The IMA Volumes in Mathematics and its Applications

Harbir Antil · Drew P. Kouri
Martin-D. Lacasse · Denis Ridzal *Editors*

Frontiers in PDE-Constrained Optimization

Springer

The IMA Volumes in Mathematics and its Applications

Volume 163

Series editor
Daniel Spirn, *University of Minnesota, MN, USA*

Institute for Mathematics and its Applications (IMA)

The Institute for Mathematics and its Applications (IMA) was established in 1982 as a result of a National Science Foundation competition. The mission of the IMA is to connect scientists, engineers, and mathematicians in order to address scientific and technological challenges in a collaborative, engaging environment, developing transformative, new mathematics and exploring its applications, while training the next generation of researchers and educators. To this end the IMA organizes a wide variety of programs, ranging from short intense workshops in areas of exceptional interest and opportunity to extensive thematic programs lasting nine months. The IMA Volumes are used to disseminate results of these programs to the broader scientific community.

The full list of IMA books can be found at the Web site of the Institute for Mathematics and its Applications:

http://www.ima.umn.edu/springer/volumes.html.

Presentation materials from the IMA talks are available at

http://www.ima.umn.edu/talks/.

Video library is at

http://www.ima.umn.edu/videos/.

Daniel Spirn, Director of the IMA

More information about this series at http://www.springer.com/series/811

Harbir Antil • Drew P. Kouri • Martin-D. Lacasse
Denis Ridzal
Editors

Frontiers in PDE-Constrained Optimization

Springer

Editors
Harbir Antil
Department of Mathematical Sciences
George Mason University
Fairfax, VA, USA

Drew P. Kouri
Center for Computing Research
Sandia National Laboratories
Albuquerque, NM, USA

Martin-D. Lacasse
Corporate Strategic Research
ExxonMobil Research and
Engineering Company
Annandale, NJ, USA

Denis Ridzal
Center for Computing Research
Sandia National Laboratories
Albuquerque, NM, USA

ISSN 0940-6573 ISSN 2198-3224 (electronic)
The IMA Volumes in Mathematics and its Applications
ISBN 978-1-4939-8635-4 ISBN 978-1-4939-8636-1 (eBook)
https://doi.org/10.1007/978-1-4939-8636-1

Library of Congress Control Number: 2018949385

Mathematics Subject Classification: 49J20, 49N05, 93E20, 80M50, 35Q93, 46N10, 65K10, 49Q10, 34A55

This Springer imprint is published by the registered company Springer Science+Business Media, LLC part of Springer Nature.
The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

Foreword

This volume contains a series of papers based on a workshop "Frontiers in PDE-Constrained Optimization" held at the Institute for Mathematics and its Applications from June 6 to 10, 2016, and organized by Harbir Antil, Drew Kouri, Martin Lacasse, and Denis Ridzal. This workshop drew together a cohort of scientists working in PDE-constrained optimization in a variety of disciplines, ranging from medical imaging to geosciences. The collection of works in this volume reflects this diversity of application and documents the recent mathematical and computational advances in the field. We would like to especially thank the workshop organizers, who have served as the editors of this volume. Finally, we acknowledge ExxonMobil, which provided funding for the workshop, and the National Science Foundation for its support of the IMA.

Minneapolis, MN, USA Daniel Spirn

Preface

Many science and engineering applications necessitate the solution of optimization problems constrained by physical laws that are described by systems of partial differential equations (PDEs). As a result, PDE-constrained optimization problems arise in a variety of disciplines including geo-physics, earth and climate science, material science, chemical and mechanical engineering, medical imaging, and physics. The goal of this volume is to provide a broad and uniform introduction of PDE-constrained optimization as well as to document a number of interesting and challenging applications.

This volume contains the proceedings of the workshop "Frontiers in PDE-Constrained Optimization" held at the Institute for Mathematics and its Applications from June 6 to 10, 2016. The workshop successfully provided a common forum for networking between leaders in PDE-constrained optimization within academia, industry, and the US national labs. The five-day workshop included two days of tutorials and three days of invited talks. The tutorials were targeted toward students and researchers interested in entering the field of PDE-constrained optimization and provided an overview of the field with special emphasis on uncertainty, variational inequalities, shape optimization, inverse problems, algorithmic development, and software implementation. The invited presentations disseminated cutting-edge developments in theory, numerics, and applications.

This volume is divided into two parts. The first part provides a comprehensive review of modern topics in PDE-constrained optimization. Chapter "A Brief Introduction to PDE Constrained Optimization" provides a basic introduction to PDE-constrained optimization. Chapter "Optimization of PDEs with Uncertain Inputs" discusses optimization problems constrained by PDEs with uncertain or random inputs. Chapter "Inexact Trust-Region Methods for PDE-Constrained Optimization" focuses on the efficient numerical solution of PDE-constrained optimization problems using inexact trust-region methods. Chapter "Numerical Optimization Methods for the Optimal Control of Elliptic Variational Inequalities" provides a theoretical and numerical overview of optimization problems constrained by elliptic variational inequalities. Chapters "Introduction to PDE-Constrained Optimization

in the Oil and Gas Industry" and "Full-Wavefield Inversion: An Extreme-Scale PDE-Constrained Optimization Problem" describe a variety of theoretically and computationally challenging inverse problems arising in the oil and gas industry. Chapters 1–6 are organized in such a way that they can be used as a reference to augment a graduate course in PDE-constrained optimization.

The second part of this volume focuses on applications of PDE-constrained optimization. Chapters "Energetically Optimal Flapping Wing Motions via Adjoint-Based Optimization and High-Order Discretizations" and "Optimization of a Fractional Differential Equation" consider PDE-constrained optimal control with applications to flapping wing machines and anomalous diffusion. Chapter "Sensitivity-Based Topology and Shape Optimization with Application to Electric Motors" discusses a sensitivity-based approach for optimal design via topology and shape optimization. Chapter "Distributed Parameter Estimation for the Time-Dependent Radiative Transfer Equation" discusses the parameter estimation in time-dependent radiative transfer equations. Following this, Chapter "On the Use of Optimal Transport Distances for a PDE-Constrained Optimization Problem in Seismic Imaging" discusses the use of optimal transport distances in seismic imaging. To conclude the volume, Chapter "Exploiting Sparsity in Solving PDE-Constrained Inverse Problems: Application to Subsurface Flow Model Calibration" describes the role of sparsity in inverse problems with applications to subsurface flow model calibration.

Acknowledgement As organizers and editors, we would like to acknowledge ExxonMobil, which provided funding for the workshop, and National Science Foundation for its support of the IMA. We are further indebted to the former and current IMA directors Fadil Santosa and Daniel Spirn for encouraging this initiative, as well as to Danielle Walker (Springer) for her help in putting together this volume.

Fairfax, VA, USA — Harbir Antil
Albuquerque, NM, USA — Drew P. Kouri
Annandale, NJ, USA — Martin-D. Lacasse
Albuquerque, NM, USA — Denis Ridzal

Contents

Part I
PDE-Constrained Optimization: Tutorials

A Brief Introduction to PDE-Constrained Optimization

Harbir Antil and Dmitriy Leykekhman

Abstract In this chapter we give a brief overview of optimization problems with partial differential equation (PDE) constraints, i.e., PDE-constrained optimization (PDECO). We start with three potentially different formulations of a general PDECO problem and focus on the so-called reduced form. We present a derivation of the optimality conditions. Later we discuss the linear and the semilinear quadratic PDECO problems. We conclude with the discretization and the convergence rates for these problems. For illustration, we make a MATLAB code available at

https://bitbucket.org/harbirantil/pde_constrained_opt

that solves the semilinear PDECO problem with control constraints.

1 Introduction

This volume is aimed at early-career graduate students or whoever is interested in entering into the field of PDE-constrained optimization (PDECO), i.e., optimization problems with PDE constraints. The purpose of this particular chapter is to provide a brief mathematical introduction to this topic with an emphasis on optimality conditions and finite-element discretization. Most of the material presented here is well-known and can be found in several textbooks. In particular, we mention [14, 53, 56, 62, 81].

PDE-constrained optimization problems arise in many applications, for instance in flow control [42] and shape optimization [43, 80]. To name a few: controlling pollutants in a river [5]; drug delivery using magnetic fields [8, 9]; optimal

H. Antil (✉)
Department of Mathematical Sciences, George Mason University, Fairfax, VA 22030, USA
e-mail: hantil@gmu.edu

D. Leykekhman
Department of Mathematics, University of Connecticut, Storrs, CT 06269, USA
e-mail: dmitriy.leykekhman@uconn.edu

H. Antil et al. (eds.), *Frontiers in PDE-Constrained Optimization*, The IMA Volumes in Mathematics and its Applications 163,
https://doi.org/10.1007/978-1-4939-8636-1_1

placement of droplets on electrowetting on dielectric (EWOD) devices [7]; shape optimization of microfluidic biochips [3, 4]. In the first two examples, the control is distributed (confined to either the entire domain or a subdomain). In the third example, the control acts on the boundary, and in the final example, the control is the shape of the domain. These problems range from macroscale to microscale and contain varying levels of difficulties due to the underlying physics.

A PDE-constrained optimization problem has four or five major components: i) a control variable z; ii) state variable u; iii) a state equation (i.e., PDE or system of PDEs) which associates a state u with every control z; iv) a cost functional J to be minimized which depends on z and u; and, possibly, v) constraints imposed on z (control constraints) or/and on u (state constraints). In the most abstract form, this amounts to

$$\text{minimize} \quad J(u, z) \tag{1}$$

subject to $u \in U_{ad}$ and $z \in Z_{ad}$ satisfying

$$e(u, z) = 0. \tag{2}$$

Thus we want to minimize the cost J which is a function of the state u and the control z. The state equation is given by (2) that represents a PDE or system of PDEs (usually in the weak form) with u in an admissible set U_{ad} and the control z in an admissible set Z_{ad}.

To fully understand PDECO problems, the following topics must be considered: (i) functional analysis of the problem; (ii) solver development; (iii) discretization and software development. We shall briefly elaborate on each next, a more detailed discussion will be given in the subsequent chapters. By *analysis*, we mean

- Existence, uniqueness, and regularity of a solution to (2).
- Existence and uniqueness of a solution to the PDECO problem (1)–(2).
- First-order necessary optimality conditions using adjoint equations.
- If possible the second-order necessary/sufficient conditions.

We can carry out the *solver development*, typically gradient based, either at the continuous (infinite dimensional) level [49, 50, 82] or at the discrete (finite dimensional) level [57, 66]. The choice depends on the two options (a) "First discretize then optimize approach" or (b) "First optimize then discretize approach." As the names suggest in case (a) one first discretizes the optimization problem and then writes the optimality conditions at the finite dimensional level. On the other hand, (b) requires first writing the optimality conditions at the continuous level and then discretizing them. These two approaches in general are different. For instance the discretization of the adjoint variables is tied to the state variables if one proceeds with (a). On the other hand, (b) provides more flexibility with respect to this particular aspect. There is no universal approach to select either of the approaches [53, Section 3.2.4]. However, a choice should be made so that the discrete system preserves the structure of the continuous problem. It is worth

noting that the optimization methods can be derived directly at the continuous level. In any event, one should avoid using off-the-shelf finite-dimensional algorithms to solve PDECO problems. This may lead to *mesh dependence,*, i.e., when the number of optimization iterations drastically increases with mesh refinement. In order to overcome such issues, we must adapt such algorithms (if we are using them) to incorporate functional-analytic concepts, for instance, suitable inner products. Nevertheless, *discretization and software development* amounts to

- Either finite-difference, finite-volume, or finite-element discretizations of the PDE and the control [19, 46, 70]. Other discretizations such as wavelets are possible as well, see [30, 31, 78] and references therein.
- Analysis of discrete PDE and optimization solvers and check for mesh independence [51, 82].
- Make the solver efficient, for instance, by using
  - Adaptive finite-element methods (AFEM) [59, 67].
  - Preconditioning techniques [15, 76, 84], time-parallel approaches [16, 32, 33, 83].
  - Model reduction: Proper orthogonal decomposition [6, 41, 54, 55], reduced-basis method [17, 45, 69].
- Software development, for instance, Rapid Optimization Library (ROL) [60], Dolfin-Adjoint [36].

Our goal for this chapter is to collect the necessary ingredients to understand basic PDECO problems. The remaining chapters of this volume provide comprehensive treatments of variational inequalities, algorithm development, optimization under uncertainty, inverse problems, shape optimization, and several applications. We introduce notation, relevant function spaces, and notions of derivatives. We provide an abstract treatment to solve (1)–(2). Moreover, we discuss the linear quadratic PDECO problem in detail (quadratic cost functional with linear elliptic PDEs and control constraints), provide a flavor of the semilinear quadratic PDECO problem, and discuss the necessary components for the numerical analysis of the linear and semilinear quadratic PDECO problems. We also provide a MATLAB implementation of the semilinear quadratic PDECO problem where the space discretization is carried out using the finite-element method (FEM). We solve the resulting optimization problem using the Newton-CG method, or LBFGS method in the absence of the control constraints ($Z_{ad} = Z$) and the semismooth Newton method in the presence of the control constraints. The code can be downloaded using the link

https://bitbucket.org/harbirantil/pde_constrained_opt

The first part of this chapter is organized as follows: in Section 2, we first consider a general optimization problem. We discuss existence of solutions to this problem in Section 2.1 and provide notions of derivatives in function spaces in Section 2.2. We conclude this section with the first-order necessary optimality conditions (cf. Section 2.3). We apply the approach discussed in Section 2 to the general PDECO

problem (1)–(2) in Section 3 and discuss the full and the reduced forms. Derivation of optimality conditions is performed using both full and reduced forms.

In the second part of this chapter, we focus on the linear and semilinear quadratic PDECO problems. We introduce basic Sobolev spaces in Section 4. In Section 5, we mention the results relating to the well-posedness (existence and uniqueness whenever possible) of linear elliptic PDEs and their regularity. Next in Section 6, we formulate the linear quadratic PDECO problem and study its well-posedness in the reduced form. Section 7 discusses the semilinear quadratic PDECO problem. We conclude by introducing a finite-element discretization for the linear and semilinear quadratic PDECO problems in Section 8.

2 Abstract Optimization Problem

The purpose of this section is to consider an abstract optimization problem and study its well-posedness (cf. Section 2.1). We derive the first-order necessary optimality conditions in Section 2.3. This requires the notions of derivatives in function spaces (cf. Section 2.2). The results presented in this section will be applied to the PDECO problems in the subsequent sections.

Let Z be a real reflexive Banach space and Z_{ad} is a closed convex subset. We consider the minimization problem

$$\min_{z \in Z_{ad}} f(z). \tag{3}$$

2.1 Existence

We first show existence of solution to the minimization problem (3) using the direct method of calculus of variations. This is also known as the Weierstrass theorem.

Theorem 1 *Suppose $f : Z \to \mathbb{R}$ is weakly lower semicontinuous with Z a reflexive Banach space and $Z_{ad} \subset Z$ is closed convex. Let the lower γ-level set $\{z \in Z_{ad} : f(z) \leq \gamma\}$ of f is nonempty and bounded for some $\gamma \in \mathbb{R}$. Then problem (3) has an optimal solution, i.e., there exists $\bar{z} \in Z_{ad}$ such that $f(\bar{z}) \leq f(z)$ for all $z \in Z_{ad}$. If f is strictly convex then the solution is unique.*

Proof The proof is based on the direct method of calculus of variations. Following the proof of [14, Theorem 3.2.1], we can construct a minimizing sequence $\{z_n\}_{n\in\mathbb{N}}$ contained in the lower γ-level set such that $f(z_n) \to \inf f(z)$ as $n \to \infty$. Since the lower γ-level set is convex and closed, therefore it is weakly sequentially closed [81, Theorem 2.11]. In addition, since Z is reflexive and the lower γ-level set is bounded, therefore it is weakly sequentially compact [81, Theorem 2.11]. As a result, there exists a subsequence (not relabeled) such that

$$z_n \rightharpoonup \bar{z} \quad \text{in } Z$$

with $\bar{z}$ in the lower γ-level set. It then remains to show that $\bar{z}$ is the optimal solution. Due to the weak lower semicontinuity of f, we conclude that

$$f(\bar{z}) \leq \liminf f(z_n) = \inf_{z \in Z_{ad}} f(z).$$

Finally, in order to show the uniqueness let us assume that z_1 and z_2 be two optimal solutions. Using the definition of strict convexity, we have

$$f\left(\frac{z_1 + z_2}{2}\right) < \frac{1}{2} f(z_1) + \frac{1}{2} f(z_2) = \inf f(z)$$

which is a contradiction. This completes the proof.

After showing the existence of minimizers, a natural question to ask is what the first-order optimality conditions are. However, we need to understand the notions of derivatives in function spaces before we can proceed further.

2.2 *Differentiation in Banach Spaces*

We introduce the notions of derivatives in function spaces [18, 79]. As an example, we shall apply the ideas to a quadratic cost functional. We will also derive the first-order optimality conditions for the problem (3) based on the derivatives introduced in this section.

Let $\mathscr{L}(A, B)$ denote the space of bounded linear operators from Banach space A to B. Let $(Z, \|\cdot\|_Z)$, $(V, \|\cdot\|_V)$ be real Banach spaces, $\mathscr{Z} \subset Z$, open and $F : \mathscr{Z} \to V$. Moreover, let $z \in \mathscr{Z}$.

Definition 1 (Directional Derivative) F is said to be directionally differentiable at z if the limit $\lim_{t \downarrow 0} \frac{1}{t}(F(z + th) - F(z))$ exists in V for all $h \in Z$. If such limit exists, we denote

$$F'(z, h) := \lim_{t \downarrow 0} \frac{1}{t}(F(z + th) - F(z))$$

and say that $F'(z, h)$ is the directional derivative of F at z in the direction h.

Notice that for a given z, $h \mapsto F'(z, h)$ is not necessarily a linear mapping but it is positive homogeneous, we refer to [79, 81] for examples.

Definition 2 (Gâteaux Derivative) F is said to be Gâteaux differentiable at z if it is directionally differentiable and $F'(z, h) = F'(z)h$ for $F'(z) \in \mathscr{L}(Z, V)$. We refer to $F'(z)$ as the Gâteaux derivative at z.

We next introduce a stronger notion.

Definition 3 (Fréchet Derivative) F is said to be Fréchet differentiable at z if and only if F is Gâteaux differentiable at z and the following holds:

$$F(z+h) = F(z) + F'(z)h + r(z,h) \quad \text{with} \quad \frac{\|r(z,h)\|_V}{\|h\|_Z} \to 0 \quad \text{as} \quad \|h\|_Z \to 0.$$

We refer to $F'(z)$ as the Fréchet derivative at z.

Remark 1 (Few Facts)

(i) If the Fréchet derivative exists so does the Gâteaux derivative and they coincide. However, the converse is not true in general.
(ii) We say that F is continuously Gâteaux differentiable if $F'(\cdot)$ exists and $F'(\cdot)$ is continuous. In that case F is Fréchet differentiable [18, pp. 35–36].
(iii) Let $E = G(F(z))$ where F is Gâteaux differentiable at z and G is Fréchet differentiable at $F(z)$, then E is Gâteaux differentiable.[1]

Notice that when $V = \mathbb{R}$ then $\mathscr{L}(Z, V) = Z^*$. In addition if F is Gâteaux differentiable at z then we have

$$F'(z)h = \langle F'(z), h\rangle_{Z^*,Z},$$

where Z^* is the dual space of Z and $\langle\cdot,\cdot\rangle_{Z^*,Z}$ denotes the duality pairing.

Before we conclude this subsection, we apply the above introduced definitions to two prototypical quadratic functionals. The derivatives in both these examples are Fréchet derivatives.

Example 1 Let $(H, (\cdot,\cdot)_H)$ be a real Hilbert space and $F : H \to \mathbb{R}$ defined as $F(z) := \|z\|_H^2 = (z,z)_H$, then for all $z, h \in H$ we have

$$F(z+h) - F(z) = 2(z,h)_H + \|h\|_H^2.$$

Thus,

$$F'(z)h = (2z,h)_H.$$

Using the Riesz Representation Theorem (identify H with its dual H^*), we can write

$$(\nabla F(z), h)_H = \langle F'(z), h\rangle_{H^*,H},$$

where $\nabla F(z) \in H$ is the representative of $F'(z) \in H^*$. We refer to $\nabla F(z) \in H$ as the gradient of F at z. In the above case, we have $\nabla F(z) = 2z$.

[1] The chain rule only requires the outer function to be Hadamard directionally differentiable and the inner function to be Hadamard (Gâteaux) directionally differentiable [79, Proposition 3.6].

Remark 2 (Gradient) As can be seen from the above example, the expression that we obtain by identifying $F'(z) \in H^*$ with an element of H is called the gradient of F. We will use the notation $\nabla F(z)$ to denote the gradient. We further notice that the definition of the gradient depends on the underlying inner product.

Example 2 Let $(Z, (\cdot,\cdot)_Z)$, $(H, (\cdot,\cdot)_H)$ be real Hilbert spaces and $u_d \in H$ be fixed. Let $S \in \mathscr{L}(Z, H)$. Consider $E : Z \to \mathbb{R}$,

$$E(z) = \|Sz - u_d\|_H^2.$$

Then $E(z) = G(F(z))$, where $G(v) = \|v\|_H^2$ and $F(z) = Sz - u_d$.

Next using the chain rule, we obtain that

$$\begin{aligned}\langle E'(z), h\rangle_{Z^*,Z} &= \langle G'(F(z)), F'(z)h\rangle_{H^*,H} = (2v, F'(z)h)_H \\ &= 2(Sz - u_d, Sh)_H = 2\langle S^*(Sz - u_d), h\rangle_{Z^*,Z},\end{aligned}$$

where $S^* \in \mathscr{L}(H^*, Z^*)$ is the adjoint of S. Here we have assumed that S^*S and S^*u_d are well defined. Thus $E'(z) = S^*(Sz - u_d) \in Z^*$. Since Z is a Hilbert space, similar to the previous example, we can again apply the Riesz representation theorem to get the representative $\nabla E(z) \in Z$ of $E'(z)$.

2.3 First-Order Necessary Optimality Conditions

We conclude this section with the following result on the first-order necessary optimality conditions.

Theorem 2 *Let Z be a real Banach space (not necessarily reflexive). Let $f : \mathscr{Z} \to \mathbb{R}$ be Gâteaux differentiable in $\mathscr{Z}$, where $Z_{ad} \subset \mathscr{Z} \subset Z$, $\mathscr{Z}$ open. If $\bar{z} \in Z_{ad}$ is a solution to (3) then the first-order necessary optimality conditions are*

$$\langle f'(\bar{z}), z - \bar{z}\rangle_{Z^*,Z} \geq 0 \quad \forall z \in Z_{ad}. \tag{4}$$

In addition, if f is convex and $\bar{z} \in Z_{ad}$ solves (4) then $\bar{z}$ is a solution to (3), i.e., (4) is necessary and sufficient.

Proof The proof of the first part is a direct consequence of the definition of Gâteaux derivative. Let $z \in Z_{ad}$ be arbitrary. By the convexity of Z_{ad} we have that $\bar{z} + t(z - \bar{z}) \in Z_{ad}$ for all $t \in [0, 1]$. From the optimality of $\bar{z}$ it follows that

$$f(\bar{z} + t(z - \bar{z})) - f(\bar{z}) \geq 0 \quad \forall t \in [0, 1].$$

Dividing both sides by t and taking the limit as t approaches 0^+, we arrive at (4).

Next we use the convexity of f, i.e., for all $t \in (0, 1]$ $f(\bar{z} + t(z - \bar{z})) \leq (1 - t)$ $f(\bar{z}) + tf(z)$. By rearranging terms and taking the limit as t approaches 0^+, we arrive at

$$f(z) - f(\bar{z}) \geq \langle f'(\bar{z}), z - \bar{z} \rangle_{Z^*,Z} \quad \forall z \in Z_{ad}.$$

We then obtain the desired sufficient condition by using (4).

Remark 3 We make the following observations:

i. Theorem 2 only requires f to be directionally differentiable and it holds after replacing $\langle f'(\bar{z}), z - \bar{z} \rangle_{Z^*,Z}$ by $f'(z, z - \bar{z})$.
ii. Notice that the existence of a $\bar{z} \in Z_{ad}$ in Theorem 2 that solves (3) can be satisfied under the assumptions of Theorem 1.
iii. In general, for a nonconvex f, we cannot expect to achieve a global minimum but only a local minimum. We call $\bar{z} \in Z_{ad}$ a local minimum to (3) if there exists an $\varepsilon > 0$ such that

$$f(\bar{z}) \leq f(z) \quad \forall z \in Z_{ad} \cap B_\varepsilon(\bar{z})$$

where $B_\varepsilon(\bar{z}) \subset Z$ is a ball of radius ε centered at $\bar{z}$.
iv. Equation (4) is known as a variational inequality.

3 Application to PDE-Constrained Optimization Problems

Let Z be a real reflexive Banach space and U, Y be real Banach spaces. We begin by recalling the abstract problem (1)–(2)

$$\min_{(u,z)\in U\times Z} J(u, z) \quad \text{subject to} \quad e(u, z) = 0, \quad z \in Z_{ad},\ u \in U_{ad}, \tag{5}$$

where $J : U \times Z \to \mathbb{R}$ and $e : U \times Z_{ad} \to Y$ where $Z_{ad} \subset Z$ is closed convex. We refer to (5) as the *full-space* form. Before we proceed further, we remark that often the cost functional J has two components

$$J(u, z) = J_1(u) + J_2(z),$$

where $J_1 : U \to \mathbb{R}$ is the objective (to be attained) and $J_2 : Z \to \mathbb{R}$ is the control penalty.

Another way to write (5) is by letting $X := U \times Z$, $X_{ad} := U_{ad} \times Z_{ad}$. Then we seek $x \in X_{ad}$ such that

$$\min_{x\in X_{ad}} J(x) \quad \text{subject to} \quad e(x) = 0. \tag{6}$$

Notice that (6) does not assume splitting of the control and state variables and is a generalization of (5).

By eliminating the state variables, we obtain a third form of (5) and we call it the *reduced form*. Specifically, this requires existence of a solution operator: $S : Z \to U$, which assigns each control to a unique state

$$z \mapsto S(z) = u(z) \quad \text{where } u(z) \text{ satisfies } \quad e(u(z), z) = 0.$$

Thus we define the reduced cost functional as $\mathscr{J} : Z \to \mathbb{R}$

$$\mathscr{J}(z) := J(S(z), z).$$

Instead of (5) we then solve

$$\begin{aligned} &\min_{z \in Z_{ad}} \mathscr{J}(z) \\ &\text{subject to} \\ &S(z) \in U_{ad}. \end{aligned} \tag{7}$$

We remark that the formulations (5) and (7) are not equivalent in general. For instance, there are applications where the solution operator S does not exist, i.e., the problem (7) may not have a solution but (5) is still solvable.

We next state existence result for (7). For simplicity of presentation from here on we will assume that $U_{ad} = U$, i.e., no state constraints. However the discussion can be easily adapted to this more general situation.

Corollary 1 *Let the following assumptions hold*

(i) $Z_{ad} \subseteq Z$ *is closed and convex.*
(ii) *For each* $z \in Z_{ad}$, *there exists a unique mapping* $S(z)$ *that solves* $e(S(z), z) = 0$.
(iii) S *is weakly continuous, i.e., if* $z_n \rightharpoonup z$ *in* Z_{ad} *then* $S(z_n) \rightharpoonup S(z)$ *in* U.
(iv) *The lower* γ*-level set* $\{z \in Z_{ad} : \mathscr{J}(z) \leq \gamma\}$ *of* $\mathscr{J}$ *is nonempty and bounded for some* $\gamma \in \mathbb{R}$.
(v) J_1 *is continuous and convex and* J_2 *is weakly lower semicontinuous.*

Then there exists a solution to (7).

Proof We notice that since J_1 is continuous and convex, therefore it is weakly lower semicontinuous [81, Theorem 2.12]. The weak continuity of S combined with the weak lower semicontinuity of J_2 implies that $\mathscr{J}$ is weakly lower semicontinuous. The proof then follows using Theorem 1.

We remark that in many applications we replace the lower γ-level set by either boundedness of Z_{ad} or the coercivity of J_2. More details will be provided in the subsequent sections.

At times it is more suitable to directly work with the full-space (5) form as the reduced form (7) may not even exist. This requires us to use the Lagrangian functional; we will discuss this in Section 3.2. Another advantage of using Lagrangian

formulation is the ease with which it allows us to derive the first- and second-order derivatives. This will be discussed in Section 3.2. We first consider the derivation of first-order optimality conditions for the reduced form in Section 3.1.

3.1 Reduced Form: First-Order Necessary Optimality Conditions

Corollary 2 *Let all the assumptions of Corollary 1 hold except Z being reflexive. Let $\mathscr{Z}$ be an open set in Z such that $Z_{ad} \subset \mathscr{Z}$ such that $z \mapsto S(z) : \mathscr{Z} \to U$ is Gâteaux differentiable with derivative*

$$S'(z) \in \mathscr{L}(Z, U),$$

$(u, z) \mapsto J(u, z) : U \times Z \to \mathbb{R}$ is Fŕechet differentiable with

$$J'(u, z) \in \mathscr{L}(U \times Z, \mathbb{R}).$$

If $\bar{z}$ is minimizer of (5) over Z_{ad} then the first-order necessary optimality conditions are given by

$$\langle S'(\bar{z})^* J_u(S(\bar{z}), \bar{z}) + J_z(S(\bar{z}), \bar{z}), z - \bar{z} \rangle_{Z^*,Z} \geq 0 \quad \forall z \in Z_{ad}, \tag{8}$$

where J_u and J_z are the partial derivatives of J. If $\mathscr{J}$ is convex then the condition (8) is sufficient.

Proof The proof is a consequence of Theorem 2. Let $\bar{z}$ be a solution of (5) then from Theorem 2 we have that $\langle \mathscr{J}'(\bar{z}), z - \bar{z} \rangle_{Z^*,Z} \geq 0$ for all $z \in Z_{ad}$. Combining this with the directional derivative and setting $h := z - \bar{z}$, we obtain that

$$\langle \mathscr{J}'(\bar{z}), h \rangle_{Z^*,Z} = J'(S(z), z)h = \langle J_u(S(z), z), S'(z)h \rangle_{U^*,U} + \langle J_z(S(z), z), h \rangle_{Z^*,Z}$$

where we have used the chain rule for derivatives which holds under the stated differentiability assumptions on S and J. Since $S'(z)^*$ is well defined, we conclude that

$$\langle \mathscr{J}'(\bar{z}), h \rangle_{Z^*,Z} = \langle S'(z)^* J_u(S(z), z), h \rangle_{Z^*,Z} + \langle J_z(S(z), z), h \rangle_{Z^*,Z}$$

This completes the proof.

To further understand the structure of $S'(\bar{z})$, we assume that the PDE constraint function e is sufficiently smooth and the conditions of the implicit function theorem hold. Upon differentiating the state equation, we obtain that

$$e_u(S(\bar{z}), \bar{z}) S'(\bar{z})h = -e_z(S(\bar{z}), \bar{z})h.$$

Subsequently, we arrive at

$$S'(\bar{z})h = -e_u(S(\bar{z}), \bar{z})^{-1} \left(e_z(S(\bar{z}), \bar{z})h\right). \tag{9}$$

Substituting this in (8), we obtain that

$$-\left\langle e_z(S(\bar{z}), \bar{z})^* \left((e_u(S(\bar{z}), \bar{z})^{-1})^* J_u(S(\bar{z}), \bar{z})\right), z - \bar{z}\right\rangle_{Z^*,Z} + \langle J_z(S(\bar{z}), \bar{z}), z - \bar{z}\rangle_{Z^*,Z} \geq 0.$$

Introducing the adjoint variable $\bar{p}$ and solving the adjoint equation,

$$e_u(S(\bar{z}), \bar{z})^* \bar{p} = J_u(S(\bar{z}), \bar{z}), \tag{10}$$

we arrive at the following reformulation of (8)

$$-\left\langle e_z(S(\bar{z}), \bar{z})^* \bar{p}, z - \bar{z}\right\rangle_{Z^*,Z} + \langle J_z(S(\bar{z}), \bar{z}), z - \bar{z}\rangle_{Z^*,Z} \geq 0. \tag{11}$$

Notice that

$$\mathscr{J}'(z) = -e_z(S(z), z)^* p + J_z(S(z), z) \in Z^*, \tag{12}$$

is the derivative of $\mathscr{J}$ at z. We summarize the computation of $\mathscr{J}'(z)$ in Algorithm 1.

Algorithm 1 requires two PDE solvers (possibly nonlinear in Step 1, linear PDE in Step 2).

In order to get the gradient, we can again apply the Riesz representation theorem (under the assumption that Z is a Hilbert space or admits a representation) to get a representative $\nabla \mathscr{J}(z)$ satisfying

$$(\nabla \mathscr{J}(z), v)_Z = \langle \mathscr{J}'(z), v\rangle_{Z^*,Z} \quad \forall v \in Z.$$

Having the expression of the gradient in hand, we can develop a gradient-based optimization solver. We will provide another derivation of the first-order optimality conditions in Section 3.2 using the Lagrangian approach.

In order to design Newton-based methods, it is desirable to have the second-order derivative information of the reduced functional. This can be obtained by using either the reduced functional approach or the Lagrangian approach. We will provide a brief discussion in the next section as well. More details will be given in subsequent chapters, we also refer to [44, 53].

Algorithm 1 Derivative computation using adjoints

1: Given z, solve $e(u, z) = 0$ for the state u.
2: Solve the adjoint equation $e_u(u(z), z)^* p = J_u(u(z), z)$ for p.
3: Compute $\mathscr{J}'(z) = J_u(u(z), z) - e_z(u(z), z)^* p(z)$.

3.2 Lagrangian Formulation

3.2.1 First-Order Optimality Conditions

The full-space form requires us to introduce Lagrangian functional: $L : U \times Z_{ad} \times Y^* \to \mathbb{R}$,

$$L(u, z, p) = J(u, z) - \langle e(u, z), p\rangle_{Y,Y^*}. \tag{13}$$

where Y^* is the dual space of Y (recall that $e : U \times Z_{ad} \to Y$). Notice that if we set $u = S(z)$ in (13) then $e(S(z), z) = 0$ and we obtain that

$$\mathscr{J}(z) = J(S(z), z) = L(S(z), z, p) \quad \text{for any } p \in Y^*. \tag{14}$$

Now if $(\bar{u}, \bar{z}, \bar{p})$ denotes a stationary point, then the partial derivatives of $L(u, z, p)$ with respect to u, z, and p at $(\bar{u}, \bar{z}, \bar{p})$ vanish and as a result we obtain

$$L_p(\bar{u}, \bar{z}, \bar{p}) = 0,$$

which reduces to the state equation

$$e(\bar{u}, \bar{z}) = 0.$$

Also

$$L_u(\bar{u}, \bar{z}, \bar{p}) = 0, \tag{15}$$

which is just the adjoint equation (10). Indeed

$$\begin{aligned}\langle L_u(\bar{u}, \bar{z}, \bar{p}), \xi\rangle_{U^*,U} &= \langle J_u(\bar{u}, \bar{z}), \xi\rangle_{U^*,U} - \langle \bar{p}, e_u(\bar{u}, \bar{z})\xi\rangle_{Y^*,Y} \\ &= \langle J_u(\bar{u}, \bar{z}) - e_u(\bar{u}, \bar{z})^*\bar{p}, \xi\rangle_{U^*,U}.\end{aligned}$$

In other words

$$L_u(\bar{u}, \bar{z}, \bar{p}) = J_u(\bar{u}, \bar{z}) - e_u(\bar{u}, \bar{z})^*\bar{p}.$$

Finally,

$$\langle L_z(\bar{u}, \bar{z}, \bar{p}), z - \bar{z}\rangle_{Z^*,Z} \geq 0 \quad \forall z \in Z_{ad},$$

which is equivalent to the variational inequality for the control (11). Indeed we have that the gradient of the reduced function $\mathscr{J}$ at z (12) is

$$\mathscr{J}'(z) = L_z(u, z, p),$$

where u and p solve the state and the adjoint equations, respectively.

Few comments are in order, first of all the above approach provides an elegant way to derive the first-order necessary optimality conditions and is highly recommended. As we will discuss below, the above approach also allows us to easily derive the second-order derivatives for the reduced functional. Secondly, even though the above introduced Lagrangian L is rigorous, however, we have not yet addressed the question of existence of the Lagrange multiplier p which makes the approach "formal." The existence of Lagrange multiplier p can be shown by using the Karush–Kuhn–Tucker (KKT) theory in function spaces [81, Chapter 6]. This theory requires the Robinson's regularity condition [71] or the Zowe–Kurcyusz constraint qualification [85], see [53, Chapter 1]. In certain cases, in particular, in the absence of control constraints and linear PDE constraints, one can also use the inf-sup theory for saddle point problems [39] to show existence of the Lagrange multipliers.

3.2.2 Second-Order Derivatives

Next we focus on deriving the expression of the second-order derivative of the reduced functional $\mathscr{J}$. The second-order derivative information can significantly improve the convergence rates for optimization algorithms. For instance in the absence of control constraints, the first-order necessary optimality conditions are $\mathscr{J}'(\bar{z}) = 0$ in Z^*. In order to solve for z, one can use Newton's method which is quadratically convergent (locally). Each iteration of Newton's method requires solving

$$\mathscr{J}''(z)v = -\mathscr{J}'(z) \quad \text{in } Z^* \tag{16}$$

for a direction $v \in Z$. In general (for large problems), it is often too expensive to store and factorize the Hessian. Instead it is often more practical to solve (16) using iterative methods that only require Hessian-times-vector products. We will discuss the computation of Hessian-times-vector product next. We remark that in case of bound constraints on the control one can use a superlinearly (locally) convergent semismooth Newton method [48, 56, 82].

We will proceed by using the Lagrangian approach. We operate under the assumption that J and e are twice continuously differentiable.

From (14), we recall that

$$\mathscr{J}(z) = J(u(z), z) = L(u(z), z, p)$$

$u(z) = S(z)$ solves the state equation and $p \in Y^*$ is arbitrary. After differentiating this expression in a direction h_1, we obtain that

$$\langle \mathscr{J}'(z), h_1 \rangle_{Z^*,Z} = \langle L_u(u(z), z, p), u'(z)h_1 \rangle_{U^*,U} + \langle L_z(u(z), z, p), h_1 \rangle_{Z^*,Z}.$$

Again differentiating this expression in a direction h_2 and choosing a particular p that solves the adjoint equation (15), we arrive at

$$\mathscr{J}''(z) = T(S(z), z)^* H(S(z), z, p) T(S(z), z),$$

where

$$T(u, z) = \begin{pmatrix} -e_u(u, z)^{-1} e_z(u, z) \\ I_Z \end{pmatrix}$$

with I_Z denoting the identity map on Z and $H(u, z, p)$ is given by

$$H(u, z, p) = \begin{pmatrix} L_{uu}(u, z, p) & L_{uz}(u, z, p) \\ L_{zu}(u, z, p) & L_{zz}(u, z, p) \end{pmatrix}.$$

Then one can compute the Hessian vector product by using Algorithm 2. Notice that Algorithm 2 requires two linear PDE solvers in Steps 3 and 4. We refer to [53, Chapter 1] and [44] for further details.

So far our approach has been general. For the chapter remainder, we will focus on two particular examples where the cost functional is quadratic and the PDE constraints are linear and semilinear elliptic PDEs. In order to develop the notion of solutions to these PDEs, we first introduce Sobolev spaces.

4 Sobolev Spaces

In this section, we introduce the necessary function spaces to be used throughout the chapter remainder. Let $\Omega \subset \mathbb{R}^n$ be an open, bounded domain with Lipschitz boundary $\partial\Omega$. For $1 \leq p < \infty$, we denote by $L^p(\Omega)$ the Banach space

$$L^p(\Omega) := \left\{ v : \Omega \to \mathbb{R} \ : \ v \text{ is measurable and } \int_\Omega |v(x)|^p \, dx < \infty \right\}$$

Algorithm 2 Hessian-Times-Vector Computation

1: Given z, solve $e(u, z) = 0$ for u (if not done already).
2: Solve adjoint equation: $e_u(u, z)^* p = J_u(u, z)$ for p (if not done already).
3: Solve $e_u(u, z)w = -e_z(u, z)v$.
4: Solve $e_u(u, z)^* q = L_{uu}(u, z, p)w + L_{uz}(u, z, p)v$.
5: Compute

$$\mathscr{J}''(z)v = -e_z(u, z)^* q + L_{uz}(u, z, p)w + L_{zz}(u, z, p)v.$$

with the norm $\|v\|_{L^p(\Omega)} := \left(\int_\Omega |v(x)|^p \, dx\right)^{\frac{1}{p}}$. These spaces are equivalence classes of functions equal up to a set of measure zero [37]. In particular when $p = 2$, we obtain $L^2(\Omega)$ which is a Hilbert space with inner product $(u, v)_{L^2(\Omega)} = \int_\Omega u(x)v(x) \, dx$. When $p = \infty$, we obtain $L^\infty(\Omega)$, a Banach space with norm $\|v\|_{L^\infty(\Omega)} := \text{ess sup}_\Omega |v|$.

Moving forward, we use multi-index notation to define partial derivatives. For a multi-index $\gamma = (\gamma_1, \ldots, \gamma_n) \in \mathbb{N}_0^n := (\mathbb{N} \cup \{0\})^n$, we let its order to be $|\gamma| := \sum_{i=1}^n \gamma_i$. The associated $|\gamma|$-th order partial derivative of a function u at x is

$$D^\gamma u(x) := \frac{\partial^{|\gamma|} u}{\partial x_1^{\gamma_1} \ldots \partial x_n^{\gamma_n}}(x).$$

Then we denote by $W^{k,p}(\Omega)$ the Sobolev spaces with the norm

$$\|u\|_{W^{k,p}(\Omega)} := \begin{cases} \left(\sum_{|\gamma| \le k} \int_\Omega |D^\gamma u|^p \, dx\right)^{1/p} & 1 \le p < \infty \\ \sum_{|\gamma| \le k} \text{ess sup}_\Omega |D^\gamma u| & p = \infty. \end{cases}$$

If $p = 2$, we write

$$H^k(\Omega) = W^{k,2}(\Omega), \quad k = 0, 1, \ldots$$

which is a Hilbert space with inner product

$$(u, v)_{H^k(\Omega)} = \sum_{|\gamma| \le k} (D^\gamma u, D^\gamma v)_{L^2(\Omega)}.$$

Notice that $H^0(\Omega) = L^2(\Omega)$.

We denote by $W_0^{k,p}(\Omega)$ the closure of $C_0^\infty(\Omega)$ with respect to $W^{k,p}(\Omega)$-norm. Thus $u \in W_0^{k,p}(\Omega)$ if and only if there exist functions $u_m \in C_0^\infty(\Omega)$ such that $u_m \to u$ in $W^{k,p}(\Omega)$. The space $H_0^1(\Omega)$ consists of functions $u \in H^1(\Omega)$ such that

$$u = 0 \quad \text{on } \partial\Omega,$$

in the trace sense. Using the Poincaré inequality $\|u\|_{L^2(\Omega)} \le C\|\nabla u\|_{L^2(\Omega)}$, where $C = C(\Omega)$, we have

$$\|u\|_{H^1(\Omega)} \le C\|\nabla u\|_{L^2(\Omega)} \quad \forall u \in H_0^1(\Omega).$$

Finally, we denote the dual of $H_0^1(\Omega)$ by $H^{-1}(\Omega)$. It is easy to see that $L^2(\Omega)$ is continuously embedded in $H^{-1}(\Omega)$. For more details, we refer to [1].

5 Second-Order Linear Elliptic PDEs

In this section, we study the second-order elliptic PDEs. In general, we cannot expect classical solutions to these PDEs, therefore we first introduce the notion of weak solutions in Section 5.1. We will also study the notion of 'strong' solutions for this problem in Section 5.2. This higher regularity (strong solutions) will help us to establish the approximability of the continuous solution using the finite-element method (cf. Section 8). Strong solutions, in addition, also play a role in other situations, for instance, in studying the regularity of multipliers in state-constrained problems [20] and PDECO problems with variational inequalities [49].

5.1 Existence and Uniqueness

We begin this section by making certain uniform ellipticity assumptions.

Assumption 3 (Coefficient Matrix) *Let A be an $n \times n$ matrix with entries a_{ij} for $1 \le i, j \le n$. We assume that a_{ij} are measurable, belong to $L^\infty(\Omega)$, and are symmetric, that is, $a_{ij}(x) = a_{ji}(x)$ for all $1 \le i, j \le n$ and for a.e. $x \in \Omega$. We further assume that A is positive definite and satisfy the uniform ellipticity condition*

$$\exists\, \beta \ge \alpha > 0 \quad \text{such that} \quad \alpha|\xi|^2 \le A(x)\xi \cdot \xi \le \beta|\xi|^2 \quad \forall \xi \in \mathbb{R}^n, \ \text{a.e. } x \text{ in } \Omega. \tag{17}$$

Given f, we consider the following linear elliptic PDE

$$\begin{aligned} -\text{div}\,(A\nabla u) &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{18}$$

We understand (18) in a *weak sense*, i.e., given $f \in H^{-1}(\Omega)$, we seek a solution $u \in H_0^1(\Omega)$ that satisfies

$$\int_\Omega A\nabla u \cdot \nabla v \, dx = \langle f, v\rangle_{-1,1}, \quad \forall v \in H_0^1(\Omega), \tag{19}$$

where $\langle \cdot, \cdot \rangle_{-1,1}$ denotes the duality pairing between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$.

Theorem 4 *For every $f \in H^{-1}(\Omega)$, there exists a unique weak solution $u \in H_0^1(\Omega)$ that fulfills*

$$\|u\|_{H^1(\Omega)} \le C\|f\|_{H^{-1}(\Omega)}, \tag{20}$$

where the constant C only depends on Ω and α.

Proof The existence and uniqueness is due to the Lax–Milgram Lemma. Moreover, the bound (20) immediately follows by using the fact that A is uniformly positive definite and the Poincaré inequality [38].

Remark 4 In general, for Sobolev spaces $W^{1,p}(\Omega)$ with $p \neq 2$, we need the inf-sup conditions to prove Theorem 4. The Banach–Nečas theorem then guarantees existence and uniqueness of u [34, Theorem 2.6]. The latter is a necessary and sufficient condition. See also [67].

5.2 Regularity

It is imperative to understand the regularity of solution u to (19). For instance, such an understanding allows us to develop numerical schemes with optimal rate of convergence (see Section 8). It can assist us with proving rates of convergence for the optimization algorithms [47].

Theorem 5 *Let $u \in H_0^1(\Omega)$ be a weak solution of (19) and let the coefficient matrix A satisfy Assumption 3.*

- *If $f \in L^2(\Omega)$ and Ω is a convex polytope or $C^{1,1}$ domain in $\mathbb{R}^n$, then $u \in H^2(\Omega) \cap H_0^1(\Omega)$ and there exists a constant $C = C(\alpha, \beta, \Omega)$ such that*

$$\|u\|_{H^2(\Omega)} \leq C \|f\|_{L^2(\Omega)}.$$

- *If $f \in L^p(\Omega)$ for $1 < p < \infty$ and Ω is $C^{1,1}$, then $u \in W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)$ and there exists a constant $C = C(\alpha, \beta, \Omega, p)$ such that*

$$\|u\|_{W^{2,p}(\Omega)} \leq C \|f\|_{L^p(\Omega)}.$$

  If $p > n$, then $u \in C^{1,\alpha}(\bar{\Omega})$ with $\alpha = 1 - n/p$.

Proof If Ω is a convex polygonal/polyhedral domain, then H^2-regularity is in [40, 3.2.1.2]. When $\partial\Omega$ is $C^{1,1}$ and $f \in L^p(\Omega)$ for any $1 < p < \infty$, the result is due to [38, Theorem 9.15]. In the case $p > n$, the $C^{1,\alpha}$ regularity follows from $W^{2,p}$ regularity and the Sobolev embedding.

6 Linear Quadratic PDE-Constrained Optimization Problem

Having some basic understanding of elliptic PDEs in hand, we next apply the results of Section 3 to a linear quadratic PDECO problem (cf. Section 6.1). In Section 6.2, we formulate it as a reduced PDECO problem only in terms of the control variable z. This allows us to use the direct method of calculus of variations from Theorem 1 to show the existence of solution.

6.1 Problem Formulation

Let $u_d \in L^2(\Omega)$ and $z_a, z_b \in L^2(\Omega)$ with $z_a < z_b$ a.e. in Ω being given. Moreover, let $\lambda \geq 0$ denotes the penalty parameter. Then we are interested in minimizing

$$J(u,z) = \frac{1}{2}\|u - u_d\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\|z\|^2_{L^2(\Omega)} \tag{21}$$

subject to (in the weak form)

$$\begin{aligned} -\text{div}\,(A\nabla u) &= z \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{22}$$

and the pointwise control constraints

$$z \in Z_{ad} := \{v \in L^2(\Omega) : z_a(x) \leq v(x) \leq z_b(x)\ ,\ a.e.\ x \in \Omega\}. \tag{23}$$

Notice that in our formulation above we also allow $z_a = -\infty$ and $z_b = \infty$. We call this case as an unconstrained case, i.e., $Z_{ad} = L^2(\Omega)$.

For the above problem, we have (cf. Section 3)

$$U = H^1_0(\Omega), \quad Y = H^{-1}(\Omega), \quad Z = L^2(\Omega).$$

In order to understand the problem (21)–(23) similar to (7), first we condense the variables and write J only in terms of z. We again call this as the reduced form and the resulting cost functional as the reduced functional. We discuss this next.

6.2 Reduced PDECO Problem

For every $z \in Y$, there exists a unique solution $u = u(z) \in U$ to (22). As a result, we can define the solution operator to (22) as

$$\mathscr{S} : Y \to U, \quad z \mapsto u(z) = \mathscr{S}z,$$

which is linear and continuous. In view of the continuous embedding, $H^1_0(\Omega) \hookrightarrow L^2(\Omega) \hookrightarrow H^{-1}(\Omega)$ we may also consider $\mathscr{S}$ as a map from $L^2(\Omega)$ to $L^2(\Omega)$. In other words instead of $\mathscr{S}$, we consider the operator $S := \mathscr{E}_u \mathscr{S} \mathscr{E}_z$, where $\mathscr{E}_z : L^2(\Omega) \to H^{-1}(\Omega)$ and $\mathscr{E}_u : H^1_0(\Omega) \to L^2(\Omega)$ denote the embedding operators that assign to each $z \in L^2(\Omega)$ and $u \in H^1_0(\Omega)$ the functions $z \in H^{-1}(\Omega)$ and $u \in L^2(\Omega)$ so that when these new functions are restricted to $L^2(\Omega)$ and $H^1_0(\Omega)$ the operator yields the original functions, respectively. Notice that $S : L^2(\Omega) \to L^2(\Omega)$. One of the main advantages of using S is the fact that the adjoint operator

S^* also acts on $L^2(\Omega)$ (cf. Section 6.3). Using the solution map S, we arrive at the so-called reduced cost $\mathscr{J} : L^2(\Omega) \to \mathbb{R}$ which is defined as

$$\mathscr{J}(z) := J(Sz, z),$$

and the minimization problem (21)–(23) is equivalent to the reduced problem:

$$\min_{z \in Z_{ad}} \mathscr{J}(z). \tag{24}$$

We notice that it is more convenient to analyze (24) in comparison to (21)–(23). In fact, we have the following well-posedness result.

Corollary 3 *Let either $\lambda > 0$ or Z_{ad} be bounded. Then there exists a solution to the minimization problem (24). If in addition $\lambda > 0$ or S is injective then the solution is unique.*

Proof The proof is a consequence of Theorem 1. We will first show that $\mathscr{J}$ is weakly lower semicontinuous. Notice that $\mathscr{J}(z) = \mathscr{J}_1(Sz) + \mathscr{J}_2(z)$ where $\mathscr{J}_1(Sz) := \frac{1}{2}\|Sz - u_d\|^2_{L^2(\Omega)}$ and $\mathscr{J}_2(z) := \frac{\lambda}{2}\|z\|^2_{L^2(\Omega)}$. Clearly $\mathscr{J}_2$ is weakly lower semicontinuous (convexity and continuity imply weak lower semicontinuity [18, Theorem 2.23]). On the other hand, due to the compact embedding of $H^1_0(\Omega)$ in $L^2(\Omega)$, we have that $S : L^2(\Omega) \to L^2(\Omega)$ is completely continuous, i.e., if $z_n \rightharpoonup z$ in $L^2(\Omega)$ then $Sz_n \to Sz$ in $L^2(\Omega)$. Thus, owing to the continuity and convexity of $\mathscr{J}_1$, we conclude that $\mathscr{J}_1$ is weakly lower semicontinuous. Whence $\mathscr{J}$ is weakly lower semicontinuous.

It then remains to characterize the lower γ-level set. Here we replace the lower γ-level set condition by the coercivity of $\mathscr{J}$ (if $\lambda > 0$) or by the closed convex bounded set Z_{ad}.

Finally uniqueness is due to strict convexity of $\mathscr{J}$.

For the remainder of this section, we will consider $\lambda > 0$.

6.3 First-Order Optimality Conditions

We are now ready to derive the first-order optimality conditions by following Section 3.1 and the expression of the gradient of the reduced objective function. In Section 6.4, we follow Section 3.2 and consider an alternate strategy to derive the optimality conditions using the Lagrangian formulation.

We recall that the reduced functional is

$$\mathscr{J}(z) = \frac{1}{2}\|Sz - u_d\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\|z\|^2_{L^2(\Omega)}.$$

Using Examples 1 and 2, the gradient of $\mathscr{J}$ is given by

$$\nabla \mathscr{J}(z) = S^*(Sz - u_d) + \lambda z \in L^2(\Omega).$$

Here we have $S : L^2(\Omega) \to L^2(\Omega)$. The first-order necessary and sufficient (due to the convexity of $\mathscr{J}$) optimality condition (4) then becomes

$$(\nabla \mathscr{J}(\bar{z}), z - \bar{z})_{L^2(\Omega)} \geq 0 \quad \forall z \in Z_{ad}. \tag{25}$$

In order to efficiently evaluate $\nabla \mathscr{J}(z)$, we introduce the so-called adjoint variable $p \in H_0^1(\Omega)$ solving

$$\begin{aligned} -\text{div}\,(A\nabla p) &= u - u_d \quad \text{in } \Omega \\ p &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{26}$$

We will next show that the adjoint operator $S^* : L^2(\Omega) \to L^2(\Omega)$ can be defined by $S^*\zeta := p$ where p solves (26) with right-hand side given by ζ. Here $\zeta \in L^2(\Omega)$ is arbitrary. Let $z \in L^2(\Omega)$ be an arbitrary right-hand side of the state equation and the resulting state variable is $u = Sz \in H_0^1(\Omega)$. By testing the equation for u with p and vice versa, we obtain that $(z, p)_{L^2(\Omega)} = (\zeta, u)_{L^2(\Omega)}$. Since $u = Sz$, we deduce that $S^*\zeta = p$. As a result, $S^*(u - u_d) = p$ where p solves (26).

Thus the gradient computation reduces to evaluation of the following expression:

$$\nabla \mathscr{J}(z) = p + \lambda z \in L^2(\Omega).$$

Finally, we gather the first-order necessary and sufficient optimality system:

$$\bar{u} \in H_0^1(\Omega): \quad \int_\Omega A\nabla\bar{u} \cdot \nabla v \, dx = \int_\Omega \bar{z} v \, dx \quad \forall v \in H_0^1(\Omega) \tag{27a}$$

$$\bar{p} \in H_0^1(\Omega): \quad \int_\Omega A\nabla\bar{p} \cdot \nabla v \, dx = \int_\Omega (\bar{u} - u_d) v \, dx \quad \forall v \in H_0^1(\Omega) \tag{27b}$$

$$\bar{z} \in Z_{ad}: \quad (\bar{p} + \lambda\bar{z}, z - \bar{z})_{L^2(\Omega)} \geq 0, \quad \forall z \in Z_{ad}. \tag{27c}$$

Notice that (27) is a coupled system, namely $\bar{u}$ in (27a) depends on the unknown optimal control $\bar{z}$ which fulfills the inequality (27c). The latter depends on the adjoint variable $\bar{p}$ that solves the adjoint equation (27b). This in turn depends on $\bar{u}$. We further remark the variational inequality (27c) for the control is equivalent to the following projection formula (see [81]):

$$\bar{z}(x) = \mathscr{P}_{Z_{ad}}\left\{-\frac{1}{\lambda}\bar{p}(x)\right\} \quad \text{a.e. } x \in \Omega. \tag{28}$$

Here $\mathscr{P}_{Z_{ad}}(v)$ denotes projection of v onto Z_{ad}. In our case, we can further write this as $\mathscr{P}_{Z_{ad}}(v) := \min\{b(x), \max\{a(x), v(x)\}\}$, x a.e. in Ω.

6.4 Lagrange Method

An alternative approach to derive the first-order optimality system is using the Lagrangian as described in Section 3.2. We again emphasize that even though this approach is used formally in the following, it can be made rigorous. It provides a systematic way of deriving the optimality system, especially for tedious problems and this step is strongly recommended.

Introduce $L : H_0^1(\Omega) \times Z_{ad} \times H^1(\Omega) \to \mathbb{R}$, defined as

$$L(u, z, p) := J(u, z) - \int_\Omega (A\nabla u \cdot \nabla p - zp)\ dx.$$

If $(\bar{u}, \bar{z}, \bar{p})$ is a stationary point then

$$\begin{aligned} \langle L_p(\bar{u}, \bar{z}, \bar{p}), h\rangle_{H^{-1}(\Omega), H_0^1(\Omega)} &= 0 \quad \forall h \in H_0^1(\Omega), \\ \langle L_u(\bar{u}, \bar{z}, \bar{p}), h\rangle_{H^{-1}(\Omega), H_0^1(\Omega)} &= 0 \quad \forall h \in H_0^1(\Omega), \\ (L_z(\bar{u}, \bar{z}, \bar{p}), (z - \bar{z}))_{L^2(\Omega)} &\geq 0 \quad \forall z \in Z_{ad}. \end{aligned} \tag{29}$$

It is not hard to see that (29) leads to the same optimality system as in (27).

7 Semilinear Quadratic PDE-Constrained Optimization Problem

The focus of the previous section was on the linear quadratic PDECO problem. However, things are more delicate when we replace the linear PDE constraint by a semilinear one. In this section, we provide a brief discussion of a PDECO problem governed by a semilinear PDE.

Let $\Omega \subset \mathbb{R}^n$, with $n \geq 2$, be a Lipschitz domain and the Assumption 3 holds. Moreover, let $u_d \in L^2(\Omega)$ and $z_a, z_b \in L^\infty(\Omega)$ with $z_a(x) < z_b(x)$ a.e. $x \in \Omega$ and $\lambda \geq 0$ be given. We then consider the following semilinear optimal control problem:

$$\min J(u, z) := \frac{1}{2}\|u - u_d\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\|z\|^2_{L^2(\Omega)} \tag{30}$$

subject to $u \in L^\infty(\Omega) \cap H_0^1(\Omega)$ solving the weak form of

$$\begin{aligned} -\text{div}\,(A\nabla u) + f(\cdot, u) &= z \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega \end{aligned} \tag{31}$$

and

$$z \in Z_{ad} := \{v \in L^\infty(\Omega) : z_a(x) \leq v(x) \leq z_b(x)\ ,\ a.e.\ x \in \Omega\}. \tag{32}$$

Notice the difference between the semilinear state equation (31) and the linear state equation (22). The key difficulty in the above problem is due to the nonlinearity introduced by f. The control bounds fulfill $z_a, z_b \in L^\infty(\Omega)$. This is different than the linear case where we assumed the bounds to be in $L^2(\Omega)$. This choice enforces the control $z \in L^\infty(\Omega)$ which in turn provides additional regularity for the state u as we discuss next.

In order to establish existence of solution to the state equation (31), we make certain assumptions on the nonlinearity f.

Assumption 6 *For a function $f : \Omega \times \mathbb{R} \to \mathbb{R}$ we consider the following assumption:*

$$\begin{cases} f(x,\cdot) \text{ is strictly increasing} & \text{for a.e. } x \in \Omega, \\ f(x,0) = 0 & \text{for a.e. } x \in \Omega, \\ f(x,\cdot) \text{ is continuous} & \text{for a.e. } x \in \Omega, \\ f(\cdot,t) \text{ is measurable} & \text{for all } t \in \mathbb{R}, \\ \lim_{t\to\infty} f(x,t) = \infty & \text{for a.e. } x \in \Omega. \end{cases}$$

Remark 5 The condition $f(x,0) = 0$ in Assumption 6 is not a restriction. If this condition on f cannot be verified, then it is enough to rewrite Equation (31) in Ω as

$$-\text{div}\,(A\nabla u) + f(\cdot,u) - f(\cdot,0) = z - f(\cdot,0) \quad \text{in } \Omega.$$

A typical example of f that fulfills Assumption 6 is given next (cf. [10, 11]).

Example 3 Let $q \in [1,\infty)$ and let $b : \Omega \to (0,\infty)$ be a function in $L^\infty(\Omega)$, that is, $b(x) > 0$ for a.e. $x \in \Omega$. Define the function $f : \Omega \times \mathbb{R} \to \mathbb{R}$ by $f(x,t) = b(x)|t|^{q-1}t$.

Theorem 7 (Existence and Uniqueness for Semilinear PDE) *Let Assumptions 3 and 6 hold. Then for every $z \in L^p(\Omega)$ with $p > \frac{n}{2}$, there exists a unique weak solution $u \in L^\infty(\Omega) \cap H_0^1(\Omega)$ to the state equation (31) and there exists a constant $C = C(\alpha,\beta,\Omega) > 0$ such that*

$$\|u\|_{H^1(\Omega)} + \|u\|_{L^\infty(\Omega)} \le C\|z\|_{L^p(\Omega)}. \tag{33}$$

Proof The existence of solution is using the Browder–Minty theorem (cf [11, proposition 3.2] after setting $s = 1$). On the other hand, the $L^\infty(\Omega)$ regularity is by using a technique of Stampacchia [11, Theorem 3.5], see also [2, 21].

For the above minimization problem, we have (cf. Section 3)

$$U = H_0^1(\Omega) \cap L^\infty(\Omega), \quad Y = H^{-1}(\Omega), \quad Z = L^p(\Omega) \quad \text{with } p > n/2.$$

Notice that

$$Z_{ad} \subset L^\infty(\Omega) \subset Z.$$

As a result and owing to Theorem 7, the control to state map is well defined

$$S : L^\infty(\Omega) \to L^\infty(\Omega) \cap H_0^1(\Omega).$$

We notice that S is also well defined as a map from Z to $L^\infty(\Omega) \cap H_0^1(\Omega)$.[2] Due to S we can write the reduced problem as

$$\min_{z \in Z_{ad}} \mathscr{J}(z) := J(S(z), z). \tag{34}$$

In order to show the existence of solution to (34), typically f is assumed to be locally Lipschitz in the second argument to fulfill the assumptions on S in Corollary 1. This assumption was recently weakened in [11] and was replaced by the following growth condition on f: There exists a constant $c \in (0, 1]$ such that

$$c|f(x, \xi - \eta)| \leq |f(x, \xi) - f(x, \eta)| \tag{35}$$

for a.e. $x \in \Omega$ and for all $\xi, \eta \in \mathbb{R}$. Such a growth condition is fulfilled by Example 3 (cf. [11]). Under this condition, we have the following existence result for (34).

Corollary 4 *Let the Assumptions of Theorem 7 hold. In addition, let f fulfills the growth condition* (35) *and that*

$$f(\cdot, w(\cdot)) \in L^2(\Omega) \ \textit{for every } w \in L^\infty(\Omega). \tag{36}$$

Then there exists a solution to (34)*.*

Proof Similar to Corollary 3, the proof is again a consequence of Theorem 1. We interpret Z_{ad} as a subset of Z which is a reflexive Banach space. However, care must be taken to show the weak lower semicontinuity of $\mathscr{J}$. One has to carefully study the convergence of the state sequence $\{S(z_n)\}_{n \in \mathbb{N}}$. See for instance [11, Theorem 4.2].

Notice that the condition (36) is also fulfilled by Example 3.

Remark 6 We mention that all the results given in Corollary 4 remain true if one replaces the growth condition (35) and (36) with the following local Lipschitz continuity condition: For all $M > 0$ there exists a constant $L_M > 0$ such that f satisfies

$$|f(x, \xi) - f(x, \eta)| \leq L_M |\xi - \eta| \tag{37}$$

for a.e. $x \in \Omega$ and $\xi, \eta \in \mathbb{R}$ with $|\eta|, |\xi| \leq M$.

[2] Note both the choices of spaces for S are motivated by the theory of Nemytskii or superposition operators. Care must to taken to ensure their differentiability [81, Section 4.3].

For the remainder of this section, we will assume that $\lambda > 0$. Before we proceed further to derive the optimality conditions, we need some additional assumptions on f. Notice that the second-order derivatives are needed if one is interested in studying the second-order sufficient conditions.

Assumption 8 *We assume the following:*

(i) The function $f(x,\cdot)$ is k-times, with $k = 1, 2$, continuously differentiable for a.e. $x \in \Omega$.

(ii) For all $M > 0$, there exists a constant $L_M > 0$ such that f satisfies (37) and

$$\left| D_u^k f(x,\xi) - D_u^k f(x,\eta) \right| \le L_M |\xi - \eta|, \quad k = 1, 2,$$

for a.e. $x \in \Omega$ and $\xi, \eta \in \mathbb{R}$ with $|\xi|, |\eta| \le M$. Here D_u denotes the partial derivatives with respect to the second component.

(iii) $D_u f(\cdot, 0) \in L^\infty(\Omega)$.

(iv) $D_{uu} f(\cdot, u(\cdot)) \in L^\infty(\Omega)$ whenever $u(\cdot) \in L^\infty(\Omega)$.

Assumptions 8 help us prove that $S : L^\infty(\Omega) \to L^\infty(\Omega) \cap H_0^1(\Omega)$ is not only twice Fréchet differentiable (using the Implicit Function Theorem) but also twice continuous Fréchet differentiability of $\mathscr{J}$.

By invoking Corollary 2 the *first-order necessary optimality conditions* are given as follows: For every solution $\bar{z}$ of the problem (34), there exists a unique optimal state $\bar{u} = S(\bar{z})$ and an optimal adjoint state $\bar{p}$ such that

$$\begin{aligned}
\bar{u} \in L^\infty(\Omega) \cap H_0^1(\Omega) : \int_\Omega A\nabla\bar{u} \cdot \nabla v \, dx + \int_\Omega f(x,\bar{u}) v \, dx \\
= \int_\Omega \bar{z} v \, dx \quad \forall v \in H_0^1(\Omega) \\
\bar{p} \in H_0^1(\Omega) : \int_\Omega A\nabla\bar{p} \cdot \nabla v \, dx + \int_\Omega D_u f(x,\bar{u}) \bar{p} \, dx \\
= \int_\Omega (\bar{u} - u_d) v \, dx \quad \forall v \in H_0^1(\Omega) \\
\bar{z} \in Z_{ad} : (\bar{p} + \lambda\bar{z}, z - \bar{z})_{L^2(\Omega)} \ge 0, \quad \forall z \in Z_{ad}.
\end{aligned} \tag{38}$$

Alternatively, one can use the Lagrangian approach of Section 6.4 to derive (38).

Notice that the variational inequality in (38) again can be written using the Projection formula as

$$\bar{z}(x) = \mathscr{P}_{Z_{ad}} \left\{ -\frac{1}{\lambda} \bar{p}(x) \right\} \quad \text{a.e. } x \in \Omega. \tag{39}$$

We further remark that since $\mathscr{J}$ is non-convex, in general due to the semilinear state equation, we cannot expect a global unique solution to the PDECO problem but only a local one. This local uniqueness can be shown by studying second order sufficient conditions. Nevertheless, care must be taken to prove such a result. This

is due to the fact that the penalty term on the control in the cost functional is in $L^2(\Omega)$. However, the constraints in Z_{ad} are in $L^\infty(\Omega)$. This leads to the so-called $L^2(\Omega) - L^\infty(\Omega)$ norm discrepancy and should be taken into account before considering second-order sufficient conditions. We refer to [81, Theorem 4.29] for details. We further remark that the second-order sufficient conditions are a useful tool to derive the discretization error estimates [13]. A further discussion is provided in Theorem 12.

8 Discrete Optimal Control Problem

We will illustrate the main ideas of the finite-element approximation of the PDE-constrained optimization problems in the case of linear elliptic problem with control constraints (21)–(23). First, we briefly review the basics of the finite-element discretization just for the state equation. For what follows, it is sufficient to take $f \in L^2(\Omega)$. We consider the weak form of Equation (19),

$$(A\nabla u, \nabla v)_{L^2(\Omega)} = (f, v)_{L^2(\Omega)} \quad \forall v \in H_0^1(\Omega).$$

We partition the domain Ω into elements. For simplicity, we only discuss the case when elements are simplices. For $h \in (0, h_0]; h_0 > 0$, let $\mathcal{T}_h$ denote a quasi-uniform triangulation of Ω with mesh size h, i.e., $\mathcal{T}_h = \{\tau\}$ is a partition of Ω into triangles or tetrahedrons τ of diameter h_τ such that for $h = \max_\tau h_\tau$,

$$\text{diam}(\tau) \leq h \leq C|\tau|^{\frac{1}{n}}, \quad \forall \tau \in \mathcal{T}_h,$$

where the constant $C > 0$ independent of h. For simplicity, we assume $\cup\tau = \Omega$. Let V_h be the set of all functions in $H_0^1(\Omega)$ that are continuous on Ω and linear on each τ. V_h is usually called the space of conforming Lagrange piecewise linear elements.

Now we define the finite-element Galerkin approximation $u_h \in V_h$ of (8), as the unique solution of

$$(A\nabla u_h, \nabla v_h)_{L^2(\Omega)} = (f, v_h)_{L^2(\Omega)} \quad \forall v_h \in V_h. \tag{40}$$

Expanding u_h in terms of basis functions, it is easy to see that (40) is equivalent to a system of linear equations and since $V_h \subset H_0^1(\Omega)$ the resulting matrix is nonsingular. Notice that by construction

$$(A\nabla(u - u_h), \nabla v_h)_{L^2(\Omega)} = 0 \quad \forall v_h \in V_h. \tag{41}$$

Thus, the Galerkin solution u_h is the orthogonal projection of u onto V_h with respect to the inner-product $(A\nabla\cdot, \nabla\cdot)_{L^2(\Omega)}$. Almost immediately we obtain the following key result.

Lemma 1 (Céa Lemma) *Let u and u_h satisfy (41). Then the following estimate holds*

$$\|u - u_h\|_{H^1(\Omega)} \leq C \min_{\chi \in V_h} \|u - \chi\|_{H^1(\Omega)}.$$

The constant C depends only on ellipticity, boundedness of the matrix A, and the domain Ω.

The above result says that the Galerkin solution is the best approximation to u from V_h in $H^1(\Omega)$-norm up to a constant. We can use Cea's lemma to derive a priori error estimates. Let $I_h : H^1(\Omega) \to V_h$ be a projection with the approximation properties

$$\|u - I_h u\|_{H^s(\Omega)} \leq Ch^{2-s}\|u\|_{H^2(\Omega)}, \quad s = 0, 1, \tag{42}$$

then from Cea's lemma immediately follows

$$\|u - u_h\|_{H^1(\Omega)} \leq Ch\|u\|_{H^2(\Omega)}.$$

Notice that the constant $C > 0$ above is independent of h. The error estimates in $L^2(\Omega)$-norm are not immediate, since the Galerkin solution does not have a property of being best approximation in $L^2(\Omega)$-norm; nevertheless, one can still establish optimal error estimates in $L^2(\Omega)$ via a duality argument, also known as Nitsche's trick. This result requires $H^2(\Omega)$-regularity.

Lemma 2 *Let Ω be convex or $C^{1,1}$ and u and u_h satisfy (41). Then there exists a constant C independent of h such that*

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch \min_{\chi \in V_h} \|u - \chi\|_{H^1(\Omega)}.$$

Proof Let $e = u - u_h$ and consider a dual problem

$$(A\nabla w, \nabla v)_{L^2(\Omega)} = (e, v)_{L^2(\Omega)} \quad \forall v \in H_0^1(\Omega).$$

By setting, $v = e$, we obtain

$$\|e\|^2_{L^2(\Omega)} = (A\nabla w, \nabla e)_{L^2(\Omega)} = (A\nabla(w - w_h), \nabla e)_{L^2(\Omega)},$$

where the last equality is due to (41). Next using the Cauchy–Schwarz inequality and the fact that, under given regularity of the domain $\|w\|_{H^2(\Omega)} \leq C\|e\|_{L^2(\Omega)}$, we obtain the required result.

Combining Cea's lemma and Lemma 2, we immediately establish the optimal a priori error estimate

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^2\|u\|_{H^2(\Omega)}.$$

Notice that the above estimate does require the convexity of Ω.

Corollary 5 *H^2 regularity also allows us to express the error in terms of data. Thus from Lemmas 1 and 2 it follows*

$$\|u - u_h\|_{L^2(\Omega)} + h\|u - u_h\|_{H^1(\Omega)} \leq Ch^2\|f\|_{L^2(\Omega)}.$$

8.1 Discrete Linear Quadratic PDE-Constrained Optimization Problem

For the remainder chapter, we will assume that $\lambda > 0$. To discretize the problem, we replace the state space $H_0^1(\Omega)$ with $V_h \subset H_0^1(\Omega)$ and the control space of admissible functions Z_{ad} with $Z_{ad,h} \subset Z_{ad}$. In case of unconstrained control, we can choose $Z_{ad,h} = V_h$. Theoretically, the mesh for the discretization of the state variable and the mesh for the discretization of the control can be different. However having two different meshes adds more technical difficulties for implementation. For this reason, it is more convenient to work with the same mesh which is what we assume from now on. Thus, the discretized problem (21)–(23) becomes

$$\min_{u_h \in V_h,\ z_h \in Z_{ad,h}} J_h(u_h, z_h) = \frac{1}{2}\|u_h - u_d\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\|z_h\|^2_{L^2(\Omega)} \tag{43}$$

subject to

$$(A\nabla u_h, \nabla v_h) = (z_h, v_h) \quad \forall v_h \in V_h. \tag{44}$$

Similar to the infinite dimensional case, we define a discrete solution operator $S_h : Z_{ad,h} \to V_h$ to (44) and the reduced discrete problem becomes

$$\min_{z_h \in Z_{ad,h}} \mathscr{J}_h(z_h) := \min_{z_h \in Z_{ad,h}} J_h(S_h z_h, z_h). \tag{45}$$

Similar to the continuous problem, one can show that problem (45) has a unique solution $\bar{z}_h \in Z_{ad,h}$, the corresponding discrete optimal state is $\bar{u}_h = S_h(\bar{z}_h)$, and similar to Theorem 2 the first-order necessary and sufficient optimality condition is

$$\mathscr{J}_h'(\bar{z}_h)(z_h - \bar{z}_h) \geq 0, \quad \forall z_h \in Z_{ad,h}. \tag{46}$$

8.2 Optimization Problem Without Control Constraints

In this situation, $Z_{ad} = L^2(\Omega)$ and $Z_{ad,h} = V_h$ and as a result (25) and (46) reduce to equalities

$$\bar{z} = -\frac{1}{\lambda}\bar{p}, \quad \bar{z}_h = -\frac{1}{\lambda}\bar{p}_h, \tag{47}$$

correspondingly, and as a result the continuous and discrete PDECO problems are equivalent to the systems of equations

$$\bar{u} = S(-\frac{1}{\lambda}\bar{p})$$
$$\bar{p} = S^*(\bar{u} - u_d)$$

and

$$\bar{u}_h = S_h(-\frac{1}{\lambda}\bar{p}_h)$$
$$\bar{p}_h = S_h^*(\bar{u}_h - u_d).$$

As a result $\bar{z}, \bar{u}, \bar{p} \in H^2(\Omega)$ and we can expect second-order convergence for the optimal control in L^2 norm. Indeed, one can establish the following result.

Theorem 9 *Let $\bar{z}$ and $\bar{z}_h$ be optimal solutions to continuous and discrete PDECO problems (21) and (43), respectively, without control constraints. Assume in addition that Ω is convex or $C^{1,1}$. Then there exists a constant C independent of h such that*

$$\|\bar{z} - \bar{z}_h\|_{L^2(\Omega)} \leq Ch^2 \left(\|\bar{z}\|_{L^2(\Omega)} + \|u_d\|_{L^2(\Omega)}\right).$$

Proof We begin by recalling the unconstrained optimality conditions $\lambda\bar{z} + \bar{p} = 0 = \lambda\bar{z}_h + \bar{p}_h$, which yields $(\lambda\bar{z} + \bar{p}, \bar{z}_h - \bar{z}) = 0 = (\lambda\bar{z}_h + \bar{p}_h, \bar{z} - \bar{z}_h)$. Adding these last two equalities, we arrive at

$$\lambda\|\bar{z}-\bar{z}_h\|^2_{L^2(\Omega)} = (\bar{p}-\bar{p}_h, \bar{z}_h-\bar{z})_{L^2(\Omega)} = (S^*(S\bar{z}-u_d)-S_h^*(S_h\bar{z}_h-u_d), \bar{z}_h-\bar{z})_{L^2(\Omega)}, \tag{48}$$

where in the last equality we have used the representation of $\bar{p}$ and $\bar{p}_h$. Up on rewriting (48), we obtain

$$\lambda\|\bar{z}-\bar{z}_h\|^2_{L^2(\Omega)} = (S^*S\bar{z}-S_h^*S_h\bar{z}_h, \bar{z}_h-\bar{z})_{L^2(\Omega)}+((S^*-S_h^*)u_d, \bar{z}_h-\bar{z})_{L^2(\Omega)} = I+II. \tag{49}$$

It follows from Corollary 5 that

$$|II| \leq Ch^2\|u_d\|_{L^2(\Omega)}\|\bar{z} - \bar{z}_h\|_{L^2(\Omega)}. \tag{50}$$

It then remains to estimate I in (49). We add and subtract $(S_h^*S_h\bar{z}, \bar{z}_h - \bar{z})_{L^2(\Omega)}$ to I and arrive at

$$\begin{aligned} I &= ((S^*S - S_h^*S_h)\bar{z}, \bar{z}_h - \bar{z})_{L^2(\Omega)} + (S_h(\bar{z} - \bar{z}_h), S_h(\bar{z}_h - \bar{z}))_{L^2(\Omega)} \\ &\leq ((S^*S - S_h^*S_h)\bar{z}, \bar{z}_h - \bar{z})_{L^2(\Omega)}, \end{aligned} \tag{51}$$

where we have used the fact that $(S_h(\bar{z}-\bar{z}_h), S_h(\bar{z}_h-\bar{z}))_{L^2(\Omega)} = -\|S_h(\bar{z}-\bar{z}_h)\|^2 \leq 0$. Again adding and subtracting $(S_h^* S\bar{z}, , \bar{z}_h - \bar{z})_{L^2(\Omega)}$ to (51), we arrive at

$$\begin{aligned} |I| &\leq |((S^* - S_h^*)S\bar{z}, \bar{z}_h - \bar{z})_{L^2(\Omega)} + (S_h^*(S - S_h)\bar{z}, \bar{z}_h - \bar{z})_{L^2(\Omega)}| \\ &\leq Ch^2 \|\bar{z}\|_{L^2(\Omega)} \|\bar{z} - \bar{z}_h\|_{L^2(\Omega)}, \end{aligned} \tag{52}$$

where we have first used the triangle inequality and have estimated the first term using Corollary 5 and continuity of $S : L^2(\Omega) \to L^2(\Omega)$: $|((S^* - S_h^*)S\bar{z}, \bar{z}_h - \bar{z})_{L^2(\Omega)}| \leq Ch^2 \|S\bar{z}\|_{L^2(\Omega)} \|\bar{z} - \bar{z}_h\|_{L^2(\Omega)} \leq Ch^2 \|\bar{z}\|_{L^2(\Omega)} \|\bar{z} - \bar{z}_h\|_{L^2(\Omega)}$. The estimate of the remaining term follows again using Corollary 5 and the continuity of $S_h^* : L^2(\Omega) \to L^2(\Omega)$. Finally, substituting the estimates of I and II from (52) and (50) in (49), we arrive at the asserted result.

8.3 Optimization Problem with Control Constraints

For the rest of this section, we assume constant box constraints, i.e., $z_a, z_b \in \mathbb{R}$, with $z_a < z_b$. We remind the reader that in this situation the optimal control is given by a projection formula:

$$\bar{z}(x) = \mathscr{P}_{Z_{ad}} \left\{ -\frac{1}{\lambda} \bar{p}(x) \right\}. \tag{53}$$

If the constraints are active, $\bar{z} \notin H^2(\Omega)$. However, we can still conclude $\bar{z} \in H^1(\Omega)$ and even $\bar{z} \in W_\infty^1(\Omega)$ by using [58, Theorem A.1]. In light of this, the second-order convergence cannot in general be expected. There are several approaches to treat the problem.

8.3.1 Cell-Wise Constant Control Discretization

One idea is to consider a cellwise constant discretization of the control variable, i.e., we choose $Z_{ad,h} = Z_{ad} \cap Z_h^0$, where Z_h^0 is a space of piecewise constant functions on the partition $\mathscr{T}_h$. This idea goes back to Falk [35]. Since we consider piecewise constant discretization, only first-order convergence for the control can be expected. Indeed for such discretization, one can establish the following convergence result.

Theorem 10 *Let $\bar{z}$ and $\bar{z}_h$ be optimal solutions to continuous and discrete PDECO problems (21) and (43), respectively, with control constraints (23). Let $Z_{ad,h} = Z_{ad} \cap Z_h^0$. Assume in addition that Ω is convex or $C^{1,1}$. Then there exists a constant C independent of h such that*

$$\|\bar{z} - \bar{z}_h\|_{L^2(\Omega)} \leq Ch \left(\|\bar{z}\|_{H^1(\Omega)} + \|u_d\|_{L^2(\Omega)} \right).$$

Proof First we define a projection $\pi_h : Z_{ad} \to Z_{ad} \cap Z_h^0$ by

$$\pi_h v \mid_\tau = \frac{1}{|\tau|} \int_\tau v \, dx, \quad \forall \tau \in \mathscr{T}_h. \tag{54}$$

Thus the projection π_h is the orthogonal projection onto Z_h^0 with respect to L^2-inner-product, i.e.

$$(v - \pi_h v, w)_{L^2(\Omega)} = 0, \quad w \in Z_h^0 \tag{55}$$

and has the following approximation property

$$\|v - \pi_h v\|_{L^2(\Omega)} \leq Ch \|\nabla v\|_{L^2(\Omega)}, \quad v \in H^1(\Omega). \tag{56}$$

Then replacing z by $\bar{z}_h$ in (25) and z_h by $\pi_h \bar{z}$ in (46), we arrive at

$$(\lambda \bar{z} + \bar{p}, \bar{z}_h - \bar{z})_{L^2(\Omega)} \geq 0, \quad (\lambda \bar{z}_h + \bar{p}_h, \pi_h \bar{z} - \bar{z}_h)_{L^2(\Omega)} \geq 0.$$

Adding these inequalities, we obtain that

$$\lambda \|\bar{z} - \bar{z}_h\|^2_{L^2(\Omega)} \leq (\bar{p} - \bar{p}_h, \bar{z}_h - \bar{z})_{L^2(\Omega)} + (\lambda \bar{z}_h + \bar{p}_h, \pi_h \bar{z} - \bar{z})_{L^2(\Omega)} = I + II. \tag{57}$$

The estimate of I is exactly the same as in Theorem 9, i.e.,

$$\begin{aligned} |I| &\leq Ch^2 (\|\bar{z}\|_{L^2(\Omega)} + \|u_d\|_{L^2(\Omega)}) \|\bar{z} - \bar{z}_h\|_{L^2(\Omega)} \\ &\leq C_\lambda h^4 (\|\bar{z}\|_{L^2(\Omega)} + \|u_d\|_{L^2(\Omega)})^2 + \frac{\lambda}{4} \|\bar{z} - \bar{z}_h\|^2_{L^2(\Omega)}, \end{aligned} \tag{58}$$

where we have used the Young's inequality in addition. Next we provide estimate for II. Using the characterization of $\bar{p}_h$, followed by, adding and subtracting $(S_h^*(S_h \bar{z} - u_d), \pi_h \bar{z} - \bar{z})_{L^2(\Omega)}$, and using the continuity of S_h^*, we obtain

$$\begin{aligned} II &= (S_h^*(S_h \bar{z}_h - u_d), \pi_h \bar{z} - \bar{z})_{L^2(\Omega)} \\ &= (S_h^* S_h (\bar{z}_h - \bar{z}), \pi_h \bar{z} - \bar{z})_{L^2(\Omega)} + (S_h^*(S_h \bar{z} - u_d), \pi_h \bar{z} - \bar{z})_{L^2(\Omega)} =: II_1 + II_2. \end{aligned} \tag{59}$$

To estimate II_1, we use (56) and Young's inequality to arrive at

$$|II_1| \leq C_\lambda h^2 \|\bar{z}\|^2_{H^1(\Omega)} + \frac{\lambda}{4} \|\bar{z} - \bar{z}_h\|^2_{L^2(\Omega)}. \tag{60}$$

It then remains to estimate II_2 in (59). By adding and subtracting $(S_h^*(S\bar{z} - u_d), \pi_h\bar{z} - \bar{z})_{L^2(\Omega)}$ in II_2, we obtain that

$$\begin{aligned} |II_2| &= |(S_h^*(S_h - S)\bar{z}, \pi_h\bar{z} - \bar{z})_{L^2(\Omega)} + (S_h^*(S\bar{z} - u_d), \pi_h\bar{z} - \bar{z})_{L^2(\Omega)}| \\ &\leq Ch^3\|\bar{z}\|_{H^1(\Omega)} + |((S_h^* - S^*)(S\bar{z} - u_d), \pi_h\bar{z} - \bar{z})_{L^2(\Omega)}| \\ &\quad + |(S^*(S\bar{z} - u_d), \pi_h\bar{z} - \bar{z})_{L^2(\Omega)}| =: II_{2,1} + II_{2,2} + II_{2,3}, \end{aligned} \tag{61}$$

where we have used Corollary 5 and (56) to estimate the first term. Moreover, we have added and subtracted $(S^*(S\bar{z} - u_d), \pi_h\bar{z} - \bar{z})_{L^2(\Omega)}$ to the second term. It then remains to estimate $II_{2,2}$ and $II_{2,3}$. Again using Corollary 5 and (56), we obtain that $II_{2,2} \leq Ch^3(\|\bar{z}\|_{L^2(\Omega)} + \|u_d\|_{L^2(\Omega)})\|\bar{z}\|_{H^1(\Omega)}$. Finally, to estimate $II_{2,3}$ we first recall that $S^*(S\bar{z} - u_d) = \bar{p}$. Since π_h is L^2-orthogonal projection, we obtain

$$II_{2,3} = (\bar{p} - \pi_h\bar{p}, \pi_h\bar{z} - \bar{z})_{L^2(\Omega)} \leq Ch^3(\|\bar{z}\|_{L^2(\Omega)} + \|u_d\|_{L^2(\Omega)})\|\bar{z}\|_{H^1(\Omega)}. \tag{62}$$

Collecting all the estimates, we arrive at the asserted result.

Comparing this result with the unconstrained case, we have only first-order convergence. This is mainly due to the choice of the discrete control space which does not take the full advantage of the regularity of the optimal control, namely $\bar{z} \in W_\infty^1(\Omega)$. Moreover, away from the active constraints $\bar{z}$ is still in H^2. Taking this in consideration, there are some alternatives to increase the order of the convergence.

8.3.2 Cell-Wise Linear Control Discretization

To improve the convergence rate of the above result, we consider $Z_{ad,h} = Z_{ad} \cap V_h$, i.e., the control space consists of piecewise linear functions satisfying constraints (22). The approximation properties in this setting were investigated in a number of papers, for example [72, 74]. We will not provide all the details, only highlight the main ideas. To take advantage of the regularity for $\bar{z}$ discussed above, we consider the following sets:

$$\mathcal{T}_h^1 = \{\tau \in \mathcal{T}_h \mid \bar{z}\,|_\tau = z_a \text{ or } \bar{z}\,|_\tau = z_b\},$$

which is the set of active cells,

$$\mathcal{T}_h^2 = \{\tau \in \mathcal{T}_h \mid z_a < \bar{z}\,|_\tau < z_b\},$$

the set of not active cells, and the rest

$$\mathcal{T}_h^3 = \mathcal{T}_h \backslash (\mathcal{T}_h^1 \cup \mathcal{T}_h^2).$$

Then under the assumption that

$$meas(\mathscr{T}_h^3) \leq Ch, \tag{63}$$

which is valid, for example, if the boundary of the active set consists of a finite number of rectifiable curves, one can establish the following result.

Theorem 11 *Let $\bar{z}$ and $\bar{z}_h$ be optimal solutions to continuous and discrete PDECO problems (21) and (43), respectively, with control constraints. Assume in addition that Ω is convex or $C^{1,1}$, $u_d \in L^p(\Omega)$ for $p > n$ and assumptions (63) hold. Then there exists a constant C independent of h such that*

$$\|\bar{z} - \bar{z}_h\|_{L^2(\Omega)} \leq Ch^{\frac{3}{2}} \left(\|\bar{p}\|_{H^2(\Omega)} + \|\nabla \bar{z}\|_{L^\infty(\Omega)} \right).$$

Proof The proof of this result can be found in [73].

Remark 7 (Variational Discretization) The idea of the variational discretization approach introduced by Hinze [52] is not to discretize the control variable, i.e., to choose $Z_{ad,h} = Z_{ad}$. This approach does give second-order convergence for the control but requires a nonstandard implementation, especially for $n > 1$.

8.4 Semilinear Equations

Similar to the linear case, we discretize the problem with finite elements. Thus, we replace the state space $H_0^1(\Omega)$ with $V_h \subset H_0^1(\Omega)$ and the control space of admissible functions Z_{ad} with $Z_{ad,h} \subset Z_{ad}$. We additionally assume that $z_a, z_b \in \mathbb{R} \cup \{\pm\infty\}$, with $z_a < z_b$. In case of unconstrained control, we take $z_a = -\infty$ and $z_b = \infty$, i.e., $Z_{ad,h} = V_h$. The discretized problem (30)–(32) becomes

$$\min_{u_h \in V_h,\ z_h \in Z_{ad,h}} J_h(u_h, z_h) = \frac{1}{2}\|u_h - u_d\|_{L^2(\Omega)}^2 + \frac{\lambda}{2}\|z_h\|_{L^2(\Omega)}^2 \tag{64}$$

subject to

$$(A\nabla u_h, \nabla v_h) + (f(\cdot, u_h), v_h) = (z_h, v_h) \quad \forall v_h \in V_h. \tag{65}$$

All of the above strategies for choosing $Z_{ad,h}$ can be applied for the semilinear problem as well and similar error estimates can be obtained. Of course the arguments are more technical and we refer to [13, 22, 29] for details.

Theorem 12 *Assume that Ω is convex domain with $C^{1,1}$ boundary and let Assumptions 6 and 8 be satisfied. Let $\bar{z}$ be a strict local solution to (30)–(32) that fulfills the second-order optimality condition: There exist $\delta > 0$ and $\tau > 0$ such that*

$$\mathscr{J}''(\bar{z})(z, z) \geq \delta \|z\|_{L^2(\Omega)}^2$$

holds for all $z \in L^\infty(\Omega)$ *satisfying*

$$z(x) \begin{cases} \geq 0 & \text{if } \bar{z}(x) = z_a, \\ \leq 0 & \text{if } \bar{z}(x) = z_b, \\ = 0 & \text{if } |\lambda \bar{z} + \bar{p}| \geq \tau > 0. \end{cases} \tag{66}$$

Then:

- *[13, Thm. 5.1] (Cell-wise constant control) Let* $Z_{ad,h} = Z_{ad} \cap Z_h^0$ *and* $\{\bar{z}_h\}$ *be a sequence of locally optimal piecewise constant solutions to* (64)–(65) *that converges strongly in* $L^2(\Omega)$ *to* $\bar{z}$*. Then there exists a constant* C *independent of* h *such that*

$$\|\bar{z} - \bar{z}_h\|_{L^2(\Omega)} \leq Ch.$$

- *[29, Thm. 4.5](Cell-wise linear control) Let* $Z_{ad,h} = Z_{ad} \cap V_h$ *and* $\{\bar{z}_h\}$ *be a sequence of locally optimal piecewise linear solutions to* (64)–(65) *that converges strongly in* $L^2(\Omega)$ *to* $\bar{z}$*. If in addition* (63) *holds, then there exists a constant* C *independent of* h *such that*

$$\|\bar{z} - \bar{z}_h\|_{L^2(\Omega)} \leq Ch^{\frac{3}{2}}.$$

9 Conclusion and the Current State of the Art

In this introductory chapter, we reviewed the main ideas behind the study of PDECO. We briefly mentioned numerical approximation of such problems by the finite-element method and showed some convergence results in the case of control constraints. However, the subject is vast with many active research directions. In this final section, let us mention some topics that we skipped and some active areas of research that we did not touch.

- In the above discussion, we only touched problems with control constraints. However, problems with state constraints are equally important. In contrast to the control constraints, which basically amount to projection onto the feasible set, the state constraints require much more care since the Lagrange multipliers are not functions, only measures. We refer to [81, Chapter 6] for a nice introduction on the subject. For such problems, the error analysis is also more subtle and one often has to reserve to more technical pointwise error estimates to derive optimal-order convergence. We refer to [28, 64, 65] for some recent developments of the subject.
- All above error estimates were a priori error estimates. However, a large fraction of the literature on finite elements is devoted to a posteriori error estimates, i.e., estimates where the error between the true solution and the discrete approximated solution is expressed in terms of computable quantities. We refer to [75, 77] for more recent development of the subject.

- One can consider more complicated state equations or even systems, which can be linear or nonlinear, time dependent, variational inequalities, and so on. Currently, the theory is well developed for problems constrained by linear and semilinear elliptic problems, but the research is very much active for nonlinear, time dependent, and variational inequalities [61, 68].
- We only consider a quadratic cost functional in our error analysis. However, other choices may be desired. For example, it was observed numerically that seeking the control from the space of regular Borel measures forces the sparsity of optimal solution, meaning that the support of solution is small. This phenomenon was analyzed for elliptic and parabolic problems in a number of papers [23–25]; however, there are still some remaining open questions.
- In all our examples, we considered the distributed control, i.e., the control z was acting in the interior of the domain. However, problems where the control acts on the boundary are important in many applications. The control can enter as Dirichlet, Neumann, or Robin boundary conditions. Because of the variational structure of the problems, the Neumann and Robin boundary conditions naturally enter the variational form of the state equation and as a result Neumann boundary controls can be naturally analyzed, see [81, Chapter 2], we refer to [26, 27] for the Robin case. Dirichlet boundary conditions do not have this property and one has to use more sophisticated machinery to overcome technical difficulties and to derive optimal error estimate for the optimal solution [12]. Alternatively, one can also use the penalized Robin boundary conditions to study the Dirichlet boundary control problems [63].

References

1. R.A. Adams and J.J.F. Fournier. *Sobolev spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003.
2. J.-J. Alibert and J.-P. Raymond. Boundary control of semilinear elliptic equations with discontinuous leading coefficients and unbounded controls. *Numer. Funct. Anal. Optim.*, 18(3–4):235–250, 1997.
3. H. Antil, M. Heinkenschloss, and R. H.W. Hoppe. Domain decomposition and balanced truncation model reduction for shape optimization of the Stokes system. *Optimization Methods and Software*, 26(4–5):643–669, 2011.
4. H. Antil, M. Heinkenschloss, R.H.W. Hoppe, C. Linsenmann, and A. Wixforth. Reduced order modeling based shape optimization of surface acoustic wave driven microfluidic biochips. *Math. Comput. Simul.*, 82(10):1986–2003, June 2012.
5. H. Antil, M. Heinkenschloss, R.H.W. Hoppe, and D. C. Sorensen. Domain decomposition and model reduction for the numerical solution of PDE constrained optimization problems with localized optimization variables. *Comput. Vis. Sci.*, 13(6):249–264, 2010.
6. H. Antil, M. Heinkenschloss, and D. C. Sorensen. Application of the discrete empirical interpolation method to reduced order modeling of nonlinear and parametric systems. volume 8 of *Springer MS&A series: Reduced Order Methods for modeling and computational G. Rozza, Eds*. Springer-Verlag Italia, Milano, 2013.

7. H. Antil, M. Hintermüller, R. Nochetto, T. Surowiec, and D. Wegner. Finite horizon model predictive control of electrowetting on dielectric with pinning. *Interfaces Free Bound.*, 19(1): 1–30, 2017.
8. H. Antil, R.H. Nochetto, and P. Venegas. Controlling the Kelvin force: Basic strategies and applications to magnetic drug targeting. *arXiv preprint arXiv:1704.06872*, 2017.
9. H. Antil, R.H. Nochetto, and P. Venegas. Optimizing the Kelvin force in a moving target subdomain. *Math. Models Methods Appl. Sci.*, 28(1):95–130, 2018.
10. H. Antil, J. Pfefferer, and M. Warma. A note on semilinear fractional elliptic equation: analysis and discretization. *Math. Model. Numer. Anal. (ESAIM: M2AN)*, 51:2049–2067, 2017.
11. H. Antil and M. Warma. Optimal control of fractional semilinear PDEs. *arXiv preprint arXiv:1712.04336*, 2017.
12. T. Apel, M. Mateos, J. Pfefferer, and A. Rösch. On the regularity of the solutions of Dirichlet optimal control problems in polygonal domains. *SIAM J. Control Optim.*, 53(6):3620–3641, 2015.
13. N. Arada, E. Casas, and F. Tröltzsch. Error estimates for the numerical approximation of a semilinear elliptic control problem. *Comput. Optim. Appl.*, 23(2):201–229, 2002.
14. H. Attouch, G. Buttazzo, and G. Michaille. *Variational analysis in Sobolev and BV spaces*, volume 17 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, second edition, 2014. Applications to PDEs and optimization.
15. M. Benzi, G.H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numer.*, 14:1–137, 2005.
16. M. Berggren and M. Heinkenschloss. Parallel solution of optimal-control problems by time-domain decomposition. In *Computational science for the 21st century. Symposium*, pages 102–112, 1997.
17. P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk. Convergence rates for greedy algorithms in reduced basis methods. *SIAM J. Math. Anal.*, 43(3):1457–1472, 2011.
18. J.F. Bonnans and A. Shapiro. *Perturbation analysis of optimization problems*. Springer Series in Operations Research. Springer-Verlag, New York, 2000.
19. S.C. Brenner and L.R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008.
20. E. Casas. Control of an elliptic problem with pointwise state constraints. *SIAM J. Control Optim.*, 24(6):1309–1318, 1986.
21. E. Casas. Boundary control of semilinear elliptic equations with pointwise state constraints. *SIAM J. Control Optim.*, 31(4):993–1006, 1993.
22. E. Casas. Using piecewise linear functions in the numerical approximation of semilinear elliptic control problems. *Adv. Comput. Math.*, 26(1–3):137–153, 2007.
23. E. Casas, C. Clason, and K. Kunisch. Approximation of elliptic control problems in measure spaces with sparse solutions. *SIAM J. Control Optim.*, 50(4):1735–1752, 2012.
24. E. Casas, C. Clason, and K. Kunisch. Parabolic control problems in measure spaces with sparse solutions. *SIAM J. Control Optim.*, 51(1):28–63, 2013.
25. E. Casas and K. Kunisch. Optimal control of semilinear elliptic equations in measure spaces. *SIAM J. Control Optim.*, 52(1):339–364, 2014.
26. E. Casas, M. Mateos, and J.-P. Raymond. Penalization of Dirichlet optimal control problems. *ESAIM Control Optim. Calc. Var.*, 15(4):782–809, 2009.
27. E. Casas, M. Mateos, and F. Tröltzsch. Error estimates for the numerical approximation of boundary semilinear elliptic control problems. *Comput. Optim. Appl.*, 31(2):193–219, 2005.
28. E. Casas, M. Mateos, and B. Vexler. New regularity results and improved error estimates for optimal control problems with state constraints. *ESAIM Control Optim. Calc. Var.*, 20(3):803–822, 2014.
29. E. Casas and F. Tröltzsch. A general theorem on error estimates with application to a quasilinear elliptic optimal control problem. *Comput. Optim. Appl.*, 53(1):173–206, 2012.

30. A. Cohen, M. Hoffmann, and M. Reiß. Adaptive wavelet Galerkin methods for linear inverse problems. *SIAM J. Numer. Anal.*, 42(4):1479–1501, 2004.
31. W. Dahmen, A. Kunoth, and K. Urban. A wavelet Galerkin method for the Stokes equations. *Computing*, 56(3):259–301, 1996. International GAMM-Workshop on Multi-level Methods (Meisdorf, 1994).
32. X. Deng, X.-C. Cai, and J. Zou. Two-level space-time domain decomposition methods for three-dimensional unsteady inverse source problems. *J. Sci. Comput.*, 67(3):860–882, 2016.
33. X. Deng and M. Heinkenschloss. A parallel-in-time gradient-type method for discrete time optimal control problems. 2016.
34. A. Ern and J.L. Guermond. *Theory and practice of finite elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.
35. R.S. Falk. Approximation of a class of optimal control problems with order of convergence estimates. *J. Math. Anal. Appl.*, 44:28–47, 1973.
36. P.E. Farrell, D.A. Ham, S.W. Funke, and M.E. Rognes. Automated derivation of the adjoint of high-level transient finite element programs. *SIAM J. Sci. Comput.*, 35(4):C369–C393, 2013.
37. G.B. Folland. *Real analysis*. Pure and Applied Mathematics (New York). John Wiley & Sons, Inc., New York, second edition, 1999. Modern techniques and their applications, A Wiley-Interscience Publication.
38. D. Gilbarg and N.S. Trudinger. *Elliptic partial differential equations of second order*. Classics in Mathematics. Springer-Verlag, Berlin, 2001. Reprint of the 1998 edition.
39. V. Girault and P.-A. Raviart. *Finite element methods for Navier-Stokes equations*, volume 5 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1986. Theory and algorithms.
40. P. Grisvard. *Singularities in boundary value problems*, volume 22 of *Recherches en Mathématiques Appliquées [Research in Applied Mathematics]*. Masson, Paris, 1992.
41. M. Gubisch and S. Volkwein. Proper orthogonal decomposition for linear-quadratic optimal control. In *Model reduction and approximation*, volume 15 of *Comput. Sci. Eng.*, pages 3–63. SIAM, Philadelphia, PA, 2017.
42. M.D. Gunzburger. *Perspectives in flow control and optimization*, volume 5 of *Advances in Design and Control*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2003.
43. J. Haslinger and R. A. E. Mäkinen. *Introduction to shape optimization*, volume 7 of *Advances in Design and Control*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2003. Theory, approximation, and computation.
44. M. Heinkenschloss. Numerical solution of implicitly constrained optimization problems. 2008.
45. J.S. Hesthaven, G. Rozza, and B. Stamm. *Certified reduced basis methods for parametrized partial differential equations*. SpringerBriefs in Mathematics. Springer, Cham; BCAM Basque Center for Applied Mathematics, Bilbao, 2016. BCAM SpringerBriefs.
46. J.S. Hesthaven and T. Warburton. *Nodal discontinuous Galerkin methods*, volume 54 of *Texts in Applied Mathematics*. Springer, New York, 2008. Algorithms, analysis, and applications.
47. M. Hintermüller and M. Hinze. Moreau-Yosida regularization in state-constrained elliptic control problems: error estimates and parameter adjustment. *SIAM J. Numer. Anal.*, 47(3):1666–1683, 2009.
48. M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim.*, 13(3):865–888 (2003), 2002.
49. M. Hintermüller and I. Kopacka. Mathematical programs with complementarity constraints in function space: C- and strong stationarity and a path-following algorithm. *SIAM J. Optim.*, 20(2):868–902, 2009.
50. M. Hintermüller and I. Kopacka. A smooth penalty approach and a nonlinear multigrid algorithm for elliptic MPECs. *Comput. Optim. Appl.*, 50(1):111–145, 2011.
51. M. Hintermüller and M. Ulbrich. A mesh-independence result for semismooth Newton methods. *Math. Program.*, 101(1, Ser. B):151–184, 2004.
52. M. Hinze. A variational discretization concept in control constrained optimization: the linear-quadratic case. *Comput. Optim. Appl.*, 30(1):45–61, 2005.

