, Springer, Berlin, 2009.

[27] M. Fisher, M. Leutbecher, and G. A. Kelly, On the equivalence between Kalman smoothing and weak-constraint four-dimensional variational data assimilation, *Quarterly Journal of the Royal Meteorological Society* 131 (2006), 3235–3246.

[28] M. A. Freitag, N. K. Nichols, and C. J. Budd, Resolution of sharp fronts in the presence of model error in variational data assimilation, *Quarterly Journal of the Royal Meteorological Society* (2012), doi: 10.1002/qj.2002.

[29] M. Ghil and P. Malanotte-Rizzoli, Data assimilation in meteorology and oceanography, *Adv. Geophys* 33 (1991), 141–266.

[30] G. H. Golub, M. Heath, and G. Wahba, Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics* 21 (1979), 215–223.

[31] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed, John Hopkins University Press, Baltimore, 1996.

[32] S. Gratton, A. S. Lawless, and N. K. Nichols, Approximate Gauss-Newton methods for nonlinear least squares problems, *SIAM Journal on Optimization* 18 (2007), 106–132.

[33] J. Hadamard, *Lectures on Cauchy's problem in linear partial differential equations*, Yale University Press, New Haven, 1923.

[34] M. Hanke, A. Neubauer, and O. Scherzer, A convergence analysis of the Landweber iteration for nonlinear ill-posed problems, *Numerische Mathematik* 72 (1995), 21–37.

[35] P. C. Hansen, The truncated SVD as a method for regularization, *BIT* 27 (1987), 534–553.

[36] P. C. Hansen, Analysis of discrete ill-posed problems by means of the L-curve, *SIAM Rev.* 34 (1992), 561–580.

[37] P. C. Hansen, *Rank-deficient and discrete ill-posed problems*, SIAM Monographs on Mathematical Modeling and Computation, SIAM, Philadelphia, PA, 1998.

[38] P. C. Hansen, J. G. Nagy, and D. P. O'Leary, *Deblurring images*, Fundamentals of Algorithms 3, SIAM, Philadelphia, PA, 2006.

[39] A. C. Harvey, *Forecasting, structural time series models and the Kalman filter*, Cambridge University Press, Cambridge.

[40] B. Hofmann, *Regularization for applied inverse and ill-posed problems*, Teubner Texts in Mathematics 85, Teubner, Leipzig, 1986.

[41] P. L. Houtekamer and H. L. Mitchell, Data assimilation using an ensemble Kalman filter technique, *Monthly Weather Review* 126 (1998), 796–811.

[42] P. L. Houtekamer and H. L. Mitchell, A sequential ensemble Kalman filter for atmospheric data assimilation, *Monthly Weather Review* 129 (2001), 123–137.

[43] B. R. Hunt, E. J. Kostelich, and I. Szunyogh, Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter, *Physica D: Nonlinear Phenomena* 230 (2007), 112–126.

[44] K. Ide, P. Courtier, M. Ghil, and A. C. Lorenc, Unified notation for data assimilation: operational, sequential and variational, *Practice* 75 (1997), 181–189.

[45] C. Johnson, B. J. Hoskins, and N. K. Nichols, A singular vector perspective of 4D-Var: Filtering and interpolation, *Quarterly Journal of the Royal Meteorological Society* 131 (2006), 1–19.

[46] R. E. Kalman, A new approach to linear filtering and prediction problems, *Journal of Basic Engineering* 82 (1960), 35–45.

[47] R. E. Kalman, Mathematical description of linear dynamical systems, *Journal of the Society for Industrial & Applied Mathematics, Series A: Control* 1 (1963), 152–192.

[48] E. Kalnay, *Atmospheric modeling, data assimilation, and predictability*, Cambridge University Press, Cambridge, 2002.

[49] A. Kirsch, *An introduction to the mathematical theory of inverse problems*, Springer Verlag, 2011.

[50] F. X. Le Dimet and O. Talagrand, Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects, *Tellus. Series A, Dynamic meteorology and oceanography* 38 (1986), 97–110.

[51] J. M. Lewis, S. Lakshmivarahan, and S. Dhall, *Dynamic data assimilation*, Encyclopedia of Mathematics and its Applications 104, Cambridge University Press, Cambridge, 2006.

[52] H. Li, E. Kalnay, T. Miyoshi, and C. M. Danforth, Accounting for model errors in ensemble data assimilation, *Monthly Weather Review* 137 (2009), 3407–3419.

[53] Z. Li and I. M. Navon, Optimality of variational data assimilation and its relationship with the Kalman filter and smoother, *Quarterly Journal of the Royal Meteorological Society* 127 (2006), 661–683.

[54] E. N. Lorenz, Deterministic nonperiodic flow, *Journal of the Atmospheric Sciences* 20 (1963), 130–141.

[55] E. N. Lorenz, Predictability: A problem partly solved, in: *Proc. Seminar on Predictability*, 1, 1996.

[56] E. N. Lorenz, Designing chaotic models, *Journal of the Atmospheric Sciences* 62 (2005), 1574–1587.

[57] D. McLaughlin, Recent developments in hydrologic data assimilation, *Rev. Geophys* 33 (1995), 977–984.

[58] D. McLaughlin, An integrated approach to hydrologic data assimilation: interpolation, smoothing, and filtering, *Advances in Water Resources* 25 (2002), 1275–1286.

[59] R. Ménard and R. Daley, The application of Kalman smoother theory to the estimation of 4DVAR error statistics, *Tellus A* 48 (1996), 221–237.

[60] A. J. F. Moodey, A. S. Lawless, R. W. E. Potthast, and P. J. van Leeuwen, *Nonlinear error dynamics for cycled data assimilation*, Report no. MPS-2012-06, 2013.

[61] V. A. Morozov, *On the solution of functional equations by the method of regularization*, Soviet Math. Dokl, 7, 1966, pp. 414–417.

[62] A. Neumaier, Solving ill-conditioned and singular linear systems: A tutorial on regularization, *Siam Rev.* 40 (1998), 636–666.

[63] N. K. Nichols, *Mathematical Concepts of Data Assimilation*, Data Assimilation: Making Sense of Observations (William Lahoz, Boris Khattatov, and Richard Menard, eds.), Springer, 2010, pp. 13–39.

[64] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Classics in Applied Mathematics 30, SIAM, Philadelphia, PA, 2000.

[65] E. Ott, B. R. Hunt, I. Szunyogh, A. V. Zimin, E. J. Kostelich, M. Corazza, E. Kalnay, D. J. Patil, and J. A. Yorke, A local ensemble Kalman filter for atmospheric data assimilation, *Tellus A* 56 (2004), 415–428.

[66] S. K. Park and L. Xu, *Data assimilation for atmospheric, oceanic and hydrologic applications*, Springer, 2009.

[67] R. W. E. Potthast, A. J. F. Moodey, A. S. Lawless, and P. J. van Leeuwen, *On error dynamics and instability in data assimilation*, University of Reading, Preprint, Report no. MPS-2012-06, 2012.

[68] F. Rabier, P. Courtier, J. Pailleux, O. Talagrand, and D. Vasiljevic, A comparison between four-dimensional variational assimilation and simplified sequential assimilation relying on three-dimensional variational analysis, *Quarterly Journal of the Royal Meteorological Society* 119 (1993), 845–880.

[69] S. Reich and C. J. Cotter, *Ensemble filter techniques for intermittent data assimilation*, Large Scale Inverse Problems. Computational Methods and Applications in the Earth Sciences (M. Cullen, M. A. Freitag, S. Kindermann, and R. Scheichl, eds.), Radon Ser. Comput. Appl. Math. 13, Walter de Gruyter, Berlin, 2013, pp. 91–134.

[70] R. H. Reichle, D. B. McLaughlin, and D. Entekhabi, Hydrologic data assimilation with the ensemble Kalman filter, *Monthly Weather Review* 130 (2002), 103–114.

[71] A. R. Robinson and P. F. J. Lermusiaux, Overview of data assimilation, *Harvard Reports in Physical/Interdisciplinary (Ocean Sciences)* 62 (2000), Cambridge, Massachusetts, USA.

[72] Y. Sasaki, Some basic formalisms in numerical variational analysis, *Monthly Weather Review* 98 (1970), 875–883.

[73] E. D. Sontag, *Mathematical control theory: deterministic finite dimensional systems*, Springer, 1998.

[74] G. Strang and K. Aarikka, *Introduction to applied mathematics*, Wellesley-Cambridge Press, Wellesley, MA, 1986.

[75] A. M. Stuart, Inverse problems: a Bayesian perspective, *Acta Numerica* 19 (2010), 451–559.

[76] A. M. Stuart and A. R. Humphries, *Dynamical systems and numerical analysis*, Cambridge University Press, Cambridge, 1998.

[77] I. Szunyogh, E. J. Kostelich, G. Gyarmati, E. Kalnay, B. R. Hunt, E. Ott, E. Satterfield, and J. A. Yorke, A local ensemble transform Kalman filter data assimilation system for the NCEP global model, *Tellus A* 60 (2008), 113–130.

[78] A. N. Tikhonov and V. Y. Arsenin, *Solutions of ill-posed problems*, John Wiley & Sons, New York, 1977.

[79] M. K. Tippett, J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker, Ensemble square root filters, *Monthly Weather Review* 131 (2003), 1485–1490.

[80] Y. Trémolet, Accounting for an imperfect model in 4D-Var, *Quarterly Journal of the Royal Meteorological Society* 132 (2006), 2483–2504.

[81] K. van den Doel, U. Ascher, and E. Haber, *The lost honour of l_2-based regularization*, Large Scale Inverse Problems. Computational Methods and Applications in the Earth Sciences (M. Cullen, M. A. Freitag, S. Kindermann, and R. Scheichl, eds.), Radon Ser. Comput. Appl. Math. 13, Walter de Gruyter, Berlin, 2013, pp. 181–204.

[82] C. R. Vogel, *Computational methods for inverse problems*, SIAM, Philadelphia, PA, 1987.

[83] T. T. Warner, *Numerical weather and climate prediction*, Cambridge University Press, Cambridge, 2010.

[84] J. S. Whitaker and T. M. Hamill, Ensemble data assimilation without perturbed observations, *Monthly Weather Review* 130 (2002), 1913–1924.

[85] C. K. Wikle and L. M. Berliner, A Bayesian tutorial for data assimilation, *Physica D: Nonlinear Phenomena* 230 (2007), 1–16.

[86] X. X. Zhu and R. Bamler, Tomographic SAR Inversion by L1 Norm Regularization–The Compressive Sensing Approach, *IEEE transactions on Geoscience and Remote Sensing* 48 (2010), 3839–3846.

[87] D. Zupanski, A general weak constraint applicable to operational 4DVAR data assimilation systems, *Monthly Weather Review* 125 (1997), 2274–2292.

Amos S. Lawless

Variational data assimilation for very large environmental problems

Abstract: Variational data assimilation is commonly used in environmental forecasting to estimate the current state of the system from a model forecast and observational data. The assimilation problem can be written simply in the form of a nonlinear least squares optimization problem. However, the practical solution of the problem in large systems requires many careful choices to be made in the implementation. In this article, we present the theory of variational data assimilation and then discuss in detail how it is implemented in practice. Current solutions and open questions are discussed.

Keywords: 3DVar, 4DVar, adjoint model, background errors, error covariance, incremental formulation, nested models, observation errors, optimization, reduced order models, tangent linear model, weak-constraint

2010 Mathematics Subject Classification: 49-02, 65-02, 86-02

Amos S. Lawless: School of Mathematical and Physical Sciences, University of Reading, PO Box 220, Whiteknights, Reading, RG6 6AX, United Kingdom, a.s.lawless@reading.ac.uk

1 Introduction

Data assimilation is the process of combining a numerical model forecast with observational data in order to estimate the current state of a dynamical system. It has been an essential part of numerical weather prediction (NWP) since its beginnings in the 1940s, when it was recognized that errors in the initial model state could rapidly lead to large errors in the forecast. Early data assimilation schemes were based on a simple interpolation between the observations and the model state, with later schemes also taking account of the statistics of the errors in the data. Such schemes included smoothing splines, successive correction, optimal interpolation and analysis correction [82, 85]. The possible use of methods based on variational calculus was proposed by Sasaki [103, 104] in the late 1950s and 1960s, but at the time a practical implemen-

I would like to thank two anonymous reviewers whose detailed reading of the manuscript helped lead to many improvements in the final version.

The author is supported in part by the U.K. Natural Environment Research Council, through the National Centre for Earth Observation.

tation was not possible. A real breakthrough in the application of variational schemes to NWP came in the late 1980s with a series of papers demonstrating how the problem could be solved using techniques from the theory of optimal control, in particular the use of adjoint equations to calculate the gradient of an objective function, or cost function [77, 107]. This led to a series of papers in which the feasibility of variational data assimilation was studied on a series of different simplified atmospheric models [26, 93, 98, 108] (these experiments usually only included the large scale atmospheric dynamics and not the subgrid-scale processes of full weather prediction models).

Despite the encouraging results of these experiments, variational data assimilation remained impractical for operational use due to the high computational cost. The introduction of the incremental method of variational assimilation in 1994 [27], together with increasing computing power, opened up the possibility of an affordable implementation for operational weather prediction. Over the following decade, many weather forecasting centers began to develop variational data assimilation for operational use [42, 43, 61, 84, 99, 100]. At the same time, variational data assimilation began to be applied to other applications, such as ocean forecasting [112, 116] and atmospheric chemistry [38].

A common feature of many of these applications is that the size of the state variable being estimated is extremely large. Current numerical weather prediction models may require the initialization of the order of 10^8 variables in order to make a forecast. As computing power increases, the spatial resolution of the models tends to increase and hence so does the number of variables being represented. Furthermore, the real-time nature of environmental forecasting requires that the data assimilation problem be solved quickly. These two factors imply that when implementing variational data assimilation schemes in practice, compromises must be made. Hence, it is important to design the algorithms carefully to ensure that as accurate a solution as possible is obtained within the time available. Ideally, such design should also include knowledge of the physics of the problem, so that the final solution is physically realistic. In the remainder of this article we will discuss some of the different choices that arise in the implementation of variational data assimilation for very large systems and the practical approaches that have been developed. First, we briefly present the mathematical theory of variational data assimilation.

2 Theory of variational data assimilation

We consider a discrete nonlinear dynamical system given by the equation

$$\mathbf{x}_{i+1} = \mathcal{M}_i(\mathbf{x}_i)\,, \tag{2.1}$$

where $\mathbf{x}_i \in \mathbb{R}^n$ is the state vector at time t_i and $\mathcal{M}_i$ is the nonlinear model operator that propagates the state at time t_i to time t_{i+1} for $i = 0, 1, \ldots, N-1$. We assume that we have imperfect observations $\mathbf{y}_i \in \mathbb{R}^{p_i}$ at times $t_i, i = 0, \dots, N$ that are related to the system state through the equation

$$\mathbf{y}_i = \mathcal{H}_i(\mathbf{x}_i) + \boldsymbol{\epsilon}_i\,, \tag{2.2}$$

where $\mathcal{H}_i : \mathbb{R}^n \to \mathbb{R}^{p_i}$ is known as the observation operator and maps the state vector to observation space. The observation errors $\boldsymbol{\epsilon}_i$ are usually assumed to be unbiased, serially uncorrelated, Gaussian errors with known covariance matrices $\mathbf{R}_i$. For the numerical weather prediction problem, the vector $\mathbf{x}_i$ would contain several meteorological variables, for example, pressure, temperature and the three-dimensional wind at each grid point of the model domain. The observation operator $\mathcal{H}_i$ may just be a simple interpolation in space if the state variable is observed directly. However, it could be a much more complicated nonlinear function of the state. For example, for a satellite radiance measurement, the observation operator can include a complex radiative transfer model.

We assume that at the initial time t_0 we have an *a priori* estimate of the state, usually referred to as a *background* field, that we denote $\mathbf{x}^b$. This background field is assumed to have unbiased, Gaussian errors with known covariance matrix $\mathbf{B}$. In practice, the background field is usually a short-term forecast of the state from a previous assimilation cycle. The problem of four-dimensional variational data assimilation (4DVar)[1] is then to find the initial state that minimizes the weighted least squares distance to this background while minimizing the weighted least squares distance of the model trajectory to the observations over the time interval $[t_0, t_N]$. Mathematically, we can formulate this as an optimization problem: Find the state $\mathbf{x}_0^a$ at time t_0 that minimizes the function

$$\mathcal{J}(\mathbf{x}_0) = \frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}^b\right)^{\mathrm{T}} \mathbf{B}^{-1}\left(\mathbf{x}_0 - \mathbf{x}^b\right) + \frac{1}{2}\sum_{i=0}^{N}\left(\mathcal{H}_i(\mathbf{x}_i) - \mathbf{y}_i\right)^{\mathrm{T}}\mathbf{R}_i^{-1}\left(\mathcal{H}_i(\mathbf{x}_i) - \mathbf{y}_i\right) \tag{2.3}$$

subject to the states $\mathbf{x}_i$ satisfying the nonlinear dynamical system (2.1). In the case where $N = 0$ there is no model evolution and the scheme is referred to as three-dimensional variational data assimilation (3DVar). The solution $\mathbf{x}_0^a$ is commonly referred to as the *analysis*. In environmental data assimilation, the function $\mathcal{J}(\mathbf{x}_0)$ is usually called the *cost function*, but the terms *objective function* and *penalty function* are often used in other fields.

The minimization problem given by equation (2.3) can be interpreted in a statistical or deterministic sense. From Bayes' theorem, it can be shown that $\mathbf{x}_0^a$ gives the maximum *a posteriori* estimate of the state under the assumptions given [82]. This includes the assumption of Gaussianity of the error statistics for the background field

[1] The scheme is referred to as four-dimensional since we usually fit three spatial dimensions in time, with time being the fourth dimension.

and observations. In practice, this assumption may not always hold. For example, for variables that are inherently nonnegative, such as humidity in the atmosphere or concentrations in chemical models, Gaussian statistics may not be appropriate. In some cases these errors may be treated by assuming a lognormal distribution and using this to transform to variables whose statistics are Gaussian [13, 41]. Some allowance for non-Gaussian observation errors may also be made using the method of variational quality control, as discussed in Section 3.3. Furthermore, nonlinearity in the dynamical model implies that the background errors are likely to be non-Gaussian if the background comes from a forecast whose length is beyond the linearity regime of the model. For this reason, in numerical weather prediction the background field is usually from a forecast of only 6 or 12 hours. In some applications, such as the identification of the source of an atmospheric tracer, it may be more appropriate to specify other prior error distributions [12]. The alternative, deterministic interpretation of the minimization problem is to consider the term measuring the fit to the background state as a form of Tikhonov regularization in fitting the observations [29, 65, 90]. Each of these interpretations is able to provide different insights into the practical formulation of the problem.

It is instructive to consider the solution to the 3DVar problem under the hypothesis that the observation operator $\mathcal{H}_0$ is approximately linear, that is,

$$\mathcal{H}_0\left(\mathbf{x}^b\right) - \mathcal{H}_0(\mathbf{x}_0) \approx \mathbf{H}_0\left(\mathbf{x}^b\right)\left(\mathbf{x}^b - \mathbf{x}_0\right) \tag{2.4}$$

where $\mathbf{H}_0(\mathbf{x}^b)$ is the Jacobian of $\mathcal{H}_0$ evaluated at $\mathbf{x}^b$ (This assumption (2.4) is referred to as the tangent linear hypothesis). In this case, the minimum value of (2.3) can be written explicitly as

$$\mathbf{x}^a \ = \mathbf{x}^b + \mathbf{B}\mathbf{H}_0^T\left(\mathbf{H}_0\mathbf{B}\mathbf{H}_0^T + \mathbf{R}_0\right)^{-1}\left(\mathbf{y}_0 - \mathcal{H}_0\left(\mathbf{x}^b\right)\right) . \tag{2.5}$$

This solution is equal to the best linear unbiased estimate (or BLUE). We see then that the analysis increment, defined as the difference between the analysis and the background $\mathbf{x}^a - \mathbf{x}^b$, lies in the range space of the background error covariance matrix **B**. We return to the implications of this in Section 3.2.

The covariance of the analysis error in this case is given by

$$\mathbf{A} = \left(\mathbf{B}^{-1} + \mathbf{H}_0^T\mathbf{R}_0^{-1}\mathbf{H}_0\right)^{-1} . \tag{2.6}$$

We find that for both 3DVar and 4DVar, this is equal to the inverse of the Hessian of the cost function, that is,

$$\mathbf{A} = \left(\nabla^2 \mathcal{J}\right)^{-1} . \tag{2.7}$$

In general, an exact solution cannot be found and the cost function is minimized using iterative numerical methods, such as conjugate gradient or quasi-Newton methods. The use of these methods in data assimilation is discussed in more detail in Section 3.4. On each iteration of such methods, the value of the cost function and its gradient at the current iterate must be calculated. In order to calculate the gradient of (2.3) with respect to the initial state $\mathbf{x}_0$, we consider the discrete Euler–Lagrange equations. We introduce Lagrange multipliers $\boldsymbol{\lambda}_i$ at time t_i and define the Lagrangian by

$$\mathcal{L}(\mathbf{x}_i, \boldsymbol{\lambda}_i) = \mathcal{J}(\mathbf{x}_0) + \sum_{i=0}^{N-1} \boldsymbol{\lambda}_{i+1}^T (\mathbf{x}_{i+1} - \mathcal{M}_i(\mathbf{x}_i)) \,. \tag{2.8}$$

Then, necessary conditions for a minimum of (2.3) subject to the constraint are found by taking variations of $\mathcal{L}$ with respect to $\boldsymbol{\lambda}_i$ and $\mathbf{x}_i$. The first of these leads to the original nonlinear model equations (2.1), while the latter gives the discrete adjoint equations

$$\boldsymbol{\lambda}_i = \mathbf{M}_i^T \boldsymbol{\lambda}_{i+1} - \mathbf{H}_i^T \mathbf{R}_i^{-1} (\mathcal{H}_i(\mathbf{x}_i) - \mathbf{y}_i) \tag{2.9}$$

for $i = 1, \dots, N$ with boundary condition $\boldsymbol{\lambda}_{N+1} = 0$, where $\mathbf{H}_i$ and $\mathbf{M}_i$ are the Jacobians of the nonlinear operators $\mathcal{H}_i$ and $\mathcal{M}_i$ with respect to the state variable $\mathbf{x}_i$. In the data assimilation literature, these Jacobians are referred to as the *tangent linear operator* and the *tangent linear model* (TLM) and the operators $\mathbf{H}_i^T$ and $\mathbf{M}_i^T$ are the adjoints of the observation operator and the nonlinear model operator. From (2.8) we then have that the gradient of the Lagrangian with respect to the initial state $\mathbf{x}_0$ is given by

$$\frac{\partial \mathcal{L}}{\partial \mathbf{x}_0} = -\mathbf{M}_0^T \boldsymbol{\lambda}_1 + \mathbf{H}_0^T \mathbf{R}_0^{-1} \left(\mathcal{H}_0(\mathbf{x}_0) - \mathbf{y}_0 \right) + \mathbf{B}^{-1} \left(\mathbf{x}_0 - \mathbf{x}^b \right) \,. \tag{2.10}$$

From the theory of Lagrange multipliers, this is equal to the gradient of the function under the constraint, and thus we can write

$$\nabla \mathcal{J}(\mathbf{x}_0) = -\boldsymbol{\lambda}_0 + \mathbf{B}^{-1} \left(\mathbf{x}_0 - \mathbf{x}^b \right) \,, \tag{2.11}$$

where we have introduced the extra variable

$$\boldsymbol{\lambda}_0 = \mathbf{M}_0^T \boldsymbol{\lambda}_1 - \mathbf{H}_0^T \mathbf{R}_0^{-1} (\mathcal{H}_0(\mathbf{x}_0) - \mathbf{y}_0) \,, \tag{2.12}$$

which can be calculated from the adjoint equations (2.9) with $i = 0$. Hence, the adjoint equations provide an efficient method for calculating the gradient information needed for the minimization algorithm. Each iteration of a numerical optimization method therefore requires one run of the forward model (2.1) to calculate the value of the cost function and one run of the adjoint model (2.9) to calculate the gradient. This makes 4DVar very expensive from a computational point of view.

We note that in this derivation, we have implicitly taken the adjoint with respect to the Euclidean inner product. For a general linear operator $\mathbf{L} : X1 \to X2$ and inner products $\langle .,. \rangle_{X1}, \langle .,. \rangle_{X2}$ in the spaces $X1, X2$ respectively, the adjoint of $\mathbf{L}$ is the operator $\mathbf{L}^* : X2 \to X1$ such that

$$\langle \mathbf{L}\mathbf{x}_1, \mathbf{x}_2 \rangle_{X2} = \langle \mathbf{x}_1, \mathbf{L}^* \mathbf{x}_2 \rangle_{X1} \tag{2.13}$$

for all $\mathbf{x}_1 \in X1, \mathbf{x}2 \in X2$. In the case where the Euclidean inner product is used in both spaces, the adjoint is equal to the transpose operator, which is why we define the transpose matrices $\mathbf{H}_i^T$ and $\mathbf{M}_i^T$ as the adjoint operators. In this case, the Lagrange multipliers provide the correct gradient of the cost function with respect to the state vector, but it is difficult to interpret physically what these variables mean. For other applications of adjoint modeling, for example, generating initial perturbations for ensembles of forecasts, it may be desirable to give a physical interpretation to the gradients calculated from the Lagrange multipliers. In these applications, other inner products may be used, for example, based on the energy or enstrophy[2] of the system [95].

2.1 Incremental variational data assimilation

The possibility of implementing variational data assimilation in an operational setting came with the proposal of incremental variational data assimilation [27]. In this formulation, the solution to the nonlinear minimization problem (2.3) is approximated by a sequence of minimizations of linear quadratic cost functions. We define $\mathbf{x}_0^{(k)}$ to be the k^{th} estimate to the solution and linearize the cost function (2.3) around the model trajectory forecast from this estimate. The next estimate is then defined by

$$\mathbf{x}_0^{(k+1)} = \mathbf{x}_0^{(k)} + \delta\mathbf{x}_0^{(k)} \,, \tag{2.14}$$

where the perturbation $\delta\mathbf{x}_0^{(k)} \in \mathbb{R}^n$ is a solution of the linearized cost function

$$\begin{aligned}\tilde{\mathcal{J}}^{(k)}\left(\delta\mathbf{x}_0^{(k)}\right) &= \frac{1}{2}\left(\delta\mathbf{x}_0^{(k)} - \left[\mathbf{x}^b - \mathbf{x}_0{}^{(k)}\right]\right)^{\mathrm{T}} \mathbf{B}_0^{-1}\left(\delta\mathbf{x}_0^{(k)} - \left[\mathbf{x}^b - \mathbf{x}_0{}^{(k)}\right]\right) \\ &+ \frac{1}{2}\sum_{i=0}^{N}\left(\mathbf{H}_i\delta\mathbf{x}_i^{(k)} - \mathbf{d}_i^{(k)}\right)^{\mathrm{T}} \mathbf{R}_i^{-1}\left(\mathbf{H}_i\delta\mathbf{x}_i^{(k)} - \mathbf{d}_i^{(k)}\right) \,.\end{aligned} \tag{2.15}$$

Here, $\mathbf{d}_i^{(k)} = \mathbf{y}_i - \mathcal{H}_i(\mathbf{x}_i^{(k)})$, where $\mathbf{x}_i^{(k)}$ is the nonlinear trajectory calculated from the current estimate at the initial time using the nonlinear model equation (2.1). The perturbation $\delta\mathbf{x}_i$ satisfies the linear dynamical equation

$$\delta\mathbf{x}_{i+1} = \mathbf{M}_i\delta\mathbf{x}_i \,. \tag{2.16}$$

The linearized observation operator $\mathbf{H}_i$ and the tangent linear model operator $\mathbf{M}_i$ are evaluated at the current estimate of the nonlinear trajectory, usually called the linearization state. The minimization (2.15) is referred to as the *inner loop*, while the update of the nonlinear model trajectory $\mathbf{x}_i^{(k)}$ is the *outer loop*. On each iteration of the

[2] In fluid dynamics, the enstrophy is defined as the mean square vorticity of the fluid [58, Section 13.4].

inner loop, the TLM is integrated to calculate the evolution of the perturbation in order to calculate the cost function (2.15), and the adjoint model is integrated to provide the gradient.

The incremental method was later shown to be equivalent to an inexact Gauss–Newton method applied to the original nonlinear cost function (2.3) [72]. If we consider a general nonlinear least squares cost function

$$\phi(\mathbf{x}) = \frac{1}{2}\mathbf{f}(\mathbf{x})^T\mathbf{f}(\mathbf{x}) \tag{2.17}$$

with $\mathbf{f}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^p$ and let $\mathbf{J}(\mathbf{x})$ be the Jacobian of $\mathbf{f}(\mathbf{x})$ with respect to $\mathbf{x}$, then the Gauss–Newton method for minimizing ϕ is

Algorithm 2.1 (Gauss-Newton).

```
step 0:   choose $\mathbf{x}^{(0)}$
step 1:   repeat until convergence
          step 1.1:   compute   $\delta\mathbf{x} = -((\mathbf{J}((\mathbf{x}^{(k)})^T\mathbf{J}((\mathbf{x}^{(k)}))^{-1}\mathbf{J}((\mathbf{x}^{(k)})^T\mathbf{f}((\mathbf{x}^{(k)})$
          step 1.2:   update   $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \delta\mathbf{x}.$
```

Sufficient conditions can be found such that the algorithm will converge to a local minimum of (2.17) [34]. Step 1.1 of the algorithm is equivalent to solving the minimization problem

$$\min_{\delta\mathbf{x}} \|\mathbf{J}(\mathbf{x})\delta\mathbf{x} + \mathbf{f}(\mathbf{x})\|_2^2 \,. \tag{2.18}$$

If we define

$$\mathbf{f}(\mathbf{x}_0) = -\begin{pmatrix} \mathbf{B}^{-1}\left(\mathbf{x}_0 - \mathbf{x}^b\right) \\ \mathbf{R}_0^{-1}\,(H_0[\mathbf{x}_0] - \mathbf{y}_0^o) \\ \vdots \\ \mathbf{R}_N^{-1}\,(H_N[\mathbf{x}_N] - \mathbf{y}_n^o) \end{pmatrix} \tag{2.19}$$

subject to (2.1), then the general cost function (2.17) is equal to the 4DVar cost function (2.3). Applying the Gauss–Newton method to solve this problem, we find that the inner minimization step (2.18) is equivalent to the linearized cost function (2.15).

An advantage of using this method to solve the nonlinear problem is that each inner minimization problem is quadratic in $\delta\mathbf{x}$. Hence, whereas the nonlinear problem may have multiple minima, the inner problem has a unique solution that can be found efficiently using iterative minimization methods (we discuss these methods further in Section 3.4). Since these minimization methods are usually truncated according to some stopping criterion, the inner step of the Gauss–Newton method is not solved exactly. In this case, the outer loop iterations can be shown to be locally convergent under certain conditions, provided that the inner loop minimization is solved to sufficient accuracy [45, 71].

In practice, very few outer loop steps are performed. For example, the Met Office in the U.K. performs only one, while the European Centre for Medium-range Weather Forecasts (ECMWF) performs three [39, 100]. As for the fully nonlinear problem, the incremental method can be run as 3DVar (no model evolution) or 4DVar (including the model evolution). An alternative formulation that is often implemented is known as 3D-FGAT (First Guess at Appropriate Time). This includes the nonlinear model evolution in the calculation of the vectors $\mathbf{d}_i$, but no evolution is included for the perturbation and the TLM operator $\mathbf{M}_i$ in equation (2.16) is replaced by the identity. This ensures that the observations are compared with the nonlinear trajectory at the correct time, but approximates the perturbation in such a way that no TLM or adjoint model is needed. In this way, some of the benefits of 4DVar can be achieved without too much extra computational cost [70, 87].

A major advantage of the incremental approach is that the inner loop minimization problem may be solved in a smaller dimensional space than the outer loop step, for example, at a lower spatial resolution. In this way, the TLM and adjoint model need only be run at the lower resolution on each inner loop iteration, while the linearization trajectory from the nonlinear model is still calculated at the higher resolution on each outer loop. This is discussed further in Section 3.5. The computational savings made by implementing the inner loop in this way made incremental 4DVar feasible for operational weather and ocean forecasting.

Having presented the basic theory of variational data assimilation, we now examine some of the issues that arise in its practical implementation. For the very large systems found in environmental modeling, it is not always possible to apply the theory in an intuitive way. Many choices must be made in order to setup and solve the assimilation problem efficiently and compromises must often be made. It is the attention to detail in these choices that can determine the success or otherwise of the data assimilation scheme.

3 Practical implementation

3.1 Model development

The development of a 4DVar scheme for the large models used in operational weather and ocean forecasting is a huge undertaking. In most cases, the nonlinear model code already exists and has been developed over many years. These models are very large pieces of software, with maybe close to one million lines of code. In order to develop an incremental 4DVar scheme, the code for the TLM and adjoint model must first be written. The development of a TLM code and adjoint model code from the source code of a nonlinear model is a fairly automatic procedure. The correct code for the TLM can be found from a linearization of each statement of the nonlinear model source code based on treating the nonlinear model as a series of arithmetic operations and applying the chain rule. The adjoint model is then found by a line-by-line transpose of the TLM source code in reverse order. This method is known as automatic differentiation. We do not go into details of its application here, but refer the reader to several good introductions in the literature [10, 26, 44, 102]. The automatic nature of this procedure has led to many software tools being developed that will produce a TLM and adjoint model code from a nonlinear mode source code. These automatic differentiation tools, or automatic adjoint compilers, are now available commercially for many different programming languages. [3]

In practice, the TLM and adjoint models of many large environmental models have been developed by hand, rather than using the automatic compilers. There are several reasons for this. The first is that in many cases of operational weather and ocean forecasting, the complexity of the already existing nonlinear model codes was such that simple application of the automatic compilers was not possible. In many cases, particularly for large codes developed by many people, it is necessary to tidy the nonlinear model codes to make them suitable for use with the automatic compilers. Many centers felt that the effort to do this would have been greater than coding the TLM and adjoint model by hand.

The second reason for developing the TLM and adjoint codes by hand arises from the nature of the incremental approach to variational data assimilation. Since the TLM and adjoint are run at a lower resolution in the inner loop, the TLM is already an approximate linearization of the nonlinear model used in the outer loop. It is therefore justifiable to make further simplifications in the TLM in order to reduce the computational cost. As long as the adjoint model is derived from the approximate TLM, then the inner loop minimization will contain the correct gradient information for convergence. In coding the models by hand, it is easier to make such simplifications based on physical arguments. For example, many meteorological models contain parametrizations of subgrid-scale processes (known as the *physics* in the meteorological literature), including such things as clouds, precipitation and surface drag. The schemes used to represent these processes can be highly complex and often include nondifferentiable functions, such as on-off switches. While it is possible for automatic differentiation to deal with such functions, it is usually felt that this level of complexity is not necessary in the TLM and adjoint model. Hence, a series of simpler parametrizations have been developed solely for use in incremental 4DVar that capture the main behavior of the more complex schemes [64, 88, 99, 118].

An alternative approach, devised by the Met Office, is to start from the premise that the linear model must evolve finite and not infinitesimal perturbations so that there is no need for the linear model to be tangent to any nonlinear model. In this approach, the linear model is designed with this in mind. In particular, the resolved dynamics is approximated by a discretization of the linearized continuous equations,

[3] The term *automatic differentiation* refers to the approach itself, not just to the automatic tools.

with various simplifications in the equations and the discretization. Then simplified parametrizations can be used to represent subgrid-scale processes [74, 86]. The adjoint model is derived from this approximate linear model by the process of automatic differentiation, ensuring that it provides the exact gradient of the discrete linear cost function.

An essential part of the development of the linear and adjoint models is their testing, as any small mistakes could lead to lack of convergence of the minimization algorithms. Robust tests exist to check the coding of a TLM and adjoint model. The test for the TLM is based on comparing the evolution of a perturbation in the TLM with the evolution of the same perturbation in the nonlinear model. A Taylor series expansion of the nonlinear model operator shows that the evolutions should be closer together as the perturbation size is reduced [79, 98]. When an inexact TLM is used, the test is not able to differentiate between small coding errors and the desired inexactness. In this case, other more subjective tests must be performed [74]. The adjoint model code can be tested by a verification of the adjoint identity (2.13). If we assume that the spaces $X1$ and $X2$ are both equal to $\mathbb{R}^n$, then we must have

$$\langle \mathbf{M}_i \delta \mathbf{x}_i, \mathbf{M}_i \delta \mathbf{x}_i \rangle = \langle \delta \mathbf{x}_i, \mathbf{M}_i^* \left(\mathbf{M}_i \delta \mathbf{x}_i \right) \rangle , \tag{2.20}$$

which, in the Euclidean inner product, is equivalent to

$$\left(\mathbf{M}_i \delta \mathbf{x}_i \right)^T \left(\mathbf{M}_i \delta \mathbf{x}_i \right) = \delta \mathbf{x}_i^T \left(\mathbf{M}_i^T \left(\mathbf{M}_i \delta \mathbf{x}_i \right) \right) . \tag{2.21}$$

This identity can be tested for random perturbations $\delta \mathbf{x}_i$. If the adjoint operator $\mathbf{M}_i^T$ has been correctly coded, then this identity will hold to machine precision [93]. For large codes, each of these tests should be available for each subroutine, as well as at higher levels. A further test, also based on a Taylor expansion, is used to verify that the gradient of the cost function has been correctly coded [93].

3.2 Background error covariances

The background field $\mathbf{x}^b$ is a very important part of practical data assimilation systems in environmental forecasting. Since in many operational forecasting systems the background field is a forecast from a previous assimilation cycle, it contains information from observations assimilated at earlier times. In one of the early 4DVar systems at ECMWF, it was shown that at any assimilation time, the background field has an approximately 85% influence on the analysis, with the new observations contributing only 15% [24]. The background error covariance matrix $\mathbf{B}$ determines the relative weight between the background field and observations, and hence plays an essential role in the data assimilation algorithm. However, the calculation of these covariances for the assimilation system is a hugely complex task and very dependent on the specific system being modeled. Here, we are only able to give an outline of the main steps involved. For further details in the context of atmospheric data assimilation, the reader is referred to the comprehensive two-part review article of Bannister [6, 7].

As was seen from (2.5) in Section 2, under certain simplified assumptions the analysis increment of 3DVar can be shown to lie in the subspace spanned by the columns of the matrix **B**. In order to understand the implications of this, we consider the case where we have a single observation y of the k^{th} component of the vector **x**, with error variance σ_o^2. In this case, the observation operator is linear and is given by the k^{th} unit vector $\mathbf{e}_k$ and the analysis equation (2.5) becomes

$$\mathbf{x}^a = \mathbf{x}^b + \begin{pmatrix} b_{1,k} \\ b_{2,k} \\ \vdots \\ b_{N,k} \end{pmatrix} \frac{y - \mathbf{x}^b(k)}{b_{k,k} + \sigma_o^2}\,, \tag{2.22}$$

where $b_{i,k}, i = 1, \ldots, N$ indicates the (i, k) element of the matrix **B** and $\mathbf{x}^b(k)$ is the k^{th} component of $\mathbf{x}^b$. Hence, we see that the value of each entry $b_{i,k}$, which is the covariance between the errors in the components of the background field $\mathbf{x}^b(i)$ and $\mathbf{x}^b(k)$, determines the analysis increment to the i^{th} component of the state given an observation of the k^{th} component. As a consequence, the entries of this matrix determine how observations are used to infer information about unobserved parts of the state. Thus, this matrix is fundamental in allowing information to be inferred about unobserved physical variables or unobserved regions of space. However, it is usually impossible to represent this matrix in matrix form. If the state vector is of size n, then the matrix **B** is of size $n \times n$ and when n is of order 10^8, this matrix is impossible to calculate or store. Instead, the action of this matrix is usually represented by a variable transform.

We consider the variable transform in the context of incremental variational data assimilation since that is how it is usually implemented. We define a new variable $\delta\mathbf{z}_i \in \mathbb{R}^n$ and a transformation matrix $\mathbf{U}_i \in \mathbb{R}^{n\times n}$ such that

$$\delta\mathbf{x}_i = \mathbf{U}_i \delta\mathbf{z}_i\,, \quad i = 0, \ldots, N\,. \tag{2.23}$$

In terms of this new variable, the incremental cost function (2.15) can be written as

$$\begin{aligned} \tilde{J}^{(k)}\left(\delta\mathbf{z}_0^{(k)}\right) &= \frac{1}{2}\left(\delta\mathbf{z}_0^{(k)} - \left[\mathbf{z}^b - \mathbf{z}_0{}^{(k)}\right]\right)^{\mathrm{T}} \mathbf{U}_0^T\mathbf{B}^{-1}\mathbf{U}_0 \left(\delta\mathbf{z}_0^{(k)} - \left[\mathbf{z}^b - \mathbf{z}_0{}^{(k)}\right]\right) \\ &+ \frac{1}{2}\sum_{i=0}^{N}\left(\mathbf{H}_i\mathbf{U}_i\delta\mathbf{z}_i^{(k)} - \mathbf{d}_i^{(k)}\right)^{\mathrm{T}} \mathbf{R}_i^{-1}\left(\mathbf{H}_i\mathbf{U}_i\delta\mathbf{z}_i^{(k)} - \mathbf{d}_i^{(k)}\right)\,. \end{aligned} \tag{2.24}$$

If the variables $\delta\mathbf{z}$ are chosen in such a way that they are uncorrelated, then they have the identity covariance matrix by definition and so $\mathbf{U}_0^T\mathbf{B}^{-1}\mathbf{U}_0$ can be replaced with the identity in the cost function (2.24). In this case, the cost function no longer contains the original background error covariance matrix; instead, it is implicitly defined through the variable transform with $\mathbf{B} = \mathbf{U}_0\mathbf{U}_0^T$.

Furthermore, this variable transform is expected to lead to a better conditioned problem. To understand this, we note that the Hessian of the original inner loop cost function (2.15) is given by

$$\mathbf{G} = \mathbf{B}^{-1} + \sum_{i=0}^{N} \mathbf{M}(t_i, t_0)^T \mathbf{H}_i^T \mathbf{R}_i^{-1} \mathbf{H}_i \mathbf{M}(t_i, t_0)\,, \tag{2.25}$$

where

$$\mathbf{M}(t_i, t_0) = \mathbf{M}_{i-1}\mathbf{M}_{i-2}\ldots\mathbf{M}_0 \tag{2.26}$$

is the tangent linear model solution operator from time t_0 to time t_i. Equivalently, we can write this as

$$\mathbf{G} = \mathbf{B}^{-1} + \hat{\mathbf{H}}^T\hat{\mathbf{R}}^{-1}\hat{\mathbf{H}}\,, \tag{2.27}$$

where

$$\hat{\mathbf{H}} = \begin{pmatrix} \mathbf{H}_0 \\ \mathbf{H}_1\mathbf{M}(t_1, t_0) \\ \vdots \\ \mathbf{H}_N\mathbf{M}(t_N, t_0) \end{pmatrix} \tag{2.28}$$

and $\hat{\mathbf{R}}$ is a block diagonal matrix with blocks equal to $\mathbf{R}_i, i = 0, \ldots, N$. If the background error covariance matrix is ill-conditioned, then we expect this to dominate the conditioning of the Hessian $\mathbf{G}$. We return to an examination of this in Section 3.4. On the other hand, the Hessian of the transformed problem (2.24) is given by

$$\tilde{\mathbf{G}} = \mathbf{I} + \sum_{i=0}^{N} \mathbf{U}_i^T \mathbf{M}(t_i, t_0)^T \mathbf{H}_i^T \mathbf{R}_i^{-1} \mathbf{H}_i \mathbf{M}(t_i, t_0)\mathbf{U}_i\,. \tag{2.29}$$

Usually, the number of observations is less than the number of state variables being estimated and so the Hessian (2.29) is equal to the identity plus a low rank matrix. Then, it has a minimum eigenvalue equal to one and the condition number (in the two-norm) is equal to the largest eigenvalue. Thus, we would expect the transformed problem to be better conditioned.

Of course, this theory all relies on being able to choose appropriate variables $\delta\mathbf{z}$ that are truly uncorrelated and it is here that a knowledge of the physical problem is necessary. In presenting how the transform is designed in practice, it is easier to think about it in terms of the inverse transform, from model variables $\delta\mathbf{x}$ to uncorrelated variables $\delta\mathbf{z}$. A common approach in numerical weather prediction is to split the inverse transform into two parts. The first part, which we write as $\mathbf{U}_p^{-1}$, is known as the parameter transform and transforms to physical variables $\delta\boldsymbol{\chi}$ whose errors are assumed to be uncorrelated between themselves, but still contain spatial correlations. The spatial transform, $\mathbf{U}_s^{-1}$, then removes spatial correlations between the physical variables $\delta\boldsymbol{\chi}$. We thus have the steps

$$\delta\boldsymbol{\chi} = \mathbf{U}_p^{-1}\delta\mathbf{x} \tag{2.30}$$

$$\delta\mathbf{z} = \mathbf{U}_s^{-1}\delta\boldsymbol{\chi}\,, \tag{2.31}$$

where for ease of notation, we assume the transforms to be time-invariant. In practice, the transforms $\mathbf{U}_p$ and $\mathbf{U}_s$ may not be square and a generalization of the inverse operator is needed. We now consider each of these transforms in turn.

3.2.1 Parameter transform

In designing a suitable transform of parameters $\mathbf{U}_p^{-1}$, it is necessary to have an understanding of the particular system being modeled in order to decide which variables have errors that are likely to be uncorrelated. For atmospheric models, the transform is based on the concept of balanced variables. Balance relationships are diagnostic relationships that exist between certain atmospheric variables. For example, in midlatitudes and at large horizontal length scales, the horizontal wind is approximately in balance with the gradient of the pressure field through the relationship of geostrophic balance. This relationship can be used in the parameter transform by assuming that errors in the balanced part of the flow are uncorrelated with those in the unbalanced part [7]. This can be justified by an eigenanalysis of the linearized equation set, which shows that the balanced flow can be associated with one eigenvector and the unbalanced flow with the remaining eigenvectors. Hence, under linear evolution, these will evolve independently.

The variable that best represents the balanced flow in the atmosphere is potential vorticity (PV) [59] and so it would be natural to use this variable as the basis for the parameter transform. However, the transform from PV to the original model variables requires the solution of a three-dimensional elliptic equation as part of the application of the operator $\mathbf{U}_p$. In dynamical regimes with small characteristic horizontal length scales, the PV is well-approximated by the vorticity, which only requires the solution of a two-dimensional equation [117]. Hence, early work in this area proposed a transform based on this variable [97] and this is still the basis of the parameter transform in many operational weather forecasting systems [7]. It is recognized that this approximation is not valid in all parts of the atmosphere and it has been demonstrated on simple systems that significant correlations can remain between errors in the transformed variables [66]. For this reason, attempts are being made to implement a transformation based on PV in large scale systems [8, 28].

A similar approach may be followed in other applications, for example, in ocean forecasting, though here there has been less work on the design of appropriate transforms than in the meteorological context. In many cases, it may be assumed that errors in the model variables such as salinity and temperature are uncorrelated and only the spatial transform is needed [116], but work on defining balance relationships has allowed multivariate covariances to be introduced [114].

3.2.2 Spatial transform

Once the parameter transform has been performed, it is assumed that the errors in the resulting variables are uncorrelated between themselves. At this point, it is necessary to specify the autocovariance information for each parameter through the spatial transform. In atmospheric models, it is common to assume that the transforms in the horizontal and vertical planes are separable. In most systems, a Fourier transform is used in the horizontal and for the vertical correlations a transformation to the eigenvectors of a vertical error covariance matrix is used. The order in which these transformations are performed varies between systems. If the horizontal transform is performed first, then the horizontal spectral modes are assumed to be uncorrelated and vertical correlations are specified separately for each mode. This assumption leads to correlations that are homogeneous (independent of horizontal position) and isotropic (independent of orientation) in the transformed parameters [7]. The method allows vertical correlations that vary with horizontal scale, so that features with large horizontal scale have deeper vertical correlations [36]. However, it does not allow vertical correlations to vary with horizontal position [62]. The alternative is to first perform the vertical transform and then, assuming that these modes are independent, apply the Fourier transform to each vertical mode. This allows more variation of vertical correlations with horizontal position (for example, with latitude). However, it is more difficult to obtain an appropriate variation in horizontal correlation length scales with height [62, 84]. In both cases, a scaling transformation is also needed to ensure that the variance of the transformed variables is equal to one. In an ideal case, we would like to obtain covariances that depend on both horizontal scale and horizontal position. This has led to the development of spatial transforms based on a wavelet basis [5, 36]. Such a transform has been implemented in the operational NWP system of ECMWF.

In ocean models, the complex boundaries near the coast prohibit the simple use of a Fourier transform in the horizontal and so other methods must be used to represent spatial correlations. For example, the application of a correlation operator can be shown to be equivalent to the integration of an appropriately-constructed diffusion equation [113]. This can be used to design correlation models for use in data assimilation systems with irregular boundary conditions [115, 116].

The use of transforms for spatial covariances requires the specification of correlation length scales and variances for each of the transformed variables. Since the background field is usually a short-term forecast, these statistics must represent the structure of errors in the forecasting system being used and thus be diagnosed from that. An early method for obtaining these statistical parameters used the difference between the observations and the background field (known as the *innovations*) [57]. However, a disadvantage of this method is that it relies on having a sufficient number of observations and is therefore biased towards data-dense areas. The most popular method in atmospheric data assimilation is the “NMC method” [97][4]. In this method, the difference between two different forecasts valid at the same time is taken as a proxy for forecast errors and statistics are taken over a sample of many such forecasts. In atmospheric forecasting, usually two forecasts starting 24 hours apart are used, with the earlier one run for 48 hours and the later one for 24 hours. By using an interval of 24 hours, problems arising from modeling the diurnal variation of the atmosphere are avoided. However, this means that the differences are taken over a much longer time interval than the normal background forecast, which is usually 6 or 12 hours. As a result, the covariance structures of the forecasts differences do not necessarily reflect those of the background error and often they need to be modified for use in the assimilation system [36, 62].

This has motivated the development of ensemble methods to generate statistics from shorter forecasts. Such a method for estimating background error statistics from an ensemble of short forecasts was developed for use at ECMWF in [36]. The basis of this method is that if the inputs to the assimilation system (for example, the background, observations and physical boundary conditions) are perturbed within the statistics of their errors, then the perturbation in the resulting analysis will be drawn from the distribution of analysis error. If a short forecast is produced from this analysis, then we expect the perturbation to the forecast to be drawn from the distribution of forecast error. This perturbed forecast can then be used as a background field for the next assimilation time and the process repeated to produce the next analysis and another forecast. Suppose that we run two such cycles in parallel for l cycles, starting from two different sets of perturbations at time t_0. Then, at each assimilation time $t_i, i = 1, \dots, l$, this will produce two perturbed short forecasts $\mathbf{x}_i^{b1}$ and $\mathbf{x}_i^{b2}$. It can be shown that the statistics of the true forecast error can then be calculated from the sample covariance of the differences between these pairs [6],

$$\mathbf{B} \approx \frac{1}{2(l-1)} \sum_{i=1}^{l} \left(\mathbf{x}_i^{b1} - \mathbf{x}_i^{b2}\right)\left(\mathbf{x}_i^{b1} - \mathbf{x}_i^{b2}\right)^T , \tag{2.32}$$

under the assumption that the errors in the two forecasts are uncorrelated. The factor of $1/2$ arises since the sample covariance itself is equal to the sum of the error covariances of the two different sets of forecasts. Since the forecasts used in this method are of the same length as the forecasts used to obtain the background field in the assimilation, the error statistics produced in this way are a more realistic representation of the true error statistics.

A key assumption in the methods presented so far is that the error covariance matrix represents a statistical average over time. The computational expense of calculating these statistics means that the matrix is kept constant from day to day, perhaps

[4] So-called because it was first introduced in the National Meteorological Center of the USA, now the National Center for Environmental Prediction.

with different statistics being used with a change of season. More recently, there has been interest in developing methods for estimating statistics that vary from day to day since it is expected that the actual background errors will depend on the underlying flow. Such flow-dependent statistics arise naturally in ensemble methods of data assimilation, such as the ensemble Kalman filter. Methods are currently being designed to obtain some flow-dependent information in variational assimilation by combining information from ensembles of forecasts with the statistically-averaged error covariance matrix, for example, [15, 18].

3.3 Observation errors

As well as representing the errors in the background field, it is important to treat the errors in the observations properly within a variational data assimilation system. Observational data received into operational weather and forecasting centers can contain errors from a variety of sources, including limitations in the measuring instrument, biases in the measurements and errors simply due to human error in recording the measurement. Furthermore, other errors arise from the way the data are used within the data assimilation system, both from inaccuracies in the operators used to map the model state to observation space and from the differences in spatial resolution between the model and the observations. The theory of variational data assimilation assumes that all observational errors are random, unbiased errors with a Gaussian distribution and known covariance. It is therefore important that as many of these sources of error as possible are accounted for in the data assimilation system.