53. M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, New York, 2009.
54. M. Hinze and S. Volkwein. Proper orthogonal decomposition surrogate models for nonlinear dynamical systems: error estimates and suboptimal control. In *Dimension reduction of large-scale systems*, volume 45 of *Lect. Notes Comput. Sci. Eng.*, pages 261–306. Springer, Berlin, 2005.
55. M. Hinze and S. Volkwein. Error estimates for abstract linear-quadratic optimal control problems using proper orthogonal decomposition. *Comput. Optim. Appl.*, 39(3):319–345, 2008.
56. K. Ito and K. Kunisch. *Lagrange multiplier approach to variational problems and applications*, volume 15 of *Advances in Design and Control*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.
57. C.T. Kelley. *Iterative methods for optimization*, volume 18 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1999.
58. D. Kinderlehrer and G. Stampacchia. *An introduction to variational inequalities and their applications*, volume 31 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000. Reprint of the 1980 original.
59. K. Kohls, A. Rösch, and K.G. Siebert. A posteriori error analysis of optimal control problems with control constraints. *SIAM J. Control Optim.*, 52(3):1832–1861, 2014.
60. D. Kouri, D. Ridzal, and G.von Winckel. Webpage. https://trilinos.org/packages/rol/.
61. G. Leugering, S. Engell, A. Griewank, M. Hinze, R. Rannacher, V. Schulz, M. Ulbrich, and S. Ulbrich, editors. *Constrained optimization and optimal control for partial differential equations*, volume 160 of *International Series of Numerical Mathematics*. Birkhäuser/Springer Basel AG, Basel, 2012.
62. J.-L. Lions. *Optimal control of systems governed by partial differential equations.* Translated from the French by S. K. Mitter. Die Grundlehren der mathematischen Wissenschaften, Band 170. Springer-Verlag, New York-Berlin, 1971.
63. S. May, R. Rannacher, and B. Vexler. Error analysis for a finite element approximation of elliptic Dirichlet boundary control problems. *SIAM J. Control Optim.*, 51(3):2585–2611, 2013.
64. I. Neitzel, J. Pfefferer, and A. Rösch. Finite element discretization of state-constrained elliptic optimal control problems with semilinear state equation. *SIAM J. Control Optim.*, 53(2):874–904, 2015.
65. I. Neitzel and F. Tröltzsch. Numerical analysis of state-constrained optimal control problems for PDEs. In *Constrained optimization and optimal control for partial differential equations*, volume 160 of *Internat. Ser. Numer. Math.*, pages 467–482. Birkhäuser/Springer Basel AG, Basel, 2012.
66. J. Nocedal and S.J. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.
67. R.H. Nochetto, K.G. Siebert, and A. Veeser. Theory of adaptive finite element methods: an introduction. In *Multiscale, nonlinear and adaptive approximation*, pages 409–542. Springer, Berlin, 2009.
68. E. Rocca G. Schimperna J. Sprekels P. Colli, A. Favini, editor. *Solvability, Regularity, and Optimal Control of Boundary Value Problems for PDEs*, volume 22 of *Springer INdAM Series*. Springer, Cham, 2017. In Honour of Prof. Gianni Gilardi.
69. A. Quarteroni, A. Manzoni, and F. Negri. *Reduced basis methods for partial differential equations*, volume 92 of *Unitext*. Springer, Cham, 2016. An introduction, La Matematica per il 3+2.
70. B. Rivière. *Discontinuous Galerkin methods for solving elliptic and parabolic equations*, volume 35 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008. Theory and implementation.
71. S.M. Robinson. Stability theory for systems of inequalities. II. Differentiable nonlinear systems. *SIAM J. Numer. Anal.*, 13(4):497–513, 1976.
72. A. Rösch. Error estimates for linear-quadratic control problems with control constraints. *Optim. Methods Softw.*, 21(1):121–134, 2006.

73. A. Rösch and R. Simon. Linear and discontinuous approximations for optimal control problems. *Numer. Funct. Anal. Optim.*, 26(3):427–448, 2005.
74. A. Rösch and R. Simon. Superconvergence properties for optimal control problems discretized by piecewise linear and discontinuous functions. *Numer. Funct. Anal. Optim.*, 28(3–4): 425–443, 2007.
75. A. Rösch and D. Wachsmuth. A-posteriori error estimates for optimal control problems with state and control constraints. *Numer. Math.*, 120(4):733–762, 2012.
76. A. Schiela and S. Ulbrich. Operator preconditioning for a class of inequality constrained optimal control problems. *SIAM J. Optim.*, 24(1):435–466, 2014.
77. R. Schneider and G. Wachsmuth. A posteriori error estimation for control-constrained, linear-quadratic optimal control problems. *SIAM J. Numer. Anal.*, 54(2):1169–1192, 2016.
78. C. Schwab and C.J. Gittelson. Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs. *Acta Numer.*, 20:291–467, 2011.
79. A. Shapiro. On concepts of directional differentiability. *J. Optim. Theory Appl.*, 66(3):477–487, 1990.
80. J. Sokoł owski and J.-P. Zolésio. *Introduction to shape optimization*, volume 16 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1992. Shape sensitivity analysis.
81. F. Tröltzsch. *Optimal control of partial differential equations*, volume 112 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2010. Theory, methods and applications, Translated from the 2005 German original by Jürgen Sprekels.
82. M. Ulbrich. *Semismooth Newton methods for variational inequalities and constrained optimization problems in function spaces*, volume 11 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, 2011.
83. S. Ulbrich. Preconditioners based on "parareal" time-domain decomposition for time-dependent PDE-constrained optimization. In *Multiple Shooting and Time Domain Decomposition Methods*, pages 203–232. Springer, 2015.
84. A.J. Wathen. Preconditioning. *Acta Numer.*, 24:329–376, 2015.
85. J. Zowe and S. Kurcyusz. Regularity and stability for the mathematical programming problem in Banach spaces. *Appl. Math. Optim.*, 5(1):49–62, 1979.

Optimization of PDEs with Uncertain Inputs

Drew P. Kouri and Alexander Shapiro

Abstract Uncertainty pervades nearly all science and engineering applications including the optimal control and design of systems governed by partial differential equations (PDEs). In many applications, it is critical to determine optimal solutions that are resilient to the inherent uncertainty in unknown boundary conditions, inaccurate coefficients, and unverifiable modeling assumptions. In this tutorial, we develop a general theory for PDE-constrained optimization problems in which inputs or coefficients of the PDE are uncertain. We discuss numerous approaches for incorporating risk preference and conservativeness into the optimization problem formulation, motivated by concrete engineering applications. We conclude with a discussion of nonintrusive solution methods and numerical examples.

1 Introduction

Optimization problems constrained by partial differential equations (PDEs) arise in a number of science and engineering applications as optimal control and design problems. More often than not, the governing physical equations (PDEs) are fraught with uncertainty including uncertain coefficient, unknown boundary and initial conditions, and unverifiable modeling assumptions. When uncertainty exists, it is critical to determine optimal solutions that account for and in some sense are resilient to this uncertainty.

D. P. Kouri (✉)
Center for Computing Research, Sandia National Laboratories, Albuquerque, NM 87185-9999, USA
e-mail: dpkouri@sandia.gov

A. Shapiro
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA
e-mail: ashapiro@isye.gatech.edu

H. Antil et al. (eds.), *Frontiers in PDE-Constrained Optimization*, The IMA Volumes in Mathematics and its Applications 163,
https://doi.org/10.1007/978-1-4939-8636-1_2

Such problems arise for example in the topological design of elastic structures [5, 67, 77, 78]. Recently, topology optimization has gained increased interest due to the emergence of additive manufacturing technologies [56, 109]. There are many uncertainties associated with additively manufactured components such as random grain structures [2, 21], unknown internal forces due to, e.g., residual stresses [52], and potentially variable operating conditions such as external loads. The target then is to design a structure that is, for example, maximally stiff and in some sense reliable given the uncertain material properties and loads. Another common application is the secondary oil recovery phase in petroleum engineering. In this example, an oil company may choose to inject water or other solvents into a reservoir to increase pressure and produce more oil. Of course the subsurface rock properties are unknown but may be estimated from core samples, flow and pressure history [40, 73, 118], or seismic imaging [65, 101, 111]. The optimization problem is to determine the well locations and injection rates that maximize the net present value of the reservoir [3, 10, 102, 119]. However, the optimal rates should be resilient to the inherent uncertainties of the subsurface.

The purpose of this chapter is to review concepts from stochastic programming [25, 55, 75, 90, 108] that play fundamental roles in formulating PDE-constrained optimization problems in a rigorous and physically meaningful (application relevant) manner. In particular, we discuss the basic extension from deterministic PDE-constrained optimization to optimization of PDEs with uncertain inputs by introducing conditions on the deterministic objective function and PDE solution that ensure a well-defined stochastic problem. When the PDE has uncertain inputs, the associated state (PDE solution) becomes a random field. Substituting the random field solution into the objective function results in a random objective function. In order to solve this problem, we must replace the random objective function with a scalar quantity. There are a number of approaches for doing this. In particular, we discuss risk measures [4, 99, 115], probabilistic functions [76, 81, 93, 114], and distributionally robust optimization [15, 107, 121].

In addition to problem formulation, we discuss the challenges associated with the numerical solution of such problems. Many stochastic formulations result in nonsmooth objective functions which motivate new research on rapidly converging nonsmooth optimization algorithms that can exploit structures inherent to PDE-constrained optimization. We present three classical approaches for approximating and solving stochastic optimization problems: stochastic approximation [80, 89, 91], sample average and quadrature approximation [61, 62, 87, 106], and the progressive hedging algorithm [96].

The remainder of this chapter is structured as follows. We first discuss tensor products of Banach spaces. Such spaces play a central role in the functional analytic framework for PDE-constrained optimization under uncertainty. Next, we provide a general problem formulation and, under certain assumptions, show the existence of minimizers as well as first-order necessary optimality conditions. We demonstrate these results on the standard linear-elliptic quadratic control problem. In the following section, we discuss specific problem formulations including risk measures, probabilistic functions, and distributionally robust optimization. We

then introduce three basic numerical methods: stochastic approximation, sample average and quadrature approximation, and the progressive hedging algorithm. We briefly discuss convergence of these methods and conclude with a numerical demonstration.

2 Tensor Product Spaces

Let $(\Omega, \mathscr{F})$ be a measurable space where Ω is the set of possible outcomes and $\mathscr{F}$ is a σ-algebra of events. We denote the expected value of a random variable $X : \Omega \to \mathbb{R}$ with respect to a probability measure $P : \mathscr{F} \to [0, 1]$ defined on the measurable space $(\Omega, \mathscr{F})$ by

$$\mathbb{E}_P[X] = \int_\Omega X(\omega)\, \mathrm{d}P(\omega).$$

We denote the usual Lebesgue space of $r \in [1, \infty)$ integrable real-valued functions (defined up to a set of P-measure zero) by

$$L^r(\Omega, \mathscr{F}, P) := \left\{\theta : \Omega \to \mathbb{R} : \theta \text{ is } \mathscr{F}\text{-measurable}, \mathbb{E}_P[|\theta|^r] < \infty\right\}.$$

If $r = \infty$, then

$$L^\infty(\Omega, \mathscr{F}, P) := \left\{\theta : \Omega \to \mathbb{R} : \theta \text{ is } \mathscr{F}\text{-measurable}, \operatorname*{ess\,sup}_{\omega\in\Omega} |\theta(\omega)| < \infty\right\}.$$

The Lebesgue spaces defined on $(\Omega, \mathscr{F}, P)$ are Banach spaces and serve as natural spaces for real-valued random variables (i.e., $\mathscr{F}$-measurable functions). In the context of PDE-constrained optimization with uncertain inputs, the PDE solutions will be Sobolev space-valued random elements, which motivate the use of tensor-product vector spaces. Given any real Banach space V, the tensor-product vector space associated with $L^r(\Omega, \mathscr{F}, P)$ and V is

$$L^r(\Omega, \mathscr{F}, P) \otimes V := \operatorname{span}\left\{\theta v \,:\, \theta \in L^r(\Omega, \mathscr{F}, P),\ v \in V\right\},$$

i.e., the linear span of all products of elements of $L^r(\Omega, \mathscr{F}, P)$ and V. In general, there are many norms associated with $L^r(\Omega, \mathscr{F}, P) \otimes V$, including the natural projective and injective norms (cf. [35] and [100]). In this work, we restrict our attention to the so-called Bochner norms

$$\begin{cases} \|u\|_{L^r(\Omega,\mathscr{F},P)\otimes V} = \mathbb{E}_P[\|u\|_V^r]^{\frac{1}{r}} & \text{if } 1 \le r < \infty, \\ \|u\|_{L^\infty(\Omega,\mathscr{F},P)\otimes V} = \operatorname*{ess\,sup}_{\omega\in\Omega} \|u(\omega)\|_V & \text{if } r = \infty. \end{cases}$$

The space $L^r(\Omega, \mathscr{F}, P) \otimes V$ endowed with the corresponding Bochner norm is not complete and hence is not a Banach space. However, the completion of

$L^r(\Omega, \mathscr{F}, P) \otimes V$ with respect to its Bochner norm is isomorphic to the Bochner space

$$L^r(\Omega, \mathscr{F}, P; V) := \left\{u : \Omega \to V : u \text{ is strongly } \mathscr{F}\text{-measurable}, \mathbb{E}_P[\|u\|_V^r] < \infty\right\}$$

if $r \in [1, \infty)$ and

$$L^\infty(\Omega, \mathscr{F}, P; V) := \left\{u : \Omega \to V : u \text{ is strongly } \mathscr{F}\text{-measurable}, \operatorname*{ess\,sup}_{\omega \in \Omega} \|u(\omega)\|_V < \infty\right\}$$

if $r = \infty$ (again functions in $L^r(\Omega, \mathscr{F}, P; V)$ are defined up to a set of measure zero) [35, Sect. 7.1]. Here, a function $u : \Omega \to V$ is strongly $\mathscr{F}$-measurable if there exists a sequence of V-valued simple (piecewise constant, countably-valued) functions defined on sets in $\mathscr{F}$ that converges to u P-almost everywhere (P-a.e.) [53, Def. 3.5.4].

It is worth pointing out that the tensor-product vector space $L^r(\Omega, \mathscr{F}, P) \otimes V$ consists of functions

$$u = \sum_{i=1}^{N} \theta_i v_i, \quad \theta_i \in L^r(\Omega, \mathscr{F}, P), \quad v_i \in V, \quad i = 1, \ldots, N$$

for some $N \in \mathbb{N}$, and thus provides a natural approximation space for functions in $L^r(\Omega, \mathscr{F}, P; V)$. This fact is exploited by many uncertainty quantification methods. In particular, polynomial chaos [58, 122], stochastic Galerkin [8, 9], tensor decomposition, [47, 59] and other projection-based methods for approximating PDEs with uncertain inputs decompose the PDE solution into sums of random and spatial components. These two components are then approximated separately using, e.g., polynomial approximation in $L^r(\Omega, \mathscr{F}, P)$ and finite elements in V.

3 Problem Formulation

In this section, we provide the general formulation of our optimization problem. Let U and Z be real reflexive Banach spaces, and let Y be a real Banach space. Here U denotes the deterministic state space, Z denotes the space of optimization variables (i.e., controls, designs, etc.), and Y denotes the PDE residual space. The optimization variables $z \in Z$ are always deterministic and represent a control or design that must be implemented prior to observing the randomness in the system. Stochastic controls do however arise in time-dependent decision processes and multistage stochastic programs in which case the concept of time consistency plays a central role. Time consistency is based on the famous quotation of Bellman: *"An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the*

state resulting from the first decision" [13]. In this review, we restrict our attention to optimization problems constrained by steady (i.e., time-independent, stationary) PDEs. For a more detailed discussion of dynamic stochastic programs (without PDEs) and time consistency, we direct the interested reader to [108, Ch. 6.8].

Before describing the optimization problem, we assume that the uncertainty in the PDE constraint is represented by a finite random vector $\xi : \Omega \to \Xi$ where $\Xi := \xi(\Omega) \subseteq \mathbb{R}^m$ with $m \in \mathbb{N}$ (i.e., ξ is a $\mathscr{F}$-measurable vector-valued function). In the literature, this is called the *finite-dimensional noise assumption* [7, 83] and facilitates numerical approximations such as polynomial chaos and stochastic collocation [7, 9, 58, 83]. Such a finite-dimensional representation is often achieved using a truncated Karhunen–Loève expansion [57, 69]. More importantly, this assumption permits a change of variables in which the PDE and objective function depend only on the "deterministic" parameters $\xi \in \Xi$. This change of variables transforms our original uncertainty model defined on the probability space $(\Omega, \mathscr{F}, P)$ to a model defined on the probability space $(\Xi, \mathscr{B}, \mathbb{P})$ where $\mathscr{B} \subseteq 2^{\Xi}$ is the σ-algebra generated by the sets $\xi^{-1}(A)$ for $A \in \mathscr{F}$ and $\mathbb{P} := P \circ \xi^{-1}$ is the probability law of ξ. In this new setting, we define the Bochner and Lebesgue spaces analogously to the definitions in Section 2. Throughout, we will abuse notation and let ξ denote the random variable $\xi(\omega)$ as well as its realizations. Recently, researchers in uncertainty quantification have developed and analyzed methods for handling infinite-dimensional uncertainties, e.g., $\xi(\omega)$ is a sequence of real numbers for each $\omega \in \Omega$. For example, see [31]. Since all practical computational methods for solving PDEs with uncertain inputs and their corresponding optimization problems require a finite (i.e., computer) representation of the uncertainty, we restrict our attention to the finite-dimensional noise setting. Finally, it is worth noting that no result in this section requires the finite-dimensional noise assumption. However, we work under this assumption to simplify the presentation in the forthcoming sections.

Now, let $Z_{\text{ad}} \subseteq Z$ be a closed convex subset of optimization variables, let $e : U \times Z_{\text{ad}} \times \Xi \to Y$ denote, e.g., a PDE in weak form, and consider the equality constraint

$$e(u, z, \xi(\omega)) = 0. \tag{1}$$

The goal of this article is to understand and solve general stochastic optimization problems with the form

$$\min_{z \in Z_{\text{ad}}} \left\{ \mathfrak{J}(z) := \mathscr{R}(J(S(z;\xi),\xi)) + \wp(z) \right\} \tag{2}$$

where $\mathscr{R}$ is a functional that maps random variables on $(\Xi, \mathscr{B})$ into the real numbers, $J : U \times \Xi \to \mathbb{R}$ is the *uncertain objective function*, $\wp : Z \to \mathbb{R}$ is a control penalty, and $S(z;\cdot) : \Xi \to U$ satisfies $e(S(z;\xi), z, \xi) = 0$ for $\mathbb{P}$-almost every $\xi \in \Xi$ (or equivalently $e(S(z;\xi(\omega)), z, \xi(\omega)) = 0$ for P-almost every $\omega \in \Omega$). Throughout, we denote the *reduced uncertain objective function* by

$$\mathscr{J}(z) := J(S(z;\xi),\xi). \tag{3}$$

Note that $\mathscr{J}(z)$ is also a function of ξ and hence is viewed as a random variable mapping Z_{ad} into a space of real-valued random variables on $(\Xi, \mathscr{B})$.

To ensure the PDE constraint $e(u,z,\xi) = 0$ is well posed, we require that it is uniquely solvable and the solution is in $L^q(\Xi,\mathscr{B},\mathbb{P};U)$ for some $q \in [1,\infty]$. We make this statement rigorous in the following assumption.

Assumption 1 (Properties of the Solution Map) *For each $z \in Z_{\text{ad}}$, there exists a unique mapping $S(z;\cdot) : \Xi \to U$ that solves $e(S(z;\xi),z,\xi) = 0$ for $\mathbb{P}$-almost all $\xi \in \Xi$ and satisfies the following properties:*

1. **Measurability:** *$S(z;\cdot) : \Xi \to U$ is strongly $\mathscr{B}$-measurable for all $z \in Z_{\text{ad}}$.*
2. **Growth Condition:** *There exists $q \in [1,\infty]$, a nonnegative random variable $C \in L^q(\Xi,\mathscr{B},\mathbb{P})$, and a nonnegative increasing function $\varrho : [0,\infty) \to [0,\infty)$ such that*

$$\|S(z;\xi)\|_U \leq C(\xi)\varrho(\|z\|_Z)$$

   for $\mathbb{P}$-almost all $\xi \in \Xi$ and for all $z \in Z_{\text{ad}}$.
3. **Continuity:** *S satisfies the continuity property*

$$z_n \rightharpoonup z \quad \text{in } Z_{\text{ad}} \quad \Longrightarrow \quad S(z_n;\cdot) \rightharpoonup S(z;\cdot) \quad \text{in } U, \mathbb{P}\text{-a.e.}$$

Assumptions 1.1–2 ensure that $S : Z_{\text{ad}} \to L^q(\Xi,\mathscr{B},\mathbb{P};U)$. Additionally, Assumption 1 combined with the Lebesgue Dominated Convergence Theorem ensure S is weakly continuous from Z into $L^q(\Xi,\mathscr{B},\mathbb{P};U)$ [63, Sect. 2.2]. We similarly assume there exists $p \in [1,\infty]$ such that the reduced uncertain objective function satisfies

$$\mathscr{J}(z) \in L^p(\Xi,\mathscr{B},\mathbb{P}) \quad \forall\, z \in Z_{\text{ad}}.$$

To simplify notation, we denote the realization of $\mathscr{J}(z)$ at ξ, i.e., $[\mathscr{J}(z)](\xi)$, by $\mathscr{J}(z,\xi)$. For example, the authors in [63] postulate the following assumptions on the uncertain objective function.

Assumption 2 (Properties of the Objective Function) *There exists $1 \leq p < \infty$ such that the function $J : U \times \Xi \to \mathbb{R}$ satisfies:*

1. ***Carathéodory:*** *J is a Carathéodory function, i.e., $J(\cdot,\xi)$ is continuous for $\mathbb{P}$-almost every $\xi \in \Xi$ and $J(u,\cdot)$ is $\mathscr{B}$-measurable for all $u \in U$.*
2. ***Growth Condition:*** *If $q < \infty$, then there exists $a \in L^p(\Xi,\mathscr{B},\mathbb{P})$ with $a \geq 0$ $\mathbb{P}$-a.e. and $c > 0$ such that*

$$|J(u,\xi)| \leq a(\xi) + c\|u\|_U^{q/p} \quad \forall\, u \in U \text{ and } \mathbb{P}\text{-almost all } \xi \in \Xi$$

If $q = \infty$, then for all $c > 0$ there exists $\gamma_c \in L^p(\Xi, \mathscr{B}, \mathbb{P})$ such that

$$|J(u, \xi)| \leq \gamma_c(\xi) \quad \mathbb{P}\text{-a.e. } \xi \quad \forall\, u \in U,\ \|u\|_U \leq c.$$

3. ***Convexity:*** *$J(\cdot, \xi)$ is convex for $\mathbb{P}$-almost every $\xi \in \Xi$.*

Assumptions 2.1–2 combined with Krasnosel'skii's Theorem [116, Thm. 19.1] ensure that the uncertain objective function $u \mapsto J(u, \cdot)$ is continuous from $L^q(\Xi, \mathscr{B}, \mathbb{P}; U)$ into $L^p(\Xi, \mathscr{B}, \mathbb{P})$.

3.1 Existence of Minimizers and Optimality Conditions

In this section, we present one set of assumptions on $\mathscr{R}$ that ensure the existence of minimizers of (2). In addition, when a minimizer of (2) exists, we characterize the first-order necessary optimality conditions that it satisfies.

Theorem 1 *Let Assumptions 1 and 2 hold, and define $\mathscr{X} := L^p(\Xi, \mathscr{B}, \mathbb{P})$ where $p \in [1, \infty)$ is defined in Assumption 2. Moreover, suppose that $\wp : Z \to \mathbb{R}$ is weakly lower semicontinuous and $\mathscr{R} : \mathscr{X} \to \mathbb{R}$ is convex, and satisfies the monotonicity property: for any $X,\ X' \in \mathscr{X}$,*

$$X \leq X' \quad \mathbb{P}\text{-a.e.} \qquad \Longrightarrow \qquad \mathscr{R}(X) \leq \mathscr{R}(X'). \tag{4}$$

Finally, assume that the level set $\{z \in Z_{\mathrm{ad}} : \mathfrak{J}(z) \leq \gamma\}$ is nonempty and bounded for some $\gamma \in \mathbb{R}$. Then problem (2) has an optimal solution, i.e., there exists $z_\star \in Z_{\mathrm{ad}}$ such that $\mathfrak{J}(z_\star) \leq \mathfrak{J}(z)$ for all $z \in Z_{\mathrm{ad}}$.

Proof Since $\mathscr{R}$ is finite, convex, and satisfies (4), it is continuous and subdifferentiable [108, Prop. 6.6]. The Fenchel–Young inequality then ensures that

$$\mathscr{R}(\mathscr{J}(z)) \geq \mathbb{E}[\theta\, \mathscr{J}(z)] - \mathscr{R}^*(\theta) \quad \forall\, z \in Z_{\mathrm{ad}},\ \theta \in \operatorname{dom} \mathscr{R}^* \tag{5}$$

where

$$\mathscr{R}^*(\theta) = \sup_{X \in \mathscr{X}} \{\mathbb{E}[\theta X] - \mathscr{R}(X)\}$$

is the Legendre–Fenchel transformation of $\mathscr{R}$ and

$$\operatorname{dom} \mathscr{R}^* := \{\theta \in \mathscr{X}^* : \mathscr{R}^*(\theta) < \infty\}$$

is the effective domain of $\mathscr{R}^*$. Equality in (5) holds if and only if $\theta \in \partial \mathscr{R}(\mathscr{J}(z))$ [6, Prop. 9.5.1]. Now, owing to (4), $\theta \in \operatorname{dom} \mathscr{R}^*$ satisfies $\theta \geq 0$ $\mathbb{P}$-a.e. [108, Thm. 9.3.5]. Therefore, Assumption 2 and Krasnosel'skii's Theorem ensure that $u \mapsto J(u, \cdot)$ is continuous and hence $u \mapsto \mathbb{E}[\theta J(u, \cdot)]$ is convex and continuous.

Therefore, $u \mapsto \mathbb{E}[\theta J(u,\cdot)]$ is weakly lower semincontinuous [26, Thm. 2.23], which when combined with the weak continuity of $z \mapsto S(z;\cdot)$ ensures that $z \mapsto \mathbb{E}[\theta \mathscr{J}(z)]$ is weakly lower semicontinuous. Thus, for any sequence $\{z_n\} \subset Z_{\text{ad}}$ that weakly converges to $z \in Z_{\text{ad}}$, we have that

$$\begin{aligned}\liminf_{n\to\infty} \mathscr{R}(\mathscr{J}(z_n)) &\geq \liminf_{n\to\infty} \mathbb{E}[\theta \mathscr{J}(z_n)] - \mathscr{R}^*(\theta) \\ &\geq \mathbb{E}[\theta \mathscr{J}(z)] - \mathscr{R}^*(\theta) = \mathscr{R}(\mathscr{J}(z)) \quad \forall\, \theta \in \partial\mathscr{R}(\mathscr{J}(z)),\end{aligned}$$

which implies that $z \mapsto \mathscr{R}(\mathscr{J}(z))$ is weakly lower semicontinuous. Since $\wp$ is also weakly lower semicontinuous, $\mathfrak{J}$ is as well. Moreover, the minimization is performed over a bounded weakly closed level set in the reflexive Banach space Z, which implies the level set is weakly compact. Under these conditions, the direct method of the calculus of variations [6, Thm. 3.2.1] applies and ensures the existence of a minimizer. □

Since minimizers exist, it is natural to ask what the first-order necessary optimality conditions are. The following theorem characterizes the optimality conditions when J, $\wp$, and S are differentiable. For this result, we denote the space of bounded linear operators from a Banach space A to a Banach space B by $\mathscr{L}(A,B)$. Moreover, by $T_{Z_{\text{ad}}}(z)$ and $N_{Z_{\text{ad}}}(z)$, we denote the tangent and normal cones, respectively, to the (convex) set Z_{ad} at $z \in Z_{\text{ad}}$. We say that a function $f : Z \to \mathbb{R}$ is continuously differentiable if it possesses a derivative $f'(\cdot)$ in the sense of Gâteaux and $f'(\cdot)$ is continuous. It follows then by the mean value theorem that f is differentiable in the sense of Fréchet, e.g., [26, pp. 35–36]. It is said that f is (Gâteaux) directionally differentiable at $z \in Z$ if the directional derivative $f'(z,h) := \lim_{t\downarrow 0}[f(z+th) - f(z)]/t$ exists for all $h \in Z$. Note that if f is convex and continuous, then it is locally Lipschitz [30, Prop. 2.2.7] and $f'(z,\cdot)$ is a Hadamard directional derivative [105, Prop. 3.5].

Theorem 2 *Let the assumptions of Theorem 1 hold. In addition, suppose there exists an open set $V \subseteq Z$ with $Z_{\text{ad}} \subseteq V$ such that $z \mapsto S(z;\cdot) : V \to L^q(\Xi,\mathscr{B},\mathbb{P};U)$ is continuously differentiable with derivative*

$$S'(z;\cdot) \in \mathscr{L}(Z, L^q(\Xi,\mathscr{B},\mathbb{P};U)),$$

$u \mapsto J(u,\cdot) : L^q(\Xi,\mathscr{B},\mathbb{P};U) \to L^p(\Xi,\mathscr{B},\mathbb{P})$ is continuously differentiable with derivative

$$J'(u,\cdot) \in \mathscr{L}(L^q(\Xi,\mathscr{B},\mathbb{P};U), L^p(\Xi,\mathscr{B},\mathbb{P})),$$

and $\wp : Z \to \mathbb{R}$ is continuously differentiable with derivative $\wp'(z) \in Z^*$. Then if $z_\star \in Z_{\text{ad}}$ is a minimizer of $\mathfrak{J}$ over Z_{ad}, the following first-order optimality conditions hold: $\exists\, \theta \in \partial\mathscr{R}(\mathscr{J}(z_\star))$ such that

$$\left\langle \mathbb{E}\left[\theta\, S'(z_\star;\cdot)^* J'(S(z_\star;\cdot),\cdot)\right] + \wp'(z_\star),\, h\right\rangle_{Z^*,Z} \geq 0, \quad \forall\, h \in T_{Z_{\text{ad}}}(z_\star). \tag{6}$$

Proof Let us note that if $z_\star$ is an optimal solution of problem (2), then necessarily the directional derivatives $\mathfrak{J}'(z_\star, h) \geq 0$ for all $h \in T_{Z_{\text{ad}}}(z_\star)$. Since $\wp$ is differentiable, it follows that $\wp'(z_\star, h) = \langle \wp'(z_\star), h \rangle_{Z^*,Z}$. Also under the stated assumptions, $\mathscr{J}$ is continuously differentiable with derivative

$$\mathscr{J}'(z) = J'(S(z;\cdot),\cdot)S'(z;\cdot) \in \mathscr{L}(Z, L^p(\Xi, \mathscr{B}, \mathbb{P})).$$

Now since $\mathscr{R}$ is continuous, it is subdifferentiable and its (Hadamard) directional derivatives are given by

$$\mathscr{R}'(\mathscr{J}(z_\star), H) = \sup_{\theta \in \partial \mathscr{R}(\mathscr{J}(z_\star))} \mathbb{E}[\theta H] \quad \forall H \in \mathscr{X},$$

cf. [108, Thm. 6.10]. By the chain rule for directional derivatives, it follows that

$$\mathfrak{J}'(z_\star, h) = \sup_{\theta \in \partial \mathscr{R}(\mathscr{J}(z_\star))} \left\langle \mathbb{E}[\theta S'(z_\star;\cdot)^* J'(S(z_\star;\cdot),\cdot)] + \wp'(z_\star),\ h \right\rangle_{Z^*,Z}. \tag{7}$$

The function $\phi(\cdot) := \mathfrak{J}'(z_\star, \cdot)$ is convex and positively homogeneous. Moreover, the condition that $\phi(h) \geq 0$ for all $h \in T_{Z_{\text{ad}}}(z_\star)$ means that $h = 0$ is a minimizer of $\phi(h)$ subject to $h \in T_{Z_{\text{ad}}}(z_\star)$. This in turn means that $0 \in \partial\phi(0) + N_{Z_{\text{ad}}}(z_\star)$, which by (7) is equivalent to condition (6). □

Under appropriate differentiability assumptions on the PDE constraint function e, one can show that $\Lambda_\star = S'(z_\star;\cdot)^* J'(S(z_\star;\cdot),\cdot)$ is related to the solution to the adjoint equation. Informally, if the assumptions of the Implicit Function Theorem hold, then $\Lambda_\star = e_z(S(z_\star;\xi), z_\star, \xi)^* \lambda_\star$ where $\lambda_\star$ solves the adjoint equation

$$e_u(S(z_\star;\xi), z_\star, \xi)^* \lambda_\star(\xi) = -J_u(S(z_\star;\xi),\xi)$$

for $\mathbb{P}$-almost all $\xi \in \Xi$. See [61–64] for PDE-constrained optimization examples for which this holds.

3.2 Linear Elliptic Optimal Control

For this example, we assume Ξ is an m-fold Cartesian product of compact intervals and $\mathbb{P}$ is absolutely continuous with respect to the m-dimensional Lebesgue measure. Let $D \subset \mathbb{R}^d$ with $d \in \mathbb{N}$ be an open bounded Lipschitz domain, and define $U = H_0^1(D)$, $Y = U^* = H^{-1}(D)$, and $Z = L^2(D)$. Given the continuous matrix-valued function $A : \Xi \to \mathbb{R}^{d\times d}$ with $A(\xi) = A(\xi)^\top$ for all $\xi \in \Xi$, we define the parametrized linear elliptic PDE as the variational problem: find $u : \Xi \to U$ that solves

$$\langle e(u, z, \xi), v\rangle_{U^*,U} := \int_D (A(\xi)\nabla u(\xi, x)) \cdot \nabla v(x)\,\mathrm{d}x - \int_D z(x)v(x)\,\mathrm{d}x = 0 \tag{8}$$

for all $v \in U$ and fixed $z \in Z$. If there exist constants $0 < \underline{c} \leq \overline{c} < \infty$ such that

$$\underline{c} \leq \frac{x^\top A(\xi)x}{x^\top x} \leq \overline{c} \quad \forall\, x \in \mathbb{R}^d \setminus \{0\},\ \xi \in \Xi \tag{9}$$

then the Lax–Milgram Lemma [28] ensures the existence of a unique solution $S(z;\xi)$ to (8) for each $z \in Z$ and all $\xi \in \Xi$. Additionally, (9) and Poincaré's inequality guarantee the existence of a positive constant $C = C(D, \underline{c})$ such that

$$\|S(z;\cdot)\|_U \leq C\|z\|_Z \quad \forall \xi \in \Xi.$$

This and the linearity of the PDE then imply that $S(\cdot;\xi)$ is a bounded linear operator for all $\xi \in \Xi$ and since Z is compactly embedded into Y [1], $S(\cdot;\xi)$ is completely continuous for all $\xi \in \Xi$. Recall that an operator W mapping a Banach space X into another Banach space Y is completely continuous if

$$x_k \rightharpoonup x \ \text{ in } X \quad \Longrightarrow \quad W(x_k) \to W(x) \ \text{ in } Y.$$

In particular, all compact operators are completely continuous [33, Prop. 3.3]. Finally, $S(z;\cdot)$ is continuous and hence strongly $\mathscr{B}$-measurable since $A(\cdot)$ is continuous. Therefore, Assumption 1 is satisfied and since C is independent of $\xi \in \Xi$, we have that $S(z;\cdot) \in L^\infty(\Xi, \mathscr{B}, \mathbb{P}; U)$ for all $z \in Z$.

Now, let $\beta > 0$ and $u_d \in L^2(D)$ be a desired profile. We consider the PDE-constrained optimization problem

$$\min_{z \in Z} \mathscr{R}\left(\frac{1}{2}\|S(z;\xi) - u_d\|^2_{L^2(D)}\right) + \frac{\beta}{2}\|z\|^2_{L^2(D)} \tag{10}$$

where $S(z;\xi)$ solves (8) for fixed $\xi \in \Xi$ and $z \in Z$. The uncertain objective function and control penalty are

$$J(u,\xi) = \frac{1}{2}\|u - u_d\|^2_{L^2(D)} \quad \text{and} \quad \wp(z) = \frac{\beta}{2}\|z\|^2_{L^2(D)}.$$

J clearly satisfies Assumption 2 and therefore is continuous from $L^q(\Xi, \mathscr{B}, \mathbb{P}; U)$ into $L^p(\Xi, \mathscr{B}, \mathbb{P})$ for any $q \geq 2$ and $p \leq q/2$. Hence, Theorem 1 holds for any $\mathscr{R} : L^p(\Xi, \mathscr{B}, \mathbb{P}) \to \mathbb{R}$ that is convex and satisfies the monotonicity property (4).

In addition, since $e(\cdot,\cdot,\xi)$ is continuous and linear in u and z for all $\xi \in \Xi$, it is continuously Fréchet differentiable in u and z for all $\xi \in \Xi$, and again by the Lax–Milgram Lemma the state Jacobian is boundedly invertible for all $\xi \in \Xi$. Furthermore, the control Jacobian is a bounded linear operator for all $u \in U$, $z \in Z$ and $\xi \in \Xi$. In fact, $e_z(u,z,\xi)$ is independent of u, z, and ξ. Therefore, $S(\cdot;\xi)$ is continuously Fréchet differentiable for all $\xi \in \Xi$ and the derivative satisfies: For any $h \in Z$, $d = S'(z;\cdot)h : \Xi \to U$ solves the sensitivity equation

$$\int_D (A(\xi)\nabla d(\xi,x)) \cdot \nabla v(x)\,\mathrm{d}x - \int_D h(x)v(x)\,\mathrm{d}x = 0 \quad \forall\, v \in U$$

Since the sensitivity equation is identical to (8), we have that $d = S'(z;\cdot)h = S(h;\cdot) \in L^\infty(\Xi, \mathscr{B}, \mathbb{P}; U)$ for all $h \in Z$. Returning to the objective function, J and $\wp$ are clearly continuously Fréchet differentiable and thus Theorem 2 holds for any $\mathscr{R}$ satisfying the stated assumptions. Moreover, the adjoint equation corresponding to (10), at fixed $z \in Z$, is: find $\lambda : \Xi \to U$ such that

$$\int_D (A(\xi)\nabla\lambda(\xi,x)) \cdot \nabla v(x)\,\mathrm{d}x = -\int_D (S(z;\xi)(x) - u_d(x))v(x)\,\mathrm{d}x \quad \forall\, v \in U.$$

Note again that the above analysis ensures $\lambda \in L^\infty(\Xi, \mathscr{B}, \mathbb{P}; U)$.

4 Choosing the Functional $\mathscr{R}$

Under the assumptions of Section 3 (or similar assumptions), the stochastic PDE-constrained optimization problem

$$\min_{z \in Z_{\mathrm{ad}}} \mathscr{R}\left(\mathscr{J}(z)\right) + \wp(z) \tag{11}$$

where $\mathscr{R} : L^p(\Xi, \mathscr{B}, \mathbb{P}) \to \mathbb{R}$ is well-defined, but ambiguous since $\mathscr{R}$ is not explicitly specified. In traditional stochastic programming, $\mathscr{R}$ is taken to be the expected value, i.e., $\mathscr{R} = \mathbb{E}_{\mathbb{P}}$. This results in a *risk neutral* formulation of (11) for which the optimal solutions minimize $\mathscr{J}(z)$ on average. The risk neutral formulation is often not conservative enough for high-consequence applications because the average behavior of a system does not provide a sufficient proxy for variability or low probability and tail events. This motivates the use of *risk measures*. Another popular class of cost surrogates are the *probabilistic functions*. This class seeks to minimize the probability of undesirable events occurring. The use of the expectation, risk measures, and probabilistic functions is justified when the probability law $\mathbb{P}$ is known but can lead to nonsensical, even dangerous, results if $\mathbb{P}$ is unknown and estimated from noisy or incomplete data. In the subsequent sections, we will review both cases of known and unknown probability law. When the probability law is known, we simplify notation and denote $\mathbb{E} = \mathbb{E}_{\mathbb{P}}$.

It is worth mentioning that (11) is only one of many meaningful problem formulations for PDE-constrained optimization. In many applications, constraints in addition to the objective function are uncertain. In this case, we must handle the uncertainty in the constraints in a rigorous and physically relevant way. Popular approaches in stochastic programming include: chance (probabilistic) constraints (see, e.g., [81]) and stochastic dominance constraints (see, e.g., [38]). Chance constraints seek to ensure that the probability of an uncertain quantity of interest exceeding a prescribed threshold is below some nominal value (e.g., the probability that a bridge collapses is smaller than 10^{-3} percent). Stochastic dominance constraints, on the other hand, aim to ensure that our uncertain quantity of interest is in

some sense preferred over a predefined uncertain benchmark value. Since a rigorous treatment of these concepts in PDE-constrained optimization is still an open area of research, we restrict our attention to problems of the type (11). We do, however, introduce and discuss the notions of stochastic orders and stochastic dominance in the coming subsection.

4.1 Risk-Averse Optimization

When the probability law of the random vector ξ is known, we can use any of the multitudes of risk measures to complete the problem definition in (11). A particularly important class of risk measures is the class of *coherent* risk measures [4]. To simplify notation, we denote $\mathscr{X} := L^p(\Xi, \mathscr{B}, \mathbb{P})$. A function $\mathscr{R} : \mathscr{X} \to \mathbb{R}$ is a coherent risk measure if it satisfies:

(R1) **Subadditivity:** For all $X,\ X' \in \mathscr{X}$, $\mathscr{R}(X + X') \leq \mathscr{R}(X) + \mathscr{R}(X')$;
(R2) **Monotonicity:** If $X,\ X' \in \mathscr{X}$ satisfy $X \leq X'$ $\mathbb{P}$-a.e., then $\mathscr{R}(X) \leq \mathscr{R}(X')$;
(R3) **Translation Equivariance:** For all $X \in \mathscr{X}$ and $t \in \mathbb{R}$, $\mathscr{R}(X+t) = \mathscr{R}(X)+t$;
(R4) **Positive Homogeneity:** For all $X \in \mathscr{X}$ and $t \geq 0$, $\mathscr{R}(tX) = t\mathscr{R}(X)$.

Note that axiom (R1) and (R4) imply convexity of $\mathscr{R}$ and convexity plus (R4) imply subadditivity of $\mathscr{R}$. Therefore, axiom (R1) is typically replaced by

(R1′) **Convexity:** For all $X,\ X' \in \mathscr{X}$ and $t \in [0, 1]$

$$\mathscr{R}(tX + (1-t)X') \leq t\mathscr{R}(X) + (1-t)\mathscr{R}(X').$$

In the context of physical applications, $\mathscr{R}(X)$ should inherit the units of X. In which case, (R4) ensures that a change of the units of X results in a consistent change of the units of $\mathscr{R}(X)$. Additionally, (R3) ensures that deterministic quantities, such as the control penalty $\wp$ in (11), do not contribute to the overall risk. In fact, (R3) combined with (R4) ensure that deterministic quantities are riskless, i.e., $\mathscr{R}(t) = t$ for all $t \in \mathbb{R}$.

The axioms for coherent risk measures result in many desirable properties of $\mathscr{R}$. Any functional $\mathscr{R} : \mathscr{X} \to \mathbb{R}$ satisfying axioms (R2) and (R1′) is continuous in the norm topology of the space $\mathscr{X} = L^p(\Xi, \mathscr{B}, \mathbb{P})$ (see Proposition 6.6 in [108]). Therefore, the Fenchel–Moreau theorem [6, Thm. 9.3.5] ensures that $\mathscr{R}$ is equal to its biconjugate function,

$$\mathscr{R}(X) = \sup_{\theta \in \mathscr{X}^*} \{\mathbb{E}[\theta X] - \mathscr{R}^*(\theta)\}, \tag{12}$$

where $\mathscr{R}^* : \mathscr{X}^* \to \mathbb{R} \cup \{+\infty\}$ is the Legendre–Fenchel transformation of $\mathscr{R}$, i.e.,

$$\mathscr{R}^*(\theta) = \sup_{X \in \mathscr{X}} \{\mathbb{E}[\theta X] - \mathscr{R}(X)\}.$$

Clearly, the set $\mathscr{X}^*$ in the representation (12) can be replaced by

$$\mathrm{dom}(\mathscr{R}^*) = \{\theta \in \mathscr{X}^* : \mathscr{R}^*(\theta) < +\infty\}.$$

In this setting, one can further show that (R2) and (R3) hold if and only if for all $\theta \in \mathrm{dom}(\mathscr{R}^*)$ we have that $\theta \geq 0$ $\mathbb{P}$-a.e. and $\mathbb{E}[\theta] = 1$. That is, $\mathrm{dom}(\mathscr{R}^*)$ is a subset of the probability density functions in $\mathscr{X}^*$. Finally, (R4) holds if and only if $\mathscr{R}^*(\theta) = 0$ for all $\theta \in \mathrm{dom}(\mathscr{R}^*)$. See [108, Th. 6.5] for a proof of these results. In fact, Theorem 6.7 in [108] ensures that a risk measure $\mathscr{R}$ is coherent if and only if it has the equivalent form

$$\mathscr{R}(X) = \sup_{\theta \in \mathfrak{A}} \mathbb{E}[\theta X] \tag{13}$$

where $\mathfrak{A} \subset \mathscr{X}^*$ is a convex, bounded, and weakly* closed subset of probability density functions, i.e., $\mathfrak{A} = \mathrm{dom}(\mathscr{R}^*)$.

In addition to the axioms for coherent risk measures, a fundamentally important property of $\mathscr{R}$ is law invariance. We say that two random variables are *distributionally equivalent*, denoted $X \overset{D}{\sim} X'$, if their cumulative distribution functions (cdf) $\Psi_X(t) = \mathbb{P}(X \leq t)$ and $\Psi_{X'}(t) = \mathbb{P}(X' \leq t)$ are equal for all $t \in \mathbb{R}$. A functional $\mathscr{R} : \mathscr{X} \to \mathbb{R}$ is then said to be *law invariant* if

$$X \overset{D}{\sim} X' \quad \Longrightarrow \quad \mathscr{R}(X) = \mathscr{R}(X') \tag{14}$$

for any two random variables $X, X' \in \mathscr{X}$. In words, property (14) ensures that $\mathscr{R}$ is only a function of the cdf $\Psi_X(t) = \mathbb{P}(X \leq t)$ for any random variable X. For example, this excludes the scenario in which $\mathscr{R}(X) \neq \mathscr{R}(X')$ where X and X' are distributionally equivalent discrete random variables whose atoms are ordered differently.

Another important notion in stochastic optimization is that of *stochastic dominance*. A random variable X dominates another random variable X' with respect to the *first stochastic order* if

$$\Psi_X(t) \leq \Psi_{X'}(t) \quad \forall\, t \in \mathbb{R}. \tag{15}$$

We denote the relation (15) by $X \succeq_{(1)} X'$. Similarly, X dominates X' with respect to the *second stochastic order* if

$$\int_{-\infty}^{t} \Psi_X(\eta)\, \mathrm{d}\eta \leq \int_{-\infty}^{t} \Psi_{X'}(\eta)\, \mathrm{d}\eta \quad \forall\, t \in \mathbb{R}. \tag{16}$$

Owing to Fubini's theorem [45, Thm. 2.37], it is straightforward to show that

$$\int_{-\infty}^{t} \Psi_X(\eta)\, \mathrm{d}\eta = \mathbb{E}\left[\int_{-\infty}^{t} \mathbb{1}_{X \leq \eta}\, \mathrm{d}\eta\right] = \mathbb{E}[(t - X)_+]$$

where, for any $E \in \mathscr{B}$, $\mathbb{1}_E(\xi) = 1$ if $\xi \in E$ and $\mathbb{1}_E(\xi) = 0$ otherwise, and $(x)_+ = \max\{0, x\}$. Therefore, (16) is equivalent to the condition

$$\mathbb{E}[(t - X)_+] \leq \mathbb{E}[(t - X')_+] \quad \forall\, t \in \mathbb{R}.$$

We denote the relation (16) by $X \succeq_{(2)} X'$. If $(\Xi, \mathscr{B}, \mathbb{P})$ is nonatomic and $\mathscr{R}$ is law invariant, then the following two results hold: (i) the implication $X \succeq_{(1)} X' \implies \mathscr{R}(X) \geq \mathscr{R}(X')$ holds if and only if $\mathscr{R}$ satisfies the monotonicity condition (R2) [108, Th. 6.50]; (ii) if $\mathscr{R}$ satisfies conditions (R1′), (R2), and (R3), then $-X' \succeq_{(2)} -X$ implies $\mathscr{R}(X) \geq \mathscr{R}(X')$ [108, Th. 6.51]. These two properties demonstrate that law invariant coherent risk measures $\mathscr{R}$ *prefer dominated random variables* and thus are critical in reducing uncertainty (i.e., variability) in the optimized system. On the other hand, as previously noted, one could define risk aversion via stochastic dominance constraints instead of risk measures. For example, suppose $\bar{z}$ is known to produce an acceptable objective value $\mathscr{J}(\bar{z})$. One could then incorporate a constraint of the form

$$\mathscr{J}(\bar{z}) \succeq_{(1)} \mathscr{J}(z) \quad \text{or} \quad -\mathscr{J}(z) \succeq_{(2)} -\mathscr{J}(\bar{z}).$$

For more information of stochastic dominance constraints, see [38].

Example 1 (Mean-Plus-Deviation) A common risk measure in engineering applications, motivated by Markowitz's pioneering work in portfolio optimization [74], is the mean-plus-deviation risk measure

$$\mathscr{R}(X) = \mathbb{E}[X] + c\mathbb{E}[|X - \mathbb{E}[X]|^p]^{\frac{1}{p}}, \quad c > 0$$

for $p \in [1, \infty)$. Clearly, $\mathscr{R}$ is naturally defined and real valued on $\mathscr{X} = L^p(\Xi, \mathscr{B}, \mathbb{P})$ and is law invariant, convex, positively homogeneous, and translation equivariant. Unfortunately, $\mathscr{R}$ is not monotonic and can lead to the paradoxical scenario where one position is always smaller than another, but the larger position has smaller risk. In the context of finance, the risk measure $\mathscr{R}$ can lead to the selection of portfolios that have smaller risk and smaller returns. See [108, Ex. 6.62] for a simple example of this undesirable situation. The lack of monotonicity results from $\mathscr{R}$ equally penalizing the deviation below and above the expected value. In terms of minimization, one prefers large deviation below the expected value since this could lead to better than expected performance. A related law-invariant risk measure that is coherent is the mean-plus-upper-semideviation risk measure

$$\mathscr{R}(X) = \mathbb{E}[X] + c\mathbb{E}[(X - \mathbb{E}[X])_+^p]^{\frac{1}{p}}, \quad c \in [0, 1].$$

Note that this risk measure only penalizes deviation in excess of the expected value. Since this $\mathscr{R}$ is coherent, it can be represented as in (13) with *risk envelope*

$$\operatorname{dom}(\mathscr{R}^*) = \left\{\theta \in \mathscr{X}^* : \theta = 1 + \theta' - \mathbb{E}[\theta'],\ \|\theta'\|_{\mathscr{X}^*} \leq c,\ \theta' \geq 0\ \mathbb{P}\text{-a.e.}\right\}.$$

See [108, Ex. 6.23] for more details.

Example 2 (Conditional Value-at-Risk) The conditional value-at-risk[1] (CVaR) is a coherent risk measure that has recently received much attention [64, 94, 115]. CVaR at confidence level $\alpha \in (0, 1)$ is defined as

$$\mathscr{R}(X) = \mathrm{CVaR}_\alpha(X) := \inf_{t\in\mathbb{R}} \left\{ t + \frac{1}{1-\alpha}\mathbb{E}[(X-t)_+] \right\}, \tag{17}$$

which naturally acts on random variables in $\mathscr{X} = L^1(\Xi, \mathscr{B}, \mathbb{P})$. If the random variable X is continuously distributed, then $\mathrm{CVaR}_\alpha(X)$ is the expectation of X conditioned on the event that X is larger than its α-quantile, i.e.,

$$\mathrm{CVaR}_\alpha(X) = \mathbb{E}[X | X > \Psi_X^{-1}(\alpha)].$$

In the financial literature, the quantile $\Psi_X^{-1}(\alpha)$ is called the Value-at-Risk. Moreover, when $\alpha = 0$ we have that $\mathrm{CVaR}_0(X) = \mathbb{E}[X]$ and

$$\lim_{\alpha\uparrow 1} \mathrm{CVaR}_\alpha(X) = \operatorname{ess\,sup} X.$$

Since CVaR is coherent, it can be represented as in (13) with risk envelope

$$\mathrm{dom}(\mathscr{R}^*) = \left\{ \theta \in L^\infty(\Xi, \mathscr{B}, \mathbb{P}) \,:\, \mathbb{E}[\theta] = 1,\ 0 \le \theta \le (1-\alpha)^{-1}\ \mathbb{P}\text{-a.e.} \right\}.$$

See [108, Ex. 6.19] for more details.

Example 3 (Higher-Moment Coherent Risk) CVaR was extended in [66] to the higher-moment coherent risk measure (HMCR),

$$\mathscr{R}(X) = \inf_{t\in\mathbb{R}} \left\{ t + \frac{1}{1-\alpha}\mathbb{E}[(X-t)_+^p]^{\frac{1}{p}} \right\},$$

with $p \in (1, \infty)$. HMCR is a law-invariant coherent risk measure and is finite for random variables in $\mathscr{X} = L^p(\Xi, \mathscr{B}, \mathbb{P})$ (see [37] for a thorough analysis of HMCR). Since HMCR is coherent, it can be represented as in (13) with risk envelope

$$\mathrm{dom}(\mathscr{R}^*) = \left\{ \theta \in \mathscr{X}^* \,:\, \mathbb{E}[\theta] = 1,\ \theta \ge 0\ \mathbb{P}\text{-a.e.},\ \|\theta\|_{\mathscr{X}^*} \le \frac{1}{1-\alpha} \right\}.$$

This risk envelope was determined in [29, Sect. 5.3.1] for the more general class of *transformed norm risk measures*. Note that HMCR and CVaR coincide if $p = 1$ and thus so do their risk envelopes.

[1] Also called Average Value-at-Risk, Expected Shortfall, Expected Tail Loss and Superquantile.

Example 4 (Entropic Risk) The entropic risk measure is defined as

$$\mathscr{R}(X) = \sigma^{-1} \log\left(\mathbb{E}[\exp(\sigma X)]\right), \quad \sigma > 0,$$

and is finite for random variables in $\mathscr{X} = L^{\infty}(\Xi, \mathscr{B}, \mathbb{P})$. The entropic risk is convex, monotonic, and translation equivariant but is not positively homogeneous and therefore is not coherent. The name entropic risk comes from the Legendre–Fenchel transformation of $\mathscr{R}$. Since the topological dual space of $\mathscr{X} = L^{\infty}(\Xi, \mathscr{B}, \mathbb{P})$ is difficult to handle in practice, it is natural to view $\mathscr{X}$ and $L^1(\Xi, \mathscr{B}, \mathbb{P})$ as paired, locally convex topological vector spaces where $\mathscr{X}$ is equipped with the weak* topology and $L^1(\Xi, \mathscr{B}, \mathbb{P})$ is equipped with the norm topology (see, e.g., [108, Sect. 6.3] for a discussion of essentially bounded random variables). In this setting, one can show that the Legendre–Fenchel transformation of $\mathscr{R}$ is

$$\mathscr{R}^*(\theta) = \sup_{X \in \mathscr{X}} \{\mathbb{E}[\theta X] - \mathscr{R}(X)\} = \sigma^{-1}\mathbb{E}[\theta \log(\theta)]$$

when $\theta \in L^1(\Xi, \mathscr{B}, \mathbb{P})$ satisfying $\theta \geq 0$ $\mathbb{P}$-a.e. and $\mathbb{E}[\theta] = 1$. This is the negative of Shannon's entropy, i.e., the Kullback–Leibler divergence (up to the scaling by σ^{-1}). See [108, Ex. 6.20] for more details.

4.2 Probabilistic Optimization

As with risk measures, we assume in this section that $\mathbb{P}$ is known. In many applications, it is extremely important that an optimal control or design reduces the probability that the event

$$\left\{\xi \in \Xi \,:\, [\mathscr{J}(z)](\xi) > \tau\right\} \tag{18}$$

for some prescribed threshold $\tau \in \mathbb{R}$ occurs. For example, the event (18) could signify the failure of a structure. This naturally leads to the probabilistic objective function

$$\mathscr{R}(\mathscr{J}(z)) = \mathbb{P}(\mathscr{J}(z) > \tau) = \mathbb{E}[\mathbb{1}_{\mathscr{J}(z)>\tau}]. \tag{19}$$

Recall the definition of $\mathbb{1}_E$ from Section 4.1. Much work has been devoted to probabilistic optimization including the derivation of derivative formulas for this choice of $\mathscr{R}$ [76, 98, 113, 114, 117]. The functional $\mathscr{R}$ is only differentiable under certain assumptions which may be difficult to verify in the context of PDE-constrained optimization. For example, the authors in [117] require that $\xi \mapsto [\mathscr{J}(z)](\xi)$ is convex with respect to ξ and that the random vector ξ is Gaussian. Moreover, many differentiation formulas are stated in finite dimensions and it is unclear whether or not these formulas hold in infinite dimensions. Additional complications arise

when estimating probabilistic functions. See [93] for a detailed discussion of the challenges associated with estimation and optimization of probabilistic functions. Finally, $\mathscr{R}$ only quantifies the "number" of scenarios for which $\mathscr{J}(z) > \tau$ but ignores the magnitudes of these scenarios. This could lead to a situation where the optimal controls or designs result in a small probability of (18) occurring, but all scenarios in (18) have large magnitude. For example, (18) could represent any failure (no matter how minor) of the system whereas large-magnitude scenarios signal catastrophic failure.

For these reasons, the authors of [93] developed the concept of buffered probabilities. Roughly speaking, the buffered probability is one minus the inverse of $\alpha \mapsto \mathrm{CVaR}_\alpha(X)$. Let $X \in \mathscr{X} = L^1(\Xi, \mathscr{B}, \mathbb{P})$ be a nondegenerate (i.e., nonconstant) random variable, then $\alpha \mapsto \mathrm{CVaR}_\alpha(X)$ is continuous and nondecreasing for $\alpha \in [0, 1)$ and strictly increasing for $\alpha \in [0, 1 - \pi_\infty)$ where

$$\pi_\infty = \pi_\infty(X) = \mathbb{P}(\{\xi \in \Xi : X(\xi) = \operatorname{ess\,sup} X\})$$

[94]. Therefore, an inverse to $\alpha \mapsto \mathrm{CVaR}_\alpha(X) : [0, 1) \to [\mathbb{E}[X], \operatorname{ess\,sup} X)$ exists. Now, suppose X is degenerate, i.e., there exists $t \in \mathbb{R}$ such that $X = t$ $\mathbb{P}$-a.e., then $\mathrm{CVaR}_\alpha(X) = t$ for any $\alpha \in [0, 1)$ by axioms (R3) and (R4) in Section 4.1 and thus the inverse is not defined. Using these properties of CVaR, we define the buffered probability that a nondegenerate random variable X exceeds the threshold τ as $\bar{p}_\tau(X)$ where $\alpha = 1 - \bar{p}_\tau(X)$ solves

$$\tau = \mathrm{CVaR}_\alpha(X).$$

It is not hard to show that $\bar{p}_\tau(X) \geq \mathbb{P}(X > \tau)$. Moreover, if X is continuously distributed then the buffered probability is $\bar{p}_\tau(X) = \mathbb{P}(X > \tau_X)$ where τ_X solves

$$\mathbb{E}[X | X > \tau_X] = \tau.$$

In this case, τ_X is the $\alpha = 1 - \bar{p}_\tau(X)$ quantile of X. One can think of τ_X as defining a "buffer" or "safety" zone around the event (18) defined via the average of scenarios in the upper tail. Figure 1 contains a comparison of the buffered probability and the usual probability for a normally distributed random variable X. The blue line corresponds to the cdf Ψ_X while the red line corresponds to the inverse of $\alpha \mapsto \mathrm{CVaR}_\alpha(X)$, denoted $\overline{\Psi}_X(\tau)$.

It was shown in [71] that for $\tau < \operatorname{ess\,sup} X$ the buffered probability has the convenient optimization formulation

$$\bar{p}_\tau(X) = \inf_{t \geq 0} \mathbb{E}[(t(X - \tau) + 1)_+]. \tag{20}$$

This permits the optimization of $z \mapsto \bar{p}_\tau(\mathscr{J}(z))$ over Z_{ad} to be reformulated as the optimization of $(z, t) \mapsto \mathbb{E}[(t(\mathscr{J}(z) - \tau) + 1)_+]$ over the augmented space $Z_{\mathrm{ad}} \times [0, \infty)$. The objective function in the later expression is the composition of a convex

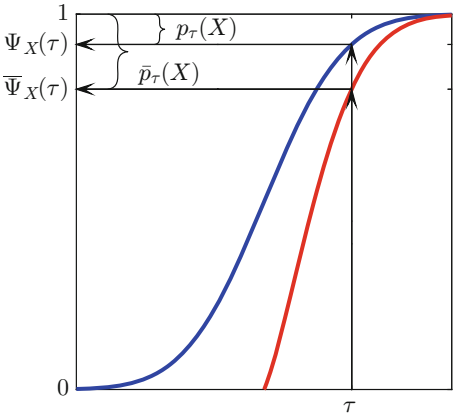

Fig. 1 A comparison of the probability that X exceeds τ, $p_\tau(X)$, and the buffered probability that X exceeds τ, $\bar{p}_\tau(X)$. The blue line is Ψ_X whereas the red line is the inverse of $\alpha \mapsto \text{CVaR}_\alpha(X)$, denoted $\overline{\Psi}_X$

function with our random variable objective function. In addition, the authors of [71] show that $X \mapsto \bar{p}_\tau(X)$ is a lower semicontinuous, quasi-convex, and monotonic function (i.e., satisfies (R2) in Section 4.1). Finally, if $X \mapsto \bar{p}_\tau(X)$ is considered as a function on $L^2(\Xi, \mathscr{B}, \mathbb{P})$, one can show that it is the minimal upper bound for $\mathbb{P}(X > \tau)$ among all quasi-convex, lower semicontinuous law-invariant functions acting on elements of $L^2(\Xi, \mathscr{B}, \mathbb{P})$ [71, Prop. 3.12]. This optimality result is related to the results in [81] in which the authors seek an optimal convex approximation for chanced constrained optimization problems.

4.3 Distributionally Robust Optimization

Often the true probability law $\mathbb{P}$ of the random inputs ξ is not known but estimated from noisy and incomplete data. In this case, making a decision based solely on an estimate of $\mathbb{P}$ can be catastrophic if the estimate does not accurately characterize the statistical behavior of the true underlying distribution. In such scenarios, we must be "averse" to the risk associated with our lack of knowledge of the true underlying probability distribution. This motivates the *distributionally robust* approach to stochastic programming of optimizing the "worst expectation"

$$\min_{z \in Z_{\text{ad}}} \left\{ \mathfrak{J}(z) := \sup_{P \in \mathfrak{M}} \mathbb{E}_P\big[\mathscr{J}(z)\big] + \wp(z) \right\}, \tag{21}$$

where $\mathfrak{M}$ is a specified set of admissible probability measures defined on the measurable space $(\Xi, \mathscr{B})$ and

$$\mathscr{R}(X) := \sup_{P \in \mathfrak{M}} \mathbb{E}_P[X] \tag{22}$$

is the associated risk functional. The set $\mathfrak{M}$ is often called the *ambiguity set*. For more information on robust optimization see, e.g., [14, 23, 41, 107] and the references therein.

In the setting of distributionally robust optimization, we often have partial information regarding the probability law $\mathbb{P}$. Using this information, we can employ Bayesian analysis to determine a single posterior distribution for ξ (see, e.g., [19]), which we can then use to formulate and solve a risk-averse (Section 4.1) or probabilistic (Section 4.2) optimization problem. Although Bayes' rule provides an analytic expression for the posterior distribution, the posterior distribution often does not have a practical (i.e., implementable on a computer) representation. Moreover, Bayesian analysis relies on subjective beliefs encoded in the chosen prior distribution for ξ. Therefore, if the prior distribution is chosen incorrectly, any decision made using the posterior distribution may result in unexpected, undesirable outcomes. There are a number of ways to circumvent this potential pitfall such as, e.g., uninformative priors or robust Bayesian analysis. Robust Bayesian analysis generates a family of posterior distributions using predefined families of noise and prior distributions [18, 20]. In the context of the distributionally robust optimization problem (21), we can incorporate this family of posterior distributions within the ambiguity set $\mathfrak{M}$.

In addition to the previously described robust Bayesian approach, there are two somewhat different methods for constructing the ambiguity set $\mathfrak{M}$. In one approach, we assume that there is a specified reference probability measure $\mathbb{P}_0$ and that the set $\mathfrak{M}$ consists of probability measures in some sense close to $\mathbb{P}_0$. If we assume further that $\mathfrak{M}$ is a set of probability measures that are absolutely continuous with respect to the reference probability measure $\mathbb{P}_0$, then as a consequence of the Radon–Nikodym theorem [45], for every $Q \in \mathfrak{M}$ there exists a $\mathscr{B}$-measurable function $\theta : \Xi \to \mathbb{R}$ such that $\mathrm{d}Q = \theta \mathrm{d}\mathbb{P}_0$. That is, with the set $\mathfrak{M}$ is associated the set of densities $\mathfrak{A} = \{\theta = \mathrm{d}Q/\mathrm{d}\mathbb{P}_0 : Q \in \mathfrak{M}\}$. Assuming that $\mathfrak{A} \subset \mathscr{X}^*$ where $\mathscr{X} = L^p(\Xi, \mathscr{B}, \mathbb{P}_0)$ with $1 \leq p < \infty$, the corresponding functional

$$\mathscr{R}(X) = \sup_{\theta \in \mathfrak{A}} \mathbb{E}[\theta X] \tag{23}$$

becomes a coherent risk measure defined on $\mathscr{X}$. By the duality relation (13), there is a one-to-one correspondence between coherent risk measures and distributionally robust functionals of the form (23).

Another common approach is to define $\mathfrak{M}$ through moment matching. This approach was pioneered by Scarf [103]. For moment matching, we assume that K moments of ξ are specified (e.g., estimated from data), and the ambiguity set is defined as

$$\mathfrak{M} := \left\{ Q : \mathscr{B} \to [0,1] \,:\, Q(\Xi) = 1,\ \mathbb{E}_Q[\psi_k(\xi)] \leq m_k,\ k = 1, \ldots, K \right\}, \tag{24}$$

where ψ_k are real-valued $\mathscr{B}$-measurable functions and $m_k \in \mathbb{R}$. For example, setting $\psi_k(\xi) := e_k^\top \xi$ where e_k denotes the m-vector of zeros with one in the kth position

(i.e., the kth component of ξ) for $k = 1, \ldots, m$ would produce the mean value in each direction of Ξ. The moment matching problem is naturally posed in the uniform closure of the space of continuous random variables with compact support, $\mathscr{X} = C_0(\Xi)$, whose topological dual space, by the Riesz Representation Theorem (see, e.g., [45, Th. 7.17] or [6, Th. 2.4.6]), is isometrically isomorphic to the Banach space of signed regular Borel measures endowed with the total variation norm (i.e., $\Xi \subseteq \mathbb{R}^m$ is a locally compact Hausdorff space). Note that if Ξ is compact, then $C_0(\Xi) = C(\Xi)$.

When the ambiguity set $\mathfrak{M}$ is defined by the moment constraints (24), evaluation of the respective functional $\mathscr{R}(X)$, defined as the optimal value of the maximization problem given by the right-hand side of (22), is known as the *problem of moments*. It is possible to show that it suffices to perform the maximization in (22) with respect to probability measures $P \in \mathfrak{M}$ with support having at most $K + 1$ points [97] (see also Proposition 6.66 and Theorem 7.37 in [108]). That is, $\mathscr{R}(\mathscr{J}(z))$ is equal to the optimal value of the following program:

$$\begin{aligned} \max_{\xi_1,\ldots,\xi_{K+1}\in\Xi,\ \alpha\in\mathbb{R}^{K+1}_+} \quad & \sum_{i=1}^{K+1} \alpha_i \mathscr{J}(z,\xi_i) \\ \text{s.t.} \quad & \sum_{i=1}^{K+1} \alpha_i \psi_k(\xi_i) \le m_k,\ k = 1,\ldots,K,\ \sum_{i=1}^{K+1} \alpha_i = 1 \end{aligned} \tag{25}$$

where $\mathbb{R}_+ := [0, +\infty)$. Furthermore, the (Lagrangian) dual of the optimization problem (25) can be written as the following semi-infinite program:

$$\begin{aligned} \min_{\mu\in\mathbb{R}\times\mathbb{R}^K_+} \quad & \mu_0 + \sum_{k=1}^{K} m_k\mu_k \\ \text{s.t.} \quad & \mu_0 + \sum_{k=1}^{K} \mu_k\psi_k(\xi) \ge \mathscr{J}(z,\xi),\ \xi \in \Xi. \end{aligned} \tag{26}$$

Under mild regularity conditions, there is no duality gap between problems (25) and (26), and hence $\mathscr{R}(\mathscr{J}(z))$ is equal to the optimal value of the dual problem (26). One such regularity condition is that the set Ξ is nonempty and compact, and the functions ψ_k, $k = 1, \ldots, K$, and $\mathscr{J}(z, \cdot)$ are continuous on Ξ. Consequently, the respective minimax problem (21) can be written as the following semi-infinite optimization problem:

$$\begin{aligned} \min_{z\in Z_{\mathrm{ad}},\ \mu\in\mathbb{R}\times\mathbb{R}^K_+} \quad & \mu_0 + \sum_{k=1}^{K} m_k\mu_k + \wp(z) \\ \text{s.t.} \quad & \mu_0 + \sum_{k=1}^{K} \mu_k\psi_k(\xi) \ge \mathscr{J}(z,\xi),\ \xi \in \Xi. \end{aligned} \tag{27}$$

In general, solving semi-infinite programs of the form (27) is not easy. In some rather specific cases, (27) can be formulated as a semi-definite programming

problem and solved efficiently [24, 36]. Also a number of specialized algorithms were suggested to solve the moment-matching problem in, e.g., [43, 44, 46].

From the point of view of risk measures $\mathscr{R} : \mathscr{X} \to \mathbb{R}$, with $\mathscr{X} = L^p(\Xi, \mathscr{B}, \mathbb{P}_0)$, the concept of law invariance is a natural one. It ensures that $\mathscr{R}(X)$ can be considered as a function of the cdf $\Psi_X(t) = \mathbb{P}_0(X \leq t)$ associated with X. In the distributionally robust setting, it makes sense to talk about law invariance when the ambiguity set consists of probability measures absolutely continuous with respect to a specified reference probability measure $\mathbb{P}_0$ and the corresponding functional $\mathscr{R}$ is defined in the form (23). It is natural to say that the respective ambiguity set $\mathfrak{A}$, of density functions, is *law invariant* (with respect to the reference probability measure $\mathbb{P}_0$) if $\theta \in \mathfrak{A}$ and $\theta' \overset{D}{\sim} \theta$ implies that $\theta' \in \mathfrak{A}$.

Theorem 3 ([107]) *Consider a set $\mathfrak{A} \subset \mathscr{X}^*$ of density functions and the respective functional $\mathscr{R}$ defined in (23). If the set $\mathfrak{A}$ is law invariant, then the functional $\mathscr{R}$ is law invariant. Conversely, if the functional $\mathscr{R}$ is law invariant and the set $\mathfrak{A}$ is convex and weakly* closed, then $\mathfrak{A}$ is law invariant.*

We can define a large class of law invariant ambiguity sets $\mathfrak{A}$ using the concept of ϕ-divergence [34, 79]. Consider a convex lower semicontinuous function $\phi : \mathbb{R} \to \mathbb{R}_+ \cup \{+\infty\}$ such that $\phi(1) = 0$ and $\phi(x) = +\infty$ for $x < 0$, and define $\mathfrak{A}$ as the set of density functions $\theta \in \mathscr{X}^*$ satisfying the constraint $\mathbb{E}_{\mathbb{P}_0}[\phi(\theta)] \leq \epsilon$ for some $\epsilon > 0$. For example, let $\phi(x) = x \ln x - x + 1$ for $x \geq 0$, and $\phi(x) = +\infty$ for $x < 0$. Then for a probability measure Q absolutely continuous with respect to $\mathbb{P}_0$ and density function $\theta = \mathrm{d}Q/\mathrm{d}\mathbb{P}_0$, we have that $\mathbb{E}_{\mathbb{P}_0}[\theta] = 1$ and hence

$$\mathbb{E}_{\mathbb{P}_0}[\phi(\theta)] = \mathbb{E}_{\mathbb{P}_0}[\theta \ln \theta] = \mathbb{E}_{\mathbb{P}_0}\left[\frac{\mathrm{d}Q}{\mathrm{d}\mathbb{P}_0} \ln \theta\right] = \mathbb{E}_Q[\ln \theta]$$

is the Kullback–Leibler divergence of Q from $\mathbb{P}_0$. As another example for $\alpha \in [0, 1)$, let $\phi(x) = 0$ for $x \in [0, (1-\alpha)^{-1}]$, and $\phi(x) = +\infty$ otherwise. Then for any $\epsilon \geq 0$, the corresponding set $\mathfrak{A}$ consists of density functions θ such that $\theta \leq (1-\alpha)^{-1}$. In that case, the corresponding functional $\mathscr{R}$ becomes the CVaR_α. For many other examples of ϕ-divergence functionals, we refer to [16, 70].

Employing Lagrange multipliers, it is possible to show that the functional $\mathscr{R}$ associated with the ϕ-divergence ambiguity set can be written as

$$\mathscr{R}(X) = \inf_{\mu \geq 0,\, \nu} \{\mu\epsilon + \nu + \mathbb{E}_{\mathbb{P}_0}[(\mu\phi)^*(X - \nu)]\}, \tag{28}$$

e.g., [16, 107]. Here $(\mu\phi)^*(y) = \sup_{x \in \mathbb{R}}\{yx - (\mu\phi)(x)\}$ is the Legendre–Fenchel transformation of $(\mu\phi)$. For the specific case of the Kullback–Leibler divergence, this can be simplified to

$$\mathscr{R}(X) = \inf_{\mu \geq 0} \left\{\mu\epsilon + \mu \ln \mathbb{E}_{\mathbb{P}_0}[\exp(\mu^{-1}X)]\right\}.$$

For the ϕ-divergence ambiguity set, the respective distributionally robust problem (21) can be written as the following stochastic programming problem:

$$\min_{z \in Z_{\text{ad}},\, \mu \geq 0,\, \nu} \mu\epsilon + \nu + \mathbb{E}_{\mathbb{P}_0}[(\mu\phi)^*(\mathscr{J}(z) - \nu)] + \wp(z). \tag{29}$$

We note that the function $(\mu\phi)^*$ is convex and hence problem (29) is convex provided that $\mathscr{J}(\cdot, \xi)$, $\wp$ and the set Z_{ad} are convex. Such problems can be solved by, e.g., Monte Carlo randomization algorithms. We will discuss this further in Section 5.

To conclude this discussion, we point out that the authors of [121] introduce a specific class of ambiguity sets that permit a reformulation of the inner maximization problem to a conic programming problem. The assumptions required for this reformulation are likely not satisfied for general nonlinear, nonconvex PDE-constrained optimization problems, motivating the need for new approximation techniques and optimization algorithms for solving (21).

5 Methods for Expectation-Based Optimization

In general, we cannot apply rapidly converging derivative-based optimization algorithms to solve (2) because the functional $\mathscr{R}$ and hence the composite function $\mathscr{R} \circ \mathscr{J}$ are often not continuously differentiable even if the underlying uncertain reduced objective function is. This issue is critical in determining the practicality of solving (2) since traditional nonsmooth optimization algorithms typically require a number of assumptions that are not satisfied in PDE-constrained optimization (e.g., convexity) and typically exhibit linear or sublinear convergence rates.

With these issues in mind, we restrict our attention to the expectation-based functionals $\mathscr{R}$ of the form

$$\mathscr{R}(X) = \inf_{t \in T} \mathbb{E}[v(X, t)]$$

where $v : \mathbb{R} \times \mathbb{R}^K \to \mathbb{R}$ and $T \subseteq \mathbb{R}^K$, $K \in \mathbb{N}$, is a closed convex set. This is a sufficiently rich class of functionals $\mathscr{R}$ that includes risk neutral $\mathscr{R} = \mathbb{E}$, the conditional value-at-risk (17), the probabilistic objective (19), the buffered probability (20), and the ϕ-divergence distributionally robust objective (28). In general, this class of functionals $\mathscr{R}$ includes the optimized certainty equivalent risk measures [17] and the expectation quadrangle risk measures [95]. To simplify notation, we denote $x = (z, t)$ for $z \in Z$ and $t \in \mathbb{R}^K$, $X = Z \times \mathbb{R}^K$ and $X_{\text{ad}} = Z_{\text{ad}} \times T$. The corresponding PDE-constrained optimization problem is

$$\min_{x=(z,t) \in X_{\text{ad}}} \mathbb{E}[v(\mathscr{J}(z), t)] + \wp(z). \tag{30}$$

For such problems, the composite objective function $h(x) := \mathbb{E}[v(\mathscr{J}(z), t)]$ inherits the differentiability properties of $v(\mathscr{J}(z), t)$ (e.g., [108, Sect. 7.2.4]). In many cases, the function v introduces nonsmoothness into the problem. For example, if $\mathscr{R} = \text{CVaR}_\alpha$, then $v(X, t) = \{t + (1 - \alpha)^{-1}(X - t)_+\}$ with $T = \mathbb{R}$ and if $\mathscr{R}$ is the buffered probability, then $v(X, t) = (t(X - \tau) + 1)_+$ with $T = [0, \infty)$. After fully discretizing (30), one could solve the resulting nonsmooth nonlinear optimization problem using, e.g., bundle methods [72]. We point out that there recently have been attempts to solve risk-averse optimization problems by smoothing CVaR (see [88] for finite-dimensional problems and [64] for PDE-constrained problems). One complication of smoothing approaches is that the gradient of the smoothed risk measure may become unstable as the smoothing is refined (i.e., as the smooth approximation approaches the original nonsmooth quantity), potentially leading to poor convergence of derivative-based optimization algorithms.

The growing interest in uncertainty quantification has led to the development of a multitude of methods for approximating the solution of PDEs with uncertain inputs. These methods can generally be partitioned into two classes: (i) intrusive methods and (ii) nonintrusive methods. Nonintrusive methods treat the deterministic PDE solver as a "black box," whereas intrusive methods require a reformulation of the deterministic PDE solver. Intrusive methods often approximate the solution of a PDE with uncertain inputs by projecting the solution or the PDE residual onto a finite-dimensional subspace such as a space of polynomials. Projection methods include, e.g., stochastic Galerkin and polynomial chaos methods [8, 9, 58, 122] (although there are nonintrusive forms of polynomial chaos [68]). On the other hand, nonintrusive approaches propagate a finite set of samples of ξ through the PDE. One then approximates the PDE solution field using interpolation or approximates integrated quantities such as moments using numerical integration. Some common choices for generating samples of ξ are (quasi) Monte Carlo [39], stochastic collocation on, e.g., sparse grids, [48, 49, 83–86, 110] and stochastic reduced order models [50, 51, 120]. In addition to these well-established methods, there has been much recent work devoted to low-rank tensor decomposition for parametrized PDE solutions [47, 59, 104]. In general, the approximation quality for polynomial-based uncertainty quantification methods is highly dependent on the choice of the approximation space, the dimension of Ξ, and the regularity of the PDE solution with respect to the random inputs.

The incorporation of uncertainty quantification methods within PDE-constrained optimization is an important and open area of research. Any feasible optimization method should be *mesh independent* in the sense that the convergence behavior does not depend on the size of the resulting discretizated problem (with respect to both the spatial domain and Ξ). Additionally, methods should exploit any structures inherent to the problem such as, e.g., adjoints, differentiability, and the optimality conditions in Theorem 2. Recently, numerous authors have applied intrusive and nonintrusive methods to approximate risk neutral optimization problems constrained by PDEs with uncertain inputs. Such problems were efficiently solved in [61, 62] using a trust-region algorithm to guide adaptive sparse grids for approximating the

objective function and its gradient. Similarly, [60] introduces a multilevel sparse grid approach that works well for some linear-quadratic and nonlinear control problems. Furthermore, the authors in [27] solve the risk neutral problem using sparse grids and reduced order models, whereas the authors of [112] solve this problem by combining nonintrusive polynomial chaos with sequential quadratic programming (SQP). Finally, the authors of [47] develop a semismooth Newton solver based on low-rank tensor decomposition to solve the risk neutral problem. Unfortunately, when v in (30) is not differentiable (e.g., minimizing CVaR or the buffered probability), the aforementioned trust-region, SQP, and semismooth Newton algorithms do not apply.

Given the myriad of possible approximations and algorithms for solving (30), we restrict our attention to three nonintrusive sampling approaches: the stochastic approximation algorithm, sample average and quadrature approximation, and the progressive hedging algorithm. We do not intend for this to be a complete list of possible solution techniques, but rather a review of classical methods in stochastic programming that may be applicable in PDE-constrained optimization. For each method, we provide an overview and highlight the challenges associated with the method in the context of PDE-constrained problems.

In the subsequent subsections, we assume X is a Hilbert space with inner product $\langle x, y\rangle_X$ and norm $\|x\|_X = \sqrt{\langle x, x\rangle_X}$. Moreover, we denote the uncertain composite objective function by $H(x,\xi) = v(\mathscr{J}(z,\xi), t)$ and the (deterministic) composite objective function by $h(x) = \mathbb{E}[H(x,\cdot)]$. We further denote the gradient or any subgradient (when $H(\cdot,\xi)$ is convex) of $H(\cdot,\xi)$ by $G(\cdot,\xi)$. To simplify the presentation, we ignore the control penalty term $\wp(z)$. However, all algorithms and results apply if $\wp(z)$ is included.

5.1 Stochastic Approximation

The *stochastic approximation* (SA) method was originally developed by Robbins and Monro in [91]. The method is based on the projected (sub)gradient method. The projection operator $\Pi : X \to X_{\text{ad}}$, onto the set $X_{\text{ad}} \subset X$, is defined as

$$\Pi(y) := \operatorname*{arg\,min}_{x \in X_{\text{ad}}} \|y - x\|_X.$$

Since X is a Hilbert space and X_{ad} is closed and convex, $\Pi(y)$ is uniquely defined for all $y \in X$ [12, Th. 3.14], and $y \mapsto \Pi(y)$ is nonexpansive [12, Prop. 4.8]. At the kth step of SA with the current iteration point x_k, the algorithm computes the next iteration point as

$$x_{k+1} = \Pi\left(x_k - \gamma_k G(x_k, \xi^k)\right). \tag{31}$$

Here $\gamma_k > 0$ are chosen step sizes and ξ^k is a realization of the random vector ξ typically generated by Monte Carlo sampling techniques. The random samples ξ^k, $k = 1, 2, \ldots$, are independent and generated according to the specified distribution of the random vector ξ. Therefore, each iteration point x_k is a random vector depending on the history of random samples $(\xi^1, \ldots, \xi^k)$. Note that each iteration requires a single state and adjoint solve corresponding to the random sample ξ^k. Although per-iteration cost of SA is low, the convergence (which is probabilistic) is heavily dependent on the convexity of $H(\cdot, \xi)$ and the choice of stepsize γ_k.

In the classical SA method, the step size is chosen to be $\gamma_k := \kappa/k$, where $\kappa > 0$ is a fixed constant. To analyze this method, we make the following assumptions:

(i) There exists a constant $M > 0$ such that

$$\mathbb{E}\left[\|G(x, \cdot)\|_X^2\right] \leq M^2, \quad x \in X_{\text{ad}}. \tag{32}$$

(ii) The function $h(x) = \mathbb{E}[H(x, \cdot)]$ is Fréchet differentiable and strongly convex, i.e., there exists $c > 0$ such that

$$h(x') \geq h(x) + \langle \nabla h(x), x' - x \rangle_X + \frac{1}{2} c \|x' - x\|_X^2 \quad \forall\, x,\ x' \in X.$$

Given these assumptions, problem (30) has a unique optimal solution $x_\star$. This result follows from the Direct Method of the Calculus of Variations (i.e., the strong convexity plus the continuity of h ensure the weak lower semicontinuity and coercivity of h). It is possible to show (cf. [80] for finite dimensional X) that for $\kappa > 1/(2c)$,

$$\mathbb{E}\left[\|x_k - x_\star\|_X^2\right] = O(k^{-1}). \tag{33}$$

That is, after k iterations, the expected error of the current solution in terms of the distance to the optimal solution $x_\star$ is of order $O(k^{-1/2})$. Moreover, if $\nabla h(x)$ is Lipschitz continuous and $x_\star \in X_{\text{ad}}$ satisfies $\nabla h(x_\star) = 0$, then (as a consequence of the Mean Value Theorem) we have

$$\mathbb{E}\left[h(x_k) - h(x_\star)\right] = O(k^{-1}). \tag{34}$$

For general convergence results of SA in Hilbert space, see [11].

Under the above assumptions (i) and (ii), the classical SA method produces iterates converging to the optimal solution. However, the method is very sensitive to choice of the step sizes and the convergence can be very slow. A simple example in [80] demonstrates that minimization of a deterministic quadratic function of one variable by the classical SA method can be extremely slow for a wrong choice of the constant κ. Moreover without strong convexity, the step sizes $\gamma_k = \kappa/k$ can result in disastrously slow convergence for any choice of the constant κ.

Another problem with (sub)gradient type algorithms is the possibility of different scales for the components of the vector x. Suppose that the space $X = \mathbb{R}^n$ is equipped with the standard Euclidean inner product $\langle x, y\rangle_X = x^\top y$ and consider the minimization of the (deterministic) quadratic function $h(x) = \frac{1}{2}x^\top Qx$ with Q being an $n \times n$ symmetric positive definite matrix. If the matrix Q is ill conditioned, then for any choice of the step sizes γ_k the SA algorithm will typically produce a zigzag trajectory, resulting in very slow convergence to the optimal solution.

Further, step sizes of order $O(k^{-1})$ could be too small to attain a reasonable rate of convergence, while taking larger step sizes, say of order $O(k^{-1/2})$, may result in no convergence of the algorithm. In order to resolve this problem, it was suggested in [82] (for finite-dimensional problems) to take larger step sizes and to use appropriate averages of the iterates x_k rather than these points themselves. It was shown in [89] that under the assumptions (i) and (ii), this strategy of taking larger step sizes and averaging automatically achieves the asymptotically optimal convergence rate. We follow [80] in analysis of this approach referred to as the robust SA method. Although the results in [80] are for finite dimensional X, it may be possible to extend them to the more general Hilbert space setting. We assume below that the function $h(x)$ is convex continuous, but not necessary strongly convex or differentiable, and that $\mathbb{E}[G(x, \cdot)]$ is a subgradient of h at x, i.e., $\mathbb{E}[G(x, \cdot)] \in \partial h(x)$. We also assume that condition (32) holds and the set X_{ad} is bounded.

For $1 < i < k$, together with the iterates x_k, consider the averages $\hat{x}_{ik} := \sum_{j=i}^k \nu_j x_j$ with weights $\nu_\ell := (\sum_{j=i}^k \gamma_j)^{-1}\gamma_\ell$. Note that $\nu_\ell > 0$ and $\sum_{j=i}^k \nu_j = 1$. We have then the following estimate: [80, p. 1580]

$$\mathbb{E}\big[h(\hat{x}_{ik}) - h(x_\star)\big] \le \frac{4D^2 + M^2 \sum_{j=i}^k \gamma_j^2}{2\sum_{j=i}^k \gamma_j} \quad \text{for } 1 < i < k, \tag{35}$$

where $D := \max_{x \in X_{\mathrm{ad}}} \|x - x_1\|_X$ (since it is assumed that the set X_{ad} is bounded, the constant D is finite). In particular, consider the strategy of fixing in advance the number of iterations N and the constant step sizes $\gamma_k = \gamma$, $k = 1, \ldots, N$. Then it follows from (35) that

$$\mathbb{E}\big[h(\hat{x}_{1N}) - h(x_\star)\big] \le \frac{4D^2 + M^2 N\gamma}{2N\gamma}. \tag{36}$$

Minimization of the right-hand side of (36) over $\gamma > 0$ suggests the optimal constant step size is

$$\gamma := \frac{2D}{M\sqrt{N}}, \tag{37}$$

providing the corresponding error estimate

$$\mathbb{E}\big[h(\hat{x}_{1N}) - h(x_\star)\big] \le \frac{2DM}{\sqrt{N}}. \tag{38}$$

Another possible strategy is to take step sizes of order $O(k^{-1/2})$, specifically

$$\gamma_k := \frac{\theta D}{M\sqrt{k}} \tag{39}$$

for some $\theta > 0$. Choosing i as a fixed fraction of N, i.e., setting $i = \lceil rN \rceil$ for some $r \in (0, 1)$, leads to the estimate

$$\mathbb{E}\big[h(\hat{x}_{iN}) - h(x_\star)\big] \leq C(r) \max\{\theta, \theta^{-1}\} \frac{DM}{\sqrt{N}}, \tag{40}$$

where $C(r)$ is a constant depending only on r.

The estimates (38) and (40) suggest the average error of the objective function to be of order $O(N^{-1/2})$. This could be compared with the estimate (34) of order $O(N^{-1})$. However, the error bounds (38) and (40) do not require differentiability or strong convexity of h. Additionally, scaling the step size in the robust SA algorithm by $\theta > 0$ has only a moderate effect on the bound (40), i.e., $\max\{\theta, \theta^{-1}\}$. Therefore, the robust SA method is considerably less sensitive to the choice of step sizes than the classical SA method. Nevertheless, the choice is still crucial for convergence of the algorithm and, unfortunately, the stepsize formulas (37) and (39) involve constants M, D, and the scaling factor θ that are often impossible to determine for PDE-constrained optimization problems.

5.2 Sample Average and Quadrature Approximation

Both the *sample average approximation* (SAA) and the deterministic quadrature approach result in approximations of the expectation in (30). As such, these methods are not algorithms for solving (30). The idea of the SAA method is to use equally probable random samples $\xi^1, \ldots, \xi^N$ to approximate the "true" optimization problem (30), whereas the quadrature approach aims to approximate the expectation in (30) using deterministic quadrature defined by N abscissae $\{\xi^1, \ldots, \xi^N\}$ and their corresponding weights $\{w^1, \ldots, w^N\}$. Both the SAA and quadrature approximations to (30) have the form

$$\min_{x \in X_{\text{ad}}} \left\{ \hat{h}_N(x) := \sum_{j=1}^{N} p^j H(x, \xi^j) \right\} \tag{41}$$

where $p^j = N^{-1}$ for SAA and $p^j = w^j$ for the quadrature approach. In the context of PDE-constrained optimization, (41) is a deterministic optimization problem with N PDE constraints. Therefore, any solution method for (41) should be mesh independent to avoid convergence issues associated with the dimension of the fully discretized problem.

There are advantages and disadvantages of the SA versus SAA or the quadrature approach. In finite dimensions, estimates of the sample size N needed to attain a specified accuracy of computed solutions are similar for both the SAA and the SA methods (cf., [108, Ch. 5]). SA is a simple algorithm requiring evaluation of a *single* (sub)gradient $G(x_j, \xi^j)$ at each iteration step, while SAA and the quadrature approach are not algorithms – the constructed problem (41) still has to be solved by a numerical procedure. Depending on the choice of algorithm for solving (41), each involved iteration can be considerably more expensive than in the SA method. For example, evaluation of the gradient (or a subgradient) of $\hat{h}_N$ at a given point x requires the calculation of *all* $G(x, \xi^j)$, $j = 1, \ldots, N$. On the other hand, SAA and the quadrature approach, combined with a good numerical optimization algorithm, may overcome the difficulties of the choice of step sizes that plagues the SA method. Also SAA and the quadrature approach are more receptive to parallelization, e.g., the (sub)gradients $G(x, \xi^j)$, $j = 1, \ldots, N$ can be computed in parallel as opposed to the sequential nature of the SA method. However, additional difficulty may arise for the quadrature approximation if the weights w^j are not all positive as with, e.g., sparse grids [48, 49, 85, 86, 110]. The presence of negative weights may adversely influence a numerical optimization solver by changing the sign associated with the objective sample $H(x, \xi^j)$.

Given the similarities between SAA and the quadrature approach, we can characterize the error committed through the approximation of (30) using the same techniques. For the subsequent analysis, we assume $x \mapsto H(x, \xi)$ is continuously Fréchet differentiable for each $\xi \in \Xi$, ensuring that h and $\hat{h}_N$ are continuously Fréchet differentiable. If h is strongly convex, then we can characterize the errors between the true optimal solution $x_\star \in X_{\text{ad}}$ and the approximate solution $x_N \in X_{\text{ad}}$. Namely, strong convexity implies there exists $c > 0$ such that

$$c\|x_\star - x_N\|_X^2 \leq \langle \nabla h(x_\star) - \nabla h(x_N), x_\star - x_N \rangle_X.$$

Similar to Theorem 2, the optimality conditions for h and $\hat{h}_N$ over X_{ad} are

$$\langle \nabla h(x_\star), x - x_\star \rangle_X \geq 0 \quad \forall x \in X_{\text{ad}} \quad \text{and} \quad \langle \nabla \hat{h}_N(x_N), x - x_N \rangle_X \geq 0 \quad \forall x \in X_{\text{ad}},$$

respectively. Since $x_\star,\ x_N \in X_{\text{ad}}$, we have that

$$\langle \nabla h(x_\star), x_\star - x_N \rangle_X \leq 0 \leq \langle \nabla \hat{h}_N(x_N), x_\star - x_N \rangle_X.$$

This relation and the Cauchy–Schwarz inequality ensure that

$$c\|x_\star - x_N\|_X \leq \|\nabla \hat{h}_N(x_N) - \nabla h(x_N)\|_X = \left\| \sum_{j=1}^{N} p^j G(x_N, \xi^j) - \mathbb{E}[G(x_N, \cdot)] \right\|_X. \tag{42}$$

Therefore, the right-hand side of (42) is simply the error associated with approximately integrating the gradient of $H(x_N, \cdot)$ and thus the error will be dictated by the

approximation quality of the points $(\xi^1, \ldots, \xi^N)$ and weights $(p^1, \ldots, p^N)$. In the context of quadrature approximation, this error depends heavily on the regularity of, e.g., the adjoint state with respect to ξ, the dimension of Ξ, and the polynomial order of the quadrature rule (see, for example, [83, 84, 86]). Thus, the convergence rate of the optimal solutions for the quadrature approximation may be algebraic, even exponential, if the gradients G are sufficiently regular with respect to ξ. On the other hand, the convergence rate for SAA is probabilistic since $(\xi^1, \ldots, \xi^N)$ are random realizations of ξ and will likely recover the Monte Carlo rate of convergence $O(N^{-1/2})$ [39].

5.3 Progressive Hedging

The progressive hedging algorithm [96], originally introduced for dynamic stochastic programs, employs a sample-based decomposition of (30). As in Section 5.2, we consider the approximate optimization problem (41) where $(\xi^1, \ldots, \xi^N)$ are fixed scenarios of the uncertain inputs ξ with associated probabilities $(p^1, \ldots, p^N)$ (i.e., $p^j \geq 0$ for all j and $p^1 + \ldots + p^N = 1$). As discussed in Section 5.2, we can exploit parallelism in (41) by evaluating $\hat{h}_N$ and its derivatives in parallel. By assigning a separate optimization variable x^j for each ξ^j (i.e., we allow x^j to *anticipate* the scenario ξ^j), the progressive hedging algorithm further exploits parallel computations at each iteration by concurrently solving a deterministic PDE-constrained optimization problem for each scenario ξ^j.

To describe the progressive hedging algorithm, we first reformulate (41) as

$$\min_{x_j, x \in X_{\text{ad}}} \sum_{j=1}^{N} p^j H(x^j, \xi^j) \quad \text{subject to} \quad x^j = x, \ \ j = 1, \ldots, N. \tag{43}$$

Here, the objective function is the sum of decoupled, scenario-specific objective functions, whereas the constraint ensures that we recover a solution to (41). We call the deterministic variable x an *implementable* solution. We then relax the equality constraint for each j using the augmented Lagrangian penalty function

$$\ell_r^j(x^j, x, \mu^j) = H(x^j, \xi^j) + \langle \mu^j, x^j \rangle_X + \frac{r}{2} \|x^j - x\|_X^2, \quad r > 0,$$

where the multipliers $\{\mu^1, \ldots, \mu^N\}$ are called an *information price system* in [96] and are required to satisfy

$$\sum_{j=1}^{N} p^j \mu^j = 0.$$

Taking the expectation of ℓ_r^j then yields the full Augmented Lagrangian for (43). In light of this, we can describe the progressive hedging algorithm as follows. Given the kth iteration points $x_k^j \in X_{\text{ad}}$ and $\mu_k^j \in X$ for $j = 1, \ldots, N$, and the current implementable solution $x_k = \sum_{j=1}^N p^j x_k^j$:

1. Compute the scenario-dependent solutions x_{k+1}^j, $j = 1, \ldots, N$ by minimizing $\ell_r^j(\cdot, x_k, \lambda_k^j)$ concurrently, i.e.,

$$x_{k+1}^j \in \underset{x^j \in X_{\text{ad}}}{\arg\min}\ \ell_r^j(x^j, x_k, \lambda_k^j),\ \ j = 1, \ldots, N; \tag{44}$$

2. Aggregate x_{k+1}^j to compute the current implementable solution x_{k+1}, i.e.,

$$x_{k+1} = \sum_{j=1}^N p^j x_{k+1}^j;$$

3. Update the multiplier estimates for fixed $x = x_{k+1}$ and $x^j = x_{k+1}^j$, $j = 1, \ldots, N$, as

$$\mu_{k+1}^j = \mu_k^j + r(x_{k+1}^j - x_{k+1}),\ \ j = 1, \ldots, N. \tag{45}$$

Clearly, all steps of this algorithm are parallelizable with the exception of the second (i.e., aggregation) step.

The convergence theory for the progressive hedging algorithm, as set fourth in [96], is restricted to finite dimensions. When $H(\cdot, \xi)$ is convex, the progressive hedging algorithm converges under specified stopping rules for approximately solving (44) (see Equation 5.35 and Theorem 5.4 in [96]). In fact, the convergence theory in the convex case is based on the convergence theory for the proximal point algorithm [92] applied to a certain saddle function. As the authors in [42] point out, the progressive hedging algorithm can be seen as a special case of Douglas–Rachford splitting and thus inherits the Hilbert space convergence theory. On the other hand, Theorem 6.1 in [96] demonstrates that if $H(\cdot, \xi)$ is not convex and X is finite dimensional, then if the sequences of iterates x_k^j and multipliers μ_k^j converge, where x_k^j are only required to be δ-locally optimal for fixed $\delta > 0$, then these sequences converge to a stationary point of the original problem (30). Given the relations between the progressive hedging and Augmented Lagrangian algorithms, it may be possible to extend the convergence analysis for Augmented Lagrangian for infinite-dimensional nonconvex problems (see, e.g., [54, Chapt. 3]).

To conclude, one potential inefficiency of the progressive hedging algorithm is the typically slow convergence rate. For example, if X is finite dimensional, $H(\cdot, \xi)$ is convex quadratic, and X_{ad} is convex polyhedral, then Theorem 5.2 in [96] ensures that the progressive hedging algorithm will converge at a linear rate. One can potentially overcome this by increasing the penalty parameter r at each

iteration (see, e.g., Theorem 2 in [92] where superlinear convergence for convex problems is shown using the proximal point algorithm). In any case, the convergence of the progressive hedging algorithm is strongly dependent on the penalty parameter r which is difficult to select a priori, especially for nonconvex problems. Another possibility to enhance the convergence rate is to replace (45) with a "second-order" multiplier update (see, e.g., [22, Ch. 2.3.2] and [54, Chapt. 6.2] for second-order multiplier updates in the context of the Augmented Lagrangian algorithm).

6 Numerical Example

To demonstrate the various stochastic programming formulations discussed in Section 4, we consider the problem of optimally mitigating a contamination by injecting chemicals at specified locations that dissolve the contaminant. We model the contaminant transport using the steady advection diffusion equation. Clearly, uncertainties arise in nearly all coefficients such as the velocity field (e.g., wind) and the contaminant source locations and magnitudes. This example was first considered in [64]. Let $D = (0, 1)^2$ denote the physical domain and $U = H^1(D)$ be the space of contaminant concentrations. The target optimization problem is

$$\min_{z \in Z_{\text{ad}}} \mathscr{R}\left(\frac{\kappa_s}{2}\int_D S(z;\cdot)^2 \,\mathrm{d}x\right) + \wp(z) \tag{46}$$

where $\kappa_s > 0$ and $S(z;\cdot) = u : \Xi \to U$ solves the weak form of the advection-diffusion equation

$$-\nabla\cdot(\epsilon(\xi)\nabla u) + \mathbb{V}(\xi)\cdot\nabla u = f(\xi) - Bz \qquad \text{in } D \tag{47a}$$

$$u = 0 \qquad \text{on } \Gamma_d \tag{47b}$$

$$-\epsilon(\xi)\nabla u \cdot n = 0 \qquad \text{on } \Gamma_n \tag{47c}$$

where the Neumann boundary is $\Gamma_n := \{1\}\times(0,1)$ and the Dirichlet boundary is $\Gamma_d := \partial D \setminus \Gamma_n$. The control space (the space of mitigating chemical concentrations) is $Z = \mathbb{R}^9$ with admissible control set $Z_{\text{ad}} := \left\{z \in \mathbb{R}^9 : 0 \le z \le 1\right\}$ and control cost

$$\wp(z) := \kappa_c \|z\|_1 = \kappa_c \sum_{k=1}^{9} |z_k|, \quad \kappa_c > 0.$$

The controls are applied using the operator $B \in \mathscr{L}(Z, L^\infty(D))$ given by

$$(Bz)(x) = \sum_{k=1}^{9} z_k \exp\left(-\frac{(x-p_k)^\top(x-p_k)}{2\sigma^2}\right)$$

Table 1 Predetermined contaminant mitigating control locations

Source	1	2	3	4	5	6	7	8	9
x_1	0.25	0.50	0.75	0.25	0.50	0.75	0.25	0.50	0.75
x_2	0.25	0.25	0.25	0.50	0.50	0.50	0.75	0.75	0.75

where p_k are predetermined control locations and $\sigma = 0.05$. That is, we model the control mechanism as Gaussians sources with magnitude dictated by z. The control locations are tabulated in Table 1.