A first essential step in an operational data assimilation system is to perform a quality control check on the data themselves. This may consist of several stages. First, a check for obvious errors in the reporting of the data is made, for example, errors in the reported position. For example, if a ship observation is reported over a land point, it will be rejected from the assimilation. Then, a so-called “background check” may be made to see how close the observation is to the forecast background field. If the difference from the background is too large when compared with its expected error variance, then the observation may be rejected and not used in the assimilation [2]. Once this check has been performed, the next step is to identify observations that may have gross errors, that is, errors that are unlikely to satisfy the assumption of being random and normally distributed. This can be done either outside or within the assimilation process. Outside the assimilation, each observation can be checked against nearby observations and any observations that largely disagree with others can be rejected [100]. Alternatively, this check can be included in the assimilation process using the variational quality control method [2, 63]. In this method, the probability density function of the observation errors are assumed to be a weighted combination of a standard Gaussian distribution and a flat probability distribution function, with the weights determined by the probability of gross error of the observation. Thus, for each single observation y with weight α_y, the probability density function of the observation error is assumed to be of the form

$$\mathcal{P}_{QC} = (1 - \alpha_y)\mathcal{P}_N + \alpha_y \mathcal{P}_F \,, \tag{2.33}$$

where $\mathcal{P}_N$ indicates the appropriate Gaussian probability density function and $\mathcal{P}_F$ is a flat distribution over a finite interval centered at zero and is equal to zero outside this interval (the size of this interval is taken to be a multiple of the observation error standard deviation). The observation part of the cost function is then taken to be equal to the negative logarithm of $\mathcal{P}_{QC}$. In the case where $\alpha_y = 0$, this corresponds to the observation term in the original nonlinear cost function (2.3). In this method, observations that have a high probability of gross error are given very little weight in the analysis. Initially, these probabilities are assigned to each observation based on a study of historical data. The probabilities are then updated on each iteration of the minimization procedure by comparison with the current estimate of the state to allow observations to be given more or less weight as the assimilation progresses. The introduction of non-Gaussianity means that variational quality control can introduce multiple minima into the cost function and so it is necessary to have a good starting point for the minimization. For this reason the minimization is first run for several iterations without the quality control term before switching it on [1].

A second important aspect of observation errors is the treatment of systematic errors, or biases, in the observations. This is particularly important for satellite radiance data where biases may occur from changes in the measuring instrument over time or from errors in the radiative transfer model needed as part of the observation operator [54]. Since the assimilation scheme assumes that the observations are unbiased, any biases in the observations can introduce biases into the analyses. As with the quality control, these biases may be treated offline or within the assimilation scheme. For each satellite channel, a bias model is assumed in such a way that we can define a new observation operator for the biased measurement

$$\tilde{\mathcal{H}}(\mathbf{x}, \boldsymbol{\beta}) = \mathcal{H}(\mathbf{x}) + \mathbf{b}(\boldsymbol{\beta}, \mathbf{x}) \,, \tag{2.34}$$

with

$$\mathbf{b}(\boldsymbol{\beta}, \mathbf{x}) = \sum_{j=0}^{N_p} \beta_j \mathbf{p}_j(\mathbf{x}) \,, \tag{2.35}$$

where $\mathbf{p}_j$ are predictors for $j = 0, \dots, N_p$ and β_j are scalar coefficients [33]. A few predictor states are chosen that may be related to the state at the observation positions. The coefficients $\boldsymbol{\beta}$ can then be estimated in an offline regression using a few weeks of data [54] or a variational procedure can be used to estimate these coefficients. This can be included directly in the assimilation procedure by including (2.34) in the cost function in place of the standard observation operator and including a background estimate $\boldsymbol{\beta}^b$ of $\boldsymbol{\beta}$ with covariance $\mathbf{B}_{\boldsymbol{\beta}}$. The 4DVar assimilation problem is then to minimize

$$\begin{aligned}\mathcal{J}_{\boldsymbol{\beta}}(\mathbf{x}_0,\boldsymbol{\beta}) = {} & \frac{1}{2}\left(\mathbf{x}_0-\mathbf{x}^b\right)^{\mathrm{T}}\mathbf{B}^{-1}\left(\mathbf{x}_0-\mathbf{x}^b\right)+\frac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^b\right)^{T}\mathbf{B}_{\boldsymbol{\beta}}^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^b\right)\\ & +\frac{1}{2}\sum_{i=0}^{N}\left(\mathcal{H}_i(\mathbf{x}_i)+\mathbf{b}(\boldsymbol{\beta},\mathbf{x}_i)-\mathbf{y}_i\right)^{\mathrm{T}}\mathbf{R}_i^{-1}\left(\mathcal{H}_i(\mathbf{x}_i)+\mathbf{b}(\boldsymbol{\beta},\mathbf{x}_i)-\mathbf{y}_i\right)\end{aligned} \tag{2.36}$$

subject to the dynamical equations in order to estimate the state $\mathbf{x}_0$ and the coefficients $\boldsymbol{\beta}$ simultaneously [33]. Alternatively, a variational procedure can be used to estimate these coefficients offline at regular intervals using the previous value as the background for the new estimate [3].

Finally, we consider the specification of the observation error covariance matrix, which represents the covariance of the random components of the observation error. It is important to note that this error is defined by the difference between the actual measurement and the model representation of the true state $\mathbf{x}^t$ mapped into observation space by the observation operator, that is, the error $\boldsymbol{\epsilon}_i^o$ at time t_i is given by

$$\boldsymbol{\epsilon}_i^o = \mathbf{y}_i - \mathcal{H}_i\left(\mathbf{x}_i^t\right) . \tag{2.37}$$

This means that the error includes different components arising from the accuracy of the measuring instrument (instrument error), errors in the observation operator $\mathcal{H}_i$ and errors due to the difference in spatial resolution between the measurement and the model state (known as the representativity error). The instrument error is the easiest to treat since the variances of this error can usually be obtained from the instrument manufacturer and it is normally safe to assume that these errors are uncorrelated. However, this may not always be the case. For example, measurements derived by preprocessing satellite data may include spatial correlations [17]. Errors in the observation operator may include such things as errors in the radiative transfer models used to model satellite data which can lead to error correlations between different satellite channels [16, 105].

Although it is recognized that observation error correlations exist, particularly with respect to satellite data, the correlations are not usually very well treated in current operational forecasting systems. Often, the correlations are ignored and it is assumed that the observation error covariance matrix is diagonal. To balance this assumption, either the error variances are inflated [56] or the data are thinned so that fewer of them are used [32]. The reasons for this are the difficulty in calculating what the error correlations should be and the difficulty in then representing these correlations within an assimilation scheme in a way that the inverse correlation matrix can easily be applied. To estimate the correlations in satellite data, the methods that have mainly been used are a comparison with independent measurements from radiosondes based on the method of [57], and the use of diagnostics calculated from the data assimilation system itself based on [35]. Various ways of representing these correlations within the data assimilation system have been proposed, including the use of a circulant matrix [55], an eigenvalue decomposition [37] and a Markov matrix [105]. However, thus far, little use of these methods exists in operational practice.

3.4 Optimization methods

The minimization of the inner loop cost function (2.15) requires the use of a suitable optimization algorithm. For the large problems of environmental modeling, there are two particularly important constraints. The first is that because of the number of variables in the system, it is not possible to obtain second derivative information. The Hessian or second derivative matrix would contain of the order 10^{16} elements, which is impossible to calculate or to store. Hence, only methods that require first derivative information can be used. The second constraint is that often these problems must be solved within a real-time forecasting system and hence the computer time that can be used to solve the problem is very limited. Hence, the methods use as few function evaluations as possible. This means that usually the problem is not allowed to run to full convergence and the use of any line search algorithms is prohibitively expensive. Traditionally, the algorithms that are most commonly used within data assimilation systems are quasi-Newton algorithms and conjugate gradient or related Lanczos algorithms, which only require first derivative information to be provided. The mathematical details of these algorithms are well explained elsewhere (e.g. [94]), and so here we limit the discussion to their implementation in data assimilation systems.

An essential aspect of the minimization procedure for variational data assimilation is an appropriate preconditioning. Experimental evidence indicates that the Hessian of the inner loop cost function (2.15) is badly conditioned and that this arises from the ill-conditioning of the background error covariance matrix [83]. This has been further confirmed by theoretical results that bound the condition number of the Hessian of the cost function in terms of the condition number of this covariance matrix [50, 51]. The first level of preconditioning that is applied is therefore to transform the problem to new variables, as described in Section 3.2. The transformed problem (2.24) can be shown in general to be better conditioned both in theory and in practice [42, 50, 51, 83]. However, even after this transformation, the problem is not very well-conditioned and can have a condition number of order 10^3–10^4 [39, 52]. Experiments in the ECMWF system showed that the ill-conditioning that remains is related to the inclusion of dense, accurate surface observations over Europe [110] and this has also been shown to be true for the system of the Met Office [52]. This can be explained by theoretical bounds obtained by [50, 52], which show that the condition number of the transformed problem increases as the spacing between observations decreases and as observations become more accurate. Hence, ideally, a second level of preconditioning is required after the variable transformation has been performed.

In order to implement a further preconditioning, it is necessary to find a preconditioning matrix $\mathbf{K}$ that is inexpensive to compute and such that the eigenvalues of $\mathbf{K}\tilde{\mathbf{G}}$ are more clustered than those of the Hessian $\tilde{\mathbf{G}}$ of the transformed problem. Often, the preconditioning matrix may be represented in the factored form $\mathbf{K} = \mathbf{P}\mathbf{P}^T$ and the preconditioning matrix $\mathbf{P}$ is then used directly, for example in the preconditioned conjugate gradient method [111]. In order to design such a preconditioner, some knowledge of the Hessian (2.29) of the transformed cost function is required. One way that this can be obtained is by using a Lanczos algorithm to perform the inner loop minimization. The Lanczos method produces estimates of the leading eigenvectors and eigenvalues of the Hessian of the function being minimized. If the first m eigenvalues λ_j and eigenvectors $\mathbf{u}_j, j = 1, \dots, m$ have sufficiently converged, then the inverse of the Hessian (2.29) can be approximated by the expression

$$\mathbf{K} = \mathbf{I} + \sum_{j=1}^{m} (\lambda_j - 1)\mathbf{u}_j\mathbf{u}_j^T \,. \tag{2.38}$$

This expression can then be used for the preconditioning of subsequent minimizations under the assumption that the Hessian does not change greatly between one minimization and another [39, 111]. This method, known as spectral preconditioning, is used in the operational forecast system of ECMWF, where three outer loops are performed for each assimilation. During the first inner loop minimization, the Lanczos vectors are stored and these are then used to precondition the minimization of the second and third inner loop cost functions [39]. It has been shown that this preconditioner belongs to a larger class of limited memory preconditioners [111]. In order to define this class, we let $\mathbf{s}_i \in \mathbb{R}^n, i = 1, \dots, l$, with $l < n$, be a set of $\tilde{\mathbf{G}}$-conjugate vectors. Then, the limited-memory preconditioning matrix is given by

$$\mathbf{K}_l = \left(\mathbf{I}_n - \sum_{i=1}^{l} \frac{\mathbf{s}_i\mathbf{s}_i^T}{\mathbf{s}_i^T\tilde{\mathbf{G}}\mathbf{s}_i}\tilde{\mathbf{G}}\right)\left(\mathbf{I}_n - \sum_{i=1}^{l} \tilde{\mathbf{G}}\frac{\mathbf{s}_i\mathbf{s}_i^T}{\mathbf{s}_i^T\tilde{\mathbf{G}}\mathbf{s}_i}\right) + \sum_{i=1}^{l} \frac{\mathbf{s}_i\mathbf{s}_i^T}{\mathbf{s}_i^T\tilde{\mathbf{G}}\mathbf{s}_i} \,. \tag{2.39}$$

If the vectors $\mathbf{s}_i$ are chosen to be the eigenvectors of $\tilde{\mathbf{G}}$, then this formula results in the spectral preconditioning matrix (2.38).

The authors of [111] propose an alternative preconditioner from the same class based on the Ritz pairs of the Hessian. Ritz pairs are approximate eigenpairs $(\theta_i, \mathbf{v}_i)$ defined in an appropriately chosen subspace. By choosing the subspace to be that spanned by the Lanczos vectors, the authors obtain the Ritz limited memory preconditioner

$$\mathbf{K}_l^{\text{Ritz}} = \left(\mathbf{I}_n - \sum_{i=1}^{l} \frac{\mathbf{v}_i\mathbf{v}_i^T}{\theta_i}\tilde{\mathbf{G}}\right)\left(\mathbf{I}_n - \sum_{i=1}^{l} \tilde{\mathbf{G}}\frac{\mathbf{v}_i\mathbf{v}_i^T}{\theta_i}\right) + \sum_{i=1}^{l} \frac{\mathbf{v}_i\mathbf{v}_i^T}{\theta_i} \,. \tag{2.40}$$

They found that the use of this preconditioner can provide an improvement over spectral preconditioning when the estimates of the Hessian eigenpairs are inaccurate.

A similar result was also found in the Regional Ocean Modeling System (ROMS) in which both of these preconditioners are implemented [91]. One drawback of both of these methods is that in order to generate the required information, the first minimization must be performed in order to generate the vectors $\mathbf{s}_i$ before any preconditioning can be applied. Thus far, little attention has been paid to preconditioning of this first minimization.

With any minimization method, it is important to specify appropriate stopping criteria and this is also the case in variational data assimilation. As discussed in Section 2.1, it has been proved that the inner loop step of the Gauss–Newton method (step 1.1 of Algorithm 2.1) needs to be solved to sufficient accuracy in order to ensure convergence of the outer loops [45]. The theory has been used to show how it is natural to use an inner loop stopping criterion based on the relative change in the norm of the gradient, of the form

$$\frac{\|\nabla \tilde{\mathcal{J}}^{(k)}_{(l)}\|_2}{\|\nabla \tilde{\mathcal{J}}^{(k)}_{(0)}\|_2} < \epsilon\,, \tag{2.41}$$

where the subscript indicates the inner loop iteration index and ϵ is a specified tolerance [73]. The tolerance used to stop the iterations must therefore be chosen carefully. If it is too high, then there is no guarantee that the outer loop steps will converge. However, the convergence should not be pushed below the level of noise on the observations, as then small spatial scales are adjusted to fit the observational noise [68]. In many practical forecasting problems, such care is not always taken and other criteria are introduced. There are two main reasons for this. One is that in a time-critical forecasting system, it may be considered more important to solve each minimization problem using approximately the same amount of wall-clock time rather than to the same accuracy. The second reason is that the preconditioning techniques described in this section require a minimum number of iterations to be performed on the first inner loop minimization in order to acquire sufficiently accurate information about the Hessian. Hence, criteria that have been introduced include stopping the iterations when the value of the cost function is close to its expected minimum value [84] or using a fixed number of iterations, particularly for the first minimization [110].

3.5 Reduced order approaches

As was mentioned in Section 2.1, a major advantage of the incremental approach is that the inner loop problem may be solved in a smaller dimensional space than the outer loop update of the linearization trajectory. Within environmental prediction, lower spatial resolution systems have often been used in the inner loop step, with the full resolution nonlinear model being used in the outer loop. Further simplifications may also be made to the linear dynamical model used in the inner loop, such as using simplified parametrizations of subgrid-scale processes as described in Section 3.1.

While a change in resolution is certainly the simplest way to achieve a more computationally tractable inner loop problem, it does not necessarily provide the most accurate low order representation of the linearized cost function and its constraint. In order to improve on this, other reduced order approaches have been investigated in the context of incremental 4DVar. These essentially fall into two categories, methods based on principal component analysis and methods based on near-optimal reduction of dynamical systems.

Principal component analysis, which is often referred to as principal orthogonal decomposition (POD) or the method of empirical orthogonal functions (EOFs), aims to represent the solution of the assimilation problem as a linear combination of basis vectors. The basis vectors are chosen to represent the leading directions of variability in the model and are calculated using a series of model states, or "snapshots", from an integration of the nonlinear model. Such a method was used in an ocean model assimilation by [101]. From the sample of model states, the authors generate the matrix $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_l)$, where $\mathbf{X}_i$ is the difference between the model state at time t_i and the mean state. The covariance matrix $\mathbf{X}\mathbf{X}^T/(l-1)$ is then diagonalized to find a set of orthonormal eigenvectors $\mathbf{v}_i$ (EOFs) and associated eigenvalues $\lambda_i, i = 1, \ldots, l$.[5] The solution $\delta\mathbf{x}_0$ to the inner loop minimization problem (2.15) is then defined by an expansion of the leading r eigenvectors

$$\delta\mathbf{x}_0 = \sum_{i=0}^{r} w_i \mathbf{v}_i = \mathbf{V}\mathbf{w} \tag{2.42}$$

where $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_r)$ is the matrix of the leading r eigenvectors and the vector $\mathbf{w} = (w_1, \ldots, w_r)^T$ contains the weights to be determined. In this case, the matrix $\mathbf{V}$ acts as a variable transformation in a similar way to the parameter transform (2.23) and so the background term can be written in the form

$$\mathcal{J}_b(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T \mathbf{B}_w^{-1} \mathbf{w}, \tag{2.43}$$

where the covariance matrix $\mathbf{B}_w$ is taken to be the diagonal matrix of eigenvalues. The number of vectors r that are used in the expansion is chosen in order to ensure that a large fraction of the total variance is retained, where this fraction is calculated from the eigenvalues as

$$\frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{l} \lambda_i}. \tag{2.44}$$

This method has been applied to assimilation in ocean models in an idealized setting [101] and using real data [60]. It is noted that the assumption behind this method is that the variability of the system can be well described by a low-dimensional space.

[5] In practice, the eigenvalues can be found by diagonalizing the much smaller matrix $\mathbf{X}^T\mathbf{X}/(l-1)$ [11].

Although the approach reduces the size of the space in which the minimization is performed, the tangent linear model (2.16) must still be integrated at full resolution on each iteration.

An alternative approach, based on POD, was put forward by [22, 23]. In that work, the solution to the full nonlinear 4DVar problem is expressed as a perturbation from the sample mean that is expanded in terms of basis functions $\mathbf{\Phi}_i$ such that

$$\delta \mathbf{x}_0 = \sum_{i=0}^{r} w_i \mathbf{\Phi}_i \,, \tag{2.45}$$

where w_i are again weights to be determined. The basis functions are derived in a similar way to the EOFs, but by then projecting the perturbation fields $\mathbf{X}$ onto the eigenvectors $\mathbf{v}_i$, and thus

$$\mathbf{\Phi} = \{\mathbf{\Phi}_1, \dots, \mathbf{\Phi}_l\} = \mathbf{XV} \,. \tag{2.46}$$

The number of basis functions that are used in the expansion is again determined using the fractional variance (2.44). In this work, the authors solve the nonlinear 4DVar cost function (2.3) in the reduced space. As well as expressing the background term in terms of coefficients of the basis functions, they also derive a Galerkin projection of the dynamical model onto the basis functions for use in the observation term. Thus, this formulation has the advantage that the dynamical model and its adjoint are also expressed in the reduced space. Again, this method relies on the snapshots being able to capture a low-dimensional subspace that adequately describes the full system.

A disadvantage with both the EOF and POD methods is that they do not use any information about the data assimilation problem itself within the reduction procedure. There have been two approaches proposed to improve on this. The first is an adaption of the POD method, called dual-weighted POD. In this method the snapshot perturbations $\mathbf{X}$ are weighted according to the sensitivity of the cost function at the time of the snapshot, where the weights are calculated using the adjoint model [30]. The other approach, put forward in the series of papers [14, 75, 76], is to use near-optimal model order reduction methods for linear dynamical systems to derive a reduced order model and observation operator. The inner loop problem of incremental 4DVar (2.15) is subject to the dynamical system described by the evolution equation (2.16) and the output equation

$$\mathbf{d}_i = \mathbf{H}_i \delta \mathbf{x}_i \,. \tag{2.47}$$

Model reduction seeks linear restriction operators $\mathbf{S}_i^T$ and prolongation operators $\mathbf{T}_i$ that map the perturbation $\delta \mathbf{x}_i \in \mathbb{R}^n$ to $\delta \hat{\mathbf{x}}_i \in \mathbb{R}^r$ with $r \ll n$. These operators are chosen such that the output of the projected system

$$\delta \hat{\mathbf{x}}_{i+1} = \mathbf{S}_i^T \mathbf{M}_i \mathbf{T}_i \delta \hat{\mathbf{x}}_i \tag{2.48}$$

$$\hat{\mathbf{d}}_i = \mathbf{H}_i \mathbf{T}_i \delta \hat{\mathbf{x}}_i \tag{2.49}$$

approximates well the output of the full dynamical system $\mathbf{d}_i$. The inner loop problem can then be defined in the reduced space as the minimization of

$$\begin{aligned}\min \hat{\mathcal{J}}^{(k)}\left[\delta\hat{\mathbf{x}}_0^{(k)}\right] = \frac{1}{2}&\left(\delta\hat{\mathbf{x}}_0^{(k)} - \mathbf{S}_0^T\left[\mathbf{x}^b - \mathbf{x}_0{}^{(k)}\right]\right)^{\mathrm{T}}\\ &\times\left(\mathbf{S}_0^T\mathbf{B}_0\mathbf{S}_0\right)^{-1}\left(\delta\hat{\mathbf{x}}_0^{(k)} - \mathbf{S}_0^T\left[\mathbf{x}^b - \mathbf{x}_0{}^{(k)}\right]\right)\\ &+\frac{1}{2}\sum_{i=0}^{N}\left(\mathbf{H}_i\mathbf{T}_i\delta\hat{\mathbf{x}}_i^{(k)} - \mathbf{d}_i^{(k)}\right)^{\mathrm{T}}\mathbf{R}^{-1}\left(\mathbf{H}_i\mathbf{T}_i\delta\hat{\mathbf{x}}_i^{(k)} - \mathbf{d}_i^{(k)}\right),\end{aligned}$$

subject to the reduced dynamical model (2.48). The linearization state is then updated with the perturbation

$$\delta\mathbf{x}_0^{(k)} = \mathbf{T}_0\delta\hat{\mathbf{x}}_0^{(k)}. \tag{2.50}$$

The authors of these papers use the method of balanced truncation [92] to demonstrate this method in the case where the operators **M** and **H** are time-invariant. The aim of balanced truncation is to truncate the states of the system that are least affected by the inputs and have least effect on the outputs. Since these are not generally the same, the first step in the method is to transform the system into one in which these states coincide, the "balancing" step. It is first necessary to find the state covariance matrices **P** and **Q** associated with the inputs and outputs respectively. These are found by solving the Stein equations

$$\mathbf{P} = \mathbf{M}\mathbf{P}\mathbf{M}^T + \mathbf{B} \tag{2.51}$$

$$\text{and}\quad \mathbf{Q} = \mathbf{M}^T\mathbf{Q}\mathbf{M} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}. \tag{2.52}$$

The balancing transformation $\boldsymbol{\Psi}$ is then given by the matrix of eigenvectors of **PQ**, while the eigenvalues of **PQ** are equal to the Hankel singular values of the full system. The reduction step then calculates the restriction and prolongation operators from

$$\mathbf{S}^T = [\mathbf{I}_r, \mathbf{0}]\,\boldsymbol{\Psi}^{-1} \tag{2.53}$$

$$\mathbf{T} = \boldsymbol{\Psi}\begin{bmatrix}\mathbf{I}_r\\ \mathbf{0}\end{bmatrix}, \tag{2.54}$$

where the decay of the Hankel singular values is used to choose the model reduction order r. In idealized models the studies [14, 75, 76] show how this method improves the solution with respect to using low resolution models and how it is important to use information about the assimilation problem in the reduction procedure, including information about the background and observation error covariance matrices. However, whereas reduction methods based on POD can be implemented in large systems, the method of balanced truncation cannot. Although efficient numerical methods are available to apply balanced truncation to systems of moderately large size (e.g. [25, 53, 69]), these are not suitable for the very large systems found in environmental prediction. Efforts are being made to design near-optimal reduction methods for such systems based on Krylov methods [21], but these methods have not yet been tried out in data assimilation for large systems.

3.6 Issues for nested models

For very high resolution weather and ocean forecasting operational centers often use models covering only the domain of interest that are nested in a larger model, often of lower resolution, which we refer to here as the parent model. In most of the systems the nesting is a one-way nesting, whereby lateral boundary conditions for the nested model are provided by the parent model, but there is no feedback from the high resolution nested model to the parent model. This presents particular challenges for the application of variational data assimilation. For problems specific to high resolution weather forecasting we refer the reader to the review articles [96] and [31]. Here we consider only more general problems arising from using a high resolution nested grid, in particular treatment of the lateral boundary conditions and of the difference in representation of spatial scales between the parent and nested models.

With respect to the lateral boundary conditions, a decision must be made as to whether to estimate them as part of the assimilation procedure or to assume that they do not change. Both approaches have been used in practice. In the operational weather forecasting system of the Met Office the lateral boundary conditions are not updated, but are fixed by the parent model. Hence the increment $\delta\mathbf{x}$ on the boundary is set to zero. This has advantages for the practical implementation of the scheme. In particular it allows a simple sine transform to be used in the definition of the spatial background error covariances described in Section 3.2, which then enforces zero boundary increments [83]. However, observational information close to the boundaries can be difficult to use, since the nested model cannot use observations lying outside the domain and the analysis inside the domain may not be consistent with the boundary conditions provided [4, 47]. This can lead to features being artificially cut-off close to the boundaries.

The alternative approach is to estimate the boundary variables within the assimilation procedure [48, 49, 67]. This means that the state vector $\mathbf{x}$ is defined to include both the variables in the interior of the domain and on the lateral boundaries. In this way observations inside the nested domain can update the boundary values and so it is possible to ensure that the analysis is consistent throughout the domain. However in this case it is no longer possible to apply a sine transform to impose the spatial background error covariances. In order to be able to apply a spectral transformation an extension zone is created around the domain to obtain fields that are horizontally periodic. A Fourier transform can then be applied. One difficulty in analyzing the boundaries in this way is that the lateral boundary conditions are only updated during the assimilation period. During the subsequent forecast no updates are available and the values from the parent model must be used, so there is some inconsistency between the boundary conditions of the analysis and those of the forecast. However, some consistency over the assimilation window can be ensured by estimating the boundary conditions at the beginning and end of the assimilation window, with both constrained by background values from the parent model. In this case the cost function to be solved is of the form

$$\begin{aligned}\mathcal{J}\left(\mathbf{x}_0, \mathbf{x}_{lbc}\right) = {}& \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}^b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}^b) + \frac{1}{2}\left(\mathbf{x}_{lbc} - \mathbf{x}^b_{lbc}\right)^{\mathrm{T}}\mathbf{B}^{-1}_{lbc}\left(\mathbf{x}_{lbc} - \mathbf{x}^b_{lbc}\right) \\ & + \frac{1}{2}\sum_{i=0}^{N}\left(\mathcal{H}_i\left(\mathbf{x}_i\right) - \mathbf{y}_i\right)^{\mathrm{T}}\mathbf{R}_i^{-1}\left(\mathcal{H}_i\left(\mathbf{x}_i\right) - \mathbf{y}_i\right) ,\end{aligned} \tag{2.55}$$

where $\mathbf{x}_0$ represents the model variables in the interior of the domain and the lateral boundary conditions at initial time t_0, $\mathbf{x}_{lbc}$ is the lateral boundary condition at final time t_N, $\mathbf{x}^b$ is the background estimate of $\mathbf{x}_0$, with error covariance matrix $\mathbf{B}$ and $\mathbf{x}^b_{lbc}$ is the background estimate of $\mathbf{x}_{lbc}$, with error covariance matrix $\mathbf{B}_{lbc}$ [67].

The second challenge we consider is the difference in the spatial scales that can be represented in the nested and parent models. In particular, since the nested model often covers only a small domain, the assimilation scheme is not able to adequately analyze scales of the size of the domain and larger. In applications such as weather prediction, it is important to capture these larger scales since the physical system is inherently multiscale with strong feedbacks between large and small scales. Hence, attempts have been made to improve the large scale information in nested model data assimilation by providing information on these scales from a parent model analysis. For example, the Met Office experimented with a system that combined large scale increments from a parent model analysis with the small scale increments from the nested model analysis [4]. In this method, the large scales of the nested model analysis are forced to be equal to those of the parent model.

An alternative, proposed by [47], is to use the large scales of the parent analysis over the nested model domain as a weak-constraint on the variational problem. We let $\mathbf{x}^a_p$ be the analysis from the parent model and define operators $\mathcal{H}_p$ and $\mathcal{H}_n$ such that $\mathcal{H}_p(\mathbf{x}^a_p)$ represents some large scales of the parent analysis on the nested domain and $\mathcal{H}_n(\mathbf{x})$ represents the same large scales from the nested model field $\mathbf{x}$. Then, the difference between the large scales of the global analysis and those forecasted by the nested model can be constrained by adding an extra term to the cost function (2.3) of the form

$$\frac{1}{2}\left(\mathcal{H}_p\left(\mathbf{x}^a_p\right) - \mathcal{H}_n\left(\mathbf{x}\right)\right)^T \mathbf{B}_p^{-1}\left(\mathcal{H}_p\left(\mathbf{x}^a_p\right) - \mathcal{H}_n\left(\mathbf{x}\right)\right) , \tag{2.56}$$

where $\mathbf{B}_p$ is the error covariance matrix of the parent model large scales. This means that the analysis is constrained by large scales from the parent model through this additional term, and by large scales from the nested model through the background term. In theory, this should introduce another term including the cross-correlation between these two sources of information. However, in their demonstration of the method in a 3DVar scheme of the ALADIN model at Météo-France, the authors of [47] concluded that this cross-correlation could be neglected, though at the cost of some inaccuracy.

A more theoretical study of this problem was carried out by [9]. They used a spectral analysis to show how information from waves longer than the domain size is projected onto different scales in the nested model domain corresponding to the lowest wave numbers that can be represented on this domain. They demonstrated that by giving more weight to these scales in the background term of the cost function, it was possible to retain more of the large scale information from a parent model background. In this method, the large spatial scales from only the parent model are used as a constraint in the assimilation, as in [4], but they are not imposed exactly and may be altered by the assimilation process. The authors of [9] demonstrated benefit from this in an idealized system, but the method has not been tested in a realistic model.

3.7 Weak-constraint variational assimilation

The formulation of variational data assimilation presented in Section 2 assumes that the discrete dynamical model (2.1) is an exact representation of the physical system being observed. In practice, we know that the models contain errors caused by limitations in our knowledge of the physical equations and limitations in the numerical modeling, for example, the need for subgrid scale parametrizations. In theory, it is possible to account for and estimate such errors in variational data assimilation, though implementation in practice is more complicated. We assume an additive error to the model equations, and thus the true dynamical system can be written as

$$\mathbf{x}_{i+1} = \mathcal{M}_i(\mathbf{x}_i) + \boldsymbol{\eta}_i\,, \tag{2.57}$$

where $\boldsymbol{\eta}_i$ are the unknown model errors at times t_i which are assumed to be random, serially uncorrelated, Gaussian errors with covariance matrix $\mathbf{Q}_i$. Then, we can define a weak-constraint 4DVar problem in which the model equations do not have to be exactly satisfied over the assimilation window. We define a cost function of the form

$$\begin{aligned}\mathcal{J}(\mathbf{x}_0, \boldsymbol{\eta}_0, \ldots, \boldsymbol{\eta}_{N-1}) = {} & \frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}^b\right)^{\mathrm{T}} \mathbf{B}^{-1}\left(\mathbf{x}_0 - \mathbf{x}^b\right) \\ & + \frac{1}{2}\sum_{i=0}^{N}\left(\mathcal{H}_i\left(\mathbf{x}_i\right) - \mathbf{y}_i\right)^{\mathrm{T}} \mathbf{R}_i^{-1}\left(\mathcal{H}_i\left(\mathbf{x}_i\right) - \mathbf{y}_i\right) + \frac{1}{2}\sum_{i=0}^{N-1} \boldsymbol{\eta}_i^T \mathbf{Q}_i^{-1} \boldsymbol{\eta}_i\end{aligned} \tag{2.58}$$

subject to (2.57). The weak-constraint problem is then to minimize (2.58) with respect to the initial state $\mathbf{x}_0$ and all the model errors $\boldsymbol{\eta}_i$.

An alternative formulation of the weak-constraint problem (2.58) is to write it in terms of the model state $\mathbf{x}_i$ at each time t_i rather than in terms of the model errors.

This leads to the cost function

$$\begin{aligned}\mathcal{J}(\mathbf{x}_0,\mathbf{x}_1,\dots,\mathbf{x}_N) = {} & \frac{1}{2}\left(\mathbf{x}_0-\mathbf{x}^b\right)^{\mathrm{T}}\mathbf{B}^{-1}\left(\mathbf{x}_0-\mathbf{x}^b\right) \\ & + \frac{1}{2}\sum_{i=0}^{N}\left(\mathcal{H}_i(\mathbf{x}_i)-\mathbf{y}_i\right)^{\mathrm{T}}\mathbf{R}_i^{-1}\left(\mathcal{H}_i(\mathbf{x}_i)-\mathbf{y}_i\right) \\ & + \frac{1}{2}\sum_{i=0}^{N-1}\left(\mathbf{x}_{i+1}-\mathcal{M}_i(\mathbf{x}_i)\right)^{T}\mathbf{Q}_i^{-1}\left(\mathbf{x}_{i+1}-\mathcal{M}_i(\mathbf{x}_i)\right).\end{aligned} \tag{2.59}$$

In [109], both formulations were presented in the incremental version of 4DVar as possibilities for inclusion in the ECMWF system.

The inclusion of the model errors at each observation time increases the size of the argument of $\mathcal{J}$ by a factor of $N+1$, the number of observation times. One way to reduce this cost is by assuming a relationship in time between the model errors $\boldsymbol{\eta}_i$. Theoretical work by [46] used an augmented state approach to solve for the state and the model error, with a dynamical equation used to explain the evolution of the error. The authors introduced a general form for the error evolution, including both a systematic and random component of the error. Various options for the systematic evolution were proposed, including a constant bias error and simple dynamical evolutions, and the methods were illustrated on simple systems. In the context of a regional atmospheric model, [119] demonstrated a weak-constraint 4DVar system under the assumption that the model error was serially correlated and obeyed a first-order Markov process.

Since this early work, there have been several idealized studies with weak-constraint 4DVar, but the move towards operational implementations in large scale systems has been slow. One of the biggest remaining challenges is the specification of the model error covariance matrix $\mathbf{Q}_i$ for real systems. An initial idea was to take this matrix to be a scalar multiple of the background error covariance matrix $\mathbf{B}$. However, in experiments with the ECMWF atmospheric forecasting system using formulation (2.58), [110] showed that this choice implies that corrections to the model error lie in the same space as those to the background. This leads to estimates of model error that are very similar to the increments to the initial conditions. An alternative method, proposed in the same paper, is based on the use of model tendency fields, that is, fields of the change in model variables over a model time step. The statistics of $\mathbf{Q}_i$ are estimated from an ensemble of differences between model tendency fields using the NMC method in a similar way that differences between the model fields themselves are used in the estimation of the background error covariances (as explained in Section 3.2). [110] interprets differences between these tendencies as a proxy for the uncertainty in the model forcing. The statistics from this sample are then fit to the same statistical model as is used for the matrix $\mathbf{B}$. The use of a covariance matrix estimated in this way was tested in weak-constraint 4DVar experiments that assumed a constant error over the assimilation window. This was shown to give an improvement over the use of a covariance matrix defined by a scalar multiple of **B**.

The work of [80] illustrated the implementation of weak-constraint 4DVar using such a matrix, again in the ECMWF system, to estimate a constant bias error in the stratosphere where the model is known to have biases. A similar scheme has been introduced into the operational assimilation system of ECMWF [40]. In this implementation, the deviation of the error from its mean value is minimized, and thus the last term of (2.58) becomes

$$\frac{1}{2}\left(\boldsymbol{\eta}-\bar{\boldsymbol{\eta}}\right)^{T}\mathbf{Q}^{-1}\left(\boldsymbol{\eta}-\bar{\boldsymbol{\eta}}\right)\,, \tag{2.60}$$

where $\bar{\boldsymbol{\eta}}$ is the estimate of the model bias from the previous analysis cycle. In this way, the assimilation ensures that the estimated error does not vary too quickly from one analysis cycle to the next.

Despite these initial successes, much more work is needed. One particular difficulty is that it is not clear how to differentiate between model bias and observation bias since the assimilation only measures the difference between the model and the observations. [110] showed a case study of observation bias being interpreted as a model error by weak-constraint 4DVar. This problem was discussed further by [78] in the context of ocean data assimilation. They suggested that to estimate both model and observation bias, it is necessary to include information on the spatial and temporal structure of these biases in the covariance matrices.

In order to then move away from the assumption of a constant bias and treat time-varying systematic and random model errors, more sophisticated methods for describing the evolution of errors must be developed. This evolution is likely to be dependent on the specific model being used, yet general methods for representing this are also needed. At the same time, efficient and accurate representations of the covariances of these model errors must be found. The use of the weak-constraint formulation of 4DVar holds much promise to counteract the inadequacies of models, but many challenges remain open to be able to implement this in very large environmental models.

4 Summary and future perspectives

Variational data assimilation is now a well-established method for combining observational data with very large environmental models. However, as illustrated in this article, its successful implementation requires careful and judicious choices in each aspect of the assimilation scheme. In some cases, these choices are determined by the physical system being modeled or the observational data available, for example, the specification of the error covariances in the system. In other cases, the choices may be determined by the size of the problem and the need to solve it in an efficient manner, often for real-time forecasting, or by features of the numerical model itself, such as lateral boundary conditions. In each instance, the choices to be made will inevitably be a compromise between the ideal solution and what is practically feasible in a given system. We have presented some of the solutions that have been found that have allowed variational data assimilation to be implemented in large environmental forecasting systems. Nevertheless, much research continues to improve on these solutions so as to find better estimates of the state and so produce better forecasts.

One particularly active area in numerical weather prediction is the desire to use more information from ensembles of forecasts to provide time-varying covariances for the background errors, combining the advantages of ensemble filtering methods with the advantages of 4DVar. ECMWF have implemented a system in which an ensemble of 4DVar assimilations are run and the statistics from this ensemble are used to update the variances of the background errors [15]. Extensions to this method to also calculate the covariance information are being sought. An alternative approach is to use information from ensembles of forecasts to calculate covariance information throughout the whole assimilation window. This method was proposed by [81] and tested in a global weather prediction model by [19, 20]. An advantage of this method is that the tangent linear and adjoint models are not required in the 4DVar since all the evolution information comes through the ensemble of nonlinear model forecasts. Hence, this makes development of the system much easier.

Besides the many great challenges that we have discussed in this article, new challenges are arising for the future evolution of variational data assimilation systems. The advent of massively parallel computers means that the algorithms used currently to solve the assimilation problem may no longer be efficient on future computer architectures. Hence, work is needed to develop new algorithms to solve the problem, particularly with respect to efficient minimization and preconditioning methods. This may be easier as systems move to a weak-constraint form of 4DVar but, as discussed above, that introduces its own difficulties [40]. Another challenge comes from the move towards more integrated Earth-system models, with different environmental models coupled to each other. For example, for seasonal to decadal prediction, it is now common to use coupled atmosphere-ocean models, but the initialization of these models with data assimilation is still in its infancy. Particular problems arise from the very different time scales in the atmosphere and ocean system and from the model biases in atmosphere and ocean models. Some work has been done to implement 4DVar in such systems in order to estimate the ocean state and coupling parameters [89, 106], but the estimation of the complete state in coupled atmosphere-ocean models remains an open problem for the coming years.

References

[1] W. K. Anderson and V. Venkatakrishan, Aerodynamic design optimization on unstructured grids with a continuous adjoint formulation, *Computers and Fluids* 28 (1999), 443–480.

[2] E. Andersson and H. Järvinen, Variational quality control, *Quart. J. Roy. Meteor. Soc.* 125 (1999), 697–722.

[3] T. Auligné, A. P. McNally, and D. P. Dee, Adaptive bias correction for satellite data in a numerical weather prediction system, *Quart. J. Roy. Meteor. Soc.* 133 (2007), 631–642.

[4] S. Ballard, Z. Li, M. Dixon, S. Swarbrick, O. Stiller, and H. Lean, Development of 1–4km resolution data assimilation for nowcasting at the Met Office, in: *World Weather Research Program Symposium on Nowcasting and Very Short Range Forecasting (WSN05)*, Paper 3.02, 2005.

[5] R. N. Bannister, Can wavelets improve the representation of forecast error covariances in variational data assimilation?, *Mon. Wea. Rev.* 135 (2007), 387–408.

[6] R. N. Bannister, A review of forecast error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast error covariances, *Quart. J. Roy. Meteor. Soc.* 134 (2008), 1951–1970.

[7] R. N. Bannister, A review of forecast error covariance statistics in atmospheric variational data assimilation. II: Modelling the forecast error covariance statistics, *Quart. J. Roy. Meteor. Soc.* 134 (2008), 1971–1996.

[8] R. N. Bannister and M. J. P. Cullen, A regime-dependent balanced control variable based on potential vorticity, in: *Proceedings of ECMWF workshop in flow-dependent aspects of data assimilation*, pp. 1–13, 2007.

[9] G. M. Baxter, S. L. Dance, A. S. Lawless, and N. K. Nichols, Four-dimensional variational data assimilation for high resolution nested models, *Computers & Fluids* 46 (2011), 137–141.

[10] C. Bischof, A. Carle, G. Corliss, A. Griewank, and P. Hovland, ADIFOR: Generating derivative codes from Fortran programs, *Scientific Programming* 1 (1992), 11–29.

[11] J. Blum, F.-X. Le Dimet, and I. M. Navon, Data assimilation for geophysical fluids, *Handbook of Numerical Analysis: Computational Methods for the Atmosphere and the Oceans. Elsevier, Amsterdam* XIV (2008), 377–433.

[12] M. Bocquet, Reconstruction of an atmospheric tracer source using the principle of maximum entropy. I: Theory, *Quart. J. Roy. Meteor. Soc.* 131 (2005), 2191–2208.

[13] M. Bocquet, C. A. Pires, and L. Wu, Beyond Gaussian statistical modeling in geophysical data assimilation, *Mon. Wea. Rev.* 138 (2010), 2997–3023.

[14] C. Boess, A. S. Lawless, N. K. Nichols, and A. Bunse-Gerstner, State estimation using model order reduction for unstable systems, *Computers & Fluids* 46 (2011), 155–160.

[15] M. Bonavita, L. Isaksen, and E. Hólm, On the use of EDA background error variances in the ECMWF 4D-Var, *Quart. J. Roy. Meteor. Soc.* 138 (2012), 1540–1559.

[16] N. Bormann and P. Bauer, Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. I: Methods and application to ATOVS data, *Quart. J. Roy. Meteor. Soc.* 136 (2010), 1036–1050.

[17] N. Bormann, S. Saarinen, G. Kelly, and J. N. Thépaut, The spatial structure of observation errors in atmospheric motion vectors from geostationary satellite data, *Mon. Wea. Rev.* 131 (2003), 706–718.

[18] M. Buehner, Ensemble-derived stationary and flow-dependent background-error covariances: Evaluation in a quasi-operational NWP setting, *Quart. J. Roy. Meteor. Soc.* 131 (2005), 1013–1043.

[19] M. Buehner, P. L. Houtekamer, C. Charette, H. L. Mitchell, and B. He, Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP.

Part I: Description and single-observation experiments, *Mon. Wea. Rev.* 138 (2010), 1550–1566.

[20] M. Buehner, P. L. Houtekamer, C. Charette, H. L. Mitchell, and B. He, Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part II: One-month experiments with real observations, *Mon. Wea. Rev.* 138 (2010), 1567–1586.

[21] A. Bunse-Gerstner, D. Kubalinska, G. Vossen, and D. Wilczek, h2-norm optimal model reduction for large scale discrete dynamical MIMO systems, *Journal of Computational and Applied Mathematics* 233 (2010), 1202–1216.

[22] Y. Cao, J. Zhu, Z. Luo, and I. M. Navon, Reduced-order modeling of the upper tropical Pacific ocean model using proper orthogonal decomposition, *Computers and Mathematics with Applications* 52 (2006), 1373–1386.

[23] Y. Cao, J. Zhu, I. M. Navon, and Z. Luo, A reduced-order approach to four-dimensional variational data assimilation using proper orthogonal decomposition, *International Journal for Numerical Methods in Fluids* 53 (2007), 1571–1583.

[24] C. Cardinali, S. Pezzulli, and E. Andersson, Influence-matrix diagnostic of a data assimilation system, *Quart. J. Roy. Meteor. Soc.* 130 (2004), 2767–2786.

[25] Y. Chahlaoui and P. Van Dooren, *Model reduction of time-varying systems*, Dimension reduction of large-scale systems (Mehrmann-V. Benner, P. and D. Sorensen, eds.), Springer-Verlag, 2005, pp. 131–142.

[26] W. C. Chao and L-P. Chang, Development of a four-dimensional variational analysis system using the adjoint method at GLA. Part I: Dynamics, *Mon. Wea. Rev.* 120 (1992), 1661–1673.

[27] P. Courtier, J-N. Thépaut, and A. Hollingsworth, A strategy for operational implementation of 4D-Var, using an incremental approach, *Quart. J. Roy. Meteor. Soc.* 120 (1994), 1367–1387.

[28] M. J. P. Cullen, Four-dimensional variational data assimilation: A new formulation of the background-error covariance matrix based on a potential-vorticity representation, *Quart. J. Roy. Meteor. Soc.* 129 (2003), 2777–2796.

[29] M. J. P. Cullen, A demonstration of 4D-Var using a time-distributed background term, *Quart. J. Roy. Meteor. Soc.* 136 (2010), 1301–1315.

[30] D. N. Daescu and I. M. Navon, A dual-weighted approach to order reduction in 4DVAR data assimilation, *Mon. Wea. Rev.* 136 (2008), 1026–1041.

[31] S. L. Dance, Issues in high resolution limited area data assimilation for quantitative precipitation forecasting, *Physica D: Nonlinear Phenomena* 196 (2004), 1–27.

[32] M. L. Dando, A. J. Thorpe, and J. R. Eyre, The optimal density of atmospheric sounder observations in the Met Office NWP system, *Quart. J. Roy. Meteor. Soc.* 133 (2007), 1933–1943.

[33] D. P. Dee, Bias and data assimilation, *Quart. J. Roy. Meteor. Soc.* 131 (2005), 3323–3343.

[34] J. E. Dennis, Jr and R. B. Schnabel, *Numerical methods for unconstrained optimization and nonlinear equations*, Society for Industrial and Applied Mathematics, 1996.

[35] G. Desroziers, L. Berre, B. Chapnik, and P. Poli, Diagnosis of observation, background and analysis-error statistics in observation space, *Quart. J. Roy. Meteor. Soc.* 131 (2005), 3385–3396.

[36] M. Fisher, Background error covariance modelling, in: *ECMWF seminar on recent developments in data assimilation for atmosphere and ocean*, pp. 45–63, 2003.

[37] M. Fisher, *Accounting for correlated observation error in the ECMWF analysis*, ECMWF, Technical memorandum MF/05106, 2005.

[38] M. Fisher and D. J. Lary, Lagrangian four-dimensional variational data assimilation of chemical species, *Quart. J. Roy. Meteor. Soc.* 121 (1995), 1681–1704.

[39] M. Fisher, J. Nocedal, Y. Trémolet, and S. J. Wright, Data assimilation in weather forecasting: a case study in PDE-constrained optimization, *Optimization and Engineering* 10 (2009), 409–426.

[40] M. Fisher, Y. Trémolet, H. Auvinen, D. Tan, and P. Poli, *Weak-constraint and long-window 4D-Var*, ECMWF, Technical memorandum 655, 2011.

[41] S. J. Fletcher and M. Zupanski, A data assimilation method for log-normally distributed observational errors, *Quart. J. Roy. Meteor. Soc.* 132 (2007), 2505–2519.

[42] P. Gauthier, C. Charette, L. Fillion, P. Koclas, and S. Laroche, Implementation of a 3D variational data assimilation system at the Canadian Meteorological Centre. Part I: The global analysis, *Atmosphere-Ocean* 37 (1999), 103–156.

[43] P. Gauthier, M. Tanguay, S. Laroche, S. Pellerin, and J. Morneau, Extension of 3DVAR to 4DVAR: Implementation of 4DVAR at the Meteorological Service of Canada, *Mon. Wea. Rev.* 135 (2007), 2339–2354.

[44] R. Giering and T. Kaminski, Recipes for Adjoint Code Construction, *ACM Trans. On Math. Software* 24 (1998), 437–474.

[45] S. Gratton, A. S. Lawless, and N. K. Nichols, Approximate Gauss-Newton methods for nonlinear least squares problems, *SIAM J. Optim.* 18 (2007), 106–132.

[46] A. K. Griffith and N.K Nichols, Adjoint methods in data assimilation for estimating model error, *Flow, Turbulence and Combustion* 65 (2000), 469–488.

[47] V. Guidard and C. Fischer, Introducing the coupling information in a limited-area variational assimilation, *Quart. J. Roy. Meteor. Soc.* 134 (2008), 723–735.

[48] N. Gustafsson, L. Berre, S. Hörnquist, X. Y. Huang, M. Lindskog, B. Navascués, K. S. Mogensen, and S. Thorsteinsson, Three-dimensional variational data assimilation for a limited area model, *Tellus A* 53 (2001), 425–446.

[49] N. Gustafsson, X. Y. Huang, X. Yang, K. Mogensen, M. Lindskog, O. Vignes, T. Wilhelmsson, and S. Thorsteinsson, Four-dimensional variational data assimilation for a limited area model, *Tellus A* 64 (2012), 14985.

[50] S. A. Haben, *Conditioning and preconditioning of the minimisation problem in variational data assimilation*, Ph.D. thesis, Department of Mathematics and Statistics, University of Reading, 2011.

[51] S. A. Haben, A. S. Lawless, and N. K. Nichols, Conditioning and preconditioning of the variational data assimilation problem, *Computers & Fluids* 46 (2011), 252–256.

[52] S. A. Haben, A. S. Lawless, and N. K. Nichols, Conditioning of incremental variational data assimilation, with application to the Met Office system, *Tellus A* 63 (2011), 782–792.

[53] S. J. Hammarling, Numerical solution of the stable, non-negative definite Lyapunov equation, *IMA Journal of Numerical Analysis* 2 (1982), 303–323.

[54] B. A. Harris and G. Kelly, A satellite radiance-bias correction scheme for data assimilation, *Quart. J. Roy. Meteor. Soc.* 127 (2001), 1453–1468.

[55] S. B. Healy and A. A. White, Use of discrete Fourier transforms in the 1D-Var retrieval problem, *Quart. J. Roy. Meteor. Soc.* 131 (2005), 63–72.

[56] F. Hilton, N. C. Atkinson, S. J. English, and J. R. Eyre, Assimilation of IASI at the Met Office and assessment of its impact through observing system experiments, *Quart. J. Roy. Meteor. Soc.* 135 (2009), 495–505.

[57] A. Hollingsworth and P. Lönnberg, The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field, *Tellus A* 38 (1986), 111–136.

[58] J. R. Holton, *An introduction to dynamic meteorology*, Elsevier Academic Press, 2004.

[59] B. J. Hoskins, M. E. McIntyre, and A. W. Robertson, On the use and significance of isentropic potential vorticity maps, *Quart. J. Roy. Meteor. Soc.* 111 (1995), 877–946.

[60] I. Hoteit and A. Köhl, Efficiency of reduced-order, time-dependent adjoint data assimilation approaches, *Journal of Oceanography* 62 (2006), 539–550.

[61] X. Y. Huang, Q. Xiao, D. M. Barker, X. Zhang, J. Michalakes, W. Huang, T. Henderson, J. Bray, Y. Chen, Z. Ma, J. Dudhia, Y. Guo, X. Zhang, D-J. Won, H-C. Lin, and Y-H. Kuo, Four-dimensional variational data assimilation for WRF: Formulation and preliminary results, *Mon. Wea. Rev.* 137 (2009), 299–314.

[62] N. B. Ingleby, The statistical structure of forecast errors and its representation in the Met Office Global 3-Dimensional variational data assimilation scheme, *Quart. J. Roy. Meteor. Soc.* 127 (2001), 209–231.

[63] N. B. Ingleby and A. C. Lorenc, Bayesian quality control using multivariate normal distributions, *Quart. J. Roy. Meteor. Soc.* 119 (1993), 1195–1225.

[64] M. Janiskova, J-N. Thépaut, and J-F. Geleyn, Simplified and regular physical parametrizations for incremental four-dimensional variational assimilation, *Mon. Wea. Rev.* 127 (1999), 26–45.

[65] C. Johnson, B. J. Hoskins, and N. K. Nichols, A singular vector perspective of 4D-Var: Filtering and interpolation, *Quart. J. Roy. Meteor. Soc.* 131 (2005), 1–19.

[66] D. Katz, A. S. Lawless, N. K. Nichols, M. J. P. Cullen, and R. N. Bannister, Correlations of control variables in variational data assimilation, *Quart. J. Roy. Meteor. Soc.* 137 (2011), 620–630.

[67] T. Kawabata, H. Seko, K. Saito, T. Kuroda, K. Tamiya, T. Tsuyuki, Y. Honda, and Y. Wakazuki, An assimilation and forecasting experiment of the Nerima heavy rainfall with a cloud-resolving nonhydrostatic 4-dimensional variational data assimilation system, *J. Met. Soc. Japan. Ser. II* 85 (2007), 255–276.

[68] S. Laroche and P. Gauthier, A validation of the incremental formulation of 4D variational data assimilation in a nonlinear barotropic flow, *Tellus* 50A (1998), 557–572.

[69] A. Laub, M. Heath, C. Paige, and R. Ward, Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms, *Automatic Control, IEEE Transactions on* 32 (1987), 115–122.

[70] A. S. Lawless, A note on the analysis error associated with 3D-FGAT, *Quart. J. Roy. Meteor. Soc.* 136 (2010), 1094–1098.

[71] A. S. Lawless, S. Gratton, and N. K. Nichols, Approximate iterative methods for variational data assimilation, *International Journal for Numerical Methods in Fluids* 47 (2005), 1129–1135.

[72] A. S. Lawless, S. Gratton, and N. K. Nichols, An investigation of incremental 4D-Var using non-tangent linear models, *Quart. J. Roy. Meteor. Soc.* 131 (2005), 459–476.

[73] A. S. Lawless and N. K. Nichols, Inner-loop stopping criteria for incremental four-dimensional variational data assimilation, *Mon. Wea. Rev.* 134 (2006), 3425–3435.

[74] A. S. Lawless, N. K. Nichols, and S. P. Ballard, A comparison of two methods for developing the linearization of a shallow-water model, *Quart. J. Roy. Meteor. Soc.* 129 (2003), 1237–1254.

[75] A. S. Lawless, N. K. Nichols, C. Boess, and A. Bunse-Gerstner, Approximate Gauss–Newton methods for optimal state estimation using reduced-order models, *International Journal for Numerical Methods in Fluids* 56 (2008), 1367–1373.

[76] A. S. Lawless, N. K. Nichols, C. Boess, and A. Bunse-Gerstner, Using model reduction methods within incremental four-dimensional variational data assimilation, *Mon. Wea. Rev.* 136 (2008), 1511–1522.

[77] F.-X. Le Dimet and O. Talagrand, Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects, *Tellus* 38A (1986), 97–110.

[78] D. J. Lea, J. P. Drecourt, K. Haines, and M. J. Martin, Ocean altimeter assimilation with observational- and model-bias correction, *Quart. J. Roy. Meteor. Soc.* 134 (2008), 1761–1774.

[79] Y. Li, I. M. Navon, W. Yang, X. Zou, J. R. Bates, S. Moorthi, and R. W. Higgins, Four-dimensional variational data assimilation experiments with a multilevel semi-Lagrangian semi-implicit general circulation model, *Mon. Wea. Rev.* 122 (1994), 966–983.

[80] M. Lindskog, D. Dee, Y. Trémolet, E. Andersson, G. Radnóti, and M. Fisher, A weak-constraint four-dimensional variational analysis system in the stratosphere, *Quart. J. Roy. Meteor. Soc.* 135 (2009), 695–706.

[81] C. Liu, Q. Xiao, and B. Wang, An ensemble-based four-dimensional variational data assimilation scheme. Part I: Technical formulation and preliminary test, *Mon. Wea. Rev.* 136 (2008), 3363–3373.

[82] A. C. Lorenc, Analysis methods for numerical weather prediction, *Quart. J. Roy. Meteor. Soc.* 112 (1986), 1177–1194.

[83] A. C. Lorenc, Development of an operational variational assimilation scheme, *J. Met. Soc. Japan* 75 (1997), 339–346.

[84] A. C. Lorenc, S. P. Ballard, R. S. Bell, N. B. Ingleby, P. L. F. Andrews, D. M. Barker, J. R. Bray, A. M. Clayton, T. Dalby, and D. Li, The Met. Office global 3-dimensional variational data assimilation scheme, *Quart. J. Roy. Meteor. Soc* 126 (2000), 2991–3012.

[85] A. C. Lorenc, R. B. Bell, and B. Macpherson, The Meteorological Office analysis correction data assimilation scheme, *Quart. J. Roy. Meteor. Soc.* 117 (1991), 59–89.

[86] A. C. Lorenc and T. Payne, 4D-Var and the butterfly effect: Statistical four-dimensional data assimilation for a wide range of scales, *Quart. J. Roy. Meteor. Soc.* 133 (2007), 607–614.

[87] A. C. Lorenc and F. Rawlins, Why does 4D-Var beat 3D-Var?, *Quart. J. Roy. Meteor. Soc.* 131 (2005), 3247–3257.

[88] J. F. Mahfouf and F. Rabier, The ECMWF operational implementation of four-dimensional variational assimilation. II: Experimental results with improved physics, *Quart. J. Roy. Meteor. Soc.* 126 (2000), 1171–1190.

[89] T. Mochizuki, N. Sugiura, T. Awaji, and T. Toyoda, Seasonal climate modeling over the Indian Ocean by employing a 4D-VAR coupled data assimilation approach, *Journal of Geophysical Research* 114 (2009), C11003.

[90] A. J. F. Moodey, A. S. Lawless, R. W. E. Potthast, and P. J. van Leeuwen, Nonlinear error dynamics for cycled data assimilation methods, *Inverse Problems* 29 (2013), 025002.

[91] A. M. Moore, H. G. Arango, G. Broquet, B. S. Powell, A. T. Weaver, and J. Zavala-Garay, The Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems Part I–System overview and formulation, *Progress in Oceanography* 91 (2011), 34–49.

[92] B. Moore, Principal component analysis in linear systems: Controllability, observability, and model reduction, *Automatic Control, IEEE Transactions on* 26 (1981), 17–32.

[93] I. M. Navon, X. Zou, J. Derber, and J. Sela, Variational data assimilation with an adiabatic version of the NMC spectral model, *Mon. Wea. Rev.* 120 (1992), 1433–1446.

[94] J. Nocedal and S. J. Wright, *Numerical optimization*, Springer Verlag, 2006.

[95] T. N. Palmer, R. Gelaro, J. Barkmeijer, and R. Buizza, Singular vectors, metrics and adaptive observations, *J. Atmos. Sci.* 55 (1998), 633–653.

[96] S. K. Park and D. Zupanski, Four-dimensional variational data assimilation for mesoscale and storm-scale applications, *Meteorology and Atmospheric Physics* 82 (2003), 173–208.

[97] D. F. Parrish and J. C. Derber, The National Meteorological Center's spectral statistical-interpolation analysis system, *Mon. Wea. Rev.* 120 (1992), 1747–1763.

[98] F. Rabier and P. Courtier, Four-dimensional assimilation in the presence of baroclinic instability, *Quart. J. Roy. Meteor. Soc.* 118 (1992), 649–672.

[99] F. Rabier, H. Järvinen, E. Klinker, J. F. Mahfouf, and A. Simmons, The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics, *Quart. J. Roy. Meteor. Soc.* 126 (2000), 1143–1170.

[100] F. Rawlins, S. P. Ballard, K. J. Bovis, A. M. Clayton, D. Li, G. W. Inverarity, A. C. Lorenc, and T. J. Payne, The Met Office global four-dimensional variational data assimilation scheme, *Quart. J. Roy. Meteor. Soc.* 133 (2007), 347–362.

[101] C. Robert, S. Durbiano, E. Blayo, J. Verron, J. Blum, and F. X. Le Dimet, A reduced-order strategy for 4D-Var data assimilation, *Journal of Marine Systems* 57 (2005), 70–82.

[102] N. Rostaing, S. Dalmas, and A. Galligo, Automatic Differentiation in Odyssée, *Tellus* 45A (1993), 558–568.

[103] Y. Sasaki, An objective analysis based on the variational method, *J. Met. Soc. Japan* 36 (1958), 77–88.

[104] Y. Sasaki, Some basic formalisms in numerical variational analysis, *Mon. Wea. Rev.* 98 (1970), 875–883.

[105] L. M. Stewart, *Correlated observation errors in data assimilation*, Ph.D. thesis, Department of Mathematics, University of Reading, 2010.

[106] N. Sugiura, T. Awaji, S. Masuda, T. Mochizuki, T. Toyoda, T. Miyama, H. Igarashi, and Y. Ishikawa, Development of a four-dimensional variational coupled data assimilation system for enhanced analysis and prediction of seasonal to interannual climate variations, *Journal of Geophysical Research* 113 (2008), C10017.

[107] O. Talagrand and P. Courtier, Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory, *Quart. J. Roy. Meteor. Soc.* 113 (1987), 1311–1328.

[108] J-N. Thépaut and P. Courtier, Four-dimensional variational data assimilation using the adjoint of a multilevel primitive-equation model, *Quart. J. Roy. Meteor. Soc.* 117 (1991), 1225–1254.

[109] Y. Trémolet, Accounting for an imperfect model in 4D-Var, *Quart. J. Roy. Meteor. Soc.* 132 (2006), 2483–2504.

[110] Y. Trémolet, Model-error estimation in 4D-Var, *Quart. J. Roy. Meteor. Soc.* 133 (2007), 1267–1280.

[111] J. Tshimanga, S. Gratton, A. T. Weaver, and A. Sartenaer, Limited-memory preconditioners, with application to incremental four-dimensional variational data assimilation, *Quart. J. Roy. Meteor. Soc.* 134 (2008), 751–769.

[112] J. Vialard, A. T. Weaver, D. L. T. Anderson, and P. Delecluse, Three-and four-dimensional variational assimilation with a general circulation model of the tropical Pacific Ocean. Part II: Physical validation, *Mon. Wea. Rev.* 131 (2003), 1379–1395.

[113] A. T. Weaver and P. Courtier, Correlation modelling on the sphere using a generalized diffusion equation, *Quart. J. Roy. Meteor. Soc.* 127 (2001), 1815–1846.

[114] A. T. Weaver, C. Deltel, E. Machu, S. Ricci, and N. Daget, A multivariate balance operator for variational ocean data assimilation, *Quart. J. Roy. Meteor. Soc.* 131 (2005), 3605–3625.

[115] A. T. Weaver and I. Mirouze, On the diffusion equation and its application to isotropic and anisotropic correlation modelling in variational assimilation, *Quart. J. Roy. Meteor. Soc.* 139 (2012), 242–260.

[116] A. T. Weaver, J. Vialard, and D. L. T. Anderson, Three-and four-dimensional variational assimilation with a general circulation model of the tropical Pacific Ocean. Part I: Formulation, internal diagnostics, and consistency checks, *Mon. Wea. Rev.* 131 (2003), 1360–1378.

[117] M. Wlasak, N. K. Nichols, and I. Roulstone, Use of potential vorticity for incremental data assimilation, *Quart. J. Roy. Meteor. Soc.* 132 (2006), 2867–2886.

[118] Q. Xu, Generalized adjoint for physical processes with parameterized discontinuities. Part I: Basic issues and heuristic examples, *J. Atmos. Sci.* 53 (1996), 1123–1155.

[119] D. Zupanski, A general weak constraint applicable to operational 4DVAR data assimilation systems, *Mon. Wea. Rev.* 125 (1997), 2274–2292.

Sebastian Reich and Colin J. Cotter

Ensemble filter techniques for intermittent data assimilation

Abstract: This survey paper is written with the intention of giving a mathematical introduction to filtering techniques for intermittent data assimilation, and to survey some recent advances in the field. The paper is divided into three parts. The first part introduces Bayesian statistics and its application to statistical inference and estimation. Basic aspects of Markov processes, as they typically arise from scientific models in the form of stochastic differential and/or difference equations, are covered in the second part. The third and final part describes the filtering approach to estimation of model states by assimilation of observational data into scientific models. While most of the material is of survey type, very recent advances in the field of nonlinear data assimilation covered in this paper include a discussion of Bayesian inference in the context of optimal transportation and coupling of random variables, as well as a discussion of recent advances in ensemble transform filters. References and sources for further reading material will be listed at the end of each section.

Keywords: Bayesian inference, Monte Carlo method, sequential data assimilation

2010 Mathematics Subject Classification: 65C05, 62M20, 93E11, 62F15, 86A22

Sebastian Reich: Lehrstuhl für Numerische Mathematik, Institut für Mathematik, Universität Potsdam, Am Neuen Palais 10, 14469 Potsdam, Germany, and Department of Mathematics, University of Reading, Whiteknights, PO Box 220, RG6 6AX, United Kingdom, sreich@math.uni-potsdam.de
Colin J. Cotter: Department of Aeronautics, Imperial College London, London SW7 2AZ, United Kingdom, colin.cotter@imperial.ac.uk

1 Bayesian statistics

In this section, we summarize the Bayesian approach to statistical inference and estimation in which probability is interpreted as a measure of uncertainty (of the system state, for example). Contrary to closely related inverse problem formulations, all variables involved are considered to be uncertain and are described as random variables. Furthermore, uncertainty is only discussed in the context of available information, requiring the computation of conditional probabilities; Bayes' formula is used for statistical inference. We start with a short introduction to random variables.

The second author would like to acknowledge support from NERC grant NE/I02013X/1.

1.1 Preliminaries

We start with a *sample space* Ω which characterizes all possible outcomes of an experiment. An *event* is a subset of Ω and we assume that the set $\mathcal{F}$ of all events forms a σ-*algebra* (i.e. $\mathcal{F}$ is nonempty and closed over complementation and countable unions). For example, suppose that $\Omega = \mathbb{R}$. Then, events can be defined by taking all possible countable unions and complements of intervals $(a, b] \subset \mathbb{R}$; these are known as the *Borel sets*.

Definition 3.1 (Probability measure). A probability measure is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ with the following properties:

(i) *Total probability equals one:* $\mathbb{P}(\Omega) = 1$.

(ii) *Probability is additive for independent events:* If $A_1, A_2, \dots, A_n, \dots$ is a finite or countable collection of events $A_i \in \mathcal{F}$ and $A_i \cap A_j = \emptyset$ for $i \neq j$, then

$$\mathbb{P}\left(\cup_i A_i\right) = \sum_i \mathbb{P}\left(A_i\right).$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *probability space*.

Definition 3.2 (Random variable). A function $X : \Omega \to \mathbb{R}$ is called a (univariate) *random variable* if

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$$

for all $x \in \mathbb{R}$. The (cumulative) *probability distribution function* of X is given by

$$F_X(x) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \leq x\}\right).$$

The cumulative probability distribution function implies a probability measure on $\mathbb{R}$ which we denote by μ_X.

Often, when working with a random variable X, the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is not emphasized; one typically only specifies the *target space* $\mathcal{X} = \mathbb{R}$ and the probability distribution or *measure* μ_X on $\mathcal{X}$. We then say that μ_X is the *law* of X and write $X \sim \mu_X$. A probability measure μ_X introduces an integral over $\mathcal{X}$ and

$$\mathbb{E}_X[f] = \int_{\mathcal{X}} f(x)\mu_X(\mathrm{d}x)$$

is called the *expectation value* of a function $f : \mathbb{R} \to \mathbb{R}$ (f is called a *measurable function* where the integral exists). We also use the notation $\mathrm{law}(X) = \mu_X$ to indicate that μ_X is the probability measure for a random variable X. Two important choices for f are $f(x) = x$, which leads to the mean $\overline{x} = \mathbb{E}_X[x]$ of X, and $f(x) = (x - \overline{x})^2$, which leads to the variance $\sigma^2 = \mathbb{E}_X[(x - \overline{x})^2]$ of X.

Univariate random variables naturally extend to the multivariate case, i.e. $\mathcal{X} = \mathbb{R}^N$, $N > 1$. A probability measure μ_X on $\mathcal{X}$ is called *absolutely continuous* (with respect to the standard Lebesgue measure $\mathrm{d}x$ on $\mathbb{R}^N$) if there exists a *probability density function* (PDF) $\pi_X : \mathcal{X} \to \mathbb{R}$ with $\pi_X(x) \geq 0$, and

$$\mathbb{E}_X[f] = \int_{\mathcal{X}} f(x)\mu_X(\mathrm{d}x) = \int_{\mathbb{R}^N} f(x)\pi_X(x)\mathrm{d}x$$

for all measurable functions f. The shorthand $\mu_X(\mathrm{d}x) = \pi_X \mathrm{d}x$ is often adopted. The implication is that one can, for all practical purposes, work within the classical Riemann integral framework and does not need to resort to Lebesgue integration. Again, we can define the mean $\overline{x} \in \mathbb{R}^N$ of a multivariate random variable and its covariance matrix

$$P = \mathbb{E}_X\left[(x - \overline{x})\,(x - \overline{x})^{\mathrm{T}}\right] \in \mathbb{R}^{N \times N}.$$

Here, a^{T} denotes the transpose of a vector a. We now discuss a few standard distributions.

Example 3.3 (Gaussian distribution). We use the notation $X \sim \mathrm{N}(\overline{x}, \sigma^2)$ to denote a univariate Gaussian random variable with mean $\overline{x}$ and variance σ^2, with PDF given by

$$\pi_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\overline{x})^2}, \quad x \in \mathbb{R}.$$

In the multivariate case, we use the notation $X \sim \mathrm{N}(\overline{x}, \Sigma)$ to denote a Gaussian random variable with PDF given by

$$\pi_X(x) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\left(-\tfrac{1}{2}(x - \overline{x})^{\mathrm{T}}\Sigma^{-1}(x - \overline{x})\right), \quad x \in \mathbb{R}^N.$$

Example 3.4 (Laplace distribution and Gaussian mixtures). The univariate Laplace distribution has PDF

$$\pi_X(x) = \frac{\lambda}{2} e^{-\lambda|x|}, \quad x \in \mathbb{R}.$$

This may be rewritten as

$$\pi_X(x) = \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)} \frac{\lambda^2}{2} e^{-\lambda^2\sigma/2} \mathrm{d}\sigma,$$

which is a weighted Gaussian PDF with mean zero and variance σ^2 integrated over σ. By replacing the integral by a Riemann sum over a sequence of quadrature points $\{\sigma_j\}_{j=1}^J$, we obtain

$$\pi_X(x) \approx \sum_{j=1}^J \alpha_j \frac{1}{\sqrt{2\pi}\sigma_j} e^{-x^2/(2\sigma_j^2)}, \quad \alpha_j \propto \frac{\lambda^2}{2} e^{-\lambda^2\sigma_j/2}(\sigma_j - \sigma_{j-1})$$

and the constant of proportionality is chosen such that the weights α_j sum to one. This is an example of a *Gaussian mixture* distribution, namely, a weighted sum of Gaussians. In this case, the Gaussians are all centered on $x = 0$; the most general form of a Gaussian mixture is

$$\pi_X(x) = \sum_{j=1}^{J} \alpha_j \frac{1}{\sqrt{2\pi}\sigma_j} e^{-(x-x_j)^2/(2\sigma_j^2)} ,$$

with weights $\alpha_j > 0$ subject to $\sum_{j=1}^{J} \alpha_j = 1$, and locations $-\infty < x_j < \infty$. Univariate Gaussian mixtures generalize to mixtures of multivariate Gaussians in the obvious manner.

Example 3.5 (Point distribution). As a final example, we consider the point measure μ_{x_0} defined by

$$\int_{\mathcal{X}} f(x)\mu_{x_0}(\mathrm{d}x) = f(x_0) .$$

Using the Dirac delta notation $\delta(\cdot)$, this can be formally written as $\mu_{x_0}(\mathrm{d}x) = \delta(x - x_0)\mathrm{d}x$. The associated random variable X has the certain outcome $X(\omega) = x_0$ for almost all $\omega \in \Omega$. One can call such a random variable *deterministic*, and write $X = x_0$ for short. Note that the point measure is not absolutely continuous with respect to the Lebesgue measure, i.e. there is no corresponding probability density function.

We now briefly discuss pairs of random variables X_1 and X_2 over the same target space $\mathcal{X}$. Formally, we can treat them as a single random variable $Z = (X_1, X_2)$ over $\mathcal{Z} = \mathcal{X} \times \mathcal{X}$ with a *joint distribution* $\mu_{X_1X_2}(x_1, x_2) = \mu_Z(z)$.

Definition 3.6 (Marginals, independence, conditional probability distributions). Let X_1 and X_2 denote two random variables on $\mathcal{X}$ with joint PDF $\pi_{X_1X_2}(x_1, x_2)$. The two PDFs

$$\pi_{X_1}(x_1) = \int_{\mathcal{X}} \pi_{X_1X_2}(x_1, x_2)\mathrm{d}x_2$$

and

$$\pi_{X_2}(x_2) = \int_{\mathcal{X}} \pi_{X_1X_2}(x_1, x_2)\mathrm{d}x_1 ,$$

respectively, are called the *marginal PDFs*, i.e. $X_1 \sim \pi_{X_1}$ and $X_2 \sim \pi_{X_2}$. The two random variables are called *independent* if

$$\pi_{X_1X_2}(x_1, x_2) = \pi_{X_1}(x_1)\, \pi_{X_2}(x_2) .$$

We also introduce the *conditional PDFs*

$$\pi_{X_1}(x_1|x_2) = \frac{\pi_{X_1X_2}(x_1, x_2)}{\pi_{X_2}(x_2)}$$

and

$$\pi_{X_2}(x_2|x_1) = \frac{\pi_{X_1X_2}(x_1,x_2)}{\pi_{X_1}(x_1)} .$$

Example 3.7 (Gaussian joint distributions). A Gaussian joint distribution $\pi_{XY}(x,y)$, $x, y \in \mathbb{R}$, with mean $(\overline{x}, \overline{y})$ and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_{xx}^2 & \sigma_{xy}^2 \\ \sigma_{yx}^2 & \sigma_{yy}^2 \end{bmatrix}$$

leads to a Gaussian conditional distribution

$$\pi_X(x|y) = \frac{1}{\sqrt{2\pi}\sigma_c} e^{-(x-\overline{x}_c)^2/(2\sigma_c^2)}, \tag{3.1}$$

with conditional mean

$$\overline{x}_c = \overline{x} + \sigma_{xy}^2 \sigma_{yy}^{-2}(y - \overline{y})$$

and conditional variance

$$\sigma_c^2 = \sigma_{xx}^2 - \sigma_{xy}^2 \sigma_{yy}^{-2} \sigma_{yx}^2.$$

For a given y, we define $X|y$ as the random variable with conditional probability distribution $\pi_X(x|y)$ and write $X|y \sim \mathrm{N}(\overline{x}_c, \sigma_c^2)$.

1.2 Bayesian inference

We start this section by considering transformations of random variables. A typical scenario is the following one. Given a pair of independent random variables Ξ with values in $\mathcal{Y} = \mathbb{R}^K$ and X with values in $\mathcal{X} = \mathbb{R}^N$ together with a continuous map $h : \mathbb{R}^N \to \mathbb{R}^K$, we define a new random variable

$$Y = h(X) + \Xi . \tag{3.2}$$

The map h is called the *observation operator*, yielding observed quantities given a particular value x of the *state variable* X, and Ξ represents *measurement errors*.

Theorem 3.8 (PDF for transformed random variable). *Assume that both X and Ξ are absolutely continuous, then Y is absolutely continuous with PDF*

$$\pi_Y(y) = \int_{\mathcal{X}} \pi_\Xi\left(y - h(x)\right)\pi_X(x)\mathrm{d}x . \tag{3.3}$$

If X is a deterministic variable, i.e. $X = x_0$ for an appropriate $x_0 \in \mathbb{R}^N$, then the PDF simplifies to

$$\pi_Y(y) = \pi_\Xi\left(y - h\left(x_0\right)\right) .$$

Proof. We start with $X = x_0$. Then, $Y - h(x_0) = \Xi$ which immediately implies the stated result. In the general case, consider the conditional probability

$$\pi_Y(y|x_0) = \pi_\Xi\left(y - h\left(x_0\right)\right) .$$

Equation (3.3) then follows from the implied joint distribution

$$\pi_{XY}(x, y) = \pi_Y(y|x)\pi_X(x)$$

and subsequent marginalization, i.e.

$$\pi_Y(y) = \int\limits_{\mathcal{X}} \pi_{XY}(y, x)\mathrm{d}x = \int\limits_{\mathcal{X}} \pi_Y(y|x)\pi_X(x)\mathrm{d}x . \qquad \square$$

The problem of predicting the distribution π_Y of Y given a particular configuration of the state variable $X = x_0$ is called the *forward problem*. The problem of predicting the distribution of the state variable X given an *observation* $Y = y_0$ gives rise to an *inference problem*, which is defined more formally as follows.

Definition 3.9 (Bayesian inference). Given a particular value $y_0 \in \mathbb{R}^K$, we consider the associated conditional PDF $\pi_X(x|y_0)$ for the random variable X. From

$$\pi_{XY}(x, y) = \pi_Y(y|x)\pi_X(x) = \pi_X(x|y)\pi_Y(y),$$

we obtain *Bayes' formula*

$$\pi_X(x|y_0) = \frac{\pi_X(y_0|x)\pi_X(x)}{\pi_Y(y_0)}. \tag{3.4}$$

The object of *Bayesian inference* is to obtain $\pi_X(x|y_0)$.

Since $\pi_Y(y_0) \neq 0$ is a constant, equation (3.4) can be written as

$$\pi_X(x|y_0) \propto \pi_X(y_0|x)\pi_X(x) = \pi_\Xi\left(y_0 - h(x)\right)\pi_X(x) ,$$

where the constant of proportionality only depends on y_0. We denote by $\pi_X(x)$ the *prior PDF* of the random variable X and $\pi_X(x|y_0)$ the *posterior PDF*. The function $\pi(y_0|x)$ is called the *likelihood function*.

Having obtained a posterior PDF $\pi_X(x|y_0)$, it is often necessary to provide an estimate of a "most likely" value of x conditioned on y_0. Bayesian estimators for x are defined as follows.

Definition 3.10 (Bayesian estimators). Given a posterior PDF $\pi_X(x|y_0)$, we define a *Bayesian estimator* $\hat{x} \in \mathcal{X}$ by

$$\hat{x} = \arg\min_{x' \in \mathcal{X}} \int\limits_{\mathcal{X}} \mathrm{L}(x', x)\pi_X(x|y_0)\mathrm{d}x$$

where $\mathrm{L}(x', x)$ is an appropriate loss function. Popular choices include the *maximum a posteriori (MAP) estimator*, with $\hat{x}$ corresponding to the modal value of $\pi_X(x|y_0)$. The MAP estimator formally corresponds to the loss function $\mathrm{L}(x', x) = 1_{\{x' \neq x\}}$. The *posterior median estimator* corresponds to $\mathrm{L}(x', x) = \|x' - x\|$ while the *minimum mean square error estimator* (or *conditional mean estimator*)

$$\hat{x} = \int_{\mathcal{X}} x \pi_X(x|y_0) \mathrm{d}x$$

results from $\mathrm{L}(x', x) = \|x' - x\|^2$.

We now consider an important example for which the posterior can be computed analytically.

Example 3.11 (Bayes' formula for Gaussian distributions). Consider the case of a scalar observation, i.e. $K = 1$, with $\Xi \sim N(0, \sigma_{rr}^2)$. Then,

$$\pi_\Xi(h(x) - y) = \frac{1}{\sqrt{2\pi}\sigma_{rr}} e^{-\frac{1}{2\sigma_{rr}^2}(h(x)-y)^2} .$$

We also assume that $X \sim \mathrm{N}(\overline{x}, P)$ and that $h(x) = Hx$. Then, the posterior distribution of X given $y = y_0$ is also Gaussian with mean

$$\overline{x}_c = \overline{x} - PH^{\mathrm{T}} \left(HPH^{\mathrm{T}} + \sigma_{rr}^2 \right)^{-1} (H\overline{x} - y_0)$$

and covariance matrix

$$P_c = P - PH^{\mathrm{T}} \left(HPH^{\mathrm{T}} + \sigma_{rr}^2 \right)^{-1} HP .$$

These are the famous Kalman update formulas which follow from the fact that the product of two Gaussian distributions is also Gaussian, where the variance of $Y = HX + \Sigma$ is given by

$$\sigma_{yy}^2 = HPH^{\mathrm{T}} + \sigma_{rr}^2$$

and the vector of covariances between $x \in \mathbb{R}^N$ and $y = Hx \in \mathbb{R}$ is given by PH^{T}. For Gaussian random variables, the MAP, posterior median, and minimum mean square error estimators coincide and are given by $\overline{x}_c$. The case of vector-valued observations will be discussed in Section 3.3. Finally, note that $\overline{x}_c$ solves the minimization problem

$$\overline{x}_c = \arg\min_{x \in \mathbb{R}^N} \left\{ \frac{1}{2}(x - \overline{x})^{\mathrm{T}} P^{-1} (x - \overline{x}) + \frac{1}{2R}(Hx - y_0)^2 \right\} ,$$

which can be viewed as a regularization of the ill-posed inverse problem

$$y_0 = Hx, \quad x \in \mathbb{R}^N, \quad N > 1 ,$$

in the sense of Tikhonov. A standard Tikhonov regularization would be based on $P^{-1} = \delta I$ with the regularization parameter $\delta > 0$ appropriately chosen. In the Bayesian approach to inverse problems, the regularization term is instead determined by the Gaussian prior π_X.

We mention in passing that Bayes' formula has to be replaced by the Radon–Nikodym derivative in the case where the prior distribution is not absolutely continuous with respect to the Lebesgue measure (or in case the space $\mathcal{X}$ does not admit a Lebesgue measure). Consider, as an example, the case of an empirical measure μ_X centered about the M samples $x_i \in \mathcal{X}$, $i = 1, \dots, M$, i.e. a weighted sum of point measures given by

$$\mu_X(\mathrm{d}x) = \frac{1}{M}\sum_{i=1}^{M} \mu_{x_i}(\mathrm{d}x)\,.$$

Then, the resulting posterior measure $\mu_X(\cdot|y_0)$ is absolutely continuous with respect to μ_X, i.e. there exists a *Radon–Nikodym derivative* such that

$$\int_{\mathcal{X}} f(x)\mu_X(\mathrm{d}x|y_0) = \int_{\mathcal{X}} f(x)\frac{\mathrm{d}\mu_X(x|y_0)}{\mathrm{d}\mu_X(x)}\mu_X(\mathrm{d}x)$$

and the Radon–Nikodym derivative satisfies

$$\frac{\mathrm{d}\mu_X(x|y_0)}{\mathrm{d}\mu_X(x)} \propto \pi_\Xi\left(h(x) - y_0\right)\,.$$

Furthermore, the explicit expression for the posterior measure is given by

$$\mu_X(\mathrm{d}x|y_0) = \sum_{i=1}^{M} w_i\, \mu_{x_i}(\mathrm{d}x)\,,$$

with weights $w_i \geq 0$ defined by

$$w_i \propto \pi_\Xi\left(h(x_i) - y_0\right)\,,$$

and the constant of proportionality is determined by the condition $\sum_{i=1}^{M} w_i = 1$.

1.3 Coupling of random variables

We have seen that under Bayes' formula, a prior probability measure $\mu_X(\cdot)$ on $\mathcal{X}$ is transformed into a posterior probability measure $\mu_X(\cdot|y_0)$ on $\mathcal{X}$ conditioned on the observation $y_0 = Y(\omega)$. With each of the probability measures, we can associate random variables such that, e.g. $X_1 \sim \mu_X$ and $X_2 \sim \mu_X(\cdot|y_0)$. However, while Bayes' formula leads to a transformation of measures, it does not imply a specific transformation on the level of the associated random variables; many different transformations of random variables lead to the same probability measure. In this section, we will, therefore, introduce the concept of coupling two probability measures.

Definition 3.12 (Coupling). Let μ_{X_1} and μ_{X_2} denote two probability measures on a space $\mathcal{X}$. A *coupling* of μ_{X_1} and μ_{X_2} consists of a pair $Z = (X_1, X_2)$ of random variables such that $X_1 \sim \mu_{X_1}$, $X_2 \sim \mu_{X_2}$, and $Z \sim \mu_Z$. The joint measure μ_Z on the product space $\mathcal{Z} = \mathcal{X} \times \mathcal{X}$ is called the *transference plan* for this coupling. The set of all transference plans is denoted by $\Pi(\mu_{X_1}, \mu_{X_2})$.

Here, we will discuss different forms of couplings, assuming that both the source and target distributions are explicitly known, whilst applications to Bayes' formula (3.4) will be discussed in Sections 1.4 and 3. In practice, the source distribution often needs to be estimated from available realizations of the underlying random variable X_1. This is the subject of parametric and nonparametric statistics and will not be discussed in this survey paper. In the context of Bayesian statistics, knowledge of the source (prior) distribution and the likelihood implies knowledge of the target (posterior) distribution.

Since prior distributions in Bayesian inference are generally assumed to be absolutely continuous, the discussion of couplings will be restricted to the less abstract case of $\mathcal{X} = \mathbb{R}^N$ and $\mu_{X_1}(\mathrm{d}x) = \pi_{X_1}(x)\mathrm{d}x$, $\mu_{X_2}(\mathrm{d}x) = \pi_{X_2}(x)\mathrm{d}x$. In other words, we assume that the marginal measures are absolutely continuous. We cannot, however, assume that the coupling is absolutely continuous on $\mathcal{Z} = \mathcal{X} \times \mathcal{X} = \mathbb{R}^{2N}$. Clearly, couplings always exist since one can use the trivial product coupling

$$\pi_Z(x_1, x_2) = \pi_{X_1}(x_1)\pi_{X_2}(x_2),$$

in which case the associated random variables X_1 and X_2 are independent. The more interesting case is that of a deterministic coupling.

Definition 3.13 (Deterministic coupling). Assume that we have a random variable X_1 with law μ_{X_1} and a second probability measure μ_{X_2}. A diffeomorphism $T : \mathcal{X} \to \mathcal{X}$ is called a *transport map* if the induced random variable $X_2 = T(X_1)$ satisfies

$$\int_{\mathcal{X}} f(x_2)\mu_{X_2}(\mathrm{d}x_2) = \int_{\mathcal{X}} f\left(T(x_1)\right)\mu_{X_1}(\mathrm{d}x_1)$$

for all suitable functions $f : \mathcal{X} \to \mathbb{R}$. The associated coupling

$$\mu_Z(\mathrm{d}x_1, \mathrm{d}x_2) = \delta\left(x_2 - T(x_1)\right)\mu_{X_1}(\mathrm{d}x_1)\mathrm{d}x_2,$$

where $\delta(\cdot)$ is the standard Dirac distribution, is called a *deterministic coupling*. Note that μ_Z is not absolutely continuous, even if both μ_{X_1} and μ_{X_2} are.

Using

$$\int_{\mathcal{X}} f(x_2)\delta\left(x_2 - T(x_1)\right)\mathrm{d}x_2 = f\left(T(x_1)\right),$$

it indeed follows from the above definition of μ_Z that

$$\int_{\mathcal{X}} f(x_2)\mu_{X_2}(\mathrm{d}x_2) = \int_{\mathcal{Z}} f(x_2)\mu_Z(\mathrm{d}x_1, \mathrm{d}x_2) = \int_{\mathcal{X}} f\left(T(x_1)\right)\mu_{X_1}(\mathrm{d}x_1).$$

We discuss a simple example.

Example 3.14 (One-dimensional transport map). Let $\pi_{X_1}(x) \geq 0$ and $\pi_{X_2}(x) > 0$ denote two PDFs on $\mathcal{X} = \mathbb{R}$. We define the associated *cumulative distribution functions* by

$$F_{X_1}(x) = \int_{-\infty}^{x} \pi_{X_1}(x')\mathrm{d}x' , \quad F_{X_2}(x) = \int_{-\infty}^{x} \pi_{X_2}(x')\mathrm{d}x' .$$

Since F_{X_2} is monotonically increasing, it has a unique inverse $F_{X_2}^{-1}(p)$ for $p \in [0,1]$. The inverse may be used to define a transport map that transforms X_1 into X_2 as follows,

$$X_2 = T(X_1) = F_{X_2}^{-1}\left(F_{X_1}(X_1)\right) .$$

For example, consider the case where X_1 is a random variable with uniform distribution $\mathrm{U}([0,1])$, and X_2 is a random variable with standard normal distribution $\mathrm{N}(0,1)$. Then, the transport map between X_1 and X_2 is simply the inverse of the cumulative distribution function

$$F_{X_2}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(x')^2/2}\mathrm{d}x' ,$$

which provides a standard tool for converting uniformly distributed random numbers to normally distributed ones.

We now extend this transform method to random variables in $\mathbb{R}^N$ with $N = 2$.

Example 3.15 (Knothe–Rosenblatt rearrangement). Let $\pi_{X_1}(x^1, x^2)$ and $\pi_{X_2}(x^1, x^2)$ denote two PDFs on $x = (x^1, x^2) \in \mathbb{R}^2$. A transport map between π_{X_1} and π_{X_2} can be constructed in the following manner. We first find the two one-dimensional marginals $\pi_{X_1^1}(x^1)$ and $\pi_{X_2^1}(x^1)$ of the two PDFs. In the previous example, we have seen how to construct a transport map $X_2^1 = T_1(X_1^1)$ which couples these two one-dimensional marginal PDFs. Here, X_i^1 denotes the first component of the random variables X_i, $i = 1, 2$. Next, we write

$$\pi_{X_1}\left(x^1, x^2\right) = \pi_{X_1^2}\left(x^2|x^1\right)\pi_{X_1^1}\left(x^1\right) , \quad \pi_{X_2}\left(x^1, x^2\right) = \pi_{X_2^2}\left(x^2|x^1\right)\pi_{X_2^1}\left(x^1\right)$$

and find a transport map $X_2^2 = T_2(X_1^1, X_1^2)$ by considering one-dimensional couplings between $\pi_{X_1^2}(x^2|x^1)$ and $\pi_{X_2^2}(x^2|T(x^1))$ with x^1 fixed. The associated joint distribution is given by

$$\pi_Z\left(x_1^1, x_1^2, x_2^1, x_2^2\right) = \delta\left(x_2^1 - T_1\left(x_1^1\right)\right)\delta\left(x_2^2 - T_2\left(x_1^1, x_1^2\right)\right)\pi_{X_1}\left(x_1^1, x_1^2\right) .$$

This is called the *Knothe–Rosenblatt rearrangement*, also well known to statisticians under the name of *conditional quantile transforms*. It can be extended to $\mathbb{R}^N$, $N \geq 3$ in the obvious way by introducing the conditional PDFs

$$\pi_{X_1^3}\left(x^3|x^1, x^2\right) , \quad \pi_{X_2^3}\left(x^3|x^1, x^2\right) ,$$

and by constructing an appropriate map $X_2^3 = T_3(X_1^1, X_1^2, X_1^3)$ from those conditional PDFs for fixed pairs (x_1^1, x_1^2) and $(x_2^1, x_2^2) = (T_1(x_1^1), T_2(x_1^1, x_1^2))$ etc. While the Knothe–Rosenblatt rearrangement can be used in quite general situations, it has the undesirable property that the map depends on the choice of ordering of the variables, i.e. in two dimensions a different map is obtained if one instead first couples the x^2 components.

Example 3.16 (Affine transport maps for Gaussian distributions). Consider two Gaussian distributions $\mathrm{N}(\overline{x}_1, \Sigma_1)$ and $\mathrm{N}(\overline{x}_2, \Sigma_2)$ in $\mathbb{R}^N$ with means $\overline{x}_1$ and $\overline{x}_2$ and covariance matrices Σ_1 and Σ_2, respectively. We first define the *square root* $\Sigma^{1/2}$ of a symmetric positive definite matrix Σ as the unique symmetric positive definite matrix which satisfies $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$. Then, the affine transformation

$$x_2 = T(x_1) = \overline{x}_2 + \Sigma_2^{1/2}\Sigma_1^{-1/2}(x_1 - \overline{x}_1) \tag{3.5}$$

provides a deterministic coupling. Indeed, we find that

$$(x_2 - \overline{x}_2)^{\mathrm{T}}\Sigma_2^{-1}(x_2 - \overline{x}_2) = (x_1 - \overline{x}_1)^{\mathrm{T}}\Sigma_1^{-1}(x_1 - \overline{x}_1)$$

under the suggested coupling. The proposed coupling is, of course, not unique since

$$x_2 = T(x_1) = \overline{x}_2 + \Sigma_2^{1/2} Q \Sigma_1^{-1/2}(x_1 - \overline{x}_1),$$

where Q is an orthogonal matrix, and also provides a coupling. We will see in Section 3.3 that a coupling between Gaussian random variables is also at the heart of the ensemble square root filter formulations of sequential data assimilation.

Deterministic couplings can be viewed as a special case of a *Markov process* $\{X_n\}_{n\in\{1,2\}}$ defined by

$$\pi_{X_2}(x_2) = \int_{\mathcal{X}_1} \pi(x_2|x_1)\pi_{X_1}(x_1)\mathrm{d}x_1,$$

where $\pi(x_2|x_1)$ denotes an appropriate conditional PDF for the random variable X_2 given $X_1 = x_1$. Indeed, we simply have

$$\pi(x_2|x_1) = \delta(x_2 - T(x_1))$$

for deterministic couplings. We will come back to Markov processes in Section 2.

The trivial coupling $\pi_Z(x_1, x_2) = \pi_{X_1}(x_1)\pi_{X_2}(x_2)$ leads to a zero correlation between the induced random variables X_1 and X_2 since their covariance is

$$\mathrm{cov}(X_1, X_2) = \mathbb{E}_Z\left[(x_1 - \overline{x}_1)(x_2 - \overline{x}_2)^{\mathrm{T}}\right] = \mathbb{E}_Z\left[x_1 x_2^{\mathrm{T}}\right] - \overline{x}_1\overline{x}_2^{\mathrm{T}} = 0,$$

where $\overline{x}_i = \mathbb{E}_{X_i}[x]$. A transport map leads instead to the covariance matrix

$$\mathrm{cov}(X_1, X_2) = \mathbb{E}_Z\left[x_1 x_2^{\mathrm{T}}\right] - \mathbb{E}_{X_1}[x_1]\left(\mathbb{E}_{X_2}[x_2]\right)^{\mathrm{T}} = \mathbb{E}_{X_1}\left[x_1 T(x_1)^{\mathrm{T}}\right] - \overline{x}_1\overline{x}_2^{\mathrm{T}},$$

which is nonzero in general. If several transport maps exist, then one could choose the one that maximizes the covariance. Now consider, for example, univariate random variables X_1 and X_2. Maximizing their covariance for given marginal PDFs has an important geometric interpretation: it is equivalent to minimizing the mean square distance between x_1 and $T(x_1) = x_2$ given by

$$\begin{aligned}\mathbb{E}_Z\left[|x_2 - x_1|^2\right] &= \mathbb{E}_{X_1}\left[|x_1|^2\right] + \mathbb{E}_{X_2}\left[|x_2|^2\right] - 2\mathbb{E}_Z\left[x_1x_2\right] \\ &= \mathbb{E}_{X_1}\left[|x_1|^2\right] + \mathbb{E}_{X_2}\left[|x_2|^2\right] - 2\mathbb{E}_Z\left[(x_1 - \overline{x}_1)(x_2 - \overline{x}_2)\right] - 2\overline{x}_1\overline{x}_2 \\ &= \mathbb{E}_{X_1}\left[|x_1|^2\right] + \mathbb{E}_{X_2}\left[|x_2|^2\right] - 2\overline{x}_1\overline{x}_2 - 2\mathrm{cov}(X_1, X_2)\,.\end{aligned}$$

Hence, finding a joint measure μ_Z that minimizes the expectation of $(x_1 - x_2)^2$ simultaneously maximizes the covariance between X_1 and X_2. This geometric interpretation leads to the celebrated Monge–Kantorovitch problem.

Definition 3.17 (Monge–Kantorovitch problem). A transference plan $\mu_Z^* \in \Pi(\mu_{X_1}, \mu_{X_2})$ is called the solution to the *Monge–Kantorovitch problem* with cost function $c(x_1, x_2) = \|x_1 - x_2\|^2$ if

$$\mu_Z^* = \arg\inf_{\mu_Z \in \Pi(\mu_{X_1}, \mu_{X_2})} \mathbb{E}_Z\left[\|x_1 - x_2\|^2\right]. \tag{3.6}$$

The associated function $W(\mu_{X_1}, \mu_{X_2})$, defined by

$$W(\mu_{X_1}, \mu_{X_2})^2 = \mathbb{E}_Z\left[\|x_1 - x_2\|^2\right], \quad \mathrm{law}(Z) = \mu_Z^*\,,$$

is called the L^2-Wasserstein distance of μ_{X_1} and μ_{X_2}.

Theorem 3.18 (Optimal transference plan). *If the measures μ_{X_i}, $i = 1, 2$, are absolutely continuous, then the optimal transference plan that solves the Monge–Kantorovitch problem corresponds to a deterministic coupling with transfer map*

$$X_2 = T(X_1) = \nabla_x \psi(X_1)$$

for some convex potential $\psi : \mathbb{R}^N \to \mathbb{R}$.

Proof. We only demonstrate that the solution to the Monge–Kantorovitch problem is of the desired form when the infimum in (3.6) is restricted to deterministic couplings. See [33, 53] for a complete proof based on approximative couplings using linear programming, the geometric concept of cyclical monotonicity of the support of an optimal coupling, and Rockafellar's theorem.

We denote the associated PDFs by π_{X_i}, $i = 1, 2$. We also introduce the inverse transfer map $X_1 = S(X_2) = T^{-1}(X_2)$ and consider the functional

$$\begin{aligned}\mathcal{L}[S, \Psi] = \frac{1}{2}\int_{\mathbb{R}^N} &\|S(x) - x\|^2 \pi_{X_2}(x)\mathrm{d}x \\ &+ \int_{\mathbb{R}^N} \left[\Psi\left(S(x)\right)\pi_{X_2}(x) - \Psi(x)\pi_{X_1}(x)\right]\mathrm{d}x\end{aligned}$$

in S and a potential $\Psi : \mathbb{R}^N \to \mathbb{R}$. We note that

$$\int\limits_{\mathbb{R}^N} \left[\Psi\left(S(x)\right)\pi_{X_2}(x) - \Psi(x)\pi_{X_1}(x)\right] \mathrm{d}x$$

$$= \int\limits_{\mathbb{R}^N} \Psi(x)\left[\pi_{X_2}\left(T(x)\right)|DT(x)| - \pi_{X_1}(x)\right] \mathrm{d}x$$

by a simple change of variables. Here, $|DT(x)|$ denotes the determinant of the Jacobian matrix of T at x and the potential Ψ can be interpreted as a Lagrange multiplier enforcing the coupling of the two marginal PDFs under the desired transport map.

Taking variational derivatives with respect to S and Ψ, we obtain

$$\frac{\delta \mathcal{L}}{\delta S} = \pi_{X_2}(x)\left[(S(x) - x) + \nabla_x \Psi\left(S(x)\right)\right] = 0$$

and

$$\frac{\delta \mathcal{L}}{\delta \Psi} = -\pi_{X_1}(x) + \pi_{X_2}\left(T(x)\right)|DT(x)| = 0, \tag{3.7}$$

characterizing critical points of the functional $\mathcal{L}$. The first equality implies

$$x_2 = x_1 + \nabla_x \Psi\left(x_1\right) = \nabla_x\left(\frac{1}{2}x_1^{\mathrm{T}} x_1 + \Psi\left(x_1\right)\right) =: \nabla_x \psi\left(x_1\right)$$

and the second recovers our *Ansatz* that T transforms π_{X_1} into π_{X_2} as a result of the Lagrange multiplier Ψ. □

Example 3.19 (Optimal transport maps for Gaussian distributions). Consider two Gaussian distributions $\mathrm{N}(\overline{x}_1, \Sigma_1)$ and $\mathrm{N}(\overline{x}_2, \Sigma_2)$ in $\mathbb{R}^N$ with means $\overline{x}_1$ and $\overline{x}_2$, and covariance matrices Σ_1 and Σ_2, respectively. We had previously discussed the deterministic coupling (3.5). However, the induced affine transformation $x_2 = T(x_1)$ cannot not be generated from a potential ψ since the matrix $\Sigma_2^{1/2}\Sigma_1^{-1/2}$ is not symmetric. Indeed the optimal coupling in the sense of Monge–Kantorovitch with cost function $c(x_1, x_2) = \|x_1 - x_2\|^2$ is provided by

$$x_2 = T(x_1) := \overline{x}_2 + \Sigma_2^{1/2}\left[\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2}\right]^{-1/2}\Sigma_2^{1/2}\left(x_1 - \overline{x}_1\right). \tag{3.8}$$

See [41] for a derivation. The following generalization will be used in Section 3.3. Assume that a matrix $A \in \mathbb{R}^{N\times M}$ is given such that $\Sigma_2 = A\,A^{\mathrm{T}}$. Clearly, we can chose $A = \Sigma_2^{1/2}$ in which case $M = N$ and A is symmetric. However, we allow for A to be nonsymmetric and M can be different from N. An important observation is that one can replace $\Sigma_2^{1/2}$ in (3.8) by A and A^{T}, respectively, i.e.

$$T(x_1) = \overline{x}_2 + A\left[A^{\mathrm{T}}\Sigma_1 A\right]^{-1/2} A^{\mathrm{T}}\left(x_1 - \overline{x}_1\right). \tag{3.9}$$

While optimal couplings are of broad theoretical and practical interest, their computational implementation can be very demanding. In Section 3, we will discuss an embedding method originally due to Jürgen Moser [38], which leads to a generally nonoptimal, but computationally more tractable formulation in the context of Bayesian statistics and data assimilation. Alternatively, we may replace the coupling problem by an appropriate finite-dimensional linear programming problem [46].

1.4 Monte Carlo methods

Monte Carlo methods, also called particle or ensemble methods depending on the context in which they are being used, can be used to approximate statistics, namely, expectation values $\mathbb{E}_X[f]$, for a random variable X. We begin by discussing the special case $f(x) = x$, namely, the mean.

Definition 3.20 (Empirical mean). Given a sequence X_i, $i = 1, \dots, M$, of independent random variables with identical measure μ_X, the *empirical mean* is

$$\overline{x}_M = \frac{1}{M}\sum_{i=1}^{M} X_i(\omega) = \frac{1}{M}\sum_{i=1}^{M} x_i$$

with samples $x_i = X_i(\omega)$.

Of course, $\overline{x}_M$ itself is the realization of a random variable $\overline{X}_M$ and we consider the *mean squared error* (MSE)

$$\begin{aligned}\mathrm{MSE}(\overline{x}) &= \mathbb{E}_{\overline{X}_M}[(\overline{x}_M - \overline{x})^2] \\ &= (\mathbb{E}_{\overline{X}_M}[\overline{x}_M] - \overline{x})^2 + \mathbb{E}_{\overline{X}_M}\left[(\overline{x}_M - \mathbb{E}_{\overline{X}_M}[\overline{x}_M])^2\right]\end{aligned} \tag{3.10}$$

with respect to the exact mean value $\overline{x} = \mathbb{E}_X[x]$. We have broken down the MSE into two components: squared bias and variance. Such a decomposition is possible for any estimator and is known as the *bias-variance decomposition*. The particular estimator $\overline{X}_M$ is called *unbiased* since $\mathbb{E}_{\overline{X}_M}[\overline{x}_M] = \overline{x}$ for any $M \geq 1$. Furthermore, $\overline{X}_M$ converges weakly to $\overline{x}$ under the central limit theorem provided μ_X has finite second-order moments, i.e.

$$\lim_{M\to\infty} \mathbb{E}_{\overline{X}_M}\left[(\overline{x}_M - \mathbb{E}_{\overline{X}_M}[\overline{x}_M])^2\right] = 0\,.$$

It remains to generate samples $x_i = X_i(\omega)$ from the required distribution. Methods to do this include the von Neumann rejection method and Markov chain Monte Carlo methods, which we will briefly discuss in Section 2. Often, the prior distribution is assumed to be Gaussian, in which case explicit random number generators are available. We now turn to the situation where samples from the prior distribution are available, and are to be used to approximate the mean of the posterior distribution (or any other expectation value).

Importance sampling is a classical method to approximate expectation values of a random variable $X^t \sim \pi_{X^t}$ using samples from a random variable $X^p \sim \pi_{X^p}$, which requires that the target PDF π_{X^t} is absolutely continuous with respect to proposal PDF π_{X^p}. This is the case for the prior and posterior PDFs from Bayes' formula (3.4), i.e. we set the proposal distribution $\pi_{X^p}(x)$ equal to the prior distribution $\pi_X(x)$ and the posterior distribution $\pi_X(x|y_0) \propto \pi_Y(y_0|x)\pi_X(x)$ becomes the target distribution $\pi_{X^t}(x)$.

Definition 3.21 (Importance sampling for Bayesian estimation). Let x_i^{prior}, $i = 1, \dots, M$, denote samples from the prior PDF $\pi_X(x)$, then the *importance sampler* estimate of the mean of the posterior $\pi_X(x|y_0)$ is

$$\overline{x}_M^{\text{post}} = \sum_{i=1}^{M} w_i x_i^{\text{prior}} \tag{3.11}$$

with *importance weights*

$$w_i = \frac{\pi_Y\left(y_0|x_i^{\text{prior}}\right)}{\sum_{i=1}^{M} \pi_Y\left(y_0|x_i^{\text{prior}}\right)} . \tag{3.12}$$

Importance sampling becomes statistically inefficient when the weights have largely varying magnitude which becomes particularly significant for high-dimensional problems. To demonstrate this effect, consider a uniform prior on the unit hypercube $V = [0,1]^N$. Each of the M samples x_i from this prior formally represent a hypercube with volume $1/M$. However, the likelihood measures the distance of a sample x_i to the observation y_0 in the Euclidean distance and the volume of a hypersphere decreases rapidly relative to that of an associated hypercube as N increases. Within the framework of the bias-variance decomposition of a mean squared error, for example, (3.10), the curse of dimensionality manifests itself in large variances for finite M.

To counteract this curse of dimensionality, one may utilize the concept of coupling. In other words, assume that we have a transport map $x^{\text{post}} = T(x^{\text{prior}})$ which couples the prior and posterior distributions. Then, with transformed samples $x_i^{\text{post}} = T(x_i^{\text{prior}})$, $i = 1, \dots, M$, we obtain the estimator

$$\overline{x}_M^{\text{post}} = \sum_{i=1}^{M} \hat{w}_i x_i^{\text{post}}$$

with equal weights $\hat{w}_i = 1/M$.