The PDE coefficients ϵ, $\mathbb{V}$, and f are random fields. The diffusivity is given by

$$\epsilon(x, \xi) = 0.5 + c \exp(\delta(x, \xi))$$

where the specific form of δ is described in [83, Sect. 4, Eqs. 4.2–4.4]. Associated with δ are 10 random variables, $(\xi_1, \ldots, \xi_{10})$, uniformly distributed on $[-\sqrt{3}, \sqrt{3}]$. The constant $c > 0$ is chosen to be the reciprocal of the maximum of $\exp(\delta)$. Clearly, ϵ satisfies: $\exists\, 0 < \epsilon_0 \leq \epsilon \leq \epsilon_1 < \infty$ for all $x \in D$ and $\xi_i \in [-\sqrt{3}, \sqrt{3}]$, $i = 1, \ldots, 10$. Moreover, the velocity field $\mathbb{V}$ is

$$\mathbb{V}(x, \xi) = \begin{bmatrix} \xi_{12} - \xi_{11} x_1 \\ \xi_{11} x_2 \end{bmatrix}$$

where ξ_{11} is uniformly distributed on $[0, 5]$, and ξ_{12} is uniformly distributed on $[5, 10]$. The two extreme cases of $\mathbb{V}$ are depicted in Figure 2. $\mathbb{V}$ is divergence free and satisfies $\mathbb{V} \cdot n \geq 0$, where n is the outward unit normal vector on the Neumann boundary. Finally, f is the sum of five Gaussian sources whose locations, widths, and magnitudes are random, i.e., f is described by 25 uniform random variables $(\xi_{13}, \ldots, \xi_{37})$. This results in a total of 37 random variables associated with the PDE (47). As shown in [64], this example satisfies the assumptions of Theorems 1 and 2 and thus a minimizing control exists and it satisfies the first-order necessary conditions in Theorem 2.

We approximate the contaminant mitigation problem using SAA with $N = 800$ Monte Carlo samples. For $\mathscr{R}$, we chose risk neutral (RN), entropic risk (ER) with $\sigma = 1$, CVaR with $\alpha = 0.95$, a convex combination of expectation and CVaR

$$\mathscr{R}(X) = \beta \mathbb{E}[X] + (1 - \beta) \mathrm{CVaR}_\alpha(X)$$

with $\alpha = 0.95$ and $\beta = 0.5$ (MCVaR), buffered probability with threshold $\tau = 6$ (BP), and KL-divergence distributionally robust optimization with threshold $\epsilon = 0.1$ (KL). Additionally, we solved the *mean value problem* (MV) in which we replaced ξ with $\mathbb{E}[\xi]$ and solved the corresponding deterministic control problem. For RN, ER, KL, and MV, we solved the resulting nonlinear program using a trust-region Newton method [32]; while for CVaR, MCVaR, and BP, we combined the aforementioned trust-region method with an adaptation of the smoothing approach described in [64]. Figure 3 depicts the optimal control sources and Table 2 includes the optimal control magnitudes. We excluded the MV control from Figure 3 due

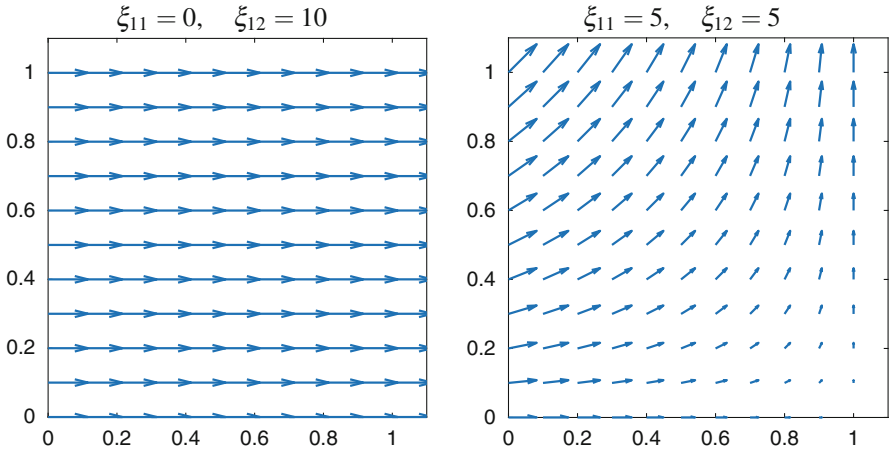

Fig. 2 Left: The vector field $\mathbb{V}$ with $\xi_{11} = 0$ and $\xi_{12} = 10$. Right: The vector field $\mathbb{V}$ with $\xi_{11} = 5$ and $\xi_{12} = 5$

Table 2 Optimal contaminant mitigating controls using different functionals $\mathscr{R}$. MV refers to the deterministic problem in which the random inputs are replaced with their expected values. RN refers to risk neutral and ER refers to the entropic risk with $\sigma = 1$. For CVaR, we set $\alpha = 0.95$ and for the "mixture of CVaRs" (MCVaR), we set $\alpha = 0.95$ and $\beta = 0.5$. For the "buffered probability of exceedance" (bPOE), we set the threshold $\tau = 6$ and for the KL-divergence distributionally robust problem, we set the threshold $\epsilon = 0.1$.

$\mathscr{R}$	1	2	3	4	5	6	7	8	9	Cost
MV	–	0.23	–	–	1.00	–	–	–	–	1.23
RN	–	0.27	–	–	1.00	–	–	–	–	1.27
CVaR	0.42	1.00	0.15	0.81	1.00	1.00	–	–	–	4.37
MCVaR	–	1.00	–	0.33	1.00	0.59	–	–	–	2.92
ER	0.33	1.00	1.00	0.55	1.00	1.00	–	–	–	4.88
BP	0.02	1.00	–	0.56	1.00	0.91	–	–	–	3.49
KL	–	1.00	–	–	1.00	–	–	–	–	2.00

to its similarity with the RN control. For the given parameter specifications, ER produced the most conservative control, whereas RN and MV produce the least conservative. However, conservativeness results in a more expensive control. This fact is depicted in Figure 4. Figure 4 includes the cdfs of the uncertain objective function $\mathscr{J}(z)$ (left) and the full objective function $\mathscr{J}(z) + \wp(z)$ (right) evaluated at the different optimal controls. The left image clearly demonstrates that more conservative approaches reduce variability and produce uncertain objective values that dominate (in the sense of the first stochastic order) those of the RN and MV approaches. On the other hand, the right image emphasizes the increased cost of being conservative. As seen in the right image, the RN and MV controls outperform the other controls in terms of total cost for more than 60% of scenarios.

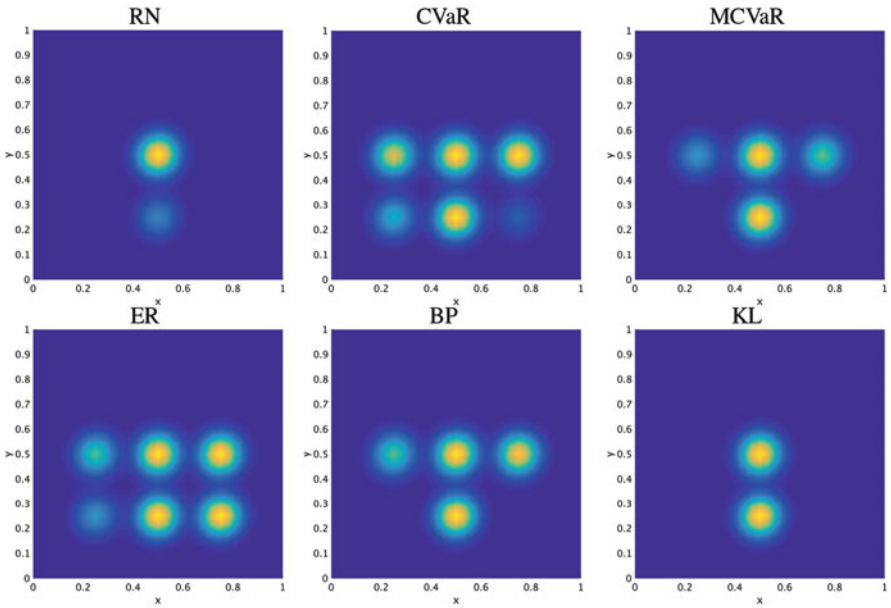

Fig. 3 The optimal controls computed using risk neutral (RN), CVaR, a mixture of expectation and CVaR (MCVaR), entropic risk (ER), buffered probability (BP) and KL-divergence distributionally robust optimization (KL)

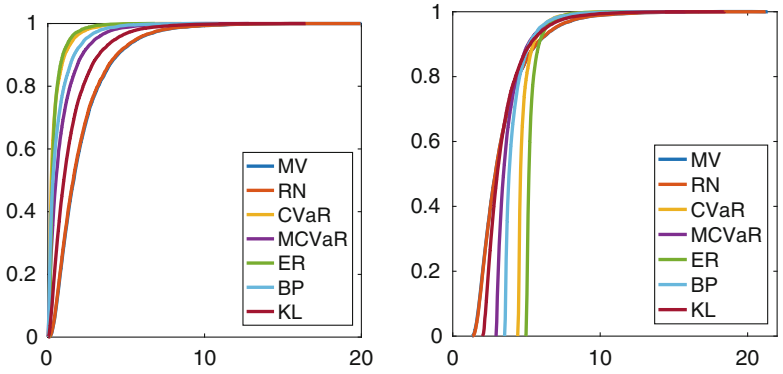

Fig. 4 Left: Cumulative distribution functions of the random variable objective function, $\mathscr{J}(z)$, evaluated at the different optimal controls. Right: Cumulative distribution functions of $\mathscr{J}(z) + \wp(z)$ evaluated at the different optimal controls

7 Conclusions

In this chapter, we reviewed a set of stochastic programming tools for formulating and solving optimization problems constrained by PDEs with uncertain coefficients. For the problem formulation, we discussed risk measures, probabilistic

optimization, and distributionally robust optimization. Each of these approaches can be justified within the context of the physical application. When the underlying probability law of the random coefficients is known, risk-averse and probabilistic optimization provide a natural foundation for incorporating conservativeness in the optimization problem formulation. However, such approaches are unjustified and may lead to arbitrarily poor solutions if the underlying probability law is unknown. In this scenario, one often has noisy, incomplete data describing the distribution of uncertain coefficients which can be used to define an ambiguity set of "feasible" distributions. This leads naturally to distributionally robust optimization in which we minimize the worst-case expectation over the ambiguity set.

For solution approaches, we discussed stochastic approximation (SA), sample average approximation (SAA), deterministic quadrature approximation, and the progressive hedging algorithm. Each approach has particular downsides. The SA approach is a simple optimization algorithm but requires convexity to guarantee convergence, which is probabilistic. The SAA approach approximates the expected value in the objective function using a sample average (e.g., Monte Carlo). The resulting approximate problem is then solved using nonlinear programming algorithms. SAA exhibits dimension-independent convergence, but the convergence is probabilistic with rate $1/\sqrt{N}$. Similar to SAA, the deterministic quadrature approach approximates the expected value using quadrature. The resulting problem is again solved with a nonlinear programming method. This approach requires sufficient regularity (with respect to the random inputs) to obtain rapidly decaying approximation error. Finally, the progressive hedging algorithm employs a sample-based decomposition of the optimization problem and the controls which permits the concurrent solution of deterministic PDE-constrained optimization problems at every iteration. For convex problems, convergence is guaranteed in Hilbert space; however, the convergence rate can be linear or worse.

Common among many stochastic optimization problems is the challenge of minimizing a nonsmooth objective function. In particular, the typical slow convergence rates of nonsmooth optimization algorithms may render the solution of PDE-constrained optimization under uncertainty computationally infeasible. Efficiently solving these nonsmooth problems is challenging and is an active research topic. Additional open research topics include the formulation and analysis for state-constrained problems; the incorporation of stochastic dominance and chance constraints for PDE-constrained optimization; and the formulation, analysis, and numerical solution of optimal control problems constrained by variational inequalities with uncertain inputs as well as optimal control problems constrained by dynamic stochastic PDEs.

Acknowledgements This work was supported by DARPA EQUiPS grant SNL 014150709.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

1. R. A. Adams. *Sobolev Spaces*. Academic Press, New York, 1975.
2. E. Andreassen, B. S. Lazarov, and O. Sigmund. Design of manufacturable 3d extremal elastic microstructure. *Mechanics of Materials*, 69(1):1–10, 2014.
3. V. Artus, J. L. Durlofsky, J. Onwunalu, and K. Aziz. Optimization of nonconventional wells under uncertainty using statistical proxies. *Computational Geosciences*, 10(4):389–404, 2006.
4. Ph. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Math. Finance*, 9(3):203–228, 1999.
5. A. Asadpoure, M. Tootkaboni, and J. K. Guest. Robust topology optimization of structures with uncertainties in stiffness – applications to trust structures. *Computers & Structures*, 89(11–12):1131–1141, 2011.
6. H. Attouch, G. Buttazzo, and G. Michaille. *Variational analysis in Sobolev and BV spaces*, volume 6 of *MPS/SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006.
7. I. Babuška, F. Nobile, and R. Tempone. A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM Rev.*, 52(2):317–355, 2010.
8. I. Babuška, R. Tempone, and G. E. Zouraris. Galerkin finite element approximations of stochastic elliptic partial differential equations. *SIAM J. Numer. Anal.*, 42(2):800–825 (electronic), 2004.
9. I. Babuška, R. Tempone, and G. E. Zouraris. Solving elliptic boundary value problems with uncertain coefficients by the finite element method: the stochastic formulation. *Comput. Methods Appl. Mech. Engrg.*, 194(12–16):1251–1294, 2005.
10. W. Bangerth, H. Klie, M. F. Wheeler, P. L. Stoffa, and M. K. Sen. On optimization algorithms for the reservoir oil well placement problem. *Computational Geosciences*, 10(3):303–319, 2006.
11. K. Barty, J.-S. Roy, and C. Strugarek. Hilbert-valued perturbed subgradient algorithms. *Mathematics of Operations Research*, 32(3):551–562, 2007.
12. H. H. Bauschke and P. L. Combettes. *Convex Analysis and Montone Operator Theory in Hilbert Space*. CMS Books in Mathematics. Springer New York, 2011.
13. R. E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.
14. A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, 2009.
15. A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.
16. A. Ben-Tal and M. Teboulle. Penalty functions and duality in stochastic programming via phi-divergence functionals. *Mathematics of Operations Research*, 12:224–240, 1987.
17. A. Ben-Tal and M. Teboulle. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007.
18. J. O. Berger. The robust Bayesian viewpoint (with discussion). *Robustness of Bayesian Analysis*, pages 63–124, 1985.
19. J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer, 1985.
20. J. O. Berger. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994.
21. J. G. Berryman and G. W. Milton. Microgeometry of random composites and porous media. *Journal of Physics D: Applied Physics*, 21(1):87, 1988.
22. D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, New York,. London, Paris, San Diego, San Francisco, 1982.
23. D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.

24. D. Bertsimas and J. Sethuraman. Moment problems and semidefinite optimization. In H. Wolkowicz, R. Saigal, and L. Vandenberghe, editors, *Handbook of Semidefinite Programming*, pages 469–510. Kluwer Academic Publishers, 2000.
25. J. R. Birge and F. Louveaux. *Introduction to stochastic programming*. Springer-Verlag, New York, 1997.
26. J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Verlag, Berlin, Heidelberg, New York, 2000.
27. A. Borzì and G. von Winckel. A POD framework to determine robust controls in PDE optimization. *Comput. Vis. Sci.*, 14:91–103, 2011.
28. S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer Verlag, Berlin, Heidelberg, New York, second edition, 2002.
29. P. Cheridito and T. Li. Risk measures on Orlicz hearts. *Mathematical Finance*, 19(2):189–214, 2009.
30. F. H. Clarke. *Nonsmooth Analysis and Control Theory*. Graduate Texts in Mathematics. Springer, 1998.
31. A. Cohen, R. DeVore, and C. Schwab. Convergence rates of best n-term Galerkin approximations for a class of elliptic sPDEs. *Foundations of Computational Mathematics*, 10(6):615–646, 2010.
32. A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust–Region Methods*. SIAM, Philadelphia, 2000.
33. J. B. Conway. *A Course in Functional Analysis*. Graduate Texts in Mathematics. Springer New York, 1985.
34. I. Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *Magyar. Tud. Akad. Mat. Kutato Int. Kozls*, 8, 1063.
35. A. Defant and K. Floret. *Tensor Norms and Operator Ideals*. North-Holland Mathematics Studies. Elsevier Science, 1993.
36. E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58:595–6127, 2010.
37. D. Dentcheva, S. Penev, and A. Ruszczyński. Kusuoka representation of higher order dual risk measures. *Annals of Operations Research*, 181(1):325–335, 2010.
38. D. Dentcheva and A. Ruszczyński. Optimization with stochastic dominance constraints. *SIAM Journal on Optimization*, 14(2):548–566, 2003.
39. I. T. Dimov. *Monte Carlo methods for applied scientists*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2008.
40. O. Dorn and R. Villegas. History matching of petroleum reservoirs using a level set technique. *Inverse Problems*, 24(3):035015, 2008.
41. J. Dupačová. Uncertainties in minimax stochastic programs. *Optimization*, 60(10–11):1235–1250, 2011.
42. J. Eckstein and D. P. Bertsekas. On the Douglas—Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1):293–318, Apr 1992.
43. Y. M. Ermoliev and A. A. Gaivoronski. Stochastic methods for solving minimax problems. *Cybernetics*, 19(4):550–559, 1983.
44. Y. M. Ermoliev, A. A. Gaivoronski, and C. Nedeva. Stochastic optimization problems with incomplete information on distribution functions. *SIAM Journal on Control and Optimization*, 23(5):697–716, 1985.
45. G. B. Folland. *Real analysis. Modern techniques and their applications*. Pure and Applied Mathematics (New York). John Wiley & Sons Inc., New York, second edition, 1999.
46. A. A. Gaivoronski. A numerical method for solving stochastic programming problems with moment constraints on a distribution function. *Annals of Operations Research*, 31(1):347–369, 1991.
47. S. Garreis and M. Ulbrich. Constrained optimization with low-rank tensors and applications to parametric problems with PDEs. *SIAM Journal on Scientific Computing*, 39(1):A25–A54, 2017.

48. T. Gerstner and M. Griebel. Numerical integration using sparse grids. *Numer. Algorithms*, 18(3–4):209–232, 1998.
49. T. Gerstner and M. Griebel. Dimension-adaptive tensor-product quadrature. *Computing*, 71(1):65–87, 2003.
50. M. Grigoriu. Reduced order models for random functions. application to stochastic problems. *Applied Mathematical Modelling*, 33(1):161–175, 2009.
51. M. Grigoriu. A method for solving stochastic equations by reduced order models and local approximations. *Journal of Computational Physics*, 231(19):6495–6513, 2012.
52. V. Hauk. *Structural and Residual Stress Analysis by Nondestructive Methods: Evaluation - Application - Assessment*. Elsevier Science, 1997.
53. E. Hille and R. S. Phillips. *Functional analysis and semi-groups*. American Mathematical Society Colloquium Publications, vol. 31. American Mathematical Society, Providence, R. I., 1957. rev. ed.
54. K. Ito and K. Kunisch. *Lagrange Multiplier Approach to Variational Problems and Applications*. Society for Industrial and Applied Mathematics, 2008.
55. P. Kall and S. W. Wallace. *Stochastic Programming*. Wiley, Chichester etc., 1994.
56. S. Kalpakjian and S. R. Schmid. *Manufacturing Engineering and Technology*. Prentice Hall, 2010.
57. K. Karhunen. Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys.*, 1947(37):79, 1947.
58. G. E. Karniadakis, C.-H. Su, D. Xiu, D. Lucor, C. Schwab, and R. A. Todor. Generalized polynomial chaos solution for differential equations with random inputs. Technical Report 2005–01, Seminar for Applied Mathematics, ETH Zurich, Zurich, Switzerland, 2005.
59. B. Khoromskij and C. Schwab. Tensor-structured Galerkin approximation of parametric and stochastic elliptic PDEs. *SIAM J. Sci. Comput.*, 33(1):364–385, 2011.
60. D. P. Kouri. A multilevel stochastic collocation algorithm for optimization of PDEs with uncertain coefficients. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):55–81, 2014.
61. D. P. Kouri, M. Heinkenschloss, D. Ridzal, and B. G. van Bloemen Waanders. A trust-region algorithm with adaptive stochastic collocation for PDE optimization under uncertainty. *SIAM Journal on Scientific Computing*, 35(4):A1847–A1879, 2013.
62. D. P. Kouri, M. Heinkenschloss, D. Ridzal, and B. G. van Bloemen Waanders. Inexact objective function evaluations in a trust-region algorithm for PDE-constrained optimization under uncertainty. *SIAM Journal on Scientific Computing*, 36(6):A3011–A3029, 2014.
63. D. P. Kouri and T. M. Surowiec. Existence and optimality conditions for risk-averse PDE-constrained optimization. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):787–815, 2018.
64. D. P. Kouri and T. M. Surowiec. Risk-averse PDE-constrained optimization using the conditional value-at-risk. *SIAM Journal on Optimization*, 26(1):365–396, 2016.
65. J. R. Krebs, J. E. Anderson, D. Hinkley, R. Neelamani, S. Lee, A. Baumstein, and M. D. Lacasse. Fast full-waveform seismic inversion using encoded sources. *Geophysics*, 74(6):177–188, 2009.
66. P. A. Krokhmal. Higher moment coherent risk measures. *Quantitative Finance*, 7(4):373–387, 2007.
67. B. Lazarov, M. Schevenels, and O. Sigmund. Topology optimization considering material and geometric uncertainties using stochastic collocation methods. *Structural and Multidisciplinary Optimization*, pages 1–16, 2012. online first.
68. O. P. Le Maitre and O. M. Knio. *Spectral Methods for Uncertainty Quantification With Applications to Computational Fluid Dynamics*. Scientific Computation. Springer-Verlag, Berlin, 2010.
69. M. Loève. *Probability theory. II*. Graduate Texts in Mathematics, Vol. 46. Springer-Verlag, New York, fourth edition, 1978.
70. D. Love and G. Bayraksan. Phi-divergence constrained ambiguous stochastic programs. Technical report, Technical report, Program in Applied Mathematics, University of Arizona, 2013.

71. A. Mafusalov and S. Uryasev. Buffered probability of exceedance: mathematical properties and optimization. *SIAM Journal on Optimization,* 28(2):1077–1103, 2018.
72. M. M. Mäkelä and N. Neittaanmäki. *Nonsmooth Optimization: Analysis And Algorithms With Applications To Optimal Control*. World Scientific Publishing Company, 1992.
73. E. M. Makhlouf, W. H. Chen, M. L. Wasserman, and J. H. Seinfeld. A general history matching algorithm for three-phase, three-dimensional petroleum reserviors. *Society of Petroleum Engineers*, 1(2), 1993.
74. H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):pp. 77–91, 1952.
75. K. Marti, editor. *Stochastic Optimization. Numerical Methods and Technical Applications*. Springer, Berlin, 1992. LN in Economics and Math. Systems 379.
76. K. Marti. Differentiation formulas for probability functions: The transformation method. *Mathematical Programming*, 75:201–220, 1996.
77. K. Maute. *Topology Optimization under Uncertainty*, pages 457–471. Springer Vienna, Vienna, 2014.
78. K. Maute and D. M. Frangopol. Reliability-based design of mems mechanisms by topology optimization. *Computers & Structures*, 81(8–11):813–824, 2003.
79. T. Morimoto. Markov processes and the h-theorem. *J. Phys. Soc. Jap.*, 18(3):328–333, 1963.
80. A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.
81. A. Nemirovski and A. Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2007.
82. A. Nemirovski and D. Yudin. On Cezari's convergence of the steepest descent method for approximating saddle point of convex-concave functions. *Soviet Math. Dokl.*, 239:1056–1059, 1978.
83. F. Nobile, R. Tempone, and C. G. Webster. An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM J. Numer. Anal.*, 46(5):2411–2442, 2008.
84. F. Nobile, R. Tempone, and C. G. Webster. A sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM Journal on Numerical Analysis*, 46(5):2309–2345, 2008.
85. E. Novak and K. Ritter. High-dimensional integration of smooth functions over cubes. *Numer. Math.*, 75(1):79–97, 1996.
86. E. Novak and K. Ritter. Simple cubature formulas with high polynomial exactness. *Constr. Approx.*, 15(4):499–522, 1999.
87. B.K. Pagnoncelli, S. Ahmed, and A. Shapiro. Sample average approximation method for chance constrained programming: theory and applications. *J. Optim. Theory Appl.*, 142(2):399–416, 2009.
88. J. S. Pang and S. Leyffer. On the global minimization of the value-at-risk. *Optimization Methods and Software*, 19(5):611–631, 2004.
89. B.T. Polyak. New method of stochastic approximation type. *Automat. Remote Control*, 51:937–946, 1990.
90. A. Prékopa. Probabilistic programming. In *Stochastic programming*, volume 10 of *Handbooks Oper. Res. Management Sci.*, pages 267–351. Elsevier Sci. B. V., Amsterdam, 2003.
91. H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 9 1951.
92. R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.
93. R. T. Rockafellar and J. O. Royset. On buffered failure probability in design and optimization of structures. *Reliability Engineering & System Safety*, 95(5):499–510, 2010.
94. R. T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.

95. R. T. Rockafellar and S. Uryasev. The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science*, 18(1–2):33–53, 2013.
96. R. T. Rockafellar and Roger J.-B. Wets. Scenarios and policy aggregation in optimization under uncertainty. *Math. Oper. Res.*, 16(1):119–147, 1991.
97. W. W. Rogosinski. Moments of non-negative mass. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 245(1240):1–27, 1958.
98. J. O. Royset and E. Polak. Extensions of stochastic optimization results to problems with system failure probability functions. *Journal of Optimization Theory and Applications*, 133(1):1–18, 2007.
99. A. Ruszczyński and A. Shapiro. Optimization of risk measures. In G. Calafiore and F. Dabbene, editors, *Probabilistic and Randomized Methods for Design Under Uncertainty*, pages 119–157, London, 2006. Springer Verlag.
100. R. A. Ryan. *Introduction to tensor products of Banach spaces*. Springer Monographs in Mathematics. Springer-Verlag London Ltd., London, 2002.
101. F. Santosa and W. W. Symes. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1307–1330, 1986.
102. P. Sarma, L. J. Durlofsky, K. Aziz, and W. H. Chen. Efficient real-time reservoir management using adjoint-based optimal control and model updating. *Computational Geosciences*, 10(1):3–36, 2006.
103. H. Scarf. A min-max solution of an inventory problem. In *Studies in the Mathematical Theory of Inventory and Production*, pages 201–209. Stanford University Press, 1958.
104. C. Schwab and C. J. Gittelson. Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs. *Acta Numer.*, 2011:291–467, 2011.
105. A. Shapiro. On concepts of directional differentiability. *J. Optim. Theory Appl.*, 66(3):477–487, 1990.
106. A. Shapiro. Monte Carlo sampling methods. In A. Ruszczynski and A. Shapiro, editors, *Stochastic Programming*, Handbooks in Operations Research and Management Science, Vol. 10, pages 353–425. Elsevier, 2003.
107. A. Shapiro. Distributionally robust stochastic programming. *SIAM J. Optimization*, 27(4):2258–2275, 2017.
108. A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory, Second Edition*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, Philadelphia, 2014.
109. O. Sigmund. Manufacturing tolerant topology optimization. *Acta Mechanica Sinica*, 25(2):227–239, 2009.
110. S. A. Smoljak. Quadrature and interpolation formulae on tensor products of certain function classes. *Soviet Math. Dokl.*, 4:240–243, 1963.
111. W. W. Symes and J. J. Carazzone. Velocity inversion by differential semblance optimization. *Geophysics*, 56(5):654–663, 1991.
112. H. Tiesler, R. M. Kirby, D. Xiu, and T. Preusser. Stochastic collocation for optimal control problems with stochastic PDE constraints. *SIAM Journal on Control and Optimization*, 50(5):2659–2682, 2012.
113. S. Uryasev. Derivatives of probability functions and integrals over sets given by inequalities. *J. Comput. Appl. Math.*, 56(1–2):197–223, 1994. Stochastic programming: stability, numerical methods and applications (Gosen, 1992).
114. S. Uryasev. Derivatives of probability functions and some applications. *Ann. Oper. Res.*, 56:287–311, 1995. Stochastic programming (Udine, 1992).
115. S. Uryasev and R. T. Rockafellar. Conditional value-at-risk: Optimization approach. In S. Uryasev and P. M. Pardalos, editors, *Stochastic optimization: algorithms and applications. Papers from the conference held at the University of Florida, Gainesville, FL, February 20–22, 2000*, volume 54 of *Appl. Optim.*, pages 411–435. Kluwer Acad. Publ., Dordrecht, 2001.

116. M. M. Vainberg. *Variational methods for the study of nonlinear operators*. Holden-Day, Inc., San Francisco, Calif.-London-Amsterdam, 1964. With a chapter on Newton's method by L. V. Kantorovich and G. P. Akilov. Translated and supplemented by Amiel Feinstein.
117. W. van Ackooij and R. Henrion. (Sub-)gradient formulae for probability functions of random inequality systems under Gaussian distribution. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):63–87, 2017.
118. B. van den Bosch and J. H. Seinfeld. History matching in two-phase petroleum reserviors: Incompressible flow. *Society of Petroleum Engineers*, 17(6), 1977.
119. G. van Essen, M. Zandvliet, P. van den Hof, O. Bosgra, and J. D. Jansen. Robust waterflooding optimization of multiple geological scenarios. *Society of Petroleum Engineers*, 14(1), 2009.
120. J. E. Warner, M. D. Grigoriu, and W. Aquino. Stochastic reduced order models for random vectors: Application to random eigenvalue problems. *Probabilistic Engineering Mechanics*, 31:1–11, 2013.
121. W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.
122. D. Xiu and G. E. Karniadakis. Modeling uncertainty in flow simulations via generalized polynomial chaos. *J. Comput. Phys.*, 187(1):137–167, 2003.

Inexact Trust-Region Methods for PDE-Constrained Optimization

Drew P. Kouri and Denis Ridzal

Abstract Numerical solution of optimization problems with partial differential equation (PDE) constraints typically requires inexact objective function and constraint evaluations, derivative approximations, and the use of iterative linear system solvers. Over the last 30 years, trust-region methods have been extended to rigorously, robustly, and efficiently handle various sources of inexactness in the optimization process. In this chapter, we review some of the recent advances, discuss their key algorithmic contributions, and present numerical examples that demonstrate how inexact computations can be exploited to enable the solution of large-scale PDE-constrained optimization problems.

1 Introduction

Numerical solution of optimization problems constrained by partial differential equations (PDEs)—and, more generally, optimization problems involving large-scale nonlinear simulations—poses a number of mathematical, algorithmic, and computational challenges. The computational challenge often lies in the sheer size of the discretized problem. Specifically, the computational expense of solving a single instance of the governing PDEs can make the solution of the optimization problem a daunting task. To make numerical solution practical, one frequently resorts to approximating the objective function and its derivatives. Similarly, one may use approximations of the constraint function, its derivatives, and their inverses. In order to ensure convergence to a solution of the original infinite-dimensional problem, however, these approximations must be intelligently managed and refined. Trust-region methods provide a robust, globally convergent framework to handle

D. P. Kouri · D. Ridzal (✉)
Center for Computing Research, Sandia National Laboratories, Albuquerque, NM 87185-9999, USA
e-mail: dpkouri@sandia.gov; dridzal@sandia.gov

H. Antil et al. (eds.), *Frontiers in PDE-Constrained Optimization*, The IMA Volumes in Mathematics and its Applications 163,
https://doi.org/10.1007/978-1-4939-8636-1_3

multiple forms of inexactness, including inexact evaluations of the objective and constraint functions and their derivatives, as well as the inexact linear system solves arising in the approximate application of constraint derivative inverses. For this reason, trust regions are a popular choice for large-scale, nonconvex multidisciplinary optimization with simulation constraints.

Early works by Moré [29], Toint [40], and Carter [10–12] pioneered the use of inexact gradients and objective function values within trust-region methods. Later in the 1990s, Alexandrov, Dennis, and Torczon analyzed trust-region algorithms as a general framework for managing approximations throughout the optimization iteration [1–4, 15, 16]. These works laid the foundation for more recent works on adaptive approximations, where the trust-region framework is used to manage the accuracy of objective and constraint function evaluations, objective gradient computations, constraint derivative operator applications, and linear system solves [5, 8, 18, 20, 21, 24–26, 43, 44]—collectively labeled *inexact trust-region methods*.

In this chapter, we review two inexact trust-region algorithms, for unconstrained and equality-constrained optimization, respectively, using PDE-constrained optimization as motivation. For each algorithm, we discuss the state of the art in terms of variable-fidelity and inexact computations. Our goal is threefold: to document the algorithms in sufficient, "ready-to-use" detail, demonstrate them on a novel numerical example, and provide a concise reference to other recent works on managing inexactness in trust-region methods for large-scale optimization. The remainder of the chapter is structured as follows. In Section 2, we introduce the notation. In Section 3, we discuss two common problem formulations used in PDE-constrained optimization, the reduced-space and the full-space formulation. In Section 4, we review an inexact trust-region algorithm for unconstrained optimization and an inexact composite-step trust-region algorithm for equality-constrained optimization. Subsequently, in Section 5 we specialize the algorithms to optimization problems constrained by PDEs with random coefficients. The discussion of full-space methods includes a novel highly parallelizable preconditioner for optimization under uncertainty where the statistics are evaluated through sampling, e.g., using sparse grids [37] or Monte-Carlo methods. The preconditioner is an extension of recent work on constraint-based "optimal" preconditioners for PDE-constrained optimization [33] and their specialization to augmented systems [36]. In Section 6, we present new numerical results for the risk-neutral optimization of thermal-fluid models with random inputs. Finally, we discuss our conclusions in Section 7.

2 Notation

The norm associated with the Banach space V is $\|\cdot\|_V$. When V is Hilbert, we denote the inner product by $\langle v, w\rangle_V$ and the norm $\|v\|_V = \sqrt{\langle v, v\rangle_V}$. If V and W are Banach spaces then we denote the space of bounded linear operators between V and W by $\mathcal{L}(V, W)$. When $V = W$, we denote $\mathcal{L}(V) := \mathcal{L}(V, V)$. Moreover, when $W = \mathbb{R}$, we denote by $V^* := \mathcal{L}(V, \mathbb{R})$ the topological dual space of V and

by $\langle f, v\rangle_{V^*,V}$ the duality pairing between $f \in V^*$ and $v \in V$. If $A \in \mathcal{L}(V, W)$, we denote its adjoint by $A^* \in \mathcal{L}(W^*, V^*)$. When V is a Hilbert space, we identify V^* with V using the Riesz Representation Theorem. We note that the algorithms described in this paper can be equivalently formulated using the notion of the Riesz isomorphism $R \in \mathcal{L}(V^*, V)$. Finally, we denote the identity operator on the Banach space V by $I_V \in \mathcal{L}(V)$.

3 Problem Formulations

An important feature of algorithms for PDE-constrained optimization is convergence in function space, leading to mesh-independent convergence after discretization. Therefore, we consider the function space setting for the formulation of PDE-constrained optimization problems. We will discuss their discrete forms, enabling numerical solution, in Section 5. Let U and Z be Hilbert spaces and C be a reflexive Banach space. U denotes the *state space* and $u \in U$ is a *state variable*. Similarly, Z is the *control space* and $z \in Z$ is a *control variable*. We wish to solve the optimization problem

$$\min_{u\in U,\ z\in Z} \quad J(u, z) \tag{1a}$$

$$\text{subject to} \quad c(u, z) = 0\,, \tag{1b}$$

where the objective $J : U \times Z \to \mathbb{R}$ and the constraint $c : U \times Z \to C$ are smooth functions, with the smoothness requirements made precise later.

In addition to the optimization problem (1), which we refer to as the *full-space* problem, we will consider its *reduced-space* companion. In general, the reduced-space problem is obtained through nonlinear elimination of the state variables. Specifically, we assume the existence of the *solution operator* $S : Z \to U$ such that

$$u = S(z) \quad \text{satisfies} \quad c(u, z) = 0\,,$$

and define the *reduced objective function* $\mathcal{J} : Z \to \mathbb{R}$ by

$$\mathcal{J}(z) := J(S(z), z)\,.$$

Instead of the optimization problem (1), we can then solve

$$\min_{z\in Z} \quad \mathcal{J}(z)\,. \tag{2}$$

We note that in general (1) and (2) are not equivalent. For example, problem (1) may have a solution even when the solution operator S does not exist, i.e., when problem (2) cannot be solved. Additionally, we will see later that certain types of inexactness in J, c, and their partial derivatives are more easily handled through the

reduced-space form, while others are better suited to the full-space form. Finally, we consider a third problem formulation, resulting from a simple change of notation in problem (1). We define $X := U \times Z$, and for $x \in X$ write

$$\min_{x \in X} \quad J(x) \tag{3a}$$

$$\text{subject to} \quad c(x) = 0\,. \tag{3b}$$

This formulation is a generalization of (1), in that it does not assume an explicit splitting of variables into state and control variables, potentially resulting in algorithmic advantages.

When discussing algorithms for the full-space problem (1), we require the Lagrangian functional

$$L(u, z, \lambda) := J(u, z) + \langle \lambda, c(u, z) \rangle_{C^*, C}.$$

Furthermore, under standard assumptions, the reduced objective function $\mathcal{J}$ of (2) is twice continuously Fréchet differentiable and the first derivative is given by

$$\nabla \mathcal{J}(z) = c_z(S(z), z)^* \Lambda(S(z), z) + \nabla_z J(S(z), z) \in Z$$

where $\Lambda(u, z) = \lambda \in C^*$ solves the adjoint equation

$$c_u(u, z)^* \lambda = -\nabla_u J(u, z). \tag{4}$$

Here, we denote the partial derivatives of c by c_u and c_z, and the partial derivatives of J by $\nabla_u J$ and $\nabla_z J$. A similar expression exists for the application of the Hessian of $\mathcal{J}$ to a vector $v \in Z$. In particular, we can represent the action of the Hessian by

$$\nabla^2 \mathcal{J}(z) v = T(S(z), z)^* H(S(z), z, \Lambda(S(z), z)) T(S(z), z) v \tag{5}$$

where the linear operator $T(u, z)$ is defined as

$$T(u, z) := \begin{pmatrix} -c_u(u, z)^{-1} c_z(u, z) \\ I_Z \end{pmatrix},$$

and the linear operator $H(u, z, \lambda)$ is defined as

$$H(u, z, \lambda) := \begin{pmatrix} \nabla_{uu} L(u, z, \lambda) & \nabla_{uz} L(u, z, \lambda) \\ \nabla_{zu} L(u, z, \lambda) & \nabla_{zz} L(u, z, \lambda) \end{pmatrix},$$

see, e.g., [22, Ch. 1]. As above, $\nabla_{uu} L$, $\nabla_{uz} L$, $\nabla_{zu} L$, and $\nabla_{zz} L$ denote the second-order partial derivatives of the Lagrangian functional.

Formulations (1), (2), and (3) of PDE-constrained optimization problems have been studied extensively in the context of inexact trust-region methods. Before

discussing the methods, we present a "menu" of possible assumptions on the full-space formulation (1), which are used to establish the applicability and global convergence of each of the methods. We note that some assumptions are shared by all methods, while others are formulation-specific and method-specific. Throughout, we assume the existence of a convex open set $\Omega \subseteq X$ such that the iterates x_k and trial steps s_k produced by the algorithms described in the subsequent sections satisfy x_k, $x_k + s_k \in \Omega$ for all k. For the reduced space formulation, we interpret $x_k = (S(z_k), z_k)$ and assume $(S(z_k + t\sigma_k), z_k + t\sigma_k) \in \Omega$ for all $t \in [0, 1]$ where $\sigma_k \in Z$ denotes the trial step produced by the reduced-space algorithm.

(A1) The functional J is bounded below and finite on Ω.
(A2) The functions J and c are twice continuously Fréchet differentiable on Ω.
(A3) The Jacobian $c_x(x)$ is uniformly bounded on Ω.
(A4) The Hessians $\nabla_{xx} J(x)$ and $c_{xx}(x)$ are uniformly bounded in $\mathcal{L}(X)$ and $\mathcal{L}(X, \mathcal{L}(X, C))$, respectively, on Ω.
(A5) The operator $c_x(x)$ is surjective for all $x \in \Omega$.
(A6) The state Jacobian $c_u(x)$ is continuously invertible for all $x \in \Omega$ and the inverse $c_u(x)^{-1}$ is uniformly bounded on Ω.
(A7) The solution operator S exists and is unique for all $z \in Z$.
(A8) The operator $T(x)^* H(x, \Lambda(x)) T(x)$ is uniformly bounded on Ω.
(A9) The following function(al)s and operators are uniformly bounded over all $x \in \Omega$: $J(x)$, $\nabla_x J(x)$, $c(x)$, and $(c_x(x) c_x(x)^*)^{-1}$.

Remark 1 When the PDE in (1) is nonlinear (e.g., semilinear), the range space of the solution operator S typically must be more regular than the space U, in order to ensure that c is Fréchet differentiable and that $c_u(u, z)$ has a bounded inverse. Although this is an important issue, we focus on the stated problem setting to simplify the presentation. We refer the interested reader to [41] for information on handling this more general setting.

4 Inexact Trust-Region Methods

We begin with the description of an inexact trust-region approach for the reduced-space formulation, (2), which was studied in [25, 26]. This is followed by the discussion of a scheme for the full-space formulation (3), originally presented in [20].

4.1 A Reduced-Space Approach

In this section, we focus on the reduced-space formulation, (2), of PDE-constrained optimization problems. Given an iterate z_k, the basic trust-region algorithm builds a smooth local model $m_k : Z \to \mathbb{R}$ of the objective function $s \mapsto \mathcal{J}(z_k + s)$ inside

the trust region $\mathcal{B}_k := \{s \in Z : \|s\| \leq \Delta_k\}$, where $\Delta_k > 0$ is the trust-region radius. The algorithm then computes a trial step s_k by approximately solving the trust-region subproblem

$$\min_{s \in Z} m_k(s) \quad \text{subject to} \quad \|s\|_Z \leq \Delta_k. \tag{6}$$

In order to ensure convergence, the trial step s_k must satisfy the fraction of Cauchy decrease condition

$$m_k(0) - m_k(s_k) \geq \kappa_{\text{fcd}} \|\nabla m_k(0)\|_Z \min\left\{\Delta_k, \frac{\|\nabla m_k(0)\|_Z}{\beta_k}\right\}, \tag{7}$$

where $\kappa_{\text{fcd}} > 0$ is fixed and $\beta_k = 1 + \sup_{s \in \mathcal{B}_k} \|\nabla^2 m_k(s)\|_{\mathcal{L}(Z)}$. When m_k in (6) is quadratic, a number of solvers exist to compute s_k that satisfies (7). For example, the Cauchy point, (double) dogleg, truncated Conjugate Gradient and truncated Lanczos algorithms all produce steps that satisfy (7), see [13, 17] and the references within for more information. Once the step s_k is computed, trust-region algorithms check whether $z_k + s_k$ is acceptable as the new iterate, and the trust-region radius Δ_k is updated accordingly. Step acceptance and the trust-region update depend on the ratio of *actual* and *predicted reduction*,

$$\text{ared}_k := (\mathcal{J}(z_k) - \mathcal{J}(z_k + s_k)) \quad \text{and} \quad \text{pred}_k := (m_k(0) - m_k(s_k)),$$

respectively. In particular, given $0 < \eta_1 < \eta_2 < 1$, the trial step s_k is accepted if the actual reduction is larger than a fraction of the predicted reduction, i.e.,

$$\text{ared}_k \geq \eta_1 \text{pred}_k.$$

If the trial step is rejected, then the trust-region radius is decreased, whereas if it is accepted and the actual reduction is sufficiently large, i.e.,

$$\text{ared}_k \geq \eta_2 \text{pred}_k,$$

then the trust-region radius is increased.

Under standard assumptions, if the objective function can be evaluated exactly for all k, the model m_k is first-order consistent with $\mathcal{J}$ in the sense that

$$\nabla \mathcal{J}(z_k) = \nabla m_k(0)$$

for all k and the trial steps s_k satisfy condition (7), then one can prove global convergence of the trust-region scheme. However, if the objective function and its gradient can only be approximated, i.e., are evaluated inexactly, additional conditions are needed to ensure global convergence. We first state the condition on gradient inexactness.

Gradient Condition For all k the model m_k must approximate the objective function $s \mapsto \mathcal{J}(z_k + s)$ so that the true and approximate gradients at $s = 0$ satisfy

$$\|\nabla m_k(0) - \nabla \mathcal{J}(z_k)\|_Z \leq \kappa_{\text{grad}} \min\{\|\nabla m_k(0)\|_Z, \Delta_k\}. \tag{8}$$

Here, $\kappa_{\text{grad}} > 0$ is independent of k. A similar condition was originally proposed by Carter in [11]. However, (8) is due to Heinkenschloss and Vicente [21]. □

The definition of the actual reduction, ared_k, involves the exact value of the objective function $\mathcal{J}$, which we often cannot compute. Instead, we consider a computable approximation $\mathcal{J}_k$, where the subscript k indicates that the approximation may change from iteration to iteration. With this approximation, we can define the *computed reduction*

$$\text{cred}_k := \mathcal{J}_k(z_k) - \mathcal{J}_k(z_k + s_k). \tag{9}$$

To ensure convergence of the trust-region algorithm, we must ensure that the difference

$$|\text{ared}_k - \text{cred}_k|,$$

is "sufficiently" small. We now state the objective function conditions.

Objective Function Conditions Assume that there exists an estimator $\theta_k = \theta(z_k, s_k)$ of the error in the objective function so that for a constant $K > 0$,

$$|\text{ared}_k - \text{cred}_k| \leq K\theta_k \quad \forall k\,. \tag{10a}$$

For a fixed $\omega \in (0, 1)$, we control the error estimator θ_k via the following bound:

$$\theta_k^{\omega} \leq \eta \min\left\{\text{pred}_k, r_k\right\}, \tag{10b}$$

where

$$\eta < \min\{\eta_1, 1 - \eta_2\} \quad \text{and} \quad \{r_k\}_{k=1}^{\infty} \subset [0, \infty) \text{ satisfies } \lim_{k\to\infty} r_k = 0. \tag{10c}$$

Condition (10b) is due to Ziems and Ulbrich [43]; also see [13, Sec. 10.6]. □

The basic trust-region algorithm, accounting for inexact objective function and gradient evaluation, is listed as Algorithm 1.

To prove convergence of the inexact unconstrained trust-region algorithm, we will use some of the problem assumptions stated in Section 3 and the following model assumptions:

(R1) For each k, $m_k : Z \to \mathbb{R}$ is twice continuously Fréchet differentiable.
(R2) For each k, $\nabla^2 m_k$ is uniformly bounded on Z.
(R3) For each k, the objective function approximation $\mathcal{J}_k$ is bounded below.

Algorithm 1 (Reduced-space trust-region algorithm)

Initialization: *Choose initial point* z_0, *initial trust-region radius* $\Delta_0 > 0$, *constants* $0 < \gamma_1 \leq \gamma_2 < 1 < \gamma_3$, $0 < \eta_1 < \eta_2 < 1$, *and* $tol > 0$.

For *k=0,1,2,...*

1. ***Model selection:*** *Choose a model* m_k *that satisfies* (8).
2. ***Convergence check:*** *If* $\|\nabla m_k(0)\|_Z < tol$, *then terminate.*
3. ***Step computation:*** *Compute an approximate solution* s_k *of* (6) *that satisfies the fraction of Cauchy decrease condition* (7).
4. ***Objective function update:*** *Determine an objective function approximation* J_k *such that the corresponding error estimate* θ_k *satisfies* (10).
5. ***Step acceptance:*** *Compute* $\varrho_k = \text{cred}_k/\text{pred}_k$.
   if $\varrho_k \geq \eta_1$ ***then*** $z_{k+1} = z_k + s_k$ ***else*** $z_{k+1} = z_k$
6. ***Trust-region update:***
   if $z_{k+1} = z_k$ ***then*** $\Delta_{k+1} \in (0, \gamma_1 \|s_k\|_Z]$
   else

$$\Delta_{k+1} \in \begin{cases} (0, \gamma_2 \|s_k\|_Z] & \textbf{if } \rho_k \leq \eta_1 \\ [\gamma_2, \|s_k\|_Z, \Delta_k] & \textbf{if } \rho_k \in (\eta_1, \eta_2) \\ [\Delta_k, \gamma_3 \Delta_k] & \textbf{if } \rho_k \geq \eta_2 \end{cases}$$

End For

Theorem 1 *Let* $\Omega = U \times Z$. *If problem assumptions (A1), (A2), (A6), (A7) and (A8), and model assumptions (R1), (R2) and (R2) hold, then the iterates* $\{z_k\}$ *generated by the inexact unconstrained trust-region algorithm, Algorithm 1, satisfy*

$$\liminf_{k\to\infty} \|\nabla m_k(0)\|_Z = \liminf_{k\to\infty} \|\nabla \mathcal{J}(z_k)\|_Z = 0.$$

Proof Assumptions (A2), (A6), and (A7) together with the Implicit Function Theorem [22, Th. 1.41] ensure that $\mathcal{J}$ is twice continuously Fréchet differentiable. Moreover, (A8) ensures that the Hessian of $\mathcal{J}$ is uniformly bounded at $z_k + ts_k$ for all $t \in [0, 1]$ and for all k. The desired result then follows from assumptions (R1), (R2), and (R3), and a slight generalization of Theorem 5.6 in [26].

4.1.1 Related Work

The above approach is described in detail in [25, 26]. The authors in [25, 26] combine and modify conditions for inexact gradient and objective function values from a number of sources. For example, Moré considers inexact gradients in [29] for the case of $Z = \mathbb{R}^n$. In this case, he requires

$$z_k \to z \quad \Longrightarrow \quad \lim_{k\to\infty} \|\nabla m_k(0) - \nabla \mathcal{J}(z_k)\|_Z = 0. \tag{11}$$

Similarly, in [40] Toint analyzes an algorithm in Hilbert space for bound-constrained optimization and requires

$$\|\nabla m_k(0) - \nabla \mathcal{J}(z_k)\|_Z \leq \min\{\kappa_1, \kappa_2 \Delta_k\}$$

for appropriately chosen $\kappa_1,\ \kappa_2 > 0$. Carter, in [10], proves global convergence of Algorithm 1 using the inexactness conditions

$$|\text{ared}_k - \text{cred}_k| \leq \zeta_{f,1} \text{pred}_k$$
$$|\text{ared}_k - \text{cred}_k| \leq \zeta_{f,2} |\text{cred}_k|$$
$$\langle \nabla m_k(0) - \nabla \mathcal{J}(z_k), \nabla m_k(0) \rangle_Z \leq \zeta_g \|\nabla m_k(0)\|_Z^2$$

for constants $\zeta_g,\ \zeta_{f,1}\ \zeta_{f,2} > 0$ satisfying

$$\zeta_g + \zeta_{f,1} < 1 - \eta_2 \quad \text{and} \quad \zeta_{f,2} < 1.$$

Carter further analyzes his approach in [11, 12]. In [13, Sec. 10.6], the authors consider objective functions with dynamic accuracy for which they require

$$\max\{|\mathcal{J}(z_k) - \mathcal{J}_k(z_k)|, |\mathcal{J}(z_k + s_k) - \mathcal{J}_k(z_k + s_k)|\} \leq \tilde{\eta}\ \text{pred}_k\,, \tag{12}$$

for some $\tilde{\eta} \leq \frac{1}{2}\eta_1$. The challenge with each of these conditions (other than Moré's very general requirement (11)) is that the constants, e.g., ζ_g, $\zeta_{f,1}$, $\zeta_{f,2}$ and $\tilde{\eta}$, depend explicitly on algorithmic parameters. Therefore, it is difficult to determine a priori if these conditions can be satisfied in practice. As a practical alternative, the inexact gradient conditions of Toint and Carter are combined by Heinkenschloss and Vicente in [21], giving rise to the condition (8), which permits an arbitrary constant scaling κ_{grad} on the error bound and enables easy implementation. In a similar vein, Ulbrich and Ziems in [43] relax the dependence of the inexact objective function condition (12) on algorithmic parameters, motivating the more practical conditions (10).

Kelley and Sachs in [24] take a different approach to that presented here. They work in the setting where the value and gradient approximations are provided by "black-box" calculations that satisfy controllable absolute and relative error tolerances. They suggest modifications to the basic trust-region algorithm so that the resulting algorithm performs as if there were no errors in the computation of the value and gradient.

The authors of [25, 26] apply Algorithm 1 to PDE-constrained optimization problems for which the governing PDE has uncertain coefficients. The inexactness conditions (8) and (10) are used to adaptively refine sparse-grid quadrature approximations of the objective function. In [5, 18], the authors employ the inexact gradient condition (8) to adaptively refine reduced-order models of the PDE constraint using proper orthogonal decomposition (POD). However, they evaluate the discretized

objective function exactly since the state equation must be solved to build the POD model of the state and adjoint variables. Similarly, the authors of [8] use inexact gradients to adaptively refine Monte Carlo sample sizes for mixed logit optimization.

4.2 A Full-Space Approach

In this section, we focus on the full-space formulation (3) of PDE-constrained optimization problems. We review a sequential quadratic programming (SQP) approach to solving (3), originating in the composite-step trust-region scheme of Byrd and Omojokun [32]. We assume that inexactness in solving (3) is due to the approximate solution of a variety of subproblems, such as quadratic optimization problems or linear systems, which comprise the SQP scheme. The handling of inexactness is based on [20].

Remark 2 To further simplify the presentation, in this section we assume that the constraint space C is a Hilbert space. As a reminder, we identify its dual space C^* with C. Finally, we note that a treatment of more general constraint spaces such as reflexive Banach spaces is possible in the context of full-space composite-step methods, see, e.g., [28].

Recall $X = U \times Z$, and write the Lagrangian functional $L : X \times C \to \mathbb{R}$ for (3),

$$L(x, \lambda) = J(x) + \langle \lambda, c(x) \rangle_C .$$

We let x_k be the k-th SQP iterate, λ_k the Lagrange multiplier estimate at x_k, and $B_k = B(x_k, \lambda_k)$ the Hessian $\nabla_{xx} L(x_k, \lambda_k)$ of the Lagrangian or a self-adjoint approximation thereof. Trust-region SQP methods compute an approximate solution of (3) by approximately solving a sequence of subproblems derived from

$$\min_s \quad \frac{1}{2} \langle B_k s, s \rangle_X + \langle \nabla_x L(x_k, \lambda_k), s \rangle_X + L(x_k, \lambda_k) \tag{13a}$$

$$\text{subject to} \quad c_x(x_k)s + c(x_k) = 0 \tag{13b}$$

$$\|s\|_X \leq \Delta_k. \tag{13c}$$

To deal with the possible incompatibility of the constraints (13b), (13c), we apply a composite-step approach, where the trial step s_k is computed as the sum of a quasi-normal step n_k and a tangential step t_k. The role of the quasi-normal step n_k is to reduce linear infeasibility. It is computed as an approximate solution of

$$\min_n \quad \|c_x(x_k)n + c(x_k)\|_C^2 \tag{14a}$$

$$\text{subject to} \quad \|n\|_X \leq \zeta \Delta_k, \tag{14b}$$

where $\zeta \in (0, 1)$ is a fixed constant. To ensure global convergence of the SQP algorithm, the quasi-normal step must satisfy two conditions.

Quasi-Normal Step Conditions The quasi-normal step, n_k, must satisfy the boundedness condition

$$\|n_k\|_X \leq \kappa_1 \|c(x_k)\|_C, \tag{15}$$

where $\kappa_1 > 0$ is independent of k, and the fraction of Cauchy decrease condition

$$\|c(x_k)\|_C^2 - \|c_x(x_k)n_k + c(x_k)\|_C^2 \geq \kappa_2 \|c(x_k)\|_C \min\{\kappa_3 \|c(x_k)\|_C, \Delta_k\}, \tag{16}$$

where $\kappa_2, \kappa_3 > 0$ are independent of k. These conditions on the quasi-normal step are derived by Dennis, El-Alem, and Maciel in [14]. They are adopted by Heinkenschloss and Vicente in [21]. A more restrictive version, requiring κ_2 and κ_3 to be in the interval $(0, 1)$, is used by Ziems and Ulbrich in [43]. □

To understand the computation of the tangential step, we consider subproblem (13), where we substitute the computed quasi-normal step, i.e., $s = t + n_k$:

$$\min_t \quad \frac{1}{2}\langle B_k(t + n_k), t + n_k\rangle_X + \langle \nabla_x L(x_k, \lambda_k), t + n_k\rangle_X \tag{17a}$$

$$\text{subject to} \quad c_x(x_k)t = 0 \tag{17b}$$

$$\|t + n_k\|_X \leq \Delta_k. \tag{17c}$$

To solve (17), one typically eliminates the constraints (17b) using a representation of the null space of $c_x(x_k)$. Several null-space representations can be considered in the development of algorithms. Let E be a Hilbert space and let $W_k : E \to X$ be a bounded linear operator such that

$$\text{Range}(W_k) = \text{Null}(c_x(x_k)).$$

For instance, under the assumption (A6) and recalling the notation from Section 3, one can use $E = Z$ and define

$$W_k = T(x_k) = \begin{pmatrix} -c_u(x_k)^{-1} c_z(x_k) \\ I_Z \end{pmatrix}.$$

This null-space representation is a natural choice for many PDE-constrained optimization problems. In the context of inexact trust-region methods, it is used by Heinkenschloss and Vicente [21] and Ziems and Ulbrich [43]. Specifically, the authors in [21, 43] study inexact applications of the operator $c_u(x_k)^{-1}$. A more general alternative, not requiring assumption (A6), i.e., the invertibility of $c_u(x_k)$, is to choose $E = X$ and $W_k = W_k^* = W_k^2$, and to compute $t = W_k w$ by solving the so-called *augmented system*

$$\begin{pmatrix} I_X & c_x(x_k)^* \\ c_x(x_k) & 0 \end{pmatrix} \begin{pmatrix} t \\ z \end{pmatrix} = \begin{pmatrix} w \\ 0 \end{pmatrix}. \tag{18}$$

We note that in either case we can set $t = W_k w$ and replace (17) by

$$\min_w \quad \frac{1}{2} \langle W_k^* B_k W_k w, w \rangle_X + \langle W_k^* g_k, w \rangle_X \tag{19a}$$

$$\text{subject to} \quad \|n_k + W_k w\|_X \leq \Delta_k, \tag{19b}$$

where $g_k = \nabla_x L(x_k, \lambda_k) + B_k n_k$. A potential benefit of the second null-space representation, (18), is that the linear system can be solved using modern iterative saddle-point and Karush-Kuhn-Tucker (KKT) system solvers, see, e.g., [9, 33]. Moreover, these methods can be used to solve not only (18), but also the full KKT system, i.e., the system where I_X in (18) is replaced by B_k, thereby circumventing the need for preconditioning of the operator $W_k^* B_k W_k$, which is typically very challenging.

For the remainder of the chapter, we focus on the second null-space representation. A key source of inexactness in the application of the null-space operator W_k is the iterative solution of the linear system (18), using Krylov methods. In this case, the vector $W_k^* g_k$ and the operator $W_k^* B_k W_k$ are no longer available exactly. It is shown in [20] that the inexact solution of linear systems like (18) leads to an approximation $\widetilde{W}_k$ of W_k. While the operator W_k is self-adjoint (that is, assuming exact linear system solves), the approximation $\widetilde{W}_k$ is in general not self-adjoint and may not even be linear. Nonetheless, it is also shown, [20, p. 1525], that there exists a fixed linear operator that replicates the action of $\widetilde{W}_k$ on the set of vectors involved in the algorithm for solving (19). Therefore, an algorithm that computes a solution w_k of (19) with inexact applications of the operator W_k also solves

$$\min_w \quad \frac{1}{2} \langle \widetilde{W}_k^* B_k \widetilde{W}_k w, w \rangle_X + \langle \widetilde{W}_k^* \widetilde{W}_k g_k, w \rangle_X \tag{20a}$$

$$\text{subject to} \quad \|n_k + \widetilde{W}_k w\|_X \leq \Delta_k \tag{20b}$$

for some fixed linear operator $\widetilde{W}_k$. We note that the vector $\widetilde{W}_k^* \widetilde{W}_k g_k$ above replaces the vector $\widetilde{W}_k^* g_k$. This modification is needed for the convergence proof.[1] Problem (20) is equivalent to

$$\min_{\widetilde{t}} \quad \frac{1}{2} \langle B_k \widetilde{t}, \widetilde{t} \rangle_X + \langle \widetilde{W}_k g_k, \widetilde{t} \rangle_X \tag{21a}$$

$$\text{subject to} \quad \widetilde{t} \in \text{Range}(\widetilde{W}_k) \tag{21b}$$

$$\|n_k + \widetilde{t}\|_X \leq \Delta_k. \tag{21c}$$

[1] Clearly, with exact linear system solves we have $W_k^* W_k g_k = W_k W_k g_k = W_k g_k$, however, in the presence of inexactness the distinction is important.

We call (21) the *tangential subproblem*. Solving (21) is the first stage of computing the tangential step. To guarantee global convergence of the SQP algorithm, the approximate solution $\widetilde{t}_k$ of (21) must satisfy the following conditions.

Tangential Subproblem Conditions First, the quantity $\widetilde{W}_k g_k$ needs to satisfy

$$\|\widetilde{W}_k g_k - W_k g_k\|_X \leq \tau_1 \min\left\{\|\widetilde{W}_k g_k\|_X, \Delta_k\right\}, \tag{22}$$

for some $\tau_1 > 0$ independent of k. Second, we define the inexact quadratic model

$$\widetilde{q}_k(t) := \frac{1}{2}\langle B_k t, t\rangle_X + \left\langle \widetilde{W}_k g_k, t\right\rangle_X,$$

and impose the fraction of Cauchy decrease condition on $\widetilde{t}_k$ as follows,

$$\widetilde{q}_k(0) - \widetilde{q}_k(\widetilde{t}_k) \geq \kappa_4 \|\widetilde{W}_k g_k\|_X \min\left\{\kappa_5 \|\widetilde{W}_k g_k\|_X, \kappa_6 \Delta_k\right\}, \tag{23}$$

for $\kappa_4, \kappa_5, \kappa_6 > 0$, independent of k. Condition (23) is analogous to the fraction of Cauchy decrease condition (7) in the reduced-space setting. It is an extension of a condition discussed by Dennis, El-Alem, and Maciel in [14]; in our context, the "inexact" quantity $\widetilde{W}_k g_k$ is introduced and the quadratic model is defined according to the discussion in the previous paragraph. Condition (22) is derived from a similar condition by Heinkenschloss and Vicente [21]. It is analogous to the inexact gradient condition (8) in the reduced-space setting. Establishing the existence of $\widetilde{W}_k$ that is compatible with (22) and (23) is an important challenge, discussed in [20]. □

With inexactness $\widetilde{t}_k = \widetilde{W}_k w_k$ is no longer in the null space of $c_x(x_k)$, and may destroy some of the linear feasibility gained by the quasi-normal step n_k. To compensate for this, we compute the tangential step t_k from $\widetilde{t}_k$ to restore linear feasibility as needed. This computation is intimately tied to the global convergence mechanisms used in SQP methods. Once the trial step $s_k = n_k + t_k$ is computed, one must decide whether to accept the step and how to update the trust-region radius Δ_k. To perform these tasks, we use the *augmented Lagrangian merit function*

$$\phi(x, \lambda; \rho) = J(x) + \langle \lambda, c(x)\rangle_C + \rho \|c(x)\|_C^2 = L(x, \lambda) + \rho \|c(x)\|_C^2. \tag{24}$$

In a conventional trust-region SQP algorithm, the step s_k is accepted or rejected and the trust-region radius Δ_k is updated based on the ratio between the actual reduction

$$\text{ared}(s_k; \rho_k) = \phi(x_k, \lambda_k; \rho_k) - \phi(x_k + s_k, \lambda_{k+1}; \rho_k) \tag{25}$$

and the predicted reduction

$$\widehat{\text{pred}}(s_k; \rho_k) = \phi(x_k, \lambda_k; \rho_k) - \Big[L(x_k, \lambda_k) + \langle \nabla_x L(x_k, \lambda_k), s_k\rangle_X + \frac{1}{2}\langle B_k s_k, s_k\rangle_X + \langle \lambda_{k+1} - \lambda_k, c_x(x_k)s_k + c(x_k)\rangle_C + \rho_k \|c_x(x_k)s_k + c(x_k)\|_C^2\Big]. \tag{26}$$

Here λ_{k+1} is a Lagrange multiplier estimate corresponding to the trial iterate $x_k + s_k$. To account for the inexactness in the constraint null space projection, the definition of the predicted reduction must be modified. Using

$$r_k^t = c_x(x_k)t_k \,,$$

which can be interpreted as an indicator of the loss of linear feasibility, the predicted reduction (26) is redefined,

$$\widehat{\text{pred}}(s_k;\, \rho_k) := \text{pred}(n_k, \widetilde{t}_k;\, \rho_k) + \text{rpred}(r_k^t;\, \rho_k),$$

with the following components:

$$\begin{aligned}
&\text{pred}(n_k, \widetilde{t}_k;\, \rho_k) \\
&= -\left\langle \widetilde{W}_k g_k, \widetilde{t}_k \right\rangle_X - \frac{1}{2}\left\langle B_k \widetilde{t}_k, \widetilde{t}_k \right\rangle_X - \left\langle \nabla_x L(x_k, \lambda_k), n_k \right\rangle_X - \frac{1}{2}\left\langle B_k n_k, n_k \right\rangle_X \\
&\quad - \left\langle \lambda_{k+1} - \lambda_k, c_x(x_k)n_k + c(x_k) \right\rangle_C \\
&\quad + \rho_k \left(\|c(x_k)\|_C^2 - \|c_x(x_k)n_k + c(x_k)\|_C^2 \right)
\end{aligned} \tag{27}$$

and

$$\text{rpred}(r_k^t;\, \rho_k) = -\left\langle \lambda_{k+1} - \lambda_k, r_k^t \right\rangle_C - \rho_k \|r_k^t\|_C^2 - 2\rho_k \left\langle r_k^t, c_x(x_k)n_k + c(x_k) \right\rangle_C . \tag{28}$$

The splitting of the predicted reduction into a term that only involves $\widetilde{t}_k$ and a term that only involves t_k is discussed in [20, p. 1514–1515]; it is partially motivated by the arguments made in [21, p. 292].[2] In our algorithm, we first compute a penalty parameter ρ_k satisfying

$$\text{pred}(n_k, \widetilde{t}_k;\, \rho_k) \geq \frac{\rho_k}{2}\left(\|c(x_k)\|_C^2 - \|c_x(x_k)n_k + c(x_k)\|_C^2 \right),$$

and then "postprocess" $\widetilde{t}_k$ to compute a tangential step t_k that satisfies the conditions stated below. The postprocessing can be performed by applying another null-space projection, similar to solving (18).

Tangential Step Conditions The tangential step t_k must satisfy the requirement

$$|\text{rpred}(r_k^t;\, \rho_k)| \leq \eta_0\, \text{pred}(n_k, \widetilde{t}_k;\, \rho_k), \tag{29}$$

where $\eta_0 \in (0,\, 1 - \eta_1)$, and $\eta_1 \in (0,\, 1)$ is the smallest acceptable ratio of the actual and predicted reduction. Additionally, to control how much the tangential step t_k can deviate from the projection $W_k\widetilde{t}_k$ of $\widetilde{t}_k$ we require

[2] For all details, see the proofs in [20, p. 1536–1538] and [21, p. 295–298].

$$\|t_k - W_k\widetilde{t}_k\|_X \leq \tau_3 \Delta_k \min\{\Delta_k, \|s_k\|_X\}, \tag{30}$$

for some $\tau_3 > 0$ independent of k. Finally, we impose the boundedness condition

$$\|\widetilde{t}_k\|_X \leq \tau_4 \|s_k\|_X, \tag{31}$$

for $\tau_4 > 0$ independent of k. These conditions are discussed in [20]. □

So far, we discussed general conditions needed for the global convergence of a composite-step trust-region SQP scheme where the null-space operator W_k is applied inexactly. Following [20], we now present a concrete instance of the algorithm, specifically designed to robustly and efficiently handle inexactness in the iterative solution of linear systems comprising the application of W_k. This algorithm is useful whenever iterative solvers, such as Krylov methods, are applied to solve linear systems based on discretizations of operators c_u and c_u^*, i.e., the state Jacobians and their adjoints. In PDE-constrained optimization, the matrices resulting from, e.g., finite element discretizations of c_u and c_u^* are often very large, prohibiting direct computation of matrix inverses and matrix factorizations. Additionally, in, e.g., optimization under uncertainty, the computational challenge is exacerbated by the dependence of the constraint equation on the random input vector ξ, resulting in enormous linear systems that cannot be formed explicitly. This prompts the need for matrix-free methods, where the action of a linear operator on a vector is specified, rather than the operator (matrix) itself.

We will formulate concrete subalgorithms for the quasi-normal step computation, the solution of the tangential subproblem, the tangential step computation, and the Lagrange multiplier update that satisfy the previously discussed conditions. After specifying the subalgorithms, we state the master algorithm in Section 4.2.5. The solution of augmented systems, which are key components of the subalgorithms, is discussed in Section 5.2, in the context of PDE-constrained optimization under uncertainty.

4.2.1 Computation of the Quasi-Normal Step

An approximate solution of (14) can be computed using the dogleg method. Let n_k^{cp} be the solution of $\min\left\{\|c_x(x_k)n + c(x_k)\|_C^2 \ : \ n = -\alpha c_x(x_k)^* c(x_k), \alpha \geq 0\right\}$, also known as the *Cauchy point*. It is easy to verify that

$$n_k^{cp} = -\frac{\|c_x(x_k)^* c(x_k)\|_X^2}{\|c_x(x_k)c_x(x_k)^* c(x_k)\|_C^2} c_x(x_k)^* c(x_k). \tag{32}$$

If $\|n_k^{cp}\|_X \geq \zeta \Delta_k$, then we set the quasi-normal step to $n_k = \zeta \Delta_k n_k^{cp} / \|n_k^{cp}\|_X$.

If $\|n_k^{cp}\|_X < \zeta \Delta_k$, to accelerate convergence we take a step toward an approximate minimum-norm solution n_k^N of $\min \|c_x(x_k)n + c(x_k)\|_C^2$, sometimes called the *Newton point*. The quasi-normal step is then computed by moving from

n_k^{cp} as far as possible toward n_k^N while staying within the trust region with radius $\zeta \Delta_k$. Specifically, we solve for $\delta n_k = n_k^N - n_k^{cp}$ the augmented system

Algorithm 2 (Dogleg method for the quasi-normal subproblem)

Initialization: Choose $0 < \zeta, \tau^{qn} < 1$.

1. Compute n_k^{cp} as defined in (32).
2. If $\|n_k^{cp}\|_X \geq \zeta \Delta_k$, then set $n_k = \zeta \Delta_k n_k^{cp} / \|n_k^{cp}\|_X$;
   Else compute δn_k via (33) so that e^1 and e^2 satisfy (34).
3. If $\|n_k^{cp} + \delta n_k\|_X \leq \zeta \Delta_k$, then set $n_k = n_k^{cp} + \delta n_k = n_k^N$;
   Else compute $\theta_k \in (0, 1)$ such that $\|n_k^{cp} + \theta_k \delta n_k\|_X = \zeta \Delta_k$, and set $n_k = n_k^{cp} + \theta_k \delta n_k$.

$$\begin{pmatrix} I_X & c_x(x_k)^* \\ c_x(x_k) & 0 \end{pmatrix} \begin{pmatrix} \delta n_k \\ y \end{pmatrix} = \begin{pmatrix} -n_k^{cp} + e^1 \\ -c_x(x_k) n_k^{cp} - c(x_k) + e^2 \end{pmatrix}. \tag{33}$$

To comply with the convergence conditions (15) and (16), the size of the residual $\begin{pmatrix} e^1 & e^2 \end{pmatrix} \in X \times C$ is restricted via

$$\|e^1\|_X^2 + \|e^2\|_C^2 \leq \left(\tau^{qn}\right)^2 \|c_x(x_k) n_k^{cp} + c(x_k)\|_C^2 , \tag{34}$$

where $0 < \tau^{qn} \leq 1$. In summary, the algorithm for computing the quasi-normal step is given as follows:

Additional algorithms that satisfy conditions (15) and (16) are discussed in [21, p. 298–299].

4.2.2 Solution of the Tangential Subproblem

The tangential subproblem (21) is solved using a modified truncated Steihaug-Toint conjugate gradient (STCG) method. Aside from handling the nonstandard objective function in (21), the modifications involve a full orthogonalization of search directions, several important tunings of the classical STCG truncation criteria and the related exit computations, and a special termination condition related to an estimate of the accumulated error in the constraint null-space. The latter is discussed next.

The algorithm for the solution of the tangential subproblem (21), Algorithm 3, repeatedly applies an inexact null-space projector $\widetilde{W}_k$ by iteratively solving augmented systems of type (18). We note that $\widetilde{W}_k$ is not explicitly available; only the results of its action on the vector g_k and its action on the STCG residuals $\widetilde{r}_i$ used in

Algorithm 3 are known.[3] We introduce the operator $R_i : \mathbb{R}^{i+1} \to X$, given by

$$R_i = [g_k, \widetilde{r}_1, \ldots, \widetilde{r}_i],$$

Algorithm 3 (STCG method with inexact null-space projections)

Initialization: Given relative tolerance $tol^{CG} \in (0, 1)$. Given iteration maximum $i_{\max}^{CG}$. Choose $0 < \tau^{pg}, \tau^{proj} \leq 1$. Let $\widetilde{t}_{k,0} = 0 \in X$. Compute $\widetilde{r}_0 = \widetilde{W}_k g_k$ by solving (36) with the linear solver tolerance (37). If $\|\widetilde{r}_0\|_X = 0$, **stop**.

For $i = 0, 1, 2, \ldots, i_{\max}^{CG}$

1. If $i = 0$ set $\widetilde{z}_0 = \widetilde{r}_0$; Else compute $\widetilde{z}_i = \widetilde{W}_k \widetilde{r}_i$ via (38) with the linear solver tolerance (39).
   If $\|\widetilde{z}_i\|_X \leq tol^{CG} \|\widetilde{r}_0\|_X$ and $i > 0$, return $\widetilde{t}_k = \widetilde{t}_{k,i}$ and $\widetilde{t}_k^{cp} = \widetilde{t}_{k,1}$, and **stop**.
2. Compute $\widehat{S}_i$ defined in (35). If $\|\widehat{S}_i\|_2 > 1/2$, return $\widetilde{t}_k = \widetilde{t}_{k,i}$ and $\widetilde{t}_k^{cp} = \widetilde{t}_{k,1}$, and **stop**.
3. Set $\widetilde{p}_i = -\widetilde{z}_i + \sum_{j=0}^{i-1} \frac{\langle \widetilde{z}_i, H_k \widetilde{p}_j \rangle_X}{\langle \widetilde{p}_j, H_k \widetilde{p}_j \rangle_X} \widetilde{p}_j$.
4. If $\langle \widetilde{r}_i, \widetilde{p}_i \rangle_X \neq 0$ and $\langle \widetilde{p}_i, H \widetilde{p}_i \rangle_X \leq 0$, compute θ such that $\text{sign}(\theta) = \text{sign}(-\langle \widetilde{r}_i, \widetilde{p}_i \rangle_X)$ and $\|n_k + \widetilde{t}_{k,i} + \theta \widetilde{p}_i\|_X = \Delta_k$, and return $\widetilde{t}_k = \widetilde{t}_{k,i+1} = \widetilde{t}_{k,i} + \theta \widetilde{p}_i$ and $\widetilde{t}_k^{cp} = \widetilde{t}_{k,1}$, and **stop**.
   If $\langle \widetilde{r}_i, \widetilde{p}_i \rangle_X = 0$ and $\langle \widetilde{p}_i, H \widetilde{p}_i \rangle_X < 0$, compute θ such that $\|n_k + \widetilde{t}_{k,i} + \theta \widetilde{p}_i\| = \Delta_k$, and return $\widetilde{t}_k = \widetilde{t}_{k,i+1} = \widetilde{t}_{k,i} + \theta \widetilde{p}_i$ and $\widetilde{t}_k^{cp} = \widetilde{t}_{k,1}$, and **stop**.
5. If $\langle \widetilde{r}_i, \widetilde{p}_i \rangle_X = 0$, return $\widetilde{t}_k = \widetilde{t}_{k,i}$ and $\widetilde{t}_k^{cp} = \widetilde{t}_{k,1}$, and **stop**.
6. Set $\widetilde{\alpha}_i = -\frac{\langle \widetilde{r}_i, \widetilde{p}_i \rangle_X}{\langle \widetilde{p}_i, H_k \widetilde{p}_i \rangle_X}$.
7. Set $\widetilde{t}_{k,i+1} = \widetilde{t}_{k,i} + \widetilde{\alpha}_i \widetilde{p}_i$.
8. If $\|n_k + \widetilde{t}_{k,i+1}\|_X \geq \Delta_k$, compute θ such that $\text{sign}(\theta) = \text{sign}(\widetilde{\alpha}_i)$ and $\|n_k + \widetilde{t}_{k,i} + \theta \widetilde{p}_i\|_X = \Delta_k$, and return $\widetilde{t}_k = \widetilde{t}_{k,i+1} = \widetilde{t}_{k,i} + \theta \widetilde{p}_i$ and $\widetilde{t}_k^{cp} = \widetilde{t}_{k,1}$, and **stop**.
9. Set $\widetilde{r}_{i+1} = \widetilde{r}_i + \widetilde{\alpha}_i H_k \widetilde{p}_i$.

End For

the operator $\widetilde{Y}_i : \mathbb{R}^{i+1} \to X$, given by

$$\widetilde{Y}_i = [\widetilde{W}_k g_k, \widetilde{W}_k \widetilde{r}_1, \ldots, \widetilde{W}_k \widetilde{r}_i],$$

and the diagonal matrix

$$D_i = \text{diag}(\|\widetilde{W}_k g_k\|_X, \|\widetilde{W}_k \widetilde{r}_1\|_X, \ldots, \|\widetilde{W}_k \widetilde{r}_i\|_X).$$

[3] Only the scope of the index k extends from Algorithm 4 to Algorithm 3 – indices i and j are independent, i.e., their scope is local to each algorithm.

Finally, we define the matrix

$$\widehat{S}_i = D_i^{-1}(\widetilde{Y}_i^T R_i - D_i^2)D_i^{-1}\,. \tag{35}$$

In [20] it is shown that $\|\widehat{S}_i\|_2$ can be used to control the cumulative effect of inexactness in the projections $\widetilde{W}_k$. The modified STCG algorithm is specified next.

The augmented system residuals related to the application of the inexact projector $\widetilde{W}_k$ are controlled as follows. In Step 3 of Algorithm 3, the inexact projected gradient $\widetilde{r}_0 = \widetilde{W}_k g_k$ is computed. The iterative linear system solver returns $\widetilde{r}_0$ satisfying

$$\begin{pmatrix} I_X & c_x(x_k)^* \\ c_x(x_k) & 0 \end{pmatrix} \begin{pmatrix} \widetilde{r}_0 \\ y \end{pmatrix} = \begin{pmatrix} g_k \\ 0 \end{pmatrix} + \begin{pmatrix} e^1 \\ e^2 \end{pmatrix}\,. \tag{36}$$

The residual $\begin{pmatrix} e^1 & e^2 \end{pmatrix} \in X \times C$ must satisfy

$$\|e^1\|_X + \|e^2\|_C \leq \tau^{pg} \min\{\|\widetilde{r}_0\|_X, \Delta_k, \|g_k\|_X\}\,, \tag{37}$$

where $0 < \tau^{pg} \leq 1$. In Step 1 in Algorithm 3, we compute $\widetilde{z}_i = \widetilde{W}_k \widetilde{r}_i$. The iterative linear system solver returns $\widetilde{z}_i$ satisfying

$$\begin{pmatrix} I_X & c_x(x_k)^* \\ c_x(x_k) & 0 \end{pmatrix} \begin{pmatrix} \widetilde{z}_i \\ y \end{pmatrix} = \begin{pmatrix} \widetilde{r}_i \\ 0 \end{pmatrix} + \begin{pmatrix} e_i^1 \\ e_i^2 \end{pmatrix}, \tag{38}$$

where the residual $\begin{pmatrix} e_i^1 & e_i^2 \end{pmatrix} \in X \times C$ is controlled by the condition

$$\|e_i^1\|_X + \|e_i^2\|_C \leq \tau^{proj} \min\{\|\widetilde{z}_i\|_X, \|\widetilde{r}_i\|_X\}\,, \tag{39}$$

with $0 < \tau^{proj} \leq 1$.

4.2.3 Computation of the Tangential Step

Once the approximate solution $\widetilde{t}_k$ of the tangential subproblem (21) has been obtained, the tangential step t_k is computed. The goal is to restore some of the linear feasibility lost in Algorithm 3. To this end, another inexact null space projection is performed,

$$\begin{pmatrix} I_X & c_x(x_k)^* \\ c_x(x_k) & 0 \end{pmatrix} \begin{pmatrix} t_k \\ y \end{pmatrix} = \begin{pmatrix} \widetilde{t}_k \\ 0 \end{pmatrix} + \begin{pmatrix} e^1 \\ e^2 \end{pmatrix}, \tag{40}$$

where the residual $\begin{pmatrix} e^1 & e^2 \end{pmatrix} \in X \times C$ must satisfy

$$\|e^1\|_X + \|e^2\|_C \leq \Delta_k \min\left\{\Delta_k, \|n_k + t_k\|_X, \tau^{tang}\|\widetilde{t}_k\|_X/\Delta_k\right\}\,, \tag{41}$$

for $0 < \tau^{tang} \leq 1$.

4.2.4 Computation of the Lagrange Multipliers

For global convergence, we require only that the sequence of Lagrange multipliers be bounded. For fast convergence, we may compute the Lagrange multiplier estimate λ_{k+1} by approximately minimizing $\|\nabla J(\widehat{x}_k) + c_x(\widehat{x}_k)^*\lambda\|_X$, where $\widehat{x}_k = x_k + n_k + t_k$. Specifically, we solve for $\Delta\lambda = \lambda_{k+1} - \lambda_k$, where λ_k is the previous Lagrange multiplier estimate, as follows:

$$\begin{pmatrix} I_X & c_x(\widehat{x}_k)^* \\ c_x(\widehat{x}_k) & 0 \end{pmatrix} \begin{pmatrix} z \\ \Delta\lambda \end{pmatrix} = \begin{pmatrix} -\nabla J(\widehat{x}_k) - c_x(\widehat{x}_k)^*\lambda_k + e^1 \\ e^2 \end{pmatrix}. \tag{42}$$

The residual $\begin{pmatrix} e^1 & e^2 \end{pmatrix} \in X \times C$ must satisfy

$$\|e^1\|_X + \|e^2\|_C \leq \min\left\{\tau^{lmg},\ \tau^{lmh}\|\nabla J(\widehat{x}_k) + c_x(\widehat{x}_k)^*\lambda_k\|_X\right\}, \tag{43}$$

for $0 < \tau^{lmh} \leq 1$ and a fixed $\tau^{lmg} > 0$ independent of k. Here τ^{lmh} governs the relative size of the linear system residual, while τ^{lmg} is used to enforce boundedness of the multipliers. Clearly, there are many other ways to compute Lagrange multipliers satisfying the boundedness condition.

4.2.5 Full-Space Trust-Region SQP Algorithm with Inexact Linear System Solves

Here we state the complete full-space trust-region SQP algorithm with inexact linear system solves.

To prove convergence of the inexact full-space trust-region algorithm, we use some of the problem assumptions from Section 3 and the following algorithmic assumptions:

(F1) The sequence $\{\lambda_k\}_{k\in\mathbb{N}}$ is bounded.
(F2) The sequence of operators $\{B_k\}_{k\in\mathbb{N}}$ is bounded.
(F3) For each k, the projection $W_k : X \to X$ onto $\mathrm{Null}(c_x(x_k))$ satisfies $\|W_k\|_{\mathcal{L}(X)} = 1$.

Theorem 2 *Let $\Omega = X = U \times Z$. If problem assumptions (A2), (A3), (A4), (A5), and (A9), and algorithmic assumptions (F1), (F2), and (F3) are satisfied, then the sequences of iterates generated by Algorithm 4 satisfy*

$$\liminf_{k\to\infty}\{\|\widetilde{W}_k g_k\|_X + \|c(x_k)\|_C\} = 0. \tag{44}$$

Additionally, we have

$$\liminf_{k\to\infty}\{\|W_k\nabla_x J(x_k)\|_X + \|c(x_k)\|_C\} = 0. \tag{45}$$

Proof In [20] it is shown that Algorithm 4 is a specific instance of a more general composite-step trust-region SQP algorithm, [20, Algorithm 3.3]. Under the given problem assumptions and algorithmic assumptions, the global convergence result follows directly from [20, Theorem 3.5].

Algorithm 4 (Trust-region SQP algorithm with inexact linear system solves)
Initialization: Choose initial point x_0, initial trust-region radius Δ_0, constants $0 < \alpha_1, \eta_1 < 1$, $0 < \eta_0 < 1 - \eta_1$, $\rho_{-1} \geq 1$, $\bar{\rho} > 0$, and $tol^{SQP} > 0$. Set $\Delta_{min}, \Delta_{max}$ so that $0 < \Delta_{min} < \Delta_{max}$. Set forcing parameters $\tau^{qn}, \tau^{pg}, \tau^{proj}, \tau^{tang}, \tau^{lmh} \in (0, 1)$, $\tau^{lmg} > 0$ and $\tau_4 > 1$. Choose initial Lagrange multiplier λ_{-1} and compute λ_0 by solving (42) with linear solver tolerance (43).

For $k = 0, 1, 2, \ldots$

1. **Convergence check:** If $\|\nabla_x L(x_k, \lambda_k)\|_X < tol^{SQP}$ and $\|c(x_k)\|_C < tol^{SQP}$, then **stop**.
2. **Step computation:**
   a. Compute quasi-normal step n_k using Algorithm 2 and linear solver tolerance (34).
   b. Compute $\widetilde{t}_k, \widetilde{t}_k^{cp}$ using Algorithm 3 and linear solver tolerances (37), (39).
3. **Step acceptance:**

   For $i = 0, 1, 2, \ldots$

   a. **For** $j = 0, 1, 2, \ldots$

      i. Compute tangential step t_k by solving (40) with linear solver tolerance (41).
      ii. Compute Lagrange multiplier estimate λ_{k+1} at $x_k + n_k + t_k$ by solving (42) with linear solver tolerance (43).
      iii. Update the penalty parameter: If

$$\text{pred}(n_k, \widetilde{t}_k; \rho_{k-1}) \geq \frac{\rho_{k-1}}{2} \left(\|c(x_k)\|_C^2 - \|c_x(x_k)n_k + c(x_k)\|_C^2 \right)$$

      then set $\rho_k = \rho_{k-1}$. Otherwise set

$$\rho_k = \frac{-2\,\text{pred}(n_k, \widetilde{t}_k; \rho_{k-1})}{\|c(x_k)\|_C^2 - \|c_x(x_k)n_k + c(x_k)\|_C^2} + 2\rho_{k-1} + \bar{\rho}.$$

      iv. If $|\text{rpred}(c_x(x_k)t_k; \rho_k)| > \eta_0\,\text{pred}(n_k, \widetilde{t}_k; \rho_k)$, set $\tau^{tang} = 10^{-3}\,\tau^{tang}$, else **break**.

   End For (j)
   Reset τ^{tang} to its value at outer iteration i prior to Step 3a.

b. If $\|\widetilde{t}_k\|_X > \tau_4 \|n_k + t_k\|_X$ and $\widetilde{t}_k = \widetilde{t}_k^{cp}$

Set $\tau^{qn} = 10^{-1}\tau^{qn}$, $\tau^{pg} = 10^{-1}\tau^{pg}$, $\tau^{proj} = 10^{-1}\tau^{proj}$, $\tau^{tang} = 10^{-1}\tau^{tang}$.
Recompute n_k using Algorithm 2 and linear solver tolerance (34).
Recompute $\widetilde{t}_k$, $\widetilde{t}_k^{cp}$ using Algorithm 3 and linear solver tolerances (37), (39).

Else If $\|\widetilde{t}_k\|_X > \tau_4 \|n_k + t_k\|_X$ and $\widetilde{t}_k \neq \widetilde{t}_k^{cp}$

Set $\widetilde{t}_k = \widetilde{t}_k^{cp}$.

Else

Optional: Reset τ^{qn}, τ^{pg}, τ^{proj}, and τ^{tang} to their values from the Initialization step.
break

End For (i)

4. **Trust-region update:**

a. Compute trial step $s_k = n_k + t_k$.
b. Compute ratio $\theta_k = \text{ared}(s_k; \rho_k)/\text{pred}(n_k, \widetilde{t}_k; \rho_k)$.
c. If $\theta_k \geq \eta_1$, set $x_{k+1} = x_k + s_k$, and choose Δ_{k+1} as follows:

If $\theta_k \geq 0.9$, set
$\Delta_{k+1} = \min\{\max\{7\,\|s_k\|_{\mathcal{X}}, \Delta_k, \Delta_{min}\}, \Delta_{max}\}$
Else If $\theta_k \geq 0.8$, set
$\Delta_{k+1} = \min\{\max\{2\,\|s_k\|_{\mathcal{X}}, \Delta_k, \Delta_{min}\}, \Delta_{max}\}$
Else set
$\Delta_{k+1} = \max\{\Delta_k, \Delta_{min}\}$;

Else set $x_{k+1} = x_k$, $\lambda_{k+1} = \lambda_k$, and $\Delta_{k+1} = \alpha_1 \|s_k\|_X$.

End For (k)

4.2.6 Related Work

The above approach is based on the general composite-step trust-region framework presented by Dennis, El-Alem, and Maciel in [14]. To accommodate inexact computations, Algorithm 4 includes several modifications of [14]: (i) the gradient condition (22); (ii) the redefinition of the predicted reduction to evaluate progress, given in (27) and (28); and (iii) the tangential step conditions (29), (30), and (31). Modification (i) is derived from Heinkenschloss and Vicente [21]. It is related to the gradient condition (8) for reduced-space (unconstrained) formulations, which is discussed in detail in Section 4.1.1. Modifications (ii) and (iii) are related to similar such conditions proposed in [21]. However, we note that Heinkenschloss and Vicente assume a decomposition of optimization variables into basic and

nonbasic (state and control) variables, i.e., solve problem (1), whereas Algorithm 4 is applied to problem (3). The latter requires a different algorithmic strategy to enforce (ii) and (iii). Specifically, in [21] the state/control decomposition assumption allows one to set the control component of the quasi-normal step to zero, to formulate and solve the tangential subproblem for the control component only, and to separately compute the state component of the tangential step. In Algorithm 4, these computations are interconnected and therefore more involved when it comes to specifying concrete subalgorithms. However, Algorithm 4 is more general in the sense that it can be extended to the case where the constraint Jacobian is rank deficient. Additionally, Algorithm 4 enables the use of efficient iterative solvers and preconditioners for linear optimality systems.

Another composite-step trust-region approach using the basic/nonbasic decomposition of the equality constraints is presented by Ziems and Ulbrich in [43], where the emphasis is on an efficient management of adaptive PDE discretizations, i.e., control of finite element discretization error. In [44], Ziems extends the approach to handle additional constraints, such as bounds, on the control (nonbasic) variables. To rigorously incorporate finite element error estimates in the trust-region algorithm, in addition to the previously reviewed concepts, Ziems and Ulbrich require the notion of inexact actual reductions and propose implementable conditions to control such inexactness.

5 Application to Risk-Neutral Optimization

In this section, we specialize the reviewed algorithms to optimization problems where the governing PDEs include uncertain or random coefficients and the objective function is an expectation. First we present a reduced-space method that enables efficient use of dimension-adaptive sparse grids in the computation of the (reduced) objective function and its gradient. Subsequently, we discuss the numerical solution of augmented systems arising in the full-space approach to PDE-constrained optimization under uncertainty.

We denote the random inputs to a PDE model by ξ, which is a random vector defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Here, Ω is the set of outcomes, $\mathcal{F} \subset 2^{\Omega}$ is a σ-algebra of all possible events, and $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a probability measure. As is common in the literature, we assume finite-dimensional noise, i.e., $\Xi := \xi(\Omega) \subset \mathbb{R}^m$ for some $m \in \mathbb{N}$. We further assume that Ξ is the m-fold tensor product of one-dimensional intervals Ξ_ℓ, $\ell = 1, \ldots, m$, and the components of ξ are independent and continuously distributed with one-dimensional Lebesgue densities $\rho_\ell : \Xi_\ell \to [0, \infty)$. In this setting, the PDE solution operator $S(z)$ is a random field with realizations in U. We denote the dependence of S on the random input ξ by $S(z; \xi)$ for a fixed control $z \in Z$ and note that $u = S(z; \xi) \in U$ solves the parametrized PDE

$$c(u, z; \xi) = 0 \quad \text{a.s.}$$

where $c : U \times Z \times \Xi \to C$. Here "a.s." is an abbreviation for "almost surely"; in other words, "up to a set of probability zero." Similarly, the objective function is parametrized as $J : U \times Z \times \Xi \to \mathbb{R}$. As in Section 3, we consider the full space problem

$$\min_{u \in U,\ z \in Z} \quad \mathbb{E}[J(u, z; \xi)] \tag{46a}$$

$$\text{subject to} \quad c(u, z; \xi) = 0 \quad \text{a.s.}, \tag{46b}$$

where $\mathbb{E}[X] := \int_\Omega X(\omega)\, \mathrm{d}\mathbb{P}(\omega)$ denotes the expectation of the random variable X. When evaluating the expectation of random variables with the form $f(\xi)$ where $f : \Xi \to \mathbb{R}$, the finite-dimensional noise assumption permits the following substitution:

$$\mathbb{E}[\, f(\xi)\,] = \int_{\Xi_1} \rho_1(\xi_1) \cdots \int_{\Xi_m} \rho_m(\xi_m) f(\xi)\, \mathrm{d}\xi_m \cdots \mathrm{d}\xi_1. \tag{47}$$

We slightly abuse notation and use $\xi = (\xi_1, \ldots, \xi_m)$ to denote both the random vector of inputs and its realizations. Substituting $S(z; \xi)$ into J produces the random-variable objective function $\widehat{J}(z; \xi) = J(S(z; \xi), z; \xi)$, leading to the reduced problem

$$\min_{z \in Z} \{\mathcal{J}(z) := \mathbb{E}[\, \widehat{J}(z; \xi)\,]\}. \tag{48}$$

To approximate the expectation in (46) and (48), we employ sparse-grid quadrature [6, 7, 19, 31, 37, 42]. Let $\{\mathbb{E}_\ell^i\}_{i \geq 1}$ be a sequence of one-dimensional quadrature operators of increasing order in the $\ell = 1, \ldots, m$ direction. That is, $\mathbb{E}_\ell^{i+1}$ is exact for higher-order monomials than $\mathbb{E}_\ell^i$. Define the 1-D difference quadrature operators

$$\delta_\ell^1 := \mathbb{E}_\ell^1 \quad \text{and} \quad \delta_\ell^i := \mathbb{E}_\ell^i - \mathbb{E}_\ell^{i-1}, \quad \text{for} \quad i \geq 2.$$

To define the m-dimensional quadrature rule on $\Xi = \Xi_1 \times \cdots \times \Xi_M$ let $\mathbf{i} = (i_1, \ldots, i_m)$ be a multi-index and let $\mathcal{I} \subset \mathbb{N}^m$ be a finite multi-index set. The general sparse-grid quadrature operator is defined as

$$\mathbb{E}_\mathcal{I} := \sum_{\mathbf{i} \in \mathcal{I}} (\delta_1^{i_1} \otimes \cdots \otimes \delta_m^{i_m}). \tag{49}$$

To ensure consistency of (49), $\mathcal{I}$ must satisfy the following condition: if $\mathbf{i} = (i_1, \ldots, i_m) \in \mathcal{I}$, $\mathbf{j} = (j_1, \ldots, j_m) \in \mathbb{N}^m$, and $j_\ell \leq i_\ell$ for all $\ell = 1, \ldots, m$, then $\mathbf{j} \in \mathcal{I}$. If $\mathcal{I}$ satisfies this condition, then we say that $\mathcal{I}$ is *admissible*. In one dimension, admissibility guarantees that (49) is a telescoping sum and recovers $\mathbb{E}_1^i$ where i denotes the maximum element of $\mathcal{I}$. An example of an admissible index set is

$$\mathcal{I} = \left\{ \mathbf{i} \in \mathbb{N}^m \, : \, |i_1| + \ldots + |i_m| \leq \ell + m - 1 \right\}$$

for $\ell \in \mathbb{N}$ which results in the standard isotropic sparse grid.

In general, $\mathbb{E}_{\mathcal{I}}[f]$ for $f : \Xi \to \mathbb{R}$ can be written as

$$\mathbb{E}_{\mathcal{I}}[f] = \sum_{k=1}^{Q} w_k f(\xi_k) \tag{50}$$

where w_k are the quadrature weights associated with the quadrature points ξ_k, $k = 1, \ldots, Q$. The form of approximation (50) is not unique to sparse grids as virtually all quadrature and sampling methods have this form. Applying $\mathbb{E}_{\mathcal{I}}$ for fixed index set $\mathcal{I}$ to (46) and (48) results in the approximate optimization problems

$$\min_{u \in U, \, z \in Z} \quad \sum_{k=1}^{Q} w_k J(u_k, z; \xi_k)] \tag{51a}$$

$$\text{subject to} \quad c(u_k, z; \xi_k) = 0, \quad k = 1, \ldots, Q \, , \tag{51b}$$

and

$$\min_{z \in Z} \sum_{k=1}^{Q} w_k \widehat{J}(z; \xi_k) \tag{52}$$

where $\widehat{J}(z; \xi_k) = J(S(z; \xi_k), z; \xi_k)$ and $S(z; \xi_k) = u_k \in U$ solves (51b) for $k = 1, \ldots, Q$. When considering the full-space algorithm, we require the Lagrangian functional associated with (51), i.e.,

$$L(u_1, \ldots, u_Q, z, \lambda_1, \ldots, \lambda_Q) := \sum_{k=1}^{Q} w_k J(u_k, z; \xi_k) + \sum_{k=1}^{Q} v_k \langle \lambda_k, c(u_k, z; \xi_k) \rangle_C \tag{53}$$

where v_k, $k = 1, \ldots, Q$, are fixed weights. We can choose v_k to be $v_k = 1$ or $v_k = w_k$ for $k = 1, \ldots, Q$. The later choice corresponds to an infinite-dimensional view of the problem since

$$\sum_{k=1}^{Q} w_k \langle \lambda_k, c(u_k, z; \xi_k) \rangle_C$$

is an approximation of the expectation

$$\mathbb{E}[\langle \lambda, c(u(\xi), z; \xi) \rangle_C].$$

5.1 Sparse-Grid Adaptivity

In the subsequent subsections, we define the adaptive sparse-grid subalgorithms used to satisfy (8) and (10). To do so, we require the definition of the forward neighborhood of the admissible index set $\mathcal{I}$. The forward neighborhood of $\mathcal{I}$ is

$$\mathcal{N}(\mathcal{I}) := \{\mathbf{i} \in \mathbb{N}^m \setminus \mathcal{I} \,:\, \mathcal{I} \cup \{\mathbf{i}\} \text{ is admissible}\}.$$

We employ dimension-adaptive sparse grids [19] in an attempt to satisfy (8) and (10). The dimension-adaptive sparse grid algorithm approximates the quadrature error on a subset $\mathcal{A}$ of the forward neighborhood of the current admissible index set $\mathcal{O}$, i.e., $\mathcal{A} \subseteq \mathcal{N}(\mathcal{O})$.

5.1.1 Computation of Inexact Gradient

Given the current iterate $z_k \in Z$, we must construct a model m_k that satisfies (8). To do so, we employ an admissible index set $\mathcal{I}_k^{\mathrm{g}} \subset \mathbb{N}^m$ and the associated quadrature approximation of $\mathcal{J}(z) = \mathbb{E}[\,\widehat{J}(z;\xi)\,]$, i.e.,

$$\mathcal{J}_{\mathcal{I}_k^{\mathrm{g}}}(z) = \sum_{\mathbf{i} \in \mathcal{I}_k^{\mathrm{g}}} (\delta_1^{i_1} \otimes \cdots \otimes \delta_m^{i_m})[\,\widehat{J}(z;\xi)\,].$$

We then choose our model m_k to satisfy the first-order condition $\nabla m_k(0) = \nabla \mathcal{J}_{\mathcal{I}_k^{\mathrm{g}}}(z_k)$. Under the assumption of convergence of (49), we can write the quadrature error associated with the index set $\mathcal{I}_k^{\mathrm{g}}$ as the sum of all differential quadrature rules $(\delta_1^{i_1} \otimes \cdots \otimes \delta_m^{i_m})$ for $\mathbf{i} \notin \mathcal{I}_k^{\mathrm{g}}$. Thus, the inexact gradient condition (8) becomes

$$\left\| \sum_{\mathbf{i} \notin \mathcal{I}_k^{\mathrm{g}}} (\delta_1^{i_1} \otimes \cdots \otimes \delta_M^{i_M})[\nabla \widehat{J}(z;\xi)] \right\|_Z \leq \kappa_{\mathrm{grad}} \min \left\{ \left\| \nabla \mathcal{J}_{\mathcal{I}_k^{\mathrm{g}}}(z_k) \right\|_Z, \Delta_k \right\}. \tag{54}$$

The goal now is to determine the smallest admissible index set $\mathcal{I}_k^{\mathrm{g}}$ such that (54) holds. Since it is not computationally feasible to explicitly evaluate the left-hand side of (54), we employ the dimension-adaptive approach presented in [19] to approximately satisfy this condition. Although there is no proof that this approach satisfies (8), numerical experience suggests that (8) is typically satisfied. The dimension-adaptive gradient computation algorithm is listed in Algorithm 5.

5.1.2 Computation of Inexact Objective Function Value

Similar to the inexact gradient computation, we define our objective function approximation for the computation of cred_k as

$$\mathcal{J}_k(z) = \mathcal{J}_{\mathcal{I}_k^{\mathrm{o}}}(z) = \sum_{\mathbf{i}\in\mathcal{I}_k^{\mathrm{o}}} (\delta_1^{i_1} \otimes \cdots \otimes \delta_m^{i_m})[\widehat{J}(z;\xi)].$$

Here, $\mathcal{I}_k^{\mathrm{o}} \subset \mathbb{N}^m$ is some admissible index set. The error associated with this approximation, in the context of (10), is

$$|\mathrm{ared}_k - \mathrm{cred}_k| = \left| \sum_{\mathbf{i}\notin\mathcal{I}_k^{\mathrm{o}}} (\delta_1^{i_1} \otimes \cdots \otimes \delta_m^{i_m})[\widehat{J}(z_k + s_k;\xi) - \widehat{J}(z_k;\xi)] \right| = \theta_k. \tag{55}$$

Algorithm 5 (Gradient computation using adaptive sparse grids)

Initialization: *Set* $\mathbf{i} = (1,\ldots,1)$, $\mathcal{A} = \{\mathbf{i}\}$, $\mathcal{O} = \emptyset$, $g_{\mathbf{i}} = (\delta_1^{i_1} \otimes \cdots \otimes \delta_M^{i_M})[\nabla\widehat{J}(z_k;\xi)]$ *and* $\beta = \beta_{\mathbf{i}} = \|g_{\mathbf{i}}\|_{\mathcal{Z}}$, $g = g_{\mathbf{i}}$, *and* $\mathrm{TOL} = \kappa_{\mathrm{grad}} \min\{\|g\|_{\mathcal{Z}}, \Delta_k\}$.

While $\beta > \mathrm{TOL}$

1. *Select* $\mathbf{i} \in \mathcal{A}$ *corresponding to the largest* $\beta_{\mathbf{i}}$
2. *Set* $\mathcal{A} \leftarrow \mathcal{A} \setminus \{\mathbf{i}\}$ *and* $\mathcal{O} \leftarrow \mathcal{O} \cup \{\mathbf{i}\}$
3. *Update the error indicator* $\beta \leftarrow \beta - \beta_{\mathbf{i}}$
4. ***For*** $\ell = 1,\ldots,m$
   a. *Set* $\mathbf{j} = \mathbf{i} + \mathbf{e}_\ell$
   b. ***If*** $\mathcal{O} \cup \{\mathbf{j}\}$ *is admissible*
      i. *Set* $\mathcal{A} \leftarrow \mathcal{A} \cup \{\mathbf{j}\}$
      ii. *Set* $g_{\mathbf{j}} = (\delta_1^{j_1} \otimes \cdots \otimes \delta_m^{j_m})[\nabla\widehat{J}(z_k;\xi)]$
      iii. *Set* $\beta_{\mathbf{j}} = \|g_{\mathbf{j}}\|_{\mathcal{Z}}$
      iv. *Update the gradient approximation* $g \leftarrow g + g_{\mathbf{j}}$
      v. *Update the error indicator* $\beta \leftarrow \beta + \beta_{\mathbf{j}}$
      vi. *Update the stopping tolerance* $\mathrm{TOL} = \kappa_{\mathrm{grad}} \min\{\|g\|_{\mathcal{Z}}, \Delta_k\}$
   c. ***EndIf***
5. ***EndFor***

EndWhile

Set $\mathcal{I}_k^{\mathrm{g}} = \mathcal{A} \cup \mathcal{O}$ *and* $\nabla m_k(0) = g$.