Sometimes, one cannot couple the prior and posterior distribution directly, or the coupling is too expensive computationally. Then, one can attempt to find a coupling between the prior PDF $\pi_X(x)$ and an approximation $\tilde{\pi}_X(x|y_0)$ to the posterior PDF $\pi_X(x|y_0) \propto \pi_Y(y_0|x)\pi_X(x)$. Given an associated transport map $X^{\text{prop}} = \tilde{T}(X^{\text{prior}})$, i.e.

$$\tilde{\pi}_X(\tilde{T}(x)|y_0) = \pi_X(x)|D\tilde{T}(x)|^{-1} ,$$

one then takes $\tilde{\pi}_X(x|y_0)$ as the proposal density $\pi_{X^p}(x)$ in an importance sampler with realizations $x_i^{\rm prop}$, $i = 1, \dots, M$, defined by

$$x_i^{\rm prop} = \tilde{T}\left(x_i^{\rm prior}\right) .$$

An asymptotically unbiased estimator for the posterior mean is now provided by

$$\overline{x}_M^{\rm post} = \sum_{i=1}^M \tilde{w}_i x_i^{\rm prop} \tag{3.13}$$

with weights

$$\tilde{w}_i \propto \frac{\pi_Y\left(y_0|x_i^{\rm prop}\right)\pi_X\left(x_i^{\rm prop}\right)}{\tilde{\pi}_X\left(x_i^{\rm prop}|y_0\right)} = \pi_Y\left(y_0|x_i^{\rm prop}\right)\left|D\tilde{T}\left(x_i^{\rm prior}\right)\right| \frac{\pi_X\left(x_i^{\rm prop}\right)}{\pi_X\left(x_i^{\rm prior}\right)} , \tag{3.14}$$

$i = 1, \dots, M$. The constant of proportionality is chosen such that $\sum_{i=1}^M \tilde{w}_i = 1$. Indeed, if $\pi_{X^p}(x) = \tilde{\pi}_X(x|y_0) = \pi_X(x|y_0)$, we recover the case of equal weights $\tilde{w}_i = 1/M$, and $\pi_{X^p}(x) = \tilde{\pi}_X(x|y_0) = \pi_X(x)$ leads to standard importance sampling using prior samples, i.e. $x_i^{\rm prop} = x_i^{\rm prior}$.

We will return to the subject of sampling from the posterior distribution in Sections 2.3 and 3.2.

References

An excellent introduction to many topics covered in this survey is [22]. Bayesian inference and a Bayesian perspective on inverse problems are discussed in [24, 31, 39]. The monographs [52, 53] provide an in depth introduction to optimal transportation and coupling of random variables. Monte Carlo methods are covered in [32]. We also point to [20] for a discussion of estimation and regression methods from a bias-variance perspective. A discussion of infinite-dimensional Bayesian inference problems can be found in [51].

2 Stochastic processes

In this section, we collect basic results concerning stochastic processes.

Definition 3.22 (Stochastic process). Let T be a set of indices. A *stochastic process* is a family $\{X_t\}_{t\in T}$ of random variables on a common space $\mathcal{X}$, i.e. $X_t(\omega) \in \mathcal{X}$.

In the context of dynamical systems, the variable t corresponds to time. We distinguish between continuous time $t \in [0, t_{\rm end}] \subset \mathbb{R}$ or discrete time $t_n = n\Delta t$, $n \in \{0, 1, 2, \dots\} = T$, with $\Delta t > 0$ a time-increment. In cases where subscript indices can be confusing, we will also use the notations $X(t)$ and $X(t_n)$, respectively.

A stochastic process can be seen as a function of two arguments: t and ω. For fixed ω, $X_t(\omega)$ becomes a function of $t \in T$, which we call a realization or trajectory of the stochastic process. We will restrict ourselves to the case where $X_t(\omega)$ is continuous in t (with probability 1) in the case of a continuous time. Alternatively, one can fix the time $t \in T$ and consider the random variable $X_t(\cdot)$ and its distribution. More generally, one can consider l-tuples $(t_1, t_2, \dots, t_l)$ and associated l-tuples of random variables $(X_{t_1}(\cdot), X_{t_2}(\cdot), \dots, X_{t_l}(\cdot))$ and their joint distributions. This leads to concepts such as temporal correlation.

2.1 Discrete time Markov processes

First, we develop the concept of Markov processes for discrete time processes.

Definition 3.23 (Discrete time Markov processes). The discrete time stochastic process $\{X_n\}_{n\in T}$ with $\mathcal{X} = \mathbb{R}^N$ and $T = \{0, 1, 2, \dots)$ is called a (time-independent) *Markov process* with transition kernel $\pi(x'|x)$ if its joint PDFs can be written as

$$\pi_n\left(x_0, x_1, \dots, x_n\right) = \pi\left(x_n|x_{n-1}\right)\pi\left(x_{n-1}|x_{n-2}\right)\cdots\pi\left(x_1|x_0\right)\pi_0\left(x_0\right)$$

for all $n \in \{0, 1, 2, \dots\} = T$. The associated marginal distributions $\pi_n = \pi_{X_n}$ satisfy the *Chapman–Kolmogorov equation*

$$\pi_{n+1}\left(x'\right) = \int_{\mathbb{R}^N} \pi\left(x'|x\right)\pi_n\left(x\right)\mathrm{d}x \tag{3.15}$$

and the process can be recursively repeated to yield a family of marginal distributions $\{\pi_n\}_{n\in T}$ for given π_0. This family can also be characterized by the linear *Frobenius–Perron operator*

$$\pi_{n+1} = \mathcal{P}\pi_n\,, \tag{3.16}$$

which is induced by (3.15).

The above definition is equivalent to the more traditional definition that a process is Markov if the conditional distributions satisfy

$$\pi_n\left(x_n|x_0, x_1, \dots, x_{n-1}\right) = \pi\left(x_n|x_{n-1}\right)\,.$$

Note that, contrary to Bayes' formula (3.4), which directly yields marginal distributions, the Chapman–Kolmogorov equation (3.15) starts from a given coupling

$$\pi_{X_{n+1}X_n}\left(x_{n+1}, x_n\right) = \pi\left(x_{n+1}|x_n\right)\pi_{X_n}\left(x_n\right)$$

followed by marginalization to derive $\pi_{X_{n+1}}(x_{n+1})$. A Markov process is called time-dependent if the conditional PDF $\pi(x'|x)$ depends on t_n. While we have considered time-independent processes in this section, we will see in Section 3 that the idea of coupling applied to Bayes' formula leads to time-dependent Markov processes.

2.2 Stochastic difference and differential equations

We start from the *stochastic difference equation*

$$X_{n+1} = X_n + \Delta t f(X_n) + \sqrt{2\Delta t} Z_n, \quad t_{n+1} = t_n + \Delta t\,, \tag{3.17}$$

where $\Delta t > 0$ is a small parameter (the step-size), f is a given (Lipschitz continuous) function, and $Z_n \sim \mathrm{N}(0, Q)$ are independent and identically distributed random variables with correlation matrix Q.

The time evolution of the associated marginal densities π_{X_n} is governed by the *Chapman–Kolmogorov equation* with conditional PDF

$$\begin{aligned} \pi(x'|x) = {} & \frac{1}{(4\pi\Delta t)^{N/2}|Q|^{1/2}} \\ & \times \exp\left(-\frac{1}{4\Delta t}\left(x' - x - \Delta t f(x)\right)^{\mathrm{T}} Q^{-1}\left(x' - x - \Delta t f(x)\right)\right). \end{aligned} \tag{3.18}$$

Proposition 3.24 (Stochastic differential and Fokker–Planck equation). *Taking the limit $\Delta t \to 0$, one obtains the stochastic differential equation (SDE)*

$$\mathrm{d}X_t = f(X_t)\,\mathrm{d}t + \sqrt{2}Q^{1/2}\mathrm{d}W_t \tag{3.19}$$

for X_t, where $\{W_t\}_{t\geq 0}$ denotes standard N-dimensional Brownian motion, and the Fokker–Planck equation

$$\frac{\partial \pi_X}{\partial t} = -\nabla_x \cdot (\pi_X f) + \nabla_x \cdot (Q\nabla_x \pi_X) \tag{3.20}$$

for the marginal density $\pi_X(x, t)$. Note that $Q = 0$ (no noise) leads to the Liouville, transport or continuity equation

$$\frac{\partial \pi_X}{\partial t} = -\nabla_x \cdot (\pi_X f)\,, \tag{3.21}$$

which implies that we may interpret f as a given velocity field in the sense of fluid mechanics.

Proof. The difference equation (3.17) is called the Euler–Maruyama method for approximating the SDE (3.19). See [21, 26] for a discussion on the convergence of (3.17) to (3.19) as $\Delta t \to 0$.

The Fokker–Planck equation (3.20) is the linear combination of a drift and a diffusion term. To simplify the discussion, we derive both terms separately from (3.17) by first considering $f = 0, Q \neq 0$ and then $Q = 0, f \neq 0$. To simplify the derivation of the diffusion term even further, we also assume $x \in \mathbb{R}$ and $Q = 1$. In other words, we show that scalar Brownian motion

$$\mathrm{d}X_t = \sqrt{2}\mathrm{d}W_t$$

leads to the heat equation

$$\frac{\partial \pi_X}{\partial t} = \frac{\partial^2 \pi_X}{\partial x^2} \,.$$

We first note that the conditional PDF (3.18) reduces to

$$\pi(x'|x) = (4\pi\Delta t)^{-1/2} \exp\left(-\frac{(x'-x)^2}{4\Delta t}\right)$$

under $f(x) = 0, Q = 1, N = 1$, and the Chapman–Kolmogorov equation (3.15) becomes

$$\pi_{n+1}(x') = \int_{\mathbb{R}} \frac{1}{\sqrt{4\pi\Delta t}} e^{-y^2/(4\Delta t)} \pi_n(x'+y)\mathrm{d}y \tag{3.22}$$

under the variable substitution $y = x - x'$. We now expand $\pi_n(x'+y)$ in y about $y = 0$, i.e.

$$\pi_n\left(x'+y\right) = \pi_n\left(x'\right) + y\frac{\partial \pi_n}{\partial x}\left(x'\right) + \frac{y^2}{2}\frac{\partial^2 \pi_n}{\partial x^2}\left(x'\right) + \cdots \,,$$

and substitute the expansion into (3.22):

$$\begin{aligned} \pi_{n+1}\left(x'\right) = &\int_{\mathbb{R}} \frac{1}{\sqrt{4\pi\Delta t}} e^{-y^2/(4\Delta t)} \pi_n\left(x'\right)\mathrm{d}y \\ &+ \int_{\mathbb{R}} \frac{1}{\sqrt{4\pi\Delta t}} e^{-y^2/(4\Delta t)} y \frac{\partial \pi_n}{\partial x}\left(x'\right)\mathrm{d}y \\ &+ \int_{\mathbb{R}} \frac{1}{\sqrt{4\pi\Delta t}} e^{-y^2/(4\Delta t)} \frac{y^2}{2}\frac{\partial^2 \pi_n}{\partial x^2}\left(x'\right)\mathrm{d}y + \cdots \,. \end{aligned}$$

The integrals correspond to the zeroth, first and second-order moments of the Gaussian distribution with mean zero and variance $2\Delta t$. Hence,

$$\pi_{n+1}\left(x'\right) = \pi_n\left(x'\right) + \Delta t \frac{\partial^2 \pi_n}{\partial x^2}\left(x'\right) + \cdots$$

and it can also easily be shown that the neglected higher-order terms contribute with $\mathcal{O}(\Delta t^2)$ terms. Therefore,

$$\frac{\pi_{n+1}\left(x'\right) - \pi_n\left(x'\right)}{\Delta t} = \frac{\partial^2 \pi_n}{\partial x^2}\left(x'\right) + \mathcal{O}\left(\Delta t\right) \,,$$

and the heat equation is obtained upon taking the limit $\Delta t \to 0$. The nonvanishing drift case, i.e. $f(x) \neq 0$, while being more technical, can be treated in the same manner.

One can also use (3.7) to derive Liouville's equation (3.21) directly. We set

$$T(x) = x + \Delta t f(x)$$

and note that

$$|DT(x)| = 1 + \Delta t \nabla_x \cdot f + \mathcal{O}(\Delta t^2)\,.$$

Hence, (3.7) implies

$$\pi_{X_1} = \pi_{X_2} + \Delta t \pi_{X_2} \nabla_x \cdot f + \Delta t \left(\nabla_x \pi_{X_2}\right) \cdot f + \mathcal{O}(\Delta t^2)$$

and

$$\frac{\pi_{X_2} - \pi_{X_1}}{\Delta t} = -\nabla_x \cdot (\pi_{X_2} f) + \mathcal{O}\left(\Delta t\right)\,.$$

Taking the limit $\Delta t \to 0$, we obtain (3.21). □

Following the work of Felix Otto (see, e.g. [42, 52]), we note that in the case of pure diffusion, i.e. $f = 0$, the Fokker–Planck equation can be rewritten as a gradient flow system. We first introduce some notation.

Definition 3.25 (differential geometric structure on manifold of probability densities). We formally introduce the *manifold of all PDFs on* $\mathcal{X} = \mathbb{R}^N$

$$\mathcal{M} = \left\{ \pi : \mathbb{R}^N \to \mathbb{R} : \pi(x) \geq 0,\ \int_{\mathbb{R}^N} \pi(x) \mathrm{d}x = 1 \right\}$$

with *tangent space*

$$T_\pi \mathcal{M} = \left\{ \phi : \mathbb{R}^N \to \mathbb{R} : \int_{\mathbb{R}^N} \phi(x) \mathrm{d}x = 0 \right\}.$$

The *variational derivative* of a functional $F : \mathcal{M} \to \mathbb{R}$ is defined as

$$\int_{\mathbb{R}^N} \frac{\delta F}{\delta \pi} \phi \, \mathrm{d}x = \lim_{\epsilon \to 0} \frac{F\left(\pi + \epsilon \phi\right) - F(\pi)}{\epsilon}$$

where ϕ is a function such that $\int_{\mathbb{R}^N} \phi \mathrm{d}x = 0$, i.e. $\phi \in T_\pi \mathcal{M}$.

Consider the potential

$$V\left(\pi_X\right) = \int_{\mathbb{R}^N} \pi_X \ln \pi_X \mathrm{d}x\,, \tag{3.23}$$

which has the functional derivative

$$\frac{\delta V}{\delta \pi_X} = \ln \pi_X$$

since

$$\begin{aligned} V\left(\pi_X + \epsilon \phi\right) &= V\left(\pi_X\right) + \epsilon \int_{\mathbb{R}^N} \left(\phi \ln \pi_X + \phi\right) \mathrm{d}x + \mathcal{O}(\epsilon^2) \\ &= V\left(\pi_X\right) + \epsilon \int_{\mathbb{R}^N} \phi \ln \pi_X \, \mathrm{d}x + \mathcal{O}(\epsilon^2)\,. \end{aligned}$$

Hence, we find that the diffusion part of the Fokker–Planck equation is equivalent to

$$\frac{\partial \pi_X}{\partial t} = \nabla_x \cdot (Q \nabla_x \pi_X) = \nabla_x \cdot \left\{ \pi_X Q \nabla_x \frac{\delta V}{\delta \pi_X} \right\} . \tag{3.24}$$

This formulation allows us to treat diffusion in form of a vector field

$$v(x,t) = -Q \nabla_x \frac{\delta V}{\delta \pi_X}$$

which, contrary to vector fields arising from the theory of ordinary differential equations, depends on the PDF π_X. See the following Section 2.3 for an application.

Proposition 3.26 (Gradient on the manifold of probability densities). *Let g_π be a metric tensor defined on $T_\pi \mathcal{M}$ as*

$$g_\pi(\phi_1, \phi_2) = \int_{\mathbb{R}^N} (\nabla_x \psi_1) \cdot (\mathrm{M} \nabla_x \psi_2)\, \pi \mathrm{d}x$$

with potentials ψ_i, $i = 1, 2$, determined by the elliptic partial differential equation (PDE)

$$-\nabla_x \cdot (\pi \mathrm{M} \nabla_x \psi_i) = \phi_i ,$$

where $\mathrm{M} \in \mathbb{R}^{N \times N}$ is a symmetric, positive-definite matrix.

Then, the gradient of a potential $F(\pi)$ under g_π satisfies

$$\mathrm{grad}_\pi F(\pi) = -\nabla_x \cdot \left(\pi \mathrm{M} \nabla_x \frac{\delta F}{\delta \pi} \right) . \tag{3.25}$$

Proof. Given the metric tensor g_π, the gradient is defined by

$$g_\pi(\mathrm{grad}_\pi F(\pi), \phi) = \int_{\mathbb{R}^N} \frac{\delta F}{\delta \pi} \phi \mathrm{d}x \tag{3.26}$$

for all $\phi \in T_\pi \mathcal{M}$. Since any element $\phi \in T_\pi \mathcal{M}$ can be written in the form

$$\phi = -\nabla_x \cdot (\pi \mathrm{M} \nabla_x \psi)$$

with suitable potential ψ, a potential $\hat{\psi}$ exists such that

$$\mathrm{grad}_\pi F(\pi) = -\nabla_x \cdot \left(\pi \mathrm{M} \nabla_x \hat{\psi} \right) \in T_\pi \mathcal{M}$$

and we need to demonstrate that

$$\hat{\psi} = \frac{\delta F}{\delta \pi}$$

is consistent with (3.26). Indeed, we find that

$$\begin{aligned}\int_{\mathbb{R}^N} \frac{\delta F}{\delta \pi}\phi \mathrm{d}x &= -\int_{\mathbb{R}^N} \frac{\delta F}{\delta \pi}\nabla_x \cdot (\pi \mathrm{M}\nabla_x \psi)\,\mathrm{d}x \\ &= \int_{\mathbb{R}^N} \pi \nabla_x \frac{\delta F}{\delta \pi} \cdot (\mathrm{M}\nabla_x \psi)\,\mathrm{d}x \\ &= \int_{\mathbb{R}^N} \left(\nabla_x \hat{\psi}\right) \cdot (\mathrm{M}\nabla_x \psi)\, \pi \mathrm{d}x \\ &= g_\pi\left(\mathrm{grad} F\left(\pi\right), \phi\right) .\end{aligned}$$ □

It follows that the diffusion part of the Fokker–Planck equation can be viewed as a gradient flow on the manifold $\mathcal{M}$. More precisely, set $F(\pi) = V(\pi_X)$ and $\mathrm{M} = Q$ to reformulate (3.24) as a gradient flow

$$\frac{\partial \pi_X}{\partial t} = -\mathrm{grad}_{\pi_X} V\left(\pi_X\right)$$

with potential (3.23). We will find in Section 3 that related geometric structures arise from Bayes' formula in the context of filtering. We finally note that

$$\frac{\mathrm{d}V}{\mathrm{d}t} = \int_{\mathbb{R}^N} \frac{\delta V}{\delta \pi_X}\frac{\partial \pi_X}{\partial t}\mathrm{d}x = -\int_{\mathbb{R}^N} \left(\nabla_x \frac{\delta V}{\delta \pi_X}\right) \cdot \left(Q\nabla_x \frac{\delta V}{\delta \pi_X}\right) \pi_X \mathrm{d}x \le 0 .$$

2.3 Ensemble prediction and sampling methods

In this section, we extend the Monte Carlo method from Section 1.4 to the approximation of the marginal PDFs $\pi_X(x,t)$, $t \ge 0$, evolving under the SDE model (3.19). Assume that we have a set of independent samples $x_i(0)$, $i = 1,\dots,M$, from the initial PDF $\pi_X(x,0)$.

Definition 3.27 (ensemble prediction). A Monte Carlo approximation to the time-evolved marginal PDFs $\pi_X(x,t)$ can be obtained from solving the SDEs

$$\mathrm{d}x_i = f\left(x_i\right)\mathrm{d}t + \sqrt{2}Q^{1/2}\mathrm{d}W_i(t) \tag{3.27}$$

for $i = 1,\dots,M$, where $\{W_i(t)\}_{i=1}^M$ denote realizations of independent standard N-dimensional Brownian motion and the initial conditions $\{x_i(0)\}_{i=1}^M$ are realizations of the initial PDF $\pi_X(x,0)$. This approximation provides an example for a *particle* or *ensemble prediction* method and it can be shown that the estimator

$$\overline{x}_M(t) = \frac{1}{M}\sum_{i=1}^{M} x_i(t) \tag{3.28}$$

provides a consistent and unbiased approximation to the mean $\mathbb{E}_{X_t}[x]$.

Alternatively, using formulation (3.24) of the Fokker–Planck equation (3.20) in the pure diffusion case, we may reformulate the random part in (3.27) and introduce particle equations

$$\begin{aligned}\frac{\mathrm{d}x_i}{\mathrm{d}t} &= f(x_i) - Q\nabla_x \frac{\delta V}{\delta \pi_X}(x_i) \\ &= f(x_i) - \frac{1}{\pi_X(x_i,t)} Q\nabla_x \pi_X(x_i,t)\,,\end{aligned} \tag{3.29}$$

$i = 1, \dots, M$. Contrary to the SDE (3.27), this formulation requires the PDF $\pi_X(x,t)$, which is not explicitly available in general. However, a Gaussian approximation can be obtained from the available ensemble $x_i(t)$, $i = 1, \dots, M$, using

$$\pi_X(x,t) \approx \frac{1}{(2\pi)^{N/2}|P|^{1/2}} \exp\left(-\frac{1}{2}(x - \overline{x}_M(t))^{\mathrm{T}} P(t)^{-1}(x - \overline{x}_M(t))\right)$$

with empirical mean (3.28) and empirical covariance matrix

$$P = \frac{1}{M-1}\sum_{i=1}^{M}(x_i - \overline{x}_M)(x_i - \overline{x}_M)^{\mathrm{T}}\,. \tag{3.30}$$

Substituting this Gaussian approximation into (3.29) yields the ensemble evolution equations

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = f(x_i) + QP^{-1}(x_i - \overline{x}_M)\,, \tag{3.31}$$

which becomes exact in case the vector field f is linear, i.e. $f(x) = Ax + u$, the initial PDF $\pi_X(x,0)$ is Gaussian and for ensemble sizes $M \to \infty$.

We finally discuss the application of a particular type of SDEs (3.19) as a way of generating samples x_i from a given PDF such as the posterior $\pi_X(x|y_0)$ of Bayesian inference. To do this, consider the SDE (3.19) with the vector field f being generated by a potential $U(x)$, i.e. $f(x) = -\nabla_x U(x)$, and $Q = I$. Then, it can easily be verified that the PDF

$$\pi_X^*(x) = Z^{-1}\exp(-U(x))\,, \quad Z = \int_{\mathbb{R}^N} \exp(-U(x))\,\mathrm{d}x$$

is stationary under the associated Fokker–Planck equation (3.20). Indeed,

$$\nabla_x \cdot (\pi_X^* \nabla_X U) + \nabla_x \cdot \nabla_x \pi_X^* = \nabla_x \cdot (\pi_X^* \nabla_x U + \nabla_x \pi_X^*) = 0\,.$$

Furthermore, it can be shown that π_X^* is the unique stationary PDF and that any initial PDF $\pi_X(t = 0)$ approaches π_X^* at an exponential rate under an appropriate assumption on the potential V. Hence, $X_t \sim \pi_X^*$ for $t \to \infty$. This allows us to use an ensemble of solutions $x_i(t)$ of (3.27) with an arbitrary initial PDF $\pi_X(x,0)$ as a method for generating ensembles from the prior or posterior Bayesian PDFs provided $U(x) = -\ln \pi_X(x)$ or $U(x) = -\ln \pi_X(x|y_0)$, respectively. Note that the temporal dynamics of the associated SDE (3.19) is not of any physical significance in this context, but instead the SDE formulation is only taken as a device for generating the desired samples. If the SDE formulation is replaced by the Euler–Maruyama method (3.17), time-stepping errors lead to sampling errors which can be corrected for by combining (3.17) with a Metropolis accept-reject criterion. The Metropolis adjusted method gives rise to particular instances of *Markov chain Monte Carlo (MCMC) methods*, for example, the *Metropolis adjusted Langevin algorithm (MALA)* or the *hybrid Monte Carlo (HMC) method*. The basic idea of MALA (as well as HMC) is to rewrite (3.17) with $f(x) = -\nabla_x U(x)$, $Q = I$ as

$$p_{n+1/2} = p_n - \frac{1}{2}\sqrt{2\Delta t}\nabla_x U(x_n), \tag{3.32}$$

$$x_{n+1} = x_n + \sqrt{2\Delta t}\, p_{n+1/2}, \tag{3.33}$$

$$p_{n+1} = p_{n+1/2} - \frac{1}{2}\sqrt{2\Delta t}\nabla_x U(p_n), \tag{3.34}$$

having introduced a dummy momentum variable p with p_n being a realization of the random variable $Z_n \sim \mathrm{N}(0, I)$. Under the Metropolis accept-reject criterion, x_{n+1} is accepted with probability

$$\min\left\{1, \exp\left(-\left(E_{n+1} - E_n\right)\right)\right\},$$

where

$$E_n = \frac{1}{2}p_n^{\mathrm{T}} p_n + U(x_n), \quad E_{n+1} = \frac{1}{2}p_{n+1}^{\mathrm{T}} p_{n+1} + U(x_{n+1})$$

are the initial and final energies. Upon rejection, one continues with x_n. The momentum value p_{n+1} is discarded after a completed time-step (regardless of its acceptance or rejection) and a new momentum value is drawn from $\mathrm{N}(0, I)$. It should however be noted that $|E_{n+1} - E_n| \to 0$ as the step-size Δt goes to zero, and, in practice, the application of the Metropolis accept-rejection step is often not necessary unless Δt is chosen too large. The HMC method differs from MALA in that several iterations of (3.32–3.34) are applied before the Metropolis accept-reject criterion is being applied.

References

A gentle introduction to stochastic processes can be found in [17] and [10]. A more mathematical treatment can be found in [8, 40] and numerical issues are discussed in [21, 26]. See [42, 52] for a discussion of the gradient flow structure of the Fokker–Planck equation. The ergodic behavior of Markov chains is covered in [34]. Markov chain Monte Carlo methods and the hybrid Monte Carlo method in particular are treated in [32]. See also [47] for the Metropolis adjusted Langevin algorithm (MALA).

3 Data assimilation and filtering

In this section, we combine Bayesian inference and stochastic processes to tackle the problem of assimilating observational data into scientific models.

3.1 Preliminaries

We select a model written as a time-discretized SDE, such as (3.17), with the initial random variable satisfying $X_0 \sim \pi_0$. In addition to the pure prediction problem of computing π_n, $n \geq 1$, for given π_0, we assume that model states $x \in \mathcal{X} = \mathbb{R}^N$ are partially observed at equally spaced instances in time. These observations are to be assimilated into the model. More generally, *intermittent data assimilation* is concerned with fixed observation intervals $\Delta t_{\mathrm{obs}} > 0$ and model time-steps Δt such that $\Delta t_{\mathrm{obs}} = L\Delta t$, $L \geq 1$, which allows one to take the limit $L \to \infty$, $\Delta t = \Delta t_{\mathrm{obs}}/L \to 0$. For simplicity, we will restrict the discussion to the case where observations $y_0(t_n) = Y_n(\omega) \in \mathbb{R}^K$ are made at every time step $t_n = n\Delta t$, $n \geq 1$ and the limit $\Delta t \to 0$ is not considered here. We will further assume that the observed random variables Y_n satisfy the model (3.2), i.e.

$$Y_n = h(X_n) + \Xi_n$$

and the measurement errors $\Xi_n \sim \mathrm{N}(0, R)$ are mutually independent with common error covariance matrix R. We introduce the notation $\mathrm{Y}_k = \{y_0(t_i)\}_{i=1,\ldots,k}$ to denote all observations up to and including time t_k.

Definition 3.28 (Data assimilation). *Data assimilation* is the estimation of marginal PDFs $\pi_n(x|\mathrm{Y}_k)$ of the random variable $X_n = X(t_n)$ conditioned on the set of observations Y_k. We distinguish three cases: (i) *filtering* $k = n$, (ii) *smoothing* $k > n$, and (iii) *prediction* $k < n$.

The subsequent discussions are restricted to the filtering problem. We have already seen that evolution of the marginal distributions under (3.17) alone is governed by the Chapman–Kolmogorov equation (3.15) with transition probability density (3.18). We denote the associated Frobenius–Perron operator (3.16) by $\mathcal{P}_{\Delta t}$. Given $X_0 \sim \pi_0$, we first obtain

$$\pi_1 = \mathcal{P}_{\Delta t}\pi_0 \,.$$

This time propagated PDF is used as the prior PDF $\pi_X = \pi_1$ in Bayes' formula (3.4) at $t = t_1$ with $y_0 = y_0(t_1)$ and likelihood

$$\pi_Y(y|x) = \frac{1}{(2\pi)^{N/2}|R|^{1/2}} \exp\left(-\frac{1}{2}\left(y - h(x)\right)^{\mathrm{T}} R^{-1}\left(y - h(x)\right)\right) \,.$$

Bayes' formula implies the posterior PDF

$$\pi_1\left(x|\mathrm{Y}_1\right) \propto \pi_Y\left(y_0\left(t_1\right)|x\right)\pi_1(x) \,,$$

where the constant of proportionality only depends on $y_0(t_1)$.

Proposition 3.29 (Sequential filtering). *The filtering problem leads to the recursion*

$$\begin{aligned}\pi_{n+1}\left(\cdot|\mathrm{Y}_n\right) &= \mathcal{P}_{\Delta t}\pi_n\left(\cdot|\mathrm{Y}_n\right) \,, \\ \pi_{n+1}\left(x|\mathrm{Y}_{n+1}\right) &\propto \pi_Y\left(y_0\left(t_{n+1}\right)|x\right)\,\pi_{n+1}\left(x|\mathrm{Y}_n\right) \,,\end{aligned} \tag{3.35}$$

$n \geq 0$, and $X_n \sim \pi_n(\cdot|\mathrm{Y}_n)$ solves the filtering problem at time t_n. The constant of proportionality only depends on $y_0(t_{n+1})$.

Proof. The recursion follows by induction. □

Recall that the Frobenius–Perron operator $\mathcal{P}_{\Delta t}$ is generated by the stochastic difference equation (3.17). On the other hand, Bayes' formula only leads to a transition from the predicted $\pi_{n+1}(x|\mathrm{Y}_n)$ to the filtered $\pi_{n+1}(x|\mathrm{Y}_{n+1})$. Following our discussion on transport maps from Section 1.3, we assume the existence of a transport map $X' = T_{n+1}(X)$, depending on $y_0(t_{n+1})$, that couples the two PDFs. The use of optimal transport maps in the context of Bayesian inference and intermittent data assimilation was first proposed in [37, 43].

Proposition 3.30 (Filtering by transport maps). *Assuming the existence of appropriate transport maps T_{n+1}, which couple $\pi_{n+1}(x|\mathrm{Y}_n)$ and $\pi_{n+1}(x|\mathrm{Y}_{n+1})$, the filtering problem is solved by the following recursion for the random variables X_{n+1}, $n \geq 0$:*

$$X_{n+1} = T_{n+1}\left(X_n + \Delta t f(X_n) + \sqrt{2\Delta t}Z_n\right) \,, \tag{3.36}$$

which gives rise to a time-dependent Markov process.

Proof. Follows trivially from (3.35). □

The rest of this section is devoted to several Monte Carlo methods for sequential filtering.

3.2 Sequential Monte Carlo method

In our framework, a standard sequential Monte Carlo method, also called *bootstrap particle filter*, may be described as an ensemble of random variables X_i and associated realizations (referred to as "particles") $x_i = X_i(\omega)$, which follow the stochastic difference equation (3.17), choosing the transport map in (3.36) to be the identity map. Observational data is taken into account using importance sampling as discussed in Section 1.4, i.e. each particle carries a weight $w_i(t_n)$, which is updated according to Bayes' formula

$$w_i(t_{n+1}) \propto w_i(t_n)\pi\left(y_0(t_{n+1})|x_i(t_{n+1})\right) \,.$$

The constant of proportionality is chosen such that the new weights $\{w_i(t_{n+1})\}_{i=1}^M$ sum to one.

Whenever the particle weights $w_i(t_n)$ start to become highly nonuniform (or possibly also after each assimilation step), resampling is necessary in order to generate a new family of random variables with equal weights.

Most available resampling methods start from the *weighted empirical measure*

$$\mu_X(\mathrm{d}x) = \sum_{i=1}^{M} w_i \mu_{x_i}(\mathrm{d}x) \tag{3.37}$$

associated with a set of weighted samples $\{x_i, w_i\}_{i=1}^M$. The idea is to replace each of the original samples x_i by $\xi_i \geq 0$ offsprings with equal weights $\hat{w}_i = 1/M$ such that $\mathbb{E}[\xi_i] = w_i M$. The distribution of offsprings is chosen to be equal to the distribution of M samples (with replacement) drawn at random from the empirical distribution (3.37). In other words, the offsprings $\{\xi_i\}_{i=1}^M$ follow a *multinomial distribution* defined by

$$\mathbb{P}\left(\xi_i = n_i, i = 1, \ldots, M\right) = \frac{M!}{\prod_{i=1}^M n_i!} \prod_{i=1}^{M} (w_i)^{n_i} \tag{3.38}$$

with $n_i \geq 0$ such that $\sum_{i=1}^M n_i = M$. In practice, independent resampling is often replaced by residual or systematic resampling. We next summarize residual resampling while we refer the reader to [3] for an algorithmic description of systematic resampling.

Definition 3.31 (Residual resampling). *Residual resampling* generates

$$\xi_i = \lfloor M w_i \rfloor + \overline{\xi}_i \,,$$

offsprings of each ensemble member x_i with weight w_i, $i = 1, \ldots, M$. Here, $\lfloor x \rfloor$ denotes the integer part of x and $\overline{\xi}_i$ follows the multinomial distribution (3.38) with weights w_i being replaced by

$$\overline{w}_i = \frac{M w_i - \lfloor M w_i \rfloor}{\sum_{j=1}^M \left(M w_j - \lfloor M w_j \rfloor\right)}$$

and with a total of

$$\sum_{i=1}^{M} n_i = \overline{M} := M - \sum_i \lfloor M w_i \rfloor$$

independent trials.

In generalization of (3.38), we introduce the notation $\mathrm{Mult}(L; \omega_1, \ldots, \omega_M)$ to denote the multinomial distribution of L independent trials, where the outcome of each trial is distributed among M possible outcomes according to probabilities $\{\omega_i\}_{i=1}^M$. The following algorithm draws random samples from $\mathrm{Mult}(L; \omega_1, \ldots, \omega_M)$. We first introduce the generalized right inverse

$$F_{\mathrm{emp}}^{-1}(u) = i \quad \Longleftrightarrow \quad u \in \left(\sum_{j=1}^{i-1} \omega_j, \sum_{j=1}^{i} \omega_j\right]$$

of the cumulative distribution function $F_{\mathrm{emp}}^{-1} : [0,1] \to \{1,\dots,M\}$ for the empirical measure (3.37). We next draw L independent samples $u_l \in [0,1]$ from the uniform distribution $\mathrm{U}([0,1])$ and initially set the number of copies $\overline{\xi}_i$, $i = 1,\dots,M$, equal to zero. For $l = 1,\dots,L$, we now increment $\overline{\xi}_{I_l}$ by one for indices $I_l \in \{1,\dots,M\}$, $l = 1,\dots,L$, defined by

$$I_l = F_{\mathrm{emp}}^{-1}(u_l) = \arg\min_{i\geq 1} \sum_{j=1}^{i} \omega_j \geq u_l\,.$$

Both independent and residual resampling can be viewed as providing a coupling between the empirical measure (3.37) with all weights being equal to $w_i = 1/M$ and the target measure (3.37) with identical samples $\{x_i\}$, but nonuniform weights. Clearly, residual resampling provides a coupling with a smaller transport cost. This can already be concluded from the trivial case of equal weights in the target measure in which case residual resampling reduces to the identity map with zero transport cost, while independent resampling remains nondeterministic and produces a nonzero transport cost. The following example outlines the optimal transportation perspective on resampling more precisely for two discrete, univariate random variables.

Example 3.32 (Coupling discrete random variables). Let us consider two discrete, univariate random variables $X_i : \Omega \to \mathcal{X}$, $i = 1, 2$, with target set

$$\mathcal{X} = \{x_1, x_2, \dots, x_M\} \in \mathbb{R}^M.$$

We furthermore assume that

$$\mathbb{P}\left(X_1(\omega) = x_i\right) = 1/M\,, \quad \mathbb{P}\left(X_2(\omega) = x_i\right) = w_i$$

for given probabilities/weights $w_i \geq 0$, $i = 1,\dots,M$. Any coupling of X_1 and X_2 is characterized by a matrix $\mathcal{T} \in \mathbb{R}^{M\times M}$ such that $t_{ij} = (\mathcal{T})_{ij} \geq 0$ and

$$\sum_{i=1}^{M} t_{ij} = 1/M\,, \quad \sum_{j=1}^{M} t_{ij} = w_i.$$

Given a coupling $\mathcal{T}$ and the mean values

$$\overline{x}_1 = \frac{1}{M}\sum_i x_i\,, \quad \overline{x}_2 = \sum_i w_i x_i,$$

the covariance between X_1 and X_2 is defined by

$$\mathrm{cov}(X_1, X_2) = \sum_{i,j} (x_i - \overline{x}_2) t_{ij} (x_j - \overline{x}_1).$$

The induced Markov transition matrix from X_1 to X_2 is simply given by $M\mathcal{T}$. Independent resampling corresponds to $t_{ij} = w_i/M$ and leads to a zero correlation between X_1 and X_2. On the other hand, maximizing the correlation results in a linear programming problem for the M^2 unknowns $\{t_{ij}\}$. Its solution then also defines the solution to the associated optimal transportation problem (3.6). Implementations of this approach for sequential data assimilation are discussed in [46].

More generally, sequential Monte Carlo methods differ by the way resampling is implemented and also in the choice of proposal step which, in our context, amounts to choosing transport maps T_{n+1} in (3.36) which are different from the identity map. See also the discussion in Section 3.5 below.

3.3 Ensemble Kalman filter (EnKF)

We now introduce an alternative to sequential Monte Carlo methods which has become hugely popular in the geophysical community in recent years. The idea is to construct a simple, but robust transport map T'_{n+1} which replaces T_{n+1} in (3.36). This transport map is based on the *Kalman update equations* for linear SDEs and Gaussian prior and posterior distributions. We recall the standard Kalman filter update equations.

Proposition 3.33 (Kalman update for Gaussian distributions). *Let the prior distribution π_X be Gaussian with mean $\overline{x}^f$ and covariance matrix P^f. Observations y_0 are assumed to follow the linear model*

$$Y = HX + \Xi\,,$$

where $\Xi \sim \mathrm{N}(0, R)$ and R is a symmetric, positive-definite matrix. Then, the posterior distribution $\pi_X(x|y_0)$ is also Gaussian with mean

$$\overline{x}^a = \overline{x}^f - P^f H^{\mathrm{T}}(HP^f H^{\mathrm{T}} + R)^{-1}(H\overline{x}^f - y_0) \tag{3.39}$$

and covariance matrix

$$P^a = P^f - P^f H^{\mathrm{T}}(HP^f H^{\mathrm{T}} + R)^{-1} HP^f\,. \tag{3.40}$$

Here, we adopt the standard meteorological notation with superscript f (forecast) denoting prior statistics, and superscript a (analysis) denoting posterior statistics.

Proof. By straightforward generalization to vector-valued observations of the case of a scalar observation already discussed in Section 1.2. □

EnKFs rely on the assumption that the predicted PDF $\pi_{n+1}(x|\mathrm{Y}_n)$ is approximately Gaussian. The ensemble $\{x_i\}_{i=1}^M$ of model states is used to estimate the mean and the covariance matrix using the empirical estimates (3.28) and (3.30), respectively. The key novel idea of EnKFs is to then interpret the posterior mean and covariance matrix in terms of appropriately adjusted ensemble positions. This adjustment can be thought of as a coupling of the underlying prior and posterior random variables of which the ensembles are realizations. The original EnKF [9] uses perturbed observations to achieve the desired coupling.

Definition 3.34 (Ensemble Kalman Filter). The *EnKF with perturbed observations* for a linear observation operator $h(x) = Hx$ is given by

$$X^f_{n+1} = X_n + \Delta t f(X_n) + \sqrt{2\Delta t} Z_n \,, \tag{3.41}$$

$$X_{n+1} = X^f_{n+1} - P^f_{n+1} H^{\mathrm{T}} \left(H P^f_{n+1} H^{\mathrm{T}} + R \right)^{-1} \left(H X^f_{n+1} - y_0 + \Sigma_{n+1} \right) , \tag{3.42}$$

where the random variables $Z_n \sim \mathrm{N}(0, Q)$, $\Sigma_{n+1} \sim \mathrm{N}(0, R)$ are mutually independent, $y_0 = y_0(t_{n+1})$, $\overline{x}^f_{n+1} = \mathbb{E}_{X^f_{n+1}}[x]$, and

$$P^f_{n+1} = \mathbb{E}_{X^f_{n+1}} \left[\left(x - \overline{x}^f_{n+1} \right) \left(x - \overline{x}^f_{n+1} \right)^{\mathrm{T}} \right] .$$

Next, we investigate the properties of the assimilation step (3.42).

Proposition 3.35 (EnKF consistency). *The EnKF update step* (3.42) *propagates the mean and covariance matrix of X in accordance with the Kalman filter equations for Gaussian PDFs.*

Proof. It is easy to verify that the ensemble mean satisfies

$$\overline{x}_{n+1} = \overline{x}^f_{n+1} - P^f_{n+1} H^{\mathrm{T}} \left(H P^f_{n+1} H^{\mathrm{T}} + R \right)^{-1} \left(H \overline{x}^f_{n+1} - y_0 \right) ,$$

which is consistent with the Kalman filter update for the ensemble mean. Furthermore, the deviation $\delta X = X - \overline{x}$ satisfies

$$\delta X_{n+1} = \delta X^f_{n+1} - P^f_{n+1} H^{\mathrm{T}} \left(H P^f_{n+1} H^{\mathrm{T}} + R \right)^{-1} \left(H \delta X^f_{n+1} + \Sigma_{n+1} \right) ,$$

which implies

$$\begin{aligned} P_{n+1} &= P^f_{n+1} - 2 P^f_{n+1} H^{\mathrm{T}} \left(H P^f_{n+1} H^{\mathrm{T}} + R \right)^{-1} H P^f_{n+1} \\ &\quad + P^f_{n+1} H^{\mathrm{T}} \left(H P^f_{n+1} H^{\mathrm{T}} + R \right)^{-1} R \left(H P^f_{n+1} H^{\mathrm{T}} + R \right)^{-1} H P^f_{n+1} \\ &\quad + P^f_{n+1} H^{\mathrm{T}} \left(H P^f_{n+1} H^{\mathrm{T}} + R \right)^{-1} H P^f_{n+1} H^{\mathrm{T}} \left(H P^f_{n+1} H^{\mathrm{T}} + R \right)^{-1} H P^f_{n+1} \\ &= P^f_{n+1} - P^f_{n+1} H^{\mathrm{T}} \left(H P^f_{n+1} H^{\mathrm{T}} + R \right)^{-1} H P^f_{n+1} \end{aligned}$$

for the update of the covariance matrix, which is also consistent with the Kalman update step for Gaussian random variables. □

Practical implementations of the EnKF with perturbed observations replace the exact mean and covariance matrix by ensemble based empirical estimates (3.28) and (3.30), respectively.

Alternatively, we can derive a transport map T under the assumption of Gaussian prior and posterior distributions as follows. Using the empirical ensemble mean $\overline{x}$, we define ensemble deviations by $\delta x_i = x_i - \overline{x} \in \mathbb{R}^N$ and an associated ensemble deviation matrix $\delta \mathrm{X} = (\delta x_1, \dots, \delta x_M) \in \mathbb{R}^{N \times M}$. Using this notation, the empirical covariance matrix of the prior ensemble at t_{n+1} is then given by

$$P_{n+1}^f = \frac{1}{M-1} \delta \mathrm{X}_{n+1}^f \left(\delta \mathrm{X}_{n+1}^f \right)^{\mathrm{T}}.$$

We next seek a matrix $S \in \mathbb{R}^{M \times M}$ such that

$$P_{n+1} = \frac{1}{M-1} \delta \mathrm{X}_{n+1}^f S S^{\mathrm{T}} \left(\delta \mathrm{X}_{n+1}^f \right)^{\mathrm{T}},$$

where the rows of S sum to zero in order to preserve the zero mean property of $\delta \mathrm{X}_{n+1} = \delta \mathrm{X}_{n+1}^f S$. Such matrices do exist (see, e.g. [15]), and give rise to the *ensemble square root filters*. More specifically, Kalman's update formula (3.40) for the posterior covariance matrix implies

$$\begin{aligned} P^a &= \frac{1}{M-1} \delta \mathrm{X}^f \left\{ I - \frac{1}{M-1} \left(\delta \mathrm{Y}^f \right)^{\mathrm{T}} \left[H P^f H^{\mathrm{T}} + R \right]^{-1} \delta \mathrm{Y}^f \right\} \left(\delta \mathrm{X}^f \right)^{\mathrm{T}} \\ &= \frac{1}{M-1} \delta \mathrm{X}^f S S^{\mathrm{T}} \left(\delta \mathrm{X}^f \right)^{\mathrm{T}}, \end{aligned}$$

where we have dropped the time index subscript and introduced the ensemble perturbations $\delta \mathrm{Y}^f = H \delta \mathrm{X}^f$ in observation space $\mathcal{Y}$. Recalling now the definition of a matrix square root from Section 1.3 and making use of the Sherman–Morrison–Woodbury formula [18], we find that

$$\begin{aligned} S &= \left\{ I - \frac{1}{M-1} \left(\delta \mathrm{Y}^f \right)^{\mathrm{T}} \left[H P^f H^{\mathrm{T}} + R \right]^{-1} \delta \mathrm{Y}^f \right\}^{1/2} \\ &= \left\{ I + \frac{1}{M-1} \left(\delta \mathrm{Y}^f \right)^{\mathrm{T}} R^{-1} \delta \mathrm{Y}^f \right\}^{-1/2}. \end{aligned} \tag{3.43}$$

The complete ensemble update of an ensemble square root filter is then given by

$$x_i(t_{n+1}) = \overline{x}_{n+1} + \delta \mathrm{X}_{n+1}^f S e_i, \tag{3.44}$$

where e_i denotes the ith basis vector in $\mathbb{R}^M$ and

$$\overline{x}_{n+1} = \overline{x}_{n+1}^f - P_{n+1}^f H^{\mathrm{T}} \left(H P_{n+1}^f H^{\mathrm{T}} + R \right)^{-1} \left(H \overline{x}_{n+1}^f - y_0(t_{n+1}) \right)$$

denotes the updated ensemble mean.

We now discuss the update (3.44) from the perspective of optimal transportation which, in our context, reduces to finding a matrix $S_{\mathrm{OT}} \in \mathbb{R}^{M \times M}$ such that the trace of

$$\operatorname{cov}\left(\delta \mathrm{X}_{n+1}^f, \delta \mathrm{X}_{n+1} \right) = \mathbb{E}\left[\delta \mathrm{X}_{n+1}^f S_{\mathrm{OT}}^{\mathrm{T}} \left(\delta \mathrm{X}_{n+1}^f \right)^{\mathrm{T}} \right]$$

is maximized.

Proposition 3.36 (Optimal update for ensemble square root filter). *The trace of the covariance matrix* $\mathrm{cov}(\delta \mathrm{X}^f_{n+1}, \delta \mathrm{X}_{n+1})$ *is maximized for*

$$\delta \mathrm{X}_{n+1} = \delta \mathrm{X}^f_{n+1} S_{\mathrm{OT}}$$

with transform matrix

$$S_{\mathrm{OT}} = \frac{1}{\sqrt{M-1}} S \left[S \left(\delta \mathrm{X}^f_{n+1}\right)^{\mathrm{T}} P^f_{n+1} \delta \mathrm{X}^f_{n+1} S \right]^{-1/2} S \left(\delta \mathrm{X}^f_{n+1}\right)^{\mathrm{T}} \delta \mathrm{X}^f_{n+1}$$

and $S \in \mathbb{R}^{M\times M}$ *given by* (3.43).

Proof. Follows from (3.9) with $A = \delta \mathrm{X}^f_{n+1} S/\sqrt{M-1}$ and $\Sigma_1 = P^f_{n+1}$. The left multiplication in (3.8) is finally rewritten as a right multiplication by $S_{\mathrm{OT}} \in \mathbb{R}^{M\times M}$ in terms of ensemble deviations $\delta \mathrm{X}^f_{n+1}$. □

We finish this section by briefly discussing a couple of practical issues. It is important to recall that the Kalman filter can be viewed as a linear minimum variance estimator [14]. At the same time, it has been noted [30, 55] that the updated ensemble mean $\overline{x}_{n+1}$ is biased in case where the prior distribution is not Gaussian. Hence, the associated mean squared error (3.10) does not vanish as $M \to \infty$ even though the variance of the estimator goes to zero. If desired, the bias can be removed by replacing $\overline{x}_{n+1}$ in (3.44) by (3.11) with weights (3.12), where $y_0 = y_0(t_{n+1})$ and $x_i^{\mathrm{prior}} = x_i^f(t_{n+1})$. Higher-order moment corrections can also be implemented [30, 55]. However, the filter performance only improves for sufficiently large ensemble sizes.

We mention the *unscented Kalman filter* [23] as an alternative extension of the Kalman filter to nonlinear dynamical systems. We also mention the *rank histogram filter* [2], which is based on first constructing an approximative coupling in the observed variable y alone followed by linear regression of the updates in y onto the state space variable x.

Practical implementations of EnKFs for high-dimensional problems rely on additional modifications, in particular, *inflation* and *localization*. While localization modifies the covariance matrix P^f_{n+1} in the Kalman update (3.42) in order to increase its rank and to localize the spatial impact of observations in physical space, inflation increases the ensemble spread $\delta x_i = x_i - \overline{x}$ by replacing x_i by $\overline{x} + \alpha(x_i - \overline{x})$ with $\alpha > 1$. Note that the second term on the right-hand side of (3.31) achieves a similar effect and ensemble inflation can be viewed as a simple parametrization of (stochastic) model errors. See [15] for more details on inflation and localization techniques.

3.4 Ensemble transform Kalman–Bucy filter

In this section, we describe an alternative implementation of ensemble square root filters based on the Kalman–Bucy filter. We first describe the Kalman–Bucy formulation of the linear filtering problem for Gaussian PDFs.

Proposition 3.37 (Kalman–Bucy equations). *The Kalman update step* (3.39)–(3.40) *can be formulated as a differential equation in artificial time $s \in [0,1]$. The Kalman–Bucy equations are*

$$\frac{\mathrm{d}\overline{x}}{\mathrm{d}s} = -PH^{\mathrm{T}}R^{-1}\left(H\overline{x} - y_0\right)$$

and

$$\frac{\mathrm{d}P}{\mathrm{d}s} = -PH^{\mathrm{T}}R^{-1}HP\,.$$

The initial conditions are $\overline{x}(0) = \overline{x}^f$ and $P(0) = P^f$, and the Kalman update is obtained from the final conditions $\overline{x}^a = \overline{x}(1)$ and $P^a = P(1)$.

Proof. We present the proof for $N = 1$ (one-dimensional state space) and $K = 1$ (a single observation). Under this assumption, the standard Kalman analysis step (3.39)–(3.40) gives rise to

$$P^a = \frac{P^f R}{P^f + R}, \quad \overline{x}^a = \frac{\overline{x}^f R + y_0 P^f}{P^f + R},$$

for a given observation value y_0.

We now demonstrate that this update is equivalent to twice the application of a Kalman analysis step with R replaced by $2R$. Specifically, we obtain

$$\hat{P}^a = \frac{2P_m R}{P_m + 2R}, \quad P_m = \frac{2P^f R}{P^f + 2R}$$

for the resulting covariance matrix $\hat{P}^a$ with intermediate value P_m. The analyzed mean $\hat{x}^a$ is provided by

$$\hat{x}^a = \frac{2\overline{x}_m R + y_0 P_m}{P_m + 2R}, \quad \overline{x}_m = \frac{2\overline{x}^f R + y_0 P^f}{P^f + 2R}.$$

We need to demonstrate that $P^a = \hat{P}^a$ and $\overline{x}^a = \hat{x}^a$. We start with the covariance matrix and obtain

$$\hat{P}^a = \frac{\frac{4P^f R}{P^f + 2R} R}{\frac{2P^f R}{P^f + 2R} + 2R} = \frac{4P^f R^2}{4P^f R + 4R^2} = \frac{P^f R}{P^f + R} = P^a\,.$$

A similar calculation for $\hat{x}^a$ yields

$$\hat{x}^a = \frac{2\frac{2\overline{x}^f R + y_0 P^f}{P^f + 2R} R + y_0 \frac{2P^f R}{P^f + 2R}}{2R + \frac{2P^f R}{P^f + 2R}} = \frac{4\overline{x}^f R^2 + 4y_0 P^f R}{4R^2 + 4RP^f} = \overline{x}^a\,.$$

Hence, by induction, we can replace the standard Kalman analysis step by $D > 2$ iterative applications of a Kalman analysis with R replaced by DR. We set $P_0 = P^f$, $\overline{x}_0 = \overline{x}^f$, and iteratively compute P_{j+1} from

$$P_{j+1} = \frac{DP_j R}{P_j + DR}, \quad \overline{x}_{j+1} = \frac{D\overline{x}_j R + y_0 P_j}{P_j + DR}$$

for $j = 0, \dots, D-1$. We finally set $P^a = P_D$ and $\overline{x}^a = \overline{x}_D$. Next, we introduce a step-size $\Delta s = 1/D$ and assume $D \gg 1$. Then,

$$\overline{x}_{j+1} = \frac{\overline{x}_j R + \Delta s y_0 P_j}{R + \Delta s P_j} = \overline{x}_j - \Delta s P_j R^{-1}(\overline{x}_j - y_0) + \mathcal{O}(\Delta s^2)$$

as well as

$$P_{j+1} = \frac{P_j R}{R + \Delta s P_j} = P_j - \Delta s P_j R^{-1} P_j + \mathcal{O}(\Delta s^2)\,.$$

Taking the limit $\Delta s \to 0$, we obtain the two differential equations

$$\frac{\mathrm{d}P}{\mathrm{d}s} = -PR^{-1}P\,, \quad \frac{\mathrm{d}\overline{x}}{\mathrm{d}s} = -PR^{-1}(\overline{x} - y_0)$$

for the covariance and mean, respectively. The equation for P can be rewritten in terms of its square root Y (i.e. $P = Y^2$) as

$$\frac{\mathrm{d}Y}{\mathrm{d}s} = -\frac{1}{2}PR^{-1}Y\,. \tag{3.45}$$

□

Upon formally setting $Y = \delta \mathrm{X}/\sqrt{M-1}$ in (3.45), the Kalman–Bucy filter equations give rise to a particular implementation of ensemble square root filters in terms of evolution equations in artificial time $s \in [0,1]$.