Again, we use dimension-adaptive sparse grids to determine $\mathcal{I}_k^{\mathrm{o}}$ where we estimate (55) only on a subset of the forward neighborhood of the current index set. Similar to the gradient computation, there is no guarantee that θ_k will satisfy (10). However, numerical experience suggest that this is often the case. The dimension-adaptive objective function approximation algorithm is listed in Algorithm 6.

5.2 *Iterative Linear System Solves*

In this section, we focus on the key computational component of the subalgorithms of Algorithm 4, namely, the numerical solution of augmented systems. As mentioned earlier, in optimization under uncertainty these systems are enormous and cannot be formed explicitly, requiring iterative, matrix-free methods.

Augmented systems are KKT systems for a special type of quadratic optimization problems. Solution methods for KKT systems have received significant attention recently in the context of PDE-constrained optimization. Efficient preconditioning

Algorithm 6 (Objective function evaluation using adaptive sparse grids)

Initialization: *Set* $\mathbf{i} = (1, \ldots, 1)$, $\mathcal{A} = \{\mathbf{i}\}$, $\mathcal{O} = \emptyset$, $\mathrm{TOL} = (\eta \min\{\mathrm{pred}_k, r_k\})^{1/\omega}$, $\tilde{\theta}_k = \vartheta_{\mathbf{i}} = (\delta_1^{i_1} \otimes \cdots \otimes \delta_M^{i_M})[\widehat{J}(z_k + s_k; \xi) - \widehat{J}(z_k; \xi)]$ *and* $\mathrm{cred}_k = \vartheta_{\mathbf{i}}$.

While $|\tilde{\theta}_k| > \mathrm{TOL}$

1. *Select* $\mathbf{i} \in \mathcal{A}$ *corresponding to the largest* $|\vartheta_{\mathbf{i}}|$
2. *Set* $\mathcal{A} \leftarrow \mathcal{A} \setminus \{\mathbf{i}\}$ *and* $\mathcal{O} \leftarrow \mathcal{O} \cup \{\mathbf{i}\}$
3. *Update the error indicator* $\tilde{\theta}_k \leftarrow \tilde{\theta}_k - \vartheta_{\mathbf{i}}$
4. **For** $\ell = 1, \ldots, m$
   a. *Set* $\mathbf{j} = \mathbf{i} + \mathbf{e}_\ell$
   b. **If** $\mathcal{O} \cup \{\mathbf{j}\}$ *is admissible*
      i. *Set* $\mathcal{A} \leftarrow \mathcal{A} \cup \{\mathbf{j}\}$
      ii. *Set* $\vartheta_{\mathbf{j}} = (\delta_1^{j_1} \otimes \cdots \otimes \delta_m^{j_m})[\widehat{J}(z_k + s_k; \xi) - \widehat{J}(z_k; \xi)]$
      iii. *Update the computed reduction* $\mathrm{cred}_k \leftarrow \mathrm{cred}_k + \vartheta_{\mathbf{j}}$
      iv. *Update the error indicator* $\tilde{\theta}_k \leftarrow \tilde{\theta}_k + \vartheta_{\mathbf{j}}$
   c. **EndIf**
5. **EndFor**

EndWhile

Return $\mathcal{I}_k^{\mathrm{o}} = \mathcal{A} \cup \mathcal{O}$ *and* cred_k.

approaches based on Schur complements in the constraint null space have been developed, see, e.g., [33–35, 38, 39]. Augmented systems are treated in [36], where it is shown that Schur-complement ideas lead to preconditioners that perform well for a variety of physics models, i.e., constraint equations, independent of the mesh size. A crucial difference between the KKT systems for the subproblem (13a)–(13b) and augmented systems is that the latter do not depend at all on the

objective function J, which lowers the bar for efficient preconditioning and affords generality.[4]

Recalling the assumption in full-space methods that C is a Hilbert space (with $C^* = C$), in PDE-constrained optimization the augmented system operator $G : X \times C \to X \times C$, in general, written as

$$G = \begin{pmatrix} I_X & c_x(x_k)^* \\ c_x(x_k) & 0 \end{pmatrix},$$

takes the form $A : U \times Z \times C \to U \times Z \times C$,

$$A = \begin{pmatrix} I_U & 0 & c_u(u_k, z_k)^* \\ 0 & I_Z & c_z(u_k, z_k)^* \\ c_u(u_k, z_k) & c_z(u_k, z_k) & 0 \end{pmatrix} =: \begin{pmatrix} I_U & 0 & C_u^* \\ 0 & I_Z & C_z^* \\ C_u & C_z & 0 \end{pmatrix},$$

where the latter notation is used as shorthand. Assuming the existence of $(c_u(u_k, z_k))^{-1}$, we consider two preconditioners for the operator A:

$$P^\star = \begin{pmatrix} I_U & 0 & 0 \\ 0 & I_Z & 0 \\ 0 & 0 & (C_u C_u^* + C_z C_z^*)^{-1} \end{pmatrix} \quad \text{and} \quad P = \begin{pmatrix} I_U & 0 & 0 \\ 0 & I_Z & 0 \\ 0 & 0 & C_u^{-*} C_u^{-1} \end{pmatrix}.$$

The preconditioner $P^\star$ is an exact Schur-complement preconditioner, in the sense that a $P^\star$-preconditioned Krylov solver for a system given by the operator A converges in at most three iterations [30]. However, in PDE-constrained optimization under uncertainty, the operator $C_u C_u^* + C_z C_z^*$ is never formed explicitly, due to its sheer size. Also, approximating the inverse of the sum of matrix products is in general very difficult. The approximate preconditioner P is a practical alternative. The application of P amounts to a "linearized state solve" followed by an "adjoint solve," which are readily available in practice. Following the notation in problem (51), with Lagrangian (53), for a given number Q of samples and weights the matrices C_u and C_z have special structure, namely that of a block-diagonal and a block-column matrix with Q nonzero blocks, respectively,

$$C_u = \begin{pmatrix} v_1 C_u^1 & 0 & \dots & 0 \\ 0 & v_2 C_u^2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & v_Q C_u^Q \end{pmatrix} \quad \text{and} \quad C_z = \begin{pmatrix} v_1 C_z^1 \\ v_2 C_z^2 \\ \vdots \\ v_Q C_z^Q \end{pmatrix}.$$

[4] Under the assumptions of this chapter, augmented systems can always be related to *strictly convex quadratic problems* of the form $\min \frac{1}{2}\langle s^1, s^1\rangle_X - \langle b^1, s^1\rangle_X$ subject to $c_x(x_k)s^1 = b_2$, where $(b^1 \;\; b^2)^T$ is the right-hand side vector of the augmented system and s^1 is the first block of the left-hand side vector.

$$\begin{pmatrix}
I_U & 0 & \dots & 0 & 0 & v_1(C_u^1)^* & 0 & \dots & 0 \\
0 & I_U & \dots & 0 & 0 & 0 & v_2(C_u^2)^* & \dots & 0 \\
\vdots & \vdots & \ddots & 0 & \vdots & \vdots & \vdots & \ddots & 0 \\
0 & 0 & \dots & I_U & 0 & 0 & 0 & \dots & v_Q(C_u^Q)^* \\
0 & 0 & \dots & 0 & I_Z & v_1(C_z^1)^* & v_2(C_z^2)^* & \dots & v_Q(C_z^Q)^* \\
v_1C_u^1 & 0 & \cdots & 0 & v_1C_z^1 & 0 & 0 & \dots & 0 \\
0 & v_2C_u^2 & \cdots & 0 & v_2C_z^2 & 0 & 0 & \dots & 0 \\
\vdots & \vdots & \ddots & 0 & \vdots & \vdots & \vdots & \ddots & 0 \\
0 & 0 & \dots & v_QC_u^Q & v_QC_z^Q & 0 & 0 & \dots & 0
\end{pmatrix}$$

Fig. 1 The augmented system in PDE-constrained optimization under uncertainty. The application of the preconditioner P to this system is highly parallelizable, due to the block-diagonal structure of the highlighted C_u and C_u^* blocks

This gives rise to the augmented system depicted in Figure 1. We note that preconditioning this system with P is highly parallelizable, due to the block-diagonal structure of C_u and C_u^*. We also note that only approximations of the inverses C_u^{-1} and C_u^{-*} are needed in the application of the preconditioner, enabling efficient iterative schemes that rely on whatever solvers are provided for the linearized forward and adjoint systems.

6 Numerical Examples

We consider a thermal fluid application motivated by the transport process in high-pressure chemical vapor deposition (CVD) reactors (see Section 5.2 in [23]). Such reactors are used to produce compound semiconductors. Reactant gases are injected into the top of the reactor and must flow down to the substrate in order to form an epitaxial film. However, the substrate is maintained at a high temperature causing vorticities due to buoyancy-driven convection. For this application, we control the thermal flux on the side walls of the reactor to minimize vorticity. Let $D = (0, 1) \times (0, 1)$ and consider the following control problem:

$$\min_{z \in Z} \frac{1}{2}\mathbb{E}\left[\int_D (\nabla \times u(z))\,\mathrm{d}x\right] + \frac{\gamma}{2}\int_{\Gamma_c} |z|^2\,\mathrm{d}x$$

where $S(z) = (u(z), p(z), T(z)) = (u, p, T) : \Omega \to U$ solves the Boussinesq flow equations,

$$
\begin{aligned}
-\nu(\xi)\nabla^2 u + (u\cdot\nabla)u + \nabla p + \eta(\xi)Tg &= 0, && \text{in } D, \text{ a.s.},\\
\nabla\cdot u &= 0, && \text{in } D, \text{ a.s.},\\
-\kappa(\xi)\Delta T + u\cdot\nabla T &= 0, && \text{in } D, \text{ a.s.},\\
u - u_i &= 0, & T &= 0, \text{ on } \Gamma_i, \text{ a.s.},\\
u - u_o &= 0, & \kappa(\xi)\tfrac{\partial T}{\partial n} &= 0, \text{ on } \Gamma_o, \text{ a.s.},\\
u &= 0, & T &= T_b(\xi), \text{ on } \Gamma_b, \text{ a.s.},\\
u &= 0, & \kappa(\xi)\tfrac{\partial T}{\partial n} + h(\xi)(z-T) &= 0, \text{ on } \Gamma_c, \text{ a.s.},
\end{aligned}
$$

where $\Gamma_i = [1/3, 2/3]\times\{1\}$, $\Gamma_o = ([0, 1/3]\cup[2/3, 1])\times\{1\}$, $\Gamma_b = [0, 1]\times\{0\}$, and $\Gamma_c = \{0, 1\}\times[0, 1]$. The inflow and outflow velocities, u_i and u_o, are deterministic while the coefficients ν, η, κ, h, and T_b are uncertain. In this model,

$$
\nu = \frac{1}{\mathrm{Re}} = \frac{100}{1+0.01\xi_{N+1}}, \quad \eta = \frac{\mathrm{Gr}}{\mathrm{Re}^2} = 0.72\frac{1+0.01\xi_{N+1}}{1+0.01\xi_{N+2}},
$$

$$
\text{and} \quad \kappa = \frac{1}{\mathrm{Re}\,\mathrm{Pr}} = 10^5\frac{1+0.01\xi_{N+3}}{(1+0.01\xi_{N+1})^2}
$$

where Re is the Reynolds number, Gr is the Grashof number, and Pr is the Prandtl number. The offset N is the total number of random variables associated with T_0 and h. The uncertainty in T_0 is modeled by the expansion

$$
T_0(x,\xi) = 1 + 0.025\sum_{k=1}^{n_b}\xi_k\frac{\sqrt{2}\sin(\pi k x)}{\pi k}.
$$

The coefficient h has a similar expansion for $x = 0$ and for $x = 1$ with n_ℓ and n_r terms, respectively. All ξ_k are uniformly distributed on $[-1, 1]$. Figure 2 depicts the computational domain including boundary conditions (left) and the scenarios of the uncertain substrate temperature (right). The curves on top of the computational domain schematic (left) are the inflow and outflow profiles of the velocity, given by

$$
u(x) = \begin{cases} 2\left(\frac{1}{3}-x\right)x & \text{if } 0\le x\le\frac{1}{3}\\ -4\left(x-\frac{1}{3}\right)\left(\frac{2}{3}-x\right) & \text{if } \frac{1}{3}<x<\frac{2}{3}\\ 2\left(x-\frac{2}{3}\right)(1-x) & \text{if } \frac{2}{3}\le x\le 1\end{cases}.
$$

We study both the stated reduced-space formulation of the control problem and the corresponding full-space formulation. We use the Trilinos package Rapid Optimization Library [27] and its PDE-OPT Application Development Kit, available in the directory

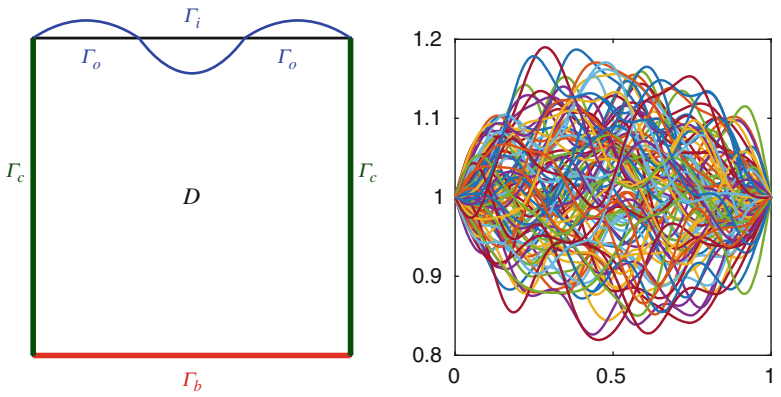

Fig. 2 Left: Computational domain for the CVD reactor. Right: Scenarios of T_0

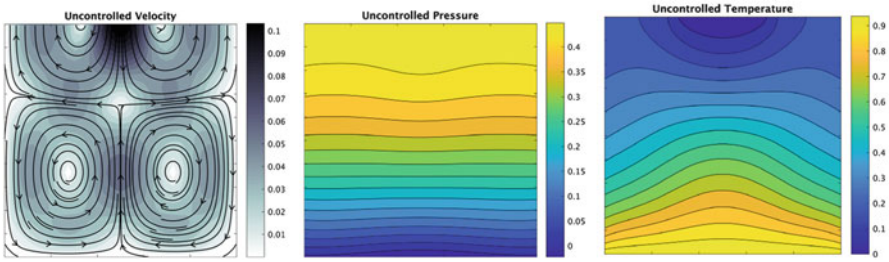

Fig. 3 Expected values of the uncontrolled velocity field (left), pressure (middle) and temperature (right)

Trilinos/packages/rol/examples/PDE-OPT,

to implement and solve the control problem. A reproducibility statement is given in Section 6.3.

To discretize the PDE, we use finite elements on a uniform mesh of 33×33 quadrilaterals. We note that to properly represent the boundary segments Γ_o and Γ_i the number of mesh cells in the horizontal direction should be divisible by three. For the velocity and pressure discretization, we use the Q2-Q1 Taylor–Hood finite element pair, and for the temperature we use the Q2 finite element. The sparse grids used to approximate the risk neutral objective function are built on one-dimensional Clenshaw–Curtis quadrature rules. We set the maximum sparse grid as the level-3 isotropic Clenshaw–Curtis sparse grid. This rule has $Q = 2245$ points. Figure 3 shows the expected values of the uncontrolled ($z = 0$) velocity field (streamlines and magnitude), pressure, and temperature.

6.1 Reduced-Space Results with Adaptive Sparse Grids

We solve the optimal control problem using the reduced-space trust-region approach with dimension-adaptive sparse grids described in Section 5.1. We terminate the algorithm when the gradient of the model at zero, $\|\nabla m_k(0)\|_Z$, is below $tol = 10^{-6}$. We set the initial guess to $z = 0$ and the initial trust-region radius to $\Delta_0 = 10$. We solve the trust-region subproblem using truncated conjugate gradients. We terminate the subproblem solve if the step has exceeded the trust-region radius, the algorithm encountered negative curvative or the residual is below the minimum of 10^{-4} and 10^{-2} times the norm of the initial residual. The problem is solved using a single computational core of a dual-socket 2.1 GHz Intel Broadwell E5-2695 compute node with 128 GB RAM. The linearized state and adjoint equations are solved using a direct solver. The solves are performed sequentially.

Figure 4 shows the expected values of the controlled velocity field (streamlines and magnitude), pressure, and temperature. We see a significant reduction of vorticity near the heated substrate. The left-wall and right-wall controls are given in Figure 5, respectively.

Table 1 displays the iteration history for the adaptive sparse grid algorithm described in Sections 4.1 and 5.1. The columns from left to right include the iteration count (iter), the computed objective function value ($\mathcal{J}_k(z_k)$), the norm of the model gradient ($\|\nabla m_k(0)\|_Z$), the trial step size ($\|s_k\|_Z$), the trust-region radius (Δ_k), the number of truncated conjugate gradient iterations (cg), a Boolean corresponding to whether the step was accepted or rejected (accept), the number of sparse-grid points for the objective function computation (obj), and the number of sparse-grid points for the gradient evaluation (grad). The algorithm starts with very few sparse-grid points (i.e., state and adjoint PDE solvers) and only refines the sparse grid as needed for global convergence. For this example, there were no "unimportant" directions resulting in a final sparse grid that is identical to the isotropic sparse grid. For examples with anisotropy, see [25, 26].

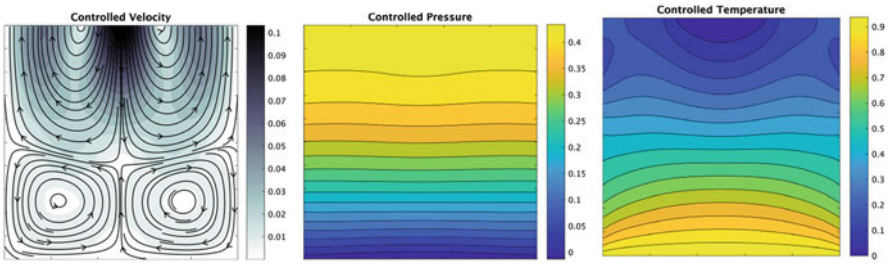

Fig. 4 Expected values of the controlled velocity field (left), pressure (middle), and temperature (right), obtained using the reduced-space method with adaptive sparse grids

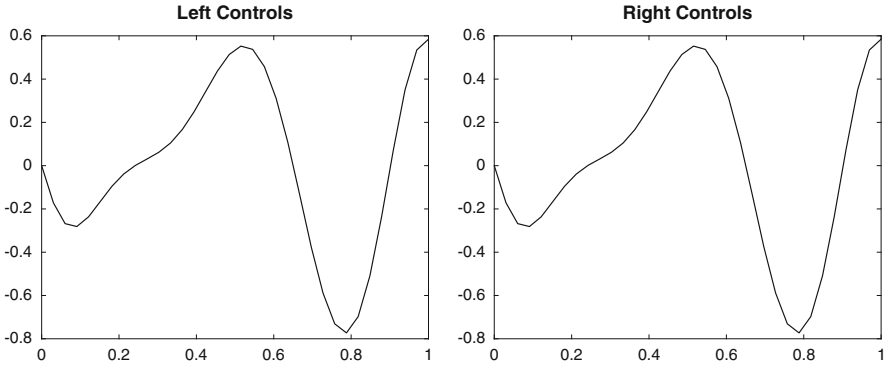

Fig. 5 Optimal controls along the left vertical side wall (left image) and the right vertical side wall (right image) of the problem domain D, obtained using the reduced-space method with adaptive sparse grids

Table 1 Iteration history for reduced-space adaptive sparse-grid approach

iter	$\mathcal{J}_k(z_k)$	$\|\nabla m_k(0)\|_Z$	$\|s_k\|_Z$	Δ_k	cg	accept	obj	grad
0	0.07457916	5.063×10^{-2}	–	10.000	–	–	1	3
1	0.07469930	5.063×10^{-2}	10.000	1.445	1	0	3	3
2	0.07469930	5.063×10^{-2}	1.445	0.361	1	0	3	3
3	0.05636707	4.875×10^{-2}	0.361	0.903	1	1	3	3
4	0.05636707	4.875×10^{-2}	0.903	0.226	1	0	3	3
5	0.04757099	2.059×10^{-2}	0.226	0.226	1	1	3	3
6	0.04680338	1.143×10^{-2}	0.226	0.226	2	1	103	117
7	0.04611002	3.468×10^{-3}	0.226	0.564	2	1	139	195
8	0.04511802	3.255×10^{-3}	0.564	1.411	2	1	117	233
9	0.04494516	1.085×10^{-3}	1.411	3.527	3	1	229	579
10	0.04499733	2.331×10^{-4}	2.838	8.818	6	1	579	949
11	0.04499338	6.211×10^{-5}	0.967	22.045	7	1	2245	1219
12	0.04499329	1.002×10^{-6}	0.127	55.113	8	1	2245	2245
13	0.04499327	7.034×10^{-9}	0.072	137.784	11	1	2245	2245

6.2 Full-Space Results with Iterative Linear System Solves

We now solve the optimal control problem using the full-space trust-region algorithm with iterative augmented system solves described in Section 5.2. The parameters for Algorithm 4 are

tol^{SQP}	tol^{CG}	ζ	Δ_0	Δ_{min}	Δ_{max}	α_1	η_1	η_0	ρ_{-1}	$\bar{\rho}$
10^{-6}	10^{-2}	0.8	10^4	10^{-10}	10^8	0.5	10^{-8}	0.5	1	10^{-8}

The nominal augmented system solver tolerances are set to $\tau^{qn} = \tau^{pg} = \tau^{proj} = \tau^{tang} = \tau^{lmh} = 10^{-6}$; however, we note that these tolerances are adjusted as needed by Algorithm 4. We set $\tau_4 = 2$ and $\tau^{lmg} = 10^4$. To solve augmented systems, we use the flexible generalized minimal residual (F-GMRES) method preconditioned with the Schur-complement preconditioner P discussed in Section 5.2. To apply the augmented system preconditioner P, for each block $v_k C_u^k$, $k = 1, \ldots, Q$, and its adjoint, we use GMRES preconditioned with a non-overlapping additive Schwarz domain-decomposition approach, where the domain D is partitioned into four horizontal strips of roughly equal size (the top strip is the largest one). For the (inner) F-GMRES stopping tolerance, we choose 10^{-4}. The linear solves on the subdomains are performed using a direct solver. As the initial guess for the control variables, we use $z = 0$. To obtain the initial guess for the state variables, we solve the nonlinear state equations with $z = 0$ for each sparse-grid point. We choose the infinite-dimensional view of the Lagrangian for the risk-neutral problem (51), i.e., $v_k = w_k$, for $k = 1, \ldots, Q$. We use a fixed level-3 Clenshaw–Curtis sparse grid with $Q = 2245$.

All studies were executed on the commodity cluster Serrano at Sandia National Labs. The results were obtained using 80 dual-socket 2.1 GHz Intel Broadwell E5-2695 nodes with 128 GB RAM. Each node has 36 cores, amounting to a total of 2880 cores. We utilized the hierarchical parallelism afforded by the Rapid Optimization Library, partitioning the cores into 720 groups, with each group using four cores for linear algebra tasks such as matrix assembly and iterative solves of the linearized state and adjoint equations needed to apply the preconditioner P. These solves are executed concurrently across the 720 groups. Considering that we process 2245 sparse-grid points, each group performs only three or four linearized state and adjoint solves per preconditioner application, enabling a high degree of concurrency in the computation. Informative studies can be performed without high-performance computing resources, using, e.g., level-2 sparse grids, which amounts to changing the "Maximum Sparse Grid Level" parameter in the input scripts described in Section 6.3 from 3 to 2.

Figure 6 shows the expected values of the controlled velocity field (streamlines and magnitude), pressure, and temperature. The left-wall and right-wall controls are given in Figure 7, respectively. As before, we see a significant reduction of vorticity near the heated substrate. In Table 2, we observe that the final objective value is different than in the reduced-space case. Figure 7 reveals that the optimal controls are also different. This is not unexpected, as our optimal control problem is nonconvex and may have multiple local minima. Substituting the full-space optimal controls into the reduced-space method, and vice versa, confirms that these are indeed locally optimal for both methods and that multiple numerical minima exist.

Table 2 shows the iteration history for the full-space approach. The columns from left to right denote the iteration count (iter), the computed objective function value ($J(x_k)$), the norm of the constraint($\|c(x_k)\|_C$), the norm of the gradient of the Lagrangian ($\|\nabla L(x_k, \lambda_k)\|_X$), the trust-region radius (Δ_k), the number of projected conjugate gradient iterations (pcg) per SQP iteration, a Boolean corresponding to whether the step was accepted or rejected (accept), the cumulative number of

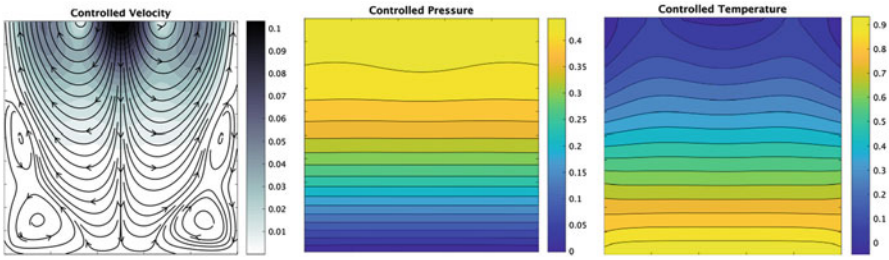

Fig. 6 Expected values of the controlled velocity field (left), pressure (middle), and temperature (right), obtained using the full-space method with iterative augmented system solves

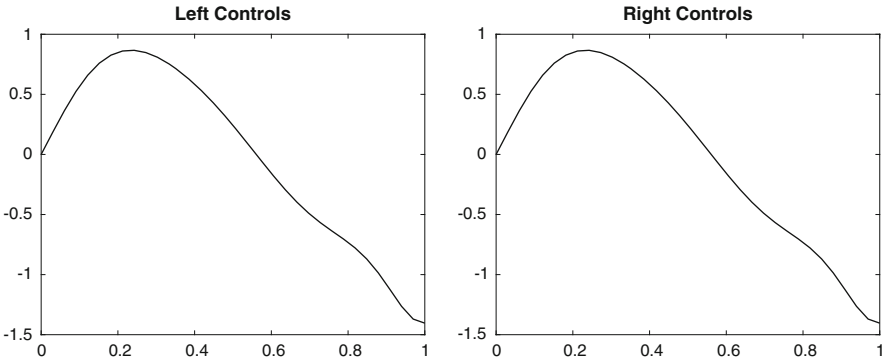

Fig. 7 Optimal controls along the left vertical side wall (left image) and the right vertical side wall (right image) of the problem domain D, obtained using the full-space method with iterative augmented system solves

Table 2 Iteration history for the full-space approach with iterative augmented system solves

iter	$J(x_k)$	$\|c(x_k)\|_C$	$\|\nabla L(x_k, \lambda_k)\|_X$	Δ_k	pcg	accept	ls calls	ls iters
0	0.07484675	7.820623×10^{-15}	8.377793×10^{-3}	1.00×10^{4}	–	–	–	–
1	0.05533699	1.661657×10^{-2}	3.641571×10^{-4}	1.00×10^{4}	11	1	16	597
2	0.03588474	3.052458×10^{-3}	9.338262×10^{-5}	1.00×10^{4}	13	1	33	1292
3	0.03515891	1.017679×10^{-4}	7.117806×10^{-5}	1.00×10^{4}	20	1	56	2303
4	0.03480817	1.444319×10^{-4}	2.439603×10^{-5}	1.00×10^{4}	15	1	75	3108
5	0.03480817	1.444319×10^{-4}	2.439321×10^{-5}	4.08×10^{0}	20	0	98	4157
6	0.03465050	2.237452×10^{-6}	4.364539×10^{-6}	3.03×10^{1}	2	1	104	4438
7	0.03464773	2.716452×10^{-7}	1.042585×10^{-7}	3.03×10^{1}	8	1	116	4989

calls to F-GMRES for augmented system solves (ls calls), and the cumulative number of F-GMRES iterations (ls iters). The first observation is that the full-space scheme converges robustly to the desired tolerance despite inexactness in the augmented system solves and inexactness in the Schur-complement preconditioner applications, i.e., linearized state and adjoint solves. Second, the average number

of P-preconditioned F-GMRES iterations per augmented system solve is roughly 43, which is encouraging considering that the size of the state space alone is 30,900,180 and that we have used fairly tight nominal tolerances for augmented system solves. Nonetheless, opportunities exist for preconditioner research in the context of optimization under uncertainty, and additional studies with larger nominal tolerances for augmented system solves are necessary.

6.3 Reproducibility

The numerical studies are contained in the directory

rol/examples/PDE-OPT/published/IMAvolumes_KouriRidzal2017

of the Rapid Optimization Library. The driver source file for the reduced-space studies is `example_RS.cpp`, with the accompanying input script `input_RS.xml`. The driver source file for the full-space studies is `example_FS.cpp`, with the accompanying input script `input_FS.xml`. The version of the Trilinos git repository used to generate all numerical results is labeled with the commit tag

`3958350daababd03f37fc422bf6a546d2d5ab5f5`,

and the branch is "`develop`." We report results with the Intel 17.0.0.098 compiler; however, we observed virtually identical results with GCC 6.1.0.

7 Conclusions

In recent years, trust-region methods have been extended to rigorously, robustly, and efficiently handle many sources of inexactness in the optimization process, including inexact evaluations of the objective and constraint functions and their derivatives as well as the inexact linear system solves arising in the approximation of constraint derivative inverses. In this chapter, we reviewed in some detail two such methods, which are particularly well suited to the solution of large-scale PDE-constrained optimization problems. The first method tackles the challenges of inexact objective function and gradient evaluations in unconstrained (reduced-space) formulations of PDE-constrained optimization problems, and is demonstrated in the context of sparse-grid adaptivity for risk-neutral optimization of thermal fluids. The second method deals with inexact linear system solves in constrained (full-space) formulations, and is demonstrated on large risk-neutral thermal-fluid optimization problems with fixed sparse grids, but with iterative, and therefore inexact, linearized state and adjoint solves. A principal remaining challenge in inexact trust-region methods for PDE-constrained optimization is in the handling of general inequality constraints on the control and state variables, with research opportunities in formulation and algorithm development, large-scale solvers for optimality systems, and efficient software implementations.

Acknowledgements This work was supported by DARPA EQUiPS grant SNL 014150709 and the DOE NNSA ASC ATDM program.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

1. N. Alexandrov. Robustness properties of a trust-region framework for managing approximation models in engineering optimization. In *Proceedings from the AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, Work-in-progress Paper AIAA–96–4102-CP*, pages 1056–1059, 1996.
2. N. Alexandrov. A trust–region framework for managing approximations in constrained optimization problems. In *Proceedings of the First ISSMO/NASA Internet Conference on Approximation and Fast Reanalysis Techniques in Engineering Optimization, June 14–27, 1998*, 1998.
3. N. Alexandrov and J. E. Dennis. Multilevel algorithms for nonlinear optimization. In J. Borggaard, J. Burkardt, M. D. Gunzburger, and J. Peterson, editors, *Optimal Design and Control*, pages 1–22, Basel, Boston, Berlin, 1995. Birkhäuser Verlag.
4. N. Alexandrov, J. E. Dennis Jr., R. M. Lewis, and V. Torczon. A trust region framework for managing the use of approximation models in optimization. *Structural Optimization*, 15: 16–23, 1998. Appeared also as ICASE report 97–50.
5. E. Arian, M. Fahl, and E. W. Sachs. Trust–region proper orthogonal decomposition for flow control. Technical Report 2000–25, ICASE, NASA Langley Research Center, Hampton VA 23681–2299, 2000.
6. I. Babuška, F. Nobile, and R. Tempone. A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM Rev.*, 52(2):317–355, 2010.
7. V. Barthelmann, E. Novak, and K. Ritter. High dimensional polynomial interpolation on sparse grids. *Adv. Comput. Math.*, 12(4):273–288, 2000. Multivariate polynomial interpolation.
8. F. Bastin, C. Cirillo, and Ph. L. Toint. An adaptive Monte Carlo algorithm for computing mixed logit estimators. *Comput. Manag. Sci.*, 3(1):55–79, 2006.
9. M. Benzi, G. H Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14(1):1–137, 2005.
10. R. G. Carter. Numerical optimization in Hilbert space using inexact function and gradient evaluations. Technical Report 89–45, ICASE, Langley, VA, 1989.
11. R. G. Carter. On the global convergence of trust region algorithms using inexact gradient information. *SIAM J. Numer. Anal.*, 28:251–265, 1991.
12. R. G. Carter. Numerical experience with a class of algorithms for nonlinear optimization using inexact function and gradient information. *SIAM Journal on Scientific Computing*, 14(2):368–388, 1993.
13. A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust–Region Methods*. SIAM, Philadelphia, 2000.
14. J. E. Dennis, M. El-Alem, and M. C. Maciel. A Global Convergence Theory for General Trust–Region–Based Algorithms for Equality Constrained Optimization. *SIAM J. Optimization*, 7:177–207, 1997.
15. J. E. Dennis and V. Torczon. Approximation model managemet for optimization. In *Proceedings from the AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, Work-in-progress Paper AIAA–96–4099-CP*, pages 1044–1046, 1996.

16. J. E. Dennis and V. Torczon. Managing approximation models in optimization. In N. Alexandrov and M. Y. Hussaini, editors, *Multidisciplinary Design Optimization. State of the Art*, pages 330–347, Philadelphia, 1997. SIAM.
17. J. E. Dennis, Jr. and R. B. Schnabel. *Numerical Methods for Nonlinear Equations and Unconstrained Optimization*. SIAM, Philadelphia, 1996.
18. M. Fahl and E.W. Sachs. Reduced order modelling approaches to PDE–constrained optimization based on proper orthogonal decompostion. In L. T. Biegler, O. Ghattas, M. Heinkenschloss, and B. van Bloemen Waanders van Bloemen Waanders, editors, *Large-Scale PDE-Constrained Optimization*, Lecture Notes in Computational Science and Engineering, Vol. 30, Heidelberg, 2003. Springer-Verlag.
19. T. Gerstner and M. Griebel. Dimension-adaptive tensor-product quadrature. *Computing*, 71(1):65–87, 2003.
20. M. Heinkenschloss and D. Ridzal. A matrix-free trust-region SQP method for equality constrained optimization. *SIAM Journal on Optimization*, 24(3):1507–1541, 2014.
21. M. Heinkenschloss and L. N. Vicente. Analysis of inexact trust–region SQP algorithms. *SIAM J. Optimization*, 12:283–302, 2001.
22. M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with Partial Differential Equations*, volume 23 of *Mathematical Modelling, Theory and Applications*. Springer Verlag, Heidelberg, New York, Berlin, 2009.
23. K. Ito and S. S. Ravindran. Optimal control of thermally convected fluid flows. *SIAM J. on Scientific Computing*, 19:1847–1869, 1998.
24. C.T. Kelley and E.W. Sachs. Truncated newton methods for optimization with inaccurate functions and gradients. *Journal of Optimization Theory and Applications*, 116(1):83–98, 2003.
25. D. P. Kouri, M. Heinkenschloss, D. Ridzal, and B. G. van Bloemen Waanders. A trust-region algorithm with adaptive stochastic collocation for PDE optimization under uncertainty. *SIAM Journal on Scientific Computing*, 35(4):A1847–A1879, 2013.
26. D. P. Kouri, M. Heinkenschloss, D. Ridzal, and B. G. van Bloemen Waanders. Inexact objective function evaluations in a trust-region algorithm for PDE-constrained optimization under uncertainty. *SIAM Journal on Scientific Computing*, 36(6):A3011–A3029, 2014.
27. D. P. Kouri, G. von Winckel, and D. Ridzal. ROL: Rapid Optimization Library. https://trilinos.org/packages/rol, 2017.
28. L. Lubkoll, A. Schiela, and M. Weiser. An affine covariant composite step method for optimization with PDEs as equality constraints. *Optimization Methods and Software*, 32(5):1132–1161, 2017.
29. J. J. Moré. Recent developments in algorithms and software for trust region methods. In A. Bachem, M. Grötschel, and B. Korte, editors, *Mathematical Programming, The State of The Art*, pages 258–287. Springer Verlag, Berlin, Heidelberg, New-York, 1983.
30. M. F. Murphy, G. H. Golub, and A. J. Wathen. A note on preconditioning for indefinite linear systems. *SIAM Journal on Scientific Computing*, 21(6):1969–1972, 2000.
31. E. Novak and K. Ritter. High-dimensional integration of smooth functions over cubes. *Numer. Math.*, 75(1):79–97, 1996.
32. E. O. Omojokun. *Trust region algorithms for optimization with nonlinear equality and inequality constraints*. PhD thesis, Department of Computer Science, University of Colorado, Boulder, Colorado, 1989.
33. T. Rees, H. S. Dollar, and A. J. Wathen. Optimal Solvers for PDE-Constrained Optimization. *SIAM Journal on Scientific Computing*, 32(1):271–298, 2010.
34. T. Rees, M. Stoll, and A. Wathen. All-at-once preconditioning in PDE-constrained optimization. *Kybernetika*, 46(2):341–360, 2010.
35. T. Rees and A. J. Wathen. Preconditioning iterative methods for the optimal control of the Stokes equation. *SIAM J. Sci. Comput*, 33(5), 2010.
36. D. Ridzal. Preconditioning of a Full-Space Trust-Region SQP Algorithm for PDE-constrained Optimization. In *Report No. 04/2013: Numerical Methods for PDE Constrained Optimization with Uncertain Data*. Mathematisches Forschungsinstitut Oberwolfach, 2013.

37. S. A. Smoljak. Quadrature and interpolation formulae on tensor products of certain function classes. *Soviet Math. Dokl.*, 4:240–243, 1963.
38. M. Stoll. One-shot solution of a time-dependent time-periodic PDE-constrained optimization problem. *IMA Journal of Numerical Analysis*, 34(4):1554–1577, 2014.
39. M. Stoll and A. Wathen. All-at-once solution of time-dependent Stokes control. *Journal of Computational Physics*, 232(1):498–515, 2013.
40. Ph. L. Toint. Global convergence of a class of trust-region methods for nonconvex minimization in Hilbert space. *IMA Journal of Numerical Analysis*, 8:231–252, 1988.
41. S. Ulbrich and J. C. Ziems. Adaptive multilevel trust-region methods for time-dependent PDE-constrained optimization. *Portugaliae Mathematica*, 74(1):37–67, 2017.
42. D. Xiu and J. S. Hesthaven. High-order collocation methods for differential equations with random inputs. *SIAM J. Sci. Comput.*, 27(3):1118–1139 (electronic), 2005.
43. J. C. Ziems and S. Ulbrich. Adaptive multilevel inexact SQP methods for PDE-constrained optimization. *SIAM Journal on Optimization*, 21(1):1–40, 2011.
44. J. Carsten Ziems. Adaptive Multilevel Inexact SQP-Methods for PDE-Constrained Optimization with Control Constraints. *SIAM Journal on Optimization*, 23(2):1257–1283, 2013.

Numerical Optimization Methods for the Optimal Control of Elliptic Variational Inequalities

Thomas M. Surowiec

Abstract The optimal control of variational inequalities introduces a number of additional challenges to PDE-constrained optimization problems both in terms of theory and algorithms. The purpose of this article is to first introduce the theoretical underpinnings and then to illustrate various types of numerical methods for the optimal control of variational inequalities. For a generic problem class, sufficient conditions for the existence of a solution are discussed and subsequently, the various types of multiplier-based optimality conditions are introduced. Finally, a number of function-space-based algorithms for the numerical solution of these control problems are presented. This includes adaptive methods based on penalization or regularization as well as non-smooth approaches based on tools from non-smooth optimization and set-valued analysis. A new type of projected subgradient method based on an approximation of limiting coderivatives is proposed. Moreover, several existing methods are extended to include control constraints. The computational performance of the algorithms is compared and contrasted numerically.

1 Introduction

The optimal control of variational inequalities is a natural extension of PDE-constrained optimization in which the forward problem or underlying PDE is replaced by a variational inequality or convex variational problem. There are a vast number of applications in which variational inequalities or convex variational problems are used. Perhaps the most well-known disciplines are continuum mechanics

TMS's research was sponsored in part by the DFG grant no. SU 963/1-1 as well as Research Center MATHEON supported by the Einstein Foundation Berlin within project OT1.

T. M. Surowiec (✉)
FB12 Mathematik und Informatik, Philipps-Universität Marburg, Marburg, Germany
e-mail: surowiec@mathematik.uni-marburg.de

H. Antil et al. (eds.), *Frontiers in PDE-Constrained Optimization*, The IMA Volumes in Mathematics and its Applications 163,
https://doi.org/10.1007/978-1-4939-8636-1_4

[32, 52] and mathematical image processing [67, 70, 72]. However, one also finds variational inequalities in such diverse areas as petroleum engineering [85], digital microfluidics [84], and mathematical finance [51].

Just as in PDE-constrained optimization, we are interested in optimizing or controlling the solution to variational inequalities through one or several parameters. For example, in a packaging problem, we might be interested in determining the optimal force distribution to apply to an elastic membrane in order to obtain a desired shape without passing through a fixed obstacle. In the petroleum engineering application mentioned above, one obtains a variational inequality when using the variational formulation for the steady laminar flow of drill mud modelled as a Bingham fluid. Here, one might wish to determine the optimal pressure needed to affect a desired flow regime.

Due to a lack of Fréchet (or even Gâteaux) differentiability of the control-to-state mapping, a reduced space optimization approach will require either concepts from variational analysis or approximation theory not only to prove existence of a solution but also to derive meaningful first-order optimality conditions and develop efficient numerical methods.

The theoretical and algorithmic challenges remain intact when considering a full-space approach, since the feasible set (after reformulating the variational inequality as a complementarity problem) is both non-convex and fails to satisfy the usual constraint qualifications. For example, in PDE-constrained optimization, one typically uses the Slater or Robinson-Zowe-Kurcyusz constraint qualifications, cf. [48, 77], to ensure the existence and boundedness of the set of Lagrange multipliers. In order to handle these degenerate constraints, one may take a penalty or relaxation approach, though recent work suggests that there are viable theoretical approaches to this "full-space" setting that do not rely on smoothing, see [81].

We now give a brief historical development. We caution the reader that this is only a sampling of the sizeable amount of work published over the past five decades. The purpose is merely to provide an understanding of the current motivations in theory and numerical methods, especially the ideas presented in this paper. Variational inequalities were first introduced by Signorini and solved by Fichera [29]. For this article, the early works by Brezis, Lions, and Stampacchia [18, 60] are perhaps the most relevant as they provide us with a sufficient existence and regularity theory. As a starting point for further study, we also suggest the well-known monographs by Kinderlehrer and Stampacchia and Rodrigues [53, 71]. Perhaps the earliest work on the optimal control of variational inequalities is due to Yvon, in which the adaptive penalty technique for variational inequalities (see Section 4.1) was used in order to approximate the control problem by a more tractable, parameter-dependent problem [86, 87]. This was later used by Lions [58, 59] and fully developed by Barbu [7]. The monograph by Barbu [7] contains many important results and techniques that are still used today, see also related contributions in [9–12, 50, 68, 73]. In 1976, Mignot offered an alternative to the smoothing approach by developing concepts of generalized differentiability [63]. There, he was able to prove and obtain an explicit formula for the Hadamard directional differentiability of the control-to-state mapping of the variational inequality and, with this result,

derive first-order optimality conditions. This result later appeared in [64] and is rederived in [45] using techniques from variational analysis. More recent work in the control of variational inequalities has focused on the development of efficient, function-space-based numerical methods and ever more complex settings as in [2, 4, 21–23, 34–36, 39, 40, 45, 49, 56, 65, 66, 80, 83].

Parallel to the infinite-dimensional developments, a great deal of progress has been made on theory and numerical methods for finite-dimensional mathematical programs with equilibrium or complementarity constraints, as evidenced by the well-known manuscripts by Luo, Pang, and Ralph [61] and Outrata, Kočvara, and Zowe [69]. In addition to the references therein, we also mention the approaches in [3, 74], which have since been extended to infinite dimensions and are in part featured in this paper.

In reference to the existing finite-dimensional literature, recent works have begun referring to the problem class considered here as elliptic mathematical programs with equilibrium constraints (or elliptic MPECs). We will henceforth do the same for the sake of brevity.

The article is structured as follows. After introducing some necessary notation, data assumptions, and the canonical example in Section 2, we discuss sufficient conditions for the existence of a solution in Section 3. Afterwards, we give several kinds of multiplier-based first-order stationarity conditions similar to classical Karush-Kuhn-Tucker (KKT) conditions. These may or may not be first-order optimality conditions depending on the regularity and data of the given problem. For a newcomer to elliptic MPECs, this potentially puzzling array of possible stationarity conditions may seem strange. However, it is necessary to understand the gap in what is theoretically the "best" type of KKT point and what type of stationary point a given numerical method can theoretically guarantee (at least asymptotically).

Following the theoretical results in Section 3, we pass to the main focus of our discussion: numerical optimization methods for elliptic MPECs. First, in Section 4, we consider what we refer to as "regularization-based" methods. In Section 4.1, we present a typical adaptive smoothing method that makes use of approximation theory for variational inequalities and which is directly linked to the derivation of first-order stationarity conditions for the elliptic MPEC. Similarly, in Section 4.2, we use a simple penalization of the complementarity condition (in weak form) to obtain a simpler PDE-constrained optimization problem with control and state constraints. The numerical solution of the subproblems in the adaptive penalty method is thoroughly discussed for a canonical example problem and the numerical behavior is illustrated. We note that the smoothed problems become increasingly difficult to solve as the smoothing parameter tends to zero, which in part motivates the desire for "non-smooth" methods.

The structure of the non-smooth numerical methods section is devised to illustrate the parallels to some popular methods for smooth PDE-constrained optimization problems, e.g., projected-gradient methods (Section 5.1), direct solvers for the KKT system (Section 5.2), and globalization of the direct solvers via a line search (Section 5.3). There are obvious (and expected) limitations to these approaches, all of which arise from the non-smooth or degenerate nature of elliptic

MPECs. Nevertheless, they indicate the possibility of inventing new efficient numerical methods beyond the adaptive smoothing/penalty paradigm, which often outperform the smooth methods in practice.

In Section 5.1, we present a new subgradient method for the reduced space problem. The choice of "subgradient" is motivated by the limiting variational calculus found, e.g., in the book by Mordukhovich [66]. We discuss the asymptotic behavior of this method and illustrate the potential for future study. Following this in Section 5.2, an active-set-based solver with feasibility restoration suggested by Hintermüller in [36] as a direct solver for the C-stationarity system is considered. Finally, in Section 5.3, we consider the so-called bundle-free implicit programming approach [46], which can be understood as a globalization of the active-set solver using a line search. We extend the latter to include control constraints. The performance of the non-smooth methods is then compared and contrasted with that of the adaptive penalty method. All numerical examples have been solved in the Julia programming language [14] version 0.5.0 on a 2016 MacBook Pro Intel(R) Core(TM) i5 @ 3.1 GHz with 16 GB 2133 MHz LPDDR3 RAM.

2 Notation, Assumptions, and Preliminary Results

In this section, we fix our notation and analytical framework. Elliptic variational inequalities and elliptic MPECs are introduced.

2.1 Norms, Inner Products, and Convergence

All spaces are based on the real number field. The Euclidean norm and scalar product on $\mathbb{R}^m$ are denoted by $|x|$ and $x \cdot y$ for $x, y \in \mathbb{R}^m$, respectively. For $r \in \mathbb{R}$, we denote $\max(0, r)$ by $(r)_+$. All other norms are denoted by $\|\cdot\|_X$ for some space X. The topological dual of X is denoted by X^* and the dual pairing by $\langle \cdot, \cdot \rangle_{X,X^*}$. The inner product on a Hilbert space H is denoted by $(\cdot, \cdot)_H$. Given a sequence $\{x_k\} \subset X$, we denote strong convergence to $\overline{x} \in X$ by $x_k \overset{X}{\to} \overline{x}$; for weak convergence, we use $x_k \overset{X}{\rightharpoonup} \overline{x}$ and for weak-star convergence, $x_k \overset{X}{\rightharpoonup}^* \overline{x}$. In all these cases, we drop the sub- or superscripts if it is clear in context.

2.2 Extended Real-Valued Functionals and Convex Analysis

For a topological vector space U, the extended real-valued functional $G : U \to \overline{\mathbb{R}}$ is proper if $G(u) > -\infty$ for all $u \in U$ and $G(w) < +\infty$ for some $w \in U$. G is said to be closed or lower-semicontinuous if its epigraph

$$\operatorname{epi} G := \{(u, \alpha) \in U \times \mathbb{R} \mid G(u) \leq \alpha\}$$

is a closed subset in the product topology on $U \times \mathbb{R}$. Moreover, G is convex provided

$$G(\lambda u + (1-\lambda)w)) \leq \lambda G(u) + (1-\lambda)G(w), \quad \forall u, w \in U, \forall \lambda \in [0,1].$$

For a Banach space V and $F : V \to \overline{\mathbb{R}}$, the (convex) subdifferential of F at some point $x \in V$ such that $-\infty < F(x) < +\infty$ is defined by the potentially empty set:

$$\partial F(x) := \left\{x^* \in V^* \,\middle|\, F(y) \geq F(x) + \langle x^*, y - x\rangle, \forall y \in V\right\}.$$

If $C \subset V$, then the indicator functional i_C of C is defined by

$$i_C(x) = 0 \text{ if } x \in C, \quad \text{and } i_C(x) = +\infty \text{ otherwise.}$$

If C is non-empty, closed, and convex, then $\mathcal{N}_C(x) := \partial i_C(x)$ is called the convex normal cone to C at $x \in C$. If $K \subset V$ is a non-empty, closed, convex cone, then

$$\xi \in \mathcal{N}_K(x) \;\Leftrightarrow\; x \in K, \xi \in K^- := \left\{\mu \in V^* \,|\, \langle \mu, y\rangle \leq 0, \forall y \in K\right\} \text{ and } \langle \xi, x\rangle = 0.$$

Recall that K^- referred to as the polar cone to K. For more on convex analysis, see, e.g., [25].

2.3 Function Spaces and 2-Capacity

Unless noted, $\Omega \subset \mathbb{R}^n$, $n \in \{1, 2, 3\}$, is a non-empty, bounded, and open subset with Lipschitz boundary $\Gamma = \partial\Omega$. The Lebesgue measure of $E \subset \mathbb{R}^n$ is denoted by meas(E). For $p \in R$ with $1 \leq p < +\infty$, we denote the usual Lebesgue space of p-integrable functions by $L^p(\Omega)$ and the Lebesgue space of essentially bounded functionals by $L^\infty(\Omega)$. The norms are given by

$$\|u\|_{L^p} = \left(\int_\Omega |u(\omega)|^p d\omega\right)^{1/p} \text{ and } \|u\|_{L^\infty} = \operatorname{ess\,sup}_{\omega\in\Omega} |u(\omega)|,$$

respectively. For $k \in \mathbb{N}$, we denote the Sobolev space of L^p-functions u with $|D^\alpha u| = |(\partial^{\alpha_1} u, \ldots, \partial^{\alpha_n} u)| \in L^p(\Omega)$ by $W^{k,p}(\Omega)$, where $\alpha = (\alpha_1, \ldots, \alpha_n)$ is a multi-index with $|\alpha_1| + \cdots + |\alpha_n| \leq k$ and $\partial^{\alpha_i} u$ is the α_i-th weak partial derivative of u with respect to x_i, $i \in \{1, 2, 3\}$. The $W^{k,p}$-norms are then defined by

$$\|u\|_{W^{k,p}} = \left(\sum_{|\alpha|\leq k} \int_\Omega |D^\alpha u(\omega)|^p d\omega\right)^{1/p} \text{ and } \|u\|_{W^{k,\infty}} = \sum_{|\alpha|\leq k} \operatorname{ess\,sup}_{\omega\in\Omega} |D^\alpha u(\omega)|.$$

For our discussion, the most important case is when $p = 2$. Here, both $L^2(\Omega)$ and $H^k(\Omega) := W^{k,2}(\Omega)$ are Hilbert spaces with inner product defined using the norms given above. We denote the space of all H^1-functions with zero trace by $H_0^1(\Omega)$ and its dual by $H^{-1}(\Omega)$. It follows from the Poincaré inequality that

$$\|u\|_{H_0^1} = \|\nabla u\|_{L^2} = (\nabla u, \nabla u)_{L^2}^{1/2}.$$

is an equivalent norm on $H_0^1(\Omega)$. See, e.g., [1] for more. Finally, recall that the 2-capacity of an arbitrary subset $E \subset \Omega$ is given by

$$\mathrm{Cap}_2(E, \Omega) := \inf \left\{ \|u\|_{H_0^1}^2 : v \in H_0^1(\Omega)\ v \geq 1 \text{ a.e. on open neighborhood } G \supset E \right\}.$$

cf. [6, Prop. 5.8.3]. Note that H_0^1-functions possess a representative that is continuous up to a set of positive capacity, cf. [26]. Moreover, there exist $E \subset \Omega$ with $\mathrm{Cap}_2(E, \Omega) > 0$ and $\mathrm{meas}(E) = 0$, e.g., the boundary of a smooth open set. Hence, an H_0^1-function in 2D is continuous across a smooth curve in the plane, but not at single points.

2.4 Elliptic Variational Inequalities

Let H be a Hilbert space and H^* its topological dual. For this subsection, $(\cdot, \cdot)$ denotes the inner product and $\|\cdot\|$ the norm on H. The pairing between H and H^* is denoted by $\langle \cdot, \cdot \rangle$. Let $a : H \times H \to \mathbb{R}$ be a bilinear form on H and $A : H \to H^*$ the associated bounded linear operator, i.e., $a(u, v) = \langle Au, v \rangle$, $u, v \in H$. We recall that a is said to be coercive/elliptic, if there exists some constant $c > 0$ such that $a(v, v) \geq c\|v\|^2$ for all $v \in H$. Clearly, the function $\|v\|_a := (a(v, v))^{1/2}$ defines an equivalent norm on H. Let $K \subset H$ be non-empty, closed, and convex and $w \in H^*$. Then, the variational problem

$$\text{Find } u \in K : a(u, v - u) \geq \langle w, v - u \rangle, \qquad \text{for all } v \in K \tag{1}$$

is called an elliptic variational inequality. Note that problems of this type are often referred to as variational inequalities of the first kind. They can be equivalently written as a generalized equation:

$$\text{Find } u \in K : Au + \mathcal{N}_K(u) = Au + \partial i_K(u) \ni w. \tag{2}$$

If we replace i_K by a subdifferentiable proper closed convex functional φ, then we obtain a variational inequality of the second kind. The following is due to Lions and Stampacchia [60], see also [53]:

Theorem 1 *Let $a : H \times H \to \mathbb{R}$ be a coercive bilinear form, $K \subset H$ closed and convex, and $w \in H^*$. Then, (1) possesses a unique solution $S(w)$. Moreover, the solution mapping $w \mapsto S(w)$ is Lipschitz continuous: For all $w_1, w_2 \in H^*$, it holds that*

$$\|S(w_1) - S(w_2)\|_H \leq (1/c)\|w_1 - w_2\|_{H^*},$$

where c is the constant of coercivity of a.

Although the solution mapping S is Lipschitz continuous, it is not necessarily differentiable. In some special cases, S is directionally differentiable. For more on variational inequalities, see, e.g., [53]. The conditions on A and K will be considered standing assumptions on the variational inequality for the rest of this article. We conclude this subsection with an example.

Example 1 In the context of (1), let $H := H_0^1(\Omega)$ and $H^* = H^{-1}(\Omega)$. Moreover, define $\Psi \in H^1(\Omega)$ with $\Psi|_\Gamma \leq 0$ and $K \subset H_0^1(\Omega)$ such that

$$K := \left\{ u \in H_0^1(\Omega) \mid u(x) \geq \Psi(x), \quad \text{for almost every (a.e.) } x \in \Omega \right\}.$$

Clearly, since $\max(0, \Psi) \in H_0^1(\Omega)$, $K \neq \emptyset$. Moreover, one readily shows that K is closed and convex in $H_0^1(\Omega)$. The setting of (1) is general enough to allow for nonsymmetric bilinear forms $a(u, v)$, e.g.,

$$a(u, v) := \int_\Omega \sum_{i,j} a_{ij} \partial_i u \partial_j v - \sum_i b_i (\partial_i u) v + cuv \mathrm{d}x, \quad u, v \in H_0^1(\Omega),$$

with appropriate assumptions on a_{ij}, b_i, c. However, for our purposes it suffices to consider

$$a(u, v) := \int_\Omega \nabla u \cdot \nabla v \mathrm{d}x, \quad u, v \in H_0^1(\Omega)$$

Letting $f \in H^{-1}(\Omega)$ and combining the above, we obtain a classical obstacle problem

$$\text{Find } u \in K : \int_\Omega \nabla u \cdot \nabla [v - u] \mathrm{d}x \geq \langle f, v - u \rangle, \quad \text{for all } v \in K.$$

If $\Psi|_\Gamma \equiv 0$, then (without altering the boundary conditions), we obtain an equivalent problem:

$$\text{Find } u \in K_0 : \int_\Omega \nabla u \cdot \nabla [v - u] \mathrm{d}x \geq \langle f + \Delta\Psi, v - u \rangle, \quad \text{for all } v \in K_0,$$

where $K_0 = \left\{ u \in H_0^1(\Omega) \mid u(x) \geq 0, \quad \text{a.e. } x \in \Omega \right\}$; see, e.g., [53, 71].

Of course, this set K represents perhaps the easiest kind of set that one can consider in a variational inequality. It is just one example of a so-called "polyhedric set," a notion introduced in the mid-1970s by Haraux [33] and Mignot [63] in the context of sensitivity analysis of variational inequalities. See the recent detailed study [82] for a state-of-the-art on polyhedricity. For more general closed, convex sets, e.g.,

$$\left\{ u \in H_0^1(\Omega) \mid |(\nabla u)(x)| \leq \Psi(x), \text{ a.e. } x \in \Omega \right\} \tag{3}$$

it is possible to prove existence, uniqueness, and continuity results under weak assumptions on Ψ. Since these sets are neither polyhedric nor cones, the differential sensitivity analysis of the solution map and subsequent derivation of optimality conditions or efficient numerical methods is extremely challenging. Elliptic MPECs with this type of constraint were considered in [45]. However, we caution the reader that the proofs for the derivation of the tangent cones are in fact erroneous. Thus, the differential sensitivity results for the solution mapping only hold under the assumption that the tangent cones do in fact have the form purported in the text.

2.5 *Elliptic Mathematical Programs with Equilibrium Constraints*

We present an abstract framework for a class of elliptic MPECs. Let V, H, and Z be separable Hilbert spaces such that the state space V is a dense subset of H and $V \subset H \subset V^*$ both algebraically and topologically. Moreover, assume that $f \in V^*$, $Z_{\text{ad}} \subset Z$ is a non-empty, closed, and convex set (the set of admissible controls/decision variables), $B : Z \to V^*$ is a bounded linear operator and $F : U \to \overline{\mathbb{R}}$ and $G : Z \to \overline{\mathbb{R}}$. We define an elliptic MPEC as follows:

$$\min J(z, u) := F(u) + G(z) \text{ over } (z, u) \in Z \times V, \tag{4a}$$

$$\text{s.t. } z \in Z_{\text{ad}}, \ u \text{ solves } (1) \text{ with } w := Bz + f. \tag{4b}$$

In light of Theorem 1, we can rewrite (4) in reduced form, analogously to standard PDE-constrained optimization problems:

$$\min \ \mathscr{J}(z) := F(S(Bz)) + G(z) \text{ over } z \in Z_{\text{ad}}. \tag{5}$$

Here, $f \in V^*$ is fixed. Therefore, we only write $S(Bz)$ for the solution mapping (instead of $S(Bz + f)$). Obtaining a meaningful full-space formulation of the optimization problem (4) is not always possible. However, if K in (1) is a cone, then we can also formulate a full-space version of the elliptic MPEC by introducing a slack variable $\xi \in V^*$:

$$\min J(z,u) \text{ over } (z,u,\xi) \in Z \times V \times V^*, \tag{6a}$$

$$\text{s.t. } z \in Z_{\text{ad}}, \tag{6b}$$

$$Au - \xi = Bz + f, \tag{6c}$$

$$u \in K,\ \xi \in K^+,\ \langle \xi, u \rangle = 0. \tag{6d}$$

Here, $K^+ := -K^-$. This follows from (2). Due to the presence of the complementarity condition (6d), problems of the type (6) are often referred to in the finite-dimensional literature (and more recently in the infinite-dimensional literature) as (elliptic) mathematical programs with complementarity constraints (MPCCs); see e.g., [61, 69, 74] (finite dimensions) and the recent work [81] (infinite dimensions).

Remark 1 We consider here the simplest case in which the control or decision variable z appears only on the "right-hand side" of the variational inequality. However, in many interesting applications, e.g., topology optimization [8], the control enters nonlinearly through the differential operator A, e.g., when $A(z)u = -\text{div}(z\nabla u)$. In the context of elliptic MPECs, this case is scarcely covered in the literature, see [36].

We finalize this section with a canonical example of an elliptic MPEC.

Example 2 In the notation of (4), we set $V = H_0^1(\Omega)$, $H = L^2(\Omega)$, $V^* = H^{-1}(\Omega)$, and $Z = L^2(\Omega)$. For some $u_d \in L^2(\Omega)$, we let F be a standard tracking-type functional and G an L^2-Tikhonov regularization:

$$J(z,u) := F(u)+G(z) = \frac{1}{2}\|u-u_d\|_{L^2}^2+\frac{\alpha}{2}\|z\|_{L^2}^2, \quad (z,u) \in L^2(\Omega) \times H_0^1(\Omega),\ \alpha > 0.$$

For the forward problem, we choose the simplest form of the obstacle problem with $\Psi \equiv 0$. Then, the following is an elliptic MPEC:

$$\min \frac{1}{2}\|u-u_d\|_{L^2}^2 + \frac{\alpha}{2}\|z\|_{L^2}^2, \text{ over } (z,u) \in L^2(\Omega) \times H_0^1(\Omega), \tag{7a}$$

$$\text{s.t. } z \in Z_{\text{ad}},\ u \in K_0 \text{ solves } : \int_\Omega \nabla u \cdot \nabla[v-u]\mathrm{d}x \geq \langle Bz+f, v-u\rangle, \quad \text{for all } v \in K_0, \tag{7b}$$

Some possibilities for Z_{ad} are local bilateral constraints: $-1 \leq w(x) \leq 1$, a.e. $x \in \Omega$ or global constraints such as $\|w\|_{L^2} \leq 1$.

3 Existence and Stationarity Conditions

We start this section in the abstract framework of Section 2.5. In order to provide insight into the deeper meaning of the various stationarity conditions, we ultimately restrict ourselves to the canonical example (Example 2).

3.1 Existence of a Solution

Under suitable conditions, on F, G, Z_{ad}, and B, we can use Weierstrass's existence theorem, see, e.g., [6, Thm. 3.2.5] to prove that the reduced form MPEC has a solution. For the following result, we appeal to the monograph by Barbu [7, Chap. 3].

Theorem 2 *In the context of problem* (4)*, we assume the following:*

1. $F : H \to \mathbb{R}$ *is locally Lipschitz and nonnegative.*
2. $G : Z \to \overline{\mathbb{R}}$ *is convex, lower-semicontinuous, and for some constants* $\kappa_1 > 0$, $\kappa_2 \in \mathbb{R}$. $G(z) \geq \kappa_1 \|z\|_Z + \kappa_2$, $\forall z \in Z$
3. B *is completely continuous.*

Then, (4)*, or equivalently* (5)*, admits a solution.*

Proof By replacing $G(z)$ with $G(z) + i_{Z_{ad}}(z)$, the assertion follows from [7, Prop. 3.1]. In particular, we note that by Barbu [7, Lem. 3.1] the mapping $(S \circ B) : Z \to V$ is completely continuous and, furthermore, the composite objective functional $(F \circ S \circ B) : Z \to \mathbb{R}_+$ is weakly lower-semicontinuous.

Since for any $z \in Z_{\text{ad}}$, (1) has a unique solution $u = S(Bz)$. There exists a unique $\xi \in \mathcal{N}_K(u)$ such that $\xi = Au - Bz - f$. Conversely, if (z, u, ξ) satisfies (6b)–(6d), then $z \in Z_{\text{ad}}$ and $u = S(Bz)$. In other words, there is a one-to-one correspondence between the feasible set of (5) and (6). Therefore, if (4), or equivalently (5), admits a solution, then so does (6).

Corollary 1 *Under the assumptions of Theorem 2,* (6) *admits a solution.*

Remark 2 Ignoring Theorem 2, we could also prove this assertion by appealing to [6, Thm. 3.2.5]. Indeed, under the hypotheses of Theorem 2, (6) can be viewed as an unconstrained problem in which the associated objective functional is radially unbounded, proper, and weakly lower-semicontinuous.

3.2 Primal Stationarity Conditions

In this subsection, we restrict ourselves to the setting of Example 2. As mentioned earlier, there are some situations in which the control-to-state mapping S is directionally differentiable. In particular, Mignot [63] demonstrated that the control-to-state mapping in Example 2 is directionally differentiable in the sense of Hadamard. The next result draws the parallel to PDE-constrained optimization problems.

Theorem 3 *Let* $(\overline{z}, \overline{u})$ *be a solution to* (7)*, then the following first-order necessary optimality condition holds:*

$$\alpha(\overline{z}, w - \overline{z}) + (\overline{u} - u_d, S'(\overline{u}; B(w - \overline{z}))) \geq 0, \quad \forall w \in Z_{\text{ad}}. \tag{8}$$

In the finite-dimensional literature, e.g., in [61, 74], primal first-order conditions of the type (8) are often referred to as B-stationarity conditions. There, the "B" comes from the fact that either the so-called Bouligand tangent cone or Bouligand differentiability is used. This motivates the next definition.

Definition 1 If $(\overline{z}, \overline{u})$ is a feasible point of (7) that satisfies (8), then $(\overline{z}, \overline{u})$ is B-stationary.

For the setting in (5), it is possible to derive B-stationarity conditions provided that F, G are smooth. Clearly, the main challenge is determining directional differentiability of S. Suppose for the sake of argument that $S'(\overline{u}; du)$ is linear in $du \in V^*$. Then, we can "dualize" the B-stationarity conditions (8):

$$\alpha(\overline{z}, w - \overline{z}) + (B^* p, w - \overline{z}) \geq 0, \quad \forall w \in Z_{\text{ad}}, \quad p = S'(\overline{u})^*(u - u_d).$$

Note that the condition "$p = S'(\overline{u})^*(u - u_d)$" essentially means (in this ideal case) that p solves a type of adjoint equation. Using this "adjoint state" p, we could now develop solution algorithms. However, $S'(\overline{u}; du)$ is in general nonlinear (albeit Lipschitz) in du and thus, the standard adjoint state p does not exist. As a result, there are several different types of dual/multiplier-based first-order stationarity conditions reminiscent of classical KKT conditions in nonlinear programming.

3.3 Dual Stationarity Conditions

In this subsection, we again restrict ourselves to the setting in Example 2. Furthermore, we assume that Γ is regular enough to guarantee that $S(Bz) \in H^2(\Omega) \cap H_0^1(\Omega)$ for any $z \in Z_{\text{ad}}$. This is possible if, e.g., $Bz + f \in L^2(\Omega)$ and Ω is a convex polyhedron, cf. [18, 71]. This regularity assumption ensures that $\xi \in L^2(\Omega)$ and, by the Sobolev embedding theorem, $S(Bz)$ is at least Hölder continuous on Ω. We now introduce two multiplier-based first-order stationarity conditions taken from [39, 40].

Definition 2 Let $z \in Z_{\text{ad}}$ and $u = S(Bz)$ with associated slack variable $\xi \in L^2(\Omega)$. The set $\Omega^0 := \{x \in \Omega \mid u(x) = 0\}$ is called the active set or coincidence set and $I := \{x \in \Omega \mid u(x) > 0\}$ the inactive set. The set $\Omega^{00} := \Omega^0 \cap \{x \in \Omega \mid \xi(x) = 0\}$ is called the biactive or weakly active set and $\Omega^{0+} := \Omega^0 \cap \{x \in \Omega \mid \xi(x) > 0\}$ the strongly active set.

Note that the biactive and strongly active sets are only defined up to a set of Lebesgue measure zero. In contrast, the active and inactive sets may be more finely defined up to sets of positive capacity even in less regular settings.

Definition 3 The point $(z, u, \xi) \in L^2(\Omega) \times H_0^1(\Omega) \times L^2(\Omega)$ is called a C-stationary point for (7) provided that there exist $p \in H_0^1(\Omega)$ and $\lambda \in H^{-1}(\Omega)$ such that the following system of equations is fulfilled:

$$\alpha(\overline{z}, w - \overline{z}) + (B^* p, w - \overline{z}) \geq 0, \quad \forall w \in Z_{\text{ad}} \tag{9a}$$

$$A^* p - \lambda = u_d - u \tag{9b}$$

$$Au - \xi = Bz + f \tag{9c}$$

$$\xi \geq 0, \text{ a.e. } x \in \Omega \tag{9d}$$

$$u \geq 0, \text{ a.e. } x \in \Omega \tag{9e}$$

$$(\xi, u) = 0 \tag{9f}$$

$$\langle \lambda, p \rangle \leq 0 \tag{9g}$$

$$p = 0, \text{ a.e. } x \in \Omega^{0+} \tag{9h}$$

$$\langle \lambda, \phi \rangle = 0, \forall \phi \in H_0^1(\Omega) : \phi = 0, \text{ a.e. } \Omega^0 \tag{9i}$$

If, in addition to (9), p and λ satisfy:

$$p \leq 0, \text{ a.e. } x \in \Omega^{00} \tag{10a}$$

$$\langle \lambda, \phi \rangle \geq 0, \forall \phi \in H_0^1(\Omega) : \phi \geq 0, \text{ a.e. } \Omega^{00}, \phi = 0, \text{ a.e. } \Omega \setminus (I \cup \Omega^{00}), \tag{10b}$$

then $(z, u, \xi) \in L^2(\Omega) \times H_0^1(\Omega) \times L^2(\Omega)$ is called an S-stationary point.

Several comments on this system are necessary. To start, we note that if K_0 is replaced by the entire space $H_0^1(\Omega)$, then ξ, λ and the conditions (9d)–(10b) vanish and we obtain the usual first-order system for a linear-quadratic PDE-constrained optimization problem with convex control constraints. Moreover, if there is no biactive set, then both (in fact all) dual stationarity concepts coincide. Thus, biactivity is essentially the source of all difficulties in the study of MPECs. We will see later that it also relates to the Gâteaux differentiability of the control-to-state map.

Perhaps the most critical point here is the usage of almost everywhere versus quasi everywhere conditions for the multipliers. As defined above, these S-stationarity conditions are not equivalent to the necessary first-order optimality conditions in the pioneering works [63, 64]; unless certain regularity assumptions on the active/biactive/inactive sets are made. In [63, 64] (see also [43, 80]), the notion of capacity (in contrast to Lebesgue measure) is used. This is because capacity is needed for the correct representation of the tangent/contingent cone and its polar. When replacing q.e. by a.e. some conditions become stronger and others finer.

In order to see this, suppose that $\text{meas}(\Omega^0) = 0$, but $\text{Cap}_2(\Omega^0, \Omega) > 0$. Then, (9i) would imply $\lambda = 0 \in H^{-1}(\Omega)$. However, when using capacity, since the test functions ϕ admit quasi-continuous representatives, requiring $\phi = 0$ "quasi-everywhere" (q.e.) on Ω^0 would shrink the set of test functions so that λ is not necessarily zero. Similar problems arise by requiring a.e. instead of q.e. conditions on the adjoint variables p. Indeed, if Ω^{0+} has measure zero, then (9h) is trivially

satisfied by any p; this would not be the case if we use q.e. since $p \in H_0^1(\Omega)$ admits a quasi-continuous representative.

Therefore, the stationarity conditions of Mignot and Puel, see (29) below, should be considered the true strong/S-stationarity conditions for our canonical MPEC as opposed to (9)–(10). For more complex constraints in which the image space of the constraint mapping is an L^p-space, as in, e.g.,[34, 35, 45], the problems with defining active sets up to sets of capacity zero seem to be absent.

Nevertheless, (9)–(10) arose naturally through the limiting process of an adaptive penalty scheme. As such, they are highly relevant for the study of numerical methods for elliptic MPECs, especially for methods utilizing smoothing plus continuation. In contrast, it is unclear how to properly include notions of capacity into efficient numerical methods (without simplifying assumptions as in [46]).

Finally, in many complex real-world applications, it might be impossible to obtain even C-stationarity conditions of this type, due to a lack of compactness and regularity properties. For example, in the optimal control of electrowetting on dielectrics[4], Allen-Cahn [27, 28], Cahn-Hilliard [20], or Cahn-Hilliard-Navier-Stokes system with obstacle potentials [47], one can usually only derive an approximation of (9g). Here, this would be equivalent to replacing (9g) by

$$\limsup_{k\to\infty} \langle \lambda_k, p_k \rangle \leq 0$$

for sequences $\{\lambda_k\}$ and $\{p_k\}$ with $\lambda_k \rightharpoonup \lambda$ and $p_k \rightharpoonup p$.

In conclusion, the various notions of dual stationarity and the theory needed to derive them are still active areas of research. The main point here is that there is a stratification of concepts that one should also be aware of when considering the design and convergence of numerical methods. That is, it is not enough to prove that a scheme converges but also to what kind of stationary point. Of course, many of the issues involving capacity, weak topologies, products of weakly convergent sequences, and weak lim-inf or lim-sup-type sign conditions will not necessarily appear in numerical experiments.

4 Regularization-Based Methods

In this and the coming sections, we present a number of numerical optimization methods for elliptic MPECs. These are split into two classes: Methods that employ adaptive smoothing, relaxation, or penalization of the forward problem or complementarity constraints (regularization-based methods) and those that do not (non-smooth methods).

We also note that the proper discretization of elliptic MPECs using (adaptive) finite elements schemes that take into account the additional difficulties due to the inherent degeneracy/non-smoothness has only been considered in a handful of papers. For example, we mention the most recent works [16, 17, 24, 37, 62]. For

our numerical study, we make use of a simple finite difference scheme to discretize the operators along with a nested grid approach to simulate mesh refinements. This allows us to easily compare the methods.

The regularization-based methods all follow a similar scheme. First, the variational inequality is approximated by a parameter-dependent semilinear elliptic PDE or the complementarity constraint is relaxed or penalized. This yields in both cases a more tractable family of approximating optimization problems. Next, the smooth PDE-constrained problems are solved, yielding a parameter-dependent KKT point. Finally, continuation is performed on the regularization parameter (passing to 0 or $+\infty$).

The non-smooth methods are rather different and there is still plenty of room for new ideas. We only mention here that with these methods the emphasis is placed on directly solving the original, non-smooth problem without changing the forward problem. We postpone further details until later.

4.1 An Adaptive Penalty Method

We begin in the abstract framework of Section 2.5 and present a general approximation result as found in [7, Thm. 2.2]. For a comprehensive study on the numerical analysis and approximation of variational inequalities, see, e.g., [31]. In the following, ϕ^ϵ refers to some penalty functional for the constraint K.

Definition 4 For any constant $\epsilon > 0$, let $\phi^\epsilon : V \to \mathbb{R}$ such that ϕ^ε is convex and Fréchet differentiable on V and satisfies

1. There exists a C, independent of ϵ, u with $\phi^\epsilon(u) \geq -C(\|u\|_V + 1)$ for all $\varepsilon > 0$, $u \in V$.
2. $\phi^\epsilon(u) \to i_K(u)$ as $\epsilon \downarrow 0$ for all $u \in V$.
3. For all $u \in V$ and for all $\{u_\epsilon\}$ such that $u_\epsilon \rightharpoonup u$ as $\epsilon \downarrow 0$, $\liminf_{\epsilon \downarrow 0} \phi^\epsilon(u_\epsilon) \geq i_K(u)$.

Remark 3 See also [7, Thm 2.4] for functionals on H, which is more relevant for (13).

In the abstract setting, we approximate (1) by

$$Au + \nabla \phi^\epsilon(u) = Bz + f \tag{11}$$

We denote the approximate solution mapping by S_ϵ. For the next result, we restrict ourselves to the case when A is symmetric (as in Example 2), see [7, Thm. 2.2].

Theorem 4 *Let $w_\epsilon \to w$ in V^* as $\epsilon \downarrow 0$, then the sequence $\{u_\epsilon\} \subset V$ with $u_\epsilon := S_\epsilon(w_\epsilon)$ converges weakly to $u = S(w)$ as $\epsilon \downarrow 0$. If in addition,*

$$\langle \nabla\phi^\epsilon(y) - \nabla\phi^\delta(v), y - v\rangle \geq -C(1 + (\|\nabla\phi^\epsilon(y)\|^2_{V^*} + \|\nabla\phi^\delta(v)\|^2_{V^*}))(\epsilon + \delta),$$
$$\forall \epsilon, \delta > 0, \forall y, v \in V, \tag{12}$$

then $u_\epsilon \to u$ *(strongly in* V*) as* $\epsilon \downarrow 0$.

In the setting of Example 2, one possibility for ϕ^ϵ is:

$$\phi^\epsilon(u) := (2\epsilon)^{-1}\|(-u)_+\|^2_{L^2}. \tag{13}$$

Redefining K_0 for L^2-functions, ϕ^ϵ in (13) is the Moreau-Yosida regularization of the indicator functional i_{K_0} with respect to the L^2-topology, cf. [5, Sec. 2.7]. However, in order to solve the smoothed MPECs numerically, we will require more smoothness of ϕ^ϵ below. We now approximate (5) by:

$$\min \mathscr{J}_\epsilon(z) := F(S_\epsilon(Bz)) + G(z) \text{ over } z \in Z_{\text{ad}}, \tag{14}$$

Thus, the existence theory, optimality conditions, and numerical methods for (14) reduces to results from PDE-constrained optimization.

Turning to Example 2, we employ a smoothed plus function in (11) characterized by:

$$\nabla\phi^\epsilon(u) = \begin{cases} \epsilon^{-1}u - 0.5, & \text{if } u \geq \epsilon, \\ \frac{u^3}{\epsilon^3} - \frac{u^4}{2\epsilon^4}, & \text{if } u \in (0,\epsilon), \\ 0, & \text{else,} \end{cases} \quad \nabla^2\phi^\epsilon(u) = \begin{cases} \epsilon^{-1}, & \text{if } u \geq \epsilon, \\ \frac{3u^2}{\epsilon^3} - \frac{2u^3}{\epsilon^4}, & \text{if } u \in (0,\epsilon), \\ 0. & \text{else.} \end{cases} \tag{15}$$

We can then prove that the approximate solution mapping S_ϵ is Fréchet differentiable. Then, for a solution (z_ϵ, u_ϵ), there exists an adjoint state $p_\epsilon \in V$ such that

$$\alpha(z_\epsilon, w - z_\epsilon) + (B^* p_\epsilon, w - z_\epsilon) \geq 0, \quad \forall w \in Z_{\text{ad}} \tag{16a}$$
$$A^* p_\epsilon - \lambda_\epsilon = u_d - u_\epsilon, \tag{16b}$$
$$Au_\epsilon - \xi_\epsilon, = Bz_\epsilon + f, \tag{16c}$$
$$\xi_\epsilon = \nabla\phi^\epsilon(-u_\epsilon), \text{ a.e. } x \in \Omega \tag{16d}$$
$$\lambda_\epsilon = -\nabla^2\phi^\epsilon(-u_\epsilon)p_\epsilon, \text{ a.e. } x \in \Omega, \tag{16e}$$

holds; cf. the techniques in [57, 77].

We note here that the case of gradient constraints mentioned in (3) can also be treated using such a penalty method. However, using a standard quadratic penalty/Moreau-Yosida-type approach yields a penalized state equation of the form:

$$-\text{div}((1 + \xi_\epsilon)\nabla u_\epsilon) = Bz_\epsilon + f, \tag{17a}$$
$$\xi_\epsilon = \frac{1}{\epsilon}\left(1 - \frac{\Psi}{|\nabla u_\epsilon|}\right)_+, \tag{17b}$$

which is considerably more challenging due to the bilinear dependence on u_ϵ in the PDE and accompanying non-smooth first-order PDE.

From (16), it is possible to derive C-stationarity conditions and, under further assumptions, even S-stationarity conditions, cf. [39, 40], by passing to the limit in ϵ. More specifically, we show that a subsequence of stationary points for (16) converges to a point that satisfies C-stationarity for (7). Therefore, if we have a numerical method that solves (16), then by performing continuation on $\epsilon \downarrow 0$ we have a means of numerically approximating C-stationary points (or better) for (7). This furnishes a convergence proof in function space for the "outer loop" of the method, depicted in Algorithm 4.1. For the interested reader, we provide a short discussion of the limiting arguments in Detail 5.

Algorithm 4.1 Adaptive penalty method: outer loop

Input: u_d, $f \in L^2(\Omega)$, $\beta \in (0,1)$;
1: Choose $\epsilon^0 > 0$, $(z^0, u^0, \xi^0, p^0, \lambda^0) \in L^2(\Omega) \times H_0^1(\Omega) \times L^2(\Omega) \times H_0^1(\Omega) \times L^2(\Omega)$ and set $k := 0$;
2: **repeat**
3: Compute a stationary point $(z^{k+1}, u^{k+1}, \xi^{k+1}, p^{k+1}, \lambda^{k+1})$ of (16) with $\epsilon = \epsilon^k$ using an iterative scheme with initial value $(z^k, u^k, \xi^k, p^k, \lambda^k)$;
4: Set $\epsilon^{k+1} = \beta\epsilon^k$
5: **until** some stopping rule is satisfied.

Detail 5 (Sketch of the Limiting Technique) We first note that (16a) is equivalent to

$$z_\epsilon = \mathrm{Proj}_{Z_{\mathrm{ad}}}(-\frac{1}{\alpha}B^* p_\epsilon). \tag{18}$$

Suppose Z_{ad} is bounded, then $\{z_\epsilon\}_{\epsilon>0}$ is bounded in Z. Let $\epsilon_k \downarrow 0$. Then, there exists a subsequence $\{z_l\}$ with $z_l = z_{\epsilon_{k_l}} \overset{Z}{\rightharpoonup} z^\star \in Z_{ad}$. Hence, $S_{\epsilon_{k_l}}(Bz_l) =: u_l \overset{V}{\to} u^\star = S(Bz^\star)$ by Theorem 4 due to the complete continuity of B. Therefore, $\xi_l \to \xi^\star := Bz^\star + f - Au^\star \in -\mathscr{N}_{K_0}(u^\star)$, i.e., $\xi^\star, u^\star$ satisfy. For the adjoint state p_ϵ and multiplier λ_ϵ, we test the adjoint equation (16b) with p_ϵ:

$$c_1\|p_\epsilon\|_V^2 \le \langle A^* p_\epsilon - \lambda_\epsilon, p_\epsilon\rangle = (u_d - u_\epsilon, p_\epsilon) \le \\ \|u_d - u_\epsilon\|_H \|p_\epsilon\|_H \le c_2\|u_d - u_\epsilon\|_H \|p_\epsilon\|_V$$

for scalars $c_1, c_2 > 0$, independent of ϵ. Since $u_l \overset{V}{\to} u^\star$, $\{p_l\}$ is bounded in V. Hence, there exists $\{p_{l_m}\}$ with $p_{l_m} \overset{V}{\rightharpoonup} p^\star$ and, by substitution, $\lambda_{l_m} \overset{V^*}{\rightharpoonup} \lambda^\star = u^\star - u_d + A^* p^\star$.

Without knowledge of the best possible system, one might stop at this point; however, the theory indicates that a much more refined system is possible. We quickly demonstrate (9g). The remaining conditions (9h), (9i) require lengthy

arguments that go beyond the scope of this work. Returning to (16b), it is clear that

$$\langle A^* p_\epsilon, p_\epsilon\rangle - (u_d - u_\epsilon, p_\epsilon) = \langle \lambda_\epsilon, p_\epsilon\rangle = -\int_\Omega \nabla^2\phi^\epsilon(-u_\epsilon)|p_\epsilon|^2 \mathrm{d}x \leq 0.$$

Since A is symmetric, the bilinear form $a(\cdot,\cdot)$ is weakly lower-semicontinuous. Moreover, due to the Rellich-Kondrachov theorem, both $u_\epsilon \xrightarrow{H} u^\star$, $p_\epsilon \xrightarrow{H} p^\star$. Thus,

$$\langle \lambda^\star, p^\star\rangle = \langle A^* p^\star, p^\star\rangle - (u_d - u^\star, p^\star) \leq \liminf_{m\to+\infty} \langle A^* p_{l_m}, p_{l_m}\rangle - (u_d - u_{l_m}, p_{l_m}) \leq 0. \tag{19}$$

In [7], a more general setting allows a similar argument. However in general, e.g., if A is not symmetric, we only have:

$$\limsup_{\epsilon\downarrow 0} \langle \lambda_\epsilon, p_\epsilon\rangle \leq 0$$

as mentioned at the end of Section 3.3. □

A possible stopping rule in Algorithm 4.1 could be the residual of the C-stationarity system (9) or when a minimum value of ϵ^{k+1} is reached, e.g., machine precision. Obviously, the most computationally demanding part here is step 3. Assuming that $\mathrm{Proj}_{Z_{\mathrm{ad}}}$ is relatively simple to calculate, e.g., for local bilateral constraints, then (16) reduces to

$$A^* p_\epsilon + \nabla^2\phi^\epsilon(-u_\epsilon)p_\epsilon = u_d - u_\epsilon, \tag{20a}$$

$$Au_\epsilon - \nabla\phi^\epsilon(-u_\epsilon) = B\mathrm{Proj}_{Z_{\mathrm{ad}}}(-\frac{1}{\alpha}B^* p_\epsilon) + f. \tag{20b}$$

One could then solve (20) using, e.g., a semismooth Newton method. Alternatively, an interior point approach for the projection might be employed. We briefly recall the semismooth Newton method in infinite-dimensional spaces as discussed in [19, 38, 79]. Let X, Y be Banach spaces, $D \subset X$ an open subset of X, and $\mathscr{F} : D \to Y$.

Definition 5 The mapping $\mathscr{F} : D \subset X \to Y$ is said to be Newton-differentiable on the open subset $U \subset D$, if there exists a family of mappings $\mathscr{G} : U \to \mathscr{L}(X, Y)$ such that

$$\|\mathscr{F}(x+h) - \mathscr{F}(x) - \mathscr{G}(x+h)h\|_Y = o(\|h\|_X), \forall x \in U.$$

$\mathscr{G}$ is called the Newton derivative for $\mathscr{F}$ on U. In [38], it is shown that

$$\mathscr{G}_\delta(y)(x) = \begin{cases} 1 \text{ if } y(x) > 0 \\ 0 \text{ if } y(x) < 0 \\ \delta \text{ if } y(x) = 0 \end{cases} \tag{21}$$

for every $y \in X$ and $\delta \in \mathbb{R}$ is a Newton derivative of the $\max(0, \cdot)$, under the condition that $\max(0, \cdot) : L^p(\Omega) \to L^q(\Omega)$ with $1 \leq q < p \leq \infty$.

Therefore, if a mapping $\mathscr{F}$ is Newton-differentiable, then using the concepts above leads to a generalized Newton step for the equation $\mathscr{F}(x) = 0$, see, e.g., [19, 38].

Theorem 6 *Suppose that $\mathscr{F}(x^*) = 0$ and that $\mathscr{F}$ is Newton-differentiable on an open neighborhood U of x^* with Newton derivative $\mathscr{G}$. If $\mathscr{G}(x)$ is nonsingular for all $x \in U$ and the set $\left\{ ||\mathscr{G}(x)^{-1}||_{\mathscr{L}(Y,X)} : x \in U \right\}$ is bounded, then the semismooth Newton iteration*

$$x_{l+1} = x_l - \mathscr{G}(x_l)^{-1}\mathscr{F}(x_l), \quad l = 0, 1, 2, \ldots \tag{22}$$

converges superlinearly to x^*, provided that $||x_0 - x^*||_X$ is sufficiently small.

Under the assumption that $Z_{\text{ad}} := \left\{ v \in L^2(\Omega) \mid a \leq v \leq b, \text{ a.e. } x \in \Omega \right\}$ with $a, b \in L^2(\Omega)$ and $a < b$, we can derive a semismooth Newton step for the solution of (20) in function space: Fix some $(u, p) \in H_0^1(\Omega) \times H_0^1(\Omega)$ and define the following subsets of Ω:

$$\Omega^a := \left\{ x \in \Omega \,\middle|\, a(x) + \alpha^{-1}(B^* p)(x) > 0 \right\},$$
$$\Omega^b := \left\{ x \in \Omega \,\middle|\, -\alpha^{-1}(B^* p)(x) - b(x) > 0 \right\}.$$

Moreover, let $\Omega^{ina} := \Omega \setminus (\Omega^a \cup \Omega^b)$ (up to a set of Lebesgue measure zero) and define the residual $\mathscr{F}_\epsilon(u, p)$ of (20) by:

$$\mathscr{F}_\epsilon^1(u, p) := Au - \nabla\phi^\epsilon(-u) - B\text{Proj}_{Z_{\text{ad}}}(-\frac{1}{\alpha}B^* p) - f,$$
$$\mathscr{F}_\epsilon^2(u, p) := A^* p + \nabla^2\phi^\epsilon(-u)p - u_d + u.$$

Since

$$z = \text{Proj}_{Z_{\text{ad}}}(-\alpha^{-1}B^* p) = -\alpha^{-1}B^* p - (-\alpha^{-1}B^* p - b)_+ + (a + \alpha^{-1}B^* p)_+, \tag{23}$$

we can use (21) to obtain a Newton derivative $\mathscr{G}$ for $\mathscr{F}$:

$$\mathscr{G}_\epsilon(u, p) = \begin{bmatrix} A + \nabla^2\phi^\epsilon(-u) & \alpha^{-1}\chi_{\Omega^{ina}}BB^* \\ I - \nabla^3\phi^\epsilon(-u)p & A^* + \nabla^2\phi^\epsilon(-u) \end{bmatrix},$$

where $\chi_{\Omega^{ina}}$ is the characteristic function for the set Ω^{ina}. If $(\delta u, \delta p)$ denotes the difference between the new iterate and the current iterate in the semismooth Newton step, then at each iteration, we solve

$$\begin{bmatrix} A + \nabla^2\phi^\epsilon(-u) & \alpha^{-1}\chi_{\Omega^{ina}} BB^* \\ I - \nabla^3\phi^\epsilon(-u)p & A^* + \nabla^2\phi^\epsilon(-u) \end{bmatrix} \begin{bmatrix} \delta u \\ \delta p \end{bmatrix} = -\mathscr{F}_\epsilon(u, p). \tag{24}$$

If we can show that $\mathscr{G}_\epsilon(u, p)$ is invertible independently of (u, p) (for fixed $\epsilon > 0$) so that the set $\left\{||\mathscr{G}_\epsilon(u, p)^{-1}|| : (u, p) \in H_0^1(\Omega) \times H_0^1(\Omega)\right\}$ is bounded, then we are guaranteed to have local superlinear convergence. This leads to Algorithm 4.2.

Algorithm 4.2 Adaptive penalty method: inner loop

Input: tol $> 0, a, b, u_d, f \in L^2(\Omega), \epsilon_k > 0, (u_0^k, p_0^k) \in H_0^1(\Omega) \times H_0^1(\Omega), l := 0;$

1: **repeat**
2: Compute a step $(\delta u_l^k, \delta p_l^k)$ by solving (24) with $(u, p) := (u_l^k, p_l^k), \epsilon = \epsilon_k$.
3: Set $u_{l+1}^k := u_l^k + \delta u_l^k, p_{l+1}^k := p_l^k + \delta p_l^k, l := l + 1;$
4: **until** $\|\mathscr{F}(u_l^k, p_l^k)\| <$ tol

We conclude this subsection with a numerical experiment. This is used in part to compare to the non-smooth numerical methods in later sections.

Example 3 Let $\Omega = (0, 1)^2$, $\alpha = 1$, $b \equiv 0.035$, $a \equiv 0$, and $A = -\Delta$ (associated with $H_0^1(\Omega)$). Defining

$$y^\dagger(\mathbf{x}_1, \mathbf{x}_2) = \begin{cases} 160(\mathbf{x}_1^3 - \mathbf{x}_1^2 + 0.25\mathbf{x}_1)(\mathbf{x}_2^3 - \mathbf{x}_2^2 + 0.25\mathbf{x}_2) & \text{in } (0, 0.5) \times (0, 0.5), \\ 0 & \text{else,} \end{cases}$$

$$\xi^\dagger(\mathbf{x}_1, \mathbf{x}_2) = \max(0, -2|\mathbf{x}_1 - 0.8| - 2|\mathbf{x}_1\mathbf{x}_2 - 0.3| + 0.5),$$

we set

$$f = -\Delta y^\dagger - y^\dagger - \xi^\dagger, \text{ and } u_d = y^\dagger + \xi^\dagger - \alpha\Delta y^\dagger.$$

The example is chosen due to the nontrivial biactive set and overlap of active sets for the control and state constraints. The only change to [40, Exp. 5.1] is the addition of control constraints, which were set to $\pm\infty$ there. Concerning the discretization and solution, we use a standard five-point stencil to discretize the negative Laplace operator with finite differences. The problem is solved on a uniform mesh with 512^2 grid points.

We start the algorithm at $(z^0, u^0, \xi^0) = (0, 0, 0)$. The stopping criterion is based on the L^2-norm of the residual with stopping tolerance of 10^{-9}. For this example, the ϵ-update in 4.1 proved to be extremely sensitive, meaning that a reasonably aggressive update strategy, e.g., $\epsilon_{k+1} = \epsilon_k/2$ failed once $\epsilon_k = \mathscr{O}(10^{-4})$. To be fair, one would normally not cold-start this algorithm on such a fine grid. Opting instead

for either an adaptive FEM strategy, multigrid scheme (as in [40]), or a nested grid strategy (as in [39, 46]) would certainly improve the performance and allow for a more aggressive update strategy for the smoothing parameter. Nevertheless, we see that as the penalized problem approaches the original non-smooth non-convex problem, the nonlinear system becomes increasingly more difficult to solve (eventually even failing). See Table 1 for the convergence history and Figure 1 for plots of the solution.

4.2 An ℓ^1 Penalty Method

In this subsection, we present a technique originating in the finite-dimensional MPEC literature [3]. The extension to infinite dimensions can be found here [43]. The idea is elegant in its simplicity and allows us to approximate the elliptic MPEC by a sequence of PDE-constrained optimization problems with control and state constraints. Moreover, instead of a semilinear elliptic PDE, as in the previous method, we have a linear elliptic PDE. We begin in the abstract framework of Section 2.5 under the assumption that K is a cone and then pass to the problem in Example 2.

Using an ℓ^1-penalty for the condition $\langle u, \xi \rangle = 0$, we approximate (6) by

$$\min J(z,u) + \frac{1}{\epsilon}|\langle u, \xi \rangle| \text{ over } (z,u,\xi) \in Z \times V \times V^*, \tag{25a}$$

$$\text{s.t. } z \in Z_{\text{ad}}, \tag{25b}$$

$$Au - \xi = Bz + f, \tag{25c}$$

$$u \in K,\ \xi \in K^+. \tag{25d}$$

By definition of K^+, $\langle u, \xi \rangle \geq 0$. Hence, for $(u, \xi) \in K \times K^+$, $\frac{1}{\epsilon}|\langle u, \xi \rangle| = \frac{1}{\epsilon}\langle u, \xi \rangle$, which yields the smooth objective $J_\epsilon(z,u,\xi) := J(z,u) + \frac{1}{\epsilon}\langle u, \xi \rangle$.

The analysis for this problem requires several technical results. Nevertheless, under appropriate regularity and boundedness assumptions, one can still show existence of a solution, consistency of the approximation, derive first-order conditions and (after passing to the limit in ϵ) obtain a (weak) form of C-stationarity. We briefly sketch the ideas here and refer the reader to [43, Section 2] for the detailed technical analysis in the context of Example 2.

Detail 7 (Sketch of Existence and Consistency Arguments) In addition to the assumptions in Theorem 2, let F be weakly lower-semicontinuous, Z_{ad} bounded, and A symmetric.

To show that (25) has a solution, we prove the boundedness of infimizing sequence $\{(z_k, u_k, \xi_k)\}$ (despite the unboundedness of K^+ and lack of coercivity of $J_\epsilon(z,u,\xi)$). Let (z_0, u_0, ξ_0) be feasible for (6). Then, for all sufficiently large $k \in N$,

Table 1 Convergence history of the adaptive penalty method Algorithm 4.1 with stopping tolerance tol = 10^{-9}, starting $\epsilon_1 = 1$ using $\epsilon_k = 2^{-(k-1)}$ with $k = 1, \ldots, 12$, l represents the iterations per inner loop of Algorithm 4.2

l \ ϵ	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}	2^{-9}	2^{-10}	2^{-11}
1	3.037e−04	2.284e−03	1.547e−02	7.270e−02	1.581e−01	2.013e−01	2.247e−01	2.381e−01
2	2.120e−04	1.169e−02	4.510e−01	5.556e+00	1.852e+01	3.929e+01	7.506e+01	1.407e+02
3	1.524e−09	4.459e−06	5.586e−03	6.919e−01	6.179e+00	1.755e+01	3.830e+01	7.712e+01
4	1.550e−12	1.856e−12	8.915e−07	8.776e−03	4.555e−01	2.449e+00	7.160e+00	1.683e+01
5			1.561e−12	1.441e−06	2.453e−03	4.973e−02	2.747e−01	9.366e−01
6				1.824e−12	8.096e−08	2.439e−05	4.920e−04	3.762e−03
7					1.817e−12	7.139e−12	2.038e−09	9.028e−08
8							4.209e−12	6.215e−06
9								4.915e−06
⋮								⋮
33								1.649e−09
34								1.193e−09
35								8.641e−10

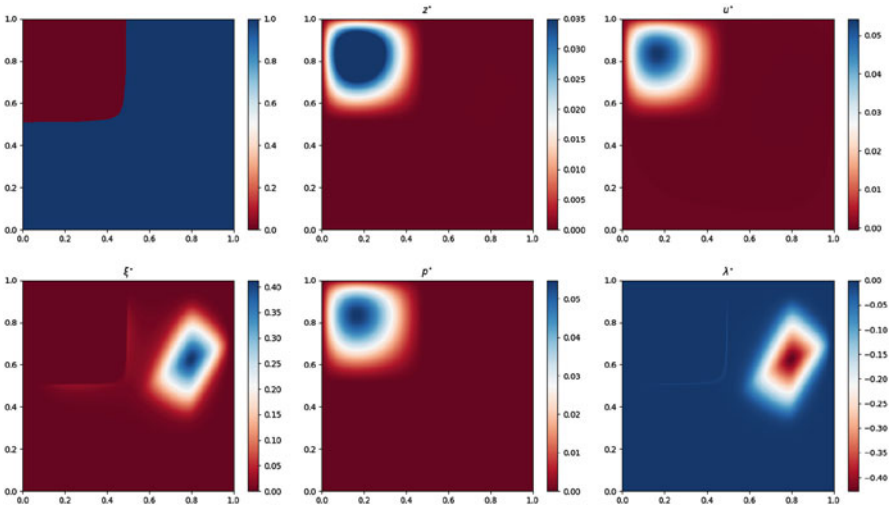

Fig. 1 Solution plots for adaptive penalty method Algorithms 4.1 and 4.2, clockwise from upper left: characteristic function $\chi_{\Omega_\star^0}$ optimal control $z^\star$, state $u^\star$, multiplier $\lambda^\star$, adjoint $p^\star$, and multiplier $\xi^\star$. Notice the nontrivial biactive set for the variational inequality and the upper and lower active sets for the control constraints

$$J(z_0, u_0) = J_\epsilon(z_0, u_0, \xi_0) \geq$$

$$J_\epsilon(z_k, u_k, \xi_k) = F(u_k) + G(z_k) + \frac{1}{\epsilon}\langle u_k, \xi_k\rangle \geq G(z_k) + \frac{1}{\epsilon}\langle u_k, \xi_k\rangle. \tag{26}$$

Hence, $J(z_0, u_0) - G(z_k) \geq \frac{1}{\epsilon}\langle u_k, \xi_k\rangle \geq 0$. Since Z_{ad} is bounded, there exists a weakly convergent subsequence $\{z_{k_l}\}$ with $z_{k_l} \rightharpoonup z^\star \in Z_{\text{ad}}$. Then, by the weak lower-semicontinuity of G, we have

$$0 \leq \liminf_l \langle u_{k_l}, \xi_{k_l}\rangle \leq \limsup_l \langle u_{k_l}, \xi_{k_l}\rangle \leq J(z_0, u_0) - G(z^\star).$$

Thus, testing (25c) with u_{k_l}, we first obtain the boundedness of $\{u_{k_l}\}$ in V and then $\{\xi_{k_l}\}$ in V^*. Given A is symmetric, we use an argument as in (19) to prove that (along a further subsequence $\{k_{l_m}\}$) $\liminf_m \langle u_{k_{l_m}}, \xi_{k_{l_m}}\rangle \geq \langle u^\star, \xi^\star\rangle$, where $u^\star$ and $\xi^\star$ are the weak limit points. Since K and K^+ are weakly closed, $u^\star \in K$ and $\xi^\star \in K^+$. This suffices to prove that (25) has a solution. A proof for consistency of the approximation, i.e., that global optimizers $\{(z_\epsilon, u_\epsilon, \xi_\epsilon)\}$ converge as $\epsilon \downarrow 0$ (at least along a subsequence) to a global optimizer $\{(z^\star, u^\star, \xi^\star)\}$ of (6) can be derived analogously using the boundedness of $\{z_\epsilon\} \subset Z_{\text{ad}}$ and subsequently, the inequality $\epsilon(J(z_0, u_0) - G(z_\epsilon)) \geq \langle u_\epsilon, \xi_\epsilon\rangle \geq 0$. □

As in Section 4.1, we restrict ourselves to the setting of Example 2 for the derivation of stationarity conditions. Here, the derivation of first-order optimality conditions requires the verification of a constraint qualification, in this case that

of Robinson-Zowe-Kurcyusz, see [88]. According to Propositions 2.5 and 2.6 in [43], we have the following ϵ-dependent first-order optimality conditions: For any solution $(z_\epsilon, u_\epsilon, \xi_\epsilon)$ to (25) (under the data assumptions of Example 2), there exists a multiplier-tuple $(p_\epsilon, \vartheta_\epsilon, \tau_\epsilon)$ such that

$$\alpha(z_\epsilon, w - z_\epsilon) + (B^* p_\epsilon, w - z_\epsilon) \geq 0, \quad \forall w \in Z_{\text{ad}} \tag{27a}$$

$$A^* p_\epsilon + \frac{1}{\epsilon}\xi_\epsilon - \vartheta_\epsilon = u_d - u_\epsilon, \tag{27b}$$

$$Au_\epsilon - \xi_\epsilon = Bz_\epsilon + f, \tag{27c}$$

$$u_\epsilon \in K_0, \ \xi_\epsilon \in K_0^+, \tag{27d}$$

$$\frac{1}{\epsilon}u_\epsilon - p_\epsilon - \tau_\epsilon = 0, \tag{27e}$$

$$\vartheta_\epsilon \in K_0^+, \ \langle \vartheta_\epsilon, u_\epsilon \rangle = 0, \ \tau_\epsilon \in K_0, \ \langle \xi_\epsilon, \tau_\epsilon \rangle = 0. \tag{27f}$$

In comparison, this system is much larger than (16) and contains additional information. Nevertheless, under certain boundedness assumptions, passing to the limit in ϵ yields, in fact, a weaker C-stationarity system, see [43, Thm. 2.9]:

$$\alpha(z^\star, w - z^\star) + (B^* p^\star, w - z^\star) \geq 0, \quad \forall w \in Z_{\text{ad}} \tag{28a}$$

$$A^* p^\star - \lambda^\star = u_d - u^\star, \tag{28b}$$

$$Au^\star - \xi^\star = Bz^\star + f, \tag{28c}$$

$$u^\star \in K_0, \ \xi^\star \in K_0^+, \ \langle u^\star, \xi^\star \rangle = 0, \tag{28d}$$

$$\langle \lambda^\star, u^\star \rangle = 0, \ \langle p^\star, \xi^\star \rangle = 0, \ \langle \lambda^\star, p^\star \rangle \leq 0. \tag{28e}$$

To see that (28e) relates to (9h), (9i) suppose for the sake of argument that $\xi^\star$, $p^\star$, $\lambda^\star$ are merely vectors of length n and the conditions in (28e) are understood as the componentwise (Hadamard) products. Then, by complementarity, $\xi^\star \geq 0$ and $\xi_i^\star = 0$ if i is an inactive or biactive index and $\langle p^\star, \xi^\star \rangle = 0$ in turn implies that the strongly active components of $p^\star$ are zero (as in (9h)). The same applies to the inactive components of $\lambda^\star$ due to $\langle \lambda^\star, u^\star \rangle = 0$ (as in (9i)).