Definition 3.38 (Ensemble transform Kalman–Bucy filter equations). The *ensemble transform Kalman–Bucy filter* equations [1, 5, 6] for the assimilation of an observation $y_0 = y_0(t_n)$ at t_n are given by

$$\frac{\mathrm{d}x_i}{\mathrm{d}s} = -\frac{1}{2}PH^{\mathrm{T}}R^{-1}\left(Hx_i + H\overline{x} - 2y_0(t_n)\right)$$

in terms of the ensemble members x_i, $i = 1, \dots, M$, and are solved over a unit time interval in artificial time $s \in [0,1]$. Here, P denotes the empirical covariance matrix (3.30) and $\overline{x}$ denotes the empirical mean (3.28) of the ensemble.

The Kalman–Bucy equations are realizations of an underlying differential equation

$$\frac{\mathrm{d}X}{\mathrm{d}s} = -\frac{1}{2}PH^{\mathrm{T}}R^{-1}\left(HX + H\overline{x} - 2y_0(t_n)\right) \tag{3.46}$$

in the random variable X with mean

$$\overline{x} = \mathbb{E}_X[x] = \int x\pi_X \mathrm{d}x$$

and covariance matrix

$$P = \mathbb{E}_X\left[(x - \overline{x})(x - \overline{x})^{\mathrm{T}}\right].$$

The associated evolution of the PDF π_X (here assumed to be absolutely continuous) is given by Liouville's equation

$$\frac{\partial \pi_X}{\partial s} = -\nabla_x \cdot (\pi_X v) \tag{3.47}$$

with vector field

$$v(x) = -\frac{1}{2} P H^{\mathrm{T}} R^{-1} \left(Hx + H\overline{x} - 2y_0(t_n) \right) . \tag{3.48}$$

Recalling the earlier discussion of the Fokker–Planck equation in Section 2.2, we note that (3.47) with vector field (3.48) also has an interesting geometric structure.

Proposition 3.39 (Ensemble transform Kalman–Bucy equations as a gradient flow). *The vector field* (3.48) *is equivalent to*

$$v(x) = -P\nabla_x \frac{\delta F}{\delta \pi_X}$$

with potential

$$\begin{aligned} F(\pi_X) = {} & \frac{1}{4} \int\limits_{\mathbb{R}^N} \left(Hx - y_0(t_n) \right)^{\mathrm{T}} R^{-1} \left(Hx - y_0(t_n) \right) \pi_X \mathrm{d}x \\ & + \frac{1}{4} \left(H\overline{x} - y_0(t_n) \right)^{\mathrm{T}} R^{-1} \left(H\overline{x} - y_0(t_n) \right) . \end{aligned} \tag{3.49}$$

Liouville's equation (3.47) *can be stated as*

$$\frac{\partial \pi_X}{\partial s} = -\nabla_x \cdot (\pi_X v) = -\mathrm{grad}_{\pi_X} F(\pi_X) ,$$

where we have used $\mathrm{M} = P$ *in the definition of the gradient* (3.25).

Proof. The result can be verified by direct calculation. □

Nonlinear forward operators can be treated in this framework by replacing the potential (3.49) by, for example,

$$\begin{aligned} F(\pi_X) = {} & \frac{1}{4} \int\limits_{\mathbb{R}^N} \left(h(x) - y_0(t_n) \right)^{\mathrm{T}} R^{-1} \left(h(x) - y_0(t_n) \right) \pi_X \mathrm{d}x \\ & + \frac{1}{4} \left(h(\overline{x}) - y_0(t_n) \right)^{\mathrm{T}} R^{-1} \left(h(\overline{x}) - y_0(t_n) \right) . \end{aligned}$$

Efficient time-stepping methods for the ensemble transform Kalman–Bucy filter equations are discussed in [1] and an application to continuous data assimilation can be found in [6].

3.5 Guided sequential Monte Carlo methods

EnKF techniques are limited by the fact that the empirical PDFs do not converge to the filter solution in the limit of ensemble sizes $M \to \infty$ unless the involved PDFs are Gaussian. Sequential Monte Carlo methods, on the other hand, can be shown to converge under fairly general assumptions, but they do not work well in high-dimensional phase spaces since importance sampling is not sufficient to guarantee good performance of a particle filter for finite ensemble sizes. In particular, the variance in the associated mean squared error (3.10) can be very large for ensemble sizes typically used in geophysical applications.

The combination of modified particle positions and appropriately adjusted particle weights appears therefore as a promising area for research and might achieve a better bias-variance trade-off than either the EnKF or traditional sequential Monte Carlo methods. In particular, combining ensemble transform techniques, such as EnKF, with sequential Monte Carlo methods appears as a natural research direction. Indeed, in the framework of Monte Carlo methods discussed in Section 1.4, the standard sequential Monte Carlo approach consists of importance sampling using proposal PDF $\pi'_X(x) = \pi_{n+1}(x|\mathrm{Y}_n)$ and subsequent reweighting of particles according to (3.12). Also, as discussed in Section 1.4, the performance of importance sampling can be improved by applying modified proposal densities with the aim of pushing the updated ensemble members $x_i(t_{n+1})$ to regions of high and nearly equal probability in the targeted posterior PDF $\pi_{n+1}(x|\mathrm{Y}_{n+1})$ (compare with equation (3.14)). We call the resulting filter algorithms *guided sequential Monte Carlo methods*.

More precisely, a guided sequential Monte Carlo method is defined by a conditional proposal PDF $\tilde{\pi}_{n+1}(x'|x, y_0(t_{n+1}))$ and an associated joint PDF

$$\tilde{\pi}_{X'X}(x', x|\mathrm{Y}_{n+1}) = \tilde{\pi}_{n+1}\left(x'|x, y_0(t_{n+1})\right)\, \pi_n(x|\mathrm{Y}_n)\,. \tag{3.50}$$

An ideal proposal density (in the sense of coupling) should be identical to the posterior PDF $\pi_{n+1}(x|\mathrm{Y}_{n+1})$. In guided sequential Monte Carlo methods, a mismatch between the proposal density and $\tilde{\pi}_{n+1}(x|\mathrm{Y}_{n+1})$ is treated by adjusted particle weights $w_i(t_{n+1})$. Following the general methodology of importance sampling, one obtains the recursion

$$w_i\left(t_{n+1}\right) \propto \frac{\pi_Y\left(y_0\left(t_{n+1}\right)|x'_i\right)\pi\left(x'_i|x_i\right)}{\tilde{\pi}_{n+1}\left(x'_i|x_i, y_0\left(t_{n+1}\right)\right)} w_i\left(t_n\right)\,.$$

Here, $\pi(x'|x)$ denotes the conditional PDF (3.18) describing the model dynamics, (x'_i, x_i), $i = 1, \dots, M$, are realizations from the joint PDF (3.50) with weights $w_i(t_n)$,

$x_i = x_i(t_n)$, and the approximation

$$\begin{aligned}\mathbb{E}_{X_{n+1}}[g] &= \frac{1}{\pi_Y(y_0(t_{n+1}))}\int_{\mathbb{R}^N}\int_{\mathbb{R}^N} f(x',x)\,\tilde{\pi}_{X'X}(x',x|\mathrm{Y}_{n+1})\,\mathrm{d}x'\mathrm{d}x \\ &\approx \frac{1}{\pi_Y(y_0(t_{n+1}))}\sum_{i=1}^{M} w_i(t_n)\, f\left(x_i', x_i\right) \\ &\propto \sum_{i=1}^{M} w_i(t_{n+1})\, g\left(x_i'\right)\end{aligned}$$

with

$$f(x',x) = g(x')\,\frac{\pi_Y(y_0(t_{n+1})|x')\,\pi(x'|x)}{\tilde{\pi}_{n+1}(x'|x, y_0(t_{n+1}))}$$

has been used. The guided sequential Monte Carlo method is continued with $x_i(t_{n+1}) = x_i'$ and new weights $w_i(t_{n+1})$.

Numerical implementations of guided sequential Monte Carlo methods have been discussed, for example, in [7, 11, 28, 36]. More specifically, a combined particle and Kalman filter is proposed in [28] to achieve almost equal particle weights (see also the discussion in [7]), while in [11, 36], new particle positions $x_i(t_{n+1})$ are defined by means of implicit equations. We emphasize that both implementation approaches give up the requirement of unbiased estimation in hopes of reduced variance at finite ensemble sizes and hence for an overall reduction of the associated mean squared error (3.10). See also [45] for a discussion of guided sequential Monte Carlo methods from a coupling and transport perspective.

Another broad class of methods is based on Gaussian mixture approximations to the prior PDF $\pi_{n+1}(x|\mathrm{Y}_n)$. Provided that the forward operator h is linear, the posterior PDF $\pi_{n+1}(x|\mathrm{Y}_{n+1})$ is then also a Gaussian mixture and several procedures have been proposed to adjust the proposals $x_i^f(t_{n+1})$ such that the adjusted $x_i(t_{n+1})$ approximately follow the posterior Gaussian mixture PDF; see, for example, [16, 49, 50]. Broadly speaking, these methods can be understood as providing approximate transport maps T_{n+1}' instead of an exact transport map T_{n+1} in (3.36). However, none of these methods avoid the need for particle reweighting and resampling. Recall that resampling can be implemented such that it corresponds to a nondeterministic optimal transference plan.

The following section is devoted to an embedding technique for constructing accurate approximations to the transport map T_{n+1} in (3.36).

3.6 Continuous ensemble transform filter formulations

The implementation of (3.36) requires the computation of a transport map T. Optimal transportation (i.e. maximizing the covariance of the transference plan), leads to

$T = \nabla_x \psi$ and the potential satisfies the highly nonlinear, elliptic Monge–Ampere equation

$$\pi_{X_2}(\nabla_x \psi)|D\nabla_x \psi| = \pi_{X_1}\,.$$

A direct numerical implementation for high-dimensional state spaces $\mathcal{X} = \mathbb{R}^N$ seems to be presently out of reach. Instead, in this section, we utilize an embedding method due to Moser [38], replacing the optimal transport map by a suboptimal transport map which is defined as the time-one flow map of a differential equation in artificial time $s \in [0,1]$. At each time instant, determining the right-hand side of the differential equation requires the solution of a linear elliptic PDE; nonlinearity is exchanged for linearity at the cost of suboptimality. In some cases, such as Gaussian PDFs and mixtures of Gaussian, the linear PDE can be solved analytically. In other cases, further approximations, for example, the mean field approach discussed later in this section, are necessary.

Inspired by the embedding method of Moser [38], we first summarize a dynamical systems formulation [43] of Bayes' formula which generalizes the continuous EnKF formulation from Section 3.4. We first note that a single application of Bayes' formula (3.4) can be replaced by an D-fold recursive application of the incremental likelihood $\hat{\pi}$:

$$\hat{\pi}(y|x) = \frac{1}{(2\pi)^{K/2}|R|^{1/2}} \exp\left(-\frac{1}{2D}\,(h(x)-y)^{\mathrm{T}} R^{-1}\,(h(x)-y)\right)\,, \qquad (3.51)$$

i.e. we first write Bayes formula as

$$\pi_X(x|y_0) \propto \pi_X(x) \prod_{j=1}^{D} \hat{\pi}(y_0|x)\,,$$

where the constant of proportionality depends only on y_0, and then consider the implied iteration

$$\pi_{j+1}(x) = \frac{\pi_j(x)\,\hat{\pi}(y_0|x)}{\int_{\mathbb{R}^N} \mathrm{d}x\,\pi_j(x)\,\hat{\pi}(y_0|x)}$$

with $\pi_0 = \pi_X$ and $\pi_X(\cdot|y_0) = \pi_D$. We may now expand the exponential function in (3.51) in the small parameter $\Delta s = 1/D$ in the limit $D \to \infty$, obtaining the evolution equation

$$\frac{\partial \pi}{\partial s} = -\frac{1}{2}\,(h(x)-y_0)^{\mathrm{T}} R^{-1}\,(h(x)-y_0)\,\pi + \mu\pi \qquad (3.52)$$

in the fictitious time $s \in [0,1]$. The scalar Lagrange multiplier μ is equal to the expectation value of the negative log likelihood function

$$L(x;y_0) = \frac{1}{2}\,(h(x)-y_0)^{\mathrm{T}} R^{-1}\,(h(x)-y_0) \qquad (3.53)$$

with respect to π and ensures that $\int_{\mathbb{R}^N}(\partial\pi/\partial s)\mathrm{d}x = 0$. We also set $\pi(x,0) = \pi_X(x)$ and obtain $\pi_X(x|y_0) = \pi(x,1)$.

We now rewrite (3.52) in the equivalent, but more compact, form

$$\frac{\partial \pi}{\partial s} = -\pi \left(L - \overline{L} \right) , \tag{3.54}$$

where $\overline{L} = \mathbb{E}_X[L]$ and $\mathbb{E}_X$ denote expectation with respect to the PDF $\pi_X = \pi(\cdot, s)$. It should be noted that the continuous embedding defined by (3.54) is not unique. Moser [38], for example, used the linear interpolation

$$\pi(x, s) = (1 - s)\pi_X(x) + s\pi_X(x|y_0) ,$$

which results in

$$\frac{\partial \pi}{\partial s} = \pi_X(x|y_0) - \pi_X(x) . \tag{3.55}$$

Yet another interpolation is given by the displacement interpolation of McCann which is based on the optimal transportation map and which has an attractive "fluid dynamics" interpretation [52, 53].

Equation (3.54) (or, alternatively, (3.55)) defines the change (or transport) of the PDF π in fictitious time $s \in [0, 1]$. Alternatively, following Moser's work [38, 52], we can view this change as being induced by a continuity (Liouville) equation

$$\frac{\partial \pi}{\partial s} = -\nabla_x \cdot (\pi g) \tag{3.56}$$

for an appropriate vector field $g(x, s) \in \mathbb{R}^N$.

At any time $s \in [0, 1]$, the vector field $g(\cdot, s)$ is not uniquely determined by (3.54) and (3.56) unless we also require that it is the minimizer of the kinetic energy

$$\mathcal{T}(v) = \frac{1}{2} \int_{\mathbb{R}^N} \pi v^{\mathrm{T}} \mathrm{M}^{-1} v \, \mathrm{d}x$$

over all admissible vector fields $v : \mathbb{R}^N \to \mathbb{R}^N$ (i.e. g satisfies (3.56) for given π and $\partial\pi/\partial s$), where $\mathrm{M} \in \mathbb{R}^{N \times N}$ is a positive definite matrix. Under these assumptions, minimization of the functional

$$\mathcal{L}[v, \phi] = \frac{1}{2} \int_{\mathbb{R}^N} \pi v^{\mathrm{T}} \mathrm{M}^{-1} v \, \mathrm{d}x + \int_{\mathbb{R}^N} \phi \left\{ \frac{\partial \pi}{\partial s} + \nabla_{\mathbf{x}} \cdot (\pi v) \right\} \mathrm{d}x$$

for given π and $\partial\pi/\partial s$ leads to the Euler–Lagrange equations

$$\pi \mathrm{M}^{-1} g - \pi \nabla_x \psi = 0 , \quad \frac{\partial \pi}{\partial s} + \nabla_x \cdot (\pi g) = 0$$

in the velocity field g and the potential ψ. Hence, provided that $\pi > 0$, the desired vector field is given by $g = \mathrm{M}\nabla_x \psi$, and we have shown the following result.

Proposition 3.40 (Transport map from gradient flow). *If the potential $\psi(x,s)$ is the solution of the elliptic PDE*

$$\nabla_x \cdot (\pi_X \mathrm{M} \nabla_x \psi) = \pi_X (L - \overline{L}) , \tag{3.57}$$

then the desired transport map $x' = T(x)$ for the random variable X with PDF $\pi_X(x,s)$ is defined by the time-one flow map of the differential equations

$$\frac{\mathrm{d}x}{\mathrm{d}s} = -\mathrm{M}\nabla_x \psi .$$

The continuous Kalman–Bucy filter equations correspond to the special case $\mathrm{M} = P$ and $\psi = \delta F / \delta \pi_X$ with the functional F given by (3.49).

The elliptic PDE (3.57) can be solved analytically for Gaussian approximations to the PDF π_X and the resulting differential equations are equivalent to the ensemble transform Kalman–Bucy equations (3.46). Appropriate analytic expressions can also be found in the case where π_X can be approximated by a Gaussian mixture and the forward operator $h(x)$ is linear (see [44] for details).

Gaussian mixtures are contained in the class of *kernel smoothers*. It should however be noted that approximating a PDF π_X over high-dimensional phase spaces $\mathcal{X} = \mathbb{R}^N$ using kernel smoothers is a challenging task, especially if only a relatively small number of realizations x_i, $i = 1, \dots, M$, from the associated random variable X are available.

In order to overcome this curse of dimensionality, we outline a modification to the above continuous formulation, which is inspired by the rank histogram filter of Anderson [2]. For simplicity of exposition, consider a single observation $y \in \mathbb{R}$ with forward operator $h : \mathbb{R}^N \to \mathbb{R}$. We augment the state vector $x \in \mathbb{R}^N$ by $y = h(x)$, i.e. we consider (x, y) and introduce the associated joint PDF

$$\pi_{XY}(x,y) = \pi_X(x|y)\pi_Y(y) .$$

We apply the embedding technique first to y alone, resulting in

$$\frac{\mathrm{d}y}{\mathrm{d}s} = f_y(y,s)$$

with

$$\partial_y(\pi_Y(y) f_y(y)) = \pi_Y(y)(L - \overline{L}) .$$

One then finds an equation in the state variable $x \in \mathbb{R}^N$ from

$$\nabla_x \cdot (\pi_X(x|y) f_x(x,y,s)) + f_y(y,s)\partial_y \pi_X(x|y) = 0$$

and

$$\frac{\mathrm{d}x}{\mathrm{d}s} = f_x(x,y,s) .$$

Next, we introduce the *mean field approximation*

$$\pi_1(x^1|y)\pi_2(x^2|y)\cdots\pi_N(x^N|y) \tag{3.58}$$

for the conditional PDF $\pi_X(x|y)$ with the components of the state vector written as $x = (x^1, x^2, \ldots, x^N)^T \in \mathbb{R}^N$. Under the mean field approximation, the vector field $f_x = (f_{x^1}, f_{x^2}, \ldots, f_{x^N})^T$ can be obtained component-wise by solving scalar equations

$$\partial_z\left(\pi_k(z|y)\, f_{x^k}(z,y)\right) + f_y(y)\,\partial_y \pi_k(z|y) = 0\,, \tag{3.59}$$

$k = 1, \ldots, N$, for $f_{x^k}(z,y)$ with $z = x^k \in \mathbb{R}$. The (two-dimensional) conditional PDFs $\pi_k(x^k|y)$ need to be estimated from the available ensemble members $x_i \in \mathbb{R}^N$ by either using parametric or nonparametric statistics.

We first discuss the case for which both the prior and the posterior distributions are assumed to be Gaussian. In this case, the resulting update equations in $x \in \mathbb{R}^N$ become equivalent to the ensemble transform Kalman–Bucy filter. This can be seen by first noting that the update in a scalar observable $y \in \mathbb{R}$ is

$$\frac{\mathrm{d}y}{\mathrm{d}s} = -\frac{1}{2}\sigma_{yy}^2 R^{-1}\left(y + \overline{y} - 2y_0\right)\,.$$

Furthermore, if the condition PDF $\pi_k(z|y)$, $z = x^k \in \mathbb{R}$, is of the form (3.1), then (3.59) leads to

$$f_{x^k}(x^k, y) = \sigma_{xy}^2 \sigma_{yy}^{-2} f_y(y)\,,$$

which, combined with the approximation (3.58), results in the continuous ensemble transform Kalman–Bucy filter formulation discussed previously.

The rank histogram filter of Anderson [2] corresponds in this continuous embedding formulation to choosing a general PDF $\pi_Y(y)$, while a Gaussian approximation is used for the conditional PDFs $\pi_k(x^k|y)$.

Other ensemble transform filters can be derived by using appropriate approximations to the marginal PDF π_Y and the conditional PDFs $\pi_k(x^k|y)$, $k = 1, \ldots, N$, from the available ensemble members x_i, $i = 1, \ldots, M$.

References

An excellent introduction to filtering and Bayesian data assimilation is [22]. The linear filter theory (Kalman filter) can, for example, be found in [48]. Fundamental issues of data assimilation in a meteorological context are covered in [25]. Ensemble filter techniques and the ensemble Kalman filter are treated in depth in [15]. Sequential Monte Carlo methods are discussed in [3, 4, 13] and by [7, 27] in a geophysical context. See also the recent monograph [19]. The transport view has been proposed in [12] for continuous filter problems and in [43] for intermittent data assimilation. Gaussian mixtures are a special class of nonparametric kernel smoothing techniques which are discussed, for example, in [54].

4 Concluding remarks

We have summarized the Bayesian perspective on sequential data assimilation and filtering in particular. Special emphasis has been put on discussing Bayes' formula in the context of coupling of random variables, which allows for a dynamical system's interpretation of the data assimilation step. Within a Bayesian framework, all variables are treated as random. While this implies an elegant mathematical treatment of data assimilation problems, any Bayesian approach should be treated with caution in the presence of sparse data, high-dimensional model problems, and limited sample sizes. It should be noted in this context that successful assimilation techniques such as 4DVar (not covered in this survey) and the EnKF lead to biased approximations to the state estimation problem. In both cases, the bias is due to the fact that the algorithms are derived under the assumption that the prior distributions are Gaussian. Nevertheless, 4DVar and EnKF often work well in terms of the observed mean squared error (3.10) since the variance of the estimator remains small, even for relatively small ensemble sizes M. On the contrary, asymptotically unbiased Bayesian approaches such as sequential Monte Carlo methods suffer from the curse of dimensionality, generally lead to large variances in the estimators for small M and have therefore not yet found systematic applications in operational forecasting, for example. To overcome this limitation, one could consider more suitable proposal steps such as guided sequential Monte Carlo methods and/or impose certain independence assumptions such as mean field approximations which lead to an improved balance between bias and variance in the mean squared error (3.10). See also the discussion of [20] on the bias-variance trade-off in the context of supervised learning. Promising results for guided particle filters have been reported very recently in [29, 35]. Alternatively, non-Bayesian approaches to data assimilation could be explored in the future, for example, (i) shadowing for partially observed reference solutions, (ii) a nonlinear control approach with transport maps as dynamic feedback laws, and (iii) derivation and analysis of ensemble filter techniques within the framework of stochastic interacting particle systems.

References

[1] J. Amezcua, E. Kalnay, K. Ide, and S. Reich, Using the Kalman–Bucy filter in an ensemble framework, *Q. J. R. Meteorological Soc.*, to appear (2013).

[2] J. L. Anderson, A non-Gaussian ensemble filter update for data assimilation, *Monthly Weather Review* 138 (2010), 4186–4198.

[3] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, *IEEE Trans. Sign. Process.* 50 (2002), 174–188.

[4] A. Bain and D. Crisan, *Fundamentals of stochastic filtering*, Stochastic modelling and applied probability 60, Springer-Verlag, New-York, 2009.

[5] K. Bergemann and S. Reich, A mollified ensemble Kalman filter, *Q. J. R. Meteorological Soc.* 136 (2010), 1636–1643.

[6] K. Bergemann and S. Reich, An ensemble Kalman–Bucy filter for continuous data assimilation, *Meteorolog. Zeitschrift* 21 (2012), 213–219.

[7] M. Bocquet, C. A. Pires, and L. Wu, Beyond Gaussian statistical modeling in geophysical data assimilation, *Mon. Wea. Rev.* 138 (2010), 2997–3022.

[8] Z. Breźniak and T. Zastawniak, *Basic Stochastic Processes*, Springer-Verlag, London, 1999.

[9] G. Burgers, P. J. van Leeuwen, and G. Evensen, On the analysis scheme in the ensemble Kalman filter, *Mon. Wea. Rev.* 126 (1998), 1719–1724.

[10] A. J. Chorin and O. H. Hald, *Stochastic tools in mathematics and science*, 2nd ed, Springer-Verlag, Berlin Heidelberg New York, 2009.

[11] A. J. Chorin, M. Morzfeld, and X. Tu, Implicit filters for data assimilation, *Comm. Appl. Math. Comp. Sc.* 5 (2010), 221–240.

[12] D. Crisan and J. Xiong, Approximate McKean–Vlasov representation for a class of SPDEs, *Stochastics* 82 (2010), 53–68.

[13] A. Doucet, N. de Freitas, and N. Gordon (eds.), *Sequential Monte Carlo methods in practice*, Springer-Verlag, Berlin Heidelberg New York, 2001.

[14] D. B. Duncan and S. D. Horn, Linear dynamic recursive estimation from the viewpoint of regression analysis, *J. American Stat. Association* 67 (1972), 815–821.

[15] G. Evensen, *Data assimilation. The ensemble Kalman filter*, Springer-Verlag, New York, 2006.

[16] M. Frei and H. R. Künsch, Mixture ensemble Kalman filters, *Computational Statistics and Data Analysis* 58 (2011), 127–138.

[17] C. W. Gardiner, *Handbook on stochastic methods*, 3rd ed, Springer-Verlag, 2004.

[18] G. H. Golub and Ch. F. Van Loan, *Matrix computations*, 3rd ed, The Johns Hopkins University Press, Baltimore, 1996.

[19] J. Harlim and A. Majda, *Filtering Complex Turbulent Systems*, Cambridge University Press, Cambridge, 2012.

[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed, Springer-Verlag, New York, 2009.

[21] D. J. Higham, An algorithmic introduction to numerical simulation of stochastic differential equations, *SIAM Review* 43 (2001), 525–546.

[22] A. H. Jazwinski, *Stochastic processes and filtering theory*, Academic Press, New York, 1970.

[23] Simon J. Julier and Jeffrey K. Uhlmann, A New Extension of the Kalman Filter to Nonlinear Systems, in: *Signal processing, sensor fusion, and target recognition. Conference No. 6*, 3068, pp. 182–193, Orlando FL, 1997.

[24] J. Kaipio and E. Somersalo, *Statistical and computational inverse problems*, Springer-Verlag, New York, 2005.

[25] E. Kalnay, *Atmospheric modeling, data assimilation and predictability*, Cambridge University Press, 2002.

[26] P. E. Kloeden and E. Platen, *Numerical solution of stochastic differential equations*, Springer-Verlag, Berlin Heidelberg New York, 1992.

[27] P. J. Van Leeuwen, Particle filtering in geophysical systems, *Monthly Weather Review* 137 (2009), 4089–4114.

[28] P. J. Van Leeuwen, Nonlinear data assimilation in the geosciences: an extremely efficient particle filter, *Q.J.R. Meteorolog. Soc.* 136 (2010), 1991–1996.

[29] P. J. Van Leeuwen and M. Ades, Efficient fully nonlinear data assimilation for geophysical fluid dynamics, *Computers and Geosciences* 55 (2013), 16–27.

[30] J. Lei and P. Bickel, A moment matching ensemble filter for nonlinear and non-Gaussian data assimilation, *Mon. Weath. Rev.* 139 (2011), 3964–3973.

[31] J.M Lewis, S. Lakshmivarahan, and S. K. Dhall, *Dynamic data assimilation: A least squares approach*, Cambridge University Press, Cambridge, 2006.

[32] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag, New York, 2001.

[33] R. J. McCann, Existence and uniqueness of monotone measure-preserving maps, *Duke Mathematical Journal* 80 (1995), 309–323.

[34] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*, Springer-Verlag, London New York, 1993.

[35] M. Morzfeld and A. J. Chorin, Implicit particle filtering for models with partial noise and an application to geomagnetic data assimilation, *Nonlinear Processes in Geophysics* 19 (2012), 365–382.

[36] M. Morzfeld, X. Tu, E. Atkins, and A. J. Chorin, A random map implementation of implicit filters, *J. Comput. Phys.* 231 (2012), 2049–2066.

[37] T. A. El Moselhy and Y. M. Marzouk, Bayesian inference with optimal maps, *J. Comput. Phys.* 231 (2012), 7815–7850.

[38] J. Moser, On the volume elements on a manifold, *Trans. Amer. Math. Soc.* 120 (1965), 286–294.

[39] R. M. Neal, *Bayesian learning for neural networks*, Springer-Verlag, New York, 1996.

[40] B. Øksendal, *Stochastic Differential Equations*, 5th ed, Springer-Verlag, Berlin-Heidelberg, 2000.

[41] I. Olkin and F. Pukelsheim, The distance between two random vectors with given dispersion matrices, *Linear Algebra and its Applications* 48 (1982), 257–263.

[42] F. Otto, The geometry of dissipative evolution equations: the porous medium equation, *Comm. Part. Diff. Eqs.* 26 (2001), 101–174.

[43] S. Reich, A dynamical systems framework for intermittent data assimilation, *BIT Numer Math* 51 (2011), 235–249.

[44] S. Reich, A Gaussian mixture ensemble transform filter, *Q. J. R. Meterolog. Soc.* 138 (2012), 222–233.

[45] S. Reich, A guided sequential Monte Carlo method for the assimilation of data into stochastic dynamical systems, *Recent Trends in Dynamical Systems*, to appear (2013).

[46] S. Reich, A non-parametric ensemble transform method for Bayesian inference, *SIAM J. Sci. Comp.*, to appear (2013).

[47] G. O. Roberts and R. L. Tweedie, Exponential convergence of Langevin distributions and their discrete approximations, *Bernoulli* 2 (1996), 341–363.

[48] D. J. Simon, *Optimal state estimation*, John Wiley & Sons, Inc., New York, 2006.

[49] K. W. Smith, Cluster ensemble Kalman filter, *Tellus* 59A (2007), 749–757.

[50] A. S. Stordal, H. A. Karlsen, G. Nævdal, H. J. Skaug, and B. Vallés, Bridging the ensemble Kalman filter and particle filters: the adaptive Gaussian mixture filter, *Comput. Geosci.* 15 (2011), 293–305.

[51] A. M. Stuart, *Inverse problems: a Bayesian perspective*, Acta Numerica, 17, Cambridge University Press, Cambridge, 2010.

[52] C. Villani, *Topics in Optimal Transportation*, American Mathematical Society, Providence, Rhode Island, NY, 2003.

[53] C. Villani, *Optimal transportation: Old and new*, Springer-Verlag, Berlin Heidelberg, 2009.

[54] M. P. Wand and M. C. Jones, *Kernel smoothing*, Chapmann and Hall, London, 1995.

[55] X. Xiong, I. M. Navon, and B. Uzungoglu, A note on the particle filter with posterior Gaussian resampling, *Tellus* 85A (2006), 456–460.

Martin Burger, Hendrik Dirks and Jahn Müller

Inverse problems in imaging

Abstract: This chapter provides an overview of inverse problems in imaging, with a particular focus on biomedical imaging applications and current developments. We discuss some basics in the mathematical modeling of images, image reconstruction, and imaging devices. Then, we proceed to three topics of high current interest, namely, problems with missing data as appearing in inpainting or imaging from surface measurements, nonlinear inverse problems created by the need to perform additional calibrations, and finally high-dimensional inverse problems in dynamic imaging.

Keywords: Inverse problems, imaging, image reconstruction, inpainting, blind deconvolution, dynamic imaging

2010 Mathematics Subject Classification: 65N21, 35R20, 92C55, 65M32

Martin Burger: Institute for Computational and Applied Mathematics, University of Münster, Einsteinstrasse 62, 48149 Münster, Germany, martin.burger@wwu.de
Hendrik Dirks: Institute for Computational and Applied Mathematics, University of Münster, Einsteinstrasse 62, 48149 Münster, Germany, hendrik.dirks@uni-muenster.de
Jahn Müller: Institute for Computational and Applied Mathematics, University of Münster, Einsteinstrasse 62, 48149 Münster, Germany, jahn.mueller@uni-muenster.de

Nowadays, life is hard to imagine without the use of images and videos, with increasing fraction in digital format. While humans conveyed a huge amount of information via audio and audio-type signals (speeches, telephone, telegraphs, radio) one-hundred years ago, we increasingly use image and video-based methods now (television, computers, internet). This also applies to many other parts of daily life, engineering, medicine, and science. As examples, consider the transition from stethoscopes to modern medical imaging devices or from ground based meteorological stations to satellite-based weather surveillance.

The increasing use of images and videos has led to a new branch of science, often called *imaging science*, where mathematical methods play an important role. Inverse problems are an important part in this area since they arise at two fundamental points:

The work of the authors has been supported by the German Research Foundation DFG through grants SFB 656 / B2, BU 2327/2-1, BU 2327/6-1, BU 2327/8-1. The authors thank Michael Fieseler and Thomas Kösters (EIMI, Münster) for providing simulated data, Otmar Schober (Nuclear medicine, Münster) and Klaus Schäfers (EIMI, Münster) for permission to use PET data, and Eric Schulze-Bahr (Cardiology, Münster) for permission to use cardiac imaging data.

- *The way to the image*: Most measurement devices are not able to automatically deliver high-quality images, but rather raw data from which images need to be reconstructed. Image reconstruction is a classical inverse problem, particularly tomographic setups are currently widely studied with different mathematical techniques.
- *The (quantitative) interpretation of the image*: In many cases, images and videos are of interest due to the quantitative information they carry. In order to benefit from the latter, an appropriate link to mathematical models is needed which can also be cast in the framework of inverse problems.

We will provide a brief overview of the mathematical issues arising from these two questions in this chapter with a focus on recent developments and open questions. Our aim is by no means to give an extensive overview of mathematical techniques nor of imaging devices and problems. Rather, we focus on variational methods which allow for a unified treatment of large classes of problems and build on a sound mathematical background, and on certain classes of problems that we think can highly benefit from further development in inverse problems. We start with the basic mathematical modeling of images and their properties in Section 1 and then proceed to some examples of imaging devices and related mathematical models in order to further motivate the subsequent investigations. In the subsequent Section 3, we review some classical mathematical problems in image reconstruction. Afterwards, we turn to three areas of high current interest: The first are effectively underdetermined problems discussed in Section 4, where the appropriate incorporation of prior information becomes of ultimate importance. The second are problems usually arising in fast measurements, when the system parameters cannot be well calibrated and need to be estimated together with the image, as we will highlight in Section 5. Finally, we will discuss dynamic problems in imaging, i.e. related to videos, and the link to mathematical models for the dynamics in Section 6.

1 Mathematical models for images

Images can be modeled as densities or intensities (gray values) on an image domain $\Omega \subset \mathbb{R}^d$, which means the image $u : \Omega \to \mathbb{R}$ is simply a nonnegative function. Frequently, the density is not *a priori* a distribution of gray values, but directly a density of some physical quantity carrying quantitative information, e.g. tracer substances in medical imaging. Thus, inverse problems dealing with images as unknowns can directly be related to the bulk of other inverse problems dealing with reconstructing functions. A major difference to the majority of such inverse problems is that images are not expected to be smooth functions, but have specific structures of particular importance:

- *Edges:* A highly important part of images are edges, which are related mainly to discontinuities in the function u. The edges and the *cartoon*, i.e. a piecewise constant approximation of the image between the edges, are often the first quantity of interest in the interpretation of the image. Hence, it is of high importance that solution methods for inverse problems do not destroy edges.
- *Textures:* In natural images, these are small scale patterns, i.e. locally structured high-frequency information. Different patterns are usually separated by the edges, and thus the interplay with edges is crucial. Since the high-frequency information is highly damped by typical forward operators, it is often out of reach to reconstruct textures in inverse problems.
- *Morphology:* Images are frequently interpreted in a morphological way, i.e. the exact gray values are of limited interest, but rather the isocontours or the level sets of the image provide the relevant information.

It has become a standard setting to consider cartoon images as functions of bounded variation $u \in BV(\Omega)$. Assuming a normalization of the image such as

$$\int_\Omega u\,dx = 1, \tag{4.1}$$

the cartoon should be characterized by a rather small total variation

$$TV(u) = \sup_{g \in C_0^\infty(\Omega;\mathbb{R}^d), \|g\|_{L^\infty} \leq 1} \int_\Omega u \nabla \cdot g\,dx\,. \tag{4.2}$$

Note that for a piecewise constant function u, i.e. an image consisting of regions with homogeneous gray values separated by sharp edges, the total variation is just the perimeter of the jump set weighted by the jump height. Another description of the cartoon comes from the work of Mumford and Shah [71], and was originally designed for image segmentation. Their description consists of an edge set $\Gamma \subset \Omega$ and a smooth component $u \in H^1(\Omega \setminus \Gamma)$. The corresponding functional that is thought to be small for good cartoon images is of the form

$$J_{MS}(u,\Gamma) = \int_{\Omega\setminus\Gamma} |\nabla u|^2\,dx + \mathcal{H}^{d-1}(\Gamma)\,, \tag{4.3}$$

where $\mathcal{H}^{d-1}$ denotes the $d-1$-dimensional Hausdorff-measure.

The texture part is more difficult to characterize, as it is usually attributed to oscillatory parts in the image, but consequently difficult to separate from potential noise. In analogy to the cartoon part, Meyer [66] proposed a dual approach and tried to characterize texture as parts v with $\|v\|_*$ rather small, where for a distribution $v \in BV(\Omega)^*$ with zero mean,

$$\|v\|_* = \sup_{\varphi \in BV(\Omega), TV(\varphi) \leq 1} \langle v, \varphi \rangle\,. \tag{4.4}$$

Other approaches are based on representations in negative Sobolev spaces ([78, 98]), nonlocal versions of total variation or Sobolev norms exploiting similarities of patches, which will be discussed below. While image decomposition into structure and texture is a highly relevant problem in many parts of image processing, it is often of less importance for inverse problems in imaging. The main reason is a smoothing property of the forward operators which are usually strongly damping the high-frequency components. Hence, in reconstructions of images, the focus is laid on the cartoon parts, which is also a reason why total variation is a very popular penalty in variational regularization methods (cf. [22] for a detailed discussion of total variation reconstruction methods).

Several bases or frames such as wavelets, curvelets, or shearlets have been proposed to efficiently represent images. They are based on multiscale decompositions, usually in a dyadic rescaling of space. ℓ^1-norms on the coefficients of such systems, in particular on wavelet coefficients, induce norms on Besov spaces. A particularly well studied case is the Besov space $B^1_{1,1}$, which is quite close to the space of functions of bounded variation. Also, the wavelet approximation of total variation functionals has been frequently studied (cf., e.g. [24, 34]).

A strong recent trend are nonlocal approaches for images motivated by the nonlocal filter introduced by Buades and coworkers [19]. Roughly speaking, the idea is to interpret an image not as a collection of single gray values, but as a collection of local patches. A corresponding continuum model is to consider the image as a function U on $\Omega \times \Sigma$, where Σ is a small neighborhood of the origin modeling a patch. The consistency is obtained by $U(x, y) = U(x + y, 0)$ for all $y \in \Sigma$. From this space of patch-functions, a set of weights $w(x, \xi)$ for $x, \xi \in \Omega$ is computed by comparing patches, i.e. the functions $U(x, \cdot)$ and $U(\xi, \cdot)$. This yields a weighted graph structure on the image which can be further analyzed ([19, 59, 85, 88]). One option is to use discrete calculus on graphs to define analogues of total variation or other functionals for these patch-functions ([48, 59, 85]). In particular, for natural images, such approaches yield superior results in many tasks such as denoising since one can exploit that similar patches appear several times within the image, e.g. in textures.

In most areas of imaging, in particular those related to inverse problems where one does not just play with given images, variational approaches (respectively Bayesian methods with particular focus on MAP estimation) have become a standard tool. There are two natural functionals involved, namely, the fidelity $D(u, f)$ (which can be interpreted as the negative log-likelihood of obtaining the data f conditioned on the image u) and the regularization functional $R(u)$. A standard solution approach is the minimization of the energy functional

$$E(u) = \lambda D(u, f) + R(u)\,, \tag{4.5}$$

with a weighting parameter $\lambda > 0$. Clearly, such approaches are an equivalent formulation to Tikhonov-type regularization in inverse problems, with regularization parameter $\alpha = \frac{1}{\lambda}$. We will discuss the detailed modeling of prior information and the relation to Bayesian models below.

2 Examples of imaging devices

In the following, we give a brief overview of the most frequently used types of devices for acquiring imaging data. We focus on the basic structures and implications for mathematical modeling and inverse problems rather than on the detailed device physics and specific application context.

2.1 Optical imaging

Optical imaging based on recording photons (usually found in CCD devices nowadays) is probably the most intuitive way of obtaining image data used in various digital camera systems (from microscopes over hand-held systems, high level movie cameras up to astronomical telescopes). In this case, one can interpret the recording directly as an image via translating the number of photons (possibly at different wavelengths) into a grayscale (or a color scale). Effects to be taken into account in forward models are certain factors that can lead to a convolution (e.g. defocus) or make it necessary to investigate the structure of the noise (e.g. low light intensity).

Active research in optical imaging is still related to denoising and deconvolution, also in the version of blind deconvolution as we shall discuss below. Since, in many cases, the recorded images or image sequences themselves are of reasonable quality, most research is rather related to processing digital images and videos. Another quite active field is to acquire three-dimensional information from stereo or other multi-camera systems.

The applications of optical imaging are ubiquitous as digital images and videos are part of almost everyone's daily life in the modern world. In addition to usual optical frequencies, an increasing number of devices use other or larger parts of the frequency band of electromagnetic waves. In particular, multi- and hyperspectral imaging is a strong trend since it can provide much better information than just the usual three primary colors we can distinguish.

2.2 Transmission tomography

In transmission tomography, rays (X-rays, electrons) are sent through the object from different positions and their attenuation is recorded on the opposite side. The classical forward model is the Radon transform, i.e. the line integrals of the object density, since the attenuation is proportional to the density along the line. The principle of

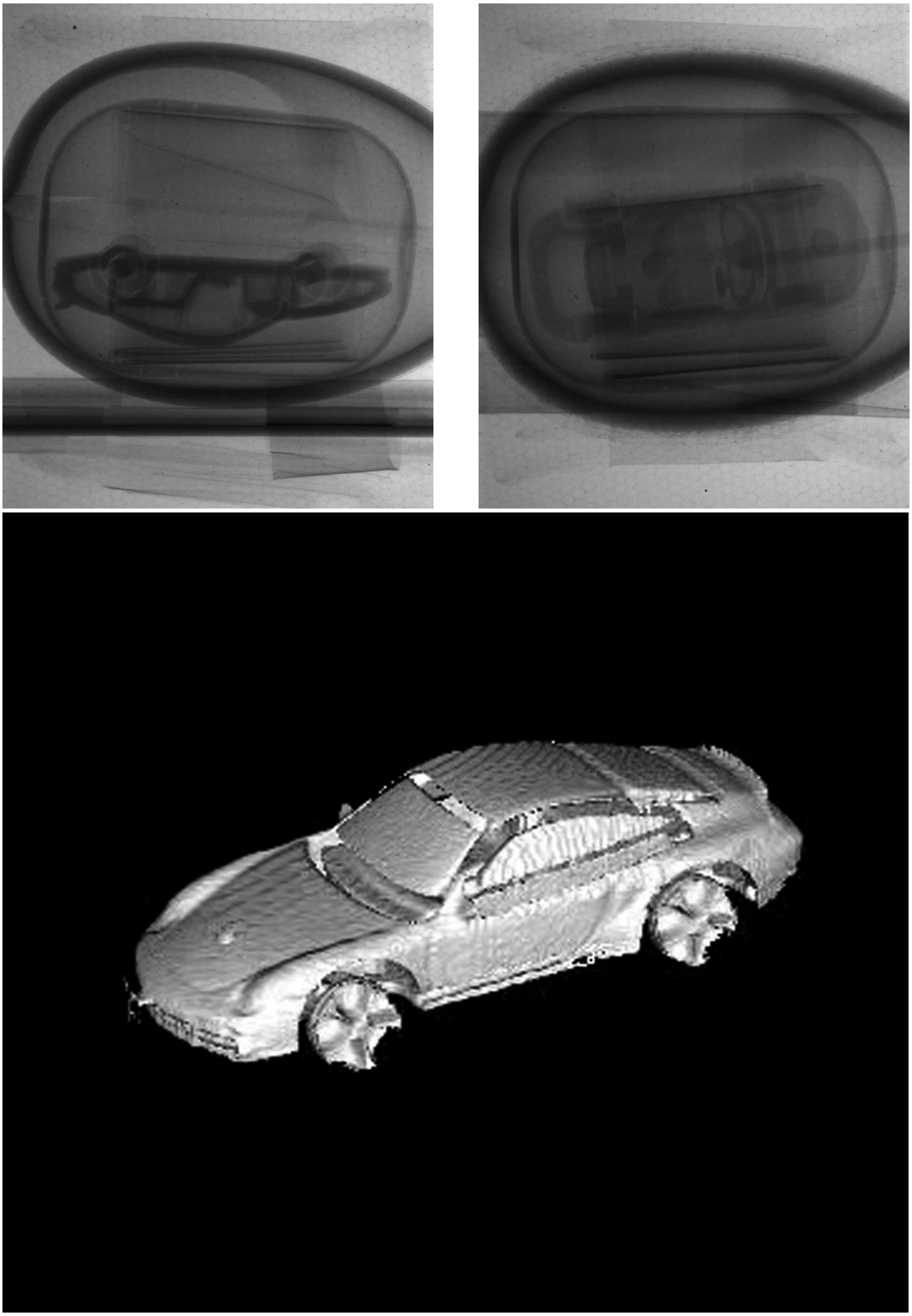


Figure 4.1: Illustration of transmission tomography: Micro-CT imaging of a kid toy. Top row: two projections from different angles. Bottom row: 3D Image reconstruction (threshold segmented). Data courtesy of European Institute for Molecular Imaging and SFB 656, Münster.

transmission tomography is illustrated by micro-CT data in Figure 4.1, with two projections from different sides and the final 3D image reconstruction.

Tomography is a very well studied mathematical topic (cf., e.g. [73, 74]). Current challenges are related to exact reconstruction formulas for three-dimensional scanning geometries and problems with limited data, e.g. limited angles which are severely ill-posed in contrast to the full data case.

Major applications of transmission tomography are in medicine and material testing. In physics and biology, there is increasing interest in electron tomography for visualizing three-dimensional structures at the nanoscale ([43, 60]).

2.3 Emission tomography

Emission tomography such as positron emission tomography (PET) and single photon emission computed tomography (SPECT) are based on recording photons emitted in the case of radioactive decay of some tracer inside the body. Since the radioactive decay is random, the forward models for emission tomography naturally need to be of stochastic nature. In PET, one uses tracers emitting photons to the opposite direction, and thus to each recorded coincidence of photons, one can attribute a decay event on the line in between. The tracers used in SPECT only emit a single photon in random direction and one uses collimators to get information about the direction, that is, the line on which the decay event has taken place. Since clearly the probability of a decay event is proportional to the tracer density along the corresponding interval, one obtains a stochastic sampling of the Radon transform in both cases. It is quite standard to use the Poisson distribution as a model for the randomness of the decay. Subtle

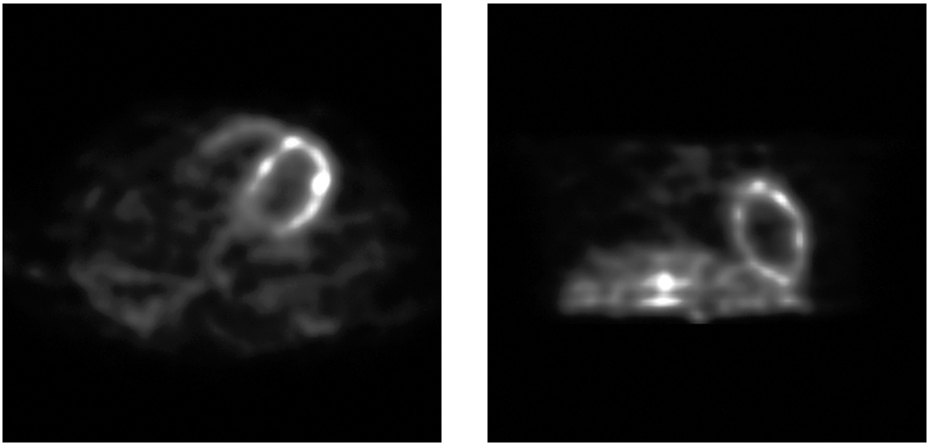


Figure 4.2: Illustration of emission tomography: Reconstruction of a cardiac PET scan in two different slice views. Data courtesy of SFB 656, Münster (subproject C1).

differences between PET and SPECT are in the detailed modeling of attenuation. We will come back to the issues arising in SPECT below.

The major application of emission tomography is nuclear medicine. Due to their potential of monitoring time-dependent physiological processes (in a quite specific way when using appropriate tracers), these techniques have received increasing attention within *in vivo* imaging. The trade-off between spatial resolution and specificity in emission tomography is illustrated in Figure 4.2 via a cardiac scan clearly displaying the concentration of the tracer in the ventricles.

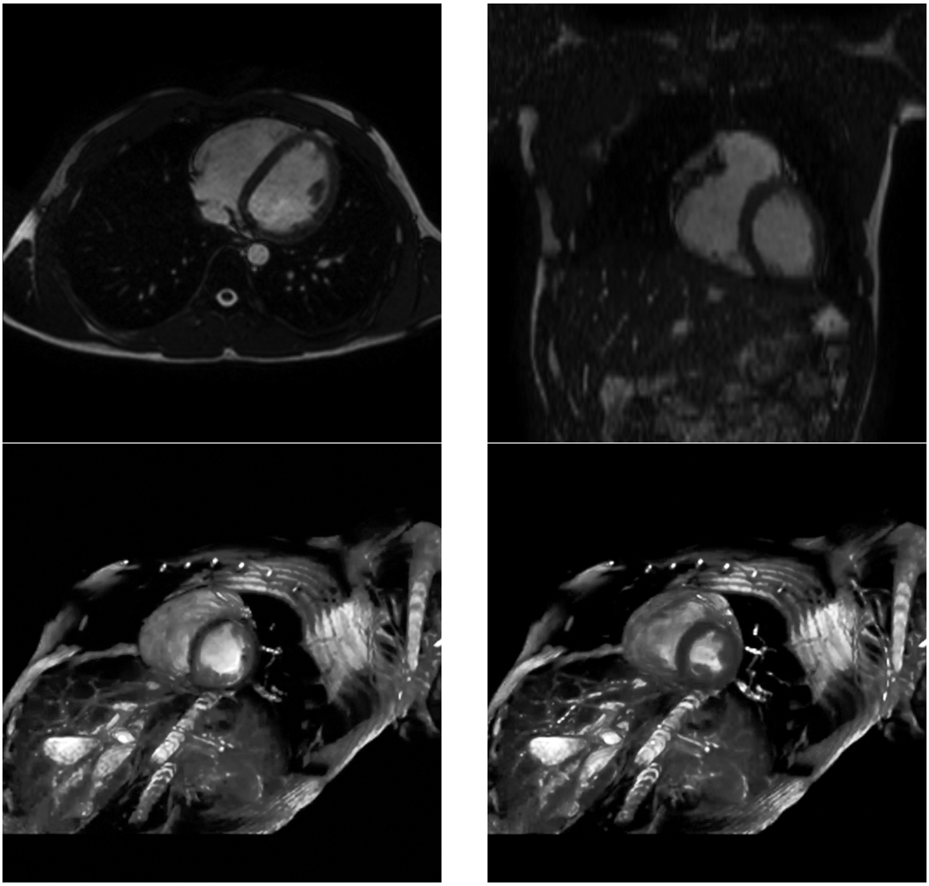


Figure 4.3: Illustration of MR imaging: Reconstruction of a cardiac MR scan of the same subject as in Figure 4.2, in two different slice views (top) and 3D visualization (bottom) of two different time frames. Data courtesy of SFB 656, Münster (subproject C1).

2.4 MR imaging

Magnetic Resonance (MR) imaging is probably the technique with the most complex physics, though on the other hand, it yields the least ill-posed reconstruction problem. The principle of MR is to induce nuclear magnetic relaxation of protons or certain molecules by applying magnetic fields. The measurements consist of electrical signals at the same frequency.

The usual forward model in MR is related to the Fourier transform, and hence the inverse problem of reconstructing the image from full data is well-posed. However, the acquisition in slices is rather time-consuming, and thus a main challenge is to make MR faster in order to obtain high resolution images with increasing time resolution. The current image quality in MR is illustrated by a cardiac MR scan in Figure 4.3.

The major application of MR is nowadays a medical one. Due to the absence of radiation exposure compared to X-ray or radioactivity based techniques, MR can be frequently used for various tasks. Functional versions of MR scans play a prominent (and recently strongly debated) role in neurosciences.

2.5 Acoustic imaging

In acoustic imaging, like in ultrasound or seismic data acquisition, an acoustic wave is usually sent into the body or earth, and the echo is recorded on the surface. The natural forward model is the wave equation with a spatially varying wave speed, and the inverse problem is to reconstruct the wave speed. Mainly for computational reasons, frequency domain formulations or several approximations (e.g. Eikonal equations or first arrival data) of the wave equation have been used in the past.