Though there is a significant gap, the derivation of (28) is related to the convergence of a function-space-based numerical method. Indeed, using known results for linear elliptic PDE-constrained optimization problems with control and state constraints, we have viable efficient algorithms that can guarantee convergence to a KKT point, which satisfies (27). By performing continuation on ϵ we can be assured to converge (along a subsequence) to a weak C-stationary point. Furthermore, Theorem 2.12 in [43] provides a more compelling argument.

Theorem 8 (Thm 2.12 [43]) *Suppose* $(z_\epsilon, u_\epsilon, \xi_\epsilon, p_\epsilon, \vartheta_\epsilon, \tau_\epsilon)$ *satisfies* (27) *and that* $(z_\epsilon, u_\epsilon, \xi_\epsilon)$ *is feasible for* (7). *Then,* $(z_\epsilon, u_\epsilon, \xi_\epsilon)$ *is strongly stationary in the sense of Mignot and Puel, i.e., conditions* (9h)–(10b) *are replaced by*

$$p = 0, \ q.e.\ x \in \Omega^{0+} \tag{29a}$$

$$\langle \lambda, \phi \rangle = 0, \forall \phi \in H_0^1(\Omega) : \phi = 0, \ q.e.\ \Omega^0 \tag{29b}$$

$$p \leq 0, \ q.e.\ x \in \Omega^{00} \tag{29c}$$

$$\langle \lambda, \phi \rangle \geq 0, \forall \phi \in H_0^1(\Omega) : \phi \geq 0, \ q.e.\ \Omega^{00}, \phi = 0, \ q.e.\ \Omega \setminus (I \cup \Omega^{00}) \tag{29d}$$

Since we are using an ℓ^1 penalty, which often amounts to an exact penalty function, there is a good chance that a stationary point $(z_\epsilon, u_\epsilon, \xi_\epsilon)$ is feasible for (7) for sufficiently small ϵ.

Algorithm 4.3 ℓ^1 penalty method: outer loop

Input: $u_d, f \in L^2(\Omega), \beta \in (0, 1)$;

1: Choose $\epsilon^0 > 0$, $(z^0, u^0, \xi^0, p^0, \vartheta^0, \tau^0) \in L^2(\Omega) \times H_0^1(\Omega) \times L^2(\Omega) \times H_0^1(\Omega) \times H^{-1}(\Omega) \times H_0^1(\Omega)$ and set $k := 0$;

2: **repeat**

3: Compute a stationary point $(z^{k+1}, u^{k+1}, \xi^{k+1}, p^{k+1}, \vartheta^{k+1}, \tau^{k+1})$ of (27) with $\epsilon = \epsilon^k$ using an iterative scheme with initial value $(z^k, u^k, \xi^k, p^k, \vartheta^k, \tau^k)$;

3: Set $\epsilon^{k+1} = \beta \epsilon^k$

4: **until** some stopping rule is satisfied.

Again we might choose as a stopping criterion the residual of C-stationarity, taking $\lambda := \epsilon^{-1}\xi - \vartheta$ and substituting $\tau = \epsilon^{-1}u - p$. Since the theory only guarantees $\vartheta \in H^{-1}(\Omega)$, one will need to treat the discrete quantities carefully, cf. [43] for more details. In addition, the solution of the subproblems (25) does not reduce to the solution of a system similar to (20). Therefore, the solution of the subproblems can be more difficult than in the adaptive penalty framework. Nevertheless, the theoretical result in Theorem 8 indicates the potential of this algorithm to generate a better stationary point.

5 Non-Smooth Numerical Methods

In this section, we present several methods that do not require a smoothing or penalization of the original MPEC. We present a new approximate projected subgradient method alongside a direct solver for the C-stationarity system presented in [36] and a recent method from [46] that may serve as a globalization of the direct solver. In all of these methods, we need to solve the variational inequality (1). This can be done using the semismooth Newton methods as in [38, 41] or special monotone multigrid methods as in [55].

5.1 An Approximate Projected Subgradient Method

The subgradient method is perhaps the simplest non-smooth optimization algorithm. Suppose X is a real separable Hilbert space. Given a proper, convex, lower-semicontinuous, and subdifferentiable functional $f : X \to \mathbb{R}$, $x^0 \in X$, $g^0 \in \partial f(x^0)$, and a sequence $\{\nu_k\}$ with $\nu_k > 0$, compute a sequence of iterates $\{x^k\}$ according to the rule

$$x^{k+1} := x^k - \nu_k g^k, \quad g^{k+1} \in \partial f(x^k).$$

Here, and unless otherwise noted below, we would need to apply the Riesz map to g^k before using it in this iteration.

Similarly, given a non-empty, closed, and convex subset $C \subset X$ the projected-gradient method replaces the previous rule by

$$x^{k+1} := \mathrm{Proj}_C\,(x^k - \nu_k g^k), \quad g^{k+1} \in \partial f(x^{k+1}). \tag{30}$$

The subgradient method was invented by N.Z. Shor in the 1970s, see [76], and although it does not guarantee descent of the objective functional and can be quite slow to converge, it still finds a wide array of applications due to its simplicity and ability to be combined with distributed algorithm techniques, as in, e.g., [78].

Consider the reduced elliptic MPEC in (5):

$$\min\ \mathscr{J}(z) := F(S(Bz)) + G(z) \text{ over } z \in Z_{\mathrm{ad}}.$$

Since S is nonlinear, the reduced objective functional is typically non-convex. Therefore, we cannot directly apply the projected subgradient method. However, there exist a number of generalized subdifferentials for non-convex functions. In our case, we will initially make use of the limiting subdifferential (also known as the Mordukhovich subdifferential) for the reduced objective. We will restrict ourselves to the framework of Example 2. First, we recall several definitions from variational analysis, see [66].

Definition 6 (Normal Cones to Arbitrary Sets) Let X be a Hilbert space and $C \subset X$. Then, the multifunction $\widehat{\mathscr{N}_C} : X \rightrightarrows X^*$ defined by

$$\widehat{\mathscr{N}_C}(x) := \left\{ x^* \in X^* \;\middle|\; \langle x^*, x'-x\rangle_X \le o(||x'-x||_X),\ \forall x' \overset{X}{\to} x, x' \in C \right\}, \quad x \in C, \tag{31}$$

and $\widehat{\mathscr{N}_C}(x) := \emptyset$ for $x \notin C$ is called the regular (Fréchet) normal cone to C. The multifunction $\mathscr{N}_C : X \rightrightarrows X^*$ defined by

$$\mathscr{N}_C(x) := \left\{ x^* \in X^* \;\middle|\; \exists x_k \overset{X}{\to} x,\ \exists x_k^* \overset{X^*}{\rightharpoonup} x^* :\ x_k^* \in \widehat{\mathscr{N}_C}(x_k),\ \forall k \in \mathbb{N} \right\} \tag{32}$$

is called the limiting (Mordukhovich) normal cone to C.

Although $\widehat{\mathcal{N}_C}$ is convex, it fails to admit a satisfactory calculus needed for most non-smooth, non-convex problems. In contrast, the limiting normal cone enjoys a robust calculus. Note that for closed convex sets C, both cones agree, and in general $\widehat{\mathcal{N}_C}(x) \subsetneq \mathcal{N}_C(x)$. We will use the limiting normal cone to define a generalized subgradient, needed in part for our proposed numerical method.

Definition 7 (Limiting Subdifferential) Let X be a Hilbert space, $\phi : X \to \overline{\mathbb{R}}$, and $x \in X$ such that $|\phi(x)| < +\infty$. The set

$$\partial\phi(x) := \left\{ x^* \in X^* \,\middle|\, (x^*, -1) \in \mathcal{N}_{\operatorname{epi}\phi}(x, \phi(x)) \right\} \tag{33}$$

is called the limiting (Mordukhovich) subdifferential. If $|\phi(x)| = \infty$, we set $\partial\phi(x) = \emptyset$.

Therefore, if we know the limiting subdifferential $\partial \mathscr{J}(z)$ in (5), then we could design a projected subgradient iteration along the lines of (30):

$$z^{k+1} := \operatorname{Proj}_{Z_{\text{ad}}}(z^k - \nu_k g^k), \quad g^{k+1} \in \partial \mathscr{J}(z^{k+1}).$$

If $\mathscr{J}$ were smooth, then $\partial \mathscr{J}$ is just the gradient of the reduced objective functional, which we usually calculate in PDE-constrained optimization by solving an adjoint equation. However here, $\mathscr{J}$ is non-smooth and non-convex. In order to obtain a generalized adjoint state for the reduced objective functional we require the so-called "coderivatives."

Definition 8 (Coderivatives) Let X be a Hilbert space, $\Phi : X \rightrightarrows X^*$, and $y \in \Phi(x)$, i.e., $(x, y) \in \operatorname{Graph}\Phi$. The regular (Fréchet) coderivative of Φ at (x, y) is the multifunction $\widehat{D}^*\Phi(x, y) \colon Y^* \rightrightarrows X^*$ defined by

$$h^* \in \widehat{D}^*\Phi(x, y)(d^*) \iff (h^*, -d^*) \in \widehat{\mathcal{N}}_{\operatorname{Graph}\Phi}(x, y). \tag{34}$$

The limiting (Mordukhovich) coderivative $D^*\Phi(x, y)$ of Φ at $(x, y) \in \operatorname{Graph}\Phi$ is similarly defined by

$$h^* \in D^*\Phi(x, y)(d^*) \iff (h^*, -d^*) \in \mathcal{N}_{\operatorname{Graph}\Phi}(x, y). \tag{35}$$

For example, if $\Phi = S_\epsilon$ from Section 4.1, then the coderivatives coincide and we have:

$$\widehat{D}^*S_\epsilon(w, u)(w^*) = D^*S_\epsilon(w, u)(w^*) = S'_\epsilon(w)^*w^*, \quad w^* \in X^*,$$

i.e., $\widehat{D}^*S_\epsilon(w, u)(w^*)$ yields the usual adjoint state p obtained by solving the associated linear elliptic PDE with $-w^*$ on the right-hand side.

For the tracking-type objective in Example 2, it was argued in [44, Prop. 1] that

$$\partial \mathscr{J}(z) \subset \alpha z + B^*D^*S(Bz, u)(u - u_d). \tag{36}$$

Here, it follows from [66, Thm 4.44] that $p \in D^*S(Bz, u)(u - u_d)$ is a solution to the generalized adjoint equation:

$$A^* p + D^* \mathcal{N}_{K_0}(u, \xi)(p) \ni u_d - u, \tag{37}$$

where $\xi = Bz + f - Au \in \mathcal{N}_{K_0}(u)$. Therefore, assuming that (37) were solvable, we could fashion our projected subgradient method as in Algorithm 5.4.

Algorithm 5.4 Limiting projected subgradient algorithm

Input: $\{\nu_k\}$ with $\nu_k > 0$; $z^0 \in Z$; $u^0 = S(Bz^0)$; $\xi^0 = Bz^0 + f - Au^0$; Find a solution p^0 to

$$A^* p + D^* \mathcal{N}_{K_0}(u^0, \xi^0)(p) \ni u_d - u^0.$$

1: **for** $k = 0, 1, \ldots$ **do**
2: Set $z^{k+1} := \mathrm{Proj}_{Z_{\mathrm{ad}}}(z^k - \nu_k g^k)$ with $g^k = \alpha z^k + B^* p^k$.
3: Set $u^{k+1} := S(Bz^{k+1})$, $\xi^{k+1} := Bz^{k+1} + f - Au^{k+1}$.
4: Find p^{k+1} a solution to

$$A^* p + D^* \mathcal{N}_{K_0}(u^{k+1}, \xi^{k+1})(p) \ni u_d - u^{k+1}.$$

5: **end for**

Since we have no efficient means of handling $D^* \mathcal{N}_{K_0}(u^{k+1}, \xi^{k+1})(p)$, Algorithm 5.4 is impractical, especially when we consider that subgradient methods potentially require many iterations even for favorable convex problems. We therefore propose an alternative in Algorithm 5.5. The formal derivation for the approximate generalized adjoint state is based on simple geometric observations for a related finite-dimensional setting.

Consider that the (convex) normal cone $\mathcal{N}_{K_0}$ is generated by the (convex) subdifferential of the indicator functional i_{K_0}. Since the functionals ϕ^ϵ in (13) converge in a variational sense to i_{K_0}, they provide a viable candidate for approximating elements of $\mathcal{N}_{K_0}$. Moreover, for any $u \in V$, $\nabla \phi^\epsilon(u) = -\epsilon^{-1}(-u)_+$. Comparing as $\epsilon \downarrow 0$, it appears (at least in finite dimensions) that Graph $\nabla \phi^\epsilon \to$ Graph $-\mathcal{N}_{K_0}$, see Figure 2. This behavior transfers to $\widehat{\mathcal{N}}_{\mathrm{Graph}\,-\mathcal{N}_{K_0}}(u, \xi)$ and $\widehat{\mathcal{N}}_{\mathrm{Graph}\,\nabla\phi^\epsilon}(u, \xi)$, cf. Figure 2 with $\Theta := \mathrm{Graph}\,-\mathcal{N}_{\mathbb{R}_+}$ and $\Lambda := \mathrm{Graph}\,\nabla\phi^{0.1}$.

Using the information in Figure 2, we can first calculate the limiting normal cones $\mathcal{N}_\Theta(u, \xi)$, from which we obtain: $D^* \mathcal{N}_\Theta(1, 0)(p) = \{0\}$, for all p, and

$$D^* \mathcal{N}_\Theta(0, 1)(p) = \mathbb{R}, \text{ if } p = 0,$$

otherwise $D^* \mathcal{N}_\Theta(0, 1)(p) = \emptyset$. The most interesting case is:

$$D^* \mathcal{N}_\Theta(0, 0)(p) = \begin{cases} \{0\}, & \text{for all } p < 0, \\ \mathbb{R}, & \text{if } p = 0 \\ \mathbb{R}_-, & \text{else.} \end{cases}$$

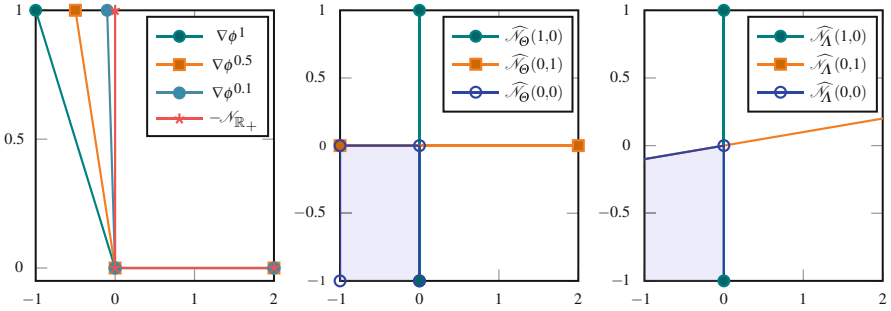

Fig. 2 From left to right: Graph $\nabla\phi^\epsilon$ ($\epsilon = 1, 0.5, 0.1$) versus Graph $\mathcal{N}_{\mathbb{R}_+}$; $\widehat{\mathcal{N}}_{\text{Graph}\,-\mathcal{N}_{\mathbb{R}_+}}(u,\xi)$ for $(u,\xi) = (1,0), (0,1), (0,0)$; $\widehat{\mathcal{N}}_{\text{Graph}\,\nabla\phi 0.1}(u,\xi)$ for $(u,\xi) = (1,0), (0,1), (0,0)$

Similarly, we have for all p

$$D^*\mathcal{N}_\Lambda(1,0)(p) = \{0\}, \quad D^*\mathcal{N}_\Lambda(0,1)(p) = -\epsilon^{-1}p;$$

and

$$D^*\mathcal{N}_\Lambda(0,0)(p) = \begin{cases} \{0\} \cup \{-\epsilon^{-1}p\}, & \text{for all } p < 0, \\ \{0\}, & \text{if } p = 0, \\ \left[-\epsilon^{-1}p, 0\right] & \text{else.} \end{cases}$$

Though certainly more tractable numerically, $D^*\mathcal{N}_\Lambda$ is still a set-valued mapping with non-convex images. We therefore suggest the following single-valued mapping:

$$\widetilde{D}^*\mathcal{N}_\Lambda(0,0)(p) := \left\{q : q = -\epsilon^{-1}\chi_{\{u=0\}}p\right\}$$

where $\chi_{\{u=0\}}$ is the characteristic function for the active set. This mapping coincides with the limiting coderivative $D^*\mathcal{N}_\Theta$ on the inactive set and strongly active set, whenever $D^*\mathcal{N}_\Theta$ is non-empty. On the biactive set, it is either contained in $D^*\mathcal{N}_\Theta$ or approaches it for $\epsilon \downarrow 0$, cf. Figure 3.

By extrapolating these ideas from this simple one-dimensional geometric study to the infinite-dimensional setting, we arrive at our proposed algorithm.

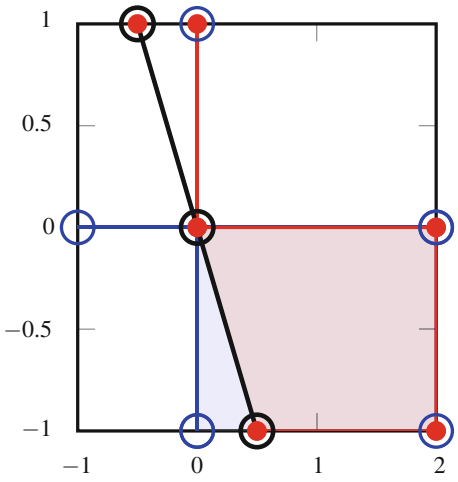

Fig. 3 Blue: $D^*\mathscr{N}_\Theta(0,0)(p)$, Red: $D^*\mathscr{N}_\Lambda(0,0)(p)$, Black: $\widetilde{D}^*\mathscr{N}_\Lambda(0,0)(p)$; $\epsilon = 0.5$

Algorithm 5.5 Approximate projected subgradient algorithm

Input: $\{\epsilon_k\}$ such that $\epsilon_k \downarrow 0$; $\{\nu_k\}$ such that $\nu_k > 0$.
Input: $z^0 \in Z$; $u^0 = S(Bz^0)$; $\xi^0 = Bz^0 + f - Au^0$; Solve for p^0:

$$A^* p + \frac{1}{\epsilon_0}\chi_{\{u^0 \le 0\}} p = u_d - u^0.$$

1: **for** $k = 0, 1, \ldots$ **do**
2: Set $z^{k+1} := \mathrm{Proj}_{Z_{\mathrm{ad}}}(z^k - \nu_k g^k)$ with $g^k = \alpha z^k - B^* p^k$.
3: Set $u^{k+1} := S(Bz^{k+1})$, $\xi^{k+1} := Bz^{k+1} + f - Au^{k+1}$
4: Solve for p^{k+1}:

$$A^* p + \frac{1}{\epsilon_{k+1}}\chi_{\{u^{k+1} \le 0\}} p = u_d - u^{k+1}.$$

5: **end for**

Remark 4 In fact, $-\chi_{\{u \le 0\}} p$ is nothing more than the adjoint of the Newton derivative for the non-smooth Nemytskii operator $(-\cdot)_+ : H_0^1(\Omega) \to L^2(\Omega)$.

Using the analytical techniques described throughout the text, we have the next result.

Theorem 9 *Suppose that Z_{ad} is bounded. Then, any sequence $\{(z^k, u^k, \xi^k, p^k, \lambda^k)\} \subset Z \times V \times V^* \times V \times V^*$ generated by Algorithm 5.5 is bounded. Here,*

$$\lambda_k := -\frac{1}{\epsilon_k}\chi_{\{u^k \le 0\}} p^k.$$

Moreover, any weak accumulation point of $(z^\star, u^\star, \xi^\star, p^\star, \lambda^\star)$ *will be feasible for* (7) *and we have*

$$\lim_{k\to\infty} \int_{\{u^k=0\}} |p^k|^2 \mathrm{d}x = 0, \quad \langle \lambda^\star, p^\star \rangle \leq 0, \quad \langle \lambda^\star, u^\star \rangle = 0. \tag{38}$$

Proof Since Z_{ad} is bounded, $\{z^k\}$ is bounded in Z. It immediately follows that $\{u^k\}$ is bounded, since

$$c\|u^k\|_V^2 \leq \langle Au^k, u^k \rangle - \langle \xi^k, u^k \rangle = \langle Bz^k + f, u^k \rangle \Rightarrow c\|u^k\|_V \leq \|Bz^k + f\|_{V^*}$$

and $\xi^k = Bz^k + f - Au^k$. Similarly, we obtain the boundedness of $\{p^k\}$ in V and $\{\lambda^k\}$ in V^*:

$$c\|p^k\|_V^2 \leq \langle A^* p^k, p^k \rangle + \langle \lambda^k, p^k \rangle = \langle u_d - u^k, p^k \rangle \Rightarrow c\|p^k\|_V \leq \|u_d - u^k\|_{V^*}$$

and $\lambda^k = u_d - u^k - A^* p^k$. Furthermore, since

$$0 \leq \epsilon_k \langle A^* p^k, p^k \rangle + \int_{\{u^k=0\}} |p^k|^2 \mathrm{d}x = \epsilon_k \langle u_d - u^k, p^k \rangle$$

the limit condition in (38) holds. Moreover, will can again show that $\langle \lambda^\star, p^\star \rangle \leq 0$ using the same argument as in (19) and by definition $\langle \lambda^k, u^k \rangle = 0$. Since B is compact, $u^k \to u^\star$ in V (along a subsequence). This yields $\langle \lambda^\star, u^\star \rangle = 0$.

The purpose of Theorem 9 is to show that Algorithm 5.5 produces a sequence with a weak accumulation point that satisfies a kind of limiting C-stationarity system. However, the lag in indices prevents us from closing the argument by proving that $\{z^k\}$ fulfils (9) (regardless of whether we choose fixed, bounded, or diminishing step sizes). Nevertheless, we will see later in the bundle-free approach that a variant of our approximate adjoint equation can under certain circumstances yield descent directions for the reduced objective.

We now demonstrate the performance of the algorithm on an example with a nontrivial biactive set: Example 3. In order to assure feasibility at every step, we solve the variational inequality with the primal-dual active set (PDAS)/semismooth Newton method from [38]. Note that although the solver for the variational inequality is mesh dependent, the majority of the linear solves are done within the first four iterations, see Table 2. For a graph of the behavior of the residuals as well as plots of the solution, see Figure 4.

We again use a uniform grid with 512^2 grid points and start the algorithm at $(z^0, u^0, \xi^0) = (0, 0, 0)$. We choose the a priori step sizes $\nu_k := (k)^{-1/2}$ and update ϵ_k according to $\epsilon_k := 10^{-4}/2^k$. The inner PDAS solver stops once the residual of the non-smooth system of equations reaches a tolerance of 10^{-10}. Though the theory does not provide a stopping criterion, we check the residual of strong stationarity, which reaches $\mathcal{O}(10^{-9})$ after 30 iterations. The residual is calculated using a discrete approximation of the following quantity:

Table 2 Outer loop k vs. inner loop iterations "iter" for PDAS with tol = 10^{-10} used in Algorithm 5.5

k	1	2	3	4	5	⋯	31	32
iter	84	9	2	1	1	⋯	1	1

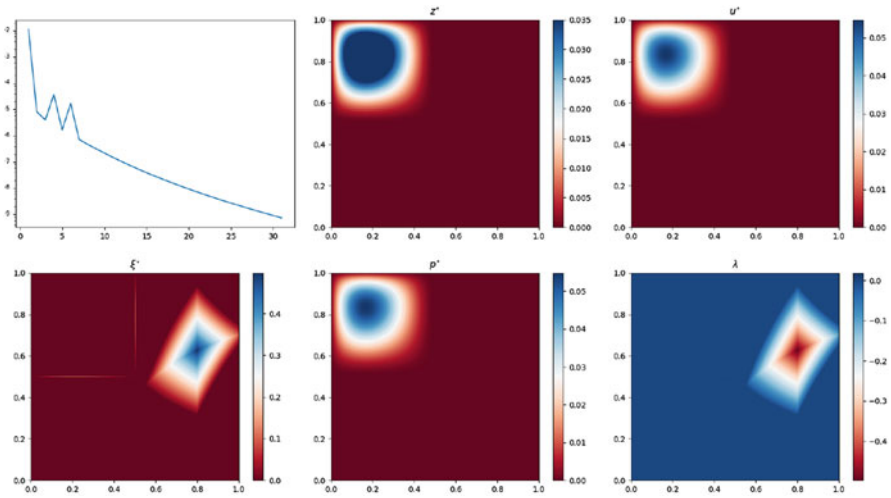

Fig. 4 Clockwise from upper left: residual of strong stationarity for Algorithm 5.5, optimal control $z^\star$, state $u^\star$, multiplier $\lambda^\star$, adjoint $p^\star$, and multiplier $\xi^\star$

$$\begin{aligned}\mathrm{res}^k := \|z^k - \mathrm{Proj}_{Z_{\mathrm{ad}}}(z^k - g^k)\|_{L^2} + \|Au^k - \xi^k - z^k - f\|_{L^2} + \|\max(0, -u^k)\|_{L^2} + \\ \|\max(0, -\xi^k)\|_{L^2} + |(\xi^k, u^k)_{L^2}| + \|Ap^k - \lambda^k + u^k - u_d\|_{L^2} + |\max(0, \langle\lambda^k, p^k\rangle)| + \\ \|\chi_{\Omega_k^{0+}} p^k\|_{L^2} + \|\chi_{\mathscr{I}_k}\lambda^k\|_{L^2} + \|\chi_{\Omega_k^{00}}\max(0, p^k)\|_{L^2} + \|\chi_{\Omega_k^{00}}\max(0, -\lambda^k)\|_{L^2}\end{aligned} \tag{39}$$

Despite lacking a convergence theory, the algorithm performs very well on this large-scale nontrivial problem. Indeed, counting all the free variables (z, u, ξ, etc.) there are over 10^6 degrees of freedom. Moreover, the presence of biactivity means that the example considered is genuinely non-smooth and non-convex. This is particularly encouraging, as the algorithm clearly outperforms the adaptive penalty method in terms of ease of implementation, size and structure of the systems of linear equations, and order of accuracy (almost reaching even strong stationarity). In particular, we note the stark contrast to the "sharpness" of the solution in comparison to the smooth method.

5.2 A Direct Solver for C-Stationarity Conditions

In this section, we adapt a method from [36] to the canonical MPEC (7). We work in the setting of Example 2 and assume that

$$Z_{\text{ad}} := \left\{ v \in L^2(\Omega) \mid a \leq v \leq b, \text{ a.e. } x \in \Omega \right\}$$

with $a, b \in L^2(\Omega)$, $a < b$. We first state the algorithm. Then, we motivate the steps of the algorithm and discuss its convergence properties.

The formal derivation of Algorithm 5.6, which we describe in detail below, follows several basic steps: Fix a control and estimate the active and inactive sets; ignoring biactivity, use (28e) to approximate (28b)–(28e) as a system of equations; solve the reduced system (including (28a)) to obtain an update; test the residual of C-stationarity, if necessary, return to 1.

Algorithm 5.6 An active-set equality-constrained newton solver w. Feasibility restoration

Input: $a, b, u_d, f \in L^2(\Omega), \alpha > 0, z^0 \in L^2(\Omega), p^0 \in H_0^1(\Omega), k := 0;$

1: **repeat**

2: Compute (u^k, ξ^k) by solving

$$Au - \xi = Bz^k + f, \quad \xi = (\xi - cu)_+, \quad c > 0,$$

and set

$$\Omega_k^0 := \left\{ x \in \Omega \;\middle|\; u^k(x) = 0 \right\},$$
$$\Omega_k^+ := \left\{ x \in \Omega \;\middle|\; u^k(x) > 0 \right\};$$

3: Compute

$$\Omega_k^a := \left\{ x \in \Omega \;\middle|\; a(x) + \alpha^{-1}(B^* p^k)(x) > 0 \right\},$$
$$\Omega_k^b := \left\{ x \in \Omega \;\middle|\; -\alpha^{-1}(B^* p^k)(x) - b(x) > 0 \right\}.$$

and $\Omega_k^{ina} := \Omega \setminus (\Omega_k^a \cup \Omega_k^b)$.

4: Compute $(\delta u^k, \delta p^k)$ by solving

$$\begin{aligned}
A^*\delta p + \delta u &= u_d - u^k - A^* p^k, && \text{on } \Omega_k^+,\\
\delta p &= 0, && \text{on } \Omega_k^0,\\
A\delta u + \alpha^{-1} B \chi_{\Omega_k^{ina}} B^* \delta p &= B\text{Proj}_{Z_{\text{ad}}}(-\alpha^{-1} B^* p^k) + f - Au^k, && \text{on } \Omega_k^+,\\
\delta u &= 0, && \text{on } \Omega_k^0,
\end{aligned}$$

and set $u^{k+1} := u^k + \delta u^k$, $p^{k+1} := p^k + \delta p^k$,

$$z^{k+1} := z^k + \delta z = \text{Proj}_{Z_{\text{ad}}}(-\alpha^{-1} B^* p^k) - \alpha^{-1} \chi_{\Omega_k^{ina}} B^* \delta p^k$$

$k := k + 1;$

5: **until** some stopping criterion is satisfied

More specifically, suppose we are in the infinite-dimensional setting. Assuming that $\xi^\star$ is sufficiently regular, then Equations (28c), (28d) can be understood as the following system of smooth and non-smooth equations:

$$Au^\star - \xi^\star = Bz^\star + f, \tag{40a}$$

$$\xi^\star = (\xi^\star - cu^\star)_+, \tag{40b}$$

where $c > 0$ is some scaling constant. Since $\xi^\star \in L^2(\Omega)$, the Newton derivative for the plus function described at the end of Section 4.1 is not valid in this setting (since both the domain and range here must be taken as $L^2(\Omega)$). On a discrete level, this is not an issue. For some fixed $z^\star$, solving (40) gives a pair $(u^\star, \xi^\star)$ along with active and inactive sets:

$$\Omega^0(u^\star) := \left\{x \in \Omega \;\middle|\; u^\star(x) = 0\right\}, \quad \Omega^+(u^\star) := \left\{x \in \Omega \;\middle|\; u^\star(x) > 0\right\}.$$

As in our discussion of (28), if we treat the variables $u^\star$, $\lambda^\star$ as finite-dimensional vectors and the complementarity condition as the pointwise product of $u^\star$ and $\lambda^\star$, then the complementarity condition would indicate that $\lambda^\star = 0$ on the inactive set. Analogously, we take $p^\star = 0$ on the (entire) active set, thus ignoring biactivity. Consequently the remaining sign condition in (28) holds. Finally, using the projection formula (23) along with a semismooth Newton step, we can handle the variational inequality (28a). We recall the sets

$$\Omega^a := \left\{x \in \Omega \;\middle|\; a(x) + \alpha^{-1}(B^* p^\star)(x) > 0\right\},$$

$$\Omega^b. := \left\{x \in \Omega \;\middle|\; -\alpha^{-1}(B^* p^\star)(x) - b(x) > 0\right\}.$$

and $\Omega^{ina} := \Omega \setminus (\Omega^a \cup \Omega^b)$.

Now, supposing we want a new approximation $(u^\star, z^\star, p^\star)$ via $u^\star + \delta u$, $z^\star + \delta z$, and $p^\star + \delta p$, we consider the reduced system by eliminating the dual variables:

$$\delta z + \alpha^{-1}\chi_{\Omega^{ina}} B^* \delta p = \mathrm{Proj}_{Z_{\mathrm{ad}}}(-\alpha^{-1} B^* p^\star) - z^\star \qquad \text{on } \Omega, \tag{41a}$$

$$A^* \delta p + \delta u = u_d - u^\star - A^* p^\star, \qquad \text{on } \Omega^+(u^\star), \tag{41b}$$

$$\delta p = 0, \qquad \text{on } \Omega^0(u^\star), \tag{41c}$$

$$A\delta u - B\delta z = Bz^\star + f - Au^\star, \qquad \text{on } \Omega^+(u^\star), \tag{41d}$$

$$\delta u = 0, \qquad \text{on } \Omega^0(u^\star), \tag{41e}$$

If we replace δz in (41d), then we get the smaller system in $(\delta u, \delta p)$:

$$A^* \delta p + \delta u = u_d - u^\star - A^* p^\star, \qquad \text{on } \Omega^+(u^\star), \tag{42a}$$

$$\delta p = 0, \qquad \text{on } \Omega^0(u^\star), \tag{42b}$$

$$A\delta u + \alpha^{-1} B\chi_{\Omega^{ina}} B^* \delta p = B\mathrm{Proj}_{Z_{\mathrm{ad}}}(-\alpha^{-1} B^* p^\star) + f - Au^\star, \qquad \text{on } \Omega^+(u^\star), \tag{42c}$$

$$\delta u = 0, \qquad \text{on } \Omega^0(u^\star), \tag{42d}$$

Table 3 Residuals of strong stationarity for Algorithm 5.6 with stopping tolerance tol = 10^{-7}

k	1	2	3	4	5
res^k	8.7287e−2	1.6175e−3	1.1190e−5	3.8548e−7	2.2375e−8

which is remarkably similar to the semismooth Newton step in the adaptive penalty method. Unlike the smooth methods, however, a proof of convergence remains elusive. The main culprit here is clearly the sequence $\{\Omega_k^+\}$, which need not converge with respect to any notion of set convergence, e.g., Painlevé-Kuratowski or in the sense of characteristic functions.

Note also that this formally derived system is well-defined in function space provided that we have enough regularity. For example, provided that the sets $\Omega^+(u^\star)$, $\Omega^0(u^\star)$ are sufficiently regular, we could reduce the search for $(\delta u, \delta p) \in \mathscr{H} \times \mathscr{H}$ with $\mathscr{H} := H_0^1(\Omega^+(u^\star))$, solve the associated weak form of (42a), (42b), and extend the solutions by zero on $\Omega^0(u^\star)$.

An obvious stopping criterion for Algorithm 5.6 would be the residual of (28) up to some user-defined tolerance (as suggested in [36]). Moreover, we see that the computational effort is roughly that of the adaptive penalty method. In fact, there is much less nonlinearity here due to a lack of the penalty terms ϕ^ε. We also mention that the MPECs in [36] are much more challenging than (7), as the controls there arise inside the differential operator. Nevertheless, the algorithm seems to perform very well, even on examples with nontrivial biactive sets.

We demonstrate the performance of Algorithm 5.6 on Example 3. As expected, Algorithm 5.6 behaves like a second-order method, see Table 3. We once again used the PDAS/semismooth Newton method in [38] to restore feasibility at every step. Moreover, since the multiplier λ is eliminated from the algorithm, we artificially reintroduce it for the calculation of the residuals.

The solutions look identical to those plotted in Figure 5. We therefore only provide images of the biactive sets for $y^\star$ and upper and lower active sets for $z^\star$. We note that this algorithm also performs quite well, reaching a residual of strong stationarity on the order of $\mathscr{O}(10^{-8})$ within $k = 5$ iterations, though it never reaches $\mathscr{O}(10^{-9})$ in contrast to the approximating subgradient algorithm. In addition, the effort to solve each step is higher, as seen in Table 4.

Our next non-smooth method seeks to overcome the theoretical deficiencies of Algorithm 5.5 and 5.6. In some sense, it takes a step towards bridging the gap between this active-set-based solver and the approximate projected subgradient method in the previous subsection.

5.3 *The Bundle-Free Implicit Programming Method*

We now present the bundle-free implicit programming approach from [46], which we extend for control constraints. In contrast to the active-set method in the previous subsection, this method is based off of B-stationarity conditions. We must therefore assume that S is directionally differentiable.

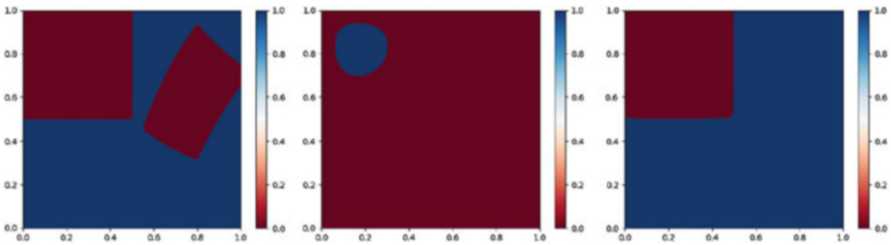

Fig. 5 Plot of solutions using Algorithm 5.6. Left to right: characteristic functions $\chi_{\Omega_\star^{00}}$(set of all indices with $u_i^\star \leq$1e−8 and $\xi_i^\star \leq 0$), $\chi_{\Omega_\star^b}$ (set of all indices with $z_i^\star \leq b_i$−1e−6), and $\chi_{\Omega_\star^a}$ (set of all indices with $z_i^\star \geq a_i$+1e−6)

Table 4 Outer loop k vs. inner loop "iter" iterations for PDAS with tol = 10^{-10} used in Algorithm 5.6

k	1	2	3	4	5
iter	84	21	19	1	1

We first state the basic assumptions and the algorithm. Afterwards, we discuss the motivations for the steps and examine the convergence properties. Throughout this subsection, let $Z_{\text{ad}} := \left\{ z \in L^2(\Omega) \mid a \leq z \leq b \right\}$ with $a, b \in L^2(\Omega)$ and $a < b$ and assume that the variational inequality is defined as in (7). We otherwise work in the abstract framework of (6), where $\mathscr{J}(z) := F(S(Bz)) + G(z)$ and F and G are continuously Fréchet differentiable. By assumption, $\xi \in H$.

For some constant $c_q > 0$, let $q(\cdot) := c_q \| \cdot \|^2/2$ and for $z \in Z_{\text{ad}}$ define the local quadratic models:

$$\mathscr{M}(\cdot) := q(\cdot) + F'(z; S'(Bz; B\cdot)) + G'(z; \cdot),$$
$$\mathscr{M}_\epsilon(\cdot) := q(\cdot) + F'(z; d_\epsilon(B\cdot)) + G'(z; \cdot), \ \epsilon > 0,$$

where $d_\epsilon(\cdot)$ is a smooth approximation of $S(Bz; \cdot)$ such that $d_\epsilon(0) = 0$. Here, we suggest letting $d_\epsilon(w) = d$, the solution to

$$Ad + \epsilon^{-1} \chi_{\Omega^{0+}} d - \chi_{\Omega^0} \nabla^2 \phi^\epsilon(-d) = w,$$

where $w \in V^*$, $\epsilon > 0$, and $u = S(Bz)$. This is related to the true formula for $d = S'(Bz; w)$ given by

$$\text{Find } d \in \mathscr{K}(u, \xi) : \int_\Omega \nabla d \cdot \nabla [v - d] \mathrm{d}x \geq \langle w, v - d \rangle, \quad \text{for all } v \in \mathscr{K}(u, \xi), \tag{43}$$

where $\mathscr{K}(u, \xi) := \left\{ u \in H_0^1(\Omega) \,\middle|\, d(x) \geq 0, \ \text{q.e. } x \in \Omega^0, \ d(x) = 0. \text{ a.e. } x \in \Omega^{0+} \right\}$; see [63] (or [15, Chap. 6] for an English summary). As this formula makes use of quasi-continuity and potentially non-negligible sets of Lebesgue measure zero, it is unclear how to make direct use of the nonnegativity condition

in a numerical method. To circumvent this issue, a certain regularity condition is assumed throughout (as in [46]):

1. $\Omega^0 = \overline{\text{int}(\Omega^0)}$,
2. If $m(\Omega^{00}) = 0$, then $\text{cap}(\Omega^{00}) = 0$ or $S'(Bu; Bh) = d = 0$ q.e. $x \in \Omega^0$.
3. If $m(\Omega^{00}) > 0$, then $\exists \nu > 0, \exists \gamma' > 0 : \forall \gamma \geq \gamma', d_{\gamma,\varepsilon} \geq 0$, $a.e.\ \mathscr{A}_\nu \setminus \Omega^0)$ where $\mathscr{A}_\nu := \left\{ x \in \Omega \,\middle|\, \text{dist}(x, \Omega^0) < \nu \right\}$.

In light of this assumption, we speak of "negligible" biactive sets whenever $m(\Omega^{00}) = 0$. Note that S is in fact Gâteaux differentiable, whenever this condition holds. This once again highlights a fundamental difference between the optimal control of PDEs and variational inequalities, even for the simplest variational inequality of interest.

We also suggest the smooth approximation of the directional derivative in order to guarantee that $p_\epsilon = d'_\epsilon(0)^* w$ solves

$$A^* p + \epsilon^{-1} \chi_{\Omega^{0+}} p = w$$

and that $\{p_\epsilon\}_{\epsilon > 0}$ is uniformly bounded in V. Notice the similarity to the approximating limiting subgradient suggested earlier.

Algorithm 5.7 A bundle-free implicit programming approach

Input: $\epsilon^0 > 0, \beta, \rho, \sigma \in (0, 1), s > 0, (z^0, u^0, \xi^0) \in Z \times V \times H, k := 0;$

repeat
  Set $\Omega_k^0 := \left\{ x \in \Omega \,\middle|\, u^k(x) = 0 \right\}$ and $\Omega_k^{00} := \left\{ x \in \Omega_k^0 \,\middle|\, \xi^k(x) = 0 \right\}$;
  if Ω_k^{00} is negligible **then**
    Compute δz^k by

$$\delta z = \text{Proj}_{\mathscr{T}_{Z_{\text{ad}}}(z^k)} \left[-c_q^{-1} \nabla \mathscr{M}^k(0) \right]; \tag{44}$$

    Compute step size τ^k using a line search;
    Set $z^{k+1} = z^k(\tau^k)$, $u^{k+1} := S(Bz^{k+1})$, $\xi^{k+1} := Au^{k+1} - Bz^{k+1} - f$, $k := k + 1$;
  else
    Compute δz^k by

$$\delta z = \text{Proj}_{\mathscr{T}_{Z_{\text{ad}}}(z^k)} \left[-c_q^{-1} \nabla \mathscr{M}^k_{\epsilon^k}(0) \right]; \tag{45}$$

    while descent criterion fails **do**
      Choose $\epsilon^k \in (0, \rho\epsilon^k)$, update model $\mathscr{M}^k_{\epsilon^k}$ go to 8:;
    end while
    Choose $\epsilon^{k+1} \in (0, \rho\epsilon^k)$;
    Compute τ^k as in 5:, update (z^k, u^k, ξ^k) as in 6:;
  end if
  (Robustification Step), $k := k + 1$;
until some stopping criterion is satisfied

In Algorithm 5.7, $\mathscr{T}_{Z_{\text{ad}}}(z)$ is the tangent cone to Z_{ad} at $z \in Z_{\text{ad}}$, which is defined by

$$\mathscr{T}_{Z_{\text{ad}}}(z) := -\left[\mathscr{N}_{\text{ad}}(z)\right]^+ = \overline{\mathbb{R}_+(Z_{\text{ad}} - z)}.$$

We use the generalized Armijo line search (cf. [13].): Set $\tau^k := \beta^{m_k} s$ where m_k is the first nonnegative integer such that

$$\mathscr{J}(z^k) - \mathscr{J}(z^k(\beta^m s)) \geq \frac{\sigma}{\beta^m s} \|z^k - z^k(\beta^m s))\|_{L^2}^2, \tag{46}$$

where $\beta, \sigma \in (0, 1)$, $s > 0$ and, given a step δz, we set

$$z(\tau) := \mathrm{Proj}_{Z_{\mathrm{ad}}}(z + \tau \delta z), \ \tau > 0.$$

Since much of Algorithm 5.7 is derived from B-stationarity conditions, we restate them here:

$$\mathscr{J}'(z; w - z) = F'(S(Bz))S'(Bz; B(w - z)) + G'(z)(w - z) \geq 0, \quad \forall w \in Z_{\mathrm{ad}}. \tag{47}$$

From (47), it is clear that we can equivalently reformulate B-stationarity as:

$$\mathscr{J}'(z; \delta z) \geq 0, \ \forall \delta z \in \mathscr{T}_{Z_{\mathrm{ad}}}(z).$$

Moreover, since $\mathscr{J}'(z; \delta z)$ is positively homogeneous in δz, it follows that $\delta z = 0$ is a minimizer of $\mathscr{J}'(z; \delta z)$ over $\mathscr{T}_{Z_{\mathrm{ad}}}(z)$, whenever z is B-stationary. Finally, given that $\mathscr{T}_{Z_{\mathrm{ad}}}(z)$ is a non-empty, closed, and convex cone, we can add the coercive quadratic form q to the definition of B-stationarity without altering the characterization, cf. the general analysis in [46, Section 2]. Therefore, if z is B-stationary, then $0 \in Z$ solves the auxiliary problem

$$\min \left\{ \mathscr{M}(\delta z) := q(\delta z) + \mathscr{J}'(z; \delta z) \text{ over } \delta z \in \mathscr{T}_{Z_{\mathrm{ad}}}(z) \right\}. \tag{48}$$

Now, if the biactive set is negligible, then S is in fact Gâteaux differentiable. In this case, (48) has a unique solution given by

$$(\delta z^* + c_q^{-1}(B^* S'(Bz)^* \nabla F'(S(Bz)) + \nabla G(z)), w - \delta z^*) \geq 0, \ \forall w \in \mathscr{T}_{Z_{\mathrm{ad}}}(z),$$

which, noting that $\nabla \mathscr{M}(0) = B^* S'(Bz)^* \nabla F'(S(Bz)) + \nabla G(z)$, is equivalent to (45). Furthermore, in this smooth setting, we can prove the following result.

Proposition 1 *Let $z \in Z_{\mathrm{ad}}$, $u = S(Bz)$, $\xi = Au - Bz - f$, and suppose that $\mathscr{J}$ is Gâteaux differentiable at z. If z is not B-stationary, then (46) stops in a finite number of steps.*

Proof Suppose δz is given by the projection formula (45). Since δz is the unique global optimum and $0 \in \mathscr{T}_{Z_{\mathrm{ad}}}(z)$, we have

$$q(\delta z) + \mathscr{J}'(z; \delta z) < 0 \Longrightarrow \mathscr{J}'(z; \delta z) \leq -\frac{c_q}{2} \|\delta z\|_{L^2}^2 \tag{49}$$

Moreover, for any $\tau > 0$, we have $z(\tau) := \mathrm{Proj}_{Z_{\mathrm{ad}}}(z - \tau\delta z)$ and since $\mathrm{Proj}_{Z_{\mathrm{ad}}}$ is non-expansive:

$$\|z - z(\tau)\|_{L^2} = \|\mathrm{Proj}_{Z_{\mathrm{ad}}}(z + \tau\delta z) - \mathrm{Proj}_{Z_{\mathrm{ad}}}(z)\|_{L^2} \leq \tau\|z\|_{L^2}. \tag{50}$$

Furthermore, since $z \in Z_{\mathrm{ad}}$, $\delta z \in \mathscr{T}_{Z_{\mathrm{ad}}}(z)$, and Z_{ad} is defined by simple pointwise bound constraints in $L^2(\Omega)$, we appeal to the proof of Lemma 6.34 [15], which shows that

$$\tau^{-1}(z(\tau) - z) \to \delta z \text{ as } \tau \downarrow 0. \tag{51}$$

Finally, suppose that (46) fails for all $\tau > 0$. Then,

$$\tau^{-1}(\mathscr{J}(z(\tau)) - \mathscr{J}(z)) > -\sigma\tau^{-2}\|z - z(\tau)\|_{L^2}^2, \quad \forall \tau > 0.$$

On the right side of the inequality, we can estimate from below using (50):

$$-\sigma\tau^{-2}\|z - z(\tau)\|_{L^2}^2 \geq -\sigma\|\delta z\|_{L^2}^2$$

Now, letting $z(\tau) = z + \tau\tau^{-1}(z(\tau) - z) = z + \tau d_\tau$, where $d_\tau \to \delta z$ by (51), we have by (49):

$$\tau^{-1}(\mathscr{J}(z(\tau)) - \mathscr{J}(z)) = \tau^{-1}(\mathscr{J}(z+\tau d_\tau) - \mathscr{J}(z)) \to \mathscr{J}'(z;\delta z) \leq -\frac{c_q}{2}\|\delta z\|_{L^2}^2.$$

But, then

$$-\frac{c_q}{2}\|\delta z\|_{L^2}^2 \geq \mathscr{J}'(z;\delta z) \geq -\sigma\|\delta z\|_{L^2}^2,$$

a contradiction, since $\sigma \in (0, c_q/2)$.

Proposition 1 provides a justification for steps 4:–6: in Algorithm 5.7. Note that the calculation of the gradient $\nabla\mathscr{M}^k(0)$ requires the solution of an adjoint equation. For example, in the setting of Example 2, $\nabla\mathscr{M}^k(0) = \alpha z^k - B^* p^k$, where p^k solves (here in strong form):

$$\begin{aligned} A^* p &= u_d - u^k \quad \text{on } \Omega_k^+, \\ p &= 0, \qquad\quad \text{on } \Omega_k^0, \end{aligned} \tag{52}$$

Turning now to the case when the biactive set is non-negligible, we can easily adjust the proof of Proposition (1) for the non-smooth setting.

Corollary 2 *Let $z \in Z_{\mathrm{ad}}$, $u = S(Bz)$, $\xi = Au - Bz - f$, and suppose that δz is minimizer for (48). If z is not B-stationary, then (46) stops in a finite number of steps.*

Nevertheless, whenever the biactive set is non-negligible, the directional derivative is nonlinear in δz. In particular, (48) is non-convex, which was a key assumption in the arguments. We therefore need an alternative procedure to calculate the step δz.

In Algorithm 5.7, we suggested the step $\delta z = \mathrm{Proj}_{\mathscr{T}_{Z_{\mathrm{ad}}}(z^k)}\left[-c_q^{-1}\nabla \mathscr{M}_{\epsilon^k}^k(0)\right]$, which is related to the smoothed auxiliary problem using the ϵ-dependent model $\mathscr{M}_\epsilon$:

$$\min\left\{\mathscr{M}_\epsilon(\delta z) := q(\delta z) + F'(z; d_\epsilon(B\delta z)) + G'(z)\delta z \text{ over } \delta z \in \mathscr{T}_{Z_{\mathrm{ad}}}(z)\right\}. \tag{53}$$

Note if the approximation was chosen so that $d_\epsilon \to S$ and 0 solves (53) for all sufficiently small $\epsilon > 0$ (or at least along some null sequence), then z must be B-stationary as 0 solves (48), as well. In fact, in the current setting, A is symmetric so we can use the approximation results in [7] (discussed above), see also [46], to argue that if $\delta z_\epsilon \to \delta z$ weakly in $L^2(\Omega)$, then

$$d_\epsilon(B\delta z_\epsilon) \to S'(Bz; B\delta z), \text{ as } \epsilon \downarrow 0,$$

provided that B is completely continuous, e.g., when B is the embedding of $L^2(\Omega)$ into $H^{-1}(\Omega)$.

On the other hand, if there exists some $\delta z \in \mathscr{T}_{Z_{\mathrm{ad}}}(z)$ such that $(\nabla \mathscr{M}_\epsilon(0), \delta z) < 0$, then by continuity, there is some $\eta_\epsilon > 0$ such that $\mathscr{M}_\epsilon(\eta_\epsilon \delta z) < 0$. Since $\mathscr{T}_{Z_{\mathrm{ad}}}(z)$ is a cone, $\eta_\epsilon \delta z \in \mathscr{T}_{Z_{\mathrm{ad}}}(z)$ and 0 does not solve (53). If this persists as $\epsilon \downarrow 0$, then z cannot be B-stationary.

In light of this, consider our δz update. Let $w := d_\epsilon'(0)^* F'(z) + G'(z) = \nabla \mathscr{M}_\epsilon(0)$ and note that

$$\mathscr{T}_{Z_{\mathrm{ad}}}(z) = \left\{\delta z \in L^2(\Omega) \;\middle|\; \delta z \geq 0, \text{ a.e. on } \Omega^a, \ \delta z \leq 0, \text{ a.e. on } \Omega^b\right\},$$

where $\Omega^a := \{x \in \Omega \mid z(x) = a(x)\}$, $\Omega^b := \{x \in \Omega \mid z(x) = b(x)\}$ and $\Omega^{ina} := \Omega \setminus (\Omega^a \cup \Omega^b)$. Then, using the basic properties of $\mathrm{Proj}_{\mathscr{T}_{Z_{\mathrm{ad}}}(z)}$, we have

$$(\nabla \mathscr{M}_\epsilon(0), \delta z) = \\ -c_q\left[\int_{\Omega^{ina}} |\delta z|^2 + \int_{\Omega^a \cap \{-c_q^{-1} w \geq 0\}} |\delta z|^2 + \int_{\Omega^b \cap \{-c_q^{-1} w \leq 0\}} |\delta z|^2\right] \leq 0.$$

Assuming that the latter term is nonzero and expanding $\mathscr{M}_\epsilon$ at zero in direction δz, we obtain:

$$\mathscr{M}_\epsilon(0 + \eta\delta z) = \eta\left(\frac{\eta c_q}{2}\|\delta z\|_{L^2}^2 + (\nabla \mathscr{M}_\epsilon(0), \delta z) + o(1)\right). \tag{54}$$

We may then choose a sufficiently small $\eta_\epsilon > 0$ such that

$$\frac{\eta_\epsilon c_q}{2}\|\delta z\|_{L^2}^2 + (\nabla \mathscr{M}_\epsilon(0), \delta z) + o(1) < 0.$$

Hence, $\mathscr{M}_\epsilon(\eta_\epsilon \delta z) \leq 0$ and by definition

$$F'(z; d_\epsilon(B(\eta_\epsilon \delta z))) + G'(z)(\eta_\epsilon \delta z) \leq -\eta_\epsilon^2 q(\delta z_\epsilon).$$

Therefore, if a uniform lower bound $\eta > 0$ with $\eta_\epsilon \geq \eta > 0$ exists, then we can prove that δz_ϵ is a descent direction for some sufficiently small $\epsilon > 0$.

Proposition 2 *Let $z \in Z_{\text{ad}}$, $u = S(Bz)$, and $\xi = Au - Bz - f$ and suppose that $\mathscr{J}$ is only directionally differentiable at z. Furthermore, let $\delta z_\epsilon = \operatorname{Proj}_{\mathscr{T}_{Z_{\text{ad}}}(z)}\left[-c_q^{-1}\nabla \mathscr{M}_\epsilon(0)\right]$ and assume that $\limsup_{\epsilon \downarrow 0}(\nabla \mathscr{M}_\epsilon(0), \delta z_\epsilon) < 0$. If there exists an $\eta > 0$ such that $\mathscr{M}_\epsilon(\eta \delta z_\epsilon) \leq 0$ for all sufficiently small $\epsilon > 0$, then there exists some $\widehat{\epsilon} > 0$ such that $\delta z_{\widehat{\epsilon}}$ is a descent direction for $\mathscr{J}$ at z.*

Proof By assumption, $F'(z; d_\epsilon(B(\eta \delta z_\epsilon)) + G'(z)(\eta \delta z_\epsilon) \leq -\eta^2 q(\delta z_\epsilon)$. Then,

$$\begin{aligned}
\mathscr{J}'(z; \eta \delta z_\epsilon) &= (\mathscr{J}'(z; \eta \delta z_\epsilon) - F'(z; d_\epsilon(B(\eta \delta z_\epsilon))) - G'(z)(\eta \delta z_\epsilon)) + \\
&\quad F'(z; d_\epsilon(B(\eta \delta z_\epsilon)) + G'(z)(\eta \delta z_\epsilon) \leq \\
&(\mathscr{J}'(z; \eta \delta z_\epsilon) - F'(z; d_\epsilon(B(\eta \delta z_\epsilon))) - G'(z)(\eta \delta z_\epsilon)) - \eta^2 q(\delta z_\epsilon) = \\
&\quad F'(z; S(Bz; B(\eta \delta z_\epsilon))) - F'(z; d_\epsilon(B(\eta \delta z_\epsilon))) - \eta^2 q(\delta z_\epsilon) = \\
&\quad \langle \nabla F(z), S(Bz; B(\eta \delta z_\epsilon)) - d_\epsilon(B(\eta \delta z_\epsilon))\rangle - \eta^2 q(\delta z_\epsilon).
\end{aligned}$$

Now, since $\{p_\epsilon\}$ is bounded, we can show that $\{\delta z_\epsilon\}$ is bounded. Hence, there is a subsequence (denoted still by ϵ) such that $\delta z_\epsilon \rightharpoonup \delta z^*$. Then, for sufficiently small $\widehat{\epsilon} > 0$

$$\langle \nabla F(z), S(Bz; B(\eta \delta z_{\widehat{\epsilon}})) - d_{\widehat{\epsilon}}(B(\eta \delta z_{\widehat{\epsilon}}))\rangle \leq \eta^2 q(\delta z_{\widehat{\epsilon}})/2, \tag{55}$$

Hence, $\mathscr{J}'(z; \delta z_{\widehat{\epsilon}}) \leq -c_q \eta \|\delta z_{\widehat{\epsilon}}\|^2/4$, as was to be shown.

Since a direct verification of the hypotheses is potentially too expensive from a computational standpoint, we suggest a heuristic. Fix a lower bound $\underline{\eta} > 0$. If $(\nabla \mathscr{M}_\epsilon(0), \delta z) < -q(\delta z)$ and (55) (or an approximation as in [46, Remark 3.13]) holds with $\eta = 1$ (or $\eta = \underline{\eta}$), we use the current δz in the line search. Thus, the "descent condition" in 9: holds. If $-q(\delta z) \leq (\nabla \mathscr{M}_\epsilon(0), \delta z) < 0$, then we choose $\eta > 0$ such that $\eta q(\delta z) + (\nabla \mathscr{M}_\epsilon(0), \delta z) < 0$, set $\delta z := \eta \delta z$, update $\underline{\eta} := \min(\eta, \underline{\eta})$, and check (55) or an approximation. Here, the descent condition also holds, but the model $\mathscr{M}_\epsilon$ might be failing. If (55) fails in the latter, then we go to step 10:.

Finally, if $(\nabla \mathscr{M}_\epsilon(0), \delta z) = 0$, then we go to step 10: . In practice, one might also attempt to circumvent the verification of (55) by using a sufficiently small $\epsilon > 0$ and decreasing at every step.

This heuristic has its limitations, e.g., it cannot prohibit $\underline{\eta} \downarrow 0$. Therefore, if it appears that this is the case, we break the while loop and go to a "robustification step" in 15:. This just means that if the quadratic model $\mathscr{M}_\epsilon$ appears to be ineffective, then we should calculate a new control z by using an alternative method that is guaranteed to converge and "restart" the algorithm. For instance, we could use the adaptive penalty method in Section 4.1 (for a fixed γ, increasing each time the robustification step is used).

Finally, we recall that in the classical projected-gradient approaches, the Lipschitz continuity of the gradient of the objective is essential for convergence proofs, cf. [13]. There, one can show that the line search will always stop provided that the step size is below a certain threshold which only depends on σ and the Lipschitz modulus L. This is, of course, not possible for the control of variational inequalities as the reduced objective is non-smooth (even when the biactive set is negligible we only have Gâteaux differentiability). Therefore, we must also monitor the behavior of the accepted step sizes at each iteration. If, as with $\underline{\eta}$, the step sizes τ^k appear to be rapidly decreasing with each iteration, then we also make use of a robustification step. For a full convergence proof in the case of the canonical MPEC (excluding control constraints), we refer the interested reader to [46].

We conclude by demonstrating the performance of the bundle-free method on two examples, Example 3 and the following example (adapted from [39, Ex. 6.1] by adding control constraints).

Example 4 Here, we let $a \equiv 0$, $b \equiv 0.8$ and set $\alpha = 1$. In addition, we define f and u_d as follows:

$$z_1(\mathbf{x}_1, \mathbf{x}_2) = -4096\mathbf{x}_1^6 + 6144\mathbf{x}_2^5 - 3072\mathbf{x}_1^4 + 512\mathbf{x}_2^3,$$

$$z_2(\mathbf{x}_1, \mathbf{x}_2) = -244.140625\mathbf{x}_1^6 + 585.9375\mathbf{x}_2^5 - 468.75\mathbf{x}_2^4 + 125\mathbf{x}_2^3,$$

$$y^*(\mathbf{x}_1, \mathbf{x}_2) = \begin{cases} z_1(\mathbf{x}_1, \mathbf{x}_2) z_2(\mathbf{x}_1, \mathbf{x}_2) & (\mathbf{x}_1, \mathbf{x}_2) \in]0, 0.5[\times]0, 0.8[, \\ 0 & \text{otherwise}, \end{cases}$$

$$u^* = y^*, \quad \xi^*(\mathbf{x}_1, \mathbf{x}_2) = 2\max(0, -|\mathbf{x}_1 - 0.8| - |\mathbf{x}_1\mathbf{x}_2 - 0.2| - 0.3 + 0.35),$$

$$f = -\Delta y^* - u^* - \xi^*, \quad u_d = y^* + \xi^* - \alpha \Delta u^*.$$

In order to compare to the other methods, we again use a uniform grid with 512^2 grid points and start the algorithm at $(z^0, u^0, \xi^0) = (0, 0, 0)$ for Example 3. In contrast, we use a random starting point when solving Example 4. We start the algorithm with $\epsilon_0 = 10^{-10}$ and subsequently set $\epsilon_{k+1} = \epsilon_k/2$.

For Example 3 we obtain the same solution as in all the previous algorithms. Likewise, the algorithm performs very well on Example 4, see Table 5 and Figure 6. We note, however, that Example 4 (when starting with a random initial guess) is

Table 5 Residuals, step sizes, and number of PDAS iterations ("iter") for Algorithm 5.7 when used to solve Examples 3 and 4

k	1	2	3	4
Example 3				
res^k	0.01018	8.7843e−7	9.2082e−11	
τ_k	1.0	1.0	1.0	
iter	84	84	84	
Example 4				
res^k	0.5388	0.00375	3.0727e−7	1.5366e−7
τ_k	1.0	1.0	0.5	0.00390625
iter	133	133	133	133

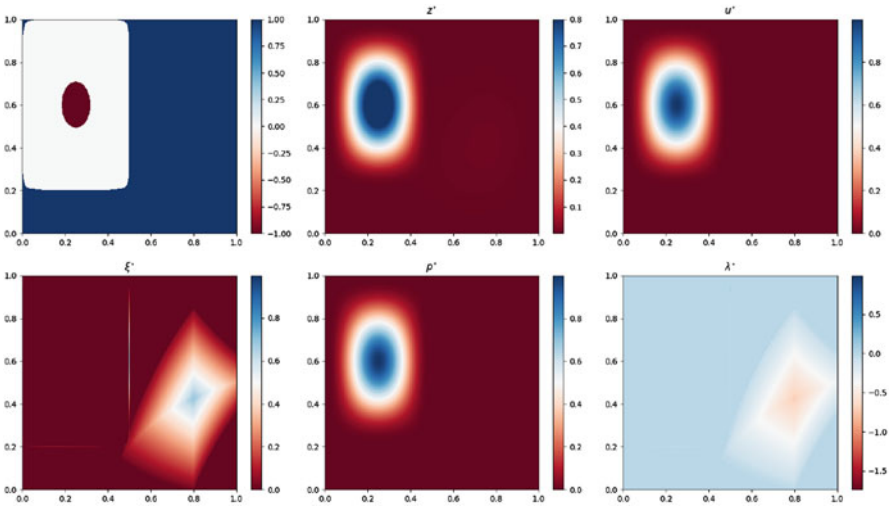

Fig. 6 Clockwise from upper left: characteristic functions $\chi_{\Omega_\star^b}$ (red) and $\chi_{\Omega_\star^a}$ (blue) (set of all indices with $z_i^\star \geq b_i - 1\text{e}{-6}$ and $z_i^\star \leq a_i + 1\text{e}{-6}$), optimal control $z^\star$, state $u^\star$, multiplier $\lambda^\star$, adjoint $p^\star$, and multiplier $\xi^\star$ for Example 4

more difficult to solve. In fact, once the difference in function values becomes negligible, i.e., on the order of $\mathcal{O}(10^{-13})$, the residual of S-stationarity stagnates at $\mathcal{O}(10^{-7})$. Finally, one important aspect of the theory for the bundle-free method was the relation $(\nabla \mathscr{M}_\epsilon(0), \delta z) < -q(\delta z)$. To see how the choice of ϵ_k influences this, see Table 6. There, we observe that far from the solution, a larger value of ϵ_k seems to yield a good approximation. However, the choice becomes critical once we close in on the solution.

Table 6 Behavior of Taylor expansion of $\mathscr{M}_\epsilon(0+\delta z_k)$ at each iteration in Algorithm 5.7 when applied to Example 3

ϵ	$q(\delta z_0)+(\nabla.\mathscr{M}_\epsilon(0),\delta z_0)$	$q(\delta z_1)+(\nabla.\mathscr{M}_\epsilon(0)\delta z_1)$	$q(\delta z_2)+(\nabla.\mathscr{M}_\epsilon(0),\delta z_2)$
1e−1	−19.1512	−0.264565	−0.264625
1e−2	−18.6723	−0.0392292	−0.0392487
1e−3	−18.5176	−0.000749805	−0.000750887
1e−4	−18.5087	−8.50068e−6	−8.37251e−6
1e−5	−18.5082	−2.36252e−7	−8.53014e−8
1e−6	−18.5082	−1.5286e−7	−8.54905e−10
1e−7	−18.5081	−1.52175e−7	−8.53968e−12
1e−8	−18.5081	−1.52185e−7	−8.5597e−14
1e−9	−18.5081	−1.52187e−7	−2.26899e−15
1e−10	−18.5081	−1.52187e−7	−1.55728e−15

6 Conclusion

Despite being a topic of interest for several decades, the optimal control of variational inequalities continues to be an active field of research. The rapidly growing interest in the past decade appears to be a result of the many theoretical, algorithmic, and computational advances in PDE-constrained optimization to date. We have seen here that both reduced space and full-space approaches are possible; however, the difficulties due to either a non-smooth control-to-state mapping or degenerate complementarity constraints persist. The various techniques for deriving optimality conditions in the presence of non-smoothness or degeneracy have led both in finite and infinite dimensions to a hierarchy of first-order optimality conditions. Some of these conditions are directly related to function-space-based numerical methods (C-stationarity) whereas other conditions (e.g., those of Mignot and Puel) are derived using concepts of generalized differentiation and a fine analysis of the regularity properties of the underlying functions and multipliers. These facts should therefore always be taken into account when developing numerical optimization algorithms.

In our numerical studies, we considered two main types of solution methods: smooth and non-smooth. Using approximation techniques for variational inequalities as in [30, 31], the smooth methods are almost always available and allow us to immediately take advantage of existing solvers for (smooth) PDE-constrained optimization problems. For the smoothed/regularized problems, we are only limited by our knowledge of the corresponding parameter-dependent PDE-constrained optimization problem. In our study, we make use of a smooth continuation approach and solve the first-order conditions directly using a semismooth Newton method. For small penalty parameters ϵ, the solution of the linear system (24) needed to calculate the updates is relatively well-behaved. However, the lower off-diagonal block becomes increasingly problematic as we attempt to approach the limiting problem for $\epsilon \downarrow 0$. Thus, we eventually pay a major price for smoothing the original problem.

As mentioned in the introduction, we present here several possible non-smooth methods that mirror related approaches in smooth PDE-constrained optimization. The first method is inspired by the classical projected subgradient methods in [54, 76]. Just as in these classical methods, the relative cost of each iteration is roughly as cheap as a standard projected-gradient approach sans line search: solve the VI, then a linear elliptic PDE, and compute a pointwise projection. As with all subgradient gradient methods, the strong convergence statements are essentially limited to convex problems. Nevertheless, the method presented here behaves quite well in practice and thus, warrants a deeper study in future research.

The active-set method has been taken from [36] and adapted to the setting of the canonical example. At every iteration, the cost of solving the nonlinear system is slightly more than in the smooth continuation approach due to the feasibility restoration step, which requires the solution of a mixed complementarity problem. However, the conditioning issues are now absent as the potentially problematic perturbation in the lower off-diagonal blocks has been removed. Just as in [36], this method performs exceptionally well, even on a problem with persistent biactivity throughout the iterations. However, as mentioned in the text, proving convergence of this method is rather difficult. One possibility would be to make regularity and monotonicity arguments on the data and active sets throughout the iterations (as was done in [42] for a related problem).

Finally, we considered the bundle-free implicit programming approach, which can be thought of as a globalization of the active-set strategy since we typically choose a step size of $\tau = 1$ at the beginning of every line search. Though the theory does contain several strong assumptions to guarantee unconditional global convergence to a stationary point, there appear to be no other genuinely non-smooth function-space-based algorithms for non-smooth non-convex problems currently in the literature. As noted in [46], if one can prove a kind of semismooth property of the reduced objective, then the convergence theory is greatly simplified. In comparison to the other non-smooth methods, it is clearly more costly due to the usage of the line search. The caveats for this method are the need to monitor the step sizes (essentially restarting if they get too small) and a meaningful heuristic for the convergence criteria. Nevertheless, the theory stands in contrast to non-convex bundle methods as outlined in, e.g., [75]. There, in the ideal non-convex setting, the algorithm terminates at a point at which 0 lies up to some tolerance $\varepsilon > 0$ in a convex hull of certain subgradients corresponding to points y_i that are not "far away" from the current iterate x_k. The connection to the various MPEC stationarity concepts is therefore unclear.

Acknowledgements This paper is an extension of a short course given by the author at the "Frontiers in PDE-constrained Optimization" workshop on June 6–10, 2016 at the Institute for Mathematics and its Applications at the University of Minnesota, Minneapolis, which was sponsored by ExxonMobil. The author would therefore like to express his gratitude for the financial support and the opportunity to write this article. In addition, the author would like to thank Harbir Antil, Patrick Farrell, and the two anonymous reviewers for their helpful comments and thought-provoking questions on the text.

References

1. R. A. Adams and J. J. F. Fournier. *Sobolev spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003.
2. S. Albrecht and M. Ulbrich. Mathematical programs with complementarity constraints in the context of inverse optimal control for locomotion. *Optimization Methods and Software*, pages 1–29, 2017.
3. M. Anitescu, P. Tseng, and S.J. Wright. Elastic-mode algorithms for mathematical programs with equilibrium constraints: global convergence and stationarity properties. *Math. Program*, 110:337–371, 2005.
4. H. Antil, M. Hintermüller, R. H. Nochetto, T. M. Surowiec, and D. Wegner. Finite horizon model predictive control of electrowetting on dielectric with pinning. *Interfaces Free Bound.*, 19(1):1–30, 2017.
5. H. Attouch. *Variational Convergence for Functions and Operators*. Pitman Advanced Publishing Program, Boston, London, Melbourne, 1984.
6. H. Attouch, G. Buttazzo, and G. Michaille. *Variational analysis in Sobolev and BV spaces*, volume 6 of *MPS/SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006. Applications to PDEs and optimization.
7. V. Barbu. *Optimal control of variational inequalities*, volume 100 of *Research Notes in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1984.
8. M. P. Bendsøe and O. Sigmund. *Topology Optimization. Theory, methods and Applications*. Springer Verlag, Berlin, Heidelberg, New York, 2003.
9. M. Bergounioux. Optimal control of variational inequalities: a mathematical programming approach. In *Modelling and optimization of distributed parameter systems (Warsaw, 1995)*, pages 123–130. Chapman & Hall, New York, 1996.
10. M. Bergounioux. Optimal control of an obstacle problem. *Appl. Math. Optim.*, 36(2):147–172, 1997.
11. M. Bergounioux. Use of augmented Lagrangian methods for the optimal control of obstacle problems. *J. Optim. Theory Appl.*, 95(1):101–126, 1997.
12. A. Bermúdez and C. Saguez. Optimal control of a Signorini problem. *SIAM J. Control Optim.*, 25(3):576–582, 1987.
13. D. P. Bertsekas. On the Goldstein-Levitin-Polyak gradient projection method. *IEEE Transactions on Automatic Control*, pages 174–184, 1976.
14. J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A Fresh Approach to Numerical Computing. *SIAM Rev.*, 59(1):65–98, 2017.
15. J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Verlag, Berlin, Heidelberg, New York, 2000.
16. M. Boukrouche and D. A. Tarzia. Convergence of distributed optimal control problems governed by elliptic variational inequalities. *Comput. Optim. Appl.*, 53(2):375–393, 2012.
17. C. Brett, C. M. Elliott, M. Hintermüller, and C. Löbhard. Mesh adaptivity in optimal control of elliptic variational inequalities with point-tracking of the state. *Interfaces Free Bound.*, 17(1):21–53, 2015.
18. H. R. Brezis and G. Stampacchia. Sur la régularité de la solution d'inéquations elliptiques. *Bull. Soc. Math. France*, 96:153–180, 1968.
19. X. Chen, Z. Nashed, and L. Qi. Smoothing methods and semismooth methods for nondifferentiable operator equations. *SIAM J. Numer. Anal.*, 38(4):1200–1216 (electronic), 2000.
20. P. Colli, M. H. Farshbaf-Shaker, G. Gilardi, and J. Sprekels. Optimal boundary control of a viscous Cahn–Hilliard system with dynamic boundary condition and double obstacle potentials. *SIAM Journal on Control and Optimization*, 53(4):2696–2721, 2015.
21. J. C. De Los Reyes. Optimal control of a class of variational inequalities of the second kind. *SIAM J. Control Optim.*, 49(4):1629–1658, 2011.
22. J. C. De los Reyes, R. Herzog, and C. Meyer. Optimal control of static elastoplasticity in primal formulation. *SIAM J. Control Optim.*, 54(6):3016–3039, 2016.

23. J. C. De los Reyes and C. Meyer. Strong stationarity conditions for a class of optimization problems governed by variational inequalities of the second kind. *J. Optim. Theory Appl.*, 168(2):375–409, 2016.
24. A. K. Dond, T. Gudi, and N. Nataraj. A nonconforming finite element approximation for optimal control of an obstacle problem. *Comput. Methods Appl. Math.*, 16(4):653–666, 2016.
25. I. Ekeland and R. Témam. *Convex analysis and variational problems*, volume 28 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, English edition, 1999. Translated from the French.
26. L. C. Evans and R. F. Gariepy. *Measure theory and fine properties of functions*. Studies in advanced mathematics. CRC Press, Boca Raton (Fla.), 1992.
27. M. H. Farshbaf-Shaker. A penalty approach to optimal control of Allen-Cahn variational inequalities: MPEC-view. *Numerical Functional Analysis and Optimization*, 33(11):1321–1349, 2012.
28. M. H. Farshbaf-Shaker and C. Hecht. Optimal control of elastic vector-valued Allen–Cahn variational inequalities. *SIAM Journal on Control and Optimization*, 54(1):129–152, 2016.
29. G. Fichera. Problemi elastostatici con vincoli unilaterali: Il problema di Signorini con ambigue condizioni al contorno. *Atti Accad. Naz. Lincei Mem. Cl. Sci. Fis. Mat. Natur. Sez. I (8)*, 7:91–140, 1963.
30. R. Glowinski. *Numerical methods for nonlinear variational problems*. Springer Series in Computational Physics. Springer-Verlag, New York, 1984.
31. R. Glowinski, J. L. Lions, and R. Trémolières. *Numerical analysis of variational inequalities*, volume 8 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1981.
32. W. Han and B. D. Reddy. *Plasticity*, volume 9 of *Interdisciplinary Applied Mathematics*. Springer, New York, second edition, 2013. Mathematical theory and numerical analysis.
33. A. Haraux. How to differentiate the projection on a convex set in Hilbert space. some applications to variational inequalities. *J. Math. Soc. Japan*, 29(4):615–631, 1977.
34. R. Herzog, C. Meyer, and G. Wachsmuth. C-stationarity for optimal control of static plasticity with linear kinematic hardening. *SIAM J. Control Optim.*, 50(5):3052–3082, 2012.
35. R. Herzog, C. Meyer, and G. Wachsmuth. B- and strong stationarity for optimal control of static plasticity with hardening. *SIAM J. Optim.*, 23(1):321–352, 2013.
36. M. Hintermüller. An active-set equality constrained Newton solver with feasibility restoration for inverse coefficient problems in elliptic variational inequalities. *Inverse Problems*, 24(3):23pp., 2008.
37. M. Hintermüller, R. H. W. Hoppe, and C. Löbhard. Dual-weighted goal-oriented adaptive finite elements for optimal control of elliptic variational inequalities. *ESAIM Control Optim. Calc. Var.*, 20(2):524–546, 2014.
38. M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth newton method. *SIAM Journal on Optimization*, 13(3):865–888, 2002.
39. M. Hintermüller and I. Kopacka. Mathematical programs with complementarity constraints in function space: C- and strong stationarity and a path-following algorithm. *SIAM J. Optim.*, 20(2):868–902, 2009.
40. M. Hintermüller and I. Kopacka. A smooth penalty approach and a nonlinear multigrid algorithm for elliptic MPECs. *Comput. Optim. Appl.*, 50(1):111–145, 2011.
41. M. Hintermüller and K. Kunisch. Path-following methods for a class of constrained minimization problems in function space. *SIAM Journal on Optimization*, 17(1):159–187, 2006.
42. M. Hintermüller and A. Laurain. Optimal shape design subject to elliptic variational inequalities. *SIAM Journal on Control and Optimization*, 49(3):1015–1047, 2011.
43. M. Hintermüller, C. Löbhard, and M. H. Tber. An ℓ^1-penalty scheme for the optimal control of elliptic variational inequalities. In Mehiddin Al-Baali, Lucio Grandinetti, and Anton Purnama, editors, *Numerical Analysis and Optimization*, volume 134 of *Springer Proceedings in Mathematics & Statistics*, pages 151–190. Springer International Publishing, 2015.

44. M. Hintermüller, B. S. Mordukhovich, and T. M. Surowiec. Several approaches for the derivation of stationarity conditions for elliptic MPECs with upper-level control constraints. *Math. Program.*, 146(1):555–582, 2014.
45. M. Hintermüller and T. Surowiec. First-order optimality conditions for elliptic mathematical programs with equilibrium constraints via variational analysis. *SIAM J. Optim.*, 21(4):1561–1593, 2011.
46. M. Hintermüller and T. Surowiec. A bundle-free implicit programming approach for a class of elliptic MPECs in function space. *Math. Program.*, 160(1):271–305, 2016.
47. M. Hintermüller and D. Wegner. Optimal control of a semidiscrete Cahn–Hilliard–Navier–Stokes system. *SIAM Journal on Control and Optimization*, 52(1):747–772, 2014.
48. M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, New York, 2009.
49. B. Horn and S. Ulbrich. Shape optimization for contact problems based on isogeometric analysis. *Journal of Physics: Conference Series*, 734(3):032008, 2016.
50. K. Ito and K. Kunisch. Optimal control of elliptic variational inequalities. *Applied Mathematics and Optimization*, 41(3):343–364, 2000.
51. P. Jaillet, D. Lamberton, and B. Lapeyre. Variational inequalities and the pricing of American options. *Acta Appl. Math.*, 21(3):263–289, 1990.
52. N. Kikuchi and J. T. Oden. *Contact problems in elasticity: a study of variational inequalities and finite element methods*, volume 8 of *SIAM Studies in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1988.
53. D. Kinderlehrer and G. Stampacchia. *An introduction to variational inequalities and their applications*, volume 88 of *Pure and Applied Mathematics*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, 1980.
54. Krzysztof C. Kiwiel. Methods of descent for nondifferentiable optimization. Lecture Notes in Mathematics. 1133. Berlin etc.: Springer-Verlag, 1985.
55. R. Kornhuber. Monotone multigrid methods for elliptic variational inequalities. I. *Numer. Math.*, 69(2):167–184, 1994.
56. K. Kunisch and T. Pock. A bilevel optimization approach for parameter learning in variational models. *SIAM J. Imaging Sci.*, 6(2):938–983, 2013.
57. J. L. Lions. *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*. Dunod, Paris, 1968.
58. J.-L. Lions. Various topics in the theory of optimal control of distributed systems. In *Optimal control theory and its applications (Proc. Fourteenth Biennial Sem. Canad. Math. Congr., Univ. Western Ontario, London, Ont., 1973), Part I*, pages 116–309. Lecture Notes in Econom. and Math. Systems, Vol. 105. Springer, Berlin, 1974.
59. J.-L. Lions. *Contrôle des systèmes distribués singuliers*, volume 13 of *Méthodes Mathématiques de l'Informatique [Mathematical Methods of Information Science]*. Gauthier-Villars, Montrouge, 1983.
60. J.-L. Lions and G. Stampacchia. Variational inequalities. *Comm. Pure Appl. Math.*, 20:493–519, 1967.
61. Z.-Q. Luo, J.-S. Pang, and D. Ralph. *Mathematical Programs with Equilibrium Constraints*. Cambridge University Press, Cambridge, 1996.
62. C. Meyer and O. Thoma. A priori finite element error analysis for optimal control of the obstacle problem. *SIAM J. Numer. Anal.*, 51(1):605–628, 2013.
63. F. Mignot. Contrôle dans les inéquations variationelles elliptiques. *J. Functional Analysis*, 22(2):130–185, 1976.
64. F. Mignot and J.-P. Puel. Optimal control in some variational inequalities. *SIAM J. Control and Optimization*, 22(3):466–476, 1984.
65. K. Mombaur, A. Truong, and J.-P. Laumond. From human to humanoid locomotion–an inverse optimal control approach. *Autonomous Robots*, 28(3):369–383, 2010.
66. B. S. Mordukhovich. *Variational analysis and generalized differentiation. I: Basic Theory*, volume 330 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, Berlin, 2006.

67. D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. Pure Appl. Math.*, 42(5):577–685, 1989.
68. P. Neittaanmäki, J. Sprekels, and D. Tiba. *Optimization of Elliptic Systems*. Springer Monographs in Mathematics. Springer, New York, 2006.
69. J. V. Outrata, M. Kočvara, and J. Zowe. *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints*, volume 28 of *Nonconvex Optimization and its Applications*. Kluwer Academic Publishers, Dordrecht, 1998.
70. P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE transactions on pattern analysis and machine intelligence*, 12(7):174–184, 1990.
71. J.-F. Rodrigues. *Obstacle Problems in Mathematical Physics*. Number 134 in North-Holland Mathematics Studies. North-Holland Publishing Co., Amsterdam, 1984.
72. L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1–4):259–268, 1992. Experimental mathematics: computational issues in nonlinear science (Los Alamos, NM, 1991).
73. C. Saguez. Optimal control of free boundary problems. In *System modelling and optimization (Budapest, 1985)*, volume 84 of *Lecture Notes in Control and Inform. Sci.*, pages 776–788. Springer, Berlin, 1986.
74. H. Scheel and S. Scholtes. Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity. *Mathematics of Operations Research*, 25(1):1–22, 2000.
75. H. Schramm and J. Zowe. A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results. *SIAM Journal on Optimization*, 2(1):121–152, 1992.
76. N. Z. Shor. *Minimization Methods for Non-differentiable Functions*. Springer-Verlag, New York, 1985.
77. F. Tröltzsch. *Optimal control of partial differential equations*, volume 112 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2010. Theory, methods and applications, Translated from the 2005 German original by Jürgen Sprekels.
78. J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Trans. Automat. Control*, 31(9):803–812, 1986.
79. M. Ulbrich. Semismooth Newton methods for operator equations in function spaces. *SIAM J. Optim.*, 13(3):805–842 (2003), 2002.
80. G. Wachsmuth. Strong stationarity for optimal control of the obstacle problem with control constraints. *SIAM J. Optim.*, 24(4):1914–1932, 2014.
81. G. Wachsmuth. Mathematical programs with complementarity constraints in Banach spaces. *J. Optim. Theory Appl.*, 166(2):480–507, 2015.
82. G. Wachsmuth. A guided tour of polyhedric sets: Basic properties, new results on intersections and applications. Technical report, TU Chemnitz, 2016.
83. G. Wachsmuth. Towards M-stationarity for optimal control of the obstacle problem with control constraints. *SIAM J. Control Optim.*, 54(2):964–986, 2016.
84. S. W. Walker, A. Bonito, and R. H. Nochetto. Mixed finite element method for electrowetting on dielectric with contact line pinning. *Interfaces Free Bound.*, 12(1):85–119, 2010.
85. Y-S. Wu, K. Pruess, and P. A. Witherspoon. Flow and displacement of Bingham non-Newtonian fluids in porous media. *SPE Reservoir Engineering*, 7(3), 1992.
86. J.-P. Yvon. *Etude de quelques problèmes de contrôle pour des systèmes distribués*. PhD thesis, Paris VI, 1973.
87. J.-P. Yvon. Optimal control of systems governed by variational inequalities. In *Fifth Conference on Optimization Techniques (Rome, 1973), Part I*, pages 265–275. Lecture Notes in Comput. Sci., Vol. 3. Springer, Berlin, 1973.
88. J. Zowe and S. Kurcyusz. Regularity and stability for the mathematical programming problem in Banach spaces. *Appl. Math. Optim.*, 5(1):49–62, 1979.

Introduction to PDE-Constrained Optimization in the Oil and Gas Industry

Jeremy Brandman, Huseyin Denli, and Dimitar Trenev

Abstract This article is an expanded version of a tutorial on applications of PDE-constrained optimization in the oil and gas industry that was given at the Frontiers in PDE-Constrained Optimization workshop. (The workshop was held at the Institute for Mathematics and its Applications June 6–10, 2016.) We begin with an overview of the oil and gas supply chain that highlights the importance of PDE-constrained optimization. Next, we take an in-depth look at two key applications: full-wavefield inversion and reservoir history matching. For each application, we introduce a PDE model, derive the gradient of the objective function using the adjoint-state method, and present simple numerical results. We conclude with a discussion of key challenges.

1 Introduction

Oil and gas[1] currently provide over half of the world's energy supply and are expected to continue to do so in the coming decades [1]. Because of its high energy density and ease of transport, oil meets close to 95% of global transportation energy demand in addition to being used to make plastics, lubricants, asphalt, and other products [1, 78]. Natural gas is the cleanest-burning major fuel; as a result, it is commonly used for power generation and is also emerging as a fuel for heavy-duty trucks and marine transportation.

The oil and gas supply chain can be broken down into multiple stages: exploration, development, production, transportation, and processing at refineries and chemical plants [78]. Exploration focuses on identifying subsurface hydrocarbon

[1]Throughout this paper, the term hydrocarbons is used interchangeably with oil and gas.

J. Brandman (✉) · H. Denli · D. Trenev
Corporate Strategic Research, ExxonMobil Research and Engineering Company, Annandale, NJ 08801, USA
e-mail: jeremy.s.brandman@exxonmobil.com

H. Antil et al. (eds.), *Frontiers in PDE-Constrained Optimization*, The IMA Volumes in Mathematics and its Applications 163,
https://doi.org/10.1007/978-1-4939-8636-1_5

reservoirs. This is an indirect process since hydrocarbon deposits are typically located thousands of meters below the ground. In this stage, multidisciplinary teams work together to identify the presence and volume of hydrocarbons. Geophysicists and applied mathematicians analyze the geophysical data sets available in an effort to infer properties of the subsurface (e.g., elastic, electric, gravitational, and magnetic properties), while geologists use these properties to interpret the large-scale geologic features essential to the creation of hydrocarbon reservoirs.

Once a potential reservoir is identified, exploratory wells are drilled to determine whether or not hydrocarbons are present. If hydrocarbons are found, we move on to the development stage. In this stage, appraisal wells may be drilled to better determine the extent of the reservoir and a depletion plan for the reservoir is formulated. The depletion plan specifies the location and operating conditions of the wells (e.g., injection and production rates) used to produce hydrocarbons and is designed to maximize profitability of the reservoir. The actual extraction of hydrocarbons occurs in the production stage.

Following production, hydrocarbons are carried by pipeline, oil tanker, truck, or railcar to refineries and chemical plants [31]. At these facilities, hydrocarbons are converted into finished products such as gasoline, kerosene, and petrochemicals used in the manufacture of plastics and agricultural fertilizers [31, 78].

PDE-constrained optimization is poised to make important contributions to some of the engineering and scientific challenges present in the oil and gas supply chain. Specific applications of PDE-constrained optimization include: subsurface inversion techniques for exploring increasingly difficult environments (e.g., deepwater offshore) - these techniques include full-wavefield inversion, electromagnetic inversion, gravity inversion, and process stratigraphy; optimal placement and operation of wells, each of which can cost tens of millions of dollars; the calibration of subsurface geology to production data; imaging techniques for flaw detection in pipelines to reduce leaks; inversion of atmospheric measurements (tower, airplane, and satellite) to identify sources of greenhouse gas emissions (e.g., carbon dioxide, and methane); and optimization of chemical plant and refinery operations [12, 16, 18, 40, 47, 55, 68, 71].

In this paper, we focus on full-wavefield inversion and reservoir history matching. Full-wavefield inversion is a PDE-constrained optimization approach to the exploration technique known as reflection seismology. Full-wavefield inversion is used to infer subsurface mechanical rock properties (e.g., elastic moduli) from surface measurements of waves traveling underground which are triggered by an applied seismic source. These rock properties are used by geologists to predict the locations and volumes of hydrocarbons.

The second problem considered in this paper - reservoir history matching - arises during the production stage. Reservoir history matching is the problem of determining a model of fluid flow within the reservoir which is consistent with the recorded production data measured at the wells (e.g., pressure and flow rates). Such a flow model is used to predict future production within the reservoir and guide changes in the depletion plan, including the placement of additional wells and modifications to existing wells' operating conditions.

For each problem, we first describe the physical setup and the corresponding optimization problem. We then explain how the gradient of the objective function is computed using the adjoint-state method and show numerical results for model problems. Numerical results for problems of industrial complexity in full-wavefield inversion can be found in the companion paper [55]. We conclude our discussion of each problem with a list of key challenges.

2 Full-Wavefield Inversion

In this section, we provide an introduction to full-wavefield inversion (FWI) designed to be accessible to applied mathematicians. For an introduction focused more on geophysics, please see [55, 87].

Hydrocarbons are typically found thousands of meters below the earth's surface. In order to locate these resources, the oil and gas industry heavily relies on reflection seismology. Reflection seismology infers properties of the subsurface rocks through the measurement of reflected waves resulting from man-made sources at the surface. The reflection of waves, along with other, more complicated wave phenomena, occurs as a result of layers present in the earth's interior formed by geologic processes such as sedimentation, faulting, and fracturing [76]. Knowledge of these layers, as obtained by reflection seismology, is ultimately used to infer the presence of hydrocarbon reservoirs.

Seismic surveys are carried out offshore and on land; in recent decades, offshore developments have become increasingly important. An example of a marine seismic survey and the accompanying reflections is shown in Figure 1(a); land surveys are configured similarly. Sources are conventionally fired sequentially and the response to each is recorded at the available receivers.

Apart from FWI, conventional workflows for analyzing seismic data suffer from two shortcomings: they typically do not use all of the information present

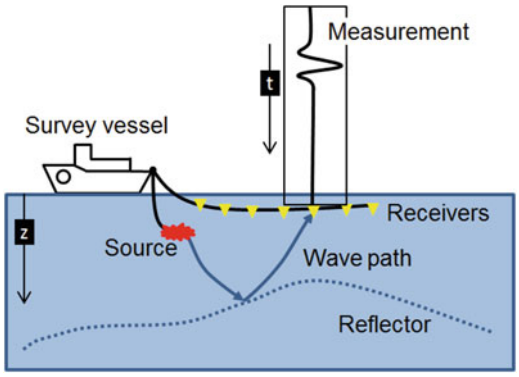

Fig. 1 Illustration of a marine seismic survey

in the measured data and they are labor intensive. In contrast, FWI is a PDE-constrained optimization approach capable of fully utilizing the available seismic data. In addition, FWI provides a framework for incorporating all prior geologic knowledge (e.g., rock physics relationships, well logs, and bounds on rock properties). As a result, it offers the promise of producing higher resolution subsurface images for more accurately determining hydrocarbon deposits [34].

In its simplest form, FWI seeks a subsurface model that matches all available seismic data to it's noise level. The selection of such a model is made through the minimization of an objective function consisting of a least-squares data misfit and a regularization term:

$$\begin{aligned}\mathcal{J}(\kappa) &= \frac{1}{2}\sum_{\substack{\text{Sources}\\ s}} \mathcal{J}_s + R(\kappa) \\ &= \frac{1}{2}\sum_{\substack{\text{Sources}\\ s}}\sum_{\substack{\text{Receivers}\\ x_r}} \int_{t_{0,s}}^{t_{0,s}+T} |u_s(\kappa; x_r, t) - \bar{u}(x_r, t)|^2 \, dt + R(\kappa).\end{aligned} \tag{1}$$

Here, $\kappa(\cdot)$ represents the unknown subsurface properties to be determined through minimization of (1); these properties can include wave velocities, anisotropy, attenuation parameters, and the source temporal signature in some cases [5, 30]. These terms are defined in Section 2.1. In addition, $t_{0,s}$ is the time at which the source s is fired, T represents the length of the time interval during which data is collected for each source, $\bar{u}(x_r, t)$ is the measured geophysical data at the receiver location x_r at time t, $u_s(\kappa; x_r, t)$ is data due to source s at the receiver location x_r at time t computed according to a PDE model, and $R(\kappa)$ is a regularization term used to mitigate the ill-posedness of the inversion (e.g., Tikhonov or total-variation regularization).

Because this paper is intended to serve as a tutorial, most of our discussion of FWI focuses on the objective function (1). However, significant modifications of (1) are required for successful application of FWI to problems of realistic complexity. This is due to two factors: the presence of critical points in the objective function (1) that are not necessarily global minima and the sheer magnitude of the computational effort required to solve the wave equation. These limitations prevented industry-scale applications of FWI for several decades [87]; surmounting them required the advent of innovative but still poorly understood continuation strategies [20] and a new generation of supercomputers. We discuss these important developments in Section 2.5.

2.1 Wave Propagation in Elastic Media

We assume that the subsurface is a linear elastic medium when modeling the propagation of subsurface waves. This is a reasonable assumption in light of the small displacements observed in seismic surveys away from the sources.

Recall that a medium is elastic if it deforms when forces are applied to it and returns to its original state when those forces are removed. Linear elasticity represents a simplification of elasticity appropriate for a medium undergoing small deformation. The key feature of linear elasticity is the generalized Hooke's law, which states that stress and strain are linearly related through a stiffness tensor. This relationship can be written as

$$\boldsymbol{\sigma} = \mathbf{C} : \boldsymbol{\epsilon} \tag{2}$$

where $\boldsymbol{\sigma}$ is the stress tensor, $\mathbf{C}$ is the fourth-order stiffness tensor, $\boldsymbol{\epsilon}$ is the strain tensor defined by

$$\boldsymbol{\epsilon} = 1/2 \left(\nabla \mathbf{u} + (\nabla \mathbf{u})^T \right), \tag{3}$$

and $\mathbf{u}$ is the vector describing the medium's displacement from its equilibrium configuration [6, 46].

Subsurface waves also satisfy the conservation of momentum, which can be expressed as

$$\rho \frac{\partial^2 \mathbf{u}}{\partial t^2} = \nabla \cdot \boldsymbol{\sigma} + \mathbf{f}, \tag{4}$$

where $\rho(x)$ denotes a time-independent density and $\mathbf{f}(x, t)$ is an external force density. The system (2), (3), (4) specifies the time-dependent deformation of a linear elastic medium once initial and boundary conditions are specified.

The stiffness tensor $\mathbf{C}$ has twenty-one independent components; this number is reduced to two when the medium is isotropic. For an isotropic medium, the stiffness tensor satisfies

$$\mathbf{C} : \boldsymbol{\epsilon} = \lambda \text{tr}(\boldsymbol{\epsilon}) \mathbf{I} + 2\mu \boldsymbol{\epsilon}, \tag{5}$$

where $\lambda(x)$ and $\mu(x)$ are the Lamé parameters. We may combine (2), (3), (4), (5), to arrive at a simplified time-dependent PDE describing the deformation of an isotropic medium:

$$\rho \frac{\partial^2 \mathbf{u}}{\partial t^2} = \nabla \left(\lambda \nabla \cdot \mathbf{u} \right) + \nabla \cdot \left(\mu \left(\nabla \mathbf{u} + (\nabla \mathbf{u})^T \right) \right) + \mathbf{f}. \tag{6}$$

For homogeneous materials, in which λ and μ are constant, the system (6) can be further simplified using vector calculus identities to arrive at

$$\begin{aligned} \rho \frac{\partial^2 \mathbf{u}}{\partial t^2} &= (\lambda + \mu) \nabla \nabla \cdot \mathbf{u} + \mu \nabla^2 \mathbf{u} + \mathbf{f} \\ &= (\lambda + 2\mu) \nabla (\nabla \cdot \mathbf{u}) - \mu \nabla \times \nabla \times \mathbf{u} + \mathbf{f}. \end{aligned} \tag{7}$$

Solutions of (7) can be expressed as $\mathbf{u} = \nabla\phi + \nabla \times \boldsymbol{\psi}$, where the scalar ϕ and vector $\boldsymbol{\psi}$ potentials satisfy wave equations [29, 76, 81]. Waves corresponding to $\boldsymbol{\phi}(x, t)$ are known as longitudinal, or compressional, since they correspond to irrotational volumetric changes ($\nabla \times \mathbf{u} = 0$, $\nabla \cdot \mathbf{u} \neq 0$) in the deformed material without shear. These travel at a speed $v_P = \sqrt{(\lambda + 2\mu)/\rho}$. Waves arising from $\boldsymbol{\psi}(x, t)$ are known as transverse, or shear, since they correspond to shear deformation. These travel at a speed $v_S = \sqrt{\mu/\rho}$ which is slower than v_P due to $\mu > 0$ and $\lambda + 2\mu/3 > 0$.[2] The interested reader may consult [29, 76, 81] for derivations of (6), (7).

Next, we consider wave propagation in a fluid. In a fluid, the shear modulus of the stress tensor $\boldsymbol{\sigma}$ is zero and the stress tensor becomes

$$\boldsymbol{\sigma} = \kappa \, (\nabla \cdot \mathbf{u}) \, \mathbf{I} \tag{8}$$

where κ is the bulk modulus. The pressure p is defined as

$$p = -\kappa \nabla \cdot \mathbf{u} \tag{9}$$

so that $\boldsymbol{\sigma} = -p\mathbf{I}$. Differentiation in time of (9) and substitution from (4) leads to an acoustic wave equation for the pressure p in which pressure disturbances propagate at a speed $c = \sqrt{\frac{\kappa}{\rho}}$ [80]:

$$\frac{1}{\kappa} \frac{\partial^2 p}{\partial t^2} = \nabla \cdot \left(\frac{1}{\rho} \nabla p \right) - \nabla \cdot \frac{\mathbf{f}}{\rho}. \tag{10}$$

The approximation (10) is sometimes used to model wave propagation in solids in order to arrive at a simpler and more computationally tractable problem. In this case, it is common practice to filter out elastic effects from the data as much as possible [86].

Additional aspects of wave propagation, namely anisotropy and attenuation, must be taken into account in order to fully explain the measured seismic data. Anisotropy is due to either the presence of materials, such as clay, which are intrinsically anisotropic, or the presence of features at sub-seismic wavelengths, such as rock layers and ordered fractures, whose measured response is equivalent to that of a homogeneous anisotropic medium due to averaging of the underlying rock properties [84, 85]. Seismic attenuation accounts for the loss of energy in the wavefield due to conversion into other forms such as heat or fluid motion. Further discussion of anisotropy and attenuation can be found in [55].

We conclude this section with a brief discussion of numerical methods for solving the acoustic wave equation (10). Adopting the convention that u denotes the solution of the acoustic wave equation on a bounded domain Ω, we have that u satisfies

[2] These two inequalities arise from the requirement that the quadratic form $\frac{1}{2}\langle \mathbf{C} : \mathbf{e}, \mathbf{e} \rangle$ defining the elastic energy of the deformed material is positive definite [46]. In particular, the latter inequality ensures that the bulk modulus, $\kappa := \lambda + \frac{2\mu}{3}$, is positive.

$$\begin{aligned} \frac{1}{\kappa}\frac{\partial^2 u}{\partial t^2} &= \nabla\cdot\frac{1}{\rho}\nabla u - \nabla\cdot\frac{\mathbf{f}}{\rho}, \\ u = \frac{\partial u}{\partial t} &= 0 \text{ at } t = 0 \text{ in } \Omega, \\ u &= 0 \text{ on } \Gamma_{\text{free}} \times [0, \mathrm{T}], \\ \sqrt{\frac{\rho}{\kappa}}\frac{\partial u}{\partial t} + \mathbf{n}\cdot\nabla u &= 0 \text{ on } \Gamma_{\text{absorbing}} \times [0, \mathrm{T}], \end{aligned} \tag{11}$$

where T is the final recording time, Γ_{free} and $\Gamma_{\text{absorbing}}$ are the free and absorbing boundaries, and $\mathbf{n}$ is the normal unit vector to $\Gamma_{\text{absorbing}}$. Typically, Γ_{free} represents the earth's surface (i.e., the air-water interface in marine environments or the air-solid interface on land) due to the approximately stress-free state there. The remaining boundaries of the computational domain typically belong to $\Gamma_{\text{absorbing}}$; the boundary condition applied there is designed to prevent spurious reflected waves due to the truncation of the physical domain [26, 53]. Finally, we approximate the source term for a shot fired at location x_s by

$$\nabla\cdot\frac{\mathbf{f}}{\rho} \approx \frac{f(t)}{\rho}\delta(x - x_s) \tag{12}$$

under the assumption that the source term is spherically symmetric and nonlinear effects in the response can be neglected. The function $f(t)$ is known as the source temporal signature or the source wavelet.

A variety of numerical techniques exist for solving the acoustic wave equation (11). The discussion in this paper focuses on time-domain methods. When solving (11) in the time domain, we typically discretize using the method of lines [57]. First, we discretize in space using finite differences [64, 82], finite volumes [56], finite elements (e.g., spectral elements [52] and the discontinuous-Galerkin method [43, 50, 83]), or some other technique. The resulting system of ODEs is then integrated using an appropriate time-stepping scheme (e.g., forward Euler, and Runge-Kutta) [89]. For a discussion of alternatives to time-domain methods, the interested reader should consult [55].

2.2 *Gradient Computation for the Acoustic Wave Equation*

We compute the gradient of the objective function (1) using the adjoint-state method. The gradient is derived in the continuous setting; afterwards, we explain how a similar approach can be taken when the forward problem is discretized. Our derivation, inspired by Marchuk [59], aims to provide intuition but does not maintain the mathematical rigor that can be found in more careful treatments such as [44].

The power of the adjoint-state method is its efficiency. In practice, $\kappa(x)$ is discretized, resulting in a vector $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_N)$ of unknowns. Computing the

gradient of the objective function using finite differences requires solving $\mathcal{O}(N)$ forward problems. In contrast, the cost of computing the gradient using the adjoint-state method is approximately that of solving two forward problems.

In order to streamline the derivation presented below, we omit regularization from the objective function (1) and compute the gradient $\nabla \mathcal{J}_s$ for the case of a single seismic source. Due to independence of the sources, it follows that

$$\nabla \mathcal{J} = \sum_{\substack{\text{Sources} \\ s}} \nabla \mathcal{J}_s.$$

In addition, we simplify the acoustic wave equation (11) by assuming $\rho \equiv 1$ and set $\Omega = \mathbb{R}^n$, $t_{0,s} = 0$. Under these assumptions, the acoustic wave equation takes the form

$$\begin{aligned} \frac{1}{\kappa} \frac{\partial^2 u}{\partial t^2} - \Delta u &= f \text{ in } \mathbb{R}^n \times [0, T], \\ u = \frac{\partial u}{\partial t} &= 0 \text{ at } t = 0 \text{ in } \mathbb{R}^n, \end{aligned} \tag{13}$$

where the source f is assumed to have compact support. We use the notation $u(\kappa; f)$ to refer to the solution of (13) determined by bulk modulus κ and source f.

Next, we define the forward and adjoint operators for a given bulk modulus $\kappa(x)$. The forward operator A is defined by

$$Af = u(\kappa; f). \tag{14}$$

Note that this is the inverse of the differential operator. The adjoint operator A^*, which plays an important role in computing $\nabla \mathcal{J}_s$, is defined by the adjoint identity

$$(Af, g)_{L^2(\mathbb{R}^n \times [0,T])} = \left(f, A^* g\right)_{L^2(\mathbb{R}^n \times [0,T])} \ \forall f, g \in L^2\left(\mathbb{R}^n \times [0, T]\right). \tag{15}$$

We will show that the adjoint operator is given by

$$A^* g = v(\kappa; g), \tag{16}$$

where $v(\kappa; g)$ solves the adjoint problem

$$\begin{aligned} \frac{1}{\kappa} \frac{\partial^2 v}{\partial t^2} - \Delta v &= g \text{ in } \mathbb{R}^n \times [0, \mathrm{T}], \\ v = \frac{\partial v}{\partial t} &= 0 \text{ at } t = T \text{ in } \mathbb{R}^n. \end{aligned} \tag{17}$$

Equation (17) is also an acoustic wave equation but, in contrast to (13), the solution's final-time value is prescribed. Equation (17) can be solved by marching backwards in time, beginning at $t = T$.

We will establish that the operator given by (16), (17) satisfies the adjoint identity (15) through integration by parts. First, by definition of the forward and adjoint problems, we have

$$(Af, g)_{L^2(\mathbb{R}^n \times [0,T])} = \int_0^T \int_{\mathbb{R}^n} u \cdot \left(\frac{1}{\kappa} \frac{\partial^2 v}{\partial t^2} - \Delta v \right) dx\, dt. \tag{18}$$

Next, we integrate the temporal derivatives of (18) by parts. Due to the initial- and final-time boundary conditions of the forward and adjoint problems, respectively, we have

$$\int_0^T \int_{\mathbb{R}^n} u \cdot \frac{1}{\kappa} \frac{\partial^2 v}{\partial t^2}\, dx\, dt = -\int_0^T \int_{\mathbb{R}^n} \frac{1}{\kappa} \frac{\partial u}{\partial t} \frac{\partial v}{\partial t}\, dx\, dt = \int_0^T \int_{\mathbb{R}^n} \frac{1}{\kappa} \frac{\partial^2 u}{\partial t^2} \cdot v\, dx\, dt. \tag{19}$$

Finally, we integrate the spatial derivatives of (18) by parts. Assuming compact support of the sources f, g, it follows from the divergence theorem that

$$\int_0^T \int_{\mathbb{R}^n} u \cdot \Delta v\, dx\, dt = -\int_0^T \int_{\mathbb{R}^n} \nabla u \cdot \nabla v\, dx\, dt = \int_0^T \int_{\mathbb{R}^n} \Delta u \cdot v\, dx\, dt. \tag{20}$$

Combining (18), (19), (20) results in (15), thus demonstrating that the adjoint operator A^* is given by Equations (16), (17).

We now use the adjoint identity (15) to compute the $L^2(\mathbb{R}^n)$ gradient $\nabla \mathcal{J}_s$ with respect to κ. The defining property of $\nabla \mathcal{J}_s$ is

$$\mathcal{J}_s(\kappa + \delta\kappa) - \mathcal{J}_s(\kappa) = (\nabla \mathcal{J}_s, \delta k)_{L^2(\mathbb{R}^n)} + \mathcal{O}\left(\|\delta\kappa\|^2_{L^2(\mathbb{R}^n)} \right). \tag{21}$$

In order to compute $\nabla \mathcal{J}_s$, we begin by linearizing the nonlinear map $u(\kappa; f)$ with respect to κ. For a small perturbation $\delta\kappa$, we define δu by

$$u(\kappa + \delta\kappa; f) \approx u(\kappa; f) + \delta u + \mathcal{O}\left(\|\delta\kappa\|^2_{L^2(\mathbb{R}^n)} \right). \tag{22}$$

Inserting the approximation (22) into (13), it follows that δu satisfies the linearized forward problem

$$\begin{aligned} \frac{1}{\kappa} \frac{\partial^2 \delta u}{\partial t^2} - \Delta \delta u &= \frac{\delta\kappa}{\kappa^2} \frac{\partial^2 u}{\partial t^2} \text{ in } \mathbb{R}^n \times [0, \mathrm{T}], \\ \delta u = \frac{\partial \delta u}{\partial t} &= 0 \text{ at } t = 0 \text{ in } \mathbb{R}^n. \end{aligned} \tag{23}$$

Observing that the right-hand side of (23) represents a source term, we rewrite (23) using the definition (14) as

$$\delta u = A\left(\frac{\delta\kappa}{\kappa^2} \frac{\partial^2 u}{\partial t^2} \right). \tag{24}$$

In particular, we set

$$\delta u_s = A\left(\frac{\delta\kappa}{\kappa^2}\frac{\partial^2 u_s}{\partial t^2}\right). \tag{25}$$

Interestingly, the quality of the linearization (24) is quite variable and depends on several factors, including the smoothness of κ and the degree of oscillation present in $\delta\kappa$ [80].

We now have all of the necessary results for computing ∇J_s. Using the linearization (22), it follows that

$$\mathcal{J}_s(\kappa+\delta\kappa)-\mathcal{J}_s(\kappa) \approx \left(\left(\sum_{\text{Receivers } x_r}\delta(x-x_r)\right)(u_s(\kappa)-\bar{u}),\delta u_s\right)_{L^2(\mathbb{R}^n\times[0,T])}. \tag{26}$$

Substitution of (25) into (26), followed by an application of the adjoint identity (15), results in

$$\begin{aligned}\mathcal{J}_s(\kappa+\delta\kappa)-\mathcal{J}_s(\kappa) &= \left(\left(\sum_{\text{Receivers } x_r}\delta(x-x_r)\right)(u_s(\kappa)-\bar{u}),\right.\\ &\quad \left.A\left(\frac{\delta\kappa}{\kappa^2}\frac{\partial^2 u_s}{\partial t^2}\right)\right)_{L^2(\mathbb{R}^n\times[0,T])}\\ &= \left(A^*\left[\left(\sum_{\text{Receivers } x_r}\delta(x-x_r)\right)(u_s(\kappa)-\bar{u})\right],\right.\\ &\quad \left.\frac{\delta\kappa}{\kappa^2}\frac{\partial^2 u_s}{\partial t^2}\right)_{L^2(\mathbb{R}^n\times[0,T])}.\end{aligned} \tag{27}$$

Recalling the definition (21) of $\nabla\mathcal{J}$, it follows from (27) that

$$\nabla\mathcal{J}_s = \int_0^T \frac{1}{\kappa^2}\frac{\partial^2 u_s}{\partial t^2}\cdot v\,dt \tag{28}$$

where v solves the adjoint problem

$$v = A^*\left[\left(\sum_{\text{Receivers } x_r}\delta(x-x_r)\right)(u_s(\kappa)-\bar{u})\right]. \tag{29}$$

Equations (28), (29) indicate that computing $\nabla\mathcal{J}_s$ requires the solution of the forward problem (13), where the source term is assumed known, and the adjoint problem (17), in which the data misfit is injected as a source term at the receivers. The computational cost of solving the adjoint problem is approximately the same as the cost of the forward, since both are wave equations in the same medium.

In practice, computing the integral (28) is more complicated than it may appear and comes at a higher cost than simply solving for u and v. One reason for this is

that u satisfies a forward-in-time wave equation but v satisfies a backward-in-time equation. It follows that values of either u or v must be stored. A second reason is that, for 3D problems, insufficient memory exists to store u and v on the entire time interval. Due to these constraints, a procedure known as check-pointing is commonly used to compute (28). Check-pointing divides the time interval $[0, T]$ into a number of subintervals, stores the values of u at the endpoints of each subinterval, and uses these values to recompute (28) separately on each subinterval. Computing (28) on each subinterval requires recomputing and storing the values of u on that subinterval [8, 38]. This leads to additional computation, but the overall cost of computing (28) using check-pointing is not greater than the solution of three forward problems.

The derivation above assumed that $\kappa(x)$ was distributed continuously in space. In practice, the bulk modulus $\kappa(x)$ is discretized, resulting in a vector $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_N)$ of unknowns. When optimizing the objective function (1) with respect to $\boldsymbol{\kappa}$, one can either solve the adjoint equation (29) using a discretization of one's choice or one can determine the discrete adjoint problem [42] which ensures that a discrete analogue of the identity (15) is satisfied . From an implementation point of view, the first approach is easier. However, the resulting approximation of the gradient may have significant errors. In contrast, using the discrete adjoint operator ensures that the discrete objective function's gradient with respect to $\boldsymbol{\kappa}$ is computed correctly. The interested reader can learn more about the trade-offs between these approaches in [39].

2.3 Optimization

Because the number of unknowns in industry-scale problems is typically large (on the order of billions) and the cost of a forward simulation is also significant (on the order of minutes to hours on high-performance computing platforms), it is prohibitive to use global optimization methods except in low-dimensional formulations [28]. Instead, local gradient-based methods are typically used to minimize the objective function (1).

Two classes of gradient-based optimization techniques are available for solving large-scale PDE-constrained optimization problems. The first class eliminates the PDE constraint and directly optimizes the set of unknown parameters appearing in the PDE; this class includes methods such as steepest descent, nonlinear conjugate gradient, and variants of Newton's method. The second class of methods does not eliminate the PDE constraint but instead solves for both the unknown parameters appearing in the PDE and the solution of the PDE itself; this class includes techniques such as the augmented-Lagrangian method and sequential quadratic programming. The reader may consult [51, 66] for a general introduction to optimization and [42, 44] for a discussion of the practical issues which arise in the context of PDE-constrained optimization.

Large-scale PDE-constrained optimization problems such as FWI and reservoir history matching rely on the first class of methods because storing the solution of the

entire forward problem, as required by the second class, places prohibitive demands on memory. This class of methods iteratively update the vector of discretized unknowns $\boldsymbol{\kappa}$ along a descent direction,

$$\boldsymbol{\kappa}^{i+1} = \boldsymbol{\kappa}^i + \alpha^i \mathbf{p}^i, \tag{30}$$

where i is the iteration index, $\mathbf{p}^i$ is the ith descent direction, and α_i is the ith step size determined by a line search algorithm. Within this class, trade-offs exist between the rate of convergence, ease of implementation, and memory requirements. Steepest descent converges at a linear rate and is straightforward to implement, the nonlinear conjugate gradient method is more complicated but generally converges faster, and variants of Newton's method can achieve superlinear convergence but may involve greater algorithmic complexity and memory requirements. A comparison of some of these methods, in the context of FWI, was carried out in [63].

Steepest descent sets $\mathbf{p}^i = -\mathbf{g}^i$, where $\mathbf{g}^i$ is the gradient of the objective function. Newton-type methods determine $\mathbf{p}^i$ by solving the linear system

$$\mathbf{H}^i \mathbf{p}^i = -\mathbf{g}^i \tag{31}$$

where $\mathbf{H}^i$ is a positive-definite approximation to the objective function's Hessian. These methods rely on an approximate Hessian because, for large-scale problems, the complete Hessian is prohibitively large to compute and store. Two of the most commonly used methods from this family are limited-memory quasi-Newton methods and the Gauss-Newton method. Limited-memory quasi-Newton methods rely on a low-rank approximation to the Hessian; the Gauss-Newton method exploits the least-squares framework present in the objective function (1) to derive an approximate Hessian for which matrix-vector products can be computed without explicitly storing the matrix [42]. Application of such methods to (1) is by no means straightforward, as a variety of complications may arise; these include sensitivity of quasi-Newton methods to the Hessian initialization and suboptimal performance of Gauss-Newton due to poor conditioning.

In the context of the objective function (1), inequality constraints arise naturally from physical considerations - for example, the bulk modulus must be positive in acoustic inversion while, in anisotropic elasticity, additional constraints must be satisfied. Due to the infinite-dimensional nature of the objective function (1), imposing bound constraints on the control parameters presents challenges not found in finite-dimensional analogues. For example, active-set methods may not converge in finitely many iterations and the projected gradient method can also lead to problems. Further discussion of these and other issues can be found in the book [51]. Enforcing nonlinear inequality control constraints as well as state constraints is even more challenging; the interested reader can find more information in [41, 44].

2.4 A Synthetic Example

In this section, we consider a synthetic example based on the Marmousi model. The example is synthetic because the data does not arise from an actual seismic survey. Instead, it is generated by solving (11) on the Marmousi model shown in Figure 2 [61]. The Marmousi model is commonly used as a benchmark to test FWI codes.

Presenting this example has two goals. The first goal is to demonstrate that, for an appropriate initial guess and source signature, acoustic FWI can successfully invert for the subsurface bulk modulus. The second goal is to highlight the sensitivity of the inversion to the choice of initial guess.

The Marmousi model extends 17 km horizontally and 3.1 km vertically and assumes a homogeneous mass density of $\rho = 1\,\text{g/cm}^3$. Synthetic seismic data is generated by performing 100 different experiments using 100 different sources and measuring the response at the same 500 receivers. The sources and receivers are equidistantly spaced horizontally at 3 m and 6 m below the surface, respectively. We use the same source temporal signature for each source (as defined by Equation (12)) - the 5 Hz peak-frequency Ricker wavelet shown in Figure 3.

The acoustic wave equation (11) is solved using a spectral-element spatial discretization and a central-difference explicit temporal discretization scheme. A discrete version of the objective function (1) is minimized using Gauss-Newton.

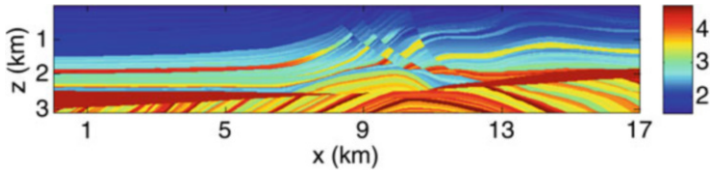

Fig. 2 Marmousi acoustic wave-velocity model. The color bar displays the velocity in km/s [61]

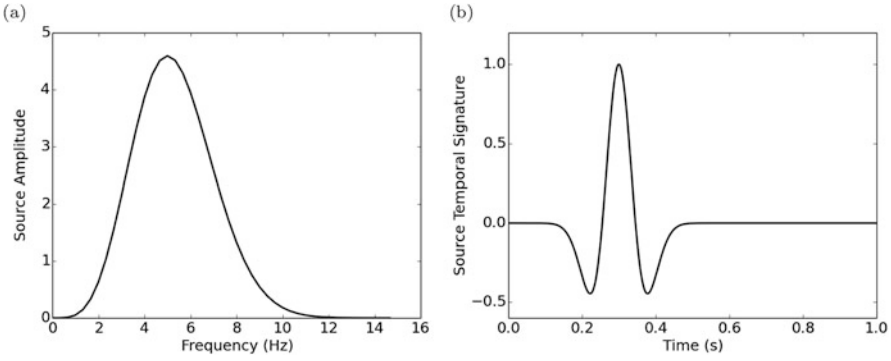

Fig. 3 Seismic Ricker source signature in frequency (a) and time (b) domains

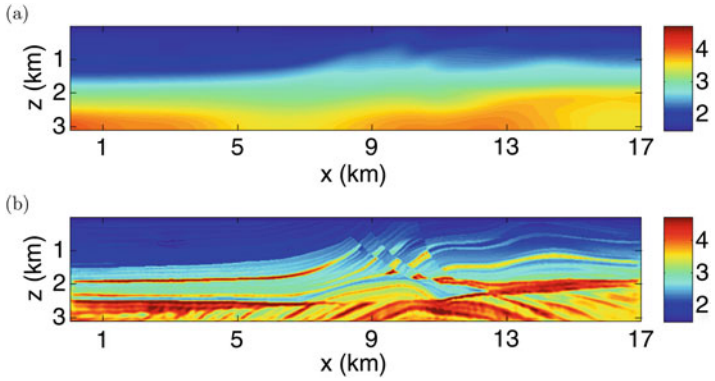

Fig. 4 (a) An initial velocity model and (b) inverted velocity model starting from the initial model given in (a)

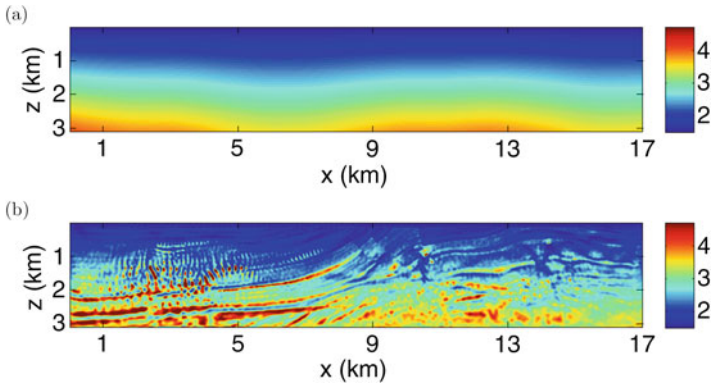

Fig. 5 (a) An initial velocity model and (b) inverted velocity model starting from the initial model given in (a)

First-order Tikhonov regularization is used. The same mesh and discretization are used to both generate the synthetic data and then invert it. In general, this can lead to misleading results and should be avoided [48].

We consider the inversion of the Marmousi model using two different initial guesses. Our first initial guess, shown in Figure 4(a), is obtained by smoothing the Marmousi model using a Gaussian filter. This initial guess preserves some of the most significant large-scale trends present in the Marmousi model. Figure 5(a) is obtained by filling in the model through linear interpolation, using only values at the top and bottom of the first initial guess. The resulting inversions are found in Figures 4(b) and 5(b), respectively. Both inversion experiments are terminated after 200 Gauss-Newton iterations, at which point the optimization appears to stall.

We can make two important observations based on these results. The first is that, for an appropriate initial guess, FWI is capable of converging to a model close to

the target model. The second is that FWI is highly sensitive to the choice of initial guess. Further discussion of the second observation can be found in Section 2.5.

2.5 Challenges in Full-Wavefield Inversion

Despite decades of research and progress, FWI in the context of oil and gas exploration continues to face significant obstacles. These include various computational challenges - the elastic wave equation is prohibitively expensive for many applications and the objective function (1) is non-convex - and the inherent nonuniqueness of FWI. We discuss these in detail below.

Computational Challenges FWI applications often contain billions of unknowns and hundreds of thousands of sources and receivers, leading to a large volume of data - please see [55] for an illustration of this on a realistic marine streamer survey. This high level of complexity dictates that many applications continue to use the acoustic approximation and precludes the use of global optimization techniques.

Each formulation of the wave equation faces its own unique challenges for large-scale problems. For example, the time-domain formulation outlined in Section 2.1 scales linearly with the number of sources, which can be in the hundreds of thousands. For problems with a large number of sources, one strategy to reduce the cost of time-domain methods is to simultaneously model a number of sources through a carefully chosen encoding scheme [13, 54, 65, 74]. Further discussion of this and other discretization issues can be found in [55, 87].

The large-scale nature of FWI dictates that local gradient-based techniques are used to optimize the objective function (1). However, it has been observed for decades that (1) is not convex due to the oscillatory nature of the measured signal and the use of a least-squares data misfit. As the example in Section 2.4 demonstrates, applying gradient-based optimization methods to the non-convex objective function (1) leads to results that are sensitive to the initial guess.

The sensitivity of FWI results to the initial guess remains unresolved, despite decades of research. Three of the main themes of this research are continuation strategies, alternative data misfit functionals, and model expansions. In essence, the goal of all of these approaches is to replace a non-convex problem by one which is closer to convex. Continuation strategies, which remain poorly understood, selectively invert the data according to some criteria [20, 77]. For example, time continuation initially inverts data on a small time window, which is gradually expanded to the entire time window. Alternative data misfit functionals replace the least-squares data misfit in the objective function (1) with a different measure of distance between the measured and computed data [10, 25, 32, 58, 73, 88]. Model expansions convert the original problem into a sequence of convex problems by expanding the parameter space of unknowns [17, 27, 79]. For more information on these and other approaches, please see [55, 87].

Nonuniqueness Nonuniqueness is a pervasive feature of inverse problems. In the context of FWI, specific sources of nonuniqueness include the source-receiver configuration (sources and receivers are typically placed on the earth's surface at a small distance from one another), physical properties of subsurface wave propagation, the range of source signatures available, and noise present in the measured data due to nearby ships, vehicles, and other sources. For further discussion of how these factors can impact FWI, please see [55].

Quantifying and mitigating the nonuniqueness present in FWI remain important challenges. Rigorous uncertainty quantification in FWI is difficult due to the nonlinear dependence of subsurface wave propagation on rock properties, which prevents the use of a simple Gaussian probability distribution, and the immense size of FWI applications, which prohibits naive application of techniques such as Markov chain Monte Carlo designed for non-Gaussian probability distributions [62].

Mitigating the uncertainty in FWI requires additional data. One potential source of such data is alternative seismic acquisition geometries; other sources include well logs, gravity data, and electromagnetic measurements [3, 14, 24, 37].

3 Reservoir History Matching

Once a hydrocarbon reservoir is identified, a production plan is formulated that is designed to maximize hydrocarbon recovery within facility limits. The production plan specifies the location and rates of the wells used to produce hydrocarbons and facilities for processing the extracted fluids. Production occurs in several stages, which we outline below [2, 72].

Prior to production, the reservoir fluids (oil, gas, and water are commonly present) are in equilibrium. This equilibrium represents a balance of various forces, such as gravity and surface tension. Primary production occurs when a production well is drilled and fluids flow out of the reservoir and into the well due to the reservoir's high pressure. The effectiveness of primary production decreases over time as fluids flow out of the reservoir and reservoir pressure decreases. As a result, primary production is typically responsible for producing only a small fraction of a reservoir's hydrocarbons.

In order to produce additional hydrocarbons, reservoir engineers rely on secondary and tertiary production strategies. During secondary production, additional fluids (e.g., gas, and water) are injected in order to both maintain the reservoir's pressure and displace additional hydrocarbons. Further enhancement of production is possible through tertiary production, which relies on chemical and thermal effects to alter the flow properties of the fluids. Examples of tertiary production include the injection of polymers, solvents, and heat into the reservoir.

Due to the complexity of the mechanisms underlying hydrocarbon production, reservoir engineers use computer simulations of reservoir fluid flow to inform production planning. The computer simulation of reservoir fluid flow, known as

reservoir simulation, is based on the numerical solution of partial differential equations [2, 7, 69]. Reservoir simulation requires, as input, a flow model of the reservoir describing key physical parameters (e.g., porosity and permeability). We define these parameters in Section 3.1.

Constructing a reliable flow model is challenging due to the nature of the available data. Seismic data is used to infer mechanical properties of the subsurface rocks, but these properties do not necessarily correlate with flow properties. In addition, the limited resolution of seismic inversion prevents the identification of reservoir features such as thin shale barriers that are critical to fluid flow. Other data sets (e.g., reservoir cores, and well logs) can be used to infer flow properties but are available at only a few locations and therefore provide sparse coverage of the reservoir.

As a result, production data measured at the wells (e.g., pressure, and well rates) may be relied on to validate a reservoir flow model. The process of adjusting the flow model to match production data is known as reservoir history matching. Potential outcomes include the drilling of new wells and the modification of existing wells' operating conditions (e.g., injection and production rates).

Gradient-based algorithms for reservoir history matching, derived using the adjoint-state method, were proposed in the 1970s [22, 23]. Since then, algorithmic and computing advances have enabled reservoir engineers to history match flow models of greater complexity and to begin quantifying the uncertainty inherent in history matching [67, 91]. We discuss these recent developments and remaining challenges in Section 3.5.

3.1 Fluid Flow in Porous Media

A hydrocarbon reservoir is a region of subsurface rock, known as reservoir rock, containing hydrocarbons in the void space between grains. The volume fraction of void space present in the reservoir rock, known as porosity, can range from 5% to 30% for conventional hydrocarbon reservoirs [72]. In order for the production of hydrocarbons to be economic, the void space within the reservoir rock must be sufficiently well connected to permit fluid flow to a well. The connectivity of the rock's void space is quantified by the concept of permeability, which we define below.

Hydrocarbons are generated when deposits of biomass (e.g., algae, and plankton) are buried in sedimentary rock layers. As these sedimentary layers accumulate and biomass is buried further below the surface, the biomass is exposed to temperature and pressure conditions that transform it to hydrocarbons. The conversion of biomass to hydrocarbons occurs in rock known as source rock. For a given reservoir, the source rock may or may not be the same as the reservoir rock. Upwards migration of hydrocarbons from source rock to reservoir rock (e.g., through a fracture network created by faulting) can occur over millions of years due to

buoyancy (since oil and gas are less dense than groundwater). A reservoir is created when further migration of hydrocarbons is impeded due to the presence of an impermeable seal (e.g., shale layer or salt body). Further discussion of hydrocarbon generation and migration can be found in [60, 72, 78].

Fluid flow under the high pressure conditions typically present in the subsurface involves complex multiphase flow behavior and fluid-structure interaction. The equations governing the key processes of such flows at the pore scale are the Navier-Stokes equations coupled to a model of rock deformation. However, such a model cannot be used in practice due to its computational complexity. In addition, our lack of knowledge of the pore-scale rock structure would make this approach impractical. Instead, we rely on Darcy's law, which models average fluid flow over a neighborhood of pores. For single-phase flow, which is the focus of this presentation, Darcy's law describes a linear relationship between the volumetric flow velocity $\mathbf{v}$ and the fluid pressure gradient ∇p expressed as

$$\mathbf{v} = -\frac{\mathbf{k}}{\mu}\nabla p, \tag{32}$$

where μ is the fluid's viscosity and $\mathbf{k}$ is the rock permeability tensor. Darcy's law represents a good approximation for weakly compressible flow and can be derived, under certain hypotheses on the pore-scale structure and fluid properties, using homogenization theory [35]. Further information on Darcy's law and other aspects of flow in porous media can be found in [15, 70].

In addition to Darcy's law, we require that the subsurface fluid satisfy conservation of mass. When modeling secondary recovery - discussed above - conservation of mass can be expressed as

$$\begin{aligned}\phi\frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\mathbf{v}) = &\sum_{\substack{\text{Injection wells}\\ x_{\text{Inj}}}} \rho q(x_{\text{Inj}}, t)\delta\left(x - x_{\text{Inj}}\right)\\ &- \sum_{\substack{\text{Production wells}\\ x_{\text{Prod}}}} \rho q(x_{\text{Prod}}, t)\delta\left(x - x_{\text{Prod}}\right).\end{aligned} \tag{33}$$

The porosity ϕ in (33) represents the volume fraction of the void space within the porous medium and ρ denotes the fluid's density. The source term in (33) models the injection and production of fluid using point sources.

Lastly, we assume that the fluid has compressibility which is independent of pressure and temperature:

$$\frac{\frac{d\rho}{dp}}{\rho} = C_f. \tag{34}$$

Substitution of (34), along with Darcy's law (32), into (33) leads to the nonlinear PDE

$$C_f \phi \frac{\partial p}{\partial t} - C_f \left(\frac{\mathbf{k}}{\mu} \nabla p \right) \cdot \nabla p - \nabla \cdot \left(\frac{\mathbf{k}}{\mu} \nabla p \right) = \sum_{\substack{\text{Injection wells} \\ x_{\text{Inj}}}} q(x_{\text{Inj}}, t) \delta \left(x - x_{\text{Inj}} \right)$$

$$- \sum_{\substack{\text{Production wells} \\ x_{\text{Prod}}}} q(x_{\text{Prod}}, t) \delta \left(x - x_{\text{Prod}} \right), \tag{35}$$

which heuristically simplifies to

$$C_f \phi \frac{\partial p}{\partial t} - \nabla \cdot \left(\frac{\mathbf{k}}{\mu} \nabla p \right) = \sum_{\substack{\text{Injection wells} \\ x_{\text{Inj}}}} q(x_{\text{Inj}}, t) \delta \left(x - x_{\text{Inj}} \right)$$

$$- \sum_{\substack{\text{Production wells} \\ x_{\text{Prod}}}} q(x_{\text{Prod}}, t) \delta \left(x - x_{\text{Prod}} \right), \tag{36}$$

under the assumption

$$C_f \left| p(x_{\text{Inj}}) - p(x_{\text{Prod}}) \right| \ll 1.$$

In order to complete Equation (36), we impose a no-flow boundary condition; due to Darcy's law, this translates to a zero Neumann boundary condition. No-flow boundary conditions are based on the assumption that the reservoir is confined by impermeable rocks that prevented fluid from migrating further upwards.

The above equations are sufficient to describe reservoirs containing a single fluid. In reality, multiple phases (i.e., liquid and vapor) and chemical components (e.g., oil, gas, and water) can be present and are modeled using a so-called black oil or compositional model [7, 69]. These models have four basic ingredients. First, they rely on multiphase extensions of Darcy's law (32). Second, they require mass conservation of each chemical component. Third, they require that all void space in the rock be occupied by the phases present. Fourth, they assume that the reservoir fluids are in thermodynamic equilibrium and use this to partition the components into phases. The resulting models are nonlinear and have greater complexity than the single-phase model presented above. In addition, such models introduce parameters (e.g., those describing the mixing of fluids and the surface tension at an interface separating phases) which are themselves uncertain.

For pedagogical purposes, we consider a second, hypothetical, inverse problem in which production data is augmented by concentration measurements throughout the reservoir of an injected passive tracer. In reality, no such tracer exists. Despite this lack of realism, the inversion of tracer data serves several purposes. Tracer data can be seen as a best-case scenario for the information ever available in reservoir history

matching. In addition, inversion of pressure and tracer data is used to illustrate the potential of the PDE-constrained optimization framework to simultaneously invert multiple data sets described by different physical processes.

Ignoring diffusive and dispersive effects, and assuming that the fluid injected into the reservoir has tracer concentration one, the tracer concentration $c(x,t)$ satisfies

$$\phi \frac{\partial \rho c}{\partial t} + \nabla \cdot (\rho c \mathbf{v}) = \sum_{\substack{\text{Injection wells} \\ x_{\text{Inj}}}} \rho q(x_{\text{Inj}}, t) \delta \left(x - x_{\text{Inj}}\right) - \sum_{\substack{\text{Production wells} \\ x_{\text{Prod}}}} \rho c q(x_{\text{Prod}}, t) \delta \left(x - x_{\text{Prod}}\right). \tag{37}$$

In order to simplify the gradient computation, we ignore density variations in (37), which are assumed to be small. This leads to

$$\phi \frac{\partial c}{\partial t} + \nabla \cdot (c \mathbf{v}) = \sum_{\substack{\text{Injection wells} \\ x_{\text{Inj}}}} q(x_{\text{Inj}}, t) \delta \left(x - x_{\text{Inj}}\right) - \sum_{\substack{\text{Production wells} \\ x_{\text{Prod}}}} c q(x_{\text{Prod}}, t) \delta \left(x - x_{\text{Prod}}\right). \tag{38}$$

The evolution of pressure, velocity, and tracer concentration within our reservoir is governed by the system of Equations (32), (36), (38). After nondimensionalization, this system takes the form

$$\begin{aligned} \mathbf{k}^{-1}\mathbf{v} + \nabla p &= 0, \\ \frac{\tau_p}{\tau_c} \frac{\partial p}{\partial t} + \nabla \cdot \mathbf{v} &= \sum_{\substack{\text{Injection wells} \\ x_{\text{Inj}}}} q(x_{\text{Inj}}, t) \delta \left(x - x_{\text{Inj}}\right) - \sum_{\substack{\text{Production wells} \\ x_{\text{Prod}}}} q(x_{\text{Prod}}, t) \delta \left(x - x_{\text{Prod}}\right), \\ \frac{\partial c}{\partial t} + \nabla \cdot (c \mathbf{v}) &= \sum_{\substack{\text{Injection wells} \\ x_{\text{Inj}}}} q(x_{\text{Inj}}, t) \delta \left(x - x_{\text{Inj}}\right) - \sum_{\substack{\text{Production wells} \\ x_{\text{Prod}}}} c q(x_{\text{Prod}}, t) \delta \left(x - x_{\text{Prod}}\right). \end{aligned} \tag{39}$$

Here, τ_p denotes an approximate timescale on which the pressure equation (36) equilibrates, τ_c denotes an approximate timescale on which tracer is transported according to (38), and all variables have been nondimensionalized.

In order to solve the system (39) numerically, we need to accurately compute both pressure and velocity. We accomplish this by discretizing the pressure-velocity equations, which are parabolic, using the mixed finite-element method [19]. The distinguishing feature of the mixed finite-element method is that it solves simultaneously for pressure and velocity. As a result, it is known to have greater accuracy than solving first for pressure and then differentiating to compute velocity - numerical differentiation of pressure leads to a one order loss of accuracy and multiplication by a possibly discontinuous permeability field contributes additional error. However, the improved accuracy of the mixed finite-element method comes at a cost: the resulting linear system is indefinite and requires preconditioning or hybridization. In addition, mixed finite-element formulations for multiphase flow have proven challenging. As a result, many large-scale industrial reservoir simulators still use finite-volume formulations [2].

The tracer transport equation is hyperbolic and is solved using an explicit-in-time upwind finite-volume discretization [56]. First-order upwinding is commonly used in reservoir simulation to solve transport equations [2]; higher-order discretizations have also been considered (e.g., ENO and the discontinuous-Galerkin method) [7, 45].

3.2 Objective Function for Reservoir History Matching

We formulate the objective function for reservoir history matching using the regularized least-squares framework introduced in Section 2. Under the assumption that the permeability tensor is isotropic

$$\mathbf{k}(x) = k(x) \cdot \mathbf{I},$$

reservoir history matching can be expressed as the following minimization problem:

$$\min_{k(x)>0} \mathcal{J}(k) := \frac{1}{2} \sum_{\text{Wells } x_w} \int_0^T |p(k; x_w, t) - \bar{p}(x_w, t)|^2 \, dt + R(k). \tag{40}$$

In the above equation, $p(k; x_w, t)$ represents reservoir pressure at the well location x_w at time t (39), $\bar{p}$ represents the measured pressure data at the wells, and $R(k)$ is a regularization term, such as Tikhonov or total variation, which is used to stabilize the inversion. The positivity constraint on the permeability comes from physical considerations and can be satisfied using, for example, an exponential transform. Ensuring that minimizers of Equation (40) are geologically realistic remains a challenge and requires either a well-designed regularization function or an effective reservoir parameterization [91]. We discuss this issue further in Section 3.5.

Incorporating tracer measurements throughout the reservoir leads to the following least-squares problem

$$\min_{k(x)>0} \mathcal{J}(k) := \frac{1}{2}\int_0^T \int_\Omega |c(k;x,t) - \bar{c}(x,t)|^2 \, dx$$
$$+\frac{\beta}{2} \sum_{\text{Wells } x_w} \int_0^T |p(k;x_w,t) - \bar{p}(x_w,t)|^2 + R(k), \tag{41}$$

where $c(k;x,t)$ is the tracer concentration at location x at time t, $\bar{c}(x,t)$ is the measured tracer concentration there, and β is used to adjust the cost of the pressure misfit relative to the tracer misfit.

Several changes to (41) are in order for performing industrial history matching; for the sake of completeness, we touch on them briefly here. First, tracer data would not be available. Second, reliable pressure data may not be available and well rates may only be available on an aggregate basis (e.g., sub-sea floor installations combine production from wells). Third, due to the limited information content of such data, as outlined in Section 3.5, history matching has traditionally focused on identifying a low-dimensional flow model that captures what are believed to be key features of the reservoir (e.g., channels, faults, and layers).

3.3 Gradient Computation for Darcy Flow

The general strategy for computing the gradient of (40) and (41) is the same as in Section 2.2. The key difference is in the determination of the appropriate adjoint problems. In the following, we neglect the regularization term and focus on the objective function with tracer data (41), which implicitly includes the objective function (40).

We begin by defining the pressure-velocity forward and adjoint operators. The pressure-velocity forward operator M can be expressed as

$$M\begin{pmatrix} f \\ \mathbf{h} \end{pmatrix} = \begin{pmatrix} p \\ \mathbf{v} \end{pmatrix}(f, \mathbf{h}) \tag{42}$$

where $\begin{pmatrix} p \\ \mathbf{v} \end{pmatrix}(f, \mathbf{h})$ solves the pressure-velocity system

$$\begin{aligned} \frac{\tau_p}{\tau_c}\frac{\partial p}{\partial t} + \nabla \cdot \mathbf{v} &= f \\ k^{-1}\mathbf{v} + \nabla p &= \mathbf{h} \end{aligned} \tag{43}$$

with source term $\begin{pmatrix} f \\ \mathbf{h} \end{pmatrix}$ subject to the initial and boundary conditions

$$\begin{aligned} p(x,0) &= 0, \\ \mathbf{v}\cdot n|_{\partial\Omega} &= 0. \end{aligned} \tag{44}$$

The adjoint operator M^* is given by

$$M^*\begin{pmatrix} g \\ \mathbf{j} \end{pmatrix} = \begin{pmatrix} m \\ \mathbf{u} \end{pmatrix}(g,\mathbf{j}) \tag{45}$$

where $\begin{pmatrix} m \\ \mathbf{u} \end{pmatrix}(g,\mathbf{j})$ solves the adjoint system

$$\begin{aligned} -\frac{\tau_p}{\tau_c}\frac{\partial m}{\partial t} - \nabla\cdot\mathbf{u} &= g \\ k^{-1}\mathbf{u} - \nabla m &= \mathbf{j} \end{aligned} \tag{46}$$

with source term $\begin{pmatrix} g \\ \mathbf{j} \end{pmatrix}$ subject to the end-time and boundary conditions

$$\begin{aligned} m(x,T) &= 0, \\ \mathbf{u}\cdot n|_{\partial\Omega} &= 0. \end{aligned} \tag{47}$$

Integration by parts shows that M and M^* satisfy the adjoint identity

$$\begin{aligned} &\left(M\begin{pmatrix} f \\ \mathbf{h} \end{pmatrix}, \begin{pmatrix} g \\ \mathbf{j} \end{pmatrix}\right)_{L^2(\mathbb{R}^n\times[0,T])} = \left(\begin{pmatrix} f \\ \mathbf{h} \end{pmatrix}, M^*\begin{pmatrix} g \\ \mathbf{j} \end{pmatrix}\right)_{L^2(\mathbb{R}^n\times[0,T])} \\ &\quad \forall f, g, \mathbf{h}, \mathbf{j} \in L^2\left(\mathbb{R}^n \times [0,T]\right). \end{aligned} \tag{48}$$

Next, we define the forward operator L describing tracer propagation:

$$Ls = c(s) \tag{49}$$

where $c(s)$ solves the PDE

$$\frac{\partial c}{\partial t} + \nabla\cdot(\mathbf{v}c) + \sum_{\substack{\text{Production wells} \\ x_{\text{Prod}}}} cq(x_{\text{Prod}},t)\delta\left(x - x_{\text{Prod}}\right) = s \tag{50}$$

subject to the initial condition $c(x,0) = 0$. The adjoint operator, L^*, is given by

$$L^*r = b(r) \tag{51}$$

where $b(r)$ solves the adjoint PDE

$$-\frac{\partial b}{\partial t} - \mathbf{v} \cdot \nabla b + \sum_{\substack{\text{Production wells} \\ x_{\text{Prod}}}} bq(x_{\text{Prod}}, t)\delta\left(x - x_{\text{Prod}}\right) = r \tag{52}$$

subject to the end-time condition $b(x, T) = 0$. Integration by parts demonstrates that L and L^* satisfy the adjoint identity

$$(Ls, r)_{L^2(\mathbb{R}^n \times [0,T])} = (s, L^* r)_{L^2(\mathbb{R}^n \times [0,T])} \; \forall r, s \in L^2\left(\mathbb{R}^n \times [0, T]\right). \tag{53}$$

Computing the gradient $\nabla \mathcal{J}$ of (41) follows the same structure as in Section 2.2. First, the forward problems (42), (49) are linearized with respect to the permeability k. Next, these linearizations are used, along with the adjoint identities (48), (53), to identify the gradient of the objective function. The final result is

$$\nabla \mathcal{J} = \int_0^T \frac{1}{k^2} \mathbf{v} \cdot \mathbf{u} \, dt, \tag{54}$$

where $\begin{pmatrix} q \\ \mathbf{u} \end{pmatrix}$ solves the adjoint problem

$$\begin{pmatrix} q \\ \mathbf{u} \end{pmatrix} = M^* \begin{pmatrix} \beta \sum\limits_{\text{Wells } x_w} \delta(x - x_w)(p - \bar{p}) \\ c \nabla b \end{pmatrix} \tag{55}$$

and b solves the adjoint problem

$$b = L^*(c - \bar{c}). \tag{56}$$

The adjoint problem (56) corresponds to injection of the tracer data misfit as a source. The pressure equation portion of the adjoint problem (55) corresponds to injection of pressure misfit at the wells; however, the source term for the velocity equation does not have a clear physical interpretation.

3.4 A Synthetic Example

We consider a reservoir history matching example in which the reservoir contains two intersecting channels of different permeabilities overlaid on a constant permeability background. Reservoir permeability, normalized by the background value of 10 mD, and well locations are illustrated in Figure 6. All data is measured until the injected tracer reaches the production well, although the pressure data equilibrates on a much shorter timescale.

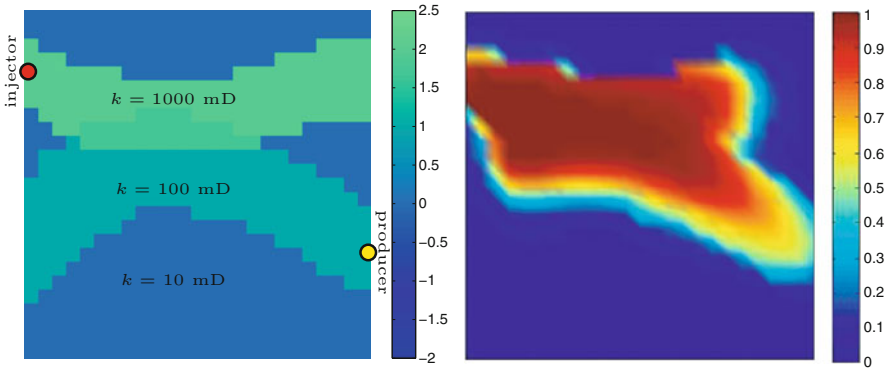

Fig. 6 Reservoir containing two intersecting channels of different permeabilities. Left: logarithm of the true reservoir permeability model, normalized by the background value of 10 mD, along with well locations. Right: tracer concentration at breakthrough time $t \approx T/2$. Notice that the tracer is concentrated in the two channels, both of which have higher permeability than the background

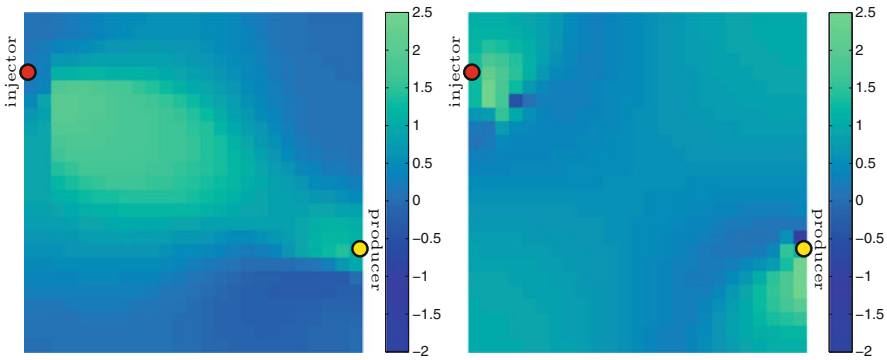

Fig. 7 Permeability parameter estimation using only pressure data. Left: estimate using an initial guess $k = 10$ mD. Right: estimate using an initial guess $k = 100$ mD. Information recovery is very limited due to the sparsity of the available data and the diffusive nature of the pressure equation

Presenting this example has two goals. The first is to highlight the nonuniqueness of reservoir history matching. The second is to demonstrate the uplift and limitations of incorporating tracer data into reservoir history matching.

We begin by considering the determination of permeability using only pressure data at the wells (i.e., without using any tracer data), as given by (40). We consider two different homogeneous initial models: $k = 10$ mD and $k = 100$ mD. The results are displayed in Figure 7.

Due to the sparsity of the data and the diffusive nature of the pressure equation, the resolution of the results is quite poor. This illustrates the nonuniqueness of reservoir history matching: despite matching the well data to a prescribed tolerance, the resulting permeability models differ substantially from the true, channelized,

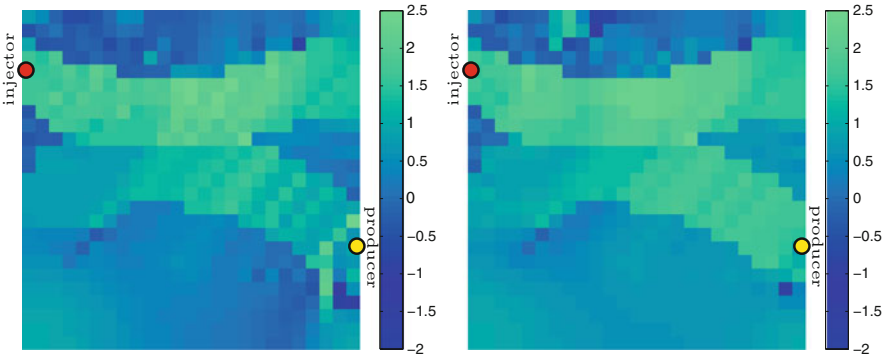

Fig. 8 Permeability parameter estimation using pressure and tracer data measured until breakthrough. Left: estimate using an initial guess $k = 10\,\text{mD}$. Right: estimate using an initial guess $k = 100\,\text{mD}$. Both estimates represent a significant improvement in comparison with those found in Figure 7

model. Improving these results requires either inverting for a reduced set of parameters (e.g., assuming a known channel geometry and inverting for the constant permeability within each region) [4] or incorporating additional data.

Next, we augment the pressure data with tracer data and consider the PDE-constrained optimization problem (41). We assume that the tracer data is available throughout the reservoir while the pressure data is available only at the wells. We consider the initial models used previously. The results are displayed in Figure 8 and demonstrate that tracer data provides an improvement in the resolution of reservoir permeability. For example, when an initial guess of $k = 10\,\text{mD}$ is used, the resulting inversion provides a good reconstruction of the reservoir's channels. This occurs because the initial guess agrees with the background permeability while the pressure and tracer data largely determine the channels' permeability. In contrast, starting with the homogeneous model $k = 100\,\text{mD}$ leads to a good estimate of the channels' permeability but a poor estimate of the background, which is 10 mD. The inability to determine the background permeability in this case is partially due to the fact that, at breakthrough, very little tracer has entered the background. This can be seen in Figure 6.

We conclude this section with a one-dimensional analytic example designed to provide intuition behind our inability to infer permeability values using pressure data only. It is important to keep in mind that this example does not capture the full complexity of the 2D numerical results presented above.

As indicated in Figure 9, we assume that an incompressible fluid flows through a one-dimensional reservoir of varying permeability $k(x)$. In addition, we assume that the inflow pressure p_0 and the outflow velocity q are prescribed. It follows, due to incompressibility, that $v(x) \equiv q$.

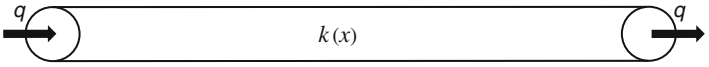

Fig. 9 One-dimensional incompressible flow through a reservoir. Due to the incompressibility of the fluid, the velocity is constant, $v(x) = q$, and the pressure obeys Darcy's law $q = -\frac{k(x)}{\mu}\frac{\partial p}{\partial x}$

The pressure-velocity system in this case can be expressed as

$$\begin{aligned} v &= -\frac{k}{\mu}\frac{\partial p}{\partial x} \text{ for } x \in (0,1), \\ \frac{\partial v}{\partial x} &= 0 \text{ for } x \in (0,1), \\ p &= p_0 \text{ at } x = 0, \\ v &= q \text{ at } x = 1. \end{aligned} \tag{57}$$

In this simplified setting, the reservoir history matching problem is the determination of permeability from measurement of the outflow pressure. It turns out that this problem is highly nonunique: we will show that any two permeability distributions $k(x)$ and $\tilde{k}(x)$ yield the same outflow pressure whenever

$$\tilde{k}(x) = \frac{k(x)}{1 + \delta k(x)} \tag{58}$$

where δk is a perturbation satisfying

$$\int_0^1 \frac{\delta k(x)}{k(x)}\, dx = 0. \tag{59}$$

In order to establish this claim, denote by $p(x)$ the pressure field corresponding to $k(x)$ and $\tilde{p}(x)$ the field arising from $\tilde{k}(x)$. The result follows from

$$\begin{aligned} \tilde{p}(1) &= p_0 + \int_0^1 \frac{\partial \tilde{p}}{\partial x} dx = p_0 - \int_0^1 \frac{q\mu}{\tilde{k}(x)} dx = p_0 - \int_0^1 \frac{q\mu(1+\delta k(x))}{k(x)} dx \\ &= p_0 - \int_0^1 \frac{q\mu}{k(x)} dx = p(1). \end{aligned} \tag{60}$$

A similar result can be found in the reservoir simulation textbook [11].

3.5 Challenges in Reservoir History Matching

Reservoir history matching faces significant challenges: the results are inherently nonunique and the complexity of the PDE models used complicates optimization and uncertainty quantification. We discuss each of these in detail below.

Nonuniqueness The results of Section 3.4 give the reader a sense of the nonuniqueness of reservoir history matching. Four distinct factors contribute to this nonuniqueness. First, the diffusive nature of the pressure equation (36) leads to a smearing of information over time. Second, there is a stark discrepancy between the small number of available well measurements and the potentially large number of unknown grid permeability values. Third, permeability can attain a wide range of values - near zero at flow barriers to extremely large at open fractures. Fourth, many of the parameters controlling fluid flow that are assumed to be known (e.g., fault locations, the initial distribution of fluids, and the mixing behavior of different fluids) are, in reality, uncertain. Incorporating additional data (e.g., geologic analogs and seismic data) [21, 36, 90] may reduce the degree of nonuniqueness but will likely not eliminate it.

The development of algorithms that quantify the uncertainty in reservoir performance represents a significant challenge. A number of algorithms have been proposed in recent years; these include Markov chain Monte Carlo, randomized maximum likelihood, and the ensemble Kalman filter [67, 68]. Due to the cost and nonlinearity of solving the forward model, the limited data available, and the fact that many parameters in reservoir flow models are highly uncertain, successful uncertainty quantification of reservoir performance may hinge on the identification of low-dimensional representations of reservoir features which ensure geologic realism and capture the most important aspects of fluid flow [91].

Complexity of Multiphase Reservoir Fluid Flow PDE models of reservoir fluid flow possess a high degree of complexity due to the nonlinear nature of multiphase flow and the heterogeneity of reservoir rock properties. As a result of this complexity, numerically solving such models is costly and computing the gradients necessary for optimization and uncertainty quantification is challenging.

In recent years, reservoir simulation has seen significant speedup due to advances in high-performance computing, novel computer architectures, and linear solver preconditioning. However, the cost of reservoir simulation continues to be a bottleneck when attempting to quantify subsurface geologic uncertainty. One way to reduce this cost is through the use of a proxy simulator that provides inexact but sufficiently accurate estimation of the flow data relatively quickly [9, 49].

Difficulty computing gradients for realistic reservoir history matching problems motivated the use of derivative-free optimization algorithms. These algorithms include the ensemble Kalman filter, the ensemble Kalman smoother, evolutionary optimization strategies, and particle swarm optimization [33, 67, 75]. In light of these developments, it is unclear whether the value of the gradient for optimization and uncertainty quantification outweighs its cost.

4 Summary and Outlook

This paper provided an overview of the oil and gas supply chain and introduced two important applications of PDE-constrained optimization: FWI and reservoir history matching. A simple model problem for each application was solved using a

common least-squares optimization framework based on the adjoint-state method. The quality of the results was dependent on the type of PDE (i.e., hyperbolic or parabolic), the experimental setup (i.e., the placement of sources and receivers and the source signature), and the initial guess.

Looking to the future, opportunities abound for the application of PDE-constrained optimization in the oil and gas industry. These include optimal source and receiver configuration in FWI, higher resolution geophysical imaging through joint inversion of seismic, electromagnetic, and gravity data sets, optimization of well placement and control, flaw detection in pipelines, and chemical reactor design and control. In particular, algorithmic advances in FWI and reservoir history matching (leveraged, perhaps, by developments in related fields such as data assimilation, medical imaging, and uncertainty quantification) will play a crucial role as the industry is confronted by ever-increasing subsurface complexity and new environments characterized by high risks and costs. The challenges, and hence opportunities, faced in discovering new hydrocarbon reservoirs and optimizing production from existing fields have never been greater.

Acknowledgements The authors thank ExxonMobil for permission to publish this work. The authors also thank Martin Lacasse, Laurent White, Rohan Panchadhara, Anatoly Baumstein, Tom Dickens, Dave Stern, Klaus Wiegand, and Xiao-Hui Wu for their feedback on an earlier version of this paper.

References

1. Outlook for energy: A view to 2040. http://corporate.exxonmobil.com/en/energy/energy-outlook, 2016.
2. Jorg E. Aarnes, Tore Gimse, and Knut-Andreas Lie. An introduction to the numerics of flow in porous media using Matlab. In Geir Hasle, Knut-Andreas Lie, and Ewald Quak, editors, *Geometric Modelling, Numerical Simulation, and Optimization*, pages 265–306. Springer, 2007.
3. A Abubakar, G Gao, T M Habashy, and J Liu. Joint inversion approaches for geophysical electromagnetic and elastic full-waveform data. *Inverse Problems*, 28(5):055016, 2012.
4. Chinedu C. Agbalaka, Dave Stern, and Dean S. Oliver. Two-stage ensemble-based history matching with multiple modes in the objective function. *Society of Petroleum Engineers Journal*, 55:28–43, 2013.
5. Volkan Akcelik, Huseyin Denli, Alex Kanevsky, Kinesh K. Patel, Laurent White, and Martin-Daniel Lacasse. Multiparameter material model and source signature full waveform inversion. *SEG Technical Program Expanded Abstracts*, pages 2406–2410, 2012.
6. Keiiti Aki and Paul G. Richards. *Quantitative Seismology*. University Science Books, 2002.
7. Myron Bartlett Allen III, Graca Alda Behie, and John Arthur Trangenstein. *Multiphase Flow in Porous Media*, volume 34 of *Lecture Notes in Engineering*. Springer, 1988.
8. John E. Anderson, Lijian Tan, and Don Wang. Time-reversal checkpointing methods for RTM and FWI. *Geophysics*, 77(4):S93–S103, 2012.
9. A. C. Antoulas, D. C. Sorensen, and S. Gugercin. A survey of model reduction methods for large-scale systems. *Contemporary Mathematics*, 280:193–219, 2001.
10. Aleksandr Aravkin, Tristan van Leeuwen, and Felix Herrmann. Robust full-waveform inversion using the student's t-distribution. pages 2669–2673, 2011.

11. Khalid Aziz and Antonin Settari. *Petroleum Reservoir Simulation*. 2002.
12. W. Bangerth, H. Klie, M. F. Wheeler, P. L. Stoffa, and M. K. Sen. On optimization algorithms for the reservoir oil well placement problem. *Comp. Geosc.*, 10(3):303–319, 2006.
13. Rishi Bansal, Jerry Krebs, Partha Routh, Sunwoong Lee, John Anderson, Anatoly Baumstein, Anoop Mullur, Spyros Lazaratos, Ivan Chikichev, and David McAdow. Simultaneous-source full-wavefield inversion. *The Leading Edge*, 32(9):1100–1108, 2013.
14. Anatoly Baumstein. Borehole-constrained multi-parameter full waveform inversion. *77th EAGE Conference and Exhibition*, 2015.
15. Jacob Bear. *Dynamics of Fluids in Porous Media*. Dover, 1972.
16. Lorenz T. Biegler. *Nonlinear Programming: Concepts, Algorithms, and Applications to Chemical Processes*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, Philadelphia, 2010.
17. Biondo Biondi and Ali Almomin. Simultaneous inversion of full data bandwidth by tomographic full-waveform inversion. *Geophysics*, 79(3):WA129–WA140, 2014.
18. Knut Bjorlykke, editor. *Petroleum Geoscience: From Sedimentary Environments to Rock Physics*. Springer, 2015.
19. Franco Brezzi and Michael Fortin. *Mixed and hybrid finite-element methods*, volume 15 of *Springer Series in Computational Mathematics*. Springer, 1991.
20. Carey Bunks, Fatimetou M. Saleck, S. Zaleski, and G. Chavent. Multiscale seismic waveform inversion. *Geophysics*, 60(5):1457–1473, 1995.
21. Jef Caers. Efficient gradual deformation using a streamline-based proxy method. *Journal of Petroleum Science and Engineering*, 39:57–83, 2003.
22. G. Chavent, M. Dupuy, and P. Lemmonier. History matching by use of optimal theory. *SPE Journal*, 15(1):74–86, 1975.
23. W.H. Chen, G.R. Gavalas, J.H. Seinfeld, and M.L. Wasserman. A new algorithm for automatic history matching. *SPE Journal*, 14(6):593–608, 1974.
24. Yangkang Chen, Hanming Chen, Kui Xiang, and Xiaohong Chen. Geological structure guided well log interpolation for high-fidelity full waveform inversion. *Geophysical Journal International*, 2016.
25. Benxin Chi, Liangguo Dong, and Yuzhu Liu. Full waveform inversion method using envelope objective function without low frequency data. *Journal of Applied Geophysics*, 109:36–46, 2014.
26. Robert Clayton and Björn Engquist. Absorbing boundary conditions for acoustic and elastic wave equations. *Bulletin of the Seismological Society of America*, 67(6):1529–1540, 1977.
27. Augustin Cosse, Stephen D. Shank, and Laurent Demanet. A short note on rank-2 relaxation for waveform inversion. *SEG Technical Program Expanded Abstracts*, pages 1344–1350, 2015.
28. Debanjan Datta and Mrinal K. Sen. Estimating a starting model for full-waveform inversion using a global optimization method. *Geophysics*, 81(4):R211–R223, 2016.
29. Laurent Demanet. Waves and imaging: Class notes - 18.367. http://math.mit.edu/icg/resources/notes367.pdf, September 2017.
30. Huseyin Denli, Volkan Akcelik, Alex Kanevsky, Dimitar Trenev, Laurent White, and Martin-Daniel Lacasse. Full-wavefield inversion for acoustic wave velocity and attenuation. *SEG Technical Program Expanded Abstracts*, pages 980–985, 2013.
31. Morgan Downey. *Oil 101*. Wooden Table Press LLC, USA, 2009.
32. Bjorn Engquist and Brittany D. Froese. Application of the Wasserstein metric to seismic signals. *Communications in Mathematical Sciences*, 12(5):979–988, 2014.
33. Geir Evensen. *Data Assimilation*. Springer, 2009.
34. ExxonMobil. *Technology: Exploration and production*, 2016. http://www.aboutnaturalgas.com/en/technology/exploration-and-production.
35. Andrew Fowler. *Mathematical Geoscience*. Interdisciplinary Applied Mathematics. Springer, 2011.
36. J. H. Seinfeld G. R. Gavalas, P. C. Shah. Reservoir history matching by Bayesian estimation. *Society of Petroleum Engineers Journal*, 16(6), 1976.

37. Michal Holtzman Gazit and Eldad Haber. Joint inversion through a level set formulation. *ASEG Extended Abstracts: 22nd Geophysical Conference*, pages 1–3, 2013.
38. Andreas Griewank and Andrea Walther. Algorithm 799: Revolve: An implementation of checkpointing for the reverse or adjoint mode of computational differentiation. *ACM Trans. Math. Softw.*, 26(1):19–45, March 2000.
39. Max Gunzburger. *Perspectives in Flow Control and Optimization*. Society for Industrial and Applied Mathematics, 2003.
40. Eldad Haber. *Computational methods in geophysical electromagnetics*. Society for Industrial and Applied Mathematics, Philadelphia, 2014.
41. Matthias Heinkenschloss. PDE Constrained Optimization. https://www.siam.org/meetings/op08/Heinkenschloss.pdf, May 2008. Presentation at the SIAM Conference on Optimization.
42. Matthias Heinkenschloss. Numerical solution of implicitly constrained optimization problems. Technical Report TR08-05, Department of Computational and Applied Mathematics, Rice University, 2013.
43. Jan S. Hesthaven and Tim Warburton. *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications*. Springer, Texts in Applied Mathematics, 2007.
44. Michael Hinze, Rene Pinnau, Michael Ulbrich, and Stefan Ulbrich. *Optimization with PDE Constraints*. Springer, 2009.
45. Hussein Hoteit and Abbas Firoozabadi. Numerical modeling of two-phase flow in heterogeneous permeable media with different capillary pressures. *Advances in Water Resources*, 31:56–73, 2008.
46. Peter Howell, Gregory Kozyreff, and John Ockendon. *Applied Solid Mechanics*. Cambridge University Press, first edition, 2009.
47. Daniel J. Jacob, Alexander J. Turner, Joannes D. Maasakkers, Jianxiong Sheng, Kang Sun, Xiong Liu, Kelly Chance, Ilse Aben, Jason McKeever, and Christian Frankenberg. Satellite observations of atmospheric methane and their value for quantifying methane emissions. *Atmos. Chem. Phys.*, 16:14371–14396, 2016.
48. Jari Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*. Applied Mathematical Sciences. Springer, 2005.
49. Małgorzata P. Kaleta, Remus G. Hanea, Arnold W. Heemink, and Jan-Dirk Jansen. Model-reduced gradient-based history matching. *Computational Geosciences*, 15(1):135–153, 2011.
50. Martin Käser, Verena Hermann, and Josep de la Puente. Quantitative accuracy analysis of the discontinuous Galerkin method for seismic wave propagation. *Geophysical Journal International*, 173(3):990–999, 2008.
51. C.T. Kelley. *Iterative Methods for Optimization*. Society for Industrial and Applied Mathematics, 1999.
52. Dimitri Komatitsch and Jeroen Tromp. The spectral element method for three-dimensional seismic wave propagation. *SEG Technical Program Expanded Abstracts*, pages 2197–2200, 2000.
53. Dimitri Komatitsch and Jeroen Tromp. A perfectly matched layer absorbing boundary condition for the second-order seismic wave equation. *Geophysical Journal International*, 154(1):146–153, 2003.
54. Jerome R. Krebs, John E. Anderson, David Hinkley, Ramesh Neelamani, Sunwoong Lee, Anatoly Baumstein, and Martin-Daniel Lacasse. Fast full-wavefield seismic inversion using encoded sources. *Geophysics*, 74(6):WCC177–WCC188, 2009.
55. Martin-Daniel Lacasse, Laurent White, Huseyin Denli, and Lingyun Qiu. Full-wavefield inversion: An extreme-scale PDE-constrained optimization problem. In Harbir Antil, Drew Kouri, Martin-Daniel Lacasse, and Denis Ridzal, editors, *Frontiers in PDE-Constrained Optimization*. Springer, 2018.
56. Randall J. Leveque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2002.
57. Randall J. Leveque. *Finite Difference Methods for Ordinary and Partial Differential Equations*. Society for Industrial and Applied Mathematics, 2007.

58. Simon Luo and Paul Sava. A deconvolution-based objective function for wave-equation inversion. *SEG Technical Program Expanded Abstracts*, pages 2788–2792, 2011.
59. Guri I. Marchuk. *Adjoint Equations and Analysis of Complex Systems*, volume 295 of *Mathematics and Its Applications*. Springer Science + Business Media, B.V., 1995.
60. Stephen Marshak. *Earth*. W.W. Norton, 2012.
61. Gary S. Martin, Robert Wiley, and Kurt J. Marfurt. Marmousi2: An elastic upgrade for Marmousi. *The Leading Edge*, 25(2):156–166, 2006.
62. James Martin, Lucas C. Wilcox, Carsten Burstedde, and Omar Ghattas. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM J. Sci Comput.*, 34:146–1487, 2012.
63. L. Métivier, R. Brossier, J. Virieux, and S. Operto. Full waveform inversion and the truncated Newton method. *SIAM Journal on Scientific Computing*, 35(2):B401–B437, 2013.
64. Peter Moczo, Johan O.A. Robertsson, and Leo Eisner. The finite-difference time-domain method for modeling of seismic wave propagation. In Valerie Maupin Ru-Shan Wu and Renata Dmowska, editors, *Advances in Wave Propagation in Heterogenous Earth*, volume 48 of *Advances in Geophysics*, pages 421–516. Elsevier, 2007.
65. Ramesh (Neelsh) Neelamani, Christine E. Krohn, Jerry R. Krebs, Justin K. Romberg, Max Deffenbaugh, and John E. Anderson. Efficient seismic forward modeling using simultaneous random sources and sparsity. *Geophysics*, 75(6):WB15–WB27, 2010.
66. Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, 2006.
67. Dean S. Oliver and Yan Chen. Recent progress on reservoir history matching: a review. *Computational Geosciences*, 15(1):185–221, 2011.
68. Dean S. Oliver, Albert C. Reynolds, and Ning Liu. *Inverse Theory for Petroleum Reservoir Characterization and History Matching*. Cambridge University Press, 2008.
69. Donald W. Peaceman. *Fundamentals of numerical reservoir simulation*. Elsevier, 1977.
70. George F. Pinder and William G. Gray. *Essentials of Multiphase Flow and Transport in Porous Media*. John Wiley and Sons, 2 2008.
71. S. J. Qina and T.A. Badgwell. A survey of industrial model predictive control technology. *Cont. Engr. Prac.*, 11:733–764, 2003.
72. Martin S. Raymond and William L. Leffler. *Oil and Gas Production in Nontechnical Language*. PennWell Corporation, 2006.
73. P. S. Routh, J. R. Krebs, S. Lazaratos, A. I. Baumstein, I. Chikichev, S. Lee, N. Downey, D. Hinkley, and J. E. Anderson. Full-wavefield inversion of marine streamer data with the encoded simultaneous source method. *73rd EAGE Conference and Exhibition*, 2011.
74. Partha Routh, Jerry Krebs, Spyros Lazaratos, Anatoly Baumstein, Sunwoong Lee, Young Ho Cha, Ivan Chikichev, Nathan Downey, Dave Hinkley, and John Anderson. Encoded simultaneous source full-wavefield inversion for spectrally shaped marine streamer data. *SEG Technical Program Expanded Abstracts*, pages 2433–2438, 2011.
75. A. Seiler, G. Evensen, J.-A. Skjervheim, J. Hove, and J. G. Vabø. *Using the Ensemble Kalman Filter for History Matching and Uncertainty Quantification of Complex Reservoir Models*, pages 247–271. John Wiley & Sons, Ltd, 2010.
76. Peter M Shearer. *Introduction to Seismology*. Cambridge University Press, second edition, 2009.
77. Laurent Sirgue and R. Gerhard Pratt. Efficient waveform inversion and imaging: a strategy for selecting temporal frequencies. *Geophysics*, 69:231–248, 2004.
78. Vaclav Smil. *Oil*. Oneworld Publications, London, 2013.
79. William W. Symes. Migration velocity analysis and waveform inversion. *Geophysical Prospecting*, 56(6):765–790, 2008.
80. William W. Symes. The seismic reflection inverse problem. *Inverse Problems*, 25(12), 2009.
81. William W. Symes. CAAM 436 notes: Partial differential equations of mathematical physics, 2012.

82. William W. Symes, Igor S. Terentyev, and Tetyana W. Vdovina. Gridding requirements for accurate finite difference simulation. *SEG Technical Program Expanded Abstracts*, pages 2077–2081, 2008.
83. Sirui Tan and Lianjie Huang. An efficient finite-difference method with high-order accuracy in both time and space domains for modelling scalar-wave propagation. *Geophysical Journal International*, 2014.
84. Leon Thomsen. *Understanding Seismic Anisotropy in Exploration and Exploitation*. Society of Exploration Geophysicists, 2014.
85. Ilya Tsvankin, James Gaiser, Vladimir Grechka, Mirko van der Baan, and Leon Thomsen. Seismic anisotropy in exploration and reservoir characterization: An overview. *Geophysics*, 75(5):A15–A29, 2010.
86. Denes Vigh, E. William Starr, and Pavan Elapavuluri. Acoustic waveform inversion vs. elastic data. *SEG Technical Program Expanded Abstracts*, pages 2298–2301, 2009.
87. J. Virieux and S. Operto. An overview of full-waveform inversion in exploration geophysics. *Geophysics*, 74(6):WCC1–WCC26, 2009.
88. Michael Warner and Lluis Guasch. Adaptive waveform inversion: Theory. *Geophysics*, 81(6):R429–R445, 2016.
89. J. H. Williamson. Low-Storage Runge-Kutta Schemes. *Journal of Computational Physics*, 35:48–56, March 1980.
90. Rachel Wood and Andrew Curtis. Geological prior information and its application to geoscientific problems. *Special Publications, London Geological Society*, 239:1–14, 2004.
91. Xiao-Hui Wu, Linfeng Bi, and Subhash Kalla. Effective parametrization for reliable reservoir performance predictions. *Int. J. Uncert. Quant.*, 2(3):259–278, 2012.

Full-Wavefield Inversion: An Extreme-Scale PDE-Constrained Optimization Problem

Martin-D. Lacasse, Laurent White, Huseyin Denli, and Lingyun Qiu

Abstract Full-wavefield inversion is a geophysical method aimed at estimating the mechanical properties of the earth subsurface. This parameter estimation problem is solved iteratively using optimization techniques aimed at minimizing some measure of misfit between computer-simulated data and real data measured in a seismic survey. This PDE-constrained optimization problem poses many challenges due to the extreme size of the surveys considered. Practical issues related to the physical fidelity and numerical accuracy of the forward problem are presented. Also, issues related to the inverse problem such as the limitations of the optimization methods employed, and the many heuristic strategies used to obtain a solution are discussed. The goal of this paper is to demonstrate some of the progress achieved over the last decades while highlighting the many areas where further investigation could bring this method to full technical maturity. It is our hope that this paper, together with other contributions in this book, will motivate a new generation of researchers to contribute to this broad and challenging research area.

1 Introduction

Reflection seismology has been instrumental to oil and gas prospecting since its first application in Oklahoma at the beginning of the last century [9]. This technology relies on triggering controlled artificial seismic sources at the surface of the earth and listening for elastic waves coming back to the surface. The acquisition methods used in reflection seismology are often referred to as multichannel seismic methods referring to the large number of simultaneous channels used for recording elastic

M.-D. Lacasse (✉) · L. White · H. Denli
Corporate Strategic Research, ExxonMobil Research and Engineering Company, Annandale, NJ 08801, USA
e-mail: martin.lacasse@exxonmobil.com

L. Qiu
Petroleum Geo-Services, Houston, TX, USA

H. Antil et al. (eds.), *Frontiers in PDE-Constrained Optimization*, The IMA Volumes in Mathematics and its Applications 163,
https://doi.org/10.1007/978-1-4939-8636-1_6

waves. Each channel is associated with one of the seismic receivers which are typically positioned at dense regular intervals. The reflections and refractions of the waves are caused by the presence of rocks of different types, which, as all materials, have characteristic sound transmission properties. The contrast in these properties causes waves to reflect and refract. Relating the reflections and refractions back to the material properties of the subsurface requires a deep understanding of the propagation of elastic waves in the earth, and the geological processes that created the different sedimentary rock layers.

The ultimate goal of seismic exploration is to provide valuable input to the identification and characterization of oil and gas reservoirs. These reservoirs are quantified in terms of properties such as fluid saturation, rock porosity, and permeability, properties that are only indirectly related to sound wave propagation. This indirect relationship is based on geological interpretations and approximate rock-physics correlations derived from field observations and lab experiments. Being able to infer more accurate values for wave velocities can help reduce the uncertainty in our knowledge of the subsurface and therefore enable us to build more accurate geological models. These better models directly improve our ability to identify and characterize hydrocarbon resources, and design optimal production strategies.

Typical seismic surveys cover hundreds of kilometers and use different hardware depending on whether the survey is performed on land or in a marine environment. The size of the physical domain to be processed is set by the extent of the region of interest and can span tens of kilometers per side and up to 10 kilometers in depth as oil deposits are located in the shallow sedimentary layers of the earth. At the smaller end of the length spectrum, the characteristic length scale of the problem is set by the wavelength of the traveling wave, which depends on the source frequency and the wave velocity. As the rock properties vary all across the physical domain, so does the wavelength, solving the wave equation on a domain with such spatial variations can only be achieved by using large calculators. Unlike simple problems where one can select a set of dimensionless parameters defined over a reduced domain of unity, the parametric representation will typically be selected to minimize the numerical bandwidth of the calculator and thus reduce numerical round-off errors.

To give an idea of the range of values of wave velocities typically encountered in geophysical surveys, the compressional wave velocity (v_p) in water is around 1.5 km/s and will typically vary between 2 and 6 km/s for sedimentary rocks. Shear-wave velocity (v_s) in rocks is roughly half those latter values, except in some soft, unconsolidated sediments (e.g., water bottoms) where it can be much lower.

Rocks are complex heterogeneous porous materials with features ranging in size from the micrometer scale to hundreds of kilometers. As elastic waves travel through them, random sub-wavelength features get homogenized and an average value emerges. Elastodynamic homogenization will not be covered here but it has a rich history during which many mathematical theories have been proposed to estimate the effective properties of composite materials [67, 79, 100]. The wavelength λ, established by the source frequency f through $\lambda = v/f$, approximately sets the

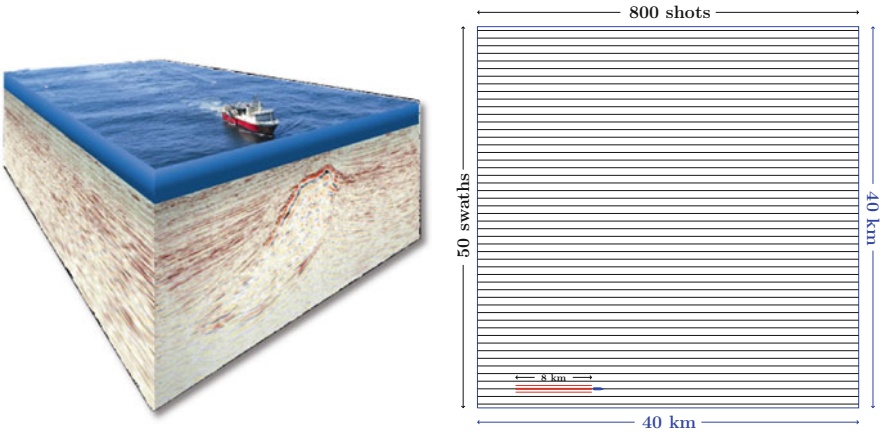

Fig. 1 A simple yet representative marine streamer seismic survey. An area of 40 km × 40 km is surveyed by a ship (indicated in blue and magnified 10×) towing 8 cables (only 4 shown in red) of length of 8 km, maintained at a distance of 100 m apart. The seismic source, an air gun array towed right behind the ship, is triggered every 50 m. Pressure sensors are located every 12.5 m along each cable. The ship has to navigate 50 swaths to sweep the entire area of interest

minimum size of heterogeneities that can be resolved. Ideally, the source frequency should be selected to resolve features at all scales down to a desired spatial resolution. In practice, however, the dominant frequency f_o of engineered sources is typically designed to be between 30 Hz and 50 Hz. This is a sweet spot that balances wave attenuation in the earth with technical and environmental limitations of generating enough energy at high frequency to travel forth to the maximum targeted depth and back to surface receivers.

The real-world constraints just described ultimately dictate the dimensions of the industrial problem to be solved. From a computational perspective, the numerical integration of the equations of motion propagating seismic waves over such a domain is an extreme-scale problem. The inversion of the rock properties through a PDE-constrained optimization approach is even more challenging as we will try to demonstrate in this paper.

To better illustrate the size of problems that the industry is facing, let us consider a realistic example of a marine streamer survey of an area of 40 km × 40 km that is shot from a ship towing 8 cables of 8 km of length, each separated by 100 m. This survey is shown in Figure 1.[1] Each shot, triggered every 50 m while the ship is moving, is captured for a total time duration T of 10 s by hydrophones located at every 12.5 m along each of the cables. Each shot therefore generates 5120 seismograms, also called traces, which are time series of measured pressure *vs.* time. Covering the full survey area will require 50 swaths each comprising 800 shots.

[1] In reality, cables will not be straight but will follow ocean currents. The hardware and controls required to keep the cables apart and monitor their locations are a real engineering accomplishment.

With a sampling period of 2 ms and a 24-bit amplitude representation, this survey will have generated 3.1 terabytes of data. Assuming a source dominant frequency of 30 Hz, our system would have a dominant wavelength λ_p of 50 m in water and could have a shear wavelength λ_s as small as 30 m for an assumed minimum v_s of 900 m/s. In comparison, if some fast rocks are present with, say, v_p of 4 km/s, the largest wavelength would be about 133 m. Rules of thumb for spatial discretization vary by numerical method but typically dictate a bare minimum between 5 to 10 points per wavelength. Considering 6 points for the sake of our argument, a regular grid of 5 meters could represent the shortest wavelength (30 m) and would require, at a minimum, 64 billion grid points for the full domain with a target depth of 5 km (i.e., $8000 \times 8000 \times 1000$). The total experimental listening time of the ship's receivers is about 4.6 days of real time ($50 \times 800 \times 10$ s). Reproducing the same experiment sequentially on a calculator by using the current generation of algorithms and supercomputers will necessitate much more than a week.

The motivation for going through this example in detail is to help the reader realize the magnitude of problem sizes that industry has to deal with, and that while great ideas might have tangible benefits for small problems, scalability of the proposed solutions must also be carefully evaluated. Moreover, when considering the full problem from the beginning, one can sometimes exploit symmetries or approximations that can be well justified under certain acquisition geometries, or subsurface conditions. This simple case also exemplifies that using a regular grid on problems with substantial variations in wave velocities might not be an optimal strategy: despite its extra computational costs, having a mesh that can adapt with the local wavelength can help solving larger problems [44].

Due to their large size, geophysical exploration problems have long been challenging to compute, and industry had to rely on approximations that allowed the problem to fit on available supercomputers. At first, only the first arrival times and so-called convolutional models were considered, then ray theory [17] allowed the computation of approximate synthetic seismograms and even enabled ray tomography when imaging from real data is posed as an inverse problem. More recently, advanced calculators allowed the computation of seismograms with better physical fidelity, finally enabling one to perform full-waveform inversion (FWI), which solves seismic imaging as a nonlinear parameter-estimation problem using the minimally modified, i.e., full-waveform, observed seismograms. An additional benefit of posing the problem of seismic tomography as an optimization problem is that it provides a rigorous framework through which additional information can be included in order to assist in getting more accurate parameter estimations.

Since the computation of industrial FWI problems is still very costly, some approximations are often used in the forward model. One common approximation involves solving the problem only for the lowest frequencies as the algorithmic complexity of the problem can easily be shown to scale as $\mathcal{O}(f_o^4)$, or even $\mathcal{O}(f_o^6)$ when denser spatial sampling of seismic sources on the earth surface

is also considered.[2] Another common approximation is the so-called *acoustic* approximation which involves filtering out the effects of shear waves to the best we can from the observed seismograms, and then assuming that the earth is a fluid. The computational cost of the resulting problem is much lower than the original problem, typically by at least one order of magnitude. Industry is actively investigating novel routes to perform FWI as robustly and economically as possible, while improving the FWI resolution over conventional imaging.

The goal of this paper is not to provide an overview of FWI; this has already been addressed by several excellent review papers and books (e.g., [34, 104]). This paper also does not focus on the computational methods used for generating synthetic seismograms; readers should consult [44], which is a very good book on that topic. Rather, our purpose is to present the reader with what we believe are still the most important unanswered questions as to how to bring FWI technology to full maturity. Also, from a more general perspective, there currently remain a large number of challenges regarding the estimation of parameters of large-scale systems using PDE-constrained optimization. It is anticipated that some of the successful methods developed for FWI will also bring impactful benefits to other PDE-constrained optimization problems, such as medical imaging, material design, and nondestructive testing applications. Recognizing that many of the challenges of FWI are common to other applications, we therefore made our best effort to avoid using field-specific jargon with the intent of making the text accessible to a broader audience.

For FWI, and other parameter-estimation problems constrained by PDEs, finding the best[3] earth parameters that can reproduce the measured data can be seen as two different activities: the forward problem of generating synthetic seismograms from a given three-dimensional (3-D) earth model, and the inverse problem of finding the optimum model that can best match the observed data. For the forward problem, current challenges can be grouped under two categories: physical fidelity and computation time. For the inverse problem, all challenges can be interpreted as "optimizing the optimizer," i.e., finding the most effective and robust optimization strategies. This quest includes finding the optimal parameter representation, the most useful misfit norm and annealing heuristics, the associated initial earth parameters that can lead to a reasonable solution, regularization, and constraints strategies, and the most efficient minimization algorithm. Many of the strategies

[2] This relation can be derived as follows. For a cubic survey of dimensions L^3, the number of spatial points N_x to compute will scale as L^3/h^3, where h is the discretization length set by $h \sim \lambda_o = v/f_o$. If using explicit time integration, the number of time integration steps $N_t = T/\Delta t$, where T is the listening time, and Δt the time step. As $\Delta t \sim h/v \sim f_o^{-1}$, then the number of operations $N_{op} \sim N_x N_t \sim f_o^4$, and therefore $\mathcal{O}(f_o^4)$. For larger f_o, the distance between the different shots at the earth surface also need to be smaller to maintain resolution, and therefore more shots are used (and need to be computed), resulting in $\mathcal{O}(f_o^6)$ when the inversion is performed separately for each source.

[3] We purposely picked the word *best* to emphasize that the problem has nonunique solutions and that the chosen solution might be the result of applying some additional measures of merit, sometimes even including some subjective domain expertise.

employed might implicitly exploit a structure of the physical problem that can guide the inversion to plausible geophysical solutions and are often based on intuition rather than on mathematical rigor.

This paper is structured along the ideas just presented. With the aim of being self-contained, we will first introduce the forward model at a level of detail sufficient for understanding the following sections. For physical fidelity, we will look in particular at the implications of using models with higher physical accuracy, and our ability to resolve the additional physical parameters. The computation time of the forward model is often related to the desired numerical accuracy, which should be minimally sufficient for the application considered. Therefore, we shall briefly present the challenge of selecting a discretization method for solving the forward problem at minimum computational costs for a given desired numerical accuracy. Then, we will discuss the inverse problem and describe the adjoint-state method used to generate the gradient of the data with respect to parameters of the earth model. Finally, we will present a few synthetic inversion cases that exemplify some of the challenges encountered in FWI. We assume that the reader has only limited exposure to seismology that can be complemented by introductory [91] or more advanced [3] textbooks. It is our hope that this paper, together with other contributions in this book, will motivate a new generation of researchers to contribute new ideas to this broad and challenging research area.

2 Forward Problem

Simulating field-scale experiments on a computer in view of generating synthetic seismograms is in itself an optimization problem: because of the large size of the problem, one has to balance accuracy with the cost of computation. For wave propagation, or any other mathematical modeling of physical phenomena, accuracy can be separated into two distinct categories: physical fidelity, i.e., having the proper representative equations to reproduce (most of) the complexity of the physical phenomena observed in the real experiment, and numerical accuracy, i.e., having the adequate numerical representation of the physical model on the calculator. Those two activities can be thought of as validation and verification of the simulation model, respectively [88, 89].[4] We will address these two issues separately.

2.1 Physical Fidelity

The physics behind the propagation of elastic waves in a heterogeneous material has been known since the nineteenth century thanks to Navier, Cauchy, and Green [93]. It assumes linear elasticity, namely, that the restoring force due to a

[4] A third element called *model qualification* determines the level of adequacy of the model for the intended application. This aspect will not be discussed here.

small deformation is accurately represented by a simple proportionality relation often referred to as a generalized Hooke's law. Except near the source where the displacements are relatively large and nonlinear effects can be important, this approximation is well justified for the small displacements typically encountered in reflection seismology. Also, when these equations are used in reflection seismology, the pressure resulting from the weight of the rocks is ignored, and linear-elasticity equations are interpreted as perturbations about an equilibrium configuration.

For a general elastic material, the stress-strain linear constitutive relationship reads[5]

$$\tau_{ij} = c_{ijkl} e_{kl}, \tag{1}$$

where $\mathbf{c}(\mathbf{x})$ is the elastic (fourth-order) tensor, and $\boldsymbol{\tau}(\mathbf{x}, t)$ is the stress (second-order) tensor. Notice that the elastic tensor is assumed to be independent of time for the duration of the experiment. The strain tensor $\mathbf{e}(\mathbf{x}, t)$ is obtained from the displacement vector $\mathbf{u}(\mathbf{x}, t)$ through

$$e_{ij} = \frac{1}{2}(\partial_i u_j + \partial_j u_i). \tag{2}$$

By construction, the strain tensor is symmetric. While $\mathbf{c}$ has 81 components, it has been shown that symmetry relations imposed by the conservation of linear and angular momenta as well as the postulation of the existence of an elastic strain-energy density function reduce the number of independent variables to 21.[6] A full derivation can be found in, e.g., [3, 56, 93]. For practical applications, it is common to rewrite Equation (1) using a 6×6 symmetric matrix $\tilde{\mathbf{c}}$. Using this so-called Voigt notation, the strain and stress tensors are now represented by 6-dimensional vectors $\tilde{\mathbf{e}}$ and $\tilde{\boldsymbol{\tau}}$, and Equation (1) is then written as

$$\tilde{\tau}_i = \tilde{c}_{ij} \tilde{e}_j, \tag{3}$$

a form much more convenient to implement on a computer. The term $\tilde{\mathbf{e}}$ defines the *engineering strain* where $\tilde{e}_i = e_{ii}$ (no sum) for $i \leq 3$ and $\tilde{e}_4 = 2e_{23}$, $\tilde{e}_5 = 2e_{13}$, and $\tilde{e}_6 = 2e_{12}$.

At the macroscopic scale, most amorphous materials display isotropic properties, and so are many sedimentary rocks, which are polycrystalline. However, rocks made of sub-wavelength layers of isotropic sediments of different elastic moduli can nevertheless exhibit anisotropy if the difference in properties between the different layers is large enough [6]. This behavior can be intuitively understood by considering stretching the material in a direction perpendicular to the layers. The emerging "average" of elastic constants will be reminiscent of springs acting

[5] We will be using the Einstein convention where repeated indices imply a sum over these indices.

[6] It is interesting to note that Green, Cauchy, and Poisson were part of a lengthy controversy in which the last two argued that the number of coefficients could not exceed 15. See [93] and references therein.

in series. If the material is stretched along those layers, however, the average will be reminiscent of springs in parallel.[7] This difference will give rise to a difference in wave speed depending on the direction of propagation. Thin-layered isotropic sedimentary rocks can therefore exhibit some degree of anisotropy. If the effects are nonnegligible, anisotropy will have to be included into the physical model. However, the additional parameters needed can introduce degeneracies preventing successful parameter estimation from the observed data, especially if all 21 parameters are considered. A common approach to this problem is to impose additional symmetries on the elastic tensor. While materials can be classified into 7 general classes of symmetry [56], only a few have broad geophysical applicability. Two classes of particular interest are the *orthorhombic* class, which has three orthogonal symmetry planes and requires 9 independent coefficients out of 12 nonzero $\tilde{c}_{ij}$ entries, and the *transversely isotropic* class in which axial symmetry reduces the elastic tensor to 5 independent coefficients out of 12 nonzero entries. In geophysics, the latter is termed *vertical transverse isotropic* (VTI), *horizontal transverse isotropic* (HTI), or *tilted transverse isotropic* (TTI) depending on the orientation of the symmetry axis with respect to the earth surface.

Thomsen [97] introduced a derived set of dimensionless parameters that provides a more intuitive connection with field-observed quantities, such as horizontal and vertical velocities in VTI environments. Instead of dealing with the elastic moduli,

$$\begin{pmatrix} \tilde{\tau}_1 \\ \tilde{\tau}_2 \\ \tilde{\tau}_3 \\ \tilde{\tau}_4 \\ \tilde{\tau}_5 \\ \tilde{\tau}_6 \end{pmatrix} = \begin{pmatrix} \tilde{c}_{11} & (\tilde{c}_{11} - 2\tilde{c}_{66}) & \tilde{c}_{13} & & & \\ (\tilde{c}_{11} - 2\tilde{c}_{66}) & \tilde{c}_{11} & \tilde{c}_{13} & & & \\ \tilde{c}_{13} & \tilde{c}_{13} & \tilde{c}_{33} & & & \\ & & & \tilde{c}_{44} & & \\ & & & & \tilde{c}_{44} & \\ & & & & & \tilde{c}_{66} \end{pmatrix} \begin{pmatrix} \tilde{e}_1 \\ \tilde{e}_2 \\ \tilde{e}_3 \\ \tilde{e}_4 \\ \tilde{e}_5 \\ \tilde{e}_6 \end{pmatrix}, \tag{4}$$

the 5 independent $\tilde{c}_{ij}$ are expressed as 2 elastic moduli ($\tilde{c}_{33}$ and $\tilde{c}_{44}$, related to v_p and v_s along the symmetry axis) coupled with three dimensionless measures of anisotropy ε, γ, and δ defined as

$$\varepsilon \equiv \frac{\tilde{c}_{11} - \tilde{c}_{33}}{2\tilde{c}_{33}}, \gamma \equiv \frac{\tilde{c}_{66} - \tilde{c}_{44}}{2\tilde{c}_{44}}, \text{ and } \delta \equiv \frac{(\tilde{c}_{13} + \tilde{c}_{44})^2 - (\tilde{c}_{33} - \tilde{c}_{44})^2}{2\tilde{c}_{33}(\tilde{c}_{33} - \tilde{c}_{44})}. \tag{5}$$

The first two parameters give a measure of anisotropy between the axial and azimuthal velocities: if there is no anisotropy, those two parameters are null. The third expression defines the ellipticity parameter δ, which has been simplified and is only exact for weak values of anisotropy. It can be associated to how axial and azimuthal velocities are changing from one to the other as the angle of propagation changes. See [97] for a full derivation.

[7] In material science, these two end-value members for binary composite materials are termed the Reuss and Voigt averages, respectively. For more details, see, for example, [66, 67].

As one Cartesian coordinate axis is typically chosen to be aligned with the vertical direction, two additional parameters are required to define the direction of the axis of symmetry for TTI. These angles can change in space, allowing for the representation of smoothly undulating layers, or tilted layered blocks. In Section 4.1.2, we show synthetic results on the inversion of two-dimensional VTI data using some of the anisotropy parameters just defined.

When material is considered macroscopically isotropic, spherical symmetry further reduces the elastic tensor to 2 independent coefficients. For such materials, the stress-strain relation thus simplifies to

$$\tau_{ij} = \lambda \delta_{ij} e_{kk} + 2\mu e_{ij}, \tag{6}$$

which is often expressed as a function of displacements u, using Equation (2),

$$\tau_{ij} = \lambda \delta_{ij} \partial_k u_k + \mu(\partial_i u_j + \partial_j u_i). \tag{7}$$

This expression, first derived by Cauchy, introduces the Lamé parameters $\lambda(\mathbf{x})$ and $\mu(\mathbf{x})$, the latter being equivalent to the shear modulus, while the bulk modulus $K = \lambda + 2\mu/3$.

When combining constitutive stress-strain relation (1) or (7) with Newton's second law of motion, stating that the rate of change in linear momentum is directly related to the forces exerted on that (infinitesimal) body, i.e.,

$$\frac{\partial}{\partial t}\left(\varrho \frac{\partial u_i}{\partial t}\right) = \partial_j \tau_{ij} + s_i, \tag{8}$$

one obtains a system of equations describing the propagation of elastic waves in a heterogeneous material. Here, $\varrho(\mathbf{x}, \mathbf{t})$ is the volumetric mass density, which is often approximated as constant in time, and $\mathbf{s}(\mathbf{x}, t)$ is a body force that could be gravity, or a seismic source term. Seismic sources are generally represented as point sources expressed using a seismic force moment $\mathbf{M}$. For each source k, a wavefield $\mathbf{u}^{(k)}(\mathbf{x}, t)$ can be computed using Equations (1) and (8) with

$$s_i^{(k)}(\mathbf{x}, t) = \delta_i^{(k)}(\mathbf{x} - \mathbf{x}^{(k)}) w^{(k)}(t) \partial_j M_{ij}^{(k)}, \tag{9}$$

where $\delta^{(k)}$ is a spatial support function, typically a spherically symmetric Dirac delta function or a modified Gaussian approximation with compact support, $\mathbf{x}^{(k)}$ is the kth-source location, and $w^{(k)}(t)$ is the source temporal signature. More complex finite sources can be modeled by an array of moment-tensor point sources [61]. The separation of spatial and temporal variables for the sources makes it possible to treat the source temporal signatures as unknowns as we shall see in Section 4.1.3.

Using those equations, one can model waves traveling in heterogeneous elastic media, thus carrying energy over distance. As waves emanate from a finite source, they expand and cover more and more volume. Because of energy conservation,

their amplitude will naturally decrease as they cover more volume through geometrical spreading. Waves traveling in rocks also exhibit intrinsic attenuation, caused by grain friction, fluid movement, and scattering at smaller scales, all of which also reduce wave amplitude. Because of these effects, a fraction of the energy stored at each oscillation cycle is not fully restituted. Attenuation will therefore be stronger for waves experiencing more cycles and transmission losses larger for waves with higher frequencies and/or traveling longer distances. This effect causes the high-frequency content of the seismic energy power spectrum to decay more rapidly than the low-frequency content, as viscoelastic waves travel.

The wave amplitudes A at two distances $x_2 > x_1$ away from the source can be related through [90]

$$A(x_2) = A(x_1) \left(\frac{x_1}{x_2} \right)^n e^{-\alpha(x_2 - x_1)}, \tag{10}$$

where α is the attenuation coefficient (in Nepers/m) and n depends on the geometry of the problem (e.g., $n = 0$ for plane waves, and $n = 1$ for spherical waves in 3-D).[8] In seismology, attenuation is commonly expressed using a dimensionless parameter Q called the *quality factor* and defined as [49]

$$\frac{2\pi}{Q} = \frac{2\alpha v}{f} = \frac{\Delta E}{E}, \tag{11}$$

where $\Delta E/E$ is the fractional energy loss per cycle, and v and f the velocity and frequency of the wave (recall $v/f = \lambda$). Values of Q ranging from 10 to several thousands have been measured experimentally for various rocks, and these values tend to be independent of frequency for dry rocks and vary as $Q \sim 1/f$ for fluids and water-saturated rocks [90, 99].[9]

For media with nearly homogeneous Q in space, it is sometimes possible to apply techniques that boost the observed signal with the intention of eliminating the effect of attenuation [108]. It is then possible to compare the modified experimental data with unattenuated synthetic seismograms using a misfit function that is sensitive to signal amplitude, such as any ℓ^p norm. Media with heterogeneous attenuation, however, will require modeling loss effects explicitly in the forward model in order to allow for a meaningful misfit measure between the synthetic and field-observed data.

Modeling attenuation in anelastic solids involves an empirical formulation that uses memory variables causing the elastic moduli $c_{ij}(t)$ to be time-dependent. The specific approach depends on whether the numerical solution method is formulated

[8] Nepers are not part of the SI units. They have dimensionless units and refer to the natural logarithm of ratios of measurements.

[9] Note that ultrasonic lab measurements are typically performed in the MHz range while the frequency bandwidth used in reflection seismology covers about 2 orders of magnitude ranging from 1 Hz to 100 Hz.

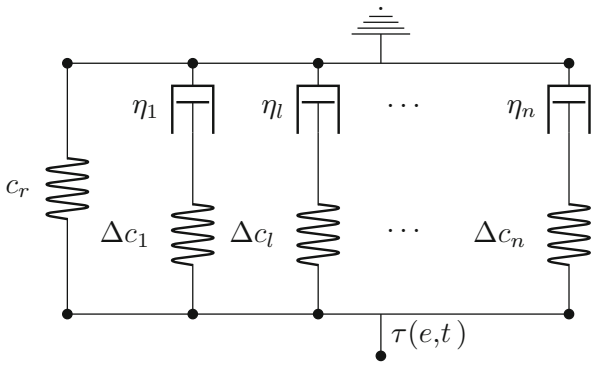

Fig. 2 Representation of the generalized Maxwell rheological model. Such models are used to account for attenuation in time-domain simulations of wave propagation. Model has n rheological mechanisms each composed of one (Hooke) spring and one (Stokes) dashpot. In this model, c_r is the relaxed modulus, while Δc_l and η_l are the elastic modulus and dashpot viscosity in the lth mechanism ($l = 1, \ldots, n$), respectively

in the space-time or space-frequency domain. For the latter, adding an imaginary part to the wave velocity is sufficient to account for attenuation effects and several models have been proposed to that effect [102]. The main difference between these phenomenological models resides in the frequency dependence of $Q(f)$. For some complex-variable velocity models [35, 48, 50], the attenuation is considered constant in the seismic frequency band, while Q varies with frequency for others [102].

In space-time formulations, a combination of simple springs and dashpots is required for modeling anelastic effects. Each one-dimensional spring is modeled with a (Hooke) linear response $\tau = ce$, similar to Equation (1), while each dashpot is a (Stokes) linear viscous term modeled as $\tau = \eta de/dt$, where η is the viscosity. See [71] for a nice introduction. The most general formulation of those combinations is termed the generalized Maxwell model [28], which, as its name indicates, includes all other variants [70]. Figure 2 shows such a rheological model. The mathematical description of this model is included in the Appendix.

In principle, each component of the elastic tensor c_{ijkl} can be represented by a different rheological model, allowing for anisotropic attenuation. In practice, however, attenuation is typically considered isotropic, and attenuation for compressional and shear waves, Q_p and Q_s, are most often considered separately. As we will demonstrate, the data sensitivity to attenuation is rather weak given that its effect is cumulative and takes place during signal transmission. Large distances between source and receivers (i.e., large offsets) should therefore be more favorable to attenuation inversion.

We have seen in this section that in order to be able to reproduce the data acquired in some surveys, it might be necessary to add effects such as anisotropy and attenuation to the forward simulation toolbox. The computational methods needed for producing synthetic traces for field-scale studies with this level of physical

accuracy are currently the state of the art in computational seismology. As we will discuss in Section 3, performing field-scale full-wavefield inversions with such high physical fidelity models is the soon-to-be-reached grand challenge given the upcoming generation of computers and algorithms.

The minimum level of physics required for reproducing the experimental data on a computer involves continuum mechanics at the macroscopic scale. The behavior of the material at the pore scale is still poorly understood and for the most part relies on phenomenological descriptions such as the attenuation models described above. On the engineering side of the acquisition, the details of the coupling of the receivers and source devices to the media is often neglected, as are the ambient and acquisition noises. Therefore, the numerical wave propagation engines that we develop are an idealized representation of the physical phenomena observed. Fortunately, they are a very good approximation and are believed to be able to account for the majority of the signal measured.

2.2 Numerical Accuracy

Solving the problem of wave propagation on a computer requires first discretizing the selected equations of motion. It is often convenient to discretize space and time independently (commonly known as the method of lines). Regardless of the level of physical fidelity selected, the wave equation can be represented by the following general equation which is obtained after space has been discretized [44],

$$\mathbf{M}\frac{\mathrm{d}^2\mathbf{u}(t)}{\mathrm{d}t^2} = \mathbf{K}\mathbf{u}(t) - \mathbf{C}\frac{\mathrm{d}\mathbf{u}(t)}{\mathrm{d}t} + \mathbf{s}(t). \tag{12}$$

Here, $\mathbf{M}$ is the so-called mass matrix, $\mathbf{K}$ the stiffness matrix, $\mathbf{C}$ the damping matrix, and $\mathbf{s}$ is a source term. The size of vector $\mathbf{u}(t)$, often termed the number of degrees of freedom, is the number of discrete points prescribed by the numerical method for the spatial discretization.[10] In this paper, we will use $\mathbf{A}s(t) = u(t)$ to express Equation (12), where $\mathbf{A}$ is referred to as the forward operator.

Equation (12) can be transformed in Fourier space, which will conveniently replace every time derivative by a mere multiplication by $\mathrm{i}\omega$, where $\mathrm{i}^2 = -1$, and ω is the angular frequency. The wave equation then becomes a linear system of equations,

$$-\omega^2\mathbf{M}\mathbf{u}(\omega) = \mathbf{K}\mathbf{u}(\omega) - \mathrm{i}\omega\mathbf{C}\mathbf{u}(\omega) + \mathbf{s}(\omega), \tag{13}$$

[10] To be more precise, in some cases the dimensionality of $\mathbf{u}$ may be higher than the number of spatial points when the numerical method introduces additional degrees of freedom, such as in the case of the discontinuous-Galerkin method.

or, for many independent sources identified by index k,

$$\mathbf{u}^{(k)}(\omega) = \mathbf{A}(\omega)\mathbf{s}^{(k)}(\omega), \tag{14}$$

where $\mathbf{A}(\omega) = (-\omega^2\mathbf{M} - \mathbf{K} + \mathrm{i}\omega\mathbf{C})^{-1}$ is defined as the forward discrete (and complex) operator in frequency.

Methods for solving Equation (14) are referred to as frequency-domain methods which, for small problems, involve performing **LU** (lower and upper) triangular decomposition on $\mathbf{A}^{-1}(\omega)$ [83]. The **LU** factorization can be reused for different sources as indicated by the lack of superscript k on operator $\mathbf{A}(\omega)$ but needs to be recomputed for each frequency [81]. For larger systems, however, this approach is not practical due to the large size of matrix $\mathbf{A}$, and one has to revert to iterative methods [85]. We refer the reader to [44, 105] for an accessible overview of the forward solvers used for seismic wave modeling, and their associated algorithmic complexity. Frequency-domain methods remain an area of active research (e.g., [45]). Some of the proposed approaches explicitly exploit the sparsity and structure of $\mathbf{A}^{-1}$ [78, 107], but these approaches still lack the ultimate scalability required by extreme-scale problems.

Because of the size limitation of the frequency-space formulation, very large problems are solved more efficiently when formulated in space and time (the so-called time-domain method). One of the differences between space-frequency and space-time methods is the trade-off between memory and CPU requirements that is representative of many other choices that have to be made when devising a computational strategy for solving an FWI problem. Both approaches can either use unstructured meshes, where the local space-discretization step size h is dictated by the smallest *local* wavelength, or regular grids, where h will be dictated by the smallest wavelength over the full domain.