While inversion is quite frequently used in geophysics, it is used less in medical ultrasound since one can usually interpret the data directly. In the latter, image pro-

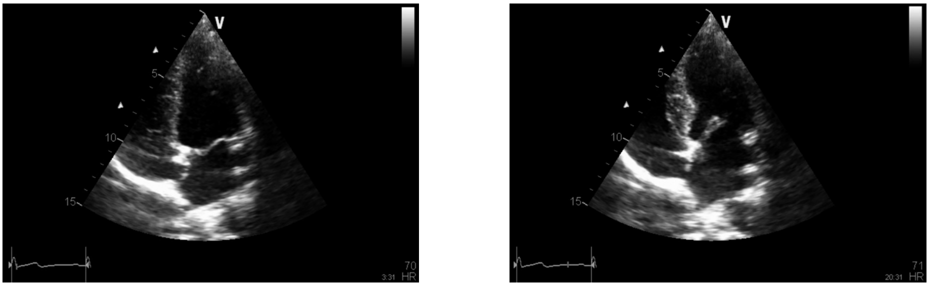


Figure 4.4: Illustration of acoustic imaging: Cardiac echo scan of the same subject as in Figure 4.2. Left: closed mitral valve during systolic phase. Right: opened mitral valve during diastolic phase. Data courtesy of SFB 656, Münster (subproject C1).

cessing and automatic image analysis techniques are of interest, particularly for dealing with the large speckle noise artifacts appearing in such image sequences ([93]). A lot of recent interest in inversion has been in the novel hybrid technique of photoacoustic imaging, where the acoustic wave is modulated by an optical laser ([100]).

A major advantage of ultrasound imaging is the inherently high time resolution which allows one to monitor relevant processes such as, e.g. heartbeat. Together with speckle artifacts, this is illustrated in Figure 4.4.

2.6 Electromagnetic imaging

In electromagnetic imaging, one records electrical potentials created from currents inside or on the boundary of the object in arrays of electrodes around the object surfaces or magnetometers in small distance to the surface. The forward models are clearly the Maxwell equations or in many cases, reductions to the Poisson equation (and the Biot–Savart law for the magnetic field). The term imaging in such applications is often debated since the forward problems are severely ill-posed and the reconstructions are hence of limited quality and mainly restricted to low frequency components.

Due to the high remaining challenges in electromagnetic imaging, this is a very active field of research in applied mathematics, particularly in inverse problems. In pure surface imaging of electrical activity created inside the object, a major challenge is the appropriate modeling of prior information to decrease or eliminate the nonuniqueness of the reconstruction problem. The technology of electrical impedance tomography ([33]), where different currents between the electrodes are sent into the body and the resulting potentials are measured, received enormous attention in inverse problems and is also known as the Calderon problem ([25]). The inversion can be formulated as reconstructing the conductivity in the Poisson equation from the knowledge of the Dirichlet-to-Neumann (or Neumann-to-Dirichlet) map, which has a rich mathematical structure.

Besides material testing and geology, applications of electromagnetic imaging have been found recently in medicine, e.g. in brain (EEG/MEG), heart (ECG/MCG) or muscle studies (EMG). The technique of EIT is mainly applied in material testing and in monitoring lung activity. Also, hybrid techniques find increasing attention ([4, 64]).

3 Basic image reconstruction

The first fundamental step is the reconstruction of images from raw data. The most prominent image reconstruction problem nowadays is related to X-ray tomography, which is based on inverting the Radon transform as we shall recall below. In optical imaging devices like photography, telescopes, or microscopy, one obtains an image directly, but it quite frequently suffers from defects or does not yet have the desired resolution. Here, the reconstruction step can also be interpreted as a correction step, e.g. of defocus or atmospheric blur.

In a canonical mathematical formulation, classical image reconstruction can be formulated as the solution of a linear operator equation

$$Ku = f\,, \tag{4.6}$$

with given (noisy) data f and a usually compact forward operator K. Due to the non-closed range of compact operators, most image reconstruction problems become ill-posed problems in the sense of Hadamard ([42]). We shall discuss some standard issues in the following.

3.1 Deblurring and point spread functions

A standard problem in imaging is blur, e.g. caused by lack of focus or motion. The mathematical model for blurring is an integral operator of the form

$$Ku(x) = \int_{\Omega} k(x-y,y)u(y)dy\,, \tag{4.7}$$

where often the Point Spread Function (PSF) k is approximated as spatially independent, i.e. k only depends on the first variable.

In many cases, blur can be approximated well by a Gaussian due to the following two reasons: On the one hand, blur is caused by diffusion-type processes, and the solution of the diffusion equation is just a convolution with a Gaussian. On the other hand, blur is sometimes caused by repeated random processes, and the central limit theorem again leads to a Gaussian PSF. For these reasons, Gaussians are also routinely used as PSFs in many tests of reconstruction algorithms and as first approximations for many devices. In such cases, only the variance of the Gaussian (often translated into the *full-width-at-half-maximum*) has to be determined, which is frequently possible using phantom measurements.

Another recent trend for many imaging devices is an experimental determination of the PSF. In such tests, very small objects (i.e. images u_z with very small support around a point z) are used, and since these approximate a Dirac-delta at z under appropriate rescaling, one obtains via

$$cKu_z(x) \approx \int_{\Omega} k(x-y,y)\delta(z-y)\,dy = k(x-z,z) \tag{4.8}$$

an approximate read-out of the PSF from the corresponding measurements. Although this is a purely experimental procedure, it creates an interesting mathematical problem due to the fact that such sources cannot be placed at an arbitrary number of positions in the device due to costs, time consumption, limited precision in placing the sources, or other issues. In practice, one obtains a rather sparse sampling of the PSF and thus, the problem of PSF interpolation occurs ([7, 11, 14, 46, 62, 63, 101, 106]).

3.2 Noise

The modeling of noise in imaging is an interesting issue and taking into account the statistics of noise can indeed yield significantly improved reconstructions in many cases. In some inherently stochastic forward problems such as emission tomography, where photons are created by random radioactive decay, the modeling of noise has a rather long tradition (cf., e.g. [97]). In other problems, noise modeling and its use in reconstruction algorithms has become a very active field of research in the last years (cf., e.g. [10, 93]). In particular, the form of the noise has consequences for the form of the data likelihood and thus on the appropriate modeling of variational or iterative reconstruction methods.

A frequently used standard model for the noise is additive Gaussian noise, i.e.

$$f|_D = g|_D + \sigma\eta_D \tag{4.9}$$

for each detector D with $\sigma > 0$ and independent normally distributed η_D. Clearly, this yields a Gaussian distribution of the noise, and the corresponding negative log-likelihood is of the form

$$L_d(Ku|f) = \frac{1}{2\sigma^2}\sum_D (f|_D - Ku|_D^2)\,. \tag{4.10}$$

Asymptotically, the negative log-likelihood converges to the squared L^2-norm

$$D(u,f) = L(Ku|f) = \frac{1}{2\sigma^2}\int (Ku-f)^2\,dx\,. \tag{4.11}$$

In imaging devices based on photon counts, different noise statistics are in place. A standard model is a Poisson distribution for the counts, i.e. the number of counts per detector D is a Poisson-distributed random variable with mean value $Ku|_D$. Here (by adding terms independent of u), the data term can be written as the Kullback–Leibler divergence

$$D(u,f) = \frac{1}{2}\int [f\log\frac{f}{Ku} - f + Ku]\,dx\,. \tag{4.12}$$

In the case of good statistics, i.e. a high number intensity, the Poisson distribution can be approximated via a Gauss distribution with the same mean and variance (both equal to Ku in the Poisson model). Thus, one obtains

$$D_{G1}(u,f) = \frac{1}{2}\int \frac{(Ku-f)^2}{Ku}\,dx\,. \tag{4.13}$$

Since this model is convex but still nonquadratic in u, frequently a further approximation based on the reasoning $Ku \approx f$ for the denominator is used, i.e.

$$D_{G2}(u,f) = \frac{1}{2}\int \frac{(Ku-f)^2}{f}\,dx\,, \tag{4.14}$$

which can also be interpreted as a second-order Taylor expansion of the Kullback–Leibler divergence in Ku around f.

Recently, a variety of different noise models have been investigated. This concerns variants of the salt-and-pepper noise ([32]), other multiplicative models (cf., e.g. [89]), and Rayleigh-type distributions for modeling speckle noise as, e.g. appearing in ultrasound ([55, 99]).

3.3 Reconstruction methods

Various reconstruction methods have been proposed over the last decades for different tasks of imaging. There is a first distinction between direct and iterative reconstruction methods. Direct reconstructions rely on exact formulas for the inverse operator of K and a numerical implementation of these. A standard example is the Radon transform, which can be inverted exactly using the Fourier transform and efficient numerical implementations can be obtained using FFT techniques. Due to the ill-posedness of the inverse problem, it does usually not work to directly use the noisy data in the inversion, but filtering has to be used before. This leads to

$$u = K^{-1}F_\alpha(f)\,, \tag{4.15}$$

where F_α is a filtering operation with parameter α (which is a regularization parameter for the inverse problem in the sense of mollification methods, cf. [72]). Currently, linear filters are mainly used, which are easy to implement and analyze, though in principle, one can think of using nonlinear filters as well.

Iterative reconstruction methods are usually based on a variational formulation. In the unregularized case, i.e. for the minimization of the functional $u \mapsto D(u,f)$, one uses appropriate early termination of iterations to receive optimal results. Examples are the simple descent method

$$u^{k+1} = u^k - \tau\partial_u D(u^k,f)\,, \tag{4.16}$$

and for nonnegative image restoration, positivity preserving schemes like the EM-type algorithm

$$u^{k+1} = u^k - \tau u^k\partial_u D(u^k,f) \tag{4.17}$$

for appropriate τ, which is specifically used for Poisson noise models as

$$u^{k+1} = u^k K^*\left(\frac{f}{Ku^k}\right)\,. \tag{4.18}$$

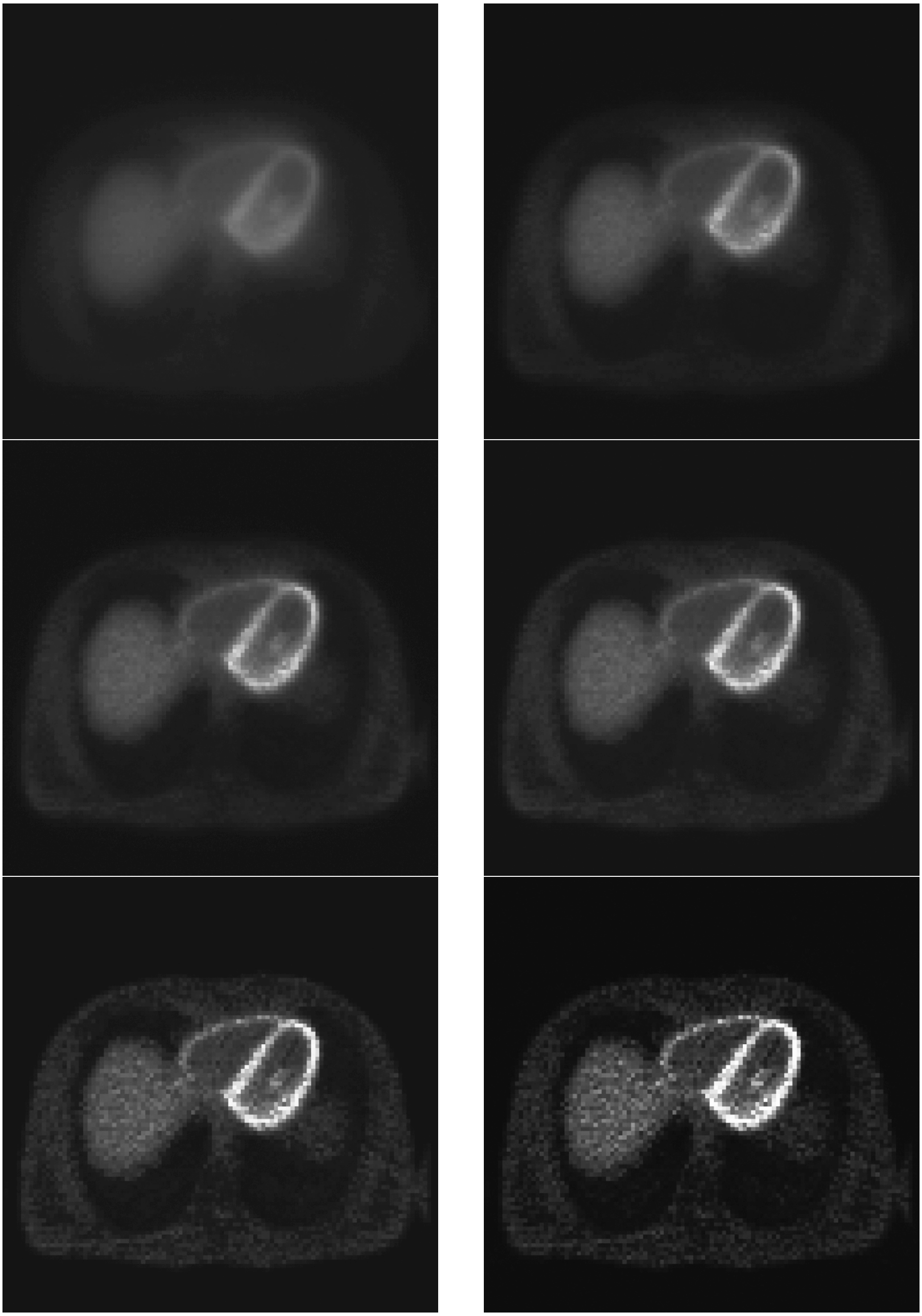


Figure 4.5: Effect of noise and iteration number in iterative regularization methods illustrated by EM iterations on the cardiac PET software Phantom XCAT (simulated data courtesy of European Institute for Molecular Imaging, Münster). First row: 5 Iterations (left) and 10 iterations (right). Second row: 15 Iterations (left) and 20 iterations (right). Third row: 60 Iterations (left) and 100 iterations (right).

Figure 4.5 is illustrating the so-called semiconvergence properties of such iterative methods, for example, here EM for Poisson noise. The iterates first approach a good reconstruction, but with too many iterations, noise effects enter the reconstruction again and distorts the image quality. In the case of additional regularization, one minimizes a functional $D(u, f) + R(u)$ instead, where R is a regularization functional. In the following, we shall always consider cases with appropriate regularization R, which prevents the instability in case of noisy data. The properties of the noise are encoded in the specific form of D (as discussed in the previous section) and the weighting between data fidelity and regularization. Depending on the specific form of R, various iterative methods to compute a minimizer have been proposed, e.g. based on splitting and augmented Lagrangian methods ([23, 35, 49, 104, 107]).

4 Missing data and prior information

A popular trend in recent years is to consider image reconstruction with missing data which is related to the main line of research in compressed sensing ([38, 41]). Ill-posedness in such problems is created in the sense of nonuniqueness of the solution in the inverse problem rather than by instability. The key idea to obtain good solutions to the inverse problem is to incorporate prior information. Again, variational methods are a standard approach and we shall discuss some favorable properties as well as some current limitations.

4.1 Prior information

Several kinds of prior information have been used to improve image reconstruction and, in particular, to enable meaningful reconstruction also with missing data. Nowadays, the standard way of modeling prior information is (at least at a formal level) Bayesian modeling. Its basis is Bayes' theorem, which yields the posterior probability density of u being the underlying image given the data f as

$$p(u|f) = \frac{p(f|u)p(u)}{p(f)} . \tag{4.19}$$

Here, $p(f|u)$ is the likelihood of the data given the image u, and $p(u)$ respectively $p(f)$ are prior probabilities for the image and data. Since f is fixed, the latter is only a scaling factor and of no particular importance. The interesting part is the prior probability of the image, which can encode relevant prior information.

Several estimates can be obtained from the posterior distribution. The most frequently used and most straightforward one to compute is the Maximum *a posteriori* Probability (MAP) estimate given by

$$\hat{u} = \arg\max_u p(u|f) . \tag{4.20}$$

Using the equivalent minimization of the negative logarithm, we find that MAP estimation can be formulated as Tikhonov-type regularization of an inverse problem, namely,

$$\hat{u} \in \arg\min_u \left(D(u,f) + R(u)\right), \tag{4.21}$$

where $R(u) = -\log p(u)$ takes the role of the regularization functional.

While Gaussian priors, respectively quadratic regularization functionals, were very popular for many years due to their computational and analytical simplicity, a different paradigm has evolved, particularly in the last decades. In many instances, it was found that ℓ^1-type regularization functionals, i.e. Laplacian prior distributions, yield superior properties. The first such approach was the ROF model for image denoising ([83]), which used the total variation as a regularization. Total variation has a nice geometric meaning via the coarea formula. For sufficiently regular BV functions ([44]), we have

$$TV(u) = \int_{\mathbb{R}} \mathcal{H}^{d-1}(\partial\{u < \alpha\})\, d\alpha, \tag{4.22}$$

where $\mathcal{H}^{d-1}$ denotes the $d-1$-dimensional Hausdorff measure. Hence, the total variation penalizes the surface area of the level sets of the image. One observes that this is also true for functions u with discontinuities, as long as the discontinuity set is Hausdorff-measurable. The latter is a key property of total variation models. While regularizations based on standard Sobolev-type norms do not allow one to obtain reconstructions with discontinuities, i.e. images with edges, the total variation model can realize reconstructions with realistic edges.

A popular alternative to MAP estimates are conditional mean (CM) estimates given by

$$\tilde{u} = \int u\, p(u|f)\, du\,. \tag{4.23}$$

A major difficulty for CM estimates is their reasonably efficient computation since a high-dimensional integration problem needs to be solved. The standard approach are Markov chain Monte Carlo methods. We refer to [27, 56, 57] for further details of those and, in general, Bayesian inversion. In the following, we shall focus on MAP estimation, namely, the corresponding variational model (4.21).

There are several ways of using prior information. In a rough classification, we can distinguish three approaches:

(1) General structure information, e.g. geometrical information such as a lack of oscillations and smoothness between reasonable edge sets as modeled by total variation and Mumford–Shah minimization. Related approaches are based on looking for sparse representations in wavelet, curvelet, shearlet, or similar frame systems. Such approaches are frequently used since prior knowledge is intuitive and quite minimal.

(2) Available dictionaries of “typical” solutions. Dictionaries are learned such that some kind of sparse representation of the image in terms of the dictionary can be expected. There are two standard approaches to sparsity in dictionaries, namely,

synthesis-based and analysis based ([39, 82]). In both cases, one frequently uses convex relaxations, respectively log-concave prior distributions like Laplace distributions on the coefficients. If a suitable dictionary is available, such approaches can be extremely efficient, but a major trouble, particularly in inverse problems, is that a dictionary of possible solutions is hardly accessible. In medical applications, a further concern about such approaches is the danger that the available dictionary does not include the pathologies arising in certain patients which might then be eliminated in the image reconstructions.

(3) Available prior information of the local content in an additional dimension. This could be possible time dynamics or spectral signatures in each pixel. Again, there is some kind of sparsity prior that can be used here. Given a dictionary of the typical local content, it is natural to assume that in each pixel, there is only a mixture (modeled as linear combination) of few elements. For example, in hyperspectral images, one can assume that in each pixel there is only a combination of a few different materials, and hence, a sparse mixture of material spectral signatures. From this example, one can also understand that the sparsity should be expected to increase with spatial resolution.

In variational models of the form (4.21), it is usually beneficial to use convex functionals R for theoretical purposes as well as to avoid computational difficulties in computing global minimizers. Thus, sparsity priors are usually relaxed from minimizing the number of nonzero coefficients (the so-called ℓ^0-norm) to minimizing the ℓ^1-norm of the coefficients. Since this step is often made in an *ad hoc* fashion in literature, we give a simple explanation why the ℓ^1-norm is a reasonable relaxation in the following. For this sake, consider the case of an operator K acting on $\ell^1(I)$ with a finite or countable index set I and assume further that some upper bound C_i on each element u_i is available, which is reasonable in most applications. Thus, we can formulate the sparsity minimization problem as a mixed integer programming problem of the form

$$D(Ku, f) + \sum_{i \in I} p_i \to \min_{u \in \ell^1(I), p_i \in \{0,1\}} \tag{4.24}$$

subject to the constraints

$$|u_i| \leq C_i, \quad (1 - p_i)u_i = 0 . \tag{4.25}$$

One observes that this constrained minimization can be equivalently rewritten as (4.24) subject to

$$|u_i| \leq C_i p_i, \tag{4.26}$$

since for $u_i \neq 0$, only $p_i = 1$ is possible and for $u_i = 0$, the minimizer will clearly satisfy $p_i = 0$. The straightforward convex relaxation is to replace $p_i \in \{0, 1\}$ by $p_i \in [0, 1]$. For the resulting problem

$$D(Ku, f) + \sum_{i} p_i \to \min_{u \in \ell^1(I), p_i \in [0,1]} \tag{4.27}$$

subject to (4.26), the optimal value of p_i can be computed as $p_i = \frac{|u_i|}{C_i}$. By eliminating this variable, we end up with minimizing the weighted ℓ^1-regularized problem ($\omega_i = \frac{1}{C_i}$)

$$D(Ku, f) + \sum_i \omega_i |u_i| \to \min_{u \in \ell^1(I), |u_i| \leq C_i} . \tag{4.28}$$

Several recent concepts have been proposed for improvements, particularly hyperprior frameworks where one has a regularization of the form $R(u; p)$ depending on additional parameters for which another prior distribution is available. The MAP estimate then amounts to solve

$$(\hat{u}, \hat{p}) \in \arg\min_{(u,p)} \left(D(u, f) + R(u; p) + S(p) \right) . \tag{4.29}$$

A related example is the idea of inf-convolution, which represents the image as a sum of two parts for which separate prior information is available. The corresponding variational model is of the form

$$(\hat{u}, \hat{v}) \in \arg\min_{(u,v)} \left(D(u, f) + R_1(u - v) + R_2(v) \right) , \tag{4.30}$$

which has been of particular interest in total variation combined with a higher-order functional for smooth parts in the image ([15, 16, 87]).

A systematic error of MAP estimates is to underestimate $R(u)$, which, e.g. results in contrast losses in the case of total variation regularization. In order to cure this, the Bregman iteration has been proposed ([77]), which, instead of a single solution of a variational model, constructs a sequence

$$u^{k+1} \in \arg\min_u \left(\tau D(u, f) + R(u) - \langle p^k, u \rangle \right) \tag{4.31}$$

for $p^k \in \partial R(u^k)$ and small τ. The behavior is the one of an iterative regularization method, and hence appropriate termination is necessary for optimal results.

4.2 Undersampling and superresolution

A frequently studied issue is the case of undersampled data, i.e. one tries to achieve a higher spatial resolution than the Shannon sampling theorem allows for the number of measurements. Obviously, this is possible only with the use of strong prior information. Let S denote a sampling operator. Then, the inverse problem can be reformulated as

$$SKu = g = Sf. \tag{4.32}$$

In this case, the most important issue is not necessarily the noise (for rather low-dimensional range of S the instability decreases and, e.g. the Moore–Penrose inverse becomes continuous), but rather the null space. Hence, it is important to understand

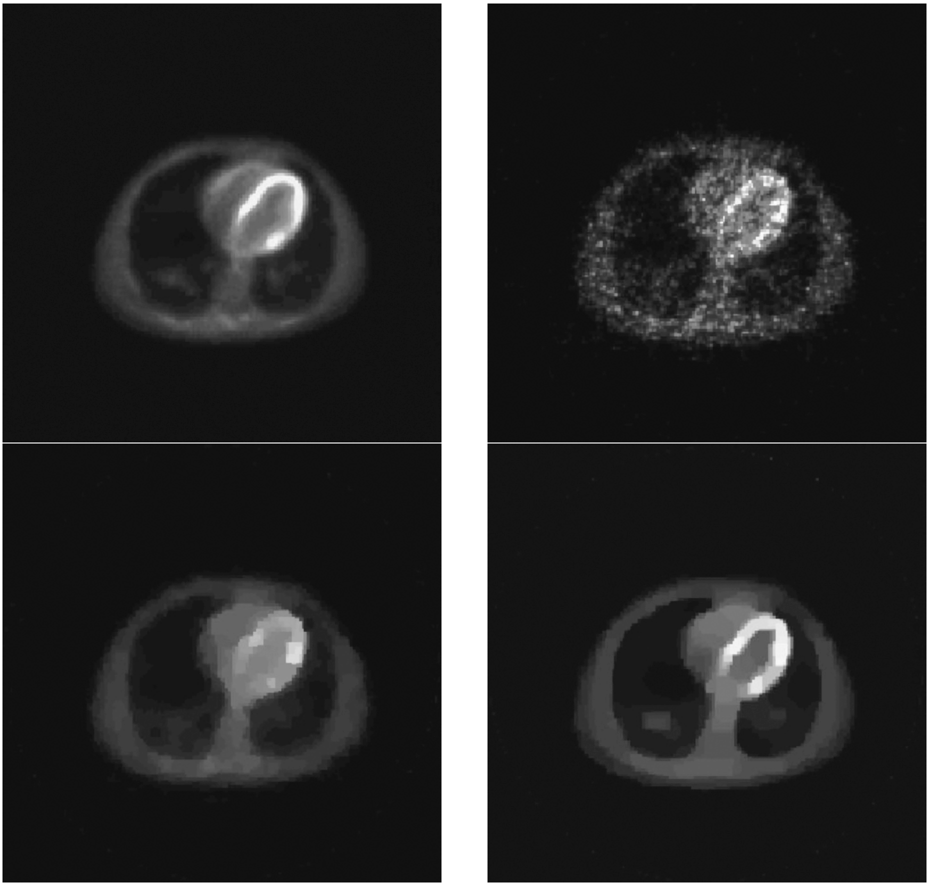


Figure 4.6: Illustration of the impact of using prior knowledge in noisy image reconstruction on a cardiac PET scan (measurement data courtesy of European Institute for Molecular Imaging, Münster). Top row: EM-reconstruction with 20 minutes of data, i.e. low noise level, representing ground truth (left) and EM-reconstruction with five seconds of data, i.e. high noise level (right). Bottom row: Reconstruction from a variational model with TV-regularization (left) and improvement by Bregman iteration (right), both with five seconds of data.

how one can favor the type of solutions corresponding well to the prior knowledge. For this sake, it is of particular interest to study the problem

$$R(u) \to \min_{u \text{ satisfying (4.32)}} . \tag{4.33}$$

In the finite-dimensional sparsity case with R being the ℓ^1-norm, the study of this problem has led to celebrated results of compressed sensing. Under certain conditions on the operator $A = SK$, which is then just a matrix with large null space, one can indeed uniquely reconstruct very sparse solutions of (4.32) in this way, even for the low-dimensional range of S. Classical conditions, by now, are low incoherence in the matrix A, i.e. for two different rows A_i and A_j of A, one should have

$$|A_i \cdot A_j| \ll 1\,, \tag{4.34}$$

and the celebrated restricted isometry property, which requests A to be close to an isometry on subspaces of k-sparse inputs u. The set of conditions needed for this sake has been refined in the last years (cf., e.g. [28, 30, 37, 47, 95]) and extended also to exact reconstruction in the case of the unconstrained problem (4.21) and to related problems such as the reconstruction of low-rank matrices ([29]) or matrices with certain sparsity patterns in rows or lines (cf., e.g. [31, 108]) by convex variational methods.

From the inverse problems point of view, a potential issue in applying the compressed sensing theory is the fact that it is formulated in a strictly finite-dimensional setting, while it is more natural to study the limit of infinite dimensions in the inversion. In particular, it would be desirable to have a theory that works at different image resolutions in this context, but it is neither well studied how sparsity priors apply at different resolutions nor how the conditions for exact recovery change when refining the image resolution. The first issue has recently been studied ([1, 2]) and led to several interesting results. The second issue is more severe in the case of ill-posed inverse problems. Due to the ill-posedness, it is clear that the coherence between some rows needs to converge to one. Also, the restricted isometry property will be difficult to satisfy for reasonable values of the sparsity level k, even for $k = 1$ one can construct simple counterexamples for compact operators based on the singular value decomposition. Indeed, it has been verified computationally that even reasonable discretizations of simple forward operators like the Radon or X-ray transform are far from satisfying restricted isometry properties ([40]). Hence, it seems that in the study of superresolution in an inverse problems setting, it is more reasonable to pose the question in a different way and to rather understand for the given operator which solutions are reconstructed nicely or even favored. In the sparsity setting, this would mean to ask on which k-dimensional subspaces conditions for exact reconstruction are satisfied. A condition that can be generalized to infinite dimensions is the exact recovery condition [47, 95] which has been used for inverse problems in [94] and [10].

For more general convex regularizations R like total variation, it is more difficult to analyze the structure of solutions. For this sake, classical concepts in regularization theory, like the source condition

$$\exists q : K^* S^* q \in \partial R(u) \tag{4.35}$$

or even more the stronger source condition

$$\exists q : K^* S^* S K w \in \partial R(u), \tag{4.36}$$

are useful. Such conditions were mainly used to obtain error estimates for variational regularizations. However, in the case of singular regularization functionals like ℓ^1 or total variation, which have large subdifferentials $\partial R(u)$, it turns out that error estimates like in [81] can indeed imply exact reconstruction. We refer to [10, 68] for further discussion and for establishing relations between source conditions and other conditions used in compressed sensing. Another recently proposed concept that allows one to study the exact solution is a generalization of singular vectors and singular values to the case of nonlinear regularization defined by

$$\lambda K^* S^* S K u_\lambda \in \partial R(u_\lambda)\,. \tag{4.37}$$

The value λ and the corresponding singular vector u_λ can also be defined to define and analyze properties at different scales in an abstract way.

4.3 Inpainting

Inpainting is a classical multidimensional interpolation problem, i.e. the restoration of an image in a subregion $\Sigma \subset \Omega$, which is generally named as the *inpainting domain*. In the predigital era, inpainting was already carried out by restorers in arts, who used their conception of the overall image and the original painter's approach to inpaint damaged regions in paintings. This history is also the reason for the nomenclature inpainting instead of interpolation. Another key difference to classical interpolation theory and methods usually taught in numerical analysis is the kind of prior knowledge and of the desired results. While classical interpolation methods work well for smooth functions and small gaps to be interpolated (which is reflected by standard error estimates), inpainting of images again needs to preserve (or continue) edges and textures, i.e. the nonsmooth components.

In a simple inverse problems formulation, we can formulate inpainting as an operator equation (4.6) with

$$K : \mathcal{U}(\Omega) \to \mathcal{U}(\Omega \setminus \Sigma), \quad u \mapsto u|_{\Omega\setminus\Sigma}\,, \tag{4.38}$$

where $\mathcal{U}(\Omega)$ is an appropriate function space, e.g. $BV(\Omega)$ in the case of cartoon images. Note that K can also be written as

$$(Ku)(x) = \chi_{\Omega\setminus\Sigma}(x)u(x)\,, \tag{4.39}$$

where χ_D is the indicator function of a set D (equal to one inside and zero outside). This immediately implies that K has a large null space (all function supported in Σ), but behaves well (like the identity) orthogonal to the null space.

Well-known models for image inpainting consist of minimizing a variational functional

$$J(u) = D(u, f) + \alpha R(u) \to \min_u \tag{4.40}$$

with a distance term $D(u, f)$ that first projects u to the smaller support $\Omega \setminus \Sigma$ and then compares it with the given data f. The regularization term $R(u)$ specifies the method of inpainting on Σ (see Figure 4.7 for an illustration).

- *Laplace inpainting:* We set the data term $D(u, f)$ as the squared difference from f and choose the regularization as the squared gradient norm to assure smooth areas inside the inpainting domain and obtain

$$\frac{1}{2}\int\limits_{\Omega\setminus\Sigma}(Ku-f)^2+\alpha\int\limits_{\Omega}|\nabla u|^2\to\min_u\,, \tag{4.41}$$

where K is the so-called *downscaling operator*. Calculating optimality conditions results in

$$K^*(Ku-f)\mid_{\Omega\setminus\Sigma}-\alpha\Delta u=0 \tag{4.42}$$

and we see that for solving the variational problem, we have to calculate the Laplace equation on the inpainting domain Σ with boundary conditions that come from the known image.

- *TV inpainting:* Again, we choose a quadratic penalty for the known data. However, for the regularizer, we would like to minimize the total variation of u, and hence

$$\frac{1}{2}\int\limits_{\Omega\setminus\Sigma}(Ku-f)^2+\alpha\int\limits_{\Omega}|\nabla u|\to\min_u\,. \tag{4.43}$$

This regularizer will fill the inpainting domain with piecewise constant areas. The optimality conditions are given by

$$K^*(Ku-f)\mid_{\Omega\setminus\Sigma}-\alpha\nabla\cdot\left[\frac{\nabla u}{|\nabla u|}\right]=0\,. \tag{4.44}$$

For the implementation by a steepest descent algorithm, we obtain a nonlinear diffusion-reaction system

$$\frac{\partial u}{\partial t}=K^*(Ku-f)\mid_{\Omega\setminus\Sigma}+\alpha\nabla\cdot\left[\frac{\nabla u}{|\nabla u|}\right]\,. \tag{4.45}$$

To avoid a singularity of $1/|\nabla u|$, the norm is usually rearranged to $|\nabla u|_\epsilon = \sqrt{\epsilon^2+|\nabla u|^2}$ with ϵ being a small positive constant.

- *TV-H^{-1} inpainting:* Both Laplace and TV inpainting belong to the class of second-order inpainting methods, where the order is given by the highest derivative in the corresponding Euler–Lagrange optimality scheme. Second-order methods generally have two important drawbacks. First, they are not able to connect edges over large distances and secondly, a continuous curvature is not propagated from the image into the inpainting domain. Methods of higher order are able to fix these drawbacks. A method of particular interest is called TV-H^{-1} inpainting. The image is inpainted via

$$\frac{\partial u}{\partial t}=K^*(Ku-f)\mid_{\Omega\setminus\Sigma}+\alpha\Delta p\,,\quad p\in\partial TV(u) \tag{4.46}$$

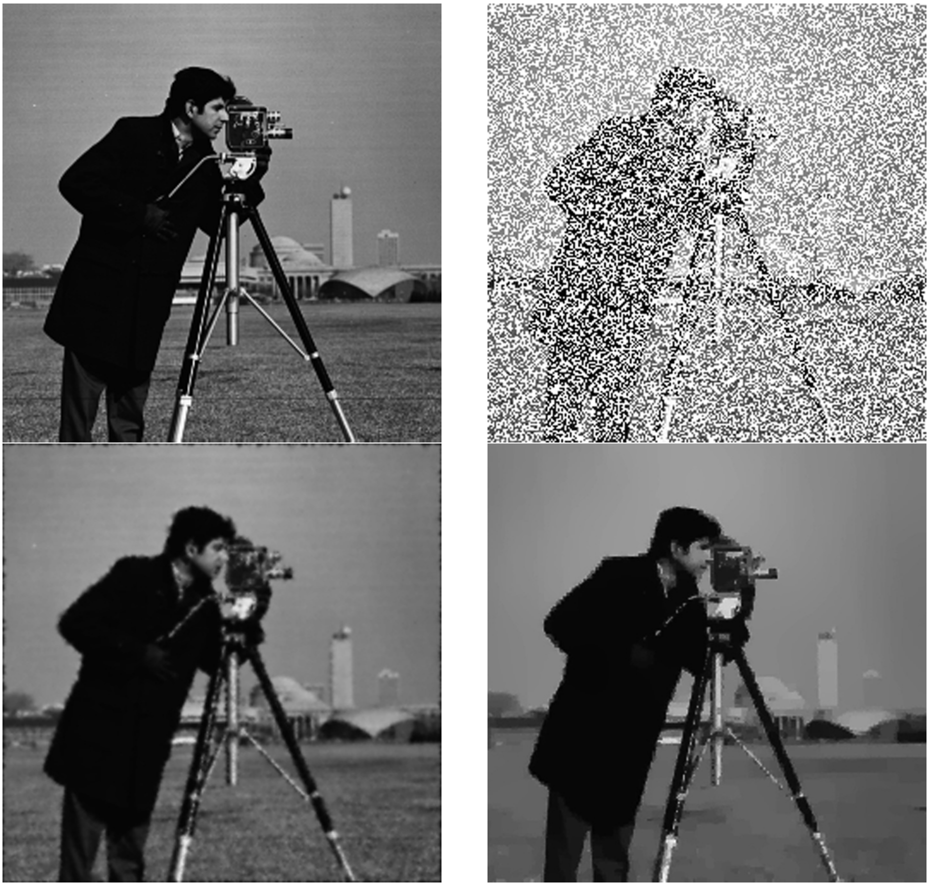


Figure 4.7: Results of different inpainting methods. First row: Original image (left) and damaged image with 50% missing pixels (right). Second row: Restoration with Laplace-inpainting (left) and restoration with TV-inpainting (right).

where $\partial TV(u)$ denotes the subdifferential of $TV(u)$. The element p is approximated by $\nabla \cdot [\nabla u/|\nabla u|_\epsilon]$ where $|\nabla u|_\epsilon$ is again a smoothed version of $|\nabla u|$ (see above). For the implementation of the resulting PDE

$$\frac{\partial u}{\partial t} = K^*(Ku - f)\;|_{\Omega\setminus\Sigma} + \alpha\Delta\nabla \cdot \left[\frac{\nabla u}{|\nabla u|_\epsilon}\right], \tag{4.47}$$

we refer to ([86]).

An even more recent problem is the inpainting of videos. It poses further challenges on computation and modeling, but also offers more prior information. An example is the inpainting of damaged parts in single frames of a video, where clearly the previous and subsequent frames can be used to gain information. We refer to [20, 36] for further information.

4.4 Surface imaging

As already mentioned in the case of tomography, it is only possible to acquire measurements related to images on outside surfaces in most cases, particularly in medical imaging devices. This means the data are effectively taken on a surface, while the unknown image is a function in the inside volume. In tomography, the dimensionality of the data is matched with the one of the image by the additional rotational degree of freedom, but in some modalities, this is not the case. Limited-angle tomography, e.g. with C-bow devices or electron tomography, are already borderline cases being severely ill-posed, that is, underdetermined. Extreme cases are optical tomography (fluorescence or bioluminescence) or electromagnetical imaging (EEG/MEG, that is, ECG/MCG). In the optical case, data can be acquired only for a few different frequencies and a few different angles, if at all. In the case of electrical and magnetic data, one has no further options to obtain data. Additional prior information can be due to physiological considerations on the one hand, e.g. sparsity of sources in space in some optical investigations or in EEG/MEG, and anatomical prior information from X-ray, CT, or MR information on the other hand. The latter can particularly restrict the support of the unknown image to the relevant structures.

As a simplified example that well reflects the mathematical issues in such problems, let us consider a source reconstruction problem for the Poisson equation, i.e.

$$-\Delta v = u \qquad \text{in } \Omega \subset \mathbb{R}^3, \tag{4.48}$$

with homogeneous Neumann boundary conditions $\frac{\partial v}{\partial n} = 0$ on $\partial\Omega$. The forward operator $K : L^p_\diamond(\Omega) \to L^2_\diamond(\partial\Omega)$, given by the map $u \mapsto v$, where v is the unique solution of (4.48) with

$$\int_{\partial\Omega} v \, d\sigma = 0 \,. \tag{4.49}$$

By $L^p_\diamond(\Omega)$, we denote the subspace of $L^p(\Omega)$ consisting of those functions with

$$\int_{\Omega} u \, dx = 0 \,. \tag{4.50}$$

Obviously, the forward operator K has a huge null space, including, in particular, the Laplacian of any compactly supported smooth function. In order to understand how the null space is affected by a variational regularization used to incorporate prior information, we again consider the problem

$$R(u) \to \min_u \quad \text{subject to } Ku = f \,. \tag{4.51}$$

This problem can be formulated as a constrained minimization problem with the Lagrangian

$$\mathcal{L}(u,p,q) = R(u) + \int_{\partial\Omega} (v - f)p \, d\sigma + \int_{\Omega} (\nabla v \cdot \nabla q - fu) \, dx \, . \tag{4.52}$$

If there exist Lagrange-multipliers p and q, which corresponds to the so-called *source condition* in regularization theory ([21, 42]), then they solve

$$q \in \partial R(u) \tag{4.53}$$

$$-\Delta q = 0 \qquad \text{in } \Omega \tag{4.54}$$

$$p + \frac{\partial q}{\partial n} = 0 \qquad \text{on } \partial\Omega \, . \tag{4.55}$$

Now, let us consider some regularization functionals R and their impact. We shall denote by w a weight representing anatomical prior knowledge, i.e. $w(x)$ is large if x is likely to be an element of the support of u and $w(x) \approx 0$. We consider the following cases:

- *Minimum-Norm Solutions:* This case, usually used if no specific prior information is available, corresponds to

$$R(u) = \frac{1}{2} \int_{\Omega} u^2 \, dx. \tag{4.56}$$

One easily checks $\partial R(u) = \{u\}$ and thus (4.53)–(4.55) is satisfied if u is a harmonic function in Ω. Due to elliptic regularity, the reconstruction will be smooth inside Ω and cannot have compact support. Moreover, note that by the maximum principle for harmonic functions, u attains its maximum on $\partial\Omega$. One observes that the latter explains the so-called *depth bias* frequently observed in such problems ([26, 61, 70]), i.e. the minimum norm solution shifts the mass of the reconstruction towards the surface.

- *Weighted-Norm Solutions:* Including the prior by the weight w in the L^2-norm, we have

$$R(u) = \frac{1}{2} \int_{\Omega} \frac{u^2}{w} \, dx. \tag{4.57}$$

One easily checks $\partial R(u) = \{\frac{u}{w}\}$ and thus (4.53)–(4.55) is satisfied if $u = wq$ for a harmonic function q in Ω. The weighting can clearly reduce the depth bias depending on w. Inside regions of homogeneous weights w, the reconstruction u is still harmonic, and thus the maximum principle holds for u. Hence, the mass is still shifted towards the outside surface as much as possible.

- *Sparse Solutions:* In order to obtain solutions with very small support, it is natural to use L^1-type priors, i.e.

$$R(u) = \int_{\Omega} |u| \, dx. \tag{4.58}$$

Formally, the subdifferential is given by $\partial R(u) = \{s\}$, where

$$s(x) \begin{cases} = 1 & \text{if } u(x) > 0 \\ = -1 & \text{if } u(x) < 0 \\ \in [0,1] & \text{if } u(x) = 0 \end{cases} \tag{4.59}$$

is the multivalued sign.

Condition (4.53)–(4.55) is satisfied if s is a harmonic function in Ω. Again, by the strong maximum principle, s is either constant or attains its maximum and minimum only at the boundary. In the first case, we need to further distinguish three cases: If the absolute value of the constant is less than one, then u vanishes everywhere. If the constant equals one, then u is nonnegative everywhere and needs to be equal to zero again due to the vanishing mean value. If the constant equals minus one, an analogous argument holds. Clearly, the absolute value of s cannot be larger than one, and thus in any case, $u \equiv 0$ if s is constant. If s is not constant, it attains its maximum and minimum on $\partial\Omega$ and thus, the absolute value of s is necessary for less than one in the interior of Ω, which means again u vanishes there. The consequence is that u needs to be concentrated on $\partial\Omega$, which of course does not work with an L^1-theory, but can be made rigorous in a usual way by considering Radon measures and their total variation instead of L^1-functions and their norm. This way, we observe the extreme consequences of the depth bias on sparsity, that is, the solution will always be concentrated at zero depth.

- *Weighted Sparse Solutions:* Again, L^1-type priors can be weighted using anatomical prior information, i.e.

$$R(u) = \int_\Omega \frac{|u|}{w}\, dx. \tag{4.60}$$

Formally, the subdifferential is given by $\partial R(u) = \{\frac{s}{w}\}$, where s is a multivalued sign of u. Now, (4.53)–(4.55) is satisfied if $s = wq$ for a harmonic function q in Ω. For appropriate weights, s can achieve its minimum and maximum inside the domain Ω such that the support is not necessarily on the outer surface. However, the possible maximum still strongly depends on the properties of w. For homogeneous regions, a maximum will again be on the part closer to the outer surface, and thus a depth bias prevails. Some depth bias can be eliminated if w is scaled with the operator, respectively.

- *Total Variation Regularization:* If we use total variation regularization, formally

$$R(u) = \int_\Omega |\nabla u|\, dx, \tag{4.61}$$

then the subdifferential contains elements of the form $\nabla \cdot g$, with $g = \frac{\nabla u}{|\nabla u|}$ on smooth parts with nonvanishing gradients. Basic arguments in differential geometry imply that indeed $\nabla \cdot g$ is the mean curvature of level sets of u. Now, since $\nabla \cdot g$ is harmonic, we need to expect that the maximal curvature is attained on $\partial\Omega$. Hence, we cannot expect to reconstruct small features (with large curvature) correctly with increasing distance from the measurement surface.

From the above discussion, one observes that all standard approaches to incorporate prior knowledge suffer from severe shortcomings in the application to imaging from (underdetermined) surface data. It remains an important future challenge to develop improved approaches that can provably reconstruct structures corresponding to the available prior knowledge.

5 Calibration problems

In the last decades and years, significant technical improvements have been made in existing devices and new modalities have been invented such that resolution is continuously improving. Novel devices to increase spatial resolution and fast measurements to increase time resolution lead, however, to a novel kind of mathematical problems which we want to summarize under the term *calibration problems* in the following. The major issues are that a good characterization of the device properties (e.g. the PSF of a microscope) is not (yet) possible or depends on the subject to be imaged, or that the need to take fast measurements does not leave enough time to calibrate the device well (e.g. coils in fast MR imaging).

The resulting mathematical structure is typically of the form

$$K(p)u = f, \tag{4.62}$$

now with $K(p)$ a linear operator depending (possibly in a nonlinear way) on the parameter (functions) encoded by p. Even if the dependence on p is linear, the overall inverse problem becomes nonlinear, often a bilinear problem, which is clearly more difficult to solve than the linear inversion for given p. Moreover, even if the problem for given p can be overdetermined, the joint reconstruction of u and p may be underdetermined and thus again enforces the use of appropriate prior information. Clearly, the prior information on the parameters is quite different than the one on the image. Usually, good mean values are available for p as well as a strong perception of spatial smoothness, which means that such functions are usually modeled as elements in Sobolev spaces (often of high order), with a small distance to the given prior value. Moreover, other structural constraints such as nonnegativity can be available.

5.1 Blind deconvolution

A classical problem of the above type is blind deconvolution, in its original version with p being the point spread function itself, i.e. the bilinear forward operator

$$[K(p)u](x) = \int_\Omega p(x-y)u(y)\,dy\,, \tag{4.63}$$

where u is a single image or even a vector of images. Due to the obvious underdetermination of the nonlinear inverse problem, it is essential to use appropriate prior information on the kernel p and the image u. An obvious constraint is nonnegativity and a scaling property of p such as $\int_{\mathbb{R}^d} p(x)\,dx = 1$. Further knowledge is usually introduced via appropriate regularization, particularly by minimizing functionals like

$$D(K(p)u, f) + \alpha_1 R_1(u) + \alpha_2 R_2(p), \tag{4.64}$$

where D is a standard distance functional such as the squared L^2-norm or the Kullback–Leibler divergence. The functionals R_i are different regularization terms with regularization parameter α_i.

In many instances, the blind deconvolution problem can be modified with additional modeling of the point spread function. Prominent examples are the phase effects appearing in various optical imaging modalities from astronomy down to nanoscopy. For the phase being the parameter to be determined, we have

$$[K(p)u](x) = \int_\Omega k(x-y, p(x))u(y)\,dy\,, \tag{4.65}$$

with a given form of the kernel k, usually

$$k(x,p) = k_0(x)\left|E_1(x) - e^{ip}E_2(x)\right|^2\,, \tag{4.66}$$

where E_1 and E_2 become the counterpropagating fields. Using variational approaches like (4.64) even with quadratic priors for the image u and (in higher Sobolev spaces) the phase p, a sufficiently good estimate of the phase can be found. It has been demonstrated that this phase can be used to obtain reconstructions of superior quality by advanced total variation reconstruction methods ([90]). Such a two-step approach is indeed tempting for many calibration problems since one can here benefit from the fact that the forward operator is not too sensitive with respect to the parameter p. Hence, even a rough estimate of p is sufficient for strong improvements in the estimation of u, which, in the end, is indeed the quantity of interest. A thorough mathematical analysis of such a two-step procedure is still missing.

5.2 Nonlinear MR imaging

A class of calibration problems that has gained high interest recently are those arising from fast measurements in MRI. In the standard setting, one obtains MR-data from the model

$$f(t) = \int_{\Omega} u(x)s(x)e^{-2\pi ik(t)\cdot x}\,dx\,, \tag{4.67}$$

where s is the known (precalibrated) coil profile and $k(t)$ encodes the specific trajectory used for MR-measurements. If one tries to obtain fast measurements, there is often not enough time to calibrate the coils accurately, and hence $p = s$ is to be treated as an unknown. Thus, one ends up with a bilinear inverse problem, however, with a good prior $s_0(x)$ from the precalibration. Deviations of s from s_0 can be expected to be small and smooth, and hence one can use a smoothness prior with high regularization parameter on $s - s_0$.

In some cases, further effects such as relaxation or field inhomogeneities become relevant, a more appropriate forward model is then given by

$$f(t) = \int_{\Omega} u(x)s(x)e^{i\omega(x)t}e^{-R_2^*(x)t}e^{-2\pi ik(t)\cdot x}\,dx\,, \tag{4.68}$$

where R_2^* is a relaxation time and ω models the local field inhomogeneity. Potential candidates for the parameter p are the coil sensitivities ([96]), the field inhomogeneity ([92]), and the relaxation time ([76]). So far, there exist few, rather practical, approaches to the solution of these nonlinear underdetermined inverse problems. A detailed analysis highlighting the potential and limitations of the joint reconstruction is an important future task.

5.3 Attenuation correction in SPECT

As mentioned in Section 2.3, SPECT imaging has a challenging structure with respect to attenuation. The forward operator is of the form

$$(Ku)(z,\theta) = \int_{L(z,\theta)} e^{-\int_{L_x(z,\theta)}\rho(y)\,dy}u(x)\,dx\,, \tag{4.69}$$

where by $L(z,\theta)$, we denote the line starting at z in direction θ and by $L_x(z,\theta)$, the line segment between z and x. Note that the tracer density u is different from the (scaled) physical density ρ, and thus the latter is usually determined by an X-ray scan before the SPECT measurement.

In several instances, e.g. in the case of patient movement or dynamic imaging of moving objects, the attenuation determined initially does not remain valid. Thus, it becomes necessary to reconstruct the attenuation density ρ together with u, i.e. the inverse problem becomes nonlinear and of the form (4.62) with $\rho = p$. The additional degrees of freedom do not necessarily lead to an underdetermined problem since the original SPECT data set from all directions indeed overdetermines u. However, the solution of the problem remains challenging from a theoretical perspective ([6, 84]) as well as from a computational point of view. For the latter, various iterative techniques have been investigated ([79]), the most natural being of course an alternating minimization technique of a functional like (4.64).

A different approach is to partly use the previously measured function ρ and instead measure the deformation that appeared before the PET scan. Thus, the parameter p is a vectorial quantity, with a strong prior of p being close to the identity. The forward model thus becomes

$$[K(p)u](z,\theta) = \int\limits_{L(z,\theta)} e^{-\int_{L_x(z,\theta)} \rho(p(y))\,dy} u(x)\,dx\,, \tag{4.70}$$

with ρ given. Already with simple parameterizations, one can obtain significant improvements ([102]). With advanced techniques of nonlinear image registration ([67]), further steps have been made recently ([8]).

5.4 Blind spectral unmixing

With the recent advances in multi- and hyperspectral imaging, the unmixing of spectral signals into basic components has received increasing attention. Blind spectral unmixing is a classical problem in audio applications. Some striking examples are the decomposition of party talk into single person statements and the decomposition of an orchestra recording into the different instruments. In the imaging context, one usually seeks a decomposition of the spectrum into spectra of basic materials to obtain a good characterization of the content of a certain region.

In discrete modeling, the spectral image is a matrix $F \in \mathbb{R}^{N\times M}$, where N is the number of pixels (or voxels) and M is the number of spectral points. The spectral unmixing looks for a coefficient matrix $U \in \mathbb{R}^{N\times K}$, where U_{ij} is the coefficient with respect to the j-th spectral basis function in pixel i. By collecting the basis spectra in a matrix $B \in \mathbb{R}^{K\times M}$, one thus has to solve the matrix equation

$$UB = F. \tag{4.71}$$

While B is given in the classical unmixing problem, it is an unknown itself in the blind unmixing or blind separation problem. Since the data F as well as U and B have naturally nonnegative elements, solving (4.71) can also be cast in the framework of nonnegative matrix factorization. In the above framework, we have $u = U$, $p = B$ and $K(p)$ being the multiplication operator. One also observes the relation to blind deconvolution problems, whose discrete version is a special form of blind unmixing with B restricted to the class of Toeplitz matrices.

A particular property in hyperspectral imaging is a spatial correlation between the pixels which can be modeled in the regularization in order to decrease the nonuniqueness in unmixing. With the popular TV prior, this naturally leads to

$$\|UB - F\|^2 + \alpha \sum_j TV(U_{\cdot j}) + \beta \sum_i R(U_{i\cdot}) + \gamma S(B)\,, \tag{4.72}$$

where $U_{\cdot j}$ denotes the j-th column and $U_{i\cdot}$ the i-th row of U. Moreover, TV denotes the total variation of a discrete image, the functional R is a local prior in each pixel, e.g. the ℓ^1-norm to enforce sparsity with respect to the basis, and S is a functional that models specific prior knowledge on the basis elements, e.g. an ℓ^1-type norm to enforce sparsity in a certain basis.

So far, most of the analysis of blind unmixing is carried out in finite dimension, thus rather for ill-conditioned than ill-posed inverse problems. However, with increasing spatial and spectral resolution of imaging devices, it becomes interesting to study the asymptotics of unmixing problems as N and K tend to infinity (independently or in appropriate relative scaling). Useful reconstruction approaches and algorithms certainly should be characterized by a robust behavior with respect to the asymptotics. Such modeling of the asymptotics and different spatial resolution is also relevant if hybrid imaging is used. In several cases, the hyperspectral data are acquired with low spatial resolution at the same time as a conventional color image at high spatial resolution. The superresolution in the hyperspectral image based on the correlation with the color image is a challenging inverse problems; we refer to [69] for further details.

In addition to the pure unmixing, an interesting inverse problem is to study joint image reconstruction and unmixing. With a forward operator acting on the pixel dimension, the problem becomes

$$AUB = F, \tag{4.73}$$

with a given matrix A.

6 Model-based dynamic imaging

An ultimate goal in a variety of modern imaging approaches is to obtain (quantitative) information about dynamics instead of only still images. Roughly speaking, this means that instead of a single image u, a whole sequence $u(t)$ for varying time t needs to be reconstructed and its dynamics needs to be analyzed. The inverse problem of reconstructing dynamic images can usually be formulated as

$$Ku(t) = f(t), \qquad t \in [0, T] \tag{4.74}$$

since the forward operator is hardly changing with the dynamics. In several instances, the time resolution is so low that it seems more appropriate to consider a time discrete model

$$Ku(t_i) = f(t_i), \qquad i = 1, \ldots, M\,. \tag{4.75}$$

It is obvious from the fact that K does not depend on time that the image reconstruction problem can be split into several stationary reconstruction steps at different times, making all standard methods applicable. However, important information is lost this way. In general, the images $u(t)$ and consequently also the data $f(t)$ are strongly correlated in time since they are typically generated by a smooth time evolution rather than arbitrary changes. One way to incorporate this kind of prior information is to use regularization functionals that penalize large changes in time. Frequently used examples (mainly due to their simplicity) are

$$R(u) = \int_0^T R_0(u(t))\, dt + \frac{1}{2} \int_0^T \|\partial_t u(t)\|^2 \, dt\,, \tag{4.76}$$

where R_0 is a regularization functional in space, respectively

$$R(u) = \sum_{i=1}^{M} R_0(u(t_i)) + \frac{1}{2} \sum_{i=1}^{M-1} \|u(t_{i+1}) - u(t_i)\|^2 \,. \tag{4.77}$$

Besides such all-purpose approaches, a different paradigm taking into account the mathematical modeling of the underlying dynamics has evolved. The correlation is guaranteed by using ODE (ordinary differential equation) or PDE (partial differential equation) models appropriately describing the dynamics, usually with unknown parameter functions to be reconstructed. The image sequence is obtained implicitly by solving the forward model with reconstructed parameters. Since either good priors for those parameter functions exist or they are of lower dimensionality (e.g. independent of time), thus making the inversion overdetermined, improved reconstructions can be gained from such approaches. The main bottlenecks are the mathematical difficulty and computational challenges compared to separate reconstructions at different time steps. Instead of reconstructing a series of images from a linear stationary forward problem, one now has to identify parameters in nonlinear time-dependent differential equations, which, as a further complication to well-known parameter identification problems, have to be combined with the forward operator of the imaging system. For these reasons, the majority of such approaches, with some exceptions for reasonably simple forward problems, are rather at the level of basic mathematical research, but they have high potential to lead to practical advances.

6.1 Kinetic models

Kinetic models are used to model biochemical effects or also as coarse descriptions of the diffusion and exchange of blood traced in examinations with emission tomography ([103]). The majority of such models uses first-order kinetics, i.e. the dynamics of the image is given by

$$u(x,t) = \sum_{j=1}^{k} \omega_j(x,P(x))u_j(x,t) + \omega_0(x,P(x))I(t), \tag{4.78}$$

where ω_0 and ω_j are weights modeling the fraction of the components in different subregions. Here, the vector $U = (u_j)$ of different states follows an ODE system of the form

$$\partial_t U(x,t) = A(x,P(x))U(x,t) + B(x,P(x))I(t), \tag{4.79}$$

where P is the vector of unknown parameters, A and B are matrices in $\mathbb{R}^{k\times k}$ depending linearly on P, and I is a vector of input functions, which we assume to be given here (in practice, they are sometimes estimated from data first, which is a different issue).

A simple example is the one-compartment model for perfusion used in positron emission tomography (PET) with radioactive water ($H_2{}^{15}O$), that is, a tracer which follows the blood flow. The tracer activity in the heart, which is the image in PET, can be written as

$$u(x,t) = \omega_0(x)I_0(t) + \omega_1(x)u_1(x,t)\,, \tag{4.80}$$

with a concentration of the tracer in tissue u_1 and the (homogeneous) arterial concentration I_0. The weights ω_0 and ω_1 correspond to the respective volume fractions of arteries and tissues and can be written as

$$\omega_0(x) = \chi(x)p_3(x), \qquad \omega_1(x) = \chi(x)(1 - p_3(x)) \tag{4.81}$$

where χ is an indicator function of the heart, namely, the region containing blood, which we again consider as given. The ODE system describing the dynamics is given by

$$\partial_t u_1(x,t) = -p_1(x)u_1(x,t) + p_2(x)I(t). \tag{4.82}$$

The model-based inversion now looks for the parameter vector $P = (p_1, p_2, p_3)$ related to the perfusion of tissue (in ml blood per second per ml tissue) and the tissue fraction using the above model equations. In this case, the parameters themselves are more interesting than the image sequence anyway. The current state of the art is to first reconstruct the image sequence $u(t)$ and then extract parameters in regions of interest in order to obtain a quantitative analysis of perfusion. Due to the inherently high noise in time-resolved PET, the reconstructed images are of rather low quality, which limits the success of subsequent parameter estimation, in particular the spatial resolution. By directly inverting for the parameters ([12]), one obtains significantly less degrees of freedom than by inverting for the image sequence, which allows one to increase the spatial resolution. Let us also mention that the above time-continuous modeling seems appropriate for data acquisition in a list-mode format, i.e. for all decay events, the exact time is recorded and saved such that the data can be interpreted as a Poisson sampling from the time-continuous forward projected image sequence. If one works with rebinned and gated data, the discrete modeling in time is more appropriate. One only obtains information about K time intervals, and the model of the image at time t_i instead becomes

$$u(x,t_i) = \int_{t_{i-1}}^{t_i} (\omega_0(x)I(t) + \omega_1(x)u_1(x,t))\,dt\,. \tag{4.83}$$

Similar problems arise in dynamic SPECT and MR ([3, 50]).

At a first glance, it is natural to solve for the parameters in (4.74), (4.78), (4.79) in the framework of a nonlinear parameter identification problem. On the other hand, using prior information for possible values of P, it is often possible to partially discretize the parameters, and since (4.79) is usually a simple system of linear ODEs, it allows for explicit solutions in many cases. Using these explicit solutions for a discrete set of parameters is the basis to rewrite the identification as a basis pursuit problem, which we discuss as an alternative approach in Section 6.3.

6.2 Parameter identification

The variational formulation of the nonlinear inversion as a parameter identification problem is rather straightforward. In the case of noisy data, we can minimize a combination of the log-likelihood with regularization functionals R acting on the parameters P, i.e.

$$\lambda \int_0^T L(f(t)|Ku(t))\,dt + R(P) \tag{4.84}$$

subject to (4.78), (4.79). Standard priors for the parameters again lead to spatial smoothness, possibly with edges, such that Dirichlet energies or total variation are useful choices. Several authors have used such approaches in studies in emission tomography ([13, 58, 105]) and the numerical results confirm significant improvements in results.

The two major questions related to analysis and numerical solution are the following:

- *Analysis*: Provide estimates (in dependence on the noise level) confirming and quantifying the gain of quality in reconstructions when using the nonlinear inversion scheme instead of linear reconstructions of u with subsequent parameter estimation in every point x.
- *Numerical Solution*: Construct numerical schemes to solve the inverse problem efficiently in three spatial and one time dimension.

So far, the first issue is completely open. Although several approaches to error estimation for regularization methods for nonlinear inverse problems exist, the application to the above problem is not straightforward due to the combination of the spatial operator K and the time dynamics. However, an even more severe issue is the comparison to the simpler approach of first reconstructing an image sequence and then estimating parameters for which no advanced concepts in inverse problems are available. The study of such questions, possibly also combined with statistical noise models as in emission tomography, is highly relevant for future research however.

With respect to the efficient numerical solution, further advance has been made recently. In designing computational algorithms, several goals and also limitations have to be taken into account. First of all, the operator K, respectively its discretization as a matrix, is rather complex and can usually neither be stored nor inverted efficiently. Thus, an algorithm for solving the inverse problem should be based mainly on the application of K and its adjoint K^* instead of solving large linear systems including K. Secondly, the problem dimension will be huge if space- and time-dependence are taken into account simultaneously. Thus, it seems more appropriate to use splitting algorithms which can iterate in an alternating way between an image reconstruction and a parameter identification step. A further complication can arise due to the spatial regularization on the parameters which additionally couples the parameter estimation step and might enforce further splitting.

In order to highlight the structure and couplings, let us derive the first-order optimality conditions for the inverse problem in a constrained formulation. Thus, we look for saddle points of the Lagrangian

$$\begin{aligned}\mathcal{L}(u,U,P;v,w) = \lambda \int_0^T L(f(t)|Ku(t))\,dt + R(P) \\ + \int_0^T \int_\Omega \Bigg(u(x,t) - \sum_{j=1}^K \omega_j(x,P(x))u_j(x,t) \\ + \omega_0(x,P(x))I(t) \Bigg) v(x,t)\,dx\,dt \\ + \int_0^T \int_\Omega (\partial_t U(x,t) - A(x,P(x))U(x,t) \\ + B(x,P(x))I(t)) \cdot w(x,t)\,dx\,dt\,.\end{aligned}$$

First-order optimality is given by vanishing first derivatives of the Lagrangian, i.e.

$$
\begin{aligned}
0 &= \partial_u \mathcal{L} = K^* \partial_{Ku} L\big(f(t)|Ku(t)\big) + v(t) \\
0 &= \partial_U \mathcal{L} = -\omega v - \partial_t w + A\, w \\
0 &= \partial_P \mathcal{L} = -\int_0^T \left(\sum_{j=1}^{K} (\partial_P \omega_j u_j + \partial_P \omega_0 I) v + (\partial_P A\, U - \partial_P B I) \cdot w \right) dt + R'(P)\,,
\end{aligned}
$$

where $\omega = (\omega_1(x, P(x)), \dots, \omega_K(x, P(x)))$ is the vector of weights. One observes that the way we introduced constraints naturally separates the image reconstruction and the parameter identification steps: The optimality with respect to u can be considered as an image reconstruction problem for u at each time step t. The forward equation for U together with the adjoint equation arising from the optimality with respect to U and the optimality with respect to P constitute a parameter identification for an ordinary differential equation in a Banach space. Thus, in algorithms, it is natural to split these two subproblems, e.g. by Augmented Lagrangian methods (ADDM). If done appropriately, this usually permits one to use available methods for static image reconstruction at each time step in the first part. For the parameter identification, one observes that all parts, except potentially $R'(P)$, are purely local on each pixel, and hence one obtains systems of decoupled ODEs in each pixel, which can be solved efficiently and parallelized in a trivial way. If R' is local, like for L^p-penalties, then one directly obtains algorithms with reasonable efficiency this way. If R' is a differential operator in space, like in regularization with total variation or Sobolev norms, then a further splitting based on doubling the parameter, i.e. a novel constraint $Q = P$ seems reasonable. If the splitting is performed such that P appears in $R'(P)$, but the coupling to the U and w is via Q, one again obtains efficient algorithms, leaving the ODEs local in space.

6.3 Basis pursuit

As an alternative approach that receives increasing attention in emission tomography, we consider a basis pursuit solution which we discuss for simplicity in the special case of the single compartment model (4.82), which we rewrite for simpler notation as

$$\partial_t v(x,t) = -a(x) v(x,t) + b(x) I(t)\,, \tag{4.85}$$

subject to initial conditions $v(x, 0) = 0$ (usually modeling injection of the tracer at time 0) and overall concentration given by

$$u(x,t) = c(x) v(x,t) + \big(1 - c(x)\big) I(t)\,. \tag{4.86}$$

The differential equation can be solved easily, yielding

$$u(x,t) = c(x)b(x)\int_0^t e^{a(x)(s-t)}I(s)\,ds + (1-c(x))I(t)\,. \tag{4.87}$$

Now, a further key step is to discretize the parameter a into a set of possible values $a_1,\dots,a_N \in \mathbb{R}_+$. Thus, we can write the image as

$$u(x,t) = \sum_{j=0}^{N} \alpha_j(x)\varphi_j(t), \tag{4.88}$$

with unknown coefficients $\alpha_j(x)$ and time basis functions

$$\varphi_0(t) = I(t), \qquad \varphi_j(t) = \int_0^t e^{a_j(s-t)}I(s)\,ds\,, \tag{4.89}$$

which can be precomputed. We need to keep in mind, however, that (4.88) is equivalent to the previous form only if the following conditions are met for each $x \in \Omega$:

$$\alpha_j(x) \geq 0, \quad \alpha_0(x) < 1 \quad \|(\alpha_1,\dots,\alpha_N)\|_{l^0} = 1\,. \tag{4.90}$$

If these conditions are met, we can reconstruct the parameters in the original model via

$$a(x) = a_{J(x)} \quad b(x) = \frac{\alpha_J(x)}{1-\alpha_0(x)} \quad c(x) = 1-\alpha_0(x)\,, \tag{4.91}$$

where $J(x)$ is the index such that $\alpha_{J(x)} \neq 0$.

A particularly attractive feature of the basis pursuit formulation is that the forward model is now linear and has some separation of spatial and temporal features, i.e.

$$f(t) = \sum_{j=0}^{N}(K\alpha_j)\varphi_j(t). \tag{4.92}$$

The major challenges – as usual in basis pursuit – come from the sparsity constraint in (4.90). One heuristic approach is to ignore the constraint and consider nonsparse decompositions or subsequent thresholding (cf., e.g. [80]). For low data quality, one however loses the disadvantages of the modeling approach and the reconstructions can become rather arbitrary. An alternative is to investigate convex relaxations, as frequently used in compressed sensing. This means that the sparsity constraint is usually formulated as a penalty (respectively regularization) and then relaxed from the nonconvex ℓ^0 to the convex ℓ^1-norm. The case of coefficient vectors with only one nonzero entry is usually the easiest one to deal with in compressed sensing. Exact reconstruction is possible, even with some data noise if the basis functions φ_j are normalized in a Hilbert space scalar product ([51]), which is easy to achieve (e.g. by redefining the coefficients).

Unfortunately the latter argument does not apply directly to the inverse problem in (4.92) since one has to solve many inverse problems with sparsity constraints for every x, which are coupled by the operator K. The appropriate sparsity prior is thus of the form

$$\|\alpha\|_{\infty,0} = \sup_{x\in\Omega} \|\alpha(x)\|_{\ell^0}\,, \tag{4.93}$$

where $\alpha(x)$ is the vector of coefficients. A convex relaxation is given by

$$\|\alpha\|_{\infty,1} = \sup_{x\in\Omega} \|\alpha(x)\|_{\ell^1}\,, \tag{4.94}$$

which motivates one to further study problems of the form

$$\lambda \int_0^T L(f|Ku)\,dt + \|\alpha\|_{\infty,1}\,, \tag{4.95}$$

subject to (4.88).

6.4 Motion and deformation models

A particularly important process in many applications is motion, and thus also its modeling receives growing attention in image reconstruction. There are two main aspects of motion in imaging: It can either simply cause disturbances of the images, e.g. as motion blur or by motion of the imaged subject between two time frames, or it can be the process of interest itself, e.g. in quantifying flow behavior. In any case, it is important to use appropriate models for motion, namely, deformations introduced by them. Using motion models, the problem is related to classical motion estimation in image sequences, e.g. via the celebrated optical flow ([5, 54]). Using deformation models between different time frames is related to image registration or fusion ([67]).

Let us start with motion models corresponding to flow dynamics. If the image is modeled via its evolving density in three spatial dimensions (e.g. tracers in fluorescence microscopy, emission tomography, or MR), then it is appropriately modeled via the transport equation

$$\partial_t u + \nabla \cdot (Vu) = 0 \tag{4.96}$$

in $\Omega \times [0, T]$, where V is a velocity vector field to be determined. We mention that in the case of an incompressible substance, the standard relation $\nabla \cdot V = 0$ holds, which reduces the degrees of freedom.

A variational reconstruction scheme including the motion model is then of the form

$$\lambda \int_0^T D(Ku(t), f(t))\,dt + \int_0^T R(u(t), V(t))\,dt \to \min_{(u,V)}\,, \tag{4.97}$$

subject to (4.96), where R is a regularization functional for density and velocity. Such a formulation is related to several classical problems, e.g. the fluid-dynamic formulation of optimal transport ([9]) or optimal control formulations of optical flow ([18, 75]) in the incompressible case. At first glance, the use of (4.96) seems an unnecessary complication in the problem since it increases the degrees of freedom from a scalar function to a vector field and it makes the overall reconstruction nonlinear due to the bilinear constraint in u and V. However, the transport formulation yields important advantages: First of all, the correlation between different time steps is appropriately modeled and using regularization functionals that prevent overly large velocities, it is possible to coherently follow the motion. Moreover, by determining u and V, one obtains a quantification of the flow together with the image reconstruction.

In [17], a reconstruction approach using optimal transport regularizers has been developed using

$$R(u,V) = \frac{|V|^2}{2u} + \left(\int_\Omega |\nabla u|^p \right)^q , \tag{4.98}$$

with particular focus on the total variation case $p = 1$. For $pq > 1$, the existence of a minimizer can be shown and the minimization is convex for standard choices of L.

An alternative to flow models are deformation models which rather correspond to the usual Lagrangian approach in solid mechanics. In this case, we use a deformation $y : \Omega \times [0,T] \to \mathbb{R}^3$ instead of the velocity field V and obtain a solution of (4.96) as

$$u(x,t) = u_0(y(x,t)) \det\left(\nabla y(x,t)\right), \tag{4.99}$$

where $u_0 = u(x,0)$ is the initial value. The variational reconstruction scheme in this case can be formulated as

$$\lambda \int_0^T D(Ku(t), f(t))\, dt + \int_0^T R(u(t), y(t))\, dt \to \min_{(u,y)} \tag{4.100}$$

with u given by (4.99). Since u is given by an explicit formula in terms of u_0 and y, the minimization can be carried out with respect to the latter two variables. Such an approach was taken by [65] for PET. We also refer to [91] for a recent study with hyperelastic regularization of the deformation.

Note that the main difference in the properties of minimizers in the Eulerian (4.97) and Lagrangian (4.100) approaches comes from the way regularization and thus prior knowledge is introduced. In the Eulerian approach, velocities are penalized over time, and hence the goal is an efficient flow as in fluids. In the Lagrangian approach, deformations are penalized, e.g. by elastic or hyperelastic energies, which rather correspond to typical situations in solids.

6.5 Advanced PDE models

So far, advanced fluid models like Navier–Stokes equations for fluids or reaction-diffusion systems are rarely used. Although, in several cases, advanced models are available. The reasons for not using such models in imaging are twofold: First of all, the computational complexity of inverse problems strongly increases with the complexity of the forward models and it is often not clear if the results can be so significantly improved that it justifies a strong increase in computation time. The second reason is that including more complex models also potentially increases the model uncertainty. The reason is that with each new model part, additional parameters and modeling assumptions are introduced. Take, as a simple example, the quantification of intracellular fluid flow from 4D fluorescence microscopy data. Standard flow estimation algorithms simply use the transport equation for the density of the fluorescence tracer with some regularization on the velocity field. One could, however, use the incompressible Stokes model for the fluid flow and estimate the force field, for which good prior knowledge is available. However, using this advanced model, one needs a further assumption of incompressibility and introduces the viscosity as a further uncertain parameter, and hence the overall uncertainty of the forward model is increased.

Besides these issues, there is still growing interest in using advanced PDE models in several fields of image reconstruction and analysis (cf., e.g. [45, 52, 53]), and a large amount of research in this direction is to be expected in the next years.

References

[1] B. Adcock and A. C. Hansen, *Generalized Sampling and Infinite Dimensional Compressed Sensing*, DAMTP, Cambridge University, Preprint, 2013.

[2] B. Adcock, A. C. Hansen, E. Herrholz, and G. Teschke, Generalized Sampling: Extensions to Frames and Inverse and Ill-Posed Problems, *Inverse Problems* 29 (2013), 015008.

[3] G. Adluru, S. P. Awate, T. Tasdizen, R. T. Whitaker, and E. V. R. DiBella, Temporally constrained reconstruction of dynamic cardiac perfusion MRI, *Magnetic Resonance in Medicine* 57 (2007), 1027–1036.

[4] H. Ammari, E. Bossy, J. Garnier, and L. Seppecher, *Acousto-electromagnetic Tomography*, Ecole Polytechnique, Paris, Preprint, 2012.

[5] G. Aubert, R. Deriche, and P. Kornprobst, Computing optical flow via variational techniques, *SIAM J. Appl. Math.* 60 (2000), 156–182.

[6] G. Bal, On the attenuated Radon transform with full and partial measurements, *Inverse Problems* 20 (2004), 399–418.

[7] J. M. Bardsley and A. Luttman, Total variation-penalized Poisson likelihood estimation for ill-posed problems, *Advances in Computational Mathematics* 31 (2009), 35–59.

[8] S. Barendt and J. Modersitzki, A Variational Model for SPECT Reconstruction with a Nonlinearly Transformed Attenuation Prototype, *International Journal of Computer Mathematics* 90 (2012), 82–91.

[9] J.-D. Benamou and Y. Brenier, A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem, *Numer. Math.* 84 (2000), 375–393.

[10] M. Benning, *Singular Regularization of Inverse Problems*, Ph.D. thesis, Institute for Computational and Applied Mathematics, University of Münster, Germany, May 2011.

[11] M. Benning and M. Burger, Error Estimates for General Fidelities, *Electronic Transactions on Numerical Analysis* 38 (2011), 44–68.

[12] M. Benning, T. Kösters, and F. Lamare, *Combined Correction and Reconstruction Methods*, ch. 9, pp. 185–206, CRC Press, April 2012.

[13] M. Benning, T. Kösters, F. Wübbeling, K. Schäfers, and M. Burger, A Nonlinear Variational Method for Improved Quantification of Myocardial Blood Flow Using Dynamic H215O PET, in: *Nuclear Science Symposium Conference Record, 2008. NSS '08. IEEE*, November 2008.

[14] J. Bergé, S. Price, A. Amara, and J. Rhodes, On point spread function modelling: towards optimal interpolation, *Monthly Notices of the Royal Astronomical Society* 419 (2012), 2356–2368.

[15] K. Bredies, K. Kunisch, and T. Pock, Total generalized variation, *SIAM J. Img. Sci.* 3 (2010), 492–526.

[16] K. Bredies, K. Kunisch, and T. Valkonen, Properties of $L^1 - TGV^2$: The one-dimensional case, *J. Math. Anal. Appl.* 398 (2013), 438–454.

[17] C. Brune, *4D Imaging in Tomography and Optical Nanoscopy*, Ph.D. thesis, Institute for Computational and Applied Mathematics, University of Münster, Germany, July 2010.

[18] C. Brune, H. Maurer, and M. Wagner, Detection of Intensity and Motion Edges within Optical Flow via Multidimensional Control, *SIAM J. Imag. Sci.* 2 (2009), 1190–1210.

[19] A. Buades, B. Coll, and J. M. Morel, Image denoising methods. A new nonlocal principle, *SIAM Rev.* 52 (2010), 113–147.

[20] A. Bugeau, P. Gargallo, O. D'Hondt, A. Hervieu, N. Papadakis, and V. Caselles, Coherent Background Video Inpainting through Kalman Smoothing along Trajectories., in: *Vision, Modeling, and Visualization*, 2010.

[21] M. Burger and S. J. Osher, Convergence rates of convex variational regularization, *Inverse Problems* 20 (2004), 1411–1421.

[22] M. Burger and S. J. Osher, *A guide to the TV zoo*, University of Münster, Preprint, Germany, 2012.

[23] J. Cai, S. Osher, and Z. Shen, Split Bregman Methods and Frame Based Image Restoration, *Multiscale Model. Simul.* 8 (2010), 337–369.

[24] J.-F. Cai, B. Dong, S. Osher, and Z. Shen, Image restoration: Total variation, wavelet frames, and beyond, *J. Amer. Math. Soc.* 25 (2012), 1033–1089.

[25] A.-P. Calderón, *On an inverse boundary value problem*, Seminar on Numerical Analysis and its Applications to Continuum Physics, Soc. Brasil. Mat., Rio de Janeiro, 1980, pp. 65–73.

[26] D. Calvetti, L. Homa, and E. Somersalo, Bayesian mixture models for source separation in MEG, *Inverse Problems* 27 (2011), 115001.

[27] D. Calvetti and E. Somersalo, *Introduction to Bayesian scientific computing. Ten Lectures on Subjective Computing*, Surveys and Tutorials in the Applied Mathematical Sciences 2, Springer, New York, 2007.

[28] E. J. Candès, The restricted isometry property and its implications for compressed sensing, *C. R. Math. Acad. Sci. Paris* 346 (2008), 589–592.

[29] E. J. Candès and B. Recht, Exact matrix completion via convex optimization, *Found. Comput. Math.* 9 (2009), 717–772.

[30] E. J. Candes and T. Tao, The Dantzig selector: statistical estimation when p is much larger than n, *Ann. Statist.* 35 (2007), 2313–2351.

[31] A. Castrodad, I. Ramirez, G. Sapiro, P. Sprechmann, and G. Yu, *Second-generation sparse modeling: structured and collaborative signal analysis*, Compressed sensing, Cambridge Univ. Press, Cambridge, 2012, pp. 65–87.

[32] T. F. Chan and J. Shen, *Image processing and analysis*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2005.

[33] M. Cheney, D. Isaacson, and J. C. Newell, Electrical impedance tomography, *SIAM Rev.* 41 (1999), 85–101.

[34] A. Cohen, W. Dahmen, I. Daubechies, and R. DeVore, Harmonic analysis of the space BV, *Rev. Mat. Iberoamericana* 19 (2003), 235–263.

[35] P. Combettes and V. Wajs, Signal Recovery by Proximal Forward-Backward Splitting, *Multiscale Model. Simul.* 4 (2005), 1168–1200.

[36] Tao D., M. Sznaier, and O. I. Camps, A Rank Minimization Approach to Video Inpainting, in: *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8, Oct. 2007.

[37] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok, *Introduction to compressed sensing*, Compressed sensing, Cambridge Univ. Press, Cambridge, 2012, pp. 1–64.

[38] D. L. Donoho, Compressed sensing, *IEEE Trans. Inform. Theory* 52 (2006), 1289–1306.

[39] D. L. Donoho and M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization, *Proc. Natl. Acad. Sci. USA* 100 (2003), 2197–2202.

[40] C. Dossal, G. Peyré, and J. Fadili, A numerical exploration of compressed sampling recovery, *Linear Algebra Appl.* 432 (2010), 1663–1679.

[41] Yonina C. Eldar and Gitta Kutyniok (eds.), *Compressed sensing. Theory and applications*, Cambridge University Press, Cambridge, 2012.

[42] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of inverse problems*, Mathematics and its Applications 375, Kluwer Academic Publishers Group, Dordrecht, 1996.

[43] D. Fanelli and O. Öktem, Electron tomography: a short overview with an emphasis on the absorption potential model for the forward problem, *Inverse Problems* 24 (2008), 013001.

[44] H. Federer, *Geometric measure theory*, Die Grundlehren der mathematischen Wissenschaften, Band 153, Springer-Verlag, New York, 1969.

[45] I. N. Figueiredo, P. N. Figueiredo, and N. Almeida, Image-driven parameter estimation in absorption-diffusion models of chromoscopy, *SIAM J. Img. Sci.* 4 (2011), 884–904.

[46] J. Flemming and B. Hofmann, A New Approach to Source Conditions in Regularization with General Residual Term, *Numerical Functional Analysis and Optimization* 31 (2010), 254–284.

[47] J. J. Fuchs, Recovery of exact sparse representations in the presence of bounded noise, *IEEE Trans. Inform. Theory* 51 (2005), 3601–3608.

[48] G. Gilboa and S. Osher, Nonlocal operators with applications to image processing, *Multiscale Model. Simul.* 7 (2008), 1005–1028.

[49] T. Goldstein and S. J. Osher, The Split Bregman Method for L1-Regularized Problems, *SIAM J. Img. Sci.* 2 (2009), 323–343.

[50] G. T. Gullberg, B. W. Reutter, A. Sitek, J. S. Maltz, and T. F. Budinger, Dynamic single photon emission computed tomography-basic principles and cardiac applications, *Physics in Medicine and Biology* 55 (2010), R111.

[51] P. Heins, *Sparse model-based reconstruction in dynamic positron emission tomography*, Diploma thesis, Institute for Computational and Applied Mathematics, University of Münster, Germany, March 2011.

[52] C. Hogea, C. Davatzikos, and G. Biros, Brain-tumor interaction biophysical models for medical image registration, *SIAM J. Sci. Comput.* 30 (2008), 3050–3072.

[53] C. Hogea, C. Davatzikos, and G. Biros, An image-driven parameter estimation problem for a reaction-diffusion glioma growth model with mass effects, *J. Math. Biol.* 56 (2008), 793–825.

[54] B. K. Horn and B. G. Schunk, Determining Optical Flow, *Artificial Intelligence* 17 (1981), 185–203.

[55] Z. Jin and X. Yang, A variational model to remove the multiplicative noise in ultrasound images, *J. Math. Imaging Vision* 39 (2011), 62–74.

[56] J. Kaipio and E. Somersalo, *Statistical and computational inverse problems*, Applied Mathematical Sciences 160, Springer-Verlag, New York, 2005.

[57] J. Kaipio and E. Somersalo, Statistical inverse problems: discretization, model reduction and inverse crimes, *J. Comput. Appl. Math.* 198 (2007), 493–504.

[58] M. E. Kamasak, C. A. Bouman, E. D. Morris, and K. Sauer, Direct reconstruction of kinetic parameter images from dynamic PET data, *Medical Imaging, IEEE Transactions on* 24 (2005), 636–650.

[59] S. Kindermann, S. Osher, and P. W. Jones, Deblurring and denoising of images by nonlocal functionals, *Multiscale Model. Simul.* 4 (2005), 1091–1115.

[60] H. Kohr and A. K. Louis, Fast and high-quality reconstruction in electron tomography based on an enhanced linear forward model, *Inverse Problems* 27 (2011), 045008.

[61] Z. J. Koles, Trends in EEG source localization, *Electroencephalography and Clinical Neurophysiology* 106 (1998), 127–137.

[62] S. Lasanen, Posterior convergence for approximated unknowns in non-Gaussian statistical inverse problems, *ArXiv e-prints* (2011).

[63] T. Le, R. Chartrand, and T. J. Asaki, A Variational Approach to Reconstructing Images Corrupted by Poisson Noise, *J. Math. Imaging Vision* 27 (2007), 257–263.

[64] A. Legros, J. Cates, J. Robertson, J. Modolo, N. Juen, J. Miller, F. Prato, and A. Thomas, Simultaneous EEG/EMG/fMRI: A powerful hybrid-imaging window on brain activation patterns during and following time-varying magnetic stimuli, in: *General Assembly and Scientific Symposium, 2011 XXXth URSI*, pp. 1–4, Aug.

[65] B. A. Mair, D. R. Gilland, and Z. Cao, Simultaneous motion estimation and image reconstruction from gated data, in: *ISBI'02*, pp. 661–664, 2002.

[66] Y. Meyer, *Oscillating patterns in image processing and nonlinear evolution equations. The fifteenth Dean Jacqueline B. Lewis memorial lectures*, University Lecture Series 22, American Mathematical Society, Providence, RI, 2001.

[67] J. Modersitzki, *FAIR: flexible algorithms for image registration*, Fundamentals of Algorithms 6, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2009.

[68] M. Möller, *Multiscale Methods for (Generalized) Sparse Recovery and Applications in High Dimensional Imaging*, Ph.D. thesis, Institute for Computational and Applied Mathematics, University of Münster, Germany, July 2012.

[69] M. Möller, T. Wittman, Andrea L. Bertozzi, and Martin Burger, A variational approach for sharpening high dimensional images, *SIAM J. Img. Sci.* 5 (2012), 150–178.

[70] J. C. Mosher, M. E. Spencer, R. M. Leahy, and P. S. Lewis, Error bounds for EEG and MEG dipole source localization, *Electroencephalography and Clinical Neurophysiology* 86 (1993), 303–321.

[71] D. Mumford and J. Shah, Optimal approximations by piecewise smooth functions and associated variational problems, *Communications on Pure and Applied Mathematics* 42 (1989), 577–685.

[72] D. A. Murio, *The mollification method and the numerical solution of ill-posed problems*, A Wiley-Interscience Publication, John Wiley & Sons Inc., New York, 1993.

[73] F. Natterer, *The mathematics of computerized tomography*, Classics in Applied Mathematics 32, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001, Reprint of the 1986 original.

[74] F. Natterer, *X-ray tomography*, Inverse problems and imaging, Lecture Notes in Math. 1943, Springer, Berlin, 2008, pp. 17–34.

[75] M. Niethammer, G. L. Hart, and C. Zach, An optimal control approach for the registration of image time-series, in: *CDC'09*, pp. 2427–2434, 2009.

[76] V. T. Olafsson, D. C. Noll, and J. A. Fessler, Fast joint reconstruction of dynamic R_2^* and field maps in functional MRI, *IEEE Trans. Med. Imag.* 27 (2008), 1177–1188.

[77] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, An iterative regularization method for total variation-based image restoration, *Multiscale Model. Simul.* 4 (2005), 460–489.

[78] S. Osher, A. Solé, and L. Vese, Image decomposition and restoration using total variation minimization and the H^{-1} norm, *Multiscale Model. Simul.* 1 (2003), 349–370.

[79] R. Ramlau, R. Clackdoyle, F. Noo, and G. Bal, Accurate attenuation correction in SPECT imaging using optimization of bilinear functions and assuming an unknown spatially-varying attenuation distribution, *Z. Angew. Math. Mech.* 80 (2000), 613–621.

[80] A. J. Reader, F. C. Sureau, C. Comtat, R. Trebossen, and I. Buvat, Direct Fully 4D List-Mode Reconstruction with Temporal Prior Basis Functions, *Nuclear Science Symposium and Medical Imaging Conference Record* 4 (1992), 1955–1959.

[81] E. Resmerita, Regularization of ill-posed problems in Banach spaces: convergence rates, *Inverse Problems* 21 (2005), 1303–1314.

[82] R. Rubinstein, M. Zibulevsky, and M. Elad, Double sparsity: learning sparse dictionaries for sparse signal approximation, *IEEE Trans. Signal Process.* 58 (2010), 1553–1564.

[83] L. I. Rudin, S. Osher, and E. Fatemi, Nonlinear total variation based noise removal algorithms, *Physica D* 60 (1992), 259–268.

[84] H. Rullgård, Stability of the inverse problem for the attenuated Radon transform with 180° data, *Inverse Problems* 20 (2004), 781–797.

[85] A. Sawatzky, *(Nonlocal) Total Variation in Medical Imaging*, Ph.D. thesis, Institute for Computational and Applied Mathematics, University of Münster, Germany, July 2011.

[86] C.-B. Schönlieb and A. Bertozzi, Unconditionally stable schemes for higher order inpainting, *Commun. Math. Sci.* 9 (2011), 413–457.

[87] S. Setzer, G. Steidl, and T. Teuber, Infimal convolution regularizations with discrete ℓ_1-type functionals, *Commun. Math. Sci.* 9 (2011), 797–827.

[88] A. Singer, Y. Shkolnisky, and B. Nadler, Diffusion interpretation of nonlocal neighborhood filters for signal denoising, *SIAM J. Img. Sci.* 2 (2009), 118–139.

[89] G. Steidl and T. Teuber, Removing multiplicative noise by Douglas-Rachford splitting methods, *J. Math. Imaging Vision* 36 (2010), 168–184.

[90] R. Stück, M. Burger, and Th. Hohage, The iteratively regularized Gauss–Newton method with convex constraints and applications in 4Pi microscopy, *Inverse Problems* 28 (2012), 015012.

[91] S. Suhr, *Bewegungskorrigierte PET-Rekonstruktion*, Master thesis, Institute for Computational and Applied Mathematics, University of Münster, Germany, March 2012.

[92] B. P. Sutton, D. C. Noll, and J. A. Fessler, Fast iterative image reconstruction for MRI in the presence of field inhomogeneities, *IEEE Trans. Med. Imag.* 22 (2008), 17–188.

[93] D. Tenbrinck, A. Sawatzky, X. Jiang, M. Burger, W. Haffner, P. Willems, M. Paul, and J. Stypmann, Impact of Physical Noise Modeling on Image Segmentation in Echocardiography, in: *Proceedings of the 3rd Eurographics Workshop on Visual Computing for Biology and Medicine (VCBM) 2012* (T. Ropinski, A. Ynnerman, C. Botha, and J. Roerdink, eds.), 2012.

[94] D. Trede, Exact support recovery for linear inverse problems with sparsity constraints, *Methods Appl. Anal.* 18 (2011), 105–110.

[95] J. A. Tropp, Just relax: convex programming methods for identifying sparse signals in noise, *IEEE Trans. Inform. Theory* 52 (2006), 1030–1051.

[96] M. Uecker, T. Hohage, K. T. Block, and J. Frahm, Image Reconstruction by Regularized Nonlinear Inversion–Joint Estimation of Coil Sensitivities and Image Content, *Magnetic Resonance in Medicine* 60 (2008), 674–682.

[97] Y. Vardi, L. A. Shepp, and L. Kaufman, A statistical model for positron emission tomography, *J. Amer. Statist. Assoc.* 80 (1985), 8–37.

[98] L. A. Vese and S. J. Osher, Image denoising and decomposition with total variation minimization and oscillatory functions, *J. Math. Imaging Vision* 20 (2004), 7–18.

[99] S. W. Wagner, R. F. and Smith, J. M. Sandrik, and H. H. Lopez, Statistics of speckle in ultrasound B-scans, *IEEE Trans. Sonic Ultrason.* 30 (1983), 156–163.

[100] K. Wang and M. A. Anastasio, *Photoacoustic and Thermoacoustic Tomography: Image Formation Principles*, Handbook of Mathematical Methods in Imaging (Otmar Scherzer, ed.), Springer New York, 2011, pp. 781–815.

[101] L. Wang, L. Xiao, L. Huang, and Zh. Wei, Nonlocal total variation based speckle noise removal method for ultrasound image, in: *Image and Signal Processing (CISP), 2011 4th International Congress on*, 2, pp. 709 –713, Oct. 2011.

[102] A. Welch, R. Clack, F. Natterer, and G. T. Gullberg, Towards accurate attenuation correction in SPECT without transmission scan, *IEEE Trans. Med. Imag.* 16 (1997), 532–541.

[103] M. N. Wernick and J. N. Aarsvold (eds.), *Emission Tomography: The Fundamentals of PET and SPECT*, Elsevier Science, 2004.

[104] C. Wu and X. Tai, Augmented Lagrangian Method, Dual Methods, and Split Bregman Iteration for ROF, Vectorial TV, and High Order Models, *SIAM J. Img. Sci.* 3 (2010), 300–339.

[105] J. Yan, B. Planeta-Wilson, and R. E. Carson, Direct 4D list mode parametric reconstruction for PET with a novel EM algorithm, in: *Nuclear Science Symposium Conference Record, 2008. NSS '08. IEEE*, pp. 3625–3628, Oct. 2008.

[106] R. Zanella, P. Boccacci, L. Zanni, and M. Bertero, Efficient gradient projection methods for edge-preserving removal of Poisson noise, *Inverse Problems* 25 (2009), 045010.

[107] X. Zhang, M. Burger, and S. Osher, A Unified Primal-Dual Algorithm Framework Based on Bregman Iteration, *Journal of Scientific Computing* 46 (2011), 20–46.

[108] H. Zou, T. Hastie, and R. Tibshirani, Sparse principal component analysis, *J. Comput. Graph. Statist.* 15 (2006), 265–286.

Kees van den Doel, Uri M. Ascher and Eldad Haber

The lost honor of ℓ_2-based regularization

Abstract: In the past two decades, regularization methods based on the ℓ_1 norm, including sparse wavelet representations and total variation, have become immensely popular. So much so, that we were led to consider the question whether ℓ_1-based techniques ought to altogether replace the simpler, faster and better known ℓ_2-based alternatives as the default approach to regularization techniques.

The occasionally tremendous advances of ℓ_1-based techniques are not in doubt. However, such techniques also have their limitations. This article explores advantages and disadvantages compared to ℓ_2-based techniques using several practical case studies. Taking into account the considerable added hardship in calculating solutions of the resulting computational problems, ℓ_1-based techniques must offer substantial advantages to be worthwhile. In this light, our results suggest that in many applications, though not all, ℓ_2-based recovery may still be preferred.

Keywords: Inverse problems, image deblurring, image reconstruction, ℓ_1-regularization, ℓ_2-regularization, compressed sensing

2010 Mathematics Subject Classification: 65F22, 65N21, 35R20, 92C55, 65M32

Kees van den Doel: Department of Computer Science, University of British Columbia, Vancouver, Canada, kvdoel@cs.ubc.ca
Uri M. Ascher: Department of Computer Science, University of British Columbia, Vancouver, Canada, ascher@cs.ubc.ca
Eldad Haber: Departments of Earth & Ocean Science and Mathematics, University of British Columbia, Vancouver, Canada, haber@math.ubc.ca

1 Introduction

Ill-posed problems typically require some regularization in order to compute a credible approximate solution in a stable, well-defined manner. In this article, we consider such problems where the objective is to recover a function $u(\mathbf{x})$, with $\mathbf{x} \in \Omega \subset \mathbb{R}^d$ (typically $d = 2$ or $d = 3$), from observed and discrete data b. Given is a forward operator, $F(u)$, which predicts data for any suitable function u, and the challenge is to find u such that the predicted data match the observed data to within a reasonable tolerance.

This work was supported in part by NSERC Discovery Grant 84306.

It is convenient for our discussion at this point to consider a linear forward operator, with u discretized on some mesh in Ω and reshaped as a vector of unknowns $\mathbf{u}$, and with the observed and predicted data likewise written as $\mathbf{b}$ and $F(\mathbf{u}) = J\mathbf{u}$, respectively. Here, J is an $m \times n$ sensitivity matrix, $m \leq n$, which often has a nontrivial null space. Then, we write down the Tikhonov-type regularized problem [25, 60, 61]

$$\min_{\mathbf{u}} \frac{1}{2} \|J\mathbf{u} - \mathbf{b}\|_2^2 + \beta R(\mathbf{u}) , \tag{5.1}$$

where $\|\cdot\|_p$ denotes the usual vector ℓ_p norm, $\beta > 0$ is a parameter, and R is a regularization operator. We focus on the following possibilities for R:

(1) Consider

$$R(\mathbf{u}) = \frac{1}{p} \|W\mathbf{u}\|_p^p , \tag{5.2}$$

for the choices $p = 1$ (referred to as L1) or $p = 2$ (referred to as L2). Here, W is an $n \times n$ weight matrix, e.g. some wavelet or curvelet transform, or just the identity [10, 24, 33]. For notational purposes, we stipulate that W is not a discretized gradient operator.[1]

(2) Recalling that $\mathbf{u}$ represents a discretization of a function $u(\mathbf{x})$ on Ω, choose $R(\mathbf{u})$ to be an appropriate discretization of

$$\mathcal{R}(u) = \frac{1}{p} \int_{\Omega} |\nabla u|^p , \tag{5.3}$$

again considering the cases $p = 2$ or $p = 1$. The case $p = 2$ leads to a discretization of the Laplacian operator on Ω when considering necessary conditions for the minimization (5.1): denote this by L2G. The case $p = 1$ leads to total variation [51, 53]: denote this by L1G.[2]

For many years, the almost automatic choices of regularization in (5.2) and (5.3) have been based on the ℓ_2-norm, i.e. $p = 2$. This yields a straightforward linear least squares problem that can be effectively solved even when the problem is very large (see, e.g. [32, 55]). Large computational problems are manageable even if F is nonlinear in u, and R is more complex but still ℓ_2-based (see, e.g. [16, 17, 29]). Furthermore, the ℓ_2-based regularization enjoys a favorable statistical interpretation for models

[1] Of course, wavelet function bases do approximate derivatives as well. For instance, our distinction as such is particularly blurred by tight frame wavelets [7]. However, the distinction of L1 from L1G should be intuitively clear. Note also that one can always transform L1 and L2 by a change of variables into a form where W becomes the identity. However, we retain our notational redundancy for convenience.

[2] Note that the gradient magnitude $|\nabla u|$ is the ℓ_2 norm of ∇u. Thus, the L1G expression is one of a discrete ℓ_1 norm only if $d = 1$. Also, a further regularization is required when using L1G upon considering necessary conditions for (5.1); see, e.g. [1].

with a prior that is normally distributed [8, 41, 58, 61]. In the past two decades, however, regularization methods based on the ℓ_1-norm (i.e. $p = 1$ in (5.2) and (5.3)) have become immensely popular; see, e.g. the books [24, 46, 51]. In fact, we have been led to consider the idea that ℓ_1-based techniques should altogether replace the simpler, faster and better known ℓ_2-based alternatives. There are two essential motivations for this exciting trend.

- It is natural to choose for the regularization term a penalty function as in (5.3), thus expressing the *a priori* information that $u(\mathbf{x})$ ought to be smooth. However, if $u(\mathbf{x})$ has jump discontinuities, then using L2G essentially smears out such discontinuities because the Dirac δ-function is not square integrable. On the other hand, the δ-function is integrable, and thus using L1G better accommodates jump discontinuities.
- Whether the term R is aimed at penalizing the magnitude of the gradient or the solution itself, the ℓ_1-based regularization tends to produce sparse approximations. In the L1G context, this is expressed in the observation that the reconstruction tends toward being piecewise constant, so the gradient is mostly zero and thus sparse. In the L1 wavelet (or DCT) approximation context, where $W\mathbf{u}$ in (5.2) corresponds to coefficients of different wavelet (or cosine) basis functions, a compressed approximation involving only a few basis functions often results (unlike the case when using $p = 2$).

The rather fundamental importance of the above two reasons for using $p = 1$ is not in doubt. Among many other researchers, we have contributed to this volume of work [1, 30, 36]. We have found that for well-conditioned problems with sufficient high-quality data,[3] ℓ_1-based regularization can, in many cases, "deliver on its promise." However, for problems with poor data, or ill-conditioned problems typically resulting from discretizations of highly ill-posed problems, we have found that this is often not the case. To demonstrate and motivate the ensuing discussion, let us consider the following example.

Example 5.1 (Image Deblurring). Let J be a discretization of a known image blurring operator and $\mathbf{u}$ be an image reshaped into a vector. Our goal is to recover the clean image given noisy blurred data. For the following numerical experiments, we have used three codes: (i) RestoreTools [33], which employs an $L2$-type recovery strategy (viz. $p = 2$ and $W = I$ in (5.2)); (ii) the GPSR package [26], which employs a wavelet L1 recovery algorithm; and (iii) a straightforward total variation (L1G) code. The above two packages, in our opinion, are both excellent representations of good software for the problems they aim to solve. However, the $L2$ code requires, comparatively speaking,

[3] We further explain in Section 3 what we mean by the intuitive terms "high-quality" versus "poor" data.

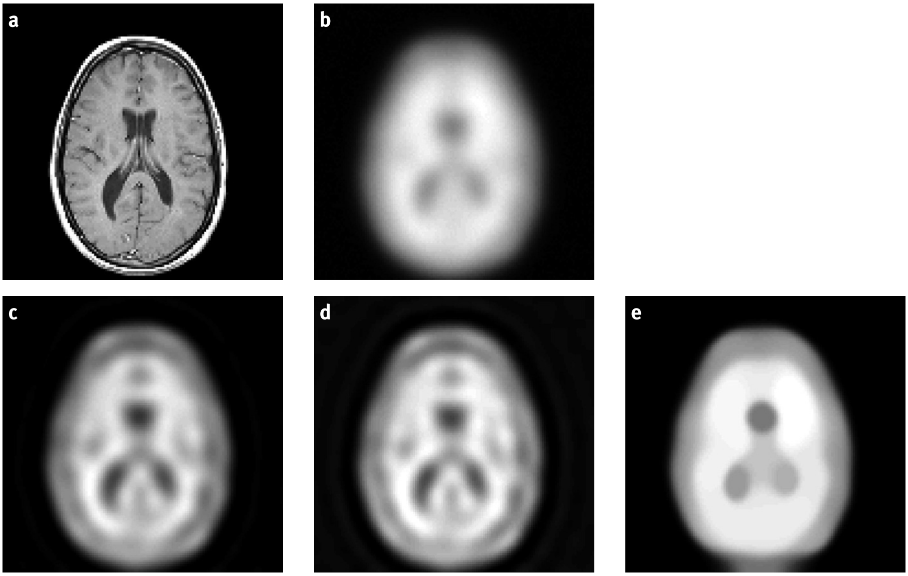


Figure 5.1: The ground truth image (a) is blurred and corrupted by noise to create the data (b). Recovered solutions obtained for this data by RestoreTools (L2), GPSR (L1) and total variation (L1G) are displayed in (c–e), respectively.

only a small fraction of computational time to terminate successfully, and hence it is to be preferred unless the $L1$ reconstructions are demonstrably better.

The "true image," or ground truth is a 128×128 MRI image from Matlab's collection. The blurring kernel is $e^{-\|\mathbf{x}\|_2^2/2\sigma}$ with $\sigma = 0.01$ and the blurred data is further corrupted by 1 % white noise. In all three methods, the data is fit to an accuracy of 1 % by tuning the regularization parameter β (see, e.g. [61]). The results are presented in Figure 5.1.

It is apparent that, at least for this problem, the ℓ_1-based reconstructions do not yield more pleasing results than the simple ℓ_2-based one. The L1G image is typically blocky, and in the present context, it may be considered the worst of the three: indeed, sparsity of the surface gradient is not a good regularization objective here. The first two recoveries are more comparable in terms of quality. In fact, it may be argued that the $L2$ result is altogether better than the ℓ_1-based ones.

Image deblurring is a favorite application in the literature for discussing and comparing both L1 and L1G techniques. Indeed, in many such examples, ℓ_1-based regularization is to be preferred (see, e.g. [11, 24, 36]). However, Example 5.1 is by no means esoteric. Furthermore, similar comparative observations arise when working on certain nonlinear ill-posed problems such as electrical impedance tomography (EIT) and direct current (DC) resistivity [1]; we return to this in Section 4.

The goals of this paper are therefore to explore, bearing in mind the occasionally impressive advances of ℓ_1-based regularization techniques, also some of their limitations. Taking into account the often considerable added hardship in calculating solutions of the resulting computational problems, ℓ_1-based techniques must offer substantial advantages to be worthwhile. In this light, our results suggest that in many applications, ℓ_2-based recovery may be preferred. To this end, we provide the following cautionary notes:

(1) Only the left term in the objective function of (5.1) is really mandated by the stated data fitting problem. The choice of regularization is discretionary: different choices may generally yield different solutions that as such must all be considered acceptable. The further specification of regularization reflects a prior which depends on additional knowledge that may or may not be truly available.

(2) It is not true that one must always seek a sparse approximate solution, especially if an appropriate basis to span the solution is not known.

(3) Codes such as those reported in [3, 4, 26, 42], which perform well when applied in the context of using wavelets for denoising or deblurring, may occasionally perform relatively poorly when applied in a wider context.

(4) In our experience, if the data is not of sufficiently high quality, in the sense that there is too much noise, then ℓ_1-based methods may occasionally perform worse than the corresponding ℓ_2-based methods.

(5) If the data is not of sufficiently high-quality, in the sense that it is too sparse or rare, then ℓ_1-based methods may occasionally perform worse than the corresponding ℓ_2-based methods.

(6) If the computational problem is highly ill-conditioned, then ℓ_1-based methods may occasionally perform worse than the corresponding ℓ_2-based methods.

In this paper, we explore examples, or case studies, which demonstrate the claims above and explain when ℓ_2-based methods merit prime consideration. Some analysis is also provided. We group our discussion into two classes: problems with poor data, considered in Section 3, and highly ill-conditioned problems, considered in Section 4. The latter section is far longer and more involved than the others, and Theorem 5.4, as well as the analysis in Section 4.1, are new. Before these, Section 2 provides a quick review of ℓ_1-based regularization. We review the theory and the requisite assumptions necessary for ℓ_1-based recovery to perform well.

Finally, we summarize the paper in Section 5.

2 ℓ_1-based regularization

Several books, e.g. [11, 24, 46, 51, 56], contain descriptions of ℓ_1-based regularization methods in the context mentioned earlier, and it is not our intention to reproduce them here. We only touch upon a few items. For early efforts in geophysics and data assimilation, see [14, 59]. For advanced uses of such methods in machine learning, see, e.g. [47, 49].

In the context of a discrete cosine or a wavelet-type transform, the problem (5.1) may be viewed as a noisy version of the problem

$$\min_{\mathbf{u}} R(\mathbf{u}) , \tag{5.4a}$$

$$\text{s.t. } J\mathbf{u} = \mathbf{b} , \tag{5.4b}$$

where J has a full row rank $m < n$. Note that this can be a well-conditioned problem for both choices of p in (5.2). For L1 (i.e. $p = 1$ in (5.2)), problem (5.4) can be cast as a linear programming problem, and linear programming theory already guarantees that there is an optimal basic feasible solution which is m-sparse (i.e. with only at most m nonzero components) [50, 62]. In contrast, when using L2, all components of the optimal $\mathbf{u}$ are typically nonzero.

This has been well known since at least the 1960s. Moreover, though, since the above transforms utilize elaborate basis functions, it is reasonable to expect that much fewer than m basis functions may suffice, corresponding to a much sparser solution. The discovery [13, 20, 22] that using L1 often yields such a sparse solution, effectively solving a very hard combinatorial problem, is much newer and constitutes a major breakthrough.