In the time-domain approach, time is most commonly integrated explicitly with (typically) a global and constant time step Δt dictated by a conditional stability criterion such as CFL[11] [44]. While implicit time integration is possible (e.g., [75]), it is generally more costly than simple time interpolation schemes and more difficult to implement. The stability of implicit methods is better than the one associated with explicit methods, but it often comes at the cost of losing the high-frequency details of the solution. Moreover, implicit methods do not offer the same scalability when implemented on parallel computing architectures due to the need of solving large linear systems. Our strategy is to use the most economical method for solving the problem with a fit-for-purpose accuracy. For these reasons, time integration is most often performed through simple second-order methods such as Newmark-β, leap-frog, or Runge-Kutta methods when higher-order schemes are required. These schemes offer just enough precision to match the spatial discretization and can be easily adjusted as the spatial discretization is modified. Another way to keep the time and space discretizations consistent when higher-order spatial discretizations

[11] Stability criterion named after Courant, Friedrichs, and Lewy stating that $v\Delta t \leq h$.

are used is to use the so-called ADER (Arbitrary high-orDER) methods [27, 98] in which the required high-order time derivatives are cleverly derived from spatial derivatives.

The spatial discretization of Equation (12) can be performed using a variety of numerical methods. Finite-difference methods, however, have long been the workhorse for geophysicists. Most common variants derive from the seminal work by Virieux [103] who proposed using a stress-velocity system of two first-order equations on a staggered grid for stability purposes. Many other approaches have been proposed over the years, but only the spectral-element [51, 52, 101] and discontinuous-Galerkin methods [27, 31, 47, 76] seem to have gained broad acceptance and have emerged as competitive candidates for use in reflection seismology. The main appeal of spectral-element and discontinuous-Galerkin methods is their ability to easily accommodate unstructured meshes, the use of which is compelling in the presence of complex topography and when local refinement is necessary for accuracy purposes, such as near a free surface or near sources. Unstructured meshes also have the flexibility to adapt the mesh size to local wave speeds. In doing so, the regions of larger wave speeds would be discretized by larger elements while still sampling the wave form richly enough (the wavelength is larger in these high-velocity regions). In addition, the maximum allowable time step (as dictated by the CFL stability criterion) would not be penalized as much as it would when using regular grids.

In spite of their appeal, generating unstructured meshes over complex domains in three dimensions is difficult and costly, and implementing numerical methods on unstructured meshes is more complex than for their regular-grid counterparts. The increased complexity comes from the need for extra bookkeeping and from the requirement of computing extra terms that account for the change of coordinates between a reference cell and any deformed cell within the mesh. Those extra terms also increase the run-time cost of unstructured-mesh methods because the number of floating-point operations required to compute a spatial derivative is larger than for the case of regular grids. Parallel computations, which rely on domain-decomposition approaches where each process is responsible for computations on a sub-domain, bring about additional complexity when using unstructured meshes. These sub-domains are much simpler to define, and optimal load balancing is much easier to guarantee, when the physical domain is discretized by a regular grid. In contrast, unstructured-mesh domain partitioning is typically based on heuristic approaches and nonoptimal distributions of partition shapes and sizes are common and can hamper parallel efficiency (i.e., a relatively large amount of time may be spent communicating across partitions or waiting for processes to finish their computations). Finally, regular grids might be much more appealing to a practitioner who is used to performing operations that are straightforward on those grids, such as Fourier spatial analysis and filtering. The use of unstructured meshes does not prevent such operations, yet it requires using injection and projection operators to enable those regular-sampling benefits.

The above discussion highlights the dilemma that one faces when deciding on a numerical method. That dilemma is characterized by a trade-off between ease

of implementation, raw speed, and convenience on the one hand, and accuracy on the other hand. Therefore, and again, choosing the appropriate numerical method is in itself an optimization problem. Its solution depends on the implementation details of the selected algorithm, the available hardware, the type of problem being considered, and the background of the person choosing the method. While a geophysicist would most likely opt for ease of implementation and convenience, and might therefore lean towards the finite-difference method, a numerical analyst would likely choose a different method and might focus more on the convergence properties of the method with less focus on convenience or raw speed. A more pragmatic approach lies in between, where one seeks the most efficient method or, in other words, the fastest method that can achieve a given level of accuracy.

Comparing the costs and benefits of the different methods is not an easy task. In Figure 3, we show the wall time against the obtained accuracy for a series of runs using different numerical methods. A simple standing-wave problem in a homogeneous domain with reflecting boundaries was chosen as the test problem. A very small time step (5×10^{-6} s) is used to integrate the acoustic wave equation forward to final time 0.5 s. Central-difference, second-order Runge-Kutta and leap-frog time-stepping schemes are used for the spectral-element, discontinuous-Galerkin, and finite-difference methods, respectively. The time step is small enough that spatial errors dominate time errors, even on fine meshes, which simplifies the analysis of spatial-error considerations. For practical applications, time errors could be significant as time steps are pushed towards their stability limits for

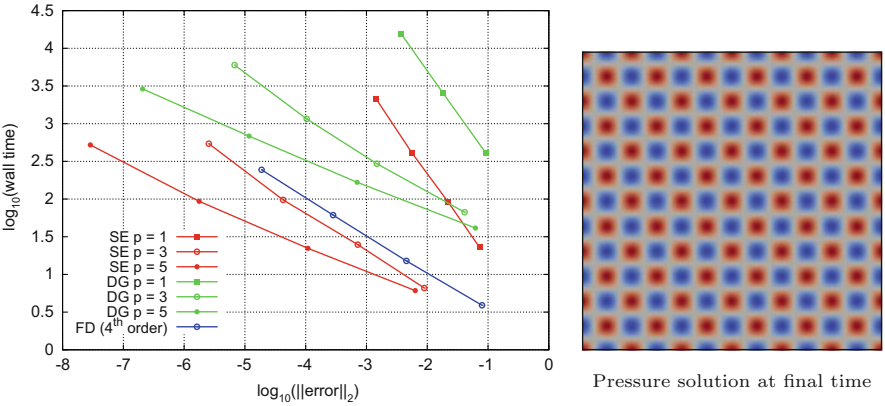

Fig. 3 Wall time on one processor against error (left figure) for a variety of numerical methods based on the solution to a two-dimensional standing-wave problem over a homogeneous domain with reflecting boundary conditions and for a fixed time step (pressure shown in the right figure). The error is the ℓ^2 norm of the difference between the modeled data and the analytical solution summed over the whole domain. Results for the spectral-element (SE), discontinuous-Galerkin (DG), and finite-difference (FD) methods are presented in red, green, and blue, respectively. Markers are used to categorize orders of accuracy, with second-order, fourth-order, and sixth-order methods represented by filled squares, empty circles, and filled circles, respectively

speed reasons. An alternative analysis would consist in running the same numerical methods on several meshes at constant CFL number. Time errors would quickly dominate and all schemes would eventually converge quadratically (second-order time-stepping schemes are used). However, high-order methods in space remain useful to mitigate grid-dispersion errors, independently of their spatial-convergence properties [1].

A few important lessons can be learned from the efficiency analysis shown in Figure 3. The most efficient method depends on the desired accuracy. For example, if one were to choose between a second-order (using polynomials of degree $p = 1$) spectral-element method and a fourth-order ($p = 3$) discontinuous-Galerkin method, the spectral-element method would be more efficient when lower accuracy is required while the discontinuous-Galerkin method rapidly becomes more efficient when higher accuracy is sought. For a given numerical method, high-order schemes ($p > 1$) are more efficient, which is visible for both the spectral-element and the discontinuous-Galerkin methods. The underlying reason comes from the way that the spatial error behaves upon mesh refinement. When the mesh resolution doubles, the run-time roughly quadruples in two spatial dimensions (the time step is fixed). The accuracy of a second-order method also quadruples and, therefore, no efficiency gain can be achieved. In two dimensions, any order of accuracy higher than two leads to accuracy gains that more than offset the cost increases upon refinement, which leads to flatter curves and the property that high-order methods achieve high accuracies more cheaply. However, diminishing returns are observed upon reaching orders of accuracy beyond four or five, leading to a sweet spot around $p = 3$ (fourth-order accuracy) for both the spectral-element and discontinuous-Galerkin methods on classical CPUs (conclusions may be different on alternative computing architectures). In addition, orders of accuracy higher than four would lead to meshes that could be inappropriately coarse for representing the inverted medium, unless material properties are also represented by higher-order polynomials, which is not simple to do. For the spectral-element and discontinuous-Galerkin methods, and for a given error, we observe that the cost reduction incurred by switching the polynomial degree from $p = 1$ to $p = 3$ is much larger than the cost reduction obtained by switching from $p = 3$ to $p = 5$. A final, though less definite, feature of the graph presented in Figure 3 is the high overall cost of the discontinuous-Galerkin method compared with both the spectral-element and finite-difference methods. The higher cost is mostly driven by the discontinuous polynomial representation, which leads to more degrees of freedom and the need for inter-element fluxes. These two attributes, however, also constitute strengths of the discontinuous-Galerkin method in terms of very good parallel efficiency and better dispersion properties via the use of numerical fluxes that are formulated in terms of the characteristic variables of the wave equation [41, 58]. Additional details on these methods can be found in [44].

An efficiency analysis, as presented in Figure 3, and the conclusions that can be drawn therefrom heavily depend on implementation details, available hardware, on the selected test case, and on whether runs are conducted serially or in parallel. Even if all methods are implemented using the same software best practices while re-using common parts of the algorithms, and using the same programming language

(our situation), the comparison could still be unfair. Undoubtedly, a developer could spend more time optimizing some methods and be less resolute about others (the finite-difference method in our case). In addition, the analysis also depends on the case considered. For a case where land topography reduces the order of accuracy of finite-difference schemes to one because of the staircase representation of topographical features [94], unstructured-mesh methods could easily be more efficient due to their second-order (at least) representation of topography. And again, we focus here on spatial errors by choosing a single time step that is small enough (and certainly smaller than any practical time step) for time errors to be negligible. Conclusions would be different under a regime where time-stepping schemes are taken into consideration. Finally, all runs are conducted on one processor and scalability properties of the methods have not been taken into account. Because of the simplicity of the test case considered above, the efficiency analysis presented in Figure 3 is just a starting point in the pursuit of a more thorough analysis that would include more realistic cases. This simple case is nevertheless presented here to illustrate the difficulty of determining the best numerical method for generating seismograms. Due to this difficulty, many research groups maintain several codes, each implementing a different numerical approach. Not only are these codes useful for verification purposes, they are also important for generating synthetic data sets to be inverted by another numerical method.

This section presented the many choices available for the forward simulation method powering the FWI inversion. This is an important decision as the largest fraction of the computational costs of FWI lies in running the forward (and adjoint) wave engine. Yet, the methods used are often dated, and very few special-purpose methods, or dedicated hardware, are available for performing this task. In most cases reported in the literature, hardware accelerators such as GPU's and FPGA's[12] provided speed benefits of less than an order of magnitude when considering problems of industrial relevance. Having access to algorithms that can provide orders of magnitude in performance gains would be a game changer for FWI, global seismology, and the many other engineering fields heavily relying on large-scale simulations of wave propagation. Besides computation, one can also envision miniature technologies and additive manufacturing methods that could use physical analogs to reproduce the experimental results.

3 Inverse Problem

In its simplest form, full-wavefield inversion formulates seismic imaging as an unconstrained optimization problem driven by an objective functional that includes some measure of misfit between the observed and computer-simulated data. Due to the extreme size of this inverse problem, only gradient-based optimization

[12] Graphics Processing Units and Field-Programmable Gate Arrays.

approaches are computationally affordable. A key enabling step for performing FWI is the utilization of the adjoint-state method for efficiently computing the gradient of the cost function with respect to the parameters of the PDE. While the adjoint-state method for inverse problems had been known since the early 1960s [18, 64], its application to the inverse problem posed by seismic wave propagation was first proposed by Lailly [55], and Tarantola [95] in the 1980s. Considered too expensive at the time, this approach generated increasing interest as the power of supercomputers continued to double regularly (following Moore's law) and after limited feasibility was demonstrated for two-dimensional problems, first in the time domain [36, 73], and then in the frequency domain [82]. Another critical step for the success of FWI is the use of annealing strategies such as multi-scale methods [15], in which the inverse problem is solved by gradually introducing higher frequencies from the data, thus bringing more resolution to the earth parameter model, and avoiding convergence to a "bad" local minimum.

In its most complex form, FWI is a family of PDE-constrained optimization problems linked together in ways that allow the inclusion of additional information, such as physical insights, in an orderly fashion. Solution strategies involving a hierarchy of inversions are required for guiding the optimization towards the most realistic solutions. Multi-scale and other strategies for partitioning and ordering the information (e.g., offset continuation, wave decomposition, etc.), preconditioning, regularization methods, constraints, exotic misfit functionals, are examples of the many ways in which this can be achieved. This scientific area is very active as more problems are amenable to solution given the current generation of high-performance computers and the availability of high-quality open-source software. Covering all aspects of the inverse problem is far beyond the scope of this paper. In the following, however, we will first introduce the basics of the optimization problem and then discuss some of the most commonly used strategies.

3.1 The Adjoint-State Method

The two main components of a gradient-based minimization algorithm are 1) a means for obtaining a descent direction, and 2) an iterative method to drive the minimization. For obtaining the descent direction, one could use a simple approach in which a small perturbation is introduced around each parameter of the PDE and its effect measured on the solution. For the acoustic wave equation, for example, that would entail perturbing the value of the speed of sound in one specific cell of the medium and simulating the resulting displacement field at the receivers. Given an earth model $\mathbf{m}$ containing the values of the speed of sound at each cell, we use the forward operator of the wave equation to generate the entire wavefield, or a subset of it represented by the seismograms measured only at specific receiver locations.

It is often more convenient to represent the relationship between the data and the model through the nonlinear mapping $\mathbf{F}(\mathbf{m}) = \mathbf{d}$, where $\mathbf{d}$ is the simulated data at all receivers and for all sources used. Note that $\mathbf{F}$ hides all the details about

the acquisition geometry and the sources used, including the multiple solutions to the wave equation, one for each of the prescribed source-receiver acquisition geometries. For a single perturbation ε at cell i of the model, a new solution is obtained as $\mathbf{F}(\mathbf{m} + \hat{\mathbf{i}}\varepsilon) = \mathbf{d}_{(i)}$, where $\hat{\mathbf{i}}$ is a unit vector in the ith direction. Then, $\mathbf{d}_{(i)} - \mathbf{d}$ is the resulting change from perturbation $\hat{\mathbf{i}}\varepsilon$. The finite-difference approximation

$$\frac{\mathbf{F}(\mathbf{m} + \hat{\mathbf{i}}\varepsilon) - \mathbf{F}(\mathbf{m})}{\varepsilon} \approx \frac{\partial \mathbf{d}}{\partial m_i} \tag{15}$$

can then be used to construct a gradient. If one were to use this approach to compute the gradient $\nabla_{\mathbf{m}}\mathbf{F}$, it would involve solving the wave equation as many times as the number of independent parameters in the earth model, and this for each source used in the survey. This approach is clearly not realistic when the earth model contains millions of cells, each with a few parameters.[13]

For an inverse problem such as FWI, one is interested in perturbations that will optimize a certain figure of merit. Typically, we seek earth models that minimize a distance from some field-observed values $\mathbf{d}^{\dagger}$. In its simplest form, an objective functional

$$\mathscr{J}(\mathbf{m}) = \frac{1}{2}\left|\mathbf{F}(\mathbf{m}) - \mathbf{d}^{\dagger}\right|^2 \tag{16}$$

is introduced and a simple chain-rule derivative is used to assemble the gradient $\mathfrak{g} = \nabla_{\mathbf{m}}\mathscr{J}(\mathbf{m})$ of the objective functional.

The adjoint-state method is a method to compute the gradient of $\mathscr{J}$ with respect to the model, i.e., $\nabla_{\mathbf{m}}\mathscr{J}(\mathbf{m})$. For completeness, we will extend the presentation from the joint tutorial paper [13] and derive the gradient for the elastic wave equations (7) and (8) using the adjoint-state method. We start with the two-dimensional isotropic elastic wave equation expressed as a system of first-order equations. As we shall see shortly, this system brings additional complexity as only 2 of the 5 state variables are given as observables.

By introducing the time derivative of the displacement $v_i = \partial_t u_i$, and using Equations (7) and (8), one obtains

$$\begin{aligned}
\partial_t \tau_{xx} &= (\lambda + 2\mu)\partial_x v_x + \lambda \partial_z v_z + f_{\tau_{xx}}, \\
\partial_t \tau_{zz} &= (\lambda + 2\mu)\partial_z v_z + \lambda \partial_x v_x + f_{\tau_{zz}}, \\
\partial_t \tau_{xz} &= \mu(\partial_z v_x + \partial_x v_z) + f_{\tau_{xz}}, \\
\partial_t v_x &= (1/\rho)\,(\partial_x \tau_{xx} + \partial_z \tau_{xz}) + f_{v_x},
\end{aligned} \tag{17}$$

[13] This simple approach is very useful, however, for providing test cases for verifying gradient computations.

$$\partial_t v_z = (1/\rho)\,(\partial_z \tau_{zz} + \partial_x \tau_{xz}) + f_{v_z}.$$

Here, we refer to the solution of this PDE system on a bounded domain Ω over a time span T with $\mathbf{w}(\lambda, \mu; \mathbf{f}) = (\tau_{xx}, \tau_{zz}, \tau_{xz}, v_x, v_z) \in L^2(\mathbb{R}^2 \times [0, \mathrm{T}]; \mathbb{R}^5)$ with a given source term $\mathbf{f} = \left(f_{\tau_{xx}}, f_{\tau_{zz}}, f_{\tau_{xz}}, f_{v_x}, f_{v_z}\right) \in L^2(\mathbb{R}^2 \times [0, \mathrm{T}]; \mathbb{R}^5)$.[14] Notice that we have described a general source term $\mathbf{f}$, which includes both acceleration and stress-rate components in Equation (17). The first three components of the source term, $f_{\tau_{ij}}$, can be used to represent the stress rate generated by force couples, such as pure explosive sources, while the last two, f_{v_i}, represent the acceleration resulting from a single point force.

For further simplifying the derivation of the gradient, we have assumed that the volume density $\rho(\mathbf{x})$ is fixed and known. This particular case still provides an example of a model with multiple parameters, $\mathbf{m} = (\lambda(\mathbf{x}), \mu(\mathbf{x}))$. It is left as an exercise to the reader to extend this case to include density, i.e., $\mathbf{m} = (\lambda(\mathbf{x}), \mu(\mathbf{x}), \rho(\mathbf{x}))$, and derive $\nabla_\rho \mathscr{J}(\mathbf{m})$.

We refer to the solution obtained for the kth source by $\mathbf{w}_{(k)}(\mathbf{x}, t) = (\tau_{xx}^{(k)}, \tau_{zz}^{(k)}, \tau_{xz}^{(k)}, v_x^{(k)}, v_z^{(k)})$. The simulated data generated by triggering the kth source is denoted by

$$\mathbf{d}_{(k)}(t) = \left(v_x(\mathbf{x}_1^{(k)}, t), v_z(\mathbf{x}_1^{(k)}, t), \ldots, v_x(\mathbf{x}_{n_r^{(k)}}^{(k)}, t), v_z(\mathbf{x}_{n_r^{(k)}}^{(k)}, t)\right) \in L^2([0, \mathrm{T}]; \mathbb{R}^{2n_r^{(k)}}),$$

as it is computed at all $n_r^{(k)}$ receivers located at $\mathbf{x}_1^{(k)}, \ldots, \mathbf{x}_{n_r^{(k)}}^{(k)}$, i.e., at the same locations as those where measurements were performed in the original survey.[15] The complete data set simulated for a survey of a number of n_s sources is denoted by

$$\mathbf{d}(t) = \left(\mathbf{d}_{(1)}, \mathbf{d}_{(2)}, \ldots \mathbf{d}_{(k)}, \ldots \mathbf{d}_{(n_s)}\right).$$

Simulated data $\mathbf{d}_{(k)}(t)$ is obtained by solving Equation (17) and then extracting the subset of state variables $\{v_x, v_z\}$ at the $n_r^{(k)}$ receiver locations. This is achieved by applying a restriction operator $\mathbf{B}$ on $\mathbf{w}$, i.e.,

$$\mathbf{B}_{(k)}\mathbf{w} = \begin{pmatrix} v_x \\ v_z \end{pmatrix}_{\left\{\left(\mathbf{x}_1^{(k)}, \ldots, \mathbf{x}_{n_r^{(k)}}^{(k)}\right) \times [0, \mathrm{T}]\right\}}, \tag{18}$$

[14] By the notation $L^2(\mathbb{R}^a \times [0, \mathrm{T}]; \mathbb{R}^b)$, we express a b-dimensional vector defined over an a-dimensional space over a time interval [0, T], with a norm that is square-integrable over $\mathbb{R}^a \times [0, \mathrm{T}]$.

[15] Current generation of geophones uses accelerometers that can measure the three orthogonal components of acceleration. Such receivers are termed multicomponents or 3-C.

for the kth source. Notice that operator $\mathbf{B}_{(k)} : L^2(\mathbb{R}^2 \times [0, \mathrm{T}]; \mathbb{R}^5) \rightarrow L^2([0, \mathrm{T}]; \mathbb{R}^{2n_r^{(k)}})$ simply has the role of extracting the time series for only the displacement rates, and only at the receiver locations.

Initial conditions are $v_x(\mathbf{x}, 0) = v_z(\mathbf{x}, 0) = 0$ and $\tau_{xx}(\mathbf{x}, 0) = \tau_{zz}(\mathbf{x}, 0) = \tau_{xz}(\mathbf{x}, 0) = 0$, indicating that the medium is in mechanical equilibrium. Dealing with the subtleties of the boundary conditions is beyond the scope of this paper and therefore we will assume an infinite domain boundary onto which both the stress and the velocity vanish.

Given the previous assumptions and notation, the objective functional given by Equation (16) takes the following explicit form

$$\begin{aligned} \mathscr{J}(\lambda, \mu) &= \sum_{\substack{\text{sources} \\ k}}^{n_s} \mathscr{J}_{(k)}(\lambda, \mu) \\ &= \frac{1}{2} \sum_{\substack{\text{sources} \\ k}}^{n_s} \int_{t_{0(k)}}^{t_{0(k)}+\mathrm{T}} \left| \mathbf{d}_{(k)}(\lambda, \mu; t) - \mathbf{d}_{(k)}^{\dagger}(t) \right|^2 dt. \end{aligned} \tag{19}$$

Notice how all traces for source k share the same time interval.

We begin by deriving the gradient $\nabla_{\mathbf{m}} \mathscr{J}_{(k)}(\mathbf{m})$, i.e., the gradient for only one of the seismic sources. The final gradient is assembled by adding the contributions from all sources. Indeed, due to the independence of the sources, and as already exploited in (19), it follows that

$$\nabla_{\mathbf{m}} \mathscr{J}(\mathbf{m}) = \sum_{\substack{\text{sources} \\ k}}^{n_s} \nabla_{\mathbf{m}} \mathscr{J}_{(k)}(\mathbf{m}).$$

Following the same approach as in [13], we define the forward and adjoint operators for the Lamé parameters $\lambda(\mathbf{x})$ and $\mu(\mathbf{x})$. The forward operator $\mathbf{A}$ is defined by

$$\mathbf{A}\mathbf{f} = \mathbf{w}(\lambda, \mu), \tag{20}$$

where we temporarily dropped the subscript notation (k) for clarity. Again, notice that forward operator $\mathbf{A}$ is the inverse of the differential operator as it operates on the source term $\mathbf{f}$ to yield solution $\mathbf{w}$. The adjoint operator $\mathbf{A}^*$, which plays an important role in computing $\nabla_{\mathbf{m}} \mathscr{J}_{(k)}(\mathbf{m})$, is defined using the adjoint identity [62, 64],[16]

$$\langle \mathbf{A}\mathbf{f}, \mathbf{g} \rangle_{L^2(\mathbb{R}^2 \times [0,\mathrm{T}]; \mathbb{R}^5)} = \langle \mathbf{f}, \mathbf{A}^* \mathbf{g} \rangle_{L^2(\mathbb{R}^2 \times [0,\mathrm{T}]; \mathbb{R}^5)} \quad \forall\, \mathbf{f}, \mathbf{g} \in L^2\left(\mathbb{R}^2 \times [0, \mathrm{T}]; \mathbb{R}^5\right), \tag{21}$$

[16] This relation is also referred to as the Lagrange identity.

where $\langle\,,\rangle$ denotes a scalar product in $L^2\left(\mathbb{R}^2 \times [0, \mathrm{T}]; \mathbb{R}^5\right)$. The adjoint operator $\mathbf{A}^*$ defines the adjoint variable $\bar{\mathbf{w}}(\lambda, \mu; \mathbf{g})$ through the solution of

$$\mathbf{A}^*\mathbf{g} = \bar{\mathbf{w}}(\lambda, \mu), \tag{22}$$

where $\mathbf{g}$ is an associated source to be determined. In our particular case, the adjoint variable, $\bar{\mathbf{w}}(\lambda, \mu; \mathbf{g}) = (\bar{\tau}_{xx}, \bar{\tau}_{zz}, \bar{\tau}_{xz}, \bar{v}_x, \bar{v}_z)$, can be shown to be the solution to the adjoint problem

$$\begin{aligned}
\partial_t \bar{\tau}_{xx} &= (1/\rho)\,(\partial_x \bar{v}_x + \partial_z \bar{v}_z) + g_{\bar{\tau}_{xx}},\\
\partial_t \bar{\tau}_{zz} &= (1/\rho)\,(\partial_z \bar{v}_z + \partial_x \bar{v}_x) + g_{\bar{\tau}_{zz}},\\
\partial_t \bar{\tau}_{xz} &= (1/\rho)\,(\partial_z \bar{v}_x + \partial_x \bar{v}_z) + g_{\bar{\tau}_{xz}},\\
\partial_t \bar{v}_x &= \partial_x\left((\lambda + 2\mu)\bar{\tau}_{xx} + \lambda \bar{v}_x\right) + \partial_z\left(\mu \bar{\tau}_{xz}\right) + g_{v_x},\\
\partial_t \bar{v}_z &= \partial_z\left(\lambda \bar{\tau}_{xx} + (\lambda + 2\mu)\bar{v}_z\right) + \partial_x\left(\mu \bar{\tau}_{xz}\right) + g_{v_z},\\
\bar{\tau}_{xx} &= \bar{\tau}_{zz} = \bar{\tau}_{xz} = \bar{v}_x = \bar{v}_z = 0 \text{ at } t = \mathrm{T}.
\end{aligned} \tag{23}$$

Equation (23) is also an elastic wave equation but, in contrast to Equation (17), the solution's final-time value is prescribed, and therefore Equation (23) must be solved by marching backwards in time, beginning at $t = \mathrm{T}$. It is also left as an exercise to the reader to establish that the operators given by Equations (17) and (23) and expressed as Equations (20) and (22) satisfy the adjoint identity (21). The proof can be demonstrated through integration by parts as outlined in the joint tutorial paper [13].

We now use the adjoint identity, Equation (21), to compute the gradient $\nabla_{\mathbf{m}} \mathscr{J}_{(k)}(\mathbf{m})$ with respect to model parameters $\mathbf{m} = (\lambda, \mu)$. Using the L^2 norm to project the gradient $\nabla_{\mathbf{m}} \mathscr{J}_{(k)}(\mathbf{m})$ on a small model perturbation $\delta\mathbf{m}$, one gets the following relationship

$$\mathscr{J}_{(k)}(\mathbf{m} + \delta\mathbf{m}) - \mathscr{J}_{(k)}(\mathbf{m}) = \left\langle \nabla_{\mathbf{m}} \mathscr{J}_{(k)}, \delta\mathbf{m} \right\rangle_{L^2(\mathbb{R}^2;\mathbb{R}^2)} + \mathscr{O}\left(\|\delta\mathbf{m}\|^2_{L^2(\mathbb{R}^2;\mathbb{R}^2)}\right). \tag{24}$$

In order to compute $\nabla_{\mathbf{m}} \mathscr{J}_{(k)}$, we begin by linearizing the nonlinear map $\mathbf{w}(\mathbf{m}; \mathbf{f})$ with respect to $\mathbf{m}$. For a small perturbation $\delta\mathbf{m} = (\delta\lambda, \delta\mu)$, we define $\delta\mathbf{w}$ by

$$\mathbf{w}(\mathbf{m} + \delta\mathbf{m}; \mathbf{f}) \approx \mathbf{w}(\mathbf{m}; \mathbf{f}) + \delta\mathbf{w} + \mathscr{O}\left(\|\delta\mathbf{m}\|^2_{L^2(\mathbb{R}^2;\mathbb{R}^2)}\right). \tag{25}$$

In particular, the same relation will hold for restricted values

$$\mathbf{d}_{(k)}(\mathbf{m} + \delta\mathbf{m}; \mathbf{f}) \approx \mathbf{d}_{(k)}(\mathbf{m}; \mathbf{f}) + \delta\mathbf{d}_{(k)} + \mathscr{O}\left(\|\delta\mathbf{m}\|^2_{L^2(\mathbb{R}^2;\mathbb{R}^2)}\right). \tag{26}$$

Inserting the approximation (25) into (17) for both Lamé parameters, it follows that $\delta\mathbf{w}$ satisfies the linearized forward problem

$$\begin{aligned}
\partial_t \delta\tau_{xx} &= (\lambda + 2\mu)\partial_x \delta v_x + \lambda\partial_z \delta v_z + (\delta\lambda + 2\delta\mu)\partial_x v_x + \delta\lambda\partial_z v_z, \\
\partial_t \delta\tau_{zz} &= (\lambda + 2\mu)\partial_z \delta v_z + \lambda\partial_x \delta v_x + (\delta\lambda + 2\delta\mu)\partial_z v_z + \delta\lambda\partial_x v_x, \\
\partial_t \delta\tau_{xz} &= \mu(\partial_z \delta v_x + \partial_x \delta v_z) + \delta\mu(\partial_z v_x + \partial_x v_z), \\
\partial_t \delta v_x &= (1/\rho)\left(\partial_x \delta\tau_{xx} + \partial_z \delta\tau_{xz}\right), \\
\partial_t \delta v_z &= (1/\rho)\left(\partial_z \delta\tau_{zz} + \partial_x \delta\tau_{xz}\right).
\end{aligned} \tag{27}$$

Using definition (20), we rewrite (27) as

$$\delta\mathbf{w}_{(k)} = \mathbf{A}\begin{pmatrix} (\delta\lambda + 2\delta\mu)\partial_x v_x^{(k)} + \delta\lambda\partial_z v_z^{(k)} \\ (\delta\lambda + 2\delta\mu)\partial_z v_z^{(k)} + \delta\lambda\partial_x v_x^{(k)} \\ \delta\mu(\partial_z v_x^{(k)} + \partial_x v_z^{(k)}) \\ 0 \\ 0 \end{pmatrix}, \tag{28}$$

where we reintroduced our notation for source k, and where $\delta\mathbf{w}_{(k)} \in L^2\left(\mathbb{R}^2 \times [0, \mathrm{T}]; \mathbb{R}^5\right)$. Now, we have all the necessary results for computing $\nabla J_{(k)}$. Using the linearization (26) in (19), it follows that

$$\mathscr{J}_{(k)}(\mathbf{m} + \delta\mathbf{m}) - \mathscr{J}_{(k)}(\mathbf{m}) \approx \int_{t_{0(k)}}^{t_{0(k)}+\mathrm{T}} \left(\mathbf{B}_{(k)}\mathbf{w}_{(k)}(\mathbf{m}) - \mathbf{d}_{(k)}^{\dagger}\right)^T \left(\mathbf{B}_{(k)}\delta\mathbf{w}_{(k)}(\mathbf{m})\right), \tag{29}$$

as $\delta\mathbf{d}_{(k)} = \mathbf{B}_{(k)}\delta\mathbf{w}_{(k)}$. In order to bring Equation (29) to the desired function space, we need to introduce two new transformations. The first transformation brings the measured traces in $\mathbb{R}^5$ by defining

$$\mathbf{w}_{(k)}^{\dagger} = \left(0, 0, 0, v_x^{(k)\dagger}, v_z^{(k)\dagger}\right),$$

while the second one defines a projection operator $\mathbf{P}$ such that

$$\mathbf{P}\mathbf{w}_{(k)} = \left(0, 0, 0, v_x^{(k)}, v_z^{(k)}\right).$$

Notice that $\mathbf{w}_{(k)}^{\dagger}$ is only defined at the receiver locations. We can then rewrite Equation (29) as

$$\mathscr{J}_{(k)}(\mathbf{m}+\delta\mathbf{m}) - \mathscr{J}_{(k)}(\mathbf{m}) \approx \left\langle \sum_{\substack{\text{receivers}\\ r}}^{n_r^{(k)}} \delta\left(\mathbf{x}-\mathbf{x_r}^{(k)}\right)\left(\mathbf{P}\mathbf{w}_{(k)} - \mathbf{w}^{\dagger}_{(k)}\right), \delta\mathbf{w}_{(k)} \right\rangle_{L^2(\mathbb{R}^2\times[0,\mathrm{T}];\mathbb{R}^5)}. \tag{30}$$

Substitution of (28) into (30), followed by an application of the adjoint identity (21), results in

$$\begin{aligned}
&\mathscr{J}_{(k)}(\mathbf{m}+\delta\mathbf{m}) - \mathscr{J}_{(k)}(\mathbf{m}) \\
&= \left\langle \sum_{r=1}^{n_r^{(k)}} \delta\left(\mathbf{x}-\mathbf{x_r}^{(k)}\right)\left(\mathbf{P}\mathbf{w}_{(k)} - \mathbf{w}^{\dagger}_{(k)}\right), \right. \\
&\left. \mathbf{A}\begin{pmatrix} (\delta\lambda+2\delta\mu)\partial_x v_x^{(k)} + \delta\lambda\partial_z v_z^{(k)} \\ (\delta\lambda+2\delta\mu)\partial_z v_z^{(k)} + \delta\lambda\partial_x v_x^{(k)} \\ \delta\mu(\partial_z v_x^{(k)} + \partial_x v_z^{(k)}) \\ 0 \\ 0 \end{pmatrix} \right\rangle_{L^2(\mathbb{R}^2\times[0,\mathrm{T}];\mathbb{R}^5)} \\
&= \left\langle \mathbf{A}^*\left(\sum_{r=1}^{n_r^{(k)}} \delta\left(\mathbf{x}-\mathbf{x_r}^{(k)}\right)\left(\mathbf{P}\mathbf{w}_{(k)} - \mathbf{w}^{\dagger}_{(k)}\right)\right), \right. \\
&\left. \begin{pmatrix} (\delta\lambda+2\delta\mu)\partial_x v_x^{(k)} + \delta\lambda\partial_z v_z^{(k)} \\ (\delta\lambda+2\delta\mu)\partial_z v_z^{(k)} + \delta\lambda\partial_x v_x^{(k)} \\ \delta\mu(\partial_z v_x^{(k)} + \partial_x v_z^{(k)}) \\ 0 \\ 0 \end{pmatrix} \right\rangle_{L^2(\mathbb{R}^2\times[0,\mathrm{T}];\mathbb{R}^5)}.
\end{aligned} \tag{31}$$

Notice how adjoint equation $\bar{\mathbf{w}}_{(k)} = \mathbf{A}^*\mathbf{g}_{(k)}$ contains an adjoint source term that is composed of the misfit in the measured displacement rates as

$$\mathbf{g}_{(k)} = \sum_{r=1}^{n_r^{(k)}} \delta\left(\mathbf{x}-\mathbf{x_r}^{(k)}\right)\left(0,0,0,\left[v_x^{(k)}(\mathbf{x},t) - v_x^{(k)\dagger}(\mathbf{x},t)\right],\left[v_z^{(k)}(\mathbf{x},t) - v_z^{(k)\dagger}(\mathbf{x},t)\right]\right). \tag{32}$$

Recalling the definition (24) of $\nabla_{\mathbf{m}}\mathscr{J}_{(k)}$, it follows from (31) that

$$\nabla_{\lambda}\mathscr{J}_{(k)} = \int_0^{\mathrm{T}} \bar{\tau}_{xx}^{(k)}\left(\partial_x v_x^{(k)} + \partial_z v_z^{(k)}\right) + \bar{\tau}_{zz}^{(k)}\left(\partial_z v_z^{(k)} + \partial_x v_x^{(k)}\right) dt,$$

$$\nabla_{\mu} \mathscr{J}_{(k)} = \int_0^{\mathrm{T}} 2\bar{\tau}_{xx}^{(k)} \left(\partial_x v_x^{(k)}\right) + 2\bar{\tau}_{zz}^{(k)} \left(\partial_z v_z^{(k)}\right) + \bar{\tau}_{xz}^{(k)} \left(\partial_z v_x^{(k)} + \partial_x v_z^{(k)}\right) \, dt \tag{33}$$

where $\bar{\boldsymbol{\tau}}_{ij}$ solves the adjoint problem

$$\bar{\mathbf{w}}_{(k)} = \begin{pmatrix} \bar{\tau}_{xx}^{(k)} \\ \bar{\tau}_{zz}^{(k)} \\ \bar{\tau}_{xz}^{(k)} \\ \bar{v}_x^{(k)} \\ \bar{v}_z^{(k)} \end{pmatrix} = \mathbf{A}^* \mathbf{g}_{(k)}. \tag{34}$$

Equation (33) indicate that computing $\nabla \mathscr{J}_{(k)}$ involves solving two separate, adjoint problems. First, one must find the solution of the forward problem (17), where the source term $\mathbf{f}_{(k)}$ is given. Then, the solution of the adjoint problem, Equation (23), must be computed, with source term $\mathbf{g}_{(k)}$ as defined in Equation (32), which injects the data misfit at the receiver locations. The computational cost of estimating the gradient then becomes about twice the cost of a forward solution, since the adjoint problem is approximately the same as the cost of the forward problem, and both involve solving an elastic wave equation in the same medium. This represents a substantial advantage over the simple finite-difference approach outlined in Equation (15).

In certain situations (see [60] and [14]), it is possible to use a formulation that makes the operator $\mathbf{A}$ self-adjoint such that $\mathbf{A} = \mathbf{A}^*$. These formulations have the advantage of using a single forward operator to compute the direct and back-propagated wavefields using only one forward modeling operator by only applying a correction term to the source.

3.2 Gradient in the Time Domain

We first discuss methods for computing the gradient in the time domain as this is the most commonly used approach for large-scale problems. After a level of physical fidelity has been selected and implemented according to the numerical method of one's choice, the first step required for performing FWI is the implementation of a solver for the associated adjoint equation. As the adjoint equation for the wave equation is also a wave equation [19, 63], this step can reuse many of the elements implemented in the original forward modeling engine. The mathematical procedure required for computing the gradient using the adjoint-state method is described in detail using continuous operators in the previous section and in a joint tutorial paper [13]. Implementing a continuous adjoint on a computer, however, involves additional decisions. A fair amount of practical experience in optimal control suggests the type of adjoint to implement [10], i.e., should one implement a discretization of the continuous adjoint, or implement the adjoint of the discrete

forward operator? It turns out that the latter is a more robust approach as it guarantees that the gradient remains accurate even when the numerical accuracy of the selected forward approach is diminished due to, e.g., using too coarse a grid for the problem.

There have been many research advances on techniques aimed at using calculators to automatically compute derivatives (e.g., [8, 39, 74]). It is only recently [32] that powerful algorithms for generating the adjoint of transient finite-element codes were developed and made available for a specific software platform [59]. For many practitioners, however, the most common approach for computing the discrete adjoint operator remains a manual procedure in which the operator is constructed by going, block by block, through the code implementation of the forward algorithm.

In the time domain, each gradient computation requires the full wavefield of the forward and adjoint problems for the whole duration of the simulation in d spatial dimensions. For each source k, the continuous variable $\mathbf{u}^{(k)}(\mathbf{x}, t) \in \mathbb{R}^d \times [0, \mathrm{T}]$ and its adjoint $\bar{\mathbf{u}}^{(k)}(\mathbf{x}, t) \in \mathbb{R}^d \times [0, \mathrm{T}]$ have been spatially discretized in N points in space and N_t time steps to give variable $\mathbf{u}^{(k)}(t) \in \mathbb{R}^{Nd} \times N_t$, and its adjoint $\bar{\mathbf{u}}^{(k)}(t) \in \mathbb{R}^{Nd} \times N_t$, obtained from the time-reversed wave propagation of the misfit residuals at the receivers which are acting as sources. In other words, the adjoint equation is a wave equation where the receivers are acting as sources with a time function built from the difference between the observed signal at that receiver and the one that was computer generated with the current model.[17] Notice how the velocity $\mathbf{v}$ variables in time integral equation (33) are forward propagated while the adjoint stress $\bar{\boldsymbol{\tau}}$ needs to be computed from a backward propagation. In discrete form, $2N_t$ copies of the spatial wavefields have to be stored in order to compute the gradient, $\mathfrak{g} = \boldsymbol{\nabla}_{\mathrm{m}} \mathscr{J}(\mathrm{m})$, of the objective functional $\mathscr{J}$, which depends on model $\mathbf{m}$. By definition, we have

$$\mathscr{J}(\mathbf{m} + \boldsymbol{\delta}\mathbf{m}) - \mathscr{J}(\mathbf{m}) \approx \boldsymbol{\nabla}_{\mathbf{m}} \mathscr{J}(\mathbf{m}) \cdot \boldsymbol{\delta}\mathbf{m}. \tag{35}$$

For example, in the case of a simple acoustic equation, contribution to the gradient coming from source k is obtained from[18]

$$\mathfrak{g}^{(k)} = \sum_{n_t=0}^{N_t} \mathbf{R}[\mathbf{u}^{(k)}(n_t \Delta t)]\, \bar{\mathbf{u}}^{(k)}(n_t \Delta t), \tag{36}$$

[17] See, e.g., Equation (32). A similar result is obtained for the acoustic wave equation as derived in joint tutorial [13].

[18] Some authors refer to this operation as a zero-lag correlation.

where Δt is the time step, and $\mathbf{R}(\mathbf{m})$ is an operator that derives from a linear perturbation $\boldsymbol{\delta}\mathbf{m}$ for the particular PDE of interest.[19] We refer the reader to [13, 64] for additional detail.

For three-dimensional problems, the storage needed for saving two wavefields for all time steps is generally too large to fit on the current generation of fast-access memory. By trading data storage for computing time, one can keep fewer snapshots and reuse these as initial values for recomputing the wavefields. For a given amount of available storage, optimal strategies for checkpointing that minimize the amount of recomputation have been derived for general [38] and special [5, 111] cases. Another approach is to save the values of the wavefield at the boundaries at all times and the single snapshot of the full wavefield at the final time T from which the simulation is time reversed using the boundaries as sources [21]. This approach is particularly advantageous when the surface-to-volume ratio of the computational domain is small (e.g., 3-D), and when the model does not have attenuation.

The full gradient is obtained by summing contributions from each of the sources. While each sequential source computation is pleasingly parallel,[20] it is also possible to trigger multiple sources simultaneously in the same forward simulation by encoding the sources and firing them simultaneously in one single or a few mega-shots [53]. Using the linearity of the wave equation, ultimately responsible for the wave superposition principle, the observed data can be transformed using the same encoding, and summed the same way, so that synthetic and observed data can still be compared meaningfully for each mega-shot. This approach provides significant speedup with the additional benefit of showing more robust convergence when initial models are poor [7]. This behavior derives from the stochastic perturbations to the objective functional resulting from selecting new random encodings at regular steps (typically $\sim$5) during the iterations of the minimization. Simultaneous-source encoding algorithms, however, require that receivers for a given mega-shot be common to all shots included. Devising a simultaneous-source encoding algorithm for data acquired with moving receivers, such as those in marine surveys, is still an open problem.

3.3 Gradient in the Frequency Domain

The frequency-domain approach to FWI has been pioneered by Pratt and Shin [81]. The computation of the gradient also involves a back propagation of the misfit residuals and is therefore bound by the same limitations as those of the space-frequency forward operator. We refer the reader to [81] for a full derivation. Space-

[19] In the derivation for a simple 1-D acoustic wave equation, one obtains (see, e.g., [13]) $\mathbf{R}(\kappa) = \frac{1}{\kappa^2}\frac{\partial^2}{\partial t^2}$. In practical computations, using the linearity of the wave equation, it is more efficient to apply this operator to the source time functions.

[20] Also less affectionately known as *trivially* or *embarrassingly* parallel problems.

frequency methods offer some attractive algorithmic benefits. Modeling attenuation, for example, is much simpler in the frequency domain, as already discussed in Section 2. Moreover, the computation of the gradient does not require checkpointing strategies such as those required for time-domain approaches. However, given the current generation of algorithms for solving large linear systems of equations on distributed-memory computers, space-frequency methods have more limited scalability than space-time methods. But, this situation could change rapidly if a new class of linear solvers becomes readily available.

3.4 Optimization Methods

We already mentioned that only local, gradient-based methods are affordable for extreme-scale problems such as FWI. While all these methods derive from Taylor's theorem,[21] there are some differences in the way the second derivative is used, if used at all. A good introduction to unconstrained optimization algorithms can be found in [74]. These algorithms require a starting point for the earth parameters $\mathbf{m_o}$ from which a series of iterates will be generated, converging to a local minimum quadratically at best. It is therefore assumed that the problem is formulated and solved in such a way that there exists a descending path for a series of iterates, going from the initial model to the "final" solution, where each iterate is linked to the next through successive local quadratic (or linear) approximations and proper step lengths. A major challenge for FWI is therefore to generate a good starting earth model and to devise a robust strategy connecting our starting model to the model containing the ultimately recoverable information.

At each iteration i, the objective functional can be represented using a general form,

$$\mathscr{J}(\mathbf{m_i}) = \left|\mathbf{F}(\mathbf{m_i}) - \mathbf{d}^{\dagger}\right| + \mathscr{R}_i(\mathbf{m_i}), \tag{37}$$

where

$$\mathbf{F}(\mathbf{m_i}) = \mathbf{d_i}, \tag{38}$$

is the nonlinear mapping between the model $\mathbf{m_i}$ and the relevant simulated data $\mathbf{d_i}$, while $\mathbf{d}^{\dagger}$ is the observed data. For synthetic studies, the observed data will have been synthesized from the test[22] model $\mathbf{m}^{\dagger}$, and in that case we can write $\mathbf{d}^{\dagger} = \mathbf{F}(\mathbf{m}^{\dagger})$. The regularization term $\mathscr{R}_i$, which can change as iterations progress as indicated by subscript i, will be discussed below. For completeness, if we choose an ℓ^2 norm for

[21] Also called *Taylor's formula with remainder* or the *mean-value theorem* when truncated to first order. See, e.g., [46]. At second order, it states that $f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T\mathbf{h} + \frac{1}{2}\mathbf{h}^T\nabla^2 f(\mathbf{x}+t\mathbf{h})\mathbf{h}$, where $t \in (0, 1)$.

[22] Sometimes also called *true* model, or *target* model.

the misfit of our FWI problem, we would write

$$\mathscr{J}(\mathbf{m_i}) = \frac{1}{2}\sum_{k}^{N_s}\sum_{r}^{N_r(k)}\int_0^T \left|\mathbf{P}_r \cdot \mathbf{u}^{(k)}(\mathbf{m_i}, t) - u_r^{\dagger(k)}(t)\right|_2^2 \mathrm{d}t + \mathscr{R}_i(\mathbf{m}_i), \tag{39}$$

where N_s is the number of sources, $N_r(k)$ is the number of receivers for source k, $\mathbf{P}_r$ is a projection operator extracting data at receiver r, $u_r^{\dagger(k)}(t)$ is the data measured at receiver r for source k, and $\mathbf{u}^{(k)}(\mathbf{m_i}, t) = \mathbf{A}(\mathbf{m_i})\mathbf{s}^{(k)}(t)$ is the simulated data as defined by Equation (12). Notice that time was left continuous only to be consistent with the formulation of Section 2, and notational convenience.

When the Newton search direction is used to generate the ith model update $\Delta\mathbf{m_i}$, it is obtained from

$$\Delta\mathbf{m_i} = -\alpha_i \mathbf{H_i}^{-1}\mathfrak{g}_i, \tag{40}$$

which involves the gradient $\mathfrak{g}_i = \nabla_{\mathbf{m}}\mathscr{J}(\mathbf{m_i})$, the Hessian matrix $\mathbf{H_i} = \nabla^2_{\mathbf{m}}\mathscr{J}(\mathbf{m_i})$ (or rather its inverse), and step length α_i. The model iterates are obtained from the cumulative changes made to the initial model, i.e., $\mathbf{m_i} = \mathbf{m_o} + \sum_{j=1}^{i}\Delta\mathbf{m_j}$. For practical problems, the full Hessian $\mathbf{H_i} = \mathbf{H}(\mathbf{m_i})$ cannot be stored in memory and one has to resort to a menagerie of approximations. These include the Gauss-Newton method, which involves additional forward and adjoint computations for the benefit of more accurate Hessian approximations, and quasi-Newton methods, such as the so-called BFGS[23] method which gathers second-order information through measuring the changes in gradients, or most likely its limited-memory version l-BFGS [74]. One could also use Hessian-free methods such as one of the variants of nonlinear conjugate-gradient methods. Comparing the costs and benefits of these methods is not an easy task and comparative studies have shown that performance can also depend on the specific problem at hand [16, 68, 72]. It is not the goal of this paper to discuss the merits of each available algorithm. However, we would like to emphasize the need for a better understanding of the conditions under which each method could have benefits. Also, we cannot overstate the fact that the availability of faster algorithms would have a major impact on the range and size of problems that could be solved. Many of the algorithms presented here follow a one-dimensional descent direction and therefore are often deterministic in nature. Having algorithms that could sense larger volumes of the solution space either through statistical or stochastic machinery could help improve the convergence speed.

[23] Named from its inventors Broyden, Fletcher, Goldfarb, and Shanno.

3.5 *Misfit Functional*

Recall that the objective functional $\mathscr{J}(\mathbf{m})$ includes some measure of misfit between time series of oscillatory signals. These functionals are notorious for having multiple minima, each forming a rather narrow neighborhood of convergence whose width depends directly on the frequency of the signal. Nevertheless, one particular choice of misfit functional, the ℓ^2 norm, offers many mathematical and numerical advantages such as smoothness and ease of computation.[24] Not surprisingly, the original formulation of FWI was derived using this norm. It has been widely used in synthetic studies, where noise is well controlled, thus avoiding its known oversensitivity to outliers.

With the desire of improving robustness of the inversion, many different misfit measures have been proposed over the last few decades. Covering the entirety of the topic is beyond the scope of this paper, and only a few will be mentioned as examples.

By construction, the ℓ^p norms for the misfit of time series implies that the amplitude of the signal is compared at each time step. If the measured signal is reduced by attenuation and amplitude can only be restored approximately, a normalized ℓ^2 norm, which multiplies signal amplitudes instead of subtracting them, is more appropriate. The authors in [87] proposed such a norm for streamer data. This norm provides more robustness to amplitude mismatch but is subject to the same multiple-minima problems that the ℓ^p misfit (subtraction) norm is subject to.

Another approach is to relate FWI to the old-standing mathematical problem of optimal transport. This approach essentially poses the problem as finding a transformation to the model $\mathbf{m}$ that would shift material properties so that the resulting model would better match the data. First proposed by Engquist and Frosse[29], Engquist et al. [30] for FWI, this so-called Vešerstěin metric[25] is given a more detailed description in a joint article of this book [69]. While the theory is very elegant, there is no physical basis on which the elastic parameters of the earth should follow such a conservation law, and this constraint is often relaxed.

As can be seen, defining a robust misfit function for FWI is currently an area of active research [109]. These activities are mostly performed through physical intuition, and trial and error. There is very limited unifying mathematical theory that can guide these investigations. Having a machinery to characterize the response hypersurface of a given objective functional could provide a means of comparing the benefits of different misfit functionals, and guide the practitioner in devising better ones.

[24] Including the ability to use Gauss-Newton methods for the inversion.

[25] Also called *Wasserstein metric*, or more descriptively *earth mover's distance*.

3.6 Regularization

One form of ill-posedness for FWI problems comes from having some elements of the test model $\mathbf{m}^\dagger$ possibly not playing any role in creating the data $\mathbf{d}^\dagger$ (i.e., the so-called null space of the inverse problem). For example, sampling a model in which the data acquisition geometry would leave a fraction of the model in the dark (i.e., not illuminated) would make these zones unresolvable. Another form of ill-posedness results from inverting a model that has been discretized too finely in space in comparison to the wavelength of the interrogating signal.

Often, both these problems get addressed by using a regularization term $\mathscr{R}_i(\mathbf{m})$ in Equation (37), which typically introduces a penalty on the spatial gradient of the model. The family of functionals defined by

$$\mathscr{R}^{(0)}(\mathbf{m_i};\beta_i) = \beta_i \int_\Omega k(\mathbf{m_i},\mathbf{x})\ \left|\mathbf{m_i}-\mathbf{m_g}\right|_2^2\, d\mathbf{x}, \tag{41}$$

where $\mathbf{m_g}$ is some given model, and

$$\mathscr{R}^{(1)}(\mathbf{m_i};\beta_i) = \beta_i \int_\Omega k(\mathbf{m_i},\mathbf{x})\ \left|\nabla_{\mathbf{x}}\mathbf{m_i}\right|\, d\mathbf{x}, \tag{42}$$

$$\mathscr{R}^{(2)}(\mathbf{m_i};\beta_i) = \beta_i \int_\Omega k(\mathbf{m_i},\mathbf{x})\ \left|\nabla_{\mathbf{x}}\mathbf{m_i}\right|_2^2\, d\mathbf{x}, \tag{43}$$

are commonly used for that purpose.[26] These functionals are given different names across the scientific literature.[27] The value of scalar β_i is critical as it needs to be large enough to dampen the free modes, and also possibly help jump over small local minima, but not so large that it biases the final solution. The sweet spot is often determined from experimentation rather than by theory: after solving several cases for $\mathscr{J}(\mathbf{m};\beta)$ for different values of β, a value near the corner of the generated so-called L-curve [40] is chosen as a good candidate for β_o.[28] To avoid biasing the final answer, it is sometimes beneficial to use a continuation strategy reducing β_i as inversion iterations are progressing, unless the problem contains unstable modes that require steady dampening. Parameter $k(\mathbf{m_i},\mathbf{x})$ is a weighting factor related to the noise and/or resolution of the model. When regularization terms as Equations (41)–(43) are used for image processing and denoising, k is set to unity as signal-to-noise ratio is considered constant throughout the image. This is clearly not the case for FWI, and penalty should be adjusted to the local length scale and noise level.

When degeneracies are caused by a discretization that is too fine for the ultimate achievable resolution, numerical methods using an unstructured mesh can better solve these problems by carefully adjusting the local mesh size as a function of

[26] Note that $\mathscr{R}^{(1)}$ is not differentiable with respect to $\mathbf{m}$. However, some differentiable approximations can be used. See [106] for more detail.

[27] For example, nth-order *Tikhonov*, *Tikhonov-Miller*, *Phillips-Twomey*, *Total Variation*, etc.

[28] These curves are log-log plots generated by plotting the misfit value $\left|\mathbf{F}(\mathbf{m_n})-\mathbf{d}^\dagger\right|$ as a function of the value of the residual $\mathscr{R}(\mathbf{m_n};\beta)/\beta$ at the "final" nth iteration obtained with different values of β.

the local wavelength. This approach achieves regularization through discretization and avoids adding a penalty term such as Equations (41)–(43) to the objective functional. Many other regularization terms are also possible depending on the additional insight that one chooses to impose on the problem. For example, inspired by lessons learned from sparsity methods, $\left|\hat{\mathbf{T}}(\mathbf{m_i})\right|_1$, where the model is transformed by $\hat{\mathbf{T}}$ to some other domain representation (Fourier, wavelet, Laplace, etc.), has often been used as a way to guide the model towards sparser solutions, and drive the deaf parts of the model towards zero.

Methods derived from Equations (41)–(43) are attractive due to the locality of the discrete operators involved: they add minimal numerical complexity to the inverse problem while providing some stability. The weights of the penalty have to be carefully chosen as to avoid creating artifacts in the solution. Their proper application requires some experimentation that makes the art of practitioners often look more magical than mathematical. A unifying mathematical theory could definitely help practitioners design the most effective approach.

3.7 Parameterization

Scientists having performed nonlinear data fitting know that the choice of parameters often alters the feasibility and the speed of convergence of the minimization [20, 83]. For example, consider acoustic waves propagating in constant-density media, where only a single scalar field is involved as parameter in the equation of motion (as the mass density is approximated to be independent of space and time). The resulting scalar equation can be written as

$$\nabla^2 u(\mathbf{x}, t) - \frac{\varrho}{K} \frac{\partial^2 u(\mathbf{x}, t)}{\partial t^2} = s(\mathbf{x}, t), \tag{44}$$

where u now represents the fluid pressure and s is the source. Despite having a single nonconstant scalar-parameter field $K(\mathbf{x})$, there are a few ways this equation could have been written. Depending on the scientific discipline, one can use, as we did, the (isentropic) bulk modulus of the fluid, $K = \rho(\partial P/\partial\rho)_s$, where the derivative is taken at fixed entropy S. One could also have used, e.g., the wave velocity, $v_p = \sqrt{K/\varrho}$. These choices would not affect the results obtained for the forward problem. For the inversion, however, it would. Performing FWI on the Marmousi test model [12], some authors reported [22] that the speed of convergence obtained while using steepest descent with either v_p or v_p^2 as the inversion parameter made a significant difference on the reduction of misfit per inversion iteration. This behavior naturally raises the question of finding the optimal parameterization for a given inverse problem. The answer could depend on the nature of the cost function, the exact problem being solved, the acquisition geometry, and the algorithm that is being used to drive the minimization. Moreover, other derived parameters, such as

the wave velocity to the nth power, v_p^n, for example, can also be used instead of the original control parameter via a simple derivative chain rule.

We will first discuss parameterization in the context of single-parameter inversions and then address the additional challenges brought by multiparameter inversions.[29] For each inversion parameter, the practitioner faces two distinct choices regarding the parameter to invert: a parametric representation, which can be a nonlinear transformation of the originally chosen parameter in the PDE (e.g., v_p vs. v_p^2), and a choice of units, which is sometimes called *parameter scaling* in the optimization literature [74], and involves only a linear transformation. The latter is not that critical for single-parameter inversions: we defer its discussion to Section 4.1 related to multiparameter inversions. The choice of parametric representation, however, can be critical as reported by Collis et al. [22].

There are a few hypotheses one can make at this point regarding the best inversion parameter to use for performing FWI using Equation (44). One could argue that K^{-1} provides a linear coefficient to the second term of Equation (44). On the other hand, Tarantola suggested [96] that one should choose the logarithm of the parameters such as $\tilde{K} = \log(K/K_o)$, where K_o is some reference value. This parameterization has the benefit of making parameters so-called Jeffreys' invariants in the sense that a distance function $d()$ between two values K_1 and K_2 becomes insensitive to change in units as $d(\tilde{K}_1, \tilde{K}_2) = |\log(K_1/K_o) - \log(K_2/K_o)|$. Inspired by the Green's function for waves in homogeneous media, one could also suggest $e^{K^{-1}}$ as a good choice for making the problem "more linear." And finally, there are those who continue to believe that having a Hessian in Equation (40) takes care of all scaling issues. While this is the case for linear scaling, we will demonstrate that using the Hessian does not solve the problem for nonlinear scaling.

An alternative and interesting approach is to use a local measure for nonlinearity as proposed by Hofmann [42, 43], and find a new parameterization that reduces its value. Simply stated, we are looking for reducing a bound on the norm of the difference between the actual value of a nonlinear mapping $\mathbf{F}(\mathbf{m}+\mathbf{h})$ and the one linearly predicted by using the first derivative (Fréchet) $\mathbf{F}'(\mathbf{m})$ near $\mathbf{m}$. Using Equation (38), and assuming a degree of nonlinearity c of mapping $\mathbf{F}(\mathbf{m})$ about model $\mathbf{m}$, we define the nonlinearity index $\mathscr{K}$ as the smallest positive real number such that

$$\left|(\mathbf{F}(\mathbf{m}+\mathbf{h}) - \mathbf{F}(\mathbf{m})) - h\mathbf{F}'(\mathbf{m})\right| \leq \mathscr{K} h^{2c}, \text{ for all } \mathbf{h} \in \mathbf{B}_R, \tag{45}$$

where $\mathbf{B}_R$ is a hypersphere of radius R centered at the origin, and $h = |\mathbf{h}|$ is a scalar norm. Using this definition, we devise a numerical method that can estimate the nonlinearity index from

[29] By multiparameter, we refer to systems being described by multiple, distinct, spatially varying physical parameters such as density, bulk, and shear moduli for a three-parameter inversion.

$$\mathscr{K} = \max_{\mathbf{h}\in\mathbf{B}_R} \frac{\left|(\mathbf{F}(\mathbf{m}+\mathbf{h}) - \mathbf{F}(\mathbf{m})) - h\mathbf{F}'(\mathbf{m})\right|}{h^{2c}}. \tag{46}$$

The degree of nonlinearity c can be determined by looking at the leading orders of the Fréchet derivatives and should remain unaffected by linear operators acting on $\mathbf{m}$ and $\mathbf{d}$.

This approach provides a few advantages: first, the effects of the model can be taken into account, and second, it can provide a quantitative measure for establishing meaningful comparisons between different parameterizations. In order to proceed however, the problem has to be greatly simplified to be solved on a computer. In [84], we propose transforming both the data $\mathbf{d}^\dagger$ and the model $\mathbf{m}$ using linear transformations that reduce the computational complexity of the problem while minimally changing the nonlinearity of the system. For that purpose, the data are Fourier transformed and only a subset (25) of the frequencies are used. Similarly, the models are decomposed into a small number (10) of Haar wavelets, as these can preserve the sharp-contrast character of the original models. A standard least-squares linear regression is used to fit this reduced system for a range of models near the test model $\mathbf{m}^\dagger$ and near some initial model $\mathbf{m_0}$. We then use the residual of the least-square regression as a representation of the numerator of Equation (46), thus estimating $\mathscr{K}$. Notice that this approach only requires access to a forward engine; no gradient computations are required. Results are shown in Figure 4. When these predictions are compared with actual inversions, the convergence speeds are in excellent agreement. This is shown in Figure 5, and these results confirm that this method can be used to compare parameterizations and help find the most effective one. It is interesting to note that one of the two most effective parameterizations corresponds to the formulation that provides a linear coefficient in Equation (44).

3.8 *Initial Model*

After reading the previous section, attentive readers might have a question on their mind: as the initial model in synthetic studies is derived from the test model using $\mathbf{m_0} = \hat{\mathbf{S}}_{\lambda_o}(\mathbf{m}^\dagger)$ where $\hat{\mathbf{S}}_{\lambda_\mathbf{o}}$ is a (possibly linear) smoothing operator such as a low-pass Butterworth filter, isn't the problem made much easier? In particular, if signal and space (i.e., data and model) are coarsened to order ω_o and λ_o, respectively, how close are low-pass filtered data

$$\mathbf{d}_{\omega_\mathbf{o}} = \hat{\mathbf{S}}_{\omega_o}(\mathbf{d}^\dagger) \tag{47}$$

to data synthesized from a smoothed test model

$$\mathbf{d_o} = \mathbf{F}_{\omega_\mathbf{o}}(\mathbf{m_0}) = \mathbf{F}_{\omega_o}(\hat{\mathbf{S}}_{\lambda_o}(\mathbf{m}^\dagger)), \tag{48}$$

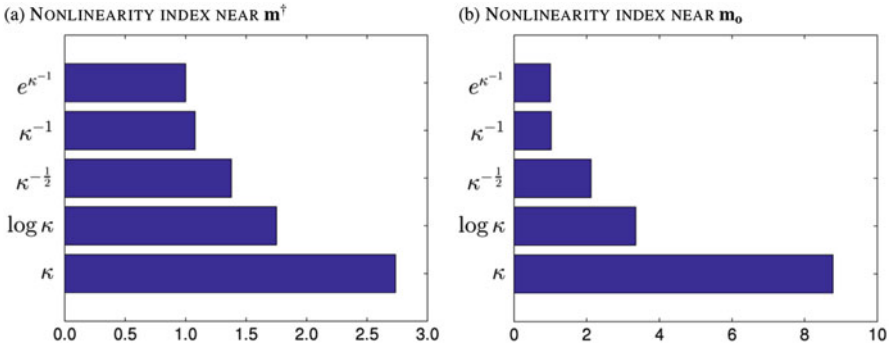

Fig. 4 Least-squares regression analysis of a single-parameter inversion. The test model $\mathbf{m}^\dagger$ used is the 2-D Marmousi model [12] with an initial $\mathbf{m_o} = \hat{\mathbf{S}}_{\boldsymbol{\lambda}_o}(\mathbf{m}^\dagger)$ where $\hat{\mathbf{S}}_{\boldsymbol{\lambda}_o}$ is a model smoothing operator. The value of nonlinearity index $\mathscr{K}$ is shown for predicting the effectiveness of different parameterizations (a) near the test model $\mathbf{m}^\dagger$, and (b) near the $\mathbf{m_o}$ initial model. A smaller index is better

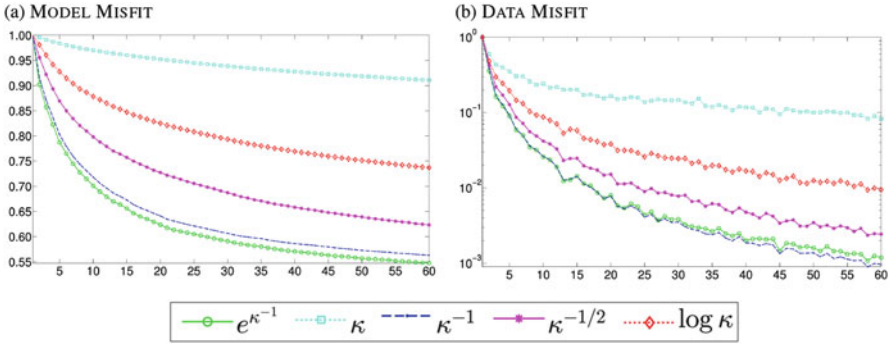

Fig. 5 Misfit reduction for FWI inversions involving different parameterizations. Panel (a) shows the reduction of the normalized model misfit, while (b) shows the reduction of the normalized data misfit as a function of Gauss-Newton iterations. Simultaneous-source encoding algorithm was used but the same encoding sequence was shared by all cases. Parameterizations involving K^{-1} perform best, while K is the worst choice amongst those shown here. See [84] for details

for well-selected values of ω_o and λ_o? Here, $\mathbf{F}_{\omega_o}(\mathbf{m_o})$ represents a model-data nonlinear mapping (including forward PDE operator Equation (12)) that has source temporal signatures filtered with operator $\hat{\mathbf{S}}_{\omega_o}(s(t))$. This problem is well known to geophysicists who devised very creative methods to extract the ultralow frequencies of the model from the data. This so-called background velocity model is critical to the success of the inversion and has received considerable attention. Guided by physical intuition, these approaches have allowed FWI practitioners to expand the envelope of convergence for the inversion of a limited set of 2-D synthetic models under very similar acquisition geometries. The key issue, however, is that there are very few rigorous nonlinear-system theories that can guide the development of more

robust methods. Homogenization theories, for example, play a role in how $\mathbf{F}_{\omega_\mathbf{o}}$, $\hat{\mathbf{S}}_{\lambda_o}$, and $\hat{\mathbf{S}}_{\omega_o}$ are related. Understanding the commutator between operators $\mathbf{F}$ and $\hat{\mathbf{S}}_{\lambda_o}$ as well as the roughness of the response hypersurface of objective functional $\mathscr{J}(\mathbf{m_i})$ could help design better algorithms. Other very interesting approaches for addressing ill-posedness and understand conditions for convergence have been proposed in [19].

As one particular example, authors in [57] use information about the spectral distribution of the wavenumbers of material properties in the earth to filter both the data and the source wavelet signatures. This so-called spectral-shaping method involves filtering the measured data (and source wavelets) using Equation (47) as a proxy to condition the model. The extension of this approach to elastic data will require a better characterization of the commutator between data filters and elastic data-model nonlinear operators.

Another approach that offers more control over the starting model is to include a distance penalty between the model at the current iteration and a given model, $\mathbf{m_g}$. More precisely, a term $\beta_i \left|\mathbf{m_i} - \mathbf{m_g}\right|$ is added to $\mathscr{J}_i(\mathbf{m_i})$ and allows one to control the prior information bias through adjusting the value of β_i. This is the regularization term in Equation (41). A different model $\mathbf{m_o}$ can then be used to bootstrap the series of iterates.

The importance of finding good starting models has been expressed multiple times by practitioners of FWI. At first glance, it might seem natural to apply more conventional and demonstrated methods to the observed data $\mathbf{d}^\dagger$ and obtain a conventional result that can then be used as a starting model for FWI. This approach, however, implicitly assumes that the only available information is already contained in the data. Therefore, a properly crafted annealing strategy could, in principle, achieve the same result. Unless the initial model building procedure uses additional data, the quest of obtaining a "good" starting model from the same data is a moot point. Moreover, the notion of "good" starting model depends on the neighborhood of convergence of the minimization procedure which, in turn, depends on the acquisition geometry, the model at hand, and the overall inversion strategy. The step of building a starting model becomes equivalent to designing a robust workflow for the inversion strategy. This conclusion naturally brings us to the next topic.

3.9 Annealing Strategies

In reflection seismology, seismic sources and receivers are typically both located at the earth surface. As a consequence, shallow reflections will arrive at the receivers before deeper ones. This sequential order creates a natural time hierarchy motivating time continuation strategies. In these approaches, the seismograms used at the i^{th} iteration are truncated at T_i such that $0 < t < T_i < T$, and $T_i < T_j$, for $i < j$. By introducing the data gradually, this ensures that the shallow part of the earth gets

resolved before inverting the deeper part. Obviously, this could also be achieved by having a moving masking filter only leaving a part of the gradient active, and sweeping forward akin to the wave propagation front. But, truncating the data is computationally much simpler and more efficient, particularly in the time domain.

A second continuation strategy is based around a hierarchy of length scales. Such multi-scale strategies were proposed early on in FWI [15] as an effective way to mitigate the nonuniqueness of the problem. The approach is based on the premise that the radius of convergence around $\mathbf{m_i}$ can be controlled by the frequency of the source (see Section 3.5). The easiest and most common way to implement a multi-scale strategy in the space-time domain is to use the linear character of the wave equation with respect to source and displacement, and filter both data and source time signatures with a low-pass filter. In the frequency-time domain, multi-scale strategies come naturally by solving in increasing order of frequency and using the output of one inversion as the initial model of the next higher frequency. These continuation strategies are sometimes referred to as frequency sweeping.

Another continuation strategy revolves around the separation of reflected and transmitted waves. As the sound velocity in the earth tends to increase as a function of depth (a result of the increasing pressure and compaction caused by the weight of the overburden), some of the waves traveling down at an angle will slowly curve and come back at the surface. As a result, these refracted waves reach receivers at larger distances from the source (i.e., larger offsets). Through their transmission path, these waves contain information about some long-range average of the earth rather than information on the specific location of reflecting interfaces. A few continuation strategies start the inversion using the large-offset information and then gradually include the shorter offset data. Many of these strategies include the name *tomography* in their description, referring to the exploitation of the information contained in refracted waves separately from those having information about reflections. These strategies can also be expanded to account for the directivity of the displacements when this type of data is available.[30]

When more than a single parameter is inverted (e.g., v_p, v_s, ϱ, δ, ε, Q_p, Q_s, etc.), continuation strategies involving the selection of a subset of the physical parameters to invert for can also be applied. This can be referred to as physics continuation as the introduction of parameters will be performed such as to improve the physical fidelity of the model as the inversion progresses. Moreover, with each physical parameter can be associated a different length-scale hierarchy and relative weight, thus opening a wide range of possibilities. In the next section, we demonstrate the complexity of multiparameter inversions with a few synthetic examples.

[30] Some receivers (termed multicomponent) can provide vectorial information on the displacement that can be exploited through a continuation strategy.

4 Inversions

In this section, we will briefly present selected 2-D synthetic cases that exemplify some of the concepts presented in the previous sections. Our goal is to highlight the current challenges encountered in FWI and the choices that practitioners have to make, rather than to propose solutions. We believe that behind these challenges lie many research opportunities involving an intimate collaboration between mathematicians, physicists, and computational scientists.

Because the real subsurface is only known through outcrops, e.g., cross-sections of the earth exposed on a cliff, or well logs, i.e., analysis of rock properties near the walls of a drilled bore hole, inversion feasibility studies need to first rely on synthetic studies. In these studies, representative earth models are built on a computer and synthetic data are generated at given numerical and physical accuracies. Then, these data, possibly combined with a priori assumptions, are used to derive a model which can be compared with the original one. One needs to be aware, however, that those synthetic models do not always reproduce the subtleties of actual data, e.g., the length-scale hierarchy that the real earth contains. Therefore, inversion success on a synthetic model is not always a guarantee of success for a real case. This statement is especially true when the physical fidelity of the synthetic model is poor in comparison to real systems, as can be the case when using the acoustic approximation. Also, many studies (including this one) rely on 2-D models. Those systems have significantly lower computational costs but do not capture additional phenomena present in 3-D surveys. While not being a sufficient condition, controlled synthetic studies are often a necessary condition for demonstrating success. Moreover, they offer additional benefits to real surveys: they are relatively less expensive, and acquisition geometry and ambient noise can be more easily controlled.

In order to reduce numerical effects in synthetic studies, it is generally advisable to use two discretization methods: one for generating accurate synthetic seismograms, and a different one, usually faster and less accurate, for driving the inverse problem. Too often, however, the amount of additional effort that this entails can lead practitioners to using a single method. The practice of using the same numerical method for both generating the synthetic data and performing the inversion is referred to as an *inverse crime* [23], which authors go on to state that "*it is crucial that the synthetic data be obtained by a forward solver which has no connection to the inverse solver.*" Consequently, in the examples presented in this paper, data were typically generated using a higher-accuracy method (discontinuous Galerkin [41] on a very fine mesh) while the inversions were performed using a more economical representation (spectral elements on a fit-for-purpose mesh).

4.1 Multiparameter Inversion

The case of multiparameter inversion is more complex for several reasons. First, one has to ensure that all inversion parameters are properly scaled and parameterized so that the valley about the sought local minimum has a similar aspect ratio along all dimensions of the inversion parameters, and that nonlinear effects are reduced. Second, the set of inversion parameters must be carefully chosen so that cross-talk[31] between the parameters, i.e., the possibility that changes in two distinct parameters produce the same effects in the data, is reduced to a minimum. Finally, the series of inversions must be designed as to introduce the physical parameters in sequences that are most favorable to successful results (i.e., multi-scale and physics continuations).

Choosing an effective parameterization for multiparameter inversions can be guided using the method described in Section 3.7. In [84], we applied this approach to the so-called elastic Marmousi test model [65] and found that a good parameterization can reduce the number of iterations required to reach a given misfit value by up to an order of magnitude. However, one must also ensure that the selected parameters have a minimum amount of cross-talk between them. One simple qualitative way to detect the cross-talk between parameters is to perform inversions using test models that had their values slightly modified by overlaying some simple geometric perturbations, such as simple shapes (e.g., squares or triangles) or regular checkerboard patterns over the whole domain. By carefully designing different patterns for each parameter, the possible leakage of one parameter into the other becomes apparent in the inverted models.

Another slightly more quantitative way to compare different parameterizations is to compute the scattering patterns caused by a Gaussian perturbation at a fixed point to a homogeneous-medium model (e.g., [110]). By visualizing and comparing the radiation patterns obtained from perturbations for each parameter at the fixed point, the practitioner selects the parameterization that has the least geometrical overlap between the patterns for desired angles of incidence and reflection. This approach has been used for a broad variety of physical fidelities (e.g., [4, 80]). For example, with only two parameters ϱ and v_p, some studies [77] suggest that using the impedance $I_p = \varrho v_p$, and v_p gives a parameterization less prone to cross-talk under certain acquisition geometries. As these studies use scattering effects of a single point in a homogeneous model, the additional effects of the model being inverted is not taken into account. In both approaches presented here, the acquisition geometry has to be carefully reproduced, as the parameter sensitivities depend on the scattering angle. Finally, it is worth mentioning that having access to the Hessian can be very helpful in understanding the cross-talk between the parameters. We refer the readers to [77] for a nice tutorial on that topic.

[31] Term borrowed from telecommunications describing signals transferring from one channel to another due to unintentional coupling (e.g., poor electromagnetic insulation).

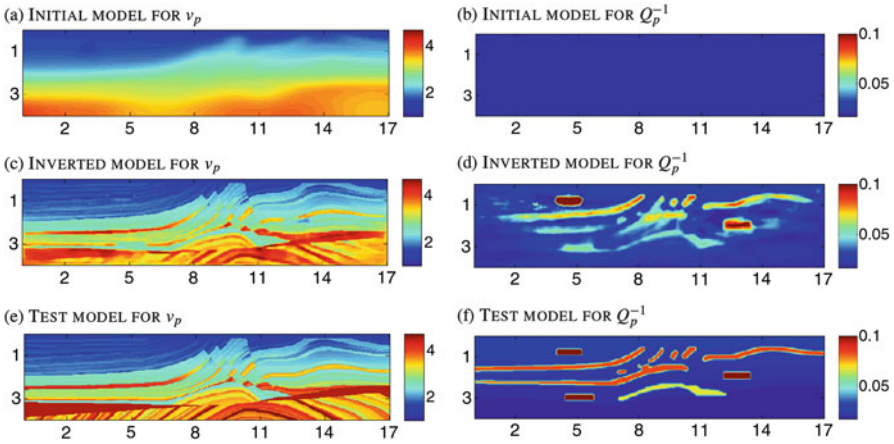

Fig. 6 Inversion of viscoacoustic media. The test model $\mathbf{m}^\dagger$ is composed of (e) a compressional velocity model [12] (v_p in km/s), and (f) an attenuation model (Q_p^{-1}, dimensionless), from which data $\mathbf{d}^\dagger$ was generated using a high-accuracy forward model $\mathbf{d}^\dagger = \mathbf{F}(\mathbf{m}^\dagger)$. Data consist of traces recorded at the free surface (top) of the model where the sources were also located. Other sides have absorbing boundary conditions. Starting from initial model $\mathbf{m_o} = \hat{\mathbf{S}}_{\boldsymbol{\lambda}_\mathbf{o}(\mathbf{m}^\dagger)}$ composed of (a) compressional velocity, and (b) a constant quality factor Q_p model, the system is inverted using a Gauss-Newton algorithm and some of the continuation strategies (time, frequency, and length scale) described in Section 3.9. Results are shown in panels (c) and (d). System dimensions are in km. The source time function used was a 10-Hz Ricker wavelet. More details can be found in [25]

The methods just described for detecting cross-talk between parameters are heuristic; rigorous, quantitative mathematical methods aimed at orthogonalizing the parameters have yet to come. And if the state of orthogonality changes as the inversion progresses, one can already envision workflows that include continuation strategies involving regular re-parameterization [24].

4.1.1 Inversions of Viscoacoustic Media

We now present an inversion example involving a viscoacoustic medium, i.e., an acoustic medium that contains heterogeneous attenuation. This medium can be described by two parameters, e.g., v_p and Q_p. Figure 6 shows the test model $\mathbf{m}^\dagger$ and the results obtained using a Gauss-Newton optimization algorithm and an ℓ^2 norm for the misfit functional. This study was performed using time-domain algorithms with three different rheological mechanisms used for representing the attenuation (see Appendix).More details on the inversion can be found in the caption.

This and other studies have demonstrated that heterogeneous attenuation can be inverted at a coarse scale provided that the acquisition geometry includes some transmitted information. In reflection seismology, this will be the case if the survey includes large offsets, allowing refracted waves to bring information to wide-offset

receivers. This behavior is apparent from the half-moon shape of the front where the resolution is highly degraded. Also, notice the three rectangles of the Q_p models showing in the inverted v_p model, resulting from the high values of attenuation in these regions that completely absorb the signal.

This test model also showed us that despite our best efforts, we could not successfully invert the model unless attenuation is explicitly taken into account in the inversion [25]. In reality, this strongly heterogeneous attenuation case is limited to zones of the subsurface where gas is often present, and this situation is well known to experimental geophysicists. The continuation strategies used for inverting the data were devised by trial and error, and the practitioner's intuition. We believe that a more unifying theory including both the physics of wave propagation and the details of the optimization would be of great help for devising inversion strategies.

4.1.2 Inversions of Anisotropic Elastic Media

We now turn to the inversion of a 4-parameter anisotropic test model derived from [33]. The inversion has been parameterized with vertical compressional velocity $v_p^{\uparrow}$, vertical shear velocity $v_s^{\uparrow}$, and anisotropy parameters δ and ε (see Equation (5)). The initial model is derived from the test model using a smoothing operator that leaves a smooth background, $\mathbf{m_0} = \hat{\mathbf{S}}_{\boldsymbol{\lambda_0}}(\mathbf{m}^{\dagger})$, while the initial values for both δ and ε are zero. Data are composed of two components, i.e., displacements in the vertical and horizontal directions. Test model $\mathbf{m}^{\dagger}$ is shown in panel (a) of Figure 7. If a conventional time continuation strategy is used in concert with a multi-scale strategy on all parameters simultaneously, the inversion fails as shown in panel (b) of the same figure. However, if one uses a multi-scale strategy in which each of the parameters inverted is given an associated length scale, the results are significantly improved. Results are shown in panel (c) of the same figure. This demonstrates that using a simple data filtering technique to achieve a multi-scale strategy is not always adequate for inversions involving multiple parameters. In our case, a different inversion mesh was used for each parameter, thus conditioning the parameters separately.

In view of validating the method and detecting potential cross-talk, a simple test model was built using a smooth model (derived from [65]) and shown in Figure 8. Each parameter was marked with a perturbation in the form of a letter mimicking the parameter to be inverted. The same inversion strategy using separate annealing for each parameter was used and the results are shown in panel (b). The inversion resolves each of the parameters and shows decreasing resolution at larger depth, as anticipated from the larger wavelengths in these areas.

This case demonstrates that state-of-the-art inversions are still relying, in good part, on trial and error and intuition. Unfortunately, there is no guarantee that what works for a given model will also work for another one [72].

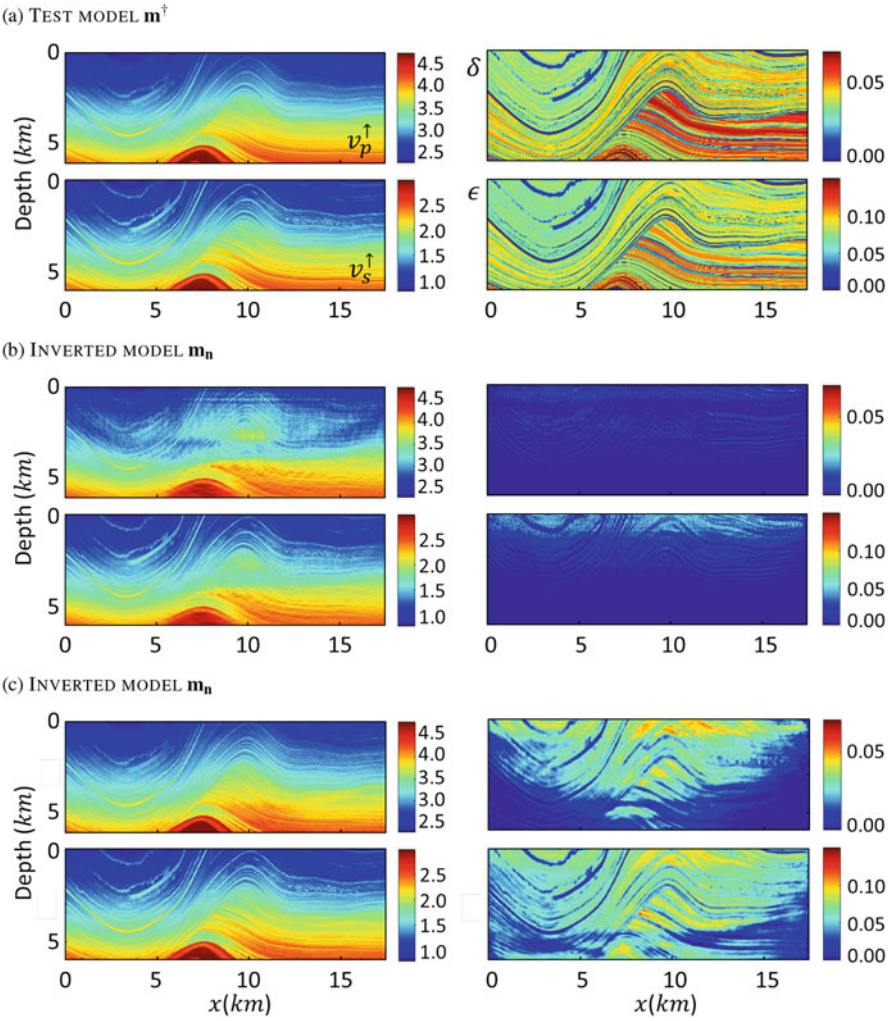

Fig. 7 Inversion of a 4-parameter anisotropic elastic medium characterized by the vertical compressional velocity $v_p^{\uparrow}$, the vertical shear-wave velocity $v_s^{\uparrow}$, and Thomsen parameters δ and ε. The test model $\mathbf{m}^\dagger$ was derived from [33] and is shown in panel (a). Panel (b) shows the results of an inversion that used time continuation and multi-scale continuation on all parameters simultaneously. In panel (c), a strategy that used different multi-scale continuation schemes on each parameter separately could resolve the parameters successfully. The starting models were the same for both inversions and were built from smoothed models $\mathbf{m_o} = \hat{\mathbf{S}}_{\boldsymbol{\lambda}_o}(\mathbf{m}^\dagger)$ for velocities and zero for δ and ε. Distances are in km, velocities in km/s, while mass density was held constant at $\varrho = 1$ kg/liter

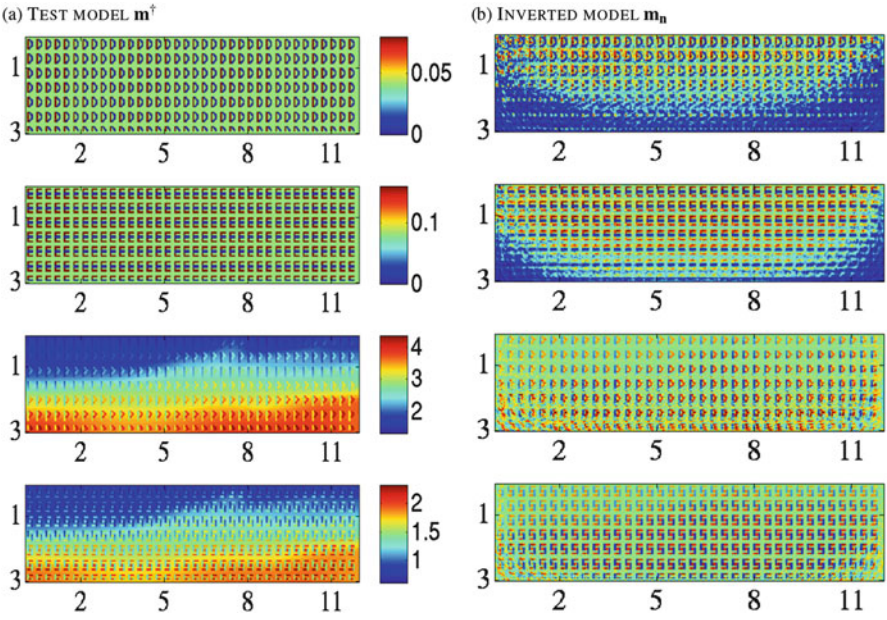

Fig. 8 Comparative inversion of 4-parameter anisotropic elastic model. Panel (a) shows test model $\mathbf{m}^\dagger$ that was built from a smooth background model (derived from [12]) and overlaid with perturbations describing the letters "D," "E," "P," and "S" for δ, ε, $v_p^{\uparrow}$, and $v_s^{\uparrow}$, respectively. Inversion results are shown in panel (b), where smooth background model has been subtracted so that only perturbations are showing. While spatial resolution is decreasing with depth, figure shows where parameters can be resolved by using the multi-scale strategies described in Figure 7. Distances are in km, and velocities in km/s

4.1.3 Source Inversion

The physics of seismic sources is complex as many nonlinear phenomena are taking place as energy is transferred into the solid. For controlled sources such as arrays of air guns used in marine surveys, service companies provide an estimate of the far-field source signature. These estimates are derived from phenomenological models and do not always account for source variability possibly taking place during the survey. For those reasons, it is sometimes desirable to estimate the source wavelet $w(t)$ (see Equation (9)) while performing FWI. Source-signature inversion can account for the variability in the source-media mechanical coupling and the source directivity between the different shots in the survey. Source estimation can also improve the physical fidelity of the model by casting finite-size effects of the real sources in the estimated wavelets triggered from a point source. Conversely, the modified source signatures can artificially account for the lack of physical fidelity and provide an additional avenue to explaining the observed data.

Estimating the source signature can be posed as a least-squares optimization problem. One approach is to assume that the model is known, and use a strategy that

alternates between source estimation and earth parameters estimation. The details of such synthetic inversions can be found in [2], in which the authors perform the joint inversion of source wavelets and material properties, while using a simultaneous-source encoding algorithm. Using combined encoded seismograms, multiple source signatures are inverted through the same iteration. Another approach is to optimize both the model and the source time functions at once. This approach has the benefit of allowing for adjusting the relative importance of the source inversion versus the model inversion. Successfully performing such inversions is impressive, given that the only information used is the locations of sources and receivers, and the observed traces.

Another approach called the *double-difference* method [112] consists in defining a misfit function built from the difference between the differences of observed and simulated signals at two different receivers. This approach, which is designed to cancel out the errors coming from possibly erroneous estimation of the material properties of the subsurface, can yield measurable benefits over existing methods, especially in land environments.

It is unclear at this time if additional dedicated receivers could provide an uplift in source-signature inversion. This question would bring us to the topic of experimental design, which we have not covered and can be the subject of an entire other paper.

5 Outlook

The progress made in computing hardware opened the possibility of solving a whole new class of FWI problems. The development of the novel algorithms necessary to solve these problems is equally important and received more attention during the recent years. As Bixby notes for linear programming, it is now more advantageous to use a 1970 computer with today's algorithms than the converse for solving these optimization problems [11]. This is not the case for FWI. However, many of the methods involving gradient-based optimization are dated and not always best suited for extreme-scale transient problems. We believe that our current needs will provide resources and purpose to revisit these algorithms, and possibly come up with a new generation of solution methods.

For applied mathematicians, novel theories are needed to better quantify and characterize the nonlinearities of the problem. A rigorous framework would help guide new algorithms and generate new insights. For computer scientists, new paradigms are needed to address the requirements of extreme-scale computing with very large amount of data, to manage the complexity of the workflows involved, and to provide better resiliency to large-scale computations. Designing new solutions will require an interdisciplinary approach involving computational scientists, computer scientists, mathematicians, and geophysicists [92].

In a very near future, the increase in availability of open-source software for FWI will likely provide an opportunity to accelerate the development of numerical tools

in addition to allowing a broader access to the technology. In particular, higher-level programming languages are emerging and offer the scalability and functionality required to explore problems in geophysics, medical imaging, and other parameter-estimation problems related to FWI. Emerging technologies such as improved machine learning and shape-based regularization techniques are promising new directions that can help improve the success of FWI.

Our goal was to provide a short and motivating introduction to FWI, and explain some of the current challenges that FWI practitioners are facing. The FWI problem is representative of many other PDE-constrained optimization problems for which we believe we are slowly but gradually getting closer to achieving practical and routine solutions. Being part of a generation of scientists that take part in solving a new class of problems is very exciting. We sincerely hope you become one of them.

Acknowledgements The authors thank ExxonMobil Research and Engineering Company for permission to publish this work. The authors would also like to thank Fadil Santosa and the Institute of Mathematics and its Applications for hosting the workshop where this work was presented. We would also like to thank Jeremy Brandman, Jerry Krebs, Anatoly Baumstein, and Dimitar Trenev for their insightful suggestions and comments on the original manuscript.

Appendix

We start the mathematical description of attenuation by considering a rheological model composed of springs and dashpots as shown in Figure 2. The effective modulus $c(\omega)$ of this mechanical model can be expressed as a function of auxiliary variables representing relaxation angular frequencies $\omega_l = 2\pi f_l$ and nondimensional anelastic coefficients a_l,

$$c(\omega) = c_u \left(1 - \sum_{l=1}^{n} \frac{a_l \omega_l}{\omega_l + \mathrm{i}\omega}\right), \tag{49}$$

where $\omega_l = \Delta c_l/\eta_l$, $a_l = \Delta c_l/c_u$ where the unrelaxed modulus $c_u = c(\omega \to \infty) = c_r + \sum_{l=1}^{n} \Delta c_l$, in contrast to the relaxed modulus $c_r = c(\omega \to 0)$. This only says that if one moves the system in Figure 2 very slowly, only spring c_r is felt as dashpots are relaxing and not transmitting force, while if one moves it very quickly all springs are fully active. Anything in between depends on the frequency according to Equation (49). This model will have an attenuation quality factor following the ratio of real and imaginary parts of the modulus [54, 71], leading to the following self-consistent relation

$$Q^{-1}(\omega) = \frac{\Im\left[c(\omega)\right]}{\Re\left[c(\omega)\right]} = \sum_{l=1}^{n} a_l \frac{\omega_l \omega + \omega_l^2 Q^{-1}(\omega)}{\omega_l^2 + \omega^2}. \tag{50}$$

The frequency dependence of $Q(\omega)$ is set by carefully picking values for ω_l and a_l. This task is usually achieved by sampling frequencies ω_l logarithmically in the band

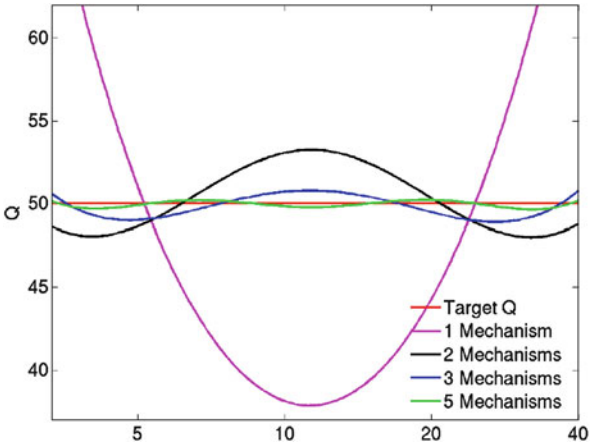

Fig. 9 Frequency response of quality factor $Q(f)$ for generalized Maxwell solids with 1, 2, 3, and 5 relaxation mechanisms over a frequency band ranging from 3 Hz to 40 Hz. The parameters of the relaxation mechanisms are optimized (least-squares) to mimic a constant target quality factor $Q = 50$ in the frequency band

of interest and fitting the anelastic coefficients a_l using a least-squares method [28, 37, 86]. In order to obtain a constant-Q attenuation, i.e., $Q(\omega) = Q_o$, we have shown [25] that at least three such relaxation mechanisms are required to obtain a response close to the desired behavior. Figure 9 shows the effect of using a different number of relaxation mechanisms on the frequency response $Q(f)$ of a generalized Maxwell solid. The parameters of the relaxation mechanisms are optimally tuned over a frequency band ranging from 3 Hz to 40 Hz in view of obtaining a constant target quality factor of $Q_o = 50$.

Each spring and dashpot added to the relaxation model introduce an additional anelastic function $\zeta_l(t)$ (sometimes called memory variable) that has to be solved as part of the governing equations. Each equation in (1) is then replaced by

$$\tau(e,t) = c_u e(t) - \sum_{l=1}^{n} a_l \zeta_l(t), \tag{51}$$

which are coupled to n additional equations for the anelastic functions,

$$\frac{d\zeta_l(t)}{dt} + \omega_l \zeta_l(t) = \omega_l e(t). \tag{52}$$

These equations are obtained after integrating the frequency-dependent moduli while maintaining causality (Boltzmann superposition principle). See [71] for details.

Because of this additional complexity, viscoelastic simulations are more costly by up to an order of magnitude, which can be reduced if special algorithms are used [26].

References

1. M. Ainsworth and H. A. Wajid, *Dispersive and dissipative behavior of the spectral element method*, SIAM Journal on Numerical Analysis, 47 (2009), pp. 3910–3937.
2. V. Akçelik, H. Denli, A. Kanevsky, K. K. Patel, L. White, and M.-D. Lacasse, *Multiparameter material model and source signature full waveform inversion*, in SEG Technical Program Expanded Abstracts, San Antonio, 2011, Society of Exploration Geophysics, p. 2406.
3. K. Aki and P. G. Richards, *Quantitative Seismology, Theory and Methods*, Freeman, San Francisco, 1980.
4. T. Alkalifah and R.-É. Plessix, *A recipe for practical full-waveform inversion in anisotropic media: An analytical parameter resolution study*, Geophysics, 79 (2014), p. R91.
5. J. E. Anderson, L. Tan, and D. Wang, *Time-reversal checkpointing methods for RTM and FWI*, Geophysics, 77 (2012), p. S93.
6. G. E. Backus, *Long-wave elastic anisotropy produced by horizontal layering*, J. Geophys. Res., 11 (1962), p. 4427.
7. R. Bansal, J. R. Krebs, P. Routh, S. Lee, J. E. Anderson, A. Baumstein, A. Mullur, S. Lazaratos, I. Chikichev, and D. McAdow, *Simultaneous-source full-wavefield inversion*, The Leading Edge, 32 (2013), p. 1100.
8. R. A. Bartlett, D. M. Gay, and E. T. Phipps, *Automatic differentiation of C++ codes for large-scale scientific computing*, in Computational Science – ICCS 2006, V. N. Alexandrov, G. D. van Albada, P. M. A. Sloot, and J. Dongarra, eds., Springer, 2006, pp. 525–532.
9. C. C. Bates, T. F. Gaskell, and R. B. Rice, *Geophysics in the Affair of Man: A Personalized History of exploration geophysics and its allied sciences of seismology and oceanography*, Pergamon Press, Oxford, 1982.
10. J. T. Betts and S. L. Campbell, *Discretize then optimize*, in Mathematics for industry: Challenger and Frontiers — A Process Review: Practice and Theory, D. R. Fergusson and T. J. Peters, eds., Society of Industrial and Applied Mathematics, Toronto, 2003, p. 140.
11. R. E. Bixby, *A brief history of linear and mixed-integer programming computation*, in Documenta Mathematica – Extra Volume ISMP, Berlin, 2012, 21st International Symposium on Mathematical Programming, pp. 107–121.
12. A. Bourgeois, P. Lailly, and R. Vesteeg, *The Marmousi model*, in The Marmousi experience, R. Versteeg and G. Grau, eds., Paris, 1991, IFP/Technip.
13. J. Brandman, H. Denli, and D. Trenev, *Introduction to PDE-constrained optimization in the oil and gas industry*, in Frontiers in PDE-Constrained Optimization, H. Antil, M.-D. Lacasse, D. Ridzal, and D. P. Kouri, eds., Berlin, 2017, Springer.
14. R. Brossier, L. Métivier, S. Operto, A. Ribodetti, and J. Vireux, *VTI acoustic equations: a first-order symmetrical PDE*, tech. report, 2013.
15. C. Bunks, F. M. Salek, S. Zaleski, and G. Chavent, *Multiscale seismic waveform inversion*, Geophysics, 60 (1995), p. 1457.
16. C. Burstedde and O. Ghattas, *Algorithmic strategies for full waveform inversion: 1D experiments*, Geophysics, 74 (2009), pp. WCC37–WCC46.
17. V. Cerveny, *Seismic Ray Theory*, Cambridge University Press, Cambridge, 2001.
18. G. Chavent, *Identification of functional parameters in partial differential equations*, in Identification of functional parameters in distributed systems, R. E. Goodson and M. Polis, eds., American Society of Mechanical Engineers, 1974, p. 31.
19. G. Chavent, *Nonlinear Least Squares for Inverse Problems*, Springer, Berlin, 2006.
20. J. Claerbout and D. Nichols, *Spectral preconditioning*, Stanford Exploration Project Report, 82 (1994), pp. 183–186.
21. R. Clapp, *Reverse-time migration: Saving the boundaries*, in SEP – 138, 2009, p. 29.
22. S. S. Collis, C. C. Ober, and B. G. van Bloemen Waanders, *Unstructured discontinuous Galerkin for seismic inversion*, in SEG Technical Program Expanded Abstracts, Denver, 2010, Society of Exploration Geophysics, p. 2767.

23. D. Colton and R. Kress, *Inverse acoustic and electromagnetic scattering theory*, Springer, New York, 3 ed., 2013.
24. D. Dagnino, V. Sallarès, and C. R. Ranero, *Scale- and parameter-adaptive model-based gradient pre-conditioner for elastic full-waveform inversion*, Geophysical Journal International, 198 (2014), p. 1130.
25. H. Denli, V. Akçelik, A. Kanevsky, D. Trenev, L. White, and M.-D. Lacasse, *Full-wavefield inversion of acoustic wave velocity and attenuation*, in SEG Technical Program Expanded Abstracts, Houston, 2013, Society of Exploration Geophysics, p. 980.
26. H. Denli and A. Kanevsky, *Fast viscoacoustic and viscoelastic full wavefield inversion*, Dec 2015, http://www.google.com/patents/US20150362622. US Patent App. 14/693,464.
27. M. Dumbser and M. Käser, *An arbitrary high-order discontinuous Galerkin method for elastic waves on unstructured meshes — ii. the three-dimensional isotropic case*, Geophys. J. Int., 167 (2006), p. 319.
28. H. Emmerich and M. Korn, *Incorporation of attenuation into time-domain computations of seismic wave fields*, Geophysics, 52 (1987), p. 1252.
29. B. Engquist and B. D. Frosse, *Application of the Wasserstein metric to seismic signals*, 2013. arXiv 1311.4581 [math-ph].
30. B. Engquist, B. D. Frosse, and Y. Yang, *Optimal transport for seismic full waveform inversion*, 2016. arXiv:1602.01540 [physics.geo-ph].
31. V. Étienne, E. Chaljub, J. Virieux, and N. Glinsky, *An h-p adaptive discontinuous Galerkin finite-element method for 3-D elastic wave modeling*, Geophys. J. Int., 183 (2010), p. 941.
32. P. M. Farrell, D. A. Ham, S. W. Funke, and M. E. Runkes, *Automated derivation of the adjoint of high-level transient finite element programs*, SIAM Journal of Scientific Computing, 35 (2013), p. C369.
33. M. Fehler and P. J. Keliher, *SEAM Phase I: Challenges of Subsalt Imaging in Tertiary Basins, with Emphasis on Deepwater Gulf of Mexico*, Society of Exploration Geophysicists, Tulsa, 2011.
34. A. Fichtner, *Full Seismic Waveform Modelling and Inversion*, Springer, Berlin, 2011.
35. W. I. Futterman, *Dispersive body waves*, J. Geophys. Res., 67 (1962), pp. 5279–5291.
36. O. Gauthier, J. Virieux, and A. Tarantola, *Two-dimensional nonlinear inversion of seismic waveforms: Numerical results*, Geophysics, 5 (1986), p. 1387.
37. R. W. Graves and S. M. Day, *Stability and accuracy analysis of coarse-grain viscoelastic simulations*, Bulletin Seismological Society of America, 93 (2003), p. 283.
38. A. Griewank and A. Walther, *Revolve: An implementation of checkpointing for the reverse or adjoint mode of computational differentiation*, Trans. Math. Software, 26 (2000), p. 19.
39. A. Griewank and A. Walther, *Evaluating Derivatives — Principles and Techniques of Algorithmic Differentiation*, Society of Industrial and Applied Mathematics, Philadelphia, second ed., 2008.
40. P. C. Hansen and D. P. O'Leary, *The use of the L-curve in the regularization of discrete ill-posed problems*, SIAM J. Sci. Comput., 14 (1993), p. 1487.
41. J. S. Hesthaven and T. Warburton, *Nodal Discontinuous Galerkin Methods*, Springer, Berlin, 2008.
42. B. Hofmann and O. Scherzer, *Factors influencing the ill-posedness on nonlinear problems*, Inverse Problems, 10 (1994), p. 1277.
43. B. Hofmann and M. Yamamoto, *On the interplay of source conditions and variational inequalities for nonlinear ill-posed problems*, Applicable Analysis, 89 (2010), p. 1705.
44. H. Igel, *Computational Seismology: A Practical Introduction*, Oxford University Press, Oxford, 2017.
45. M. Jakobsen and B. Ursin, *Full waveform inversion in the frequency domain using direct iterative t-matrix methods*, J. Geophys. Engineer., 12 (2015), p. 400.
46. W. Kaplan, *Advanced Calculus*, Addison Wesley, Reading, Massachusetts, second ed., 1973.