However, it is not always the case that the solution of the constrained optimization problem using the ℓ_1 norm yields a sparse solution. Furthermore, for (5.1) in general, it does not automatically follow that if such a sparse solution exists, it is an appropriate estimate of the true solution, see [19] and Section 4.1.

Much effort has been devoted to the question, namely, under what conditions the ℓ_1 solution of (5.4) produces the sparsest possible solution of (5.4b), referred to as the ℓ_0 solution. Of course, a more practical goal would probably be to seek a "sufficiently sparse" solution, but the quest for optimum in this regard sheds light on what is required more generally. The restricted isometry property (RIP) [9] and the null space property of [15, 21] both provide sufficient conditions, whereas the γ-condition of [40] is both necessary and sufficient for obtaining the sparsest solution by L1.

These conditions are of great value for understanding the design of compressed sensing methods. Unfortunately, though, for realistic instances of the matrix J, they are generally intractable (NP-hard) to verify numerically. Moreover, in Section 4.1, we show that such conditions are violated for a specific case of the inverse potential problem when attempting to recover a pair of point charges by ℓ_1-based methods.

The vector norm function $\| \cdot \|_p$ is well known to be convex only when $p \geq 1$. Thus, ℓ_1 is marginally convex. Even more sparsity-inducing is the use of a nonconvex norm with $0 < p < 1$ [12, 44, 54]. However, there is a price to pay for lack of convexity, in terms of both poorer theory and the necessity of convergent algorithms which typically apply a continuation (homotopy) procedure starting from a convex ℓ_p-norm.

Several famous codes cited earlier for solving (5.4) use methods that are based on gradient projection with acceleration (see, for instance, the extended Chapter 6 of [5] and references therein). The advantage of these methods is that they extend directly to problems with nonsmooth constraints and require the objective function gradient to be only Lipschitz continuous. However, bear in mind that for solving simple unconstrained convex quadratic problems, such methods boil down to accelerated gradient descent without preconditioning, generally thought to be unforgivably slow. These methods seem to work well for compressed sensing problems because the corresponding problems (5.4) are well-conditioned in an appropriate sense. Unfortunately, other applications involving, for instance, PDE-constrained optimization (as in Section 4), are highly ill-conditioned and therefore, similar numerical optimization methods should not be expected to be robust and efficient in the latter context.

Total variation (L1G) has been discovered and peaked earlier than sparse wavelet basis reconstruction and compressed sensing. The books [11, 51, 61] and many papers develop both theory and algorithms using this approach. In practice, some regularization such as a Huber switching function [56] is often used, and this really gives a mix of ℓ_1 with ℓ_2 elements while still retaining the L1G spirit [1]. See also [6] for another approach to round excessive L1G sharpness. Moreover, one popular iterative scheme to carry out the resulting algorithm is lagged diffusivity, which is a special case of iteratively reweighted least squares (IRLS) [1, 61].

Unlike the case for wavelet-type solutions, where a sparse representation is sought for the same high-quality surface or image approximation, here the regularization is applied directly to the surface variables to be recovered. Along with the advantage in directly penalizing piecewise smoothness, the tendency of the L1G regularization to give sparse gradients, translating into a “blocky image,” is not always what one necessarily wants (see, e.g. Figure 5.1 (e)) L1G penalizes large jumps in the solution more than small jumps, and this may introduce distortion in the reconstructed surface. Various nonconvex alternatives to L1G are listed in [56], for instance, and these occasionally yield sharper results for some applications. However, the nonconvex nature of these regularizations again leads to both theoretical and practical additional difficulties.

Our focus in this article is on exploring situations where use of the L1 or L1G regularization ($p = 1$ in (5.2), (5.3)) may reasonably be compared to use of L2 or L2G ($p = 2$ in (5.2), (5.3)). Therefore, employing any of the even sharper nonconvex options mentioned above is not under further consideration.

The above synopsis has been restricted to linear problems. There is very little ℓ_1 theory for nonlinear problems. Moreover, it is easy to see that some of the basic sparsity arguments fail for this case. Consider the problem

$$\begin{aligned} &\min_{\mathbf{u}} \|\mathbf{u}\|_1 \\ &\text{s.t. } F(\mathbf{u}) = \mathbf{b}\,, \end{aligned}$$

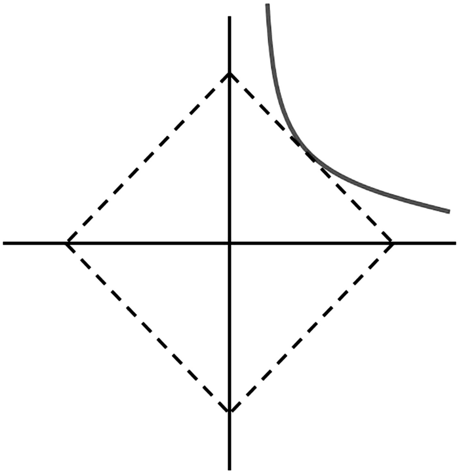


Figure 5.2: When the constraint (solid) is nonlinear, it does not need to intersect the level set of $\|\mathbf{u}\|_1$ (dashed) at a vertex, so the solution is not necessarily sparse.

where the forward mapping function $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is smooth and has significant curvature (see Figure 5.2). In such a case, the problem need not even have m-sparse solutions; indeed, the optimal solution may have n nonzero entries. Thus, the justification of using L1 for nonlinear problems is far from obvious. On the other hand, L1G is interesting because of its sharpening property. In Section 4.3, we explore the use of L1G for a particular popular nonlinear case study.

3 Poor data

The perceived quality of a given data set depends on several factors, and not simply on some idealized noise level. One of these is the inverse problem operator. For instance, in Example 5.1, the deblurring operation, which is essentially to improve contrast and sharpness of the image, counters an image smoothing operation which aims to remove noise. Thus, a noise level in the data which may otherwise be considered benign (say, in a pure denoising application) can be an important obstruction here.

In the context of data fitting, it has been known for decades that ℓ_1 data fitting is more robust than ℓ_2 against outliers in the data. See, for instance, [50] and also [45] for a recent use in the context of 3D graphics. However, such a comparative statement does not necessarily hold true for other types of noise such as white noise.

In general, bearing in mind the additional complications in carrying out ℓ_1-based regularization, the data must be of sufficiently high-quality to allow its favorable properties (when relevant) to be expressed. A common situation yielding lack of sufficiently good data is when the data is relatively rare, being given only at relatively few locations in Ω. Let us next discuss a simple example where the data locations are rare (or sparse) in the domain of the definition.

Example 5.2 (Rare data reconstruction of piecewise smooth functions). Consider the recovery of a (real) signal $u^*(t)$ on $[0,1]$ from m noisy samples $u_i \approx u^*(t_i)$, and assume we know that u^* is piecewise smooth, but may have jump discontinuities. We discretize the interval $[0,1]$ with a uniform grid of $n = 512$ points, and use in a given experiment a subset of $m \ll n$ samples taken at random points t_j from this grid. The integral appearing in (5.3) is discretized using a piecewise linear function $u(t)$ on all n grid points. Thus, the recovery problem is formulated as in (5.1), with J being the $m \times n$ matrix consisting of m columns forming an identity matrix interspersed with $n - m$ zero columns. In the limit case of no noise, the formulation (5.4) yields interpolation through the data points (t_i, u_i) of the sample.

We compare L2G and L1G regularizations. It is easy to verify that in the L2G case, these data points are connected by straight lines, whereas with L1G (total variation) regularization, the behavior is indeterminate, only restricting u to be monotone.

Figures 5.3 and 5.4 depict reconstruction results for $m = 9$ and $m = 28$ samples. The ground truth signal $u^*(t)$ contains two jumps, and we added 5 % Gaussian noise to the selected values to form the corresponding data sets. Figure 5.3 shows the result for 9 samples, with the regularization parameters tuned by the discrepancy principle to obtain a data misfit of 5 ± 0.1 %. There is little difference between the L1G and L2G reconstructions.

The reconstruction in Figure 5.4 (a) using 28 samples starts to show the advantages of L1G. Because the data contains two samples across the right discontinuity, the regularization parameter β_{L2G} now had to be decreased to $\beta_{\text{L2G}} = .002$ in order to obtain the desired misfit of roughly 5 %. As a result, the L2G reconstruction exhibits

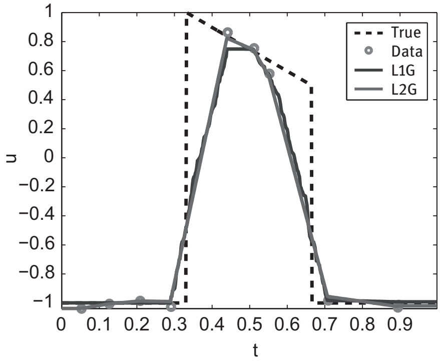


Figure 5.3: Reconstructions of a piecewise smooth function from a few noisy samples: using L2G and L1G for $m = 9$ data pairs, with $\beta_{\text{L2G}} = 0.04$, $\beta_{\text{L1G}} = 0.08$.

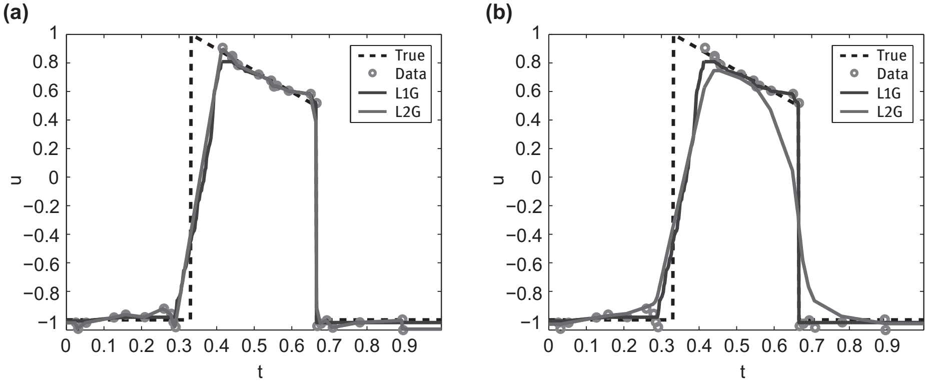


Figure 5.4: Reconstructions of a piecewise smooth function from a few noisy samples: using L2G and L1G for $m = 28$ data pairs. (a) $\beta_{\mathrm{L2G}} = 0.002$, $\beta_{\mathrm{L1G}} = 0.08$, (b) $\beta_{\mathrm{L2G}} = 0.02$, $\beta_{\mathrm{L1G}} = 0.08$

considerable oscillation in the flat sections, although note that the second jump is reproduced as well as by the L1G method. In Figure 5.4 (b), we increased β_{L2G} until the flat sections became reasonably smooth according to the "eyeball norm." Observe that the oscillation has disappeared, but the second jump is now completely blurred as well.

Example 5.2 illustrates that L1G regularization performs well when there is enough quality data to require the reconstructed model to have discontinuities. However, when the data is "too sparse", L2G regularization performs as well as L1G, even in the presence of discontinuities in the underlying ground truth function. This lesson seems perhaps obvious in hindsight. However, it extends to more complex situations where the insight is no longer so obvious. For instance, the problems considered in Section 4 have data specified only at the boundary of a given physical domain Ω, which is a lower-dimensional manifold; several examples can be found in the literature where some L1G variant is applied to such problems. For another instance, consider a point cloud in 3D, obtained as a set of somewhat noisy and not very dense 3D laser scan measurements of a body with edges, such as a desk corner. In order to obtain a good surface reconstruction, we need at each point the normal to the surface that the (cleaned) point cloud represents [38]. Since the curvature across an edge is infinite, the data can be effectively very sparse there, and indeed a global ℓ_1-reconstruction approach [2] might not work well then See Figure 10 in [39] for such an example. Poor data are often encountered in ocean and atmospheric data assimilation, as well as in other time-dependent geophysical applications [23, 27].

4 Large, highly ill-conditioned problems

In this section, we consider applying ℓ_1-based techniques to large, highly ill-conditioned problems that typically arise in applications involving PDE-constrained optimization. As a first example, we show in § 4.1 that L1 techniques may not only be expensive to carry out, but also have difficulty in producing solutions which are as sparse as a given ground truth. In § 4.2, we then supply some analytical evidence supporting this observation. Finally, in § 4.3, we show by another example that while L1G is not nearly as severely afflicted as L1, its advantage over L2G in recovering surface discontinuities requires favorable conditions to shine through.

4.1 Inverse potential problem

In the inverse potential problem, one seeks to recover an electrical source distribution in a given domain Ω from measurements of the potential on the domain's boundary. This problem arises in EEG source modeling [48] and in electromyography [18, 19]. In [19], the sought source is a combination of discrete tripoles corresponding to muscle fibers, and as such invites a sparse reconstruction. However, in 3D, the computational problem using L1 indeed became much too large and difficult to work with, and our eventual success in solving the research problem stated in [19] followed a further realization that, given the specific goals of those computations, the sparse view was not the most effective. This has left the question open regarding what an L1 reconstruction can do for such a problem (regardless of cost), a question that we now proceed to explore in a more manageable 2D context, with Ω being the unit square.

The forward model

$$-\Delta v = u(\mathbf{x}), \quad \mathbf{x} \in \Omega, \tag{5.5}$$

with Neumann boundary conditions on v, predicts the potential v for given electrical source u. The total charge must be zero due to these boundary conditions. Note that $v(\mathbf{x})$ is only determined up to an overall additive constant, reflecting the physical principle that only a potential difference is physically meaningful.

The inverse problem of finding u from values of v on the boundary does not have a unique solution, even under idealized conditions [35]. The best one can do is construct an "equivalent source" u that explains the data. Such a reconstruction gives incomplete, though still useful, information about the actual source. Hence, the role of regularization is to provide additional information leading to a distribution u that conforms to prior expectations, a rather fundamental difference from sparse signal reconstruction. Denoting the discretized Poisson operator of (5.5) by A and the data projection operator by Q (see [19] for details), we obtain a problem in the form (5.1) with

$$J = QA^{-1}. \tag{5.6}$$

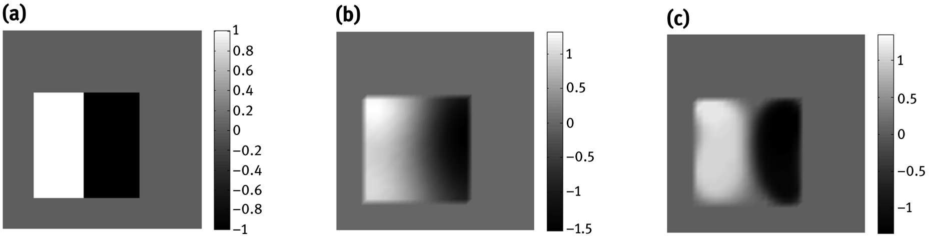


Figure 5.5: Reconstructions of a piecewise constant charge distribution from boundary data. (a) True, (b) L2G, (c) L1G.

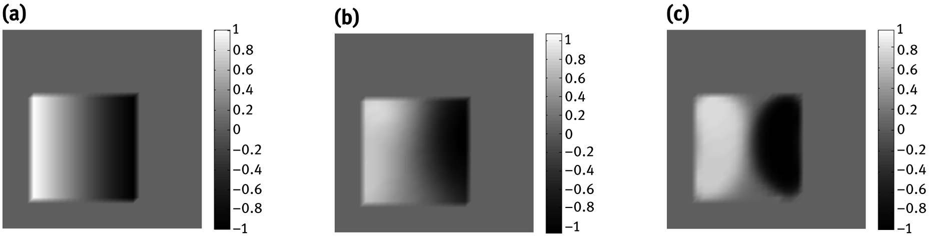


Figure 5.6: Reconstructions of a smoothed-step charge distribution from boundary data. (a) True, (b) L2G, (c) L1G

Example 5.3 (Inverse potential problem). In this numerical experiment, the support of the source u is restricted to the offset inner square (assumed to be known in the reconstruction) as depicted in Figures 5.5–5.7. The potential is measured on the boundary, taking the average boundary potential as the ground level, i.e. we subtract the average boundary potential from each datum. This is necessary as only potential differences are measurable. Figures 5.5–5.7 depict results for three different source distributions in the region. In each case, synthetic data is computed on a 64^2 grid, to which we add a 1 % Gaussian noise. The reconstruction is done with our various regularizations (5.3) and (5.2) on a 32^2 grid. The regularization constant β is tuned to obtain a resulting misfit of 1 ± 0.1 % (see, e.g. [61]).

Figures 5.5 and 5.6 serve as an appetizer We consider, respectively, piecewise constant and smoothed-step dipole distributions. Observe that the $L1G$ reconstruction results in a well-defined interface between the positively and negatively charged regions, whereas the $L2G$ reconstruction is smooth, irrespective of the true model. As such, the use of $L1G$ is especially recommended if we know *a priori* that u is piecewise smooth. However, it is not possible to determine from the reconstructions whether u has a jump or not: notice the similarity between Figures 5.5 (b) and 5.6 (b), and that between Figures 5.5 (c) and 5.6 (c).

Next, we explore the main theme of this section by considering a point charge pair. The true model (ground truth) depicted in Figure 5.7 (a) is now very sparse.

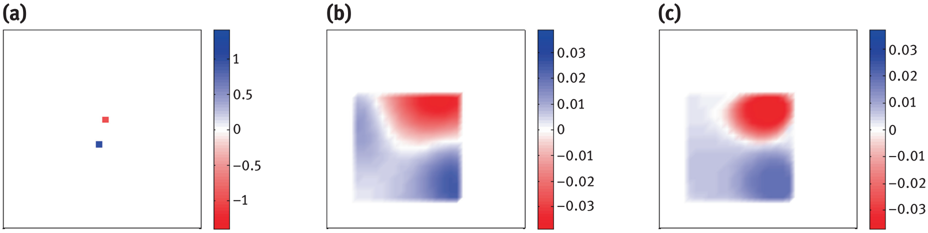


Figure 5.7: Reconstructions of a point charge pair from boundary data using gradient regularization. (a) True, (b) L2G, (c) L1G.

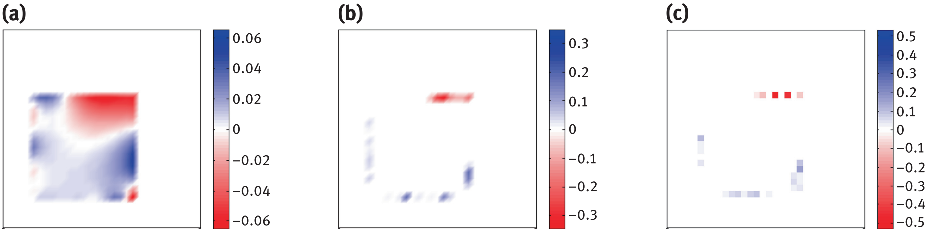


Figure 5.8: Reconstructions of a point charge pair from boundary data using regularizations (5.2). (a) L2, (b) L1, (c) Weighted L1

The results shown in Figure 5.7 are similar to those in Figures 5.5 and 5.6. The dipole structure is apparent from the $L2G$ and $L1G$ reconstructions, though not much more is. The $L1G$ reconstruction hints at a dipole pair, but may mislead one to infer an incorrect orientation.

For this last source distribution, a sparse reconstruction seems natural, and one such, obtained using an $L1$ regularization, is depicted in Figure 5.8 (b). The $L2$ reconstruction is depicted in Figure 5.8 (a) for comparison. We see that the $L1$ reconstruction is somewhat sparse, but all the reconstructed sources are on the boundary of the support of $u(\mathbf{x})$, and the $L1$ solution is not as sparse as the true model.

The reason for the observed source distribution is that sources near the detector affect the data more and are therefore favored [31]. This effect can be reduced by a location dependent reweighting of the regularization function as suggested in [28, 43], which amounts to normalizing the columns of J to unit 2-norm. Letting

$$a_j = \left(\sum_{i=1}^{m} (J_{ij})^2 \right)^{1/2} \qquad \hat{J}_{ij} = J_{ij}/a_j \quad \hat{u}_i = a_i u_i \, ,$$

we can write $J\mathbf{u} = \mathbf{b}$ as $\hat{J}\hat{\mathbf{u}} = \mathbf{b}$ and apply the $L1$ regularization to $\hat{\mathbf{u}}$. (Note though that computing a_i for large scale problems may not be practical.) The resulting reconstruction is depicted in Figure 5.8 (c). The sparsity has improved a little, but we are still far from the ℓ_0 solution.

For this example, since $\hat{J}$ has normalized columns, the famous RIP condition defined and analyzed in [9] applies. This condition requires that there be a $\delta \le \sqrt{2} - 1$ such that for all 4-sparse $\mathbf{u}$, we have

$$(1-\delta)\,\|\hat{\mathbf{u}}\|_2^2 \le \|J\mathbf{u}\|_2^2 \le (1+\delta)\,\|\hat{\mathbf{u}}\|_2^2 \,. \tag{5.7}$$

However, here it can be shown to be violated on physical grounds. Let $\mathbf{u}$ be a 4-sparse source, i.e. nonzero only for indices i in a set $\mathcal{T}$ with $|\mathcal{T}| = 4$, and further, let it have values ± 1, so

$$\|\hat{\mathbf{u}}\|_2^2 = \sum_{i \in T} a_i^2 \ge 4 \min_k \left(a_k^2\right) > 0\,.$$

(The value a_i is just the 2-norm of the boundary potential when a unit source is placed at location i.) Note that $\|J\mathbf{u}\|_2^2$ is the ℓ_2 norm of the boundary potential. By placing the positive and negative charges very close together, so that they almost cancel each other, we can make the boundary potential and thereby $\|J\mathbf{u}\|_2^2$ arbitrary small, and thus δ becomes arbitrarily close to 1. Hence, the RIP condition is violated. Note that this does not prove that the sparsest solution cannot be obtained, as the RIP is a sufficient, though not necessary condition.

The necessary and sufficient γ-condition of [40] for obtaining the ℓ_0 solution from the ℓ_1 solution relies on properties of the solution $\mathbf{y}$ to the equation

$$(J^T\mathbf{y})_i = z_i\,, \tag{5.8}$$

for selected indices i such that $z_i \neq 0$. In our case, to determine if it is possible to recover a 2-sparse source, the n-vector $\mathbf{z}$ should be 2-sparse with entries ± 1, so (5.8) has just two equations. Further, $J^T\mathbf{y} = A^{-1}Q^T\mathbf{y}$, and we can interpret $\mathbf{y}$ as describing electrical sources on the boundary only, such that the generated potential equals 1 at point $\mathbf{p}_1$ and -1 at point $\mathbf{p}_2$. These correspond to the location of the point charges described by $\mathbf{z}$. The γ-condition then implies that we can find a $\mathbf{y}$ such that the potential $J^T\mathbf{y}$ is between -1 and 1 everywhere else. Unfortunately, however, on physical grounds, we can see that this is not possible. To see this note that if we place $\mathbf{p}_1$ and $\mathbf{p}_2$ very close together, then a very large electrical field will exist between the points, which must be caused by very large boundary sources, which in turn will generate close to those sources an even larger (> 1) field. Analytically, we observe that in the continuum limit, since $\mathbf{z}$ is a harmonic function, it must take its extreme values on the boundary. Since it takes on values ± 1 inside, it must take on larger values on the boundary, and hence the γ-condition is violated.

4.2 The effect of ill-conditioning on L1 regularization

In this subsection, we consider the regularized L1 problem

$$\min_{\mathbf{u}} \frac{1}{2}\,\|J\mathbf{u} - \mathbf{b}\|_2^2 + \beta\,\|W\mathbf{u}\|_1\;, \tag{5.9}$$

and show, for a special choice of W, which in a sense favors sparsity, that in the highly ill-conditioned case and in the presence of noise, the correct sparsity of a ground truth model can be recovered only if the singular values of J and the sparsity structure combine in a beneficial manner. This helps explain the negative results of Example 5.3.

Let the singular value decomposition (SVD) of the $m \times n$ matrix J be given by

$$J = U\Sigma V^{\top},$$

where U and V are orthogonal matrices and $\Sigma = \text{diag}\{\sigma_1, \ldots, \sigma_m\}$ is $m \times n$ with the singular values ordered so that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m$. Further, consider a true model $\mathbf{u}^*$ such that $\mathbf{z}^* = V^T\mathbf{u}^*$ satisfies

$$z_i^* = \begin{cases} 1 & i \in \mathcal{T} \\ 0 & i \notin T \end{cases}. \tag{5.10}$$

The emphasis in (5.10) is on the nature of $\mathcal{T}$, i.e. the sparsity: setting the nonzero values to 1 is just for convenience. For notational simplicity, let us also assume, without loss of generality, that $U = I$, the identity. Then, it also makes sense to consider the case where $z_i^* = 0,\ i > m$. Suppose further that the data $\mathbf{b}$ is contaminated by Gaussian noise $\boldsymbol{\epsilon}$ with mean 0 and covariance $\rho^2 I$. We have

$$\mathbf{b} = \Sigma\mathbf{z}^* + \boldsymbol{\epsilon}.$$

Thus, for $i \in \mathcal{T}$, $z_i^* = (b_i - \epsilon_i)/\sigma_i = 1$.

Turning to approximate solutions and setting $\mathbf{z} = V^T\mathbf{u}$, recall first the truncated SVD method, even though it has nothing to do with L1 methods. Thus, we set $\beta = 0$ in (5.9), obtaining the least squares problem

$$\min_{\mathbf{z}} \frac{1}{2}\|\Sigma\mathbf{z} - \mathbf{b}\|_2^2, \tag{5.11}$$

and then, since the noise ϵ_i is obviously magnified by σ_i^{-1}, we set

$$z_i = \begin{cases} b_i/\sigma_i & i \leq r \\ 0 & i > r \end{cases}, \tag{5.12}$$

where the effective rank r, $r \leq m$, is such that the error term depending on σ_r^{-1} has tolerable size. Using this regularization method, it is obvious that a necessary and sufficient condition for obtaining the same sparsity for $\mathbf{z}$ and $\mathbf{z}^*$ is that $\mathcal{T} = \{1, 2, \ldots, r\}$. Indeed, no (very) small singular value index can be tolerated in the set $\mathcal{T}$ of the given true model. In particular, we cannot stably obtain the sparse approximate solution for just any true model. This requirement becomes rather restrictive in the highly ill-conditioned case, where $r \ll m$.

Of course, the truncated SVD method not only does not have L1 magic, it also requires carrying out the SVD, something we wish to avoid for the large problems considered in this section. Let us now return to the Tikhonov-type method (5.9) with $\beta > 0$, and consider the special case of the L1 approach with $W = V^\top$. This special case is in a sense the most favorable for the sparsity-inducing algorithm to work well. This is so because the subspace defined by $\Sigma \mathbf{z} = \mathbf{b}$ has the best possible orientation, with respect to the faces of the polyhedron $\|\mathbf{z}\|_1 =$ constant, to cause intersection at a face corresponding to the correct sparsity. See, for example, Figure 1 in [10] for a sparsity spoiling orientation that cannot occur in our case. So, if we encounter difficulties caused by ill-conditioning in this special case, then they will persist upon using a more general W.

Thus, we are considering the problem

$$\min_{\mathbf{z}} \frac{1}{2} \|\Sigma \mathbf{z} - \mathbf{b}\|_2^2 + \beta \|\mathbf{z}\|_1 . \tag{5.13}$$

Because (5.13) is just a sum of decoupled terms, we can solve it explicitly for each component of $\mathbf{z}$. The solution has $z_i = 0$ where the gradient of the data fitting term is bounded by the gradient of the regularization term, which gives

$$\beta \geq \left| \sigma_i(\sigma_i z_i^* + \epsilon_i) \right| .$$

Otherwise,

$$z_i = ((\sigma_i z_i^* + \epsilon_i) \pm \beta/\sigma_i)/\sigma_i ,$$

where the sign in front of β is not needed for our purposes.

In order for $\mathbf{z}$ to have the same sparsity as $\mathbf{z}^*$, we therefore must have

$$\begin{aligned} &\beta \leq |\sigma_i(\sigma_i + \epsilon_i)| \quad &&\text{for } i \in \mathcal{T} , \\ &\beta \geq |\sigma_i \epsilon_i| &&\text{for } i \notin \mathcal{T} . \end{aligned}$$

Squaring these inequalities and replacing ϵ_i^2 by its expected value ρ^2 gives the condition

$$\max_{i \notin \mathcal{T}} \rho^2 \sigma_i^2 \leq \beta^2 \leq \min_{i \in \mathcal{T}} \sigma_i^2(\sigma_i^2 + \rho^2) .$$

Thus, the regularization parameter β must satisfy

$$\rho \sigma_+ \leq \beta \leq \sigma_- \sqrt{\sigma_-^2 + \rho^2} , \tag{5.14a}$$

with

$$\sigma_+ = \max_{i \notin \mathcal{T}} \sigma_i , \quad \sigma_- = \min_{i \in \mathcal{T}} \sigma_i . \tag{5.14b}$$

From (5.14), it follows that the correct sparsity pattern can be comfortably recovered if $\sigma_+ \leq \sigma_-$, i.e. if all small singular values are not in $\mathcal{T}$ and all others are in $\mathcal{T}$, just as for the truncated SVD method.

The case where L1 may offer potential advantage over truncated SVD is when $\sigma_+ > \sigma_-$. In this case, (5.14a) yields the requirement

$$\rho \leq \frac{\sigma_-^2}{\sqrt{\sigma_+^2 - \sigma_-^2}} . \tag{5.15}$$

We summarize this as follows:

Theorem 5.4. *Consider the* $L1$ *regularization problem* (5.9). *For the specific case defined above using* (5.10), (5.13) *and* (5.14b), *the true and reconstructed models,* $\mathbf{z}^*$ *and* $\mathbf{z}$, *are expected to have the same zero structure only if either* $\sigma_+ \le \sigma_-$ *or* (5.15) *holds.*

Unfortunately, if $\sigma_- \ll 1$ and/or $\sigma_+ \gg \sigma_-$, then the condition (5.15) may be too restrictive in practice, possibly holding only for an unrealistically small noise level.

Further difficulties arise upon considering the usual practical process of selecting the regularization parameter β by the discrepancy principle (see, e.g. [61]), i.e. such that the total misfit μ satisfies

$$\mu^2 = \frac{1}{m}\sum_i \left(\sigma_i \left(z_i - z_i^*\right) - \epsilon_i\right)^2 \approx \rho^2 .$$

Let us next compute the misfit for β satisfying (5.14a), assuming ρ is such that this is possible, i.e. one of the conditions of Theorem 5.4 holds, and show that the misfit can easily be much too large in the ill-conditioned case. Conversely, this means that if β was selected by the discrepancy principle, condition (5.14a) would be violated.

Let us choose $\beta = \rho\sigma_+$, i.e. the smallest possible β satisfying (5.14a). Replacing ϵ_i^2 by its expected value, the expected misfit squared becomes

$$\mu^2 = \frac{1}{m}\left(\sum_{i\notin\mathcal{T}, i\le m} \rho^2 + \sum_{i\in\mathcal{T}} \rho^2\sigma_+^2/\sigma_i^2\right) .$$

The discrepancy principle requirement $\mu \approx \rho$ can now be written as

$$\frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}} \sigma_+^2/\sigma_i^2 \approx 1 .$$

However if J is ill-conditioned, the mean value of σ_+^2/σ_i^2 over the set $\mathcal{T}$ could be very large, implying that β (chosen to recover the correct sparsity) is too large to satisfy the discrepancy principle. Conversely, the value of β selected by the discrepancy principle will be too small to recover the correct sparsity of $\mathbf{z}^*$.

It is important to emphasize that we do not claim that L1 variants *cannot* work for highly ill-conditioned problems. Rather, they *may not necessarily* work. It all depends on how the sparsity of the true solution $\mathcal{T}$ and the singular values of J relate. Moreover, we do not know of a method that does better than L1 in the present sense. However then, our expectations regarding sparsity are lower for most other methods in the first place.

4.3 Nonlinear, highly ill-posed examples

In this subsection, we study the DC resistivity problem on the unit square. The forward problem for v, given by

$$-\nabla \cdot (\sigma(u)(\mathbf{x})\nabla v) = q(\mathbf{x}), \quad \mathbf{x} \in \Omega, \tag{5.16}$$

subject to Neumann boundary conditions, predicts the potential v for given external source q and conductivity σ (parameterized in terms of u). The inverse problem is to recover the conductivity $\sigma(u)$ from partial measurements of the potential v^i, when different current patterns q^i, $i = 1, \ldots, s$, are injected into the region.

For experiment i, q^i consists of a positive point source on the left boundary and an opposite point source on the right boundary, and thus

$$q^i(\mathbf{x}) = \delta_{\mathbf{x},\mathbf{p}_\mathrm{L}^{\mathrm{i_L}}} - \delta_{\mathbf{x},\mathbf{p}_\mathrm{R}^{\mathrm{i_R}}},$$

where $\mathbf{p}_\mathrm{L}^{\mathrm{i_L}}$ and $\mathbf{p}_\mathrm{R}^{\mathrm{i_R}}$ are located on the left and right boundaries. Different data sets are obtained by varying the positions $\mathbf{p}_\mathrm{L}^{\mathrm{i_R}}$ and $\mathbf{p}_\mathrm{R}^{\mathrm{i_R}}$ of the two opposing point sources. We place each at $\sqrt{s}$ equidistant points including the corners, in all possible combinations, which gives a total of s data sets for a perfect square. Voltage is measured on the boundary, so the number of point in each data set is the number of boundary points of the discretization mesh. See [17, 52] and references therein for details of the problem setup such as the discretization of (5.16) and the solution of the resulting optimization problem.

For this nonlinear inverse problem, it is well-known that, unlike for the inverse potential problem, increasing the number of data sets s allows a more accurate recovery of the resistivity $1/\sigma$. There is no reason to apply L1 here, and the purpose of the following experiments is to determine, for a piecewise continuous surface recovery, roughly at what point of such computational refinement the L1G regularization becomes worthwhile.

Example 5.5 (EIT and DC-resistivity). We have chosen to recover a grid approximation $\mathbf{u}$ of

$$u(\mathbf{x}) = P^{-1}(\sigma(\mathbf{x})), \tag{5.17a}$$

where the transfer function

$$P(t) = \frac{1}{2}(\sigma_{max} - \sigma_{min})\tanh(t) + \frac{1}{2}(\sigma_{max} + \sigma_{min}) \tag{5.17b}$$

enforces *a priori* known upper and lower bounds on the possible conductivity.

A synthetic conductivity model is used to compute the data $\mathbf{b}$, which is calculated on a grid that is twice as fine as the grid used for the reconstruction, and either 3 % or 1 % Gaussian noise is added to it.

The ground truth model used to synthesize data consists of an object with conductivity $\sigma = 1$ (black) placed in a background of conductivity $\sigma = 10$ (white);

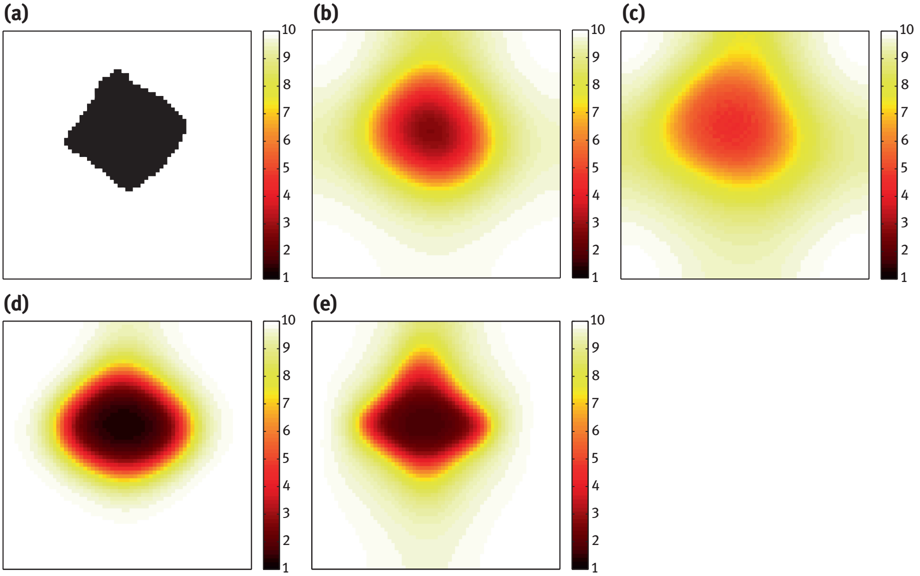


Figure 5.9: Conductivity reconstructions for different numbers s of data sets with noise level $3\,\%$. (a) True model, (b) s = 4, L2G, (c) s = 4, L1G,(d) s = 64, L2G, (e) s = 64, L1G.

see Figure 5.9 (a). In (5.17b), we set $\sigma_{min} = 1$ and $\sigma_{max} = 10$. The inverse problem involves minimizing expressions of the form (5.1), (5.3). We compare $p = 1$ (total variation, or L1G) with $p = 2$ (L2G). A 128^2 uniform grid is used in these calculations.

Figure 5.9 shows the obtained reconstructions using $s = 4$ and $s = 64$ current configurations at a noise level of $3\,\%$. The regularization parameter β was tuned to result in a misfit of $3 \pm 0.1\,\%$. Observe that in the case of rare data $s = 4$, there appears to be no advantage to using the L1G regularization, whereas with 64 data sets the L1G reconstruction is only marginally better than L2G.

Next, we use $s = 1024$ data sets at a noise level of $1\,\%$, with β correspondingly tuned. In order to accommodate so many right-hand sides, we employ the stochastic adaptive algorithm described in [17]. The results are depicted in Figure 5.10. At this increased model accuracy and resolution, the result obtained using L1G is clearly better than that obtained using L2G.

The situation described in Example 5.5 is not uncommon in practice. Often in geophysical experiments, results of the sort depicted in Figure 5.9 (d,e) are of sufficient quality and the lower noise level and larger number of experiments s required for obtaining the result in Figure 5.10 (b) is a sort of luxury that is not always attained. Moreover, the forward problem considered in this section is often indicative of what is observed numerically, also for more complex problems such as low frequency electromagnetic and seismic data inversions. Finally, weighted L2G variants that are rou-

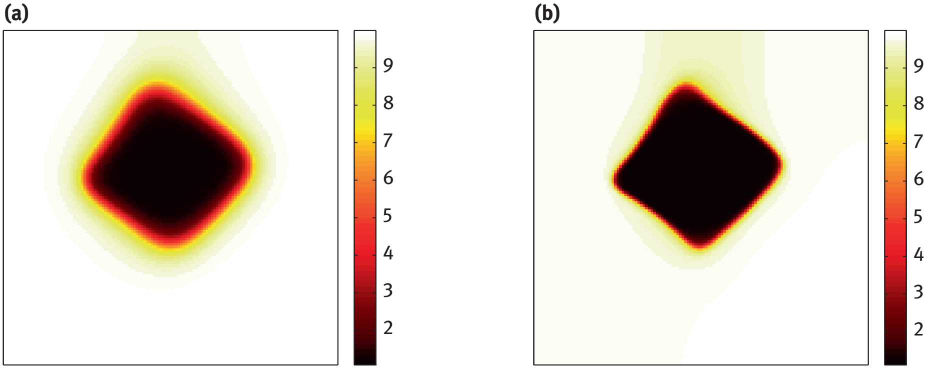


Figure 5.10: Reconstructions for a larger number of data sets $s = 1024$ and with the noise level at only $1\,\%$. Here, L1G clearly outshines L2G. (a) s = 1024, L2G, (b) s = 1024, L1G.

tinely used in geophysical applications may further improve reconstructions without resorting to ℓ_1-based regularization. In view of the occasionally significantly higher cost of computing with L1G, it cannot be automatically concluded that the latter is worthwhile for this application, although it is a viable option that we always entertain in the course of our research.

5 Summary

In this paper, we have investigated the relative performance of ℓ_1-based regularization techniques on several examples and case studies. We have shown cases where such methods are worse than ℓ_2-based ones in the sense of costing more without delivering more (Examples 5.1 and 5.5), and other cases where such methods produce better results (see Figures 5.4b and 5.10). Further, we have shown cases where the ℓ_1-based results appear to be more misleading than corresponding ℓ_2-based results (Example 5.3).

In Section 4.2, we have analyzed the effect of ill-conditioning on the ability of an L1 method to correctly recover solution sparsity. Theorem 5.4 and the arguments following it suggest severe limitations in case of extreme ill-conditioning that perhaps arises in certain inverse problems.

The results in Section 4.3 demonstrate how and when L1G becomes favored as the quality of the data improves. This in itself is intuitively expected, but less clear is where the crossover point occurs in realistic situations. Unfortunately, we had to tweak the problem beyond what may be expected in many geophysical situations in order to observe the L1G takeover.

Let us again stress our overall conviction that the swing of the pendulum in recent years towards ℓ_1-based techniques is rather important and not merely refreshing. Our purpose here, far from opposing this trend, is to simply suggest that this virtual pendulum should not swing too far and away, to realms beyond reason. To this end, we note the following.

- In many situations, ℓ_1-based regularization is well-worth using. Such techniques can provide exciting advances (e.g. in model reduction, computer graphics, image processing and reconstruction of surfaces with discontinuities).
- However, such techniques are not good for all problems, and it is dangerous (and may consume many student-years) to apply them blindly.
- In practice, we recommend to always consider first using ℓ_2-based regularization techniques because they are simpler, more easy to compute with, and do not introduce nonlinearities or lower smoothness. Only upon deciding that these are not sufficiently good for the given application, it is highly advisable to proceed to examine ℓ_1-based alternatives (when this makes sense).
- Last but not least, the possibility of combining ℓ_1- and ℓ_2-based techniques suggests itself. We have already commented on using the Huber switching function as well as IRLS techniques [1, 30, 33, 56, 61] for this purpose in the L1G–L2G context, but these ideas are also very popular in the image processing and computer vision literature in mixing the L1 and L2 approaches [34]. Another popular approach is to employ an empirical Bayesian framework in order to learn an appropriate mix [37, 57].

References

[1] U. Ascher, E. Haber, and H. Huang, On effective methods for implicit piecewise smooth surface recovery, *SIAM J. Sci. Comput.* 28 (2006), 339–358.

[2] C. Avron, A. Sharf, C. Greif, and D. Cohen-Or, ℓ_1-sparse reconstruction of sharp point set surfaces, *ACM trans. on Graphics* 29(5) (2010), 135:1–12.

[3] S. Becker, J. Bobin, and E. J. Candes, NESTA: a fast and accurate first-order method for sparse recovery, *SIAM J. on Imaging Sciences* 4 (2010), 1–39.

[4] E. van den Berg and M. Friedlander, Sparse optimization with least-squares constraints, *SIAM J. Optimization* 21 (2011), 1201–1229.

[5] D. Bertsekas, *Convex Optimization Theory*, Athena Scientific, 2009.

[6] K. Bredies, K. Kunisch, and T. Pock, Total generalized variation, *SIAM J. Imaging Sciences* 3 (2010), 492–526.

[7] J.-F. Cai, S. Osher, and Z. Shen, Split Bregman methods and frame based image restoration, *SIAM J. multiscale modeling and simulation* 8(2) (2009), 337–369.

[8] D. Calvetti and E. Somersalo, *Introduction to Bayesian Scientific Computing*, Springer, 2007.

[9] E. Candes, J. Romberg, and T. Tao, Stable signal recovery from incomplete and inaccurate measurements, *Pure Appl. Math.* 59 (2006), 1207–1223.

[10] E. J. Candes, M. B. Wakin, and S. Boyd, Enhancing Sparsity by Reweighted l1 Minimization, *Journal of Fourier Analysis and Applications* 14 (2008), 877–1905.

[11] T. Chan and J. Shen, *Image Processing and Analysis: Variational, PDE, Wavelet and Stochastic Methods*, SIAM, 2005.

[12] R. Chartrand, Fast algorithms for nonconvex compressive sensing: MRI reconstruction from very few data, *IEEE International Symposium on Biomedical Imaging (ISBI)* (2009).

[13] S. Chen, D. Donoho, and M. Saunders, Atomic Decomposition by Basis Pursuit, *SIAM Review* 43 (2001), 129–159.

[14] J. Claerbout and F. Muir, Robust modeling with erratic data, *Geophysics* 38 (1973), 826–844.

[15] A. Cohen, W. Dahmen, and R. DeVore, Compressed sensing and best k-term approximation, *Journal of the American Mathematical Society* 22 (2008), 211–231.

[16] K. van den Doel and U. Ascher, Dynamic level set regularization for large distributed parameter estimation problems, *Inverse Problems* 23 (2007), 1271–1288.

[17] K. van den Doel and U. Ascher, Adaptive and stochastic algorithms for EIT and DC resistivity problems with piecewise constant solutions and many measurements, *SIAM J. Scient. Comput.* (2012), DOI: 10.1137/110826692.

[18] K. van den Doel, U. Ascher, and D. Pai, Computed myography: three dimensional reconstruction of motor functions from surface EMG data, *Inverse Problems* 24 (2008), 065010.

[19] K. van den Doel, U. Ascher, and D. Pai, Source localization in electromyography using the inverse potential problem, *Inverse Problems* 27 (2011), 025008.

[20] D. Donoho, For most large underdetermined systems of linear equations, the minimal l1 solution is also the sparsest solution, *Comm. Pure Applied Math.* 7 (2006), 907–934.

[21] D. Donoho and X. Huo, Uncertainty principles and ideal atomic decomposition, *IEEE Transactions on Information Theory* 47 (2001), 2845–2862.

[22] D. Donoho and J. Tanner, Sparse Nonnegative Solutions of Underdetermined Linear Equations by Linear Programming, *Proc. Nat. Acad. Sciences* 102 (2005), 9446–9451.

[23] A. Ebtehaj, E. Foufoula-Georgiou, and G. Lerman, Sparse regularization for precipitation downscaling, *J. Geophys. Res.* 117 (2012), D08107 doi:10.1029/2011JD017057.

[24] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, Springer, 2010.

[25] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, 1996.

[26] M. Figueiredo, R. Nowak, and S. Wright, Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems, *IEEE J. Special Topics on Signal Processing* 1 (2007), 586–598.

[27] M. A. Freitag, N. K. Nichols, and C. J. Budd, Resolution of sharp fronts in the presence of model error in variational data assimilation, *Q.J.R. Meteorol. Soc.* (2012), doi: 10.1002/qj.2002.

[28] R. E. Greenblatt, Probabilistic reconstruction of multiple sources in the neuroelectromagnetic inverse problem, *Inverse problems* 9 (1993), 271–284.

[29] E. Haber, U. Ascher, and D. Oldenburg, Inversion of 3D Electromagnetic Data in frequency and time domain using an inexact all-at-once approach, *Geophysics* 69 (2004), 1216–1228.

[30] E. Haber, S. Heldmann, and U. Ascher, Adaptive finite volume method for distributed nonsmooth parameter identification, *Inverse Problems* 23 (2007), 1659–1676.

[31] M. S. Hämäläinen and R. J. Ilmoniemi, Interpreting magnetic fields of the brain-minimum norm estimates, *Med. Biol. Eng. Comput.* 32 (1994), 35–42.

[32] P. C. Hansen, *Rank Deficient and Ill-Posed Problems*, SIAM, Philadelphia, 1998.

[33] P.-C. Hansen, J. Nagy, and D. O'Leary, *Deblurring Images: Matrices, Spectra and Filtering*, SIAM, 2006.

[34] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, second ed, Cambridge University Press, ISBN: 0521540518, 2004.

[35] H. L. F. Helmholtz, Ueber einige Gesetze der Verheilung elektrischer Ströme in körperlicher Leitern mit Anwendung auf die thierisch-elektrischen Versuche, *Ann. Physik und Chemie* 9 (1853), 211–233.

[36] H. Huang, *Efficient Reconstruction of 2D Images and 3D Surfaces*, Ph.D. thesis, University of BC, Vancouver, 2008.

[37] H. Huang, E. Haber, and L. Horesh, Optimal estimation of l1 regularization prior from a regularized empirical Bayesian risk standpoint, *Inverse Problems and Imaging* (2013).

[38] H. Huang, D. Li, R. Zhang, U. Ascher, and D. Cohen-Or, Consolidation of unorganized point clouds for surface reconstruction, *ACM Trans. Graphics (SIGGRAPH Asia)* 29(5) (2009).

[39] H. Huang, S. Wu, M. Gong, D. Cohen-Or, U. Ascher, and H. Zhang, Edge-aware point set resampling, *ACM trans. on Graphics* (2013).

[40] A. Juditsky and A. Nemirovski, On verifiable sufficient conditions for sparse signal recovery via l1-minimization, *Mathematical Programming Ser. B* 127 (2008), 57–88.

[41] J. Kaipo and E. Somersalo, *Statistical and Computational Inverse Problems*, Springer, 2005.

[42] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinvesky, A method for large-scale l1-regularized least squares problems with applications in signal processing and statistics, *IEEE J. Select. Topics Signal Process* (2007).

[43] T. Köhler, M. Wagner, M. Fuchs, H. A. Wischmann, R. Denckkhahn, and A. Theissen, Depth Normalization in MEG/EEG Current Density Imaging, in: *18th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Amsterdam*, pp. 812–813, 2006.

[44] A. Levin, R. Fergus, F. Durand, and W. Freeman, Image and depth from a conventional camera with a coded aperture, *ACM trans. on Graphics (SIGGRAPH)* 26(3) (2007), 70.

[45] Y. Lipman, D. Cohen-Or, D. Levin, and H. Tal-Ezer, Parameterization-free projection for geometry reconstruction, *ACM trans. on Graphics (SIGGRAPH)* 26(3) (2007), 22.

[46] S. Mallat, *A Wavelet Tour of Signal Processing: the Sparse Way*, Academic Press, 2009, 3rd Ed.

[47] N. Meinshausen and P. Buehlmann, Stability selection, *J. Royal Stat. Soc.* B72 (2010), 417–473.

[48] C. M. Michel, M. M. Murray, G. Lantz, S Gonzalez, L. Spinelli, and R. G. de Peralta, EEG source imaging, *Clinical neurophysiology* 115 (2004), 2195–2222.

[49] K. Murphy, *Machine Learning: a Probabilistic Perspective*, MIT Press, 2012.

[50] M. Osborne, *Finite Algorithms in Optimization and Data Analysis*, Wiley, 1985.

[51] S. Osher and R. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces*, Springer, 2003.

[52] F. Roosta-Khorasani, K. van den Doel, and U. Ascher, Stochastic algorithms for inverse problems involving PDEs and many measurements, *Submitted* (2012).

[53] L. Rudin, S. Osher, and E. Fatemi, Nonlinear total variation based noise removal algorithms, *Physica D* 60 (1992), 259–268.

[54] R. Saab, R. Chartrand, and O. Yilmaz, Stable sparse approximations via nonconvex optimization, *33rd IEEE International Conference Acoustics, Speech and Signal Proc. (ICASSP)* (2008).

[55] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing Company, 1996.

[56] G. Sapiro, *Geometric Partial Differential Equations and Image Analysis*, Cambridge, 2001.

[57] K. Swersky, M. Ranzato, D. Buchman, B. Marlin, and N. de Freitas, On autoencoders and score matching for energy based models, in: *Proc. 28th Intl. Conf. Machine Learning, Bellevue, WA, USA*, 2011.

[58] A. Tarantola, *Inverse problem theory*, Elsevier, Amsterdam, 1987.

[59] H. Taylor, S. Banks, and J. McCoy, Deconvolution with the l1 norm, *Geophysics* 44 (1979), 39–52.

[60] A. N. Tikhonov and V. Ya. Arsenin, *Methods for Solving Ill-posed Problems*, John Wiley and Sons, Inc., 1977.

[61] C. Vogel, *Computational methods for inverse problem*, SIAM, Philadelphia, 2002.

[62] H. Wagner, Linear programming techniques for regression analysis, *Proc.* 54 (1959), 206–212.

List of contributors

Radon Series on Computational and Applied Mathematics

Volume 12
Peter Bastian, Johannes Kraus, Robert Scheichl, Mary Wheeler (Eds.), 2013
Simulation of Flow in Porous Media. Applications in Energy and Environment
ISBN 978-3-11-028221-4, e-ISBN 978-3-11-028224-5, Set-ISBN 978-3-11-028225-2

Volume 11
Pascale Charpin, Alexander Pott, Arne Winterhof (Eds.), 2013
Finite Fields and Their Applications
ISBN 978-3-11-028240-5, e-ISBN 978-3-11-028360-0, Set-ISBN 978-3-11-028361-7

Volume 10
Thomas Schuster, Barbara Kaltenbacher, Bernd Hofmann, Kamil S. Kazimierski, 2012
Regularization Methods in Banach Spaces
ISBN 978-3-11-025524-9, e-ISBN 978-3-11-025572-0, Set-ISBN 978-3-11-220450-4

Volume 9
Massimo Fornasier (Ed.), 2010
Theoretical Foundations and Numerical Methods for Sparse Recovery
ISBN 978-3-11-022614-0, e-ISBN 978-3-11-022615-7, Set-ISBN 978-3-11-174177-2

Volume 8
Hansjörg Albrecher, Wolfgang J. Runggaldier, Walter Schachermayer (Eds.), 2009
Advanced Financial Modelling
ISBN 978-3-11-021313-3, e-ISBN 978-3-11-021314-0, Set-ISBN 978-3-11-173185-8

Volume 7
Jan H. Maruhn, 2009
Robust Static Super-Replication of Barrier Options
ISBN 978-3-11-020468-1, e-ISBN 978-3-11-020851-1, Set-ISBN 978-3-11-916585-3

Volume 6
Barbara Kaltenbacher, Andreas Neubauer, Otmar Scherzer, 2008
Iterative Regularization Methods for Nonlinear Ill-Posed Problems
ISBN 978-3-11-020420-9, e-ISBN 978-3-11-020827-6, Set-ISBN 978-3-11-916135-0

www.degruyter.com