47. M. KÄSER, J. DE LA PUENTE, A.-A. GABRIEL, AND OTHER CONTRIBUTORS, SEISOL. http://www.seissol.org/, Retrieved March 1, 2018.
48. E. KJARTANSSON, *Constant Q-wave propagation and attenuation*, Journal of Geophysical Research, 84 (1979), p. 4737.
49. L. KNOPOFF, *Q*, Rev. Geophysics, 2 (1964), p. 625.
50. H. KOLSKY, *The propagation of stress pulses in viscoelastic solids*, Phys. Mag., 1 (1956), pp. 693–710.
51. D. KOMATITSCH, *Méthodes spectrales et éléments spectraux pour l'équation de l'élastodynamique 2D et 3D en milieu hétérogènes*, PhD thesis, Institut de Physique du Globe de Paris, France, 1997.
52. D. KOMATITSCH, J. TROMP, AND OTHER CONTRIBUTORS, SPECFEM3D. http://geodynamics.org/cig/software/specfem3d, Retrieved March 1, 2018.
53. J. R. KREBS, J. E. ANDERSON, D. HINKLEY, R. NEELAMANI, S. LEE, A. BAUMSTEIN, AND M.-D. LACASSE, *Fast full-wavefield seismic inversion using encoded sources*, Geophysics, 74 (2009), p. WCC177.
54. J. KRISTEK AND P. MOCZO, *Seismic wave propagation in viscoelastic media with material discontinuities — a 3D 4^{th}-order staggered-grid finite-difference modeling*, Bulletin Seismological Society of America, 93 (2003), p. 2273.
55. P. LAILLY, *The seismic inverse problem as a sequence of before-stack migrations*, in Conference on Inverse Scattering: Theory and Applications, J. B. Bednar, R. Redner, E. Robinson, and A. Weglein, eds., Philadelphia, 1983, Society of Industrial and Applied Mathematics, p. 206.
56. L. D. LANDAU AND E. M. LIFSHITZ, *Theory of Elasticity*, Pergamon, Oxford, 1959.
57. S. LAZARATOS, I. CHIKICHEV, AND Y. WANG, *Improving convergence rate of full wavefield inversion using spectral shaping*, in SEG Technical Program Expanded Abstracts, San Antonio, 2011, Society of Exploration Geophysics, p. 2428.
58. R. J. LEVEQUE, *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, New York, 2002.
59. A. LOGG, K. A. MARDAL, AND G. N. WELLS, eds., *The Fenics project*, Lecture notes in computational science and engineering, Springer, Berlin, 2012.
60. C. C. LOPEZ, *Accélération et régularisation de la méthode d'inversion des formes d'ondes complètes en exploration sismique*, PhD thesis, Université de Nice-Sophia Antipolis, 2014.
61. R. MADARIAGA, *Seismic source: Theory*, in Encyclopedia of Earth Sciences Series – Geophysics, C. W. Finkl, ed., Springer, Boston, MA, 1989, pp. 1129–1133.
62. G. MARCHUK, V. SHUTYAEV, AND G. BOCHAROV, *Adjoint equations and analysis of complex systems: Application to virus infection modelling*, J. Computational and Applied Mathematics, 184 (2005), pp. 177–204.
63. G. I. MARCHUK, *Adjoint Equations and Analysis of Complex Systems*, Springer, Netherlands, 1995.
64. G. I. MARCHUK, V. I. AGOSHKOV, AND V. P. SHUTYAEV, *Adjoint equations and perturbations algorithms in nonlinear problems*, CRC Press, Boca Raton, 1996.
65. G. S. MARTIN, R. WILEY, AND K. J. MARFURT, *Marmousi2: An elastic upgrade to Marmousi*, The Leading Edge, 25 (2006), p. 156.
66. MAVKO, *Quantitative seismic interpretation*, Springer, 2006.
67. G. MAVKO, T. MUKERJI, AND J. DVORKIN, *The Rock Physics Handbook*, Cambridge University Press, Cambridge, 1998.
68. L. MÉTIVIER, F. BRETAUDEAU, R. BROSSIER, S. OPERTO, AND J. VIRIEUX, *Full waveform inversion and the truncated Newton method: quantitative imaging of complex subsurface structures*, Geophysical Prospecting, 62 (2014), p. 1353.
69. L. METIVIER AND J. VIRIEUX, *Optimal transport theory*, in Frontiers in PDE-Constrained Optimization, H. Antil, M.-D. Lacasse, D. Ridzal, and D. P. Kouri, eds., Berlin, 2017, Springer.
70. P. MOCZO AND J. KRISTEK, *On the rheological models in the time-domain methods for seismic wave propagation*, Geophysical Review Letters, 32 (2005), p. L01306.

71. P. Moczo, J. Kristek, and P. Franek, *Lectures notes on rheological models*. http://www.fyzikazeme.sk/mainpage/stud_mat/Moczo_Kristek_Franek_Rheological_Models.pdf, 2006. retrieved March 1, 2018.
72. R. Modrak and J. Tromp, *Seismic waveform inversion best practices*, Geophysical Journal International, 206 (2016), p. 1864.
73. P. R. Mora, *Non-linear two-dimensional elastic inversion of multi-offset seismic data*, Geophysics, 52 (1987), p. 1211.
74. J. Nocedal and S. J. Wright, *Numerical optimization*, Springer Series in Operations Research and Financial Engineering, Springer, Berlin, 2006.
75. G. Noh and S. H. ans Klaus-Jürgen Bathe, *Performance of an implicit time integration scheme in the analysis of wave propagations*, Computers and Structures, 123 (2013), pp. 93–105.
76. C. C. Ober, T. M. Smith, J. R. Overfelt, S. S. Collis, G. J. von Winckel, B. G. van Bloemen Waanders, N. J. Downey, S. A. Mitchell, S. D. Bond, D. F. Aldridge, and J. R. Krebs, *Visco-TTI-elastic FWI using discontinuous Galerkin*, in SEG Technical Program Expanded Abstracts, Dallas, 2016, Society of Exploration Geophysics, p. 5654.
77. S. Operto, Y. Gholami, V. Prieux, A. Ribodetti, R. Brossier, L. Metivier, and J. Virieux, *A guided tour of multiparameter full waveform inversion with multicomponent data: from theory to practice*, The Leading Edge, 32 (2013), p. 936.
78. S. Operto, J. Virieux, P. Amestoy, J.-Y. L'Excellent, L. Giraud, and H. Ben Hadj Ali, *3D finite-difference frequency-domain modeling of visco-acoustic wave propagation using a massively parallel direct solver: A feasibility study*, Geophysics, 72 (2007), p. SM195.
79. W. J. Parnell and I. D. Abrahams, *New integral equation approach to elastodynamic homogenization*, Proceedings of the Royal Society A, 464 (2008), p. 1461.
80. R.-É. Plessix and Q. Cao, *A parametrization study for surface seismic full waveform inversion in an acoustic vertical transversely isotropic medium*, Geophys J Int, 185 (2011), p. 539.
81. R. G. Pratt, C. Shin, and G. J. Hicks, *Gauss-newton and full newton methods in frequency-space seismic waveform inversion*, Geophys. J. Int, 133 (1998), p. 341.
82. R. G. Pratt and M. H. Worthington, *Inverse theory applied to multi-source cross-hole tomography. Part I: acoustic wave-equation method*, Geophys. Prospect., 38 (1990), p. 287.
83. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes 3rd Edition: The Art of Scientific Computing*, Cambridge University Press, New York, third ed., 2007.
84. L. Qiu and M.-D. Lacasse, *Effects of parameterization on non-linear parameter estimation problems*, to be submitted.
85. C. D. Riyanti, Y. A. Erlangga, R.-É. Plessix, W. A. Mulder, C. Vuik, and C. Oosterlee, *New iterative solver for the time-harmonic wave equation*, Geophysics, 71 (2006), p. E57.
86. J. O. A. Robertsson, J. O. Blanch, and W. W. Symes, *Viscoelastic finite-difference modeling*, Geophysics, 59 (1994), p. 1444.
87. P. S. Routh, J. R. Krebs, S. Lazaratos, and J. E. Anderson, *Encoded simultaneous-source full-wavefield inversion for spectrally shaped marine streamer data*, in SEG Technical Program Expanded Abstracts, San Antonio, 2011, Society of Exploration Geophysics, p. 2433.
88. R. Sargent, *Progress in modelling and simulation*, in Verification and Validation of Simulation Models, F. Celier, ed., Academic Press, London, 1982, p. 159.
89. S. Scheslinger, R. E. Crosby, R. E. Gagné, G. S. Innis, C. S. Lalwani, J. Loch, R. J. Sylvester, R. D. Wright, N. Kheir, and D. Bartos, *Terminology for model credibility*, Simulation, (1979), pp. 103–104.
90. J. H. Schön, *Physical properties of rocks — Fundamentals and principles of petrophysics*, in Handbook of Geophysical Exploration, K. Helbig and S. Treitel, eds., vol. 18, Elsevier, 2004, p. 583.

91. P. M. SHEARER, *Introduction to Seismology*, Cambridge University Press, Cambridge, 1999.
92. SIAM WORKING GROUP ON CSE EDUCATION, *Graduate education in computational science and engineering*, SIAM Review, 43 (2001), p. 163.
93. I. S. SOKOLNIKOV, *Mathematical Theory of Elasticity*, McGraw-Hill, New York, 1956.
94. W. W. SYMES, I. S. TERENTYEV, AND T. W. VDOVINA, *Gridding requirements for accurate finite difference simulation*, in SEG Technical Program Expanded Abstracts, Las Vegas, 2008, Society of Exploration Geophysics, pp. 2077–2081.
95. A. TARANTOLA, *Inversion of seismic reflection data in the acoustic approximation*, Geophysics, 49 (1984), p. 1259.
96. A. TARANTOLA, *Inverse Problem Theory And Methods For Model Parameter Estimation*, Society of Applied and Industrial Mathematics, Philadelphia, 2005.
97. L. THOMSEN, *Weak elastic anisotropy*, Geophysics, 51 (1986), p. 1954.
98. V. A. TITAREV AND E. F. TORO, *ADER: Arbitrary high-order Godunov approach*, J. Sci. Comput., 17 (2002), pp. 609–18.
99. M. N. TOKSOZ, D. H. JOHNSTON, AND A. TIMUR, *Attenuation of seismic waves in dry and saturated rocks: I. Laboratory measurements*, Geophysics, 44 (1979), p. 681.
100. S. TORQUATO, *Random Heterogeneous Materials: Microstructure and Macroscopic Properties*, vol. 16 of Interdisciplinary applied mathematics, Springer-Verlag, New York, 2002.
101. J. TROMP, D. KOMATITSCH, AND Q. LIU, *Spectral elements and adjoint methods in seismology*, Communications in Computational Physics, 3 (2008), p. 1.
102. B. URSIN AND T. TOVERUD, *Comparison of seismic dispersion and attenuation models*, Stud. Geophys. Geod., 46 (2002), p. 293.
103. J. VIRIEUX, *P-SV wave propagation in heterogeneous media: Velocity-stress finite-difference method*, Geophysics, 51 (1986), p. 889.
104. J. VIRIEUX AND S. OPERTO, *An overview of full-waveform inversion in exploration geophysics*, Geophysics, 74 (2009), p. WCC127.
105. J. VIRIEUX, S. OPERTO, H. BEN HADJ ALI, R. BROSSIER, V. ETIENNE, F. S. AMD L. GIRAUD, AND A. HAIDAR, *Seismic wave modeling for seismic imaging*, The Leading Edge, 28 (2009), p. 538.
106. C. VOGEL, *Computational methods for inverse problems*, Society for Industrial and Applied Mathematics, Philadelphia, 2002.
107. S. WANG, M. V. DE HOOP, AND J. XIA, *On 3D modeling of seismic wave propagation via a structured parallel multifrontal direct Helmholtz solver*, Geophysical Prospecting, 59 (2011), p. 857.
108. Y. WANG, *Seismic Inverse Q Filtering*, John Wiley and Sons, New York, 2009.
109. M. WARNER AND L. GUASCH, *Adaptive waveform inversion: Theory*, Geophysics, 81 (2016), pp. R429–R445.
110. R. WU AND K. AKI, *Scattering characteristics of elastic waves by an elastic heterogeneity*, Geophysics, 50 (1985), p. 582.
111. P. YANG, R. BROSSIER, L. MÉTIVIER, AND J. VIRIEUX, *Wavefield reconstruction in attenuating media: A checkpointing-assisted reverse-forward simulation method*, Geophysics, 81 (2016), pp. R349–R362.
112. Y. O. YUAN, F. J. SIMONS, AND J. TROMP, *Double-difference adjoint seismic tomography*, Geophys. J. Int., 206 (2017), pp. 1599–1618.

Part II
PDE-Constrained Optimization: Applications

Energetically Optimal Flapping Wing Motions via Adjoint-Based Optimization and High-Order Discretizations

Matthew J. Zahr and Per-Olof Persson

Abstract A globally high-order numerical discretization of time-dependent conservation laws on deforming domains, and the corresponding fully discrete adjoint method, is reviewed and applied to determine energetically optimal flapping wing motions subject to aerodynamic constraints using a reduced space PDE-constrained optimization framework. The conservation law on a deforming domain is transformed to one on a fixed domain and discretized in space using a high-order discontinuous Galerkin method. An efficient, high-order temporal discretization is achieved using diagonally implicit Runge-Kutta schemes. Quantities of interest, such as the total energy required to complete a flapping cycle and the integrated forces produced on the wing, are discretized in a solver-consistent way, that is, via the same spatiotemporal discretization used for the conservation law. The fully discrete adjoint method is used to compute discretely consistent gradients of the quantities of interest and passed to a black-box, gradient-based nonlinear optimization solver. This framework successfully determines an energetically optimal flapping trajectory such that the net thrust of the wing is zero to within 9 digits after only 12 optimization iterations.

1 Introduction

Flapping flight has been a subject of intense interest and research over the past several decades due to its relevance in designing micro-aerial vehicles (MAVs) – unmanned aerial vehicles measuring no more than 15 cm in any dimension, envisioned in a number of civilian and military applications, including surveillance and reconnaissance [32, 43] – and in the understanding of biological systems. The

M. J. Zahr (✉) · P.-O. Persson
Mathematics Group, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

Department of Mathematics, University of California, Berkeley, Berkeley, CA 94720, USA
e-mail: mjzahr@lbl.gov; persson@berkeley.edu

H. Antil et al. (eds.), *Frontiers in PDE-Constrained Optimization*, The IMA Volumes in Mathematics and its Applications 163,
https://doi.org/10.1007/978-1-4939-8636-1_7

basic goal of any system, whether biological or manmade, that relies on flapping propulsion is to adjust the kinematics of the flapping wing, and possibly its shape, to minimize the energy required to complete a given mission. The problem of determining the flapping kinematics that lead to an energetically optimal motion, while satisfying various mission constraints, leads to a nonlinearly constrained PDE-constrained optimization problem

$$\begin{aligned} \underset{\boldsymbol{U},\,\boldsymbol{\mu}}{\text{minimize}} \quad & J(\boldsymbol{U},\,\boldsymbol{\mu}) \coloneqq \frac{1}{T}\int_0^T \int_{\boldsymbol{\Gamma}} j(\boldsymbol{U}(\boldsymbol{x},\,t),\,\boldsymbol{\mu},\,t)\,dS\,dt \\ \text{subject to} \quad & \mathbf{C}(\boldsymbol{U},\,\boldsymbol{\mu}) \coloneqq \frac{1}{T}\int_0^T \int_{\boldsymbol{\Gamma}} \mathbf{c}(\boldsymbol{U}(\boldsymbol{x},\,t),\,\boldsymbol{\mu},\,t)\,dS\,dt \leq 0 \\ & \frac{\partial \boldsymbol{U}}{\partial t} + \nabla \cdot \boldsymbol{F}(\boldsymbol{U},\,\nabla \boldsymbol{U}) = 0 \quad \text{in} \quad v(\boldsymbol{\mu},\,t), \end{aligned} \tag{1}$$

where $\boldsymbol{U}(\boldsymbol{x},\,t) \in \mathbb{R}^{N_U}$ is the spatiotemporal solution of the conservation law, i.e., the last constraint in the optimization problem, in the domain $\boldsymbol{x} \in v(\boldsymbol{\mu},\,t)$, $t \in (0,\,T]$, $\boldsymbol{\mu} \in \mathbb{R}^{N_{\boldsymbol{\mu}}}$ is a vector of parameters controlling the kinematics of the wing, T is the period of the flapping motion, $j(\boldsymbol{U},\,\boldsymbol{\mu},\,t)$ is the pointwise (in space and time) contribution to energy added to the flow and $J(\boldsymbol{U},\,\boldsymbol{\mu})$ is the corresponding quantity integrated over space and time, i.e., the time-averaged work done by the surface $\boldsymbol{\Gamma}(\boldsymbol{\mu},\,t)$ on the flow, and $\boldsymbol{c}(\boldsymbol{U},\,\boldsymbol{\mu},\,t)$ and $\mathbf{C}(\boldsymbol{U},\,\boldsymbol{\mu})$ are pointwise and integrated, respectively, mission-specific constraints. In the context of MAV design, the constraints will likely correspond to bounds on the thrust, lift, and stability of the vehicle [19, 49].

Due to the unsteady governing equations, most attempts to solve the PDE-constrained optimization problem in (1) in the context of flapping flight utilize a *reduced space* approach, also known as nested analysis and design, whereby the state variable $\boldsymbol{U}$ is treated as an implicit function of the parameters $\boldsymbol{\mu}$, i.e., $\boldsymbol{U}(\boldsymbol{\mu})$ is obtained by solving the (discretized) conservation law. This removes the state variable from the set of optimization variables and eliminates the PDE constraint to reduce the optimization problem in (1) to

$$\begin{aligned} \underset{\boldsymbol{\mu}}{\text{minimize}} \quad & \mathscr{J}(\boldsymbol{\mu}) \coloneqq J(\boldsymbol{U}(\boldsymbol{\mu}),\,\boldsymbol{\mu}) \\ \text{subject to} \quad & \boldsymbol{C}(\boldsymbol{\mu}) \coloneqq \mathbf{C}(\boldsymbol{U}(\boldsymbol{\mu}),\,\boldsymbol{\mu}) \leq 0. \end{aligned} \tag{2}$$

Due to the relatively large expense of high-fidelity methods that model the flow using the Navier-Stokes equations, a number of low- and multi-fidelity methods have been proposed to approximately solve the optimization problem in (2) or gain insight into the physics of flapping. Among these low-fidelity methods include: potential flow methods that assume that the flow is irrotational, inviscid, and incompressible such as wake only and panel methods [56], lifting line methods, and unsteady vortex-lattice methods [20, 46] that assume that the flow is inviscid and incompressible and use global vorticity circulation balance and the Biot-Savart law to construct a 3D velocity field.

While low- and multi-fidelity methods have been popular in the study of flapping flight [7, 9, 19, 20, 24, 46, 56], the need for high-fidelity computational tools has been recognized [43] due to the complex flow features that occur, and are critical for performance, in low Reynolds number flapping. In particular, these flows are highly vortical and subject to separation [2, 6, 22, 44] that will violate many of the critical assumptions of low-fidelity methods [56]. The generation and shedding of a leading-edge vortex, possibly through rapid changes in angle of attack (dynamic stall), have been shown to be important to efficient lift generation [6, 7, 42, 44] and a computational method should possess minimal dissipation to ensure that these critical structures are preserved. Furthermore, a realistic study of flapping at scales relevant to the design of MAVs should be performed in three dimensions due to the importance of three-dimensional effects such as stabilization of the leading-edge vortex [5–7, 13, 27, 50] and to include out-of-plane flapping kinematics that are relevant to thrust production and control [4].

In this work, we extend the globally high-order method and corresponding fully discrete adjoint method presented in [63] for the discretization and optimization of general nonlinear, unsteady conservation laws to address the challenges of three-dimensional flapping, such as the parametrization of three-dimensional flapping and robust deformation of the three-dimensional geometry. The conservation law on a parametrized, deforming domain is transformed to a fixed domain using an Arbitrary Lagrangian-Eulerian (ALE) formulation and the resulting equations are discretized in space and time using a discontinuous Galerkin method and diagonally implicit Runge-Kutta scheme, respectively. Relevant details are provided on using the ALE formulation to move a *curved* mesh, whereby the reference mesh is taken as straight-sided and the ALE mapping encapsulates the curving as well as the domain deformation. In contrast to most computational approaches that only integrate quantities of interest (QoIs), that will eventually define the objective and constraints of the flapping optimization problem, to second order using the trapezoidal rule, this work discretizes the QoI to exactly the same order as the governing equation using the solver-consistent approach of [63]. High-order methods are an emphasis of this work since they are well suited to model the highly vortical flow around a flapping wing due to the small amount of numerical dissipation they introduce [35]. An alternative to high-order methods that has been proposed and demonstrated in the context of flapping to limit numerical dissipation associated with low-order methods is a kinetic energy preserving finite volume scheme [2, 11]. However, we also commit to high-order methods because they have been shown to require fewer spatial degrees of freedom [54, 63] and time steps [30, 61] compared to low-order counterparts.

Given the large computational cost of objective and constraint queries that require high-order computational fluid dynamics (CFD) simulations, and the high-dimensional design space required to sufficiently parametrize three-dimensional flapping, which may include parameters for the flapping kinematics, fixed or actively morphed [20, 63] shape, flexibility of the wing [22, 42, 43, 47, 48, 65], gradient-based optimization methods are used to solve the optimization problem in (2) due to their fast convergence properties. Also, since the optimization problems

considered in this work involve more parameters than constraints, the gradients of the optimization functionals are computed via the adjoint method since the cost scales very weakly with the number of parameters. Since a black-box optimizer is used to solve the constrained optimization problem in (2) with the underlying high-order discretization, the fully discrete variant of the adjoint method is used to ensure that the computed gradients are consistent with the functionals to which they correspond.

The proposed numerical method for simulation and optimization of conservation laws on parametrized, deforming domains is used to determine energetically optimal flapping subject to a thrust constraint. The chosen optimization formulation is similar to that studied in [19], which differs from the unconstrained thrust or propulsive efficiency maximization problem that is usually chosen to study optimal flapping [40, 49, 55]. The optimization problem considered in this work is closer to the optimization problem instinctively solved in-flight by biological systems [43] and relevant in the design of MAVs.

The remainder of this document is organized as follows. Section 2 introduces the governing conservation law considered in this work, the isentropic Navier-Stokes equations, and an Arbitrary Lagrangian-Eulerian method that transforms it from a deforming, parametrized domain to a fixed one. Section 3 introduces the high-order discretization of the conservation law and its quantities of interest, with special attention paid to high-order representation of the geometry in the ALE framework and Section 4 introduces the fully discrete adjoint method that was derived in [63]. Finally, Section 5 applies this high-order simulation and optimization framework to energetically optimal, three-dimensional flapping flight under lift and thrust constraints and Section 6 offers conclusions.

2 Governing Equations

This section presents a formulation of general conservation laws on a *parametrized, deforming domain* using an Arbitrary Lagrangian-Eulerian (ALE) formulation, which summarizes the work in [38]. Given that this work is concerned with energetically optimal flapping flight, the compressible Navier-Stokes equations are taken as the governing equations; however, the primal and adjoint numerical scheme is presented for the case of a general, nonlinear, vector-valued conservation law.

2.1 Compressible Navier-Stokes Equations

The compressible Navier-Stokes equations are written as:

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x_i}(\rho u_i) = 0, \tag{3}$$

$$\frac{\partial}{\partial t}(\rho u_i) + \frac{\partial}{\partial x_i}(\rho u_i u_j + p) = \frac{\partial \tau_{ij}}{\partial x_j} \quad \text{for } i = 1, 2, 3, \tag{4}$$

$$\frac{\partial}{\partial t}(\rho E) + \frac{\partial}{\partial x_j}\left(u_j(\rho E + p)\right) = -\frac{\partial q_j}{\partial x_j} + \frac{\partial}{\partial x_j}(u_i \tau_{ij}), \tag{5}$$

where ρ is the fluid density, u_1, u_2, u_3 are the velocity components, and E is the total energy. The viscous stress tensor and heat flux are given by

$$\tau_{ij} = \mu\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} - \frac{2}{3}\frac{\partial u_k}{\partial x_k}\delta_{ij}\right) \quad \text{and} \quad q_j = -\frac{\mu}{\text{Pr}}\frac{\partial}{\partial x_j}\left(E + \frac{p}{\rho} - \frac{1}{2}u_k u_k\right). \tag{6}$$

Here, μ is the viscosity coefficient and Pr = 0.72 is the Prandtl number which we assume to be constant. For an ideal gas, the pressure p has the form

$$p = (\gamma - 1)\rho\left(E - \frac{1}{2}u_k u_k\right), \tag{7}$$

where γ is the adiabatic gas constant. In this work, the entropy is assumed constant, that is to say the flow is adiabatic and reversible. This makes the energy equation redundant and effectively reduces the square system of PDEs of size $n_{sd} + 2$ to one of size $n_{sd} + 1$, where n_{sd} is the number of spatial dimensions. It can be shown, under suitable assumptions, that the solution of the isentropic approximation of the Navier-Stokes equations converges to the solution of the incompressible Navier-Stokes equations as the Mach number goes to 0 [12, 17, 29].

2.2 Arbitrary Lagrangian-Eulerian Formulation of Conservation Laws

Consider a general system of conservation laws, defined on a parametrized, deforming domain, $v(\boldsymbol{\mu}, t)$,

$$\frac{\partial \boldsymbol{U}}{\partial t} + \nabla \cdot \boldsymbol{F}(\boldsymbol{U}, \nabla \boldsymbol{U}) = 0 \quad \text{in } v(\boldsymbol{\mu}, t) \tag{8}$$

where the physical flux is decomposed into an inviscid and a viscous part $\boldsymbol{F}(\boldsymbol{U}, \nabla \boldsymbol{U}) = \boldsymbol{F}^{inv}(\boldsymbol{U}) + \boldsymbol{F}^{vis}(\boldsymbol{U}, \nabla \boldsymbol{U})$, $\boldsymbol{U}(\boldsymbol{x}, \boldsymbol{\mu}, t)$ is the solution of the system of conservation laws, $t \in (0, T]$ represents time, and $\boldsymbol{\mu} \in \mathbb{R}^{N_{\boldsymbol{\mu}}}$ is a vector of parameters. This work will focus on the case where the *domain* is parametrized by $\boldsymbol{\mu}$.

The conservation law on the physical, deforming domain $v(\boldsymbol{\mu}, t) \subset \mathbb{R}^{n_{sd}}$ is transformed into one on a *fixed* reference domain $V \subset \mathbb{R}^{n_{sd}}$ through the introduction

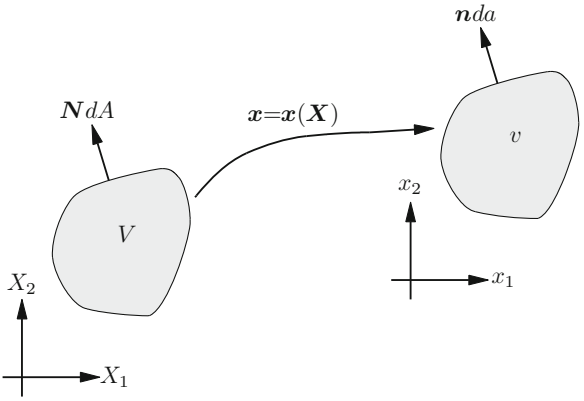

Fig. 1 Time-dependent mapping between reference and physical domains

of a time-dependent diffeomorphism between the physical and reference domains: $\boldsymbol{x}(\boldsymbol{X}, \boldsymbol{\mu}, t) = \mathscr{G}(\boldsymbol{X}, \boldsymbol{\mu}, t)$ (Figure 1). In this setting, n_{sd} is the number of spatial dimensions, $\boldsymbol{X} \in V$ is a point in the reference domain, and $\boldsymbol{x}(\boldsymbol{X}, \boldsymbol{\mu}, t) \in v(\boldsymbol{\mu}, t)$ is the corresponding point in the physical domain at time t and parameter configuration $\boldsymbol{\mu}$. The transformed system of conservation laws takes the form

$$\left.\frac{\partial \boldsymbol{U}_{\boldsymbol{X}}}{\partial t}\right|_{\boldsymbol{X}} + \nabla_{\boldsymbol{X}} \cdot \boldsymbol{F}_{\boldsymbol{X}}(\boldsymbol{U}_{\boldsymbol{X}},\ \nabla_{\boldsymbol{X}} \boldsymbol{U}_{\boldsymbol{X}}) = 0 \quad \text{in } V \tag{9}$$

where $\nabla_{\boldsymbol{X}}$ denotes spatial derivatives with respect to the reference variables, $\boldsymbol{X}$. The transformed state vector, $\boldsymbol{U}_{\boldsymbol{X}}$, and its corresponding spatial gradient with respect to the reference configuration take the form

$$\boldsymbol{U}_{\boldsymbol{X}} = g\boldsymbol{U}, \qquad \nabla_{\boldsymbol{X}} \boldsymbol{U}_{\boldsymbol{X}} = g^{-1} \boldsymbol{U}_{\boldsymbol{X}} \frac{\partial g}{\partial \boldsymbol{X}} + g \nabla \boldsymbol{U} \cdot \boldsymbol{G}, \tag{10}$$

where $\boldsymbol{G} = \nabla_{\boldsymbol{X}} \mathscr{G}$, $g = \det(\boldsymbol{G})$, $\boldsymbol{v}_{\boldsymbol{G}} = \dfrac{\partial \boldsymbol{x}}{\partial t} = \dfrac{\partial \mathscr{G}}{\partial t}$. The transformed fluxes are

$$\begin{aligned}
\boldsymbol{F}_{\boldsymbol{X}}(\boldsymbol{U}_{\boldsymbol{X}}, \nabla_{\boldsymbol{X}} \boldsymbol{U}_{\boldsymbol{X}}) &= \boldsymbol{F}_{\boldsymbol{X}}^{inv}(\boldsymbol{U}_{\boldsymbol{X}}) + \boldsymbol{F}_{\boldsymbol{X}}^{vis}(\boldsymbol{U}_{\boldsymbol{X}}, \nabla_{\boldsymbol{X}} \boldsymbol{U}_{\boldsymbol{X}}), \\
\boldsymbol{F}_{\boldsymbol{X}}^{inv}(\boldsymbol{U}_{\boldsymbol{X}}) &= g \boldsymbol{F}^{inv}(g^{-1} \boldsymbol{U}_{\boldsymbol{X}}) \boldsymbol{G}^{-T} - \boldsymbol{U}_{\boldsymbol{X}} \otimes \boldsymbol{G}^{-1} \boldsymbol{v}_{\boldsymbol{G}}, \\
\boldsymbol{F}_{\boldsymbol{X}}^{vis}(\boldsymbol{U}_{\boldsymbol{X}}, \nabla_{\boldsymbol{X}} \boldsymbol{U}_{\boldsymbol{X}}) &= g \boldsymbol{F}^{vis}\left(g^{-1} \boldsymbol{U}_{\boldsymbol{X}}, g^{-1}\left[\nabla_{\boldsymbol{X}} \boldsymbol{U}_{\boldsymbol{X}} - g^{-1} \boldsymbol{U}_{\boldsymbol{X}} \frac{\partial g}{\partial \boldsymbol{X}}\right] \boldsymbol{G}^{-1}\right) \boldsymbol{G}^{-T}.
\end{aligned} \tag{11}$$

For details regarding the derivation of the transformed equations, the reader is referred to [38].

When integrated using inexact numerical schemes, this ALE formulation does not satisfy the geometric conservation law (GCL) [15, 38]. This is overcome by introduction of an auxiliary variable $\bar{g}$, defined as the solution of

$$\frac{\partial \bar{g}}{\partial t} - \nabla_{\boldsymbol{X}} \cdot \left(g \boldsymbol{G}^{-1} \boldsymbol{v}_{\boldsymbol{G}} \right) = 0. \tag{12}$$

The auxiliary variable, $\bar{g}$, is used to modify the *transformed* conservation law according to

$$\left. \frac{\partial \boldsymbol{U}_{\bar{\boldsymbol{X}}}}{\partial t} \right|_{\boldsymbol{X}} + \nabla_{\boldsymbol{X}} \cdot \boldsymbol{F}_{\bar{\boldsymbol{X}}}(\boldsymbol{U}_{\bar{\boldsymbol{X}}}, \nabla_{\boldsymbol{X}} \boldsymbol{U}_{\bar{\boldsymbol{X}}}) = 0 \tag{13}$$

where the GCL-transformed state variables are

$$\boldsymbol{U}_{\bar{\boldsymbol{X}}} = \bar{g} \boldsymbol{U}, \qquad \nabla_{\boldsymbol{X}} \boldsymbol{U}_{\bar{\boldsymbol{X}}} = \bar{g}^{-1} \boldsymbol{U}_{\bar{\boldsymbol{X}}} \frac{\partial \bar{g}}{\partial \boldsymbol{X}} + \bar{g} \nabla \boldsymbol{U} \cdot \boldsymbol{G} \tag{14}$$

and the corresponding fluxes

$$\begin{aligned} \boldsymbol{F}_{\bar{\boldsymbol{X}}}(\boldsymbol{U}_{\bar{\boldsymbol{X}}}, \nabla_{\boldsymbol{X}} \boldsymbol{U}_{\bar{\boldsymbol{X}}}) &= \boldsymbol{F}^{inv}_{\bar{\boldsymbol{X}}}(\boldsymbol{U}_{\bar{\boldsymbol{X}}}) + \boldsymbol{F}^{vis}_{\bar{\boldsymbol{X}}}(\boldsymbol{U}_{\bar{\boldsymbol{X}}}, \nabla_{\boldsymbol{X}} \boldsymbol{U}_{\bar{\boldsymbol{X}}}), \\ \boldsymbol{F}^{inv}_{\bar{\boldsymbol{X}}}(\boldsymbol{U}_{\bar{\boldsymbol{X}}}) &= g \boldsymbol{F}^{inv}(\bar{g}^{-1} \boldsymbol{U}_{\bar{\boldsymbol{X}}}) \boldsymbol{G}^{-T} - \boldsymbol{U}_{\bar{\boldsymbol{X}}} \otimes \boldsymbol{G}^{-1} \boldsymbol{v}_{\boldsymbol{G}}, \\ \boldsymbol{F}^{vis}_{\bar{\boldsymbol{X}}}(\boldsymbol{U}_{\bar{\boldsymbol{X}}}, \nabla_{\boldsymbol{X}} \boldsymbol{U}_{\bar{\boldsymbol{X}}}) &= g \boldsymbol{F}^{vis} \left(\bar{g}^{-1} \boldsymbol{U}_{\bar{\boldsymbol{X}}}, \bar{g}^{-1} \left[\nabla_{\boldsymbol{X}} \boldsymbol{U}_{\bar{\boldsymbol{X}}} - \bar{g}^{-1} \boldsymbol{U}_{\bar{\boldsymbol{X}}} \frac{\partial \bar{g}}{\partial \boldsymbol{X}} \right] \boldsymbol{G}^{-1} \right) \boldsymbol{G}^{-T}. \end{aligned} \tag{15}$$

It was shown in [38] that the transformed equations (13) satisfy the GCL.

2.3 Uniform Flow Initial Condition

A number of initial conditions can be used to initialize an unsteady CFD simulation, including uniform flow [21, 23], the steady-state solution [26, 28, 63], and the state that leads to periodic flow [64]. In this work, the unsteady simulation is initialized from uniform flow for the sake of simplicity. Nonphysical transients that result from using an initial condition that is incompatible with the boundary conditions will be dissipated by simulating multiple periods of the flapping motion and only integrating the quantity of interest over the final period. The ALE-transformed state corresponding to uniform flow takes the form

$$\begin{aligned} \boldsymbol{U}_{\bar{\boldsymbol{X}}}(\boldsymbol{X}, \boldsymbol{\mu}, 0) &= g_0(\boldsymbol{X}, \boldsymbol{\mu}) \bar{\boldsymbol{U}}(\boldsymbol{X}) \\ \bar{g}(\boldsymbol{X}, \boldsymbol{\mu}, 0) &= g_0(\boldsymbol{X}, \boldsymbol{\mu}) \end{aligned} \tag{16}$$

where $\bar{\boldsymbol{U}}(\boldsymbol{X})$ defines the desired uniform initial condition on the reference domain and $g_0(\boldsymbol{X},\ \boldsymbol{\mu}) \coloneqq g(\boldsymbol{X},\ \boldsymbol{\mu},\ 0)$ is the determinant of the deformation gradient at time $t = 0$.

3 High-Order Numerical Discretization

This section discusses a globally high-order numerical discretization of the governing equations presented in the previous section. It summarizes the work in [1, 3, 63].

3.1 Spatial Discretization: Discontinuous Galerkin Method

To proceed, the second-order system of partial differential equations in (12)–(13) is converted to first-order form

$$\begin{aligned} \left.\frac{\partial \bar{g}}{\partial t}\right|_{\boldsymbol{X}} + \nabla_{\boldsymbol{X}} \cdot \left(g \boldsymbol{G}^{-1} \boldsymbol{v}_{\boldsymbol{G}}\right) &= 0 \\ \left.\frac{\partial \boldsymbol{U}_{\bar{\boldsymbol{X}}}}{\partial t}\right|_{\boldsymbol{X}} + \nabla_{\boldsymbol{X}} \cdot \boldsymbol{F}_{\bar{\boldsymbol{X}}}(\boldsymbol{U}_{\bar{\boldsymbol{X}}},\ \boldsymbol{Q}_{\bar{\boldsymbol{X}}}) &= 0 \\ \boldsymbol{Q}_{\bar{\boldsymbol{X}}} - \nabla_{\boldsymbol{X}} \boldsymbol{U}_{\bar{\boldsymbol{X}}} &= 0, \end{aligned} \tag{17}$$

where $\boldsymbol{Q}_{\bar{\boldsymbol{X}}}$ is introduced as an auxiliary variable to represent the spatial gradient of the $\boldsymbol{U}_{\bar{\boldsymbol{X}}}$. Equation (17) is discretized using a standard nodal discontinuous Galerkin finite element method [3, 10], which, after local elimination of the auxiliary variables $\boldsymbol{Q}_{\bar{\boldsymbol{X}}}$, leads to the following system of ODEs

$$\boldsymbol{M} \frac{\partial \boldsymbol{u}}{\partial t} = \boldsymbol{r}(\boldsymbol{u},\ \boldsymbol{\mu},\ t), \tag{18}$$

where $\boldsymbol{M}$ is the block-diagonal, symmetric, *fixed* mass matrix (state- and parameter-independent), $\boldsymbol{u}$ is the vectorization of $\left[\boldsymbol{U}_{\bar{\boldsymbol{X}}}^T\ \bar{g}\right]^T$ at all nodes in the mesh, and $\boldsymbol{r}$ is the nonlinear function defining the DG discretization of the inviscid and viscous fluxes. See [63] for an efficient treatment of $\bar{g}$ that does not lead to an *enlarged* system of ODEs.

To achieve high-order accuracy, the geometry must be represented to high-order, which calls for a curved mesh. Since a curved mesh is usually defined as a nonlinear mapping, e.g., based on nonlinear elasticity or some optimality criteria, applied to an underlying linear or straight-sided mesh, two options exist for defining the ALE mapping. First, the curved mesh can be taken as the reference domain and the ALE mapping must only account for the mapping between the curved mesh and the physical domain. In this case, the ALE mapping takes the form

$$\boldsymbol{x}(\boldsymbol{X},\ \boldsymbol{\mu},\ t) = \boldsymbol{\varphi}(\boldsymbol{X},\ \boldsymbol{\mu},\ t) \tag{19}$$

where $\boldsymbol{X}$ are coordinates in the domain defined by the *curved mesh* and $\boldsymbol{\varphi}$ maps the curved mesh into the physical domain. Alternatively, the straight-sided mesh can be taken as the reference mesh and the ALE mapping constructed as a composition of maps that takes the straight-sided mesh into the physical domain with curved boundaries to represent the geometry to high-order. In this case, the ALE mapping takes the form

$$\boldsymbol{x}(\boldsymbol{X},\ \boldsymbol{\mu},\ t) = \boldsymbol{\varphi}(\boldsymbol{\phi}(\boldsymbol{X}),\ \boldsymbol{\mu},\ t) \tag{20}$$

where $\boldsymbol{X}$ are coordinates in the domain defined by the *linear mesh*, $\boldsymbol{\phi}$ maps the linear mesh to the curved mesh, and $\boldsymbol{\varphi}$ maps the curved mesh into the physical domain. Even though these options are mathematically equivalent, the latter option is chosen in this work as it leads to a simpler implementation, particularly in the definition of derivative terms required for the adjoint method, but also because all integrals are calculated on straight-sided elements.

This section closes with a discussion of how the domain deformation terms that arise in the ALE formulation will be defined at the semi-discrete level. If the mapping from the reference to physical domain is known analytically, all domain deformation terms, i.e., $\boldsymbol{x}$, $\dot{\boldsymbol{x}}$, $\boldsymbol{G}$, g, can be computed exactly and used in (15). However, there are many cases where this is not the case, e.g., the domain deformation is the result of a numerical procedure [14, 16, 39, 58]. An alternative that closely aligns with finite element ideology is to interpolate the ALE mapping onto the finite element shape functions and compute spatial gradients by differentiating the shape functions. In this setting, the action of the mapping and its time derivative are computed on the nodal coordinates of the reference mesh, i.e.,

$$\begin{aligned}
\boldsymbol{x}^e(\boldsymbol{X},\ \boldsymbol{\mu},\ t) &\coloneqq \boldsymbol{x}(\boldsymbol{X},\ \boldsymbol{\mu},\ t)|_{\boldsymbol{X}\in\mathscr{E}_e} = \sum_{i\in\mathscr{N}(e)} N_i(\boldsymbol{X})\boldsymbol{x}_i(\boldsymbol{\mu},\ t) \\
\dot{\boldsymbol{x}}^e(\boldsymbol{X},\ \boldsymbol{\mu},\ t) &\coloneqq \dot{\boldsymbol{x}}(\boldsymbol{X},\ \boldsymbol{\mu},\ t)|_{\boldsymbol{X}\in\mathscr{E}_e} = \sum_{i\in\mathscr{N}(e)} N_i(\boldsymbol{X})\dot{\boldsymbol{x}}_i(\boldsymbol{\mu},\ t),
\end{aligned} \tag{21}$$

where $\mathscr{E}_e$ is element e in the reference mesh, $\mathscr{N}(e)$ are the nodes associated with element e, $N_i(\boldsymbol{X})$ are the DG shape functions on the reference mesh, and $\boldsymbol{x}^e$ are the coordinates of the nodes of element e in the physical domain. An implication of defining the ALE mapping with the DG shape function is that the mapping is *discontinuous* between elements, which does not present a problem for the DG method. The expression for the mapping in (21) implies that the deformation gradient and its determinant can be easily computed as

$$\begin{aligned}
\boldsymbol{G}^e(\boldsymbol{X},\ \boldsymbol{\mu},\ t) &\coloneqq \boldsymbol{G}(\boldsymbol{X},\ \boldsymbol{\mu},\ t)|_{\mathscr{E}_e} = \sum_{i\in\mathscr{N}(e)} \boldsymbol{x}_i(\boldsymbol{\mu},\ t)\frac{\partial N_i}{\partial \boldsymbol{X}}(\boldsymbol{X}), \\
g^e(\boldsymbol{X},\ \boldsymbol{\mu},\ t) &= \det \boldsymbol{G}^e(\boldsymbol{X},\ \boldsymbol{\mu},\ t).
\end{aligned} \tag{22}$$

Therefore, once the nodal coordinates of the mapping and its time derivatives are known, all the remaining terms directly follow. The implications of such a dependence in the implementation of the adjoint method were discussed in [63] and will be further detailed in Section 5.1.

3.2 Temporal Discretization: Diagonally Implicit Runge-Kutta

The system of ODEs in (18) are discretized in time using diagonally implicit Runge-Kutta (DIRK) schemes. These schemes are capable of achieving high-order accuracy with the desired stability properties (unlike high-order multistep schemes that are only stable up to second order), without requiring the solution of an enlarged system of equations like general implicit Runge-Kutta (IRK) schemes (see [36] for an efficient solver for DG-IRK discretizations). DIRK schemes are defined by a *lower triangular* Butcher tableau (Table 1) and take the following form when applied to (18)

$$\begin{aligned} \boldsymbol{u}_0 &= \bar{\boldsymbol{u}}(\boldsymbol{\mu}) \\ \boldsymbol{u}_n &= \boldsymbol{u}_{n-1} + \sum_{i=1}^{s} b_i \boldsymbol{k}_{n,i} \\ \boldsymbol{M}\boldsymbol{k}_{n,i} &= \Delta t_n \boldsymbol{r}\left(\boldsymbol{u}_{n,i},\ \boldsymbol{\mu},\ t_{n-1} + c_i \Delta t_n\right), \end{aligned} \tag{23}$$

for $n = 1, \ldots, N_t$ and $i = 1, \ldots, s$, where N_t are the number of time steps in the temporal discretization and s is the number of stages in the DIRK scheme. The initial condition, $\bar{\boldsymbol{u}}(\boldsymbol{\mu})$, corresponds to the vectorization of the ALE-transformed uniform flow state in (16). The temporal domain, $(0,\ T]$, is discretized into N_t segments with endpoints $\{t_0, t_1, \ldots, t_{N_t}\}$, with the nth segment having length $\Delta t_n = t_n - t_{n-1}$ for $n = 1, \ldots, N_t$. Additionally, in (23), $\boldsymbol{u}_{n,i}$ is used to denote the approximation of $\boldsymbol{u}_n$ at the ith stage of time step n

$$\boldsymbol{u}_{n,i} = \boldsymbol{u}_{n,i}(\boldsymbol{u}_{n-1},\ \boldsymbol{k}_{n,1}, \ldots,\ \boldsymbol{k}_{n,s}) = \boldsymbol{u}_{n-1} + \sum_{j=1}^{i} a_{ij} \boldsymbol{k}_{n,j}. \tag{24}$$

From (23), a complete time step requires the solution of a sequence of s nonlinear systems of equation of size $N_{\boldsymbol{u}}$.

Table 1 Butcher tableau for s-stage diagonally implicit Runge-Kutta scheme

c_1	a_{11}			
c_2	a_{21}	a_{22}		
$\vdots$	$\vdots$	$\vdots$	$\ddots$	
c_s	a_{s1}	a_{s2}	$\cdots$	a_{ss}
	b_1	b_2	$\cdots$	b_s

3.3 Solver-Consistent Discretization of Quantities of Interest

In this work, quantities of interest that take the form of space-time integrals of nonlinear functions that depend on the solution of the conservation law are discretized in a solver-consistent manner [63], i.e., using the same spatial and temporal discretization used for the conservation law. This ensures that the truncation error of the quantities of interest exactly match that of the governing equations.

Consider a quantity of interest of the form

$$\mathscr{F}(\boldsymbol{U}, \boldsymbol{\mu}, t) = \int_0^t \int_{\boldsymbol{\Gamma}} w(\boldsymbol{x}, \tau) f(\boldsymbol{U}(\boldsymbol{x}, \tau), \boldsymbol{\mu}, \tau) \, dS \, d\tau. \tag{25}$$

In the context of the optimization problem in (1), $\mathscr{F}$ corresponds to either the objective or a constraint function. Define f^h as the approximation of $\int_{\boldsymbol{\Gamma}} w(\boldsymbol{x}, t) f(\boldsymbol{U}(\boldsymbol{x}, t), \boldsymbol{\mu}, t) \, dS$ using the DG shape functions from the spatial discretization of the governing equations. The solver-consistent spatial discretization of (25) becomes

$$\mathscr{F}^h(\boldsymbol{u}, \boldsymbol{\mu}, t) = \int_0^t f^h(\boldsymbol{u}, \boldsymbol{\mu}, \tau) \, d\tau, \tag{26}$$

which ensures that the spatial integration error in the quantity of interest exactly matches that of the governing equations. Solver-consistent temporal discretization requires the semi-discrete functional in (26) be converted to an ODE, which is accomplished via differentiation of (26) with respect to t

$$\dot{\mathscr{F}}^h(\boldsymbol{u}, \boldsymbol{\mu}, t) = f^h(\boldsymbol{u}, \boldsymbol{\mu}, t). \tag{27}$$

Augmenting the semi-discrete governing equations with this ODE (27) yields the system of ODEs

$$\begin{bmatrix} \boldsymbol{M} & \boldsymbol{0} \\ \boldsymbol{0} & 1 \end{bmatrix} \begin{bmatrix} \dot{\boldsymbol{u}} \\ \dot{\mathscr{F}}^h \end{bmatrix} = \begin{bmatrix} \boldsymbol{r}(\boldsymbol{u}, \boldsymbol{\mu}, t) \\ f^h(\boldsymbol{u}, \boldsymbol{\mu}, t) \end{bmatrix}. \tag{28}$$

Application of the DIRK temporal discretization introduced in Section 3.2 yields the fully discrete governing equations and corresponding solver-consistent discretization of the quantity of interest (25)

$$\begin{aligned} \boldsymbol{u}_n &= \boldsymbol{u}_{n-1} + \sum_{i=1}^{s} b_i \boldsymbol{k}_{n,i} \\ \mathscr{F}_n^h &= \mathscr{F}_{n-1}^h + \Delta t_n \sum_{i=1}^{s} b_i f^h\left(\boldsymbol{u}_{n,i}, \ \boldsymbol{\mu}, \ t_{n-1} + c_i \Delta t_n\right) \\ \boldsymbol{M}\boldsymbol{k}_{n,i} &= \Delta t_n \boldsymbol{r}\left(\boldsymbol{u}_{n,i}, \ \boldsymbol{\mu}, \ t_{n-1} + c_i \Delta t_n\right), \end{aligned} \tag{29}$$

for $n = 1, \ldots, N_t$, $i = 1, \ldots, s$, and $\boldsymbol{u}_{n,i}$ is defined in (24). Finally, the functional in (25) is evaluated at time $t = T$ to yield the solver-consistent approximation of $\mathscr{F}(\boldsymbol{u}, \boldsymbol{\mu}, T)$

$$F(\boldsymbol{u}_0, \ldots, \boldsymbol{u}_{N_t}, \boldsymbol{k}_{1,1}, \ldots, \boldsymbol{k}_{N_t,s}) \coloneqq \mathscr{F}^h_{N_t} \approx \mathscr{F}(\boldsymbol{u}, \boldsymbol{\mu}, T). \tag{30}$$

Unlike most methods used in the literature for integrating quantities of interest in time, e.g., trapezoidal rule [25, 33, 34, 51, 57], the proposed method relies on the low-order, intermediate RK stages. These stages are combined in such a way that the temporal integral in (26) is approximated to high-order. The dependence of the quantity of interest on these stages must be accounted for in the adjoint equations [63, 64], which will be seen in Section 4.1.

4 Fully Discrete Adjoint Method

4.1 Fully Discrete, Time-Dependent Adjoint Equations

This section summarizes the work in [63] and begins by posing the adjoint equations corresponding to the fully discrete system of conservation laws in (23) and the adjoint method for computing the total derivative of the fully discrete quantity of interest without requiring solution sensitivities, $\frac{\partial \boldsymbol{u}_n}{\partial \boldsymbol{\mu}}$ and $\frac{\partial \boldsymbol{k}_{n,i}}{\partial \boldsymbol{\mu}}$. Each of the $N_{\boldsymbol{\mu}}$ solution sensitivities is the solution of the following linear evolution equations

$$\begin{aligned}
\frac{\partial \boldsymbol{u}_0}{\partial \boldsymbol{\mu}} &= \frac{\partial \bar{\boldsymbol{u}}}{\partial \boldsymbol{\mu}}(\boldsymbol{\mu}) \\
\frac{\partial \boldsymbol{u}_n}{\partial \boldsymbol{\mu}} &= \frac{\partial \boldsymbol{u}_{n-1}}{\partial \boldsymbol{\mu}} + \sum_{i=1}^{s} b_i \frac{\partial \boldsymbol{k}_{n,i}}{\partial \boldsymbol{\mu}} \\
\frac{\partial \boldsymbol{u}_{n,i}}{\partial \boldsymbol{\mu}} &= \frac{\partial \boldsymbol{u}_{n-1}}{\partial \boldsymbol{\mu}} + \sum_{j=1}^{i} a_{ij} \frac{\partial \boldsymbol{k}_{n,i}}{\partial \boldsymbol{\mu}} \\
\boldsymbol{M} \frac{\partial \boldsymbol{k}_{n,i}}{\partial \boldsymbol{\mu}} &= \Delta t_n \frac{\partial \boldsymbol{r}}{\partial \boldsymbol{u}} \left(\boldsymbol{u}_{n,i},\ \boldsymbol{\mu},\ t_{n-1} + c_i \Delta t_n \right) \frac{\partial \boldsymbol{u}_{n,i}}{\partial \boldsymbol{\mu}} + \frac{\partial \boldsymbol{r}}{\partial \boldsymbol{\mu}} \left(\boldsymbol{u}_{n,i},\ \boldsymbol{\mu},\ t_{n-1} + c_i \Delta t_n \right)
\end{aligned} \tag{31}$$

for $n = 1, \ldots, N_t$ and $i = 1, \ldots, s$. These equations are solved forward-in-time and therefore the sensitivity simulation can be performed simultaneously with the primal simulation, which eliminates the need to store the primal solution. However, when $N_{\boldsymbol{\mu}}$ is large, this approach becomes intractable due to the large number of linear evolution equations that must be solved. To avoid the computational burden of computing the state sensitivities, the adjoint equations corresponding to the functional F and the corresponding dual variables are introduced to eliminate the state sensitivities from the expression for the total derivative of F with respect to the

parameters, $\boldsymbol{\mu}$. From the derivation of the adjoint equations, an expression for the reconstruction of the gradient of F, independent of the state variable's sensitivities, follows naturally. At this point, it is emphasized that F represents *any* quantity of interest whose gradient is desired, such as the optimization objective function or a constraint.

Let $\boldsymbol{\lambda}_n$ for $n = 0, \ldots, N_t$ be the adjoint variables corresponding to the state update equation in (23) and let $\boldsymbol{\kappa}_{n,i}$ for $n = 1, \ldots, N_t$ and $i = 1, \ldots, s$ be those corresponding to the stage update equations in (23). The adjoint equations are

$$\begin{aligned}
\boldsymbol{\lambda}_{N_t} &= \frac{\partial F}{\partial \boldsymbol{u}_{N_t}}^T \\
\boldsymbol{\lambda}_{n-1} &= \boldsymbol{\lambda}_n + \frac{\partial F}{\partial \boldsymbol{u}_{n-1}}^T + \sum_{i=1}^{s} \Delta t_n \frac{\partial \boldsymbol{r}}{\partial \boldsymbol{u}} \left(\boldsymbol{u}_{n,i},\ \boldsymbol{\mu},\ t_{n-1} + c_i \Delta t_n\right)^T \boldsymbol{\kappa}_{n,i} \\
\boldsymbol{M}^T \boldsymbol{\kappa}_{n,i} &= \frac{\partial F}{\partial \boldsymbol{k}_{n,i}}^T + b_i \boldsymbol{\lambda}_n + \sum_{j=i}^{s} a_{ji} \Delta t_n \frac{\partial \boldsymbol{r}}{\partial \boldsymbol{u}} \left(\boldsymbol{u}_{n,j},\ \boldsymbol{\mu},\ t_{n-1} + c_j \Delta t_n\right)^T \boldsymbol{\kappa}_{n,j}
\end{aligned} \tag{32}$$

for $n = 1, \ldots, N_t$ and $i = 1, \ldots, s$ and the expression for $dF/d\boldsymbol{\mu}$, independent of state sensitivities, is

$$\frac{\mathrm{d}F}{\mathrm{d}\boldsymbol{\mu}} = \frac{\partial F}{\partial \boldsymbol{\mu}} + \boldsymbol{\lambda}_0{}^T \frac{\partial \bar{\boldsymbol{u}}}{\partial \boldsymbol{\mu}} + \sum_{n=1}^{N_t} \Delta t_n \sum_{i=1}^{s} \boldsymbol{\kappa}_{n,i}{}^T \frac{\partial \boldsymbol{r}}{\partial \boldsymbol{\mu}}(\boldsymbol{u}_{n,i},\ \boldsymbol{\mu},\ t_{n-1} + c_i \Delta t_n). \tag{33}$$

Unlike the sensitivity equations in (31), the adjoint equations must be solved backward-in-time and the adjoint simulation cannot begin until the primal simulation completes. This implies that the entire primal time history, including intermediate stages, must be stored. In our setting, this I/O cost is negligible in comparison to the cost of a linear solve with the Jacobian matrix. Furthermore, in contrast to the sensitivity equations, the derivative of the quantity of interest with respect to the state variable appears as a *forcing* term in (32), which requires a separate set of adjoint variables for each quantity of interest whose derivative is sought. In a gradient-based optimization setting, this implies $N_c + 1$, where N_c is the number of state-dependent constraints, adjoint solves are required to compute the gradient of the objective function and all constraint functions. While the number of adjoint solves depends on the number of functionals to differentiate, it is independent of the number of parameters. Since the application in this work is in the regime where $N_{\boldsymbol{\mu}} > N_c + 1$, the adjoint method is more desirable.

For the derivation of Equations (32)–(33), the reader is referred to [62, 63]. For the adjoint equations that explicitly enforce time periodicity of the solution of the partial differential equation, see [64]. From inspection of (33), it is clear that the initial condition sensitivity $\dfrac{\partial \bar{\boldsymbol{u}}}{\partial \boldsymbol{\mu}}$ is the only sensitivity term required to reconstruct $\dfrac{\mathrm{d}F}{\mathrm{d}\boldsymbol{\mu}}$. The derivation of this term for the uniform flow initial condition introduced in

Section 2.3 is provided in the next section. From the expression for the fully discrete quantity of interest in (30), it is clear that F is *independent* of $\boldsymbol{u}_{N_t}$, which implies

$$\boldsymbol{\lambda}_{N_t} = \frac{\partial F}{\partial \boldsymbol{u}_{N_t}}^T = \boldsymbol{0}. \tag{34}$$

Furthermore, the partial derivatives of the fully discrete quantities of interest are

$$\begin{aligned}
\frac{\partial F}{\partial \boldsymbol{u}_n} &= \Delta t_n \sum_{i=1}^{s} b_i \frac{\partial f^h}{\partial \boldsymbol{u}}(\boldsymbol{u}_{n,i},\ \boldsymbol{\mu},\ t_{n-1} + c_i \Delta t_n) \qquad n = 0,\ \ldots,\ N_t - 1 \\
\frac{\partial F}{\partial \boldsymbol{k}_{n,j}} &= \Delta t_n \sum_{i=j}^{s} a_{ij} b_i \frac{\partial f^h}{\partial \boldsymbol{u}}(\boldsymbol{u}_{n,i},\ \boldsymbol{\mu},\ t_{n-1} + c_i \Delta t_n) \quad n = 1,\ \ldots,\ N_t,\ \ j = 1,\ \ldots,\ s
\end{aligned} \tag{35}$$

See [63] for a discussion of the benefits of the fully discrete adjoint framework over the continuous or semi-discrete ones in the context of optimization *or* when a Runge-Kutta temporal discretization is used.

4.2 Parametrization of the Initial Condition

Recall the form of the ALE-transformed uniform flow initial condition in (16). Since the physical uniform flow state $\bar{\boldsymbol{U}}(\boldsymbol{X})$ is parameter-independent, the sensitivity of the initial condition will be due solely to the sensitivity of the determinant of the deformation gradient. That is,

$$\begin{aligned}
\frac{\partial \boldsymbol{U}_{\bar{X}}}{\partial \boldsymbol{\mu}}(\boldsymbol{X},\ \boldsymbol{\mu},\ 0) &= \bar{\boldsymbol{U}}(\boldsymbol{X}) \frac{\partial g_0}{\partial \boldsymbol{\mu}}(\boldsymbol{X},\ \boldsymbol{\mu}) \\
\frac{\partial \bar{g}}{\partial \boldsymbol{\mu}}(\boldsymbol{X},\ \boldsymbol{\mu},\ 0) &= \frac{\partial g_0}{\partial \boldsymbol{\mu}}(\boldsymbol{X},\ \boldsymbol{\mu})
\end{aligned} \tag{36}$$

The initial condition sensitivity at the semi-discrete or fully discrete level is then the appropriate vectorization of this quantity over the DG mesh.

4.3 Parametrization of the Residual and Quantities of Interest

In addition to the initial condition sensitivity, the equation to reconstruct the total derivative of F with respect to $\boldsymbol{\mu}$ requires the partial derivatives of the residual and quantity of interest with respect to the $\boldsymbol{\mu}$. For this purpose, we assume that the parameter vector $\boldsymbol{\mu}$ purely controls the domain deformation, e.g., it does not

affect the boundary conditions or material properties. Then, given the discussion in Section 3.1 that completely defines the ALE map based on its action and the action of its time derivative on the nodes of the mesh, the parameter dependence of the residual and quantity of interest can be written in terms of $\boldsymbol{x}(\boldsymbol{\mu})$ and $\dot{\boldsymbol{x}}(\boldsymbol{\mu})$. That is,

$$\begin{aligned} \frac{\partial \boldsymbol{r}}{\partial \boldsymbol{\mu}} &= \frac{\partial \boldsymbol{r}}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\mu}} + \frac{\partial \boldsymbol{r}}{\partial \dot{\boldsymbol{x}}} \frac{\partial \dot{\boldsymbol{x}}}{\partial \boldsymbol{\mu}} \\ \frac{\partial f^h}{\partial \boldsymbol{\mu}} &= \frac{\partial f^h}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\mu}} + \frac{\partial f^h}{\partial \dot{\boldsymbol{x}}} \frac{\partial \dot{\boldsymbol{x}}}{\partial \boldsymbol{\mu}} \end{aligned} \tag{37}$$

The form of the ALE map, i.e., $\boldsymbol{x}(\boldsymbol{\mu})$ and $\dot{\boldsymbol{x}}(\boldsymbol{\mu})$, will be described in the next section. Our implementation uses the Maple software [31] to compute all required partial derivatives.

5 Application to Energetically Optimal Flapping Flight

In this section, the high-order numerical discretization of the isentropic, compressible Navier-Stokes equations and corresponding adjoint method are applied to determine the energetically optimal flapping motion of a three-dimensional wing geometry using gradient-based optimization in the low Reynolds number regime of $Re = 1000$. For a physically relevant mission, a requirement is placed on the time-averaged thrust, which leads to an optimization problem with a nonlinear constraint. As a result, two adjoint equations must be solved at each optimization iteration to compute the gradient of the objective function and the nonlinear constraint.

5.1 Flapping Wing Geometry and Kinematics

The wing geometry considered in this work is an extruded NACA0012 airfoil with a rounded tip to accurately capture three-dimensional effects. In the reference configuration, the NACA0012 airfoil is contained in the $X_1 - X_3$ plane corresponding to $X_2 = 0$, facing the $-X_1$ direction (flow in the $+X_1$ direction), and extruded in the $+X_2$ direction for the span-to-chord ratio of 2. A symmetry plane is included to consider an isolated wing without a fuselage. The fluid domain is discretized using a curved mesh with tetrahedral elements of degree $p = 3$. Figure 2 visualizes the mesh and the corresponding geometry is taken as the reference domain in ALE setting.

The flapping motion is parametrized using three angles: the flapping angle $\theta(\boldsymbol{\mu}, t)$ (rotation about the X_1-axis), the pitching angle $\alpha(\boldsymbol{\mu}, t)$ (rotation about the X_2-axis), and the sweeping angle $\beta(\boldsymbol{\mu}, t)$ (rotation about the X_3-axis). The origin of the flapping angle is taken as the intersection of the $X_2 = s_1$ and $X_3 = 0$ planes, where $s_1 > 0$ is a parameter that defines a shoulder away from the symmetry plane.

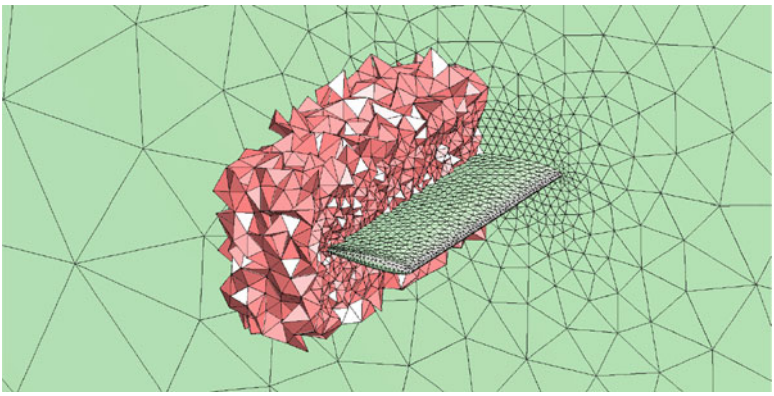

Fig. 2 Surface mesh of the wing and the symmetry plane, and some of the tetrahedral elements in the wake. All elements are curved by polynomials of degree $p = 3$

The origin of the pitching and sweeping angles are taken as the intersection of the $X_1 = 0$, $X_3 = 0$, and $X_1 = 0$, $X_2 = 0$ planes, respectively. The combination of these motions takes the form

$$\begin{aligned} x_1'(\boldsymbol{X},\, \boldsymbol{\mu},\, t) &= X_1 \cos(\alpha(\boldsymbol{\mu},\, t)) + (X_2 - s_1) \sin(\beta(\boldsymbol{\mu},\, t)) - X_3 \sin(\alpha(\boldsymbol{\mu},\, t)) \\ x_2'(\boldsymbol{X},\, \boldsymbol{\mu},\, t) &= s_1 + X_2 \cos(\theta(\boldsymbol{\mu},\, t)) \cos(\beta(\boldsymbol{\mu},\, t)) - X_3 \sin(\theta(\boldsymbol{\mu},\, t)) \\ &\quad - X_1 \sin(\beta(\boldsymbol{\mu},\, t)) \\ x_3'(\boldsymbol{X},\, \boldsymbol{\mu},\, t) &= X_3 \cos(\theta(\boldsymbol{\mu},\, t)) \cos(\alpha(\boldsymbol{\mu},\, t)) + (X_2 - s_1) \sin(\theta(\boldsymbol{\mu},\, t)) \\ &\quad + X_1 \sin(\alpha(\boldsymbol{\mu},\, t)), \end{aligned} \tag{38}$$

where we set the parameter $s_1 = 0.5$. While this kinematic description encodes exactly the desired motion of the wing itself, it cannot be applied to the entire fluid domain as it will not preserve the symmetry plane and the rotations will lead to large velocities at the far field. To avoid these issues, the domain deformation is smoothly blended to zero near the symmetry plane and away from the wing, following the work in [38, 63].

The deformation blending away from the wing is defined as a composition of a radial blending, $b_{xz}(\boldsymbol{X})$, in the $X_1 - X_3$ plane and a unidirectional blending, $b_y(\boldsymbol{X})$, in the $+X_2$ direction. These blendings take the form

$$b_{xz}(\boldsymbol{X}) = \begin{cases} 0 & d(\boldsymbol{X}) \leq r_1 \\ 1 & d(\boldsymbol{X}) \geq r_1 + r_2, \\ q\left(\frac{d(\boldsymbol{X}) - r_1}{r_2}\right) & \text{otherwise} \end{cases} \quad b_y(\boldsymbol{X}) = \begin{cases} 0 & X_2 \leq y_1 \\ 1 & X_2 \geq y_1 + y_2, \\ q\left(\frac{X_2 - y_1}{y_2}\right) & \text{otherwise} \end{cases} \tag{39}$$

where $d(\boldsymbol{X}) = \sqrt{X_1^2 + X_3^2}$ is the radial distance from the X_2 axis (the axis through the center of the wing in the spanwise direction) and $q(s) = 3s^2 - 2s^3$ is the cubic blending introduced in [38]. For smoother spatial blendings, the quintic expression $q(s) = 10s^3 - 15s^4 + 6s^5$ could be used instead. See Figure 3 for the blendings $b_{xz}(\boldsymbol{X})$ and $b_y(\boldsymbol{X})$ with the values of the blending parameters used in this work: $r_1 = 0.6$, $r_2 = 5$, $y_1 = 2.6$, $y_2 = 5$.

Suppose we want to compose two blendings, $b_1(\boldsymbol{X})$ and $b_2(\boldsymbol{X})$, in serial, that is, blend a deformed domain $\boldsymbol{x}'$ with an undeformed domain $\boldsymbol{X}$ via $b_1(\boldsymbol{X})$ and blend the result with the undeformed domain via $b_2(\boldsymbol{X})$ as follows

$$\begin{aligned} \boldsymbol{x}'' &= (1 - b_1(\boldsymbol{X}))\boldsymbol{x}' + b_1(\boldsymbol{X})\boldsymbol{X} \\ \boldsymbol{x} &= (1 - b_2(\boldsymbol{X}))\boldsymbol{x}'' + b_2(\boldsymbol{X})\boldsymbol{X}. \end{aligned} \tag{40}$$

This can be compactly expressed as a single blending $b_{12}(\boldsymbol{X})$ as $\boldsymbol{x} = (1 - b_{12}(\boldsymbol{X}))\boldsymbol{x}' + b_{12}(\boldsymbol{X})\boldsymbol{X}$, where

$$b_{12}(\boldsymbol{X}) = b_1(\boldsymbol{X}) + b_2(\boldsymbol{X}) - b_1(\boldsymbol{X})b_2(\boldsymbol{X}). \tag{41}$$

Therefore, the composition of the radial and unidirectional blending in (39) leads to a cylindrical blending that takes the form

$$b_{cyl}(\boldsymbol{X}) = b_{xz}(\boldsymbol{X}) + b_y(\boldsymbol{X}) - b_{xz}(\boldsymbol{X})b_y(\boldsymbol{X}). \tag{42}$$

To ensure that the symmetry plane remains motionless, the mapping in (38) must be smoothly blended to 0 at the $X_2 = 0$ plane. The blending at the symmetry plane, $b_{sym}(\boldsymbol{X})$, is chosen to be infinitely smooth and the rate of decay decreases with increasing radial distance from the X_2 axis to prevent mesh entanglement, i.e.,

$$b_{sym}(\boldsymbol{X}) = e^{-(X_2/(s_2+s_3 d(\boldsymbol{X})))^2}. \tag{43}$$

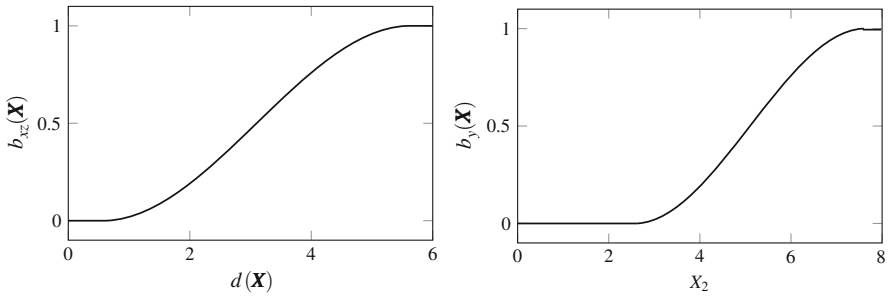

Fig. 3 *Left*: radial blending $b_{xz}(\boldsymbol{X})$ corresponding to $r_1 = 0.6$, $r_2 = 5$. *Right*: unidirectional blending $b_y(\boldsymbol{X})$ corresponding to $y_1 = 2.6$, $y_2 = 5$

The blending parameter s_2 is set to 1 for geometrical considerations since this affects the geometry of the wing during the flapping motion. The blending parameter s_3 is free in the sense that it has little effect on the wing itself and is solely used to improve mesh quality in the fluid domain. A brute force, unidimensional search is performed to determine the value of $s_3 = 0.3$ that maximizes the mesh quality. See Figure 4 for a plot of $b_{sym}(\boldsymbol{X})$ with these blending parameters at various radial positions.

The composition of the cylindrical blending $b_{cyl}(\boldsymbol{X})$ and symmetry blending $b_{sym}(\boldsymbol{X})$ using the formula in (41) leads to the final form of the spatial blending

$$b(\boldsymbol{X}) = b_{cyl}(\boldsymbol{X}) + b_{sym}(\boldsymbol{X}) - b_{cyl}(\boldsymbol{X}) b_{sym}(\boldsymbol{X}) \tag{44}$$

and the expression for the deformed domain

$$\boldsymbol{x}''(\boldsymbol{X},\ \boldsymbol{\mu},\ t) = (1 - b(\boldsymbol{\phi}(\boldsymbol{X})))\boldsymbol{x}'(\boldsymbol{\phi}(\boldsymbol{X}),\ \boldsymbol{\mu},\ t) + b(\boldsymbol{\phi}(\boldsymbol{X}))\boldsymbol{\phi}(\boldsymbol{X}). \tag{45}$$

The above expression uses $\boldsymbol{\phi}(\boldsymbol{X})$, the coordinates in the domain with curved boundaries, in place of $\boldsymbol{X}$, the coordinates in the straight-sided domain, due to the choice discussed in Section 3.1 that incorporates the curving of the domain boundaries in the ALE map. Spatial blending of this form ensures that the desired physical motion of the body, $\boldsymbol{x}'(\boldsymbol{X},\ \boldsymbol{\mu},\ t)$, is exactly achieved near the surface of the wing, there is no deformation far from the surface or at the symmetry plane, and the domain deformation smoothly varies between these extremes.

The expression for the deformed domain, $\boldsymbol{x}'(\boldsymbol{X},\ \boldsymbol{\mu},\ t)$, in (38) will have a nontrivial deformation and velocity at $t = 0$. This may cause difficulty in initializing the simulation from uniform flow as violent transients will result that may prevent convergence of the nonlinear solvers. For this reason, following the work in [51, 63], the deformation is smoothly blended to zero at $t = 0$ using the infinitely differentiable blending

$$b_t(t) = e^{-(t/T_c)^2}. \tag{46}$$

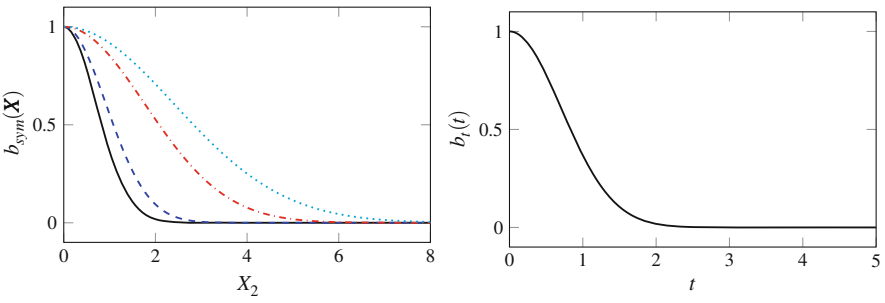

Fig. 4 *Left*: blending at symmetry plane $b_{sym}(\boldsymbol{X})$ corresponding to $s_2 = 1.0$, $s_3 = 0.3$ at radial position $d(\boldsymbol{X}) = 0$ (——), $d(\boldsymbol{X}) = 1$ (- - -), $d(\boldsymbol{X}) = 5$ (-·-·-), $d(\boldsymbol{X}) = 8$ (·····). *Right*: temporal blending $b_t(t)$ corresponding to $T_c = 1$

Temporal blendings have also been used in experimental studies involving flapping wings [18], where a quintic blending was used. The final form of the deformed domain is

$$\boldsymbol{x}(\boldsymbol{X},\ \boldsymbol{\mu},\ t) \coloneqq (1 - b_t(t))\boldsymbol{x}''(\boldsymbol{\phi}(\boldsymbol{X}),\ \boldsymbol{\mu},\ t) + b_t(t)\boldsymbol{\phi}(\boldsymbol{X}) \tag{47}$$

and the domain velocity $\dot{\boldsymbol{x}}(\boldsymbol{X},\ \boldsymbol{\mu},\ t)$ can be computed analytically. It can easily be verified that this temporal blending guarantees $\boldsymbol{x}(\boldsymbol{X},\ \boldsymbol{\mu},\ 0) = \boldsymbol{\phi}(\boldsymbol{X})$ and $\dot{\boldsymbol{x}}(\boldsymbol{X},\ \boldsymbol{\mu},\ 0) = \boldsymbol{0}$. In this work, $T_c = T/5$, where T is the period of the flapping motion, to ensure that $\boldsymbol{x}$, $\dot{\boldsymbol{x}}$ are effectively equal to $\boldsymbol{x}''$, $\dot{\boldsymbol{x}}''$ (within 0.1%) by $1/2$ a period (see Figure 4). This blending limits the transients that result from initializing the flow with incompatible boundary conditions at the viscous wall. Another implication of this temporal blending is that the sensitivity of the initial condition is zero, i.e., $\frac{\partial \boldsymbol{u}_0}{\partial \boldsymbol{\mu}} = 0$, since $\boldsymbol{x}(\boldsymbol{X},\ \boldsymbol{\mu},\ 0) = \boldsymbol{\phi}(\boldsymbol{X})$. Finally, as discussed in Section 3.1, once the ALE-mapped domain $\boldsymbol{x}(\boldsymbol{X},\ \boldsymbol{\mu},\ t)$ and velocity $\dot{\boldsymbol{x}}(\boldsymbol{X},\ \boldsymbol{\mu},\ t)$ are computed, the remaining quantities required for the ALE formulation of the governing equations, namely $\boldsymbol{G}(\boldsymbol{X},\ \boldsymbol{\mu},\ t)$ and $g(\boldsymbol{X},\ \boldsymbol{\mu},\ t)$, can be computed through differentiation of the underlying shape functions, as in (22).

Given this kinematic description of the flapping motion in (47), all that remains to completely specify the domain deformation and its parametrization is the functional form of the pitching, sweeping, and flapping angles. In this work, these angles are parametrized through a single harmonic function each as

$$\begin{aligned}
\alpha(\boldsymbol{\mu},\ t) &= \mu_1 + \mu_2 \sin(2\pi f t + \mu_3) \\
\beta(\boldsymbol{\mu},\ t) &= \mu_4 + \mu_5 \sin(2\pi f t + \mu_6) \\
\theta(\boldsymbol{\mu},\ t) &= \mu_7 + \mu_8 \sin(2\pi f t + \mu_9),
\end{aligned} \tag{48}$$

where $f = 1/T$ is the flapping frequency. Even though the flapping frequency is an important design consideration, it will not be taken as a parameter in this work as properly accounting for frequency perturbations in the fully discrete adjoint framework is still a research issue [53] and will be the subject of future work. An example of a typical flapping motion is shown in Figure 5. The same flapping motion is shown in Figures 5, 6, 7, and 8, where the various views of an unstructured volumetric mesh with 10805 $p = 3$ elements are provided to show the impact of the blending and the high-quality elements that are maintained.

5.2 *Energetically Optimal Flapping Under a Thrust Constraint*

The high-order numerical discretization of the isentropic, compressible Navier-Stokes equations and corresponding adjoint method are applied to determine the energetically optimal flapping motion of the geometry introduced in the previous section using gradient-based optimization techniques in the low Reynolds number

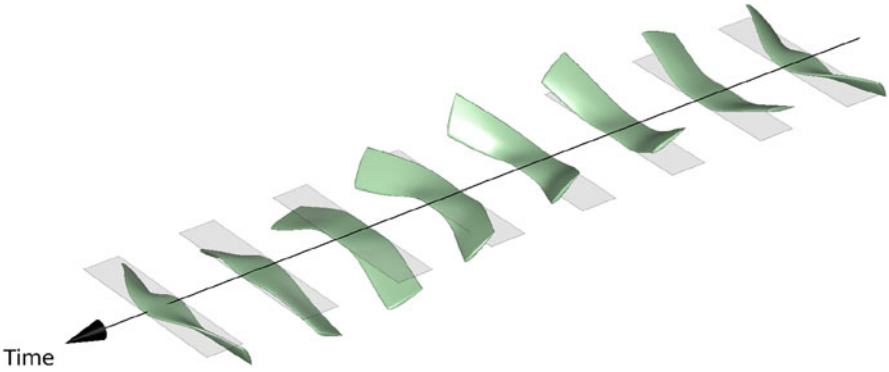

Fig. 5 Snapshots of the flapping motion in (47) with $\mu_1 = \mu_4 = \mu_7 = \mu_9 = 0$, $\mu_3 = -\mu_6 = -\pi/3$, $\mu_2 = 60°$, $\mu_5 = -\mu_8 = -25°$

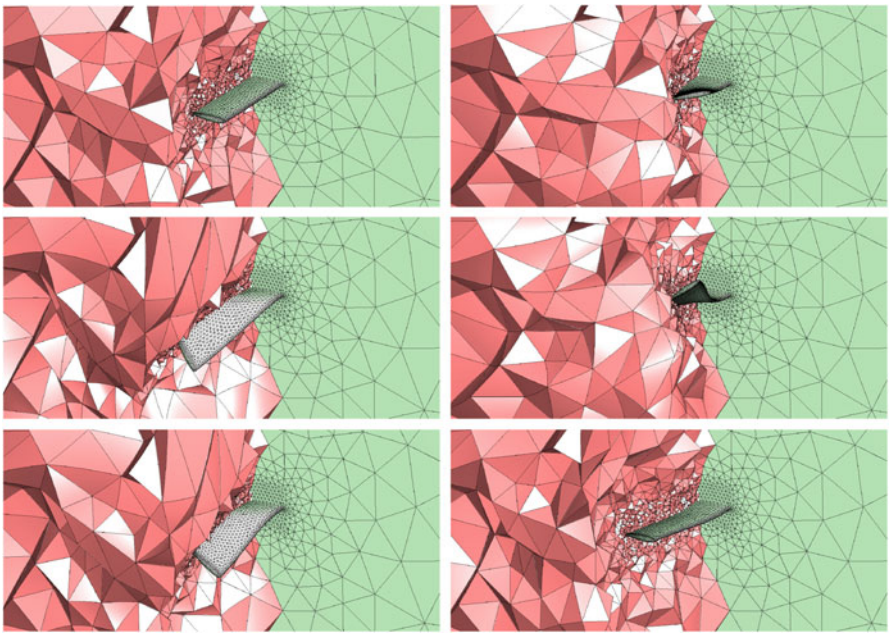

Fig. 6 Snapshots of a slice of the volumetric mesh in the $X_1 - X_3$ and $X_2 - X_3$ planes corresponding to the flapping motion in (47) with $\mu_1 = \mu_4 = \mu_7 = \mu_9 = 0$, $\mu_3 = -\mu_6 = -\pi/3$, $\mu_2 = 60°$, $\mu_5 = -\mu_8 = -25°$. The top left figure corresponds to the *curved mesh* with no other deformation applied, i.e., $\boldsymbol{x} = \boldsymbol{\phi}(\boldsymbol{X})$. The remaining figures correspond to snapshots (top to bottom, left to right) taken at equally spaced time increments during the second period that correspond to times $t = 5.0,\ 6.0,\ 7.0,\ 8.0,\ 9.0$

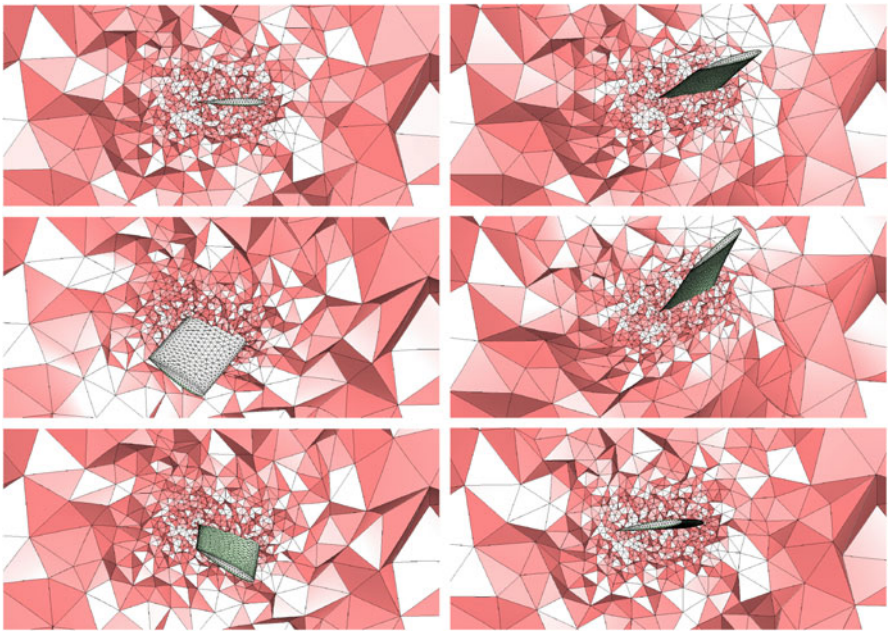

Fig. 7 Snapshots of a slice of the volumetric mesh in the $X_1 - X_3$ plane corresponding to the flapping motion in (47) with $\mu_1 = \mu_4 = \mu_7 = \mu_9 = 0$, $\mu_3 = -\mu_6 = -\pi/3$, $\mu_2 = 60°$, $\mu_5 = -\mu_8 = -25°$. The top left figure corresponds to the *curved mesh* with no other deformation applied, i.e., $\boldsymbol{x} = \boldsymbol{\phi}(\boldsymbol{X})$. The remaining figures correspond to snapshots (top to bottom, left to right) taken at equally spaced time increments during the second period that correspond to times $t = 5.0, 6.0, 7.0, 8.0, 9.0$

regime of $Re = 1000$. For a physically relevant mission, requirements are placed on the time-averaged thrust leading to an optimization problem with a nonlinear constraint. As a result, two adjoint equations must be solved at each optimization iteration to compute the gradient of the objective function and nonlinear constraint. From (32), it is clear that the linear system that arises at each stage of each time step is the same for each functional; the only difference is the right-hand side, which presents an opportunity to use some fast multiple right-hand side solver [8, 45]; however, this was not done in this work.

The DG-ALE scheme introduced in Section 2 is used for the spatial discretization of the system of conservation laws with polynomial order $p = 3$ (for both the geometry and solution representation) and a diagonally implicit Runge-Kutta scheme for the temporal discretization. The DG-ALE scheme uses the Roe flux [41] for the inviscid numerical flux and the Compact DG flux [37] for the viscous numerical flux. The Butcher tableau for the three-stage, third-order DIRK scheme considered in this work is given in Table 2. Since the present study looks to find the energetically optimal flapping motion subject to a constraint on the thrust, the quantities of interest for the optimization problem are the average work done on the fluid by the wing, $\mathscr{W}(\boldsymbol{U}, \boldsymbol{\mu})$, and thrust, $\mathscr{T}_x(\boldsymbol{U}, \boldsymbol{\mu})$, over one flapping

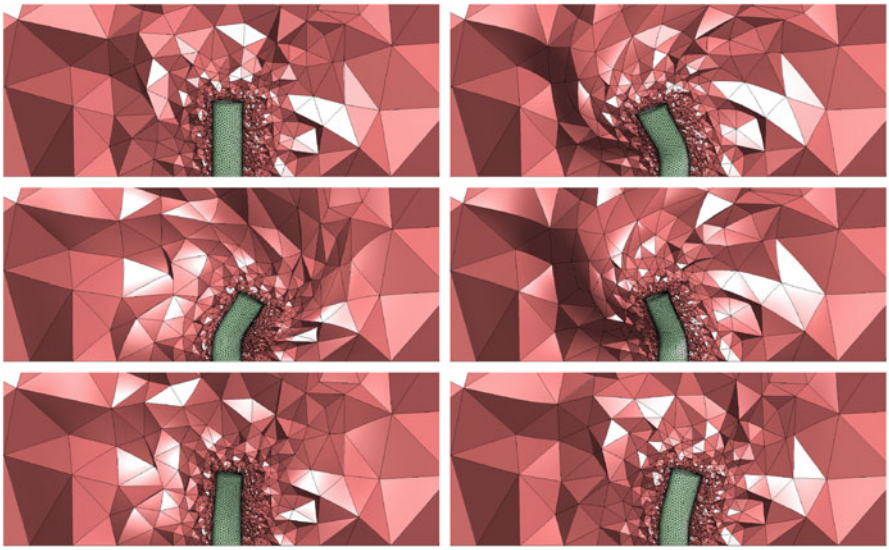

Fig. 8 Snapshots of a slice of the volumetric mesh in the $X_1 - X_2$ plane corresponding to the flapping motion in (47) with $\mu_1 = \mu_4 = \mu_7 = \mu_9 = 0$, $\mu_3 = -\mu_6 = -\pi/3$, $\mu_2 = 60°$, $\mu_5 = -\mu_8 = -25°$. The top left figure corresponds to the *curved mesh* with no other deformation applied, i.e., $\boldsymbol{x} = \boldsymbol{\phi}(\boldsymbol{X})$. The remaining figures correspond to snapshots (top to bottom, left to right) taken at equally spaced time increments during the second period that correspond to times $t = 5.0,\ 6.0,\ 7.0,\ 8.0,\ 9.0$

Table 2 Butcher tableau for 3-stage, 3rd order DIRK scheme [1] $\alpha = 0.435866521508459$, $\gamma = -\frac{6\alpha^2-16\alpha+1}{4}$, $\omega = \frac{6\alpha^2-20\alpha+5}{4}$

α	α		
$\frac{1+\alpha}{2}$	$\frac{1+\alpha}{2} - \alpha$	α	
1	γ	ω	α
	γ	ω	α

period. To ensure that the transients that result from initializing the simulation from nonperiodic flow (uniform flow in this case) do not pollute the time-averaged quantities, two full periods of the flapping motion are simulated and the quantities are averaged over only the final period. Therefore, the time-averaged quantities are defined as

$$\begin{aligned} \mathscr{W}(\boldsymbol{U}, \boldsymbol{\mu}) &= -\frac{1}{T} \int_T^{2T} \int_{\boldsymbol{\Gamma}} \boldsymbol{f}(\boldsymbol{U}, \boldsymbol{\mu}) \cdot \dot{\boldsymbol{x}} \, dS \, dt \\ \mathscr{T}_x(\boldsymbol{U}, \boldsymbol{\mu}) &= -\frac{1}{T} \int_T^{2T} \int_{\boldsymbol{\Gamma}} \boldsymbol{f}(\boldsymbol{U}, \boldsymbol{\mu}) \cdot \boldsymbol{e}_1 \, dS \, dt \end{aligned} \tag{49}$$

where $\boldsymbol{\Gamma}$ is the surface of the wing, $\boldsymbol{f} \in \mathbb{R}^3$ is the force imparted by the fluid on the body, $\boldsymbol{e}_i \in \mathbb{R}^3$ is the ith canonical basis vector, and $\dot{\boldsymbol{x}}$ is the velocity of each point

on $\boldsymbol{\Gamma}$. The negative sign in the definition of the thrust is required since the flow is in the $+X_1$-direction and, therefore, a positive thrust is directed in the $-X_1$-direction.

The initialization and integration strategy described is a commonly used and crude approximation to the ideal situation of initializing the simulation with the state that will induce a time-periodic flow, which will ensure that the simulation is completely free of unphysical initial transients. A method to initialize a simulation with this state was introduced in [64] as well as the corresponding adjoint method to allow for optimization under a time-periodicity constraint.

Finally, let the fully discrete, high-order approximation of the integrated quantities of interest (DG in space, DIRK in time) in (25) be denoted with the corresponding Roman symbol, e.g., $W(\boldsymbol{u}^{(0)}, \ldots, \boldsymbol{u}^{(N_t)}, \boldsymbol{k}_1^{(n)}, \ldots, \boldsymbol{k}_s^{(n)}, \boldsymbol{\mu})$ is the fully discrete approximation of $\mathscr{W}(\boldsymbol{U}, \boldsymbol{\mu})$ and similarly for T_x. Then, the fully discrete optimization problem of interest takes the form

$$
\begin{aligned}
\underset{\substack{\boldsymbol{u}^{(0)},\ \ldots,\ \boldsymbol{u}^{(N_t)} \in \mathbb{R}^{N_{\boldsymbol{u}}},\\ \boldsymbol{k}_1^{(1)},\ \ldots,\ \boldsymbol{k}_s^{(N_t)} \in \mathbb{R}^{N_{\boldsymbol{u}}},\\ \boldsymbol{\mu} \in \mathbb{R}^{N_{\boldsymbol{\mu}}}}}{\text{minimize}} \quad & W(\boldsymbol{u}^{(0)},\ \ldots,\ \boldsymbol{u}^{(N_t)},\ \boldsymbol{k}_1^{(1)},\ \ldots,\ \boldsymbol{k}_s^{(N_t)},\ \boldsymbol{\mu}) \\
\text{subject to} \quad & T_x(\boldsymbol{u}^{(0)},\ \ldots,\ \boldsymbol{u}^{(N_t)},\ \boldsymbol{k}_1^{(1)},\ \ldots,\ \boldsymbol{k}_s^{(N_t)},\ \boldsymbol{\mu}) \geq \bar{T}_x \\
& \boldsymbol{u}^{(0)} = \boldsymbol{u}_0 \\
& \boldsymbol{u}^{(n)} = \boldsymbol{u}^{(n-1)} + \sum_{i=1}^{s} b_i \boldsymbol{k}_i^{(n)} \\
& \boldsymbol{M}\boldsymbol{k}_i^{(n)} = \Delta t_n \boldsymbol{r}\left(\boldsymbol{u}_i^{(n)},\ \boldsymbol{\mu},\ t_{n-1} + c_i \Delta t_n\right),
\end{aligned}
\tag{50}
$$

where $\bar{T}_x$ is a lower bound on the thrust. In this work, $\bar{T}_x = 0$ is taken to ensure that the flapping motion generates sufficient thrust to overcome the induced drag on the wing. In this section, the parameters $\mu_1 = \mu_4 = \mu_7 = 0$ and $\mu_9 = \pi/2$ are frozen, which leads to a 5-parameter optimization problem in the all the amplitudes (μ_2, μ_5, μ_8) and pitch and sweep phases (μ_3, μ_6).

The optimization solver used in this work is IPOPT [52], a nonlinearly constrained interior point method. Figure 9 contains the trajectory of α, β, θ that define initial guess and solution of the optimization problem in (50). The initial guess for the optimization problem is a pure flapping motion, i.e., $\alpha(\boldsymbol{\mu}_0,\ t) = \beta(\boldsymbol{\mu}_0,\ t) = 0$. In general, a quality initial guess is important since the solution of non-convex optimization problems, such as this one, is dependent on the starting point. In a practical design setting, the goal is to improve an existing or baseline design, which will usually constitute a reasonable starting guess for the optimizer. Another strategy for generating reasonable initial guesses is to perform homotopy on the thrust constraint. The optimal solution increases the flapping amplitude from 15° to 32.1°, increases the pitch amplitude from 0° to 31.3°, and only incorporates a negligible amount of dynamic sweeping (0.02°). The optimal phase angle between the flapping and pitch motions is determined to be 87.1°.

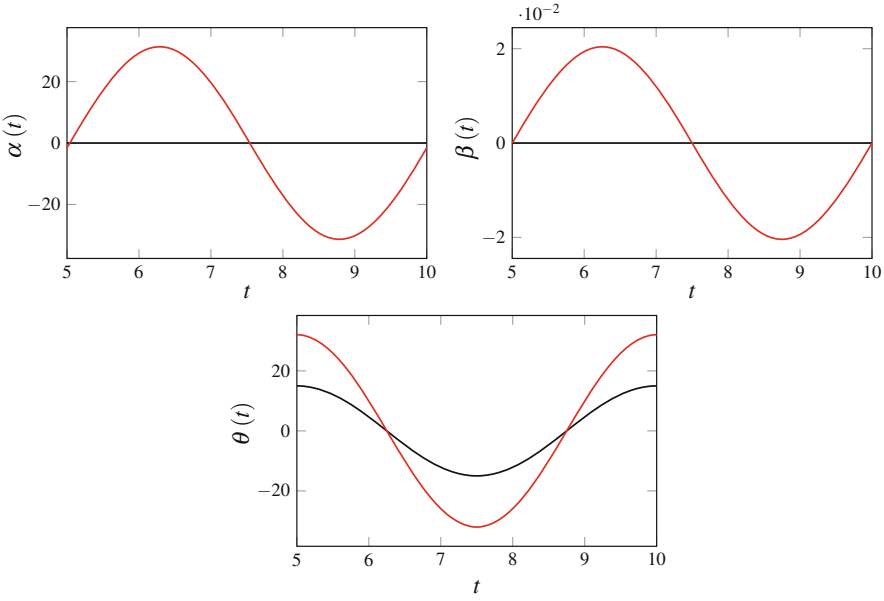

Fig. 9 Trajectories of $\alpha(t)$, $\beta(t)$, and $\theta(t)$, in degrees, at the initial guess (——) and solution (——) of the optimization problem in (50)

The instantaneous quantities of interest for the nominal motion and solution of (50) are included in Figure 10. It is clear that the optimal motion requires more work than the nominal motion to overcome the induced drag on the wing and satisfy the thrust constraint.

Figure 11 shows the convergence of the integrated quantities of interest with iterations in the optimization solver. It can be seen that, initially, the thrust constraint is violated and after only 2 optimization iterations, the flapping motion has become sufficient to overcome the induced drag and satisfy the thrust constraint, at the cost of additional energy that must be input to the system. After 10 iterations, the thrust constraint is satisfied and reduction of the work has essentially ceased. At the optimal solution, the thrust constraint is *active* and satisfied to 9 digits of accuracy.

The trajectory of the wing and isosurfaces of the surrounding flow are shown in Figure 12 (nominal) and Figure 13 (optimal). The flow around the nominal trajectory is fairly benign in that there is little flow separation, does not require much energy, and the generated thrust is not sufficient to overcome the induced drag. In contrast, the optimal trajectory flaps "harder" (larger flapping and pitching amplitudes) in order to generate sufficient thrust to satisfy the constraint. The result is more separation, even though the additional pitching helps streamline the flow, and more required energy.

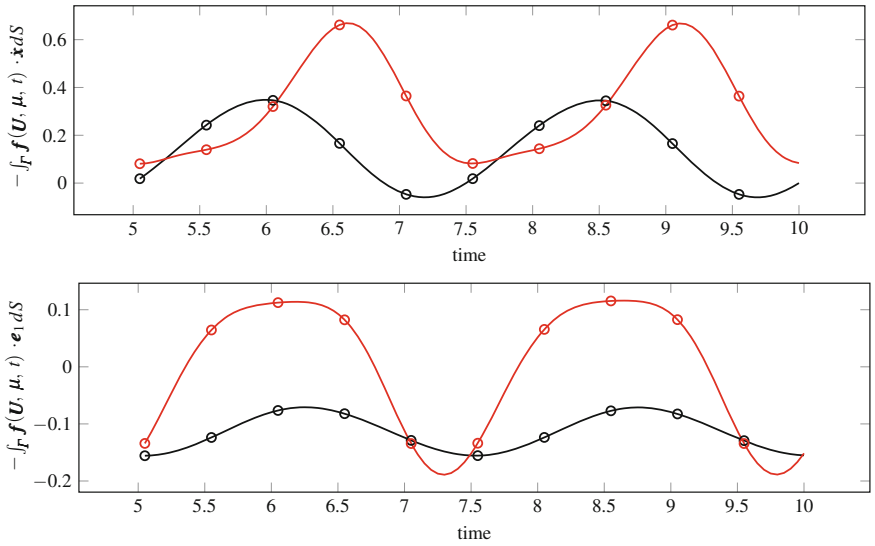

Fig. 10 Time of the total power (top) and x-directed force (bot tom) *imparted onto the fluid by the airfoil* at initial guess (—o—) and optimal solution (—o—)

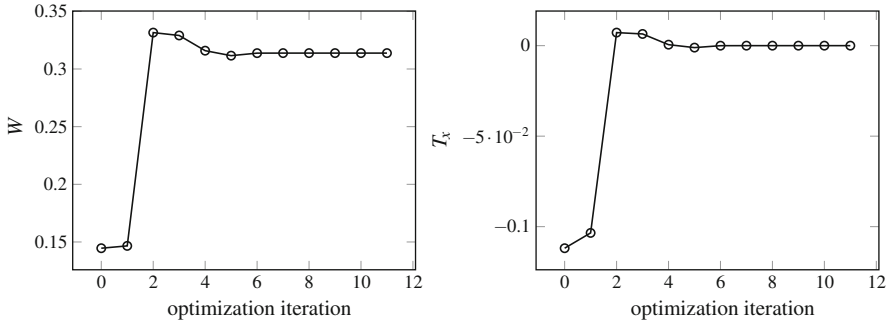

Fig. 11 Convergence of quantities of interest, W and T_x, with optimization iteration. Each iteration requires a primal and adjoint flow computation to compute the quantities of interest and their gradients, respectively

6 Conclusion

This work presents a framework for using high-order numerical discretizations to solve optimization problems constrained by deforming domain conservation laws and demonstrates its potential on the large-scale application of determining energetically optimal flapping motions of a three-dimensional wing. The high-order numerical method employs a discontinuous Galerkin spatial discretization

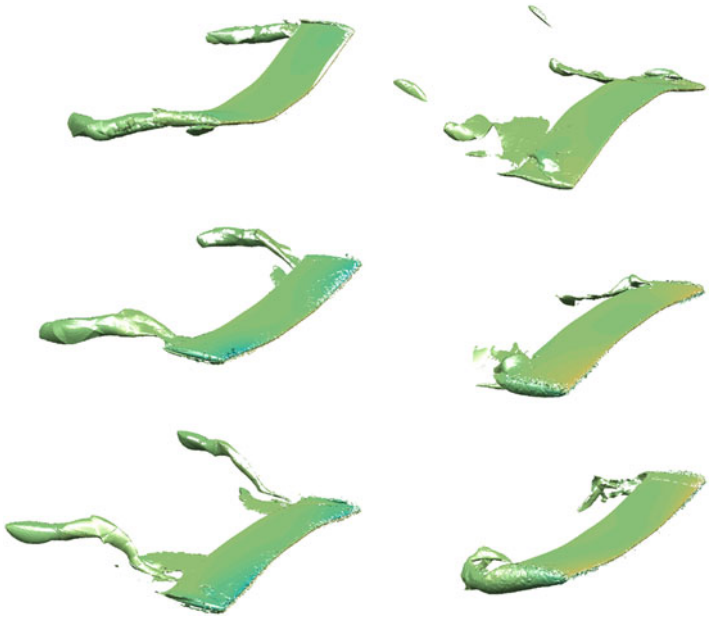

Fig. 12 Visualization of the flow field around wing with flapping motion corresponding to the *initial guess* for the optimization problem in (50). The color shows the pressure field on the wing surface as well as on an isosurface of the streamwise vorticity. Snapshots (top to bottom, left to right) taken at equally spaced time increments during the second period that correspond to times $t = 5.0,\ 5.83,\ 6.67,\ 7.5,\ 8.33,\ 9.17$

and diagonally implicit Runge-Kutta temporal discretization for both the ALE-transformed conservation law and its quantities of interest. The fully discrete adjoint method was used to compute gradients of quantities of interest to ensure that they are discretely consistent and the cost of computing them only scales weakly with the number of parameters. This framework only required 12 iterations when coupled with the nonlinear optimizer IPOPT to solve the relevant problem of finding a thrust-neutral flapping trajectory that minimizes the energy required to complete the motion.

The framework presented is sufficiently general to handle a number of relevant generalizations such as shape optimization of the wing cross-section and planform, more general spline-based parametrizations, and the inclusion of other aerodynamic constraints. The ALE framework is capable of handling completely general domain deformations, which includes static changes to the shape of the wing in a shape-only or combined shape and trajectory optimization setting. A more general parametrization can also easily be included by using a spline-based parametrization of the flapping angles in (48) and the expanded design space would likely lead to better designs. Finally, other aerodynamic constraints can easily be incorporated in the optimization problem in (50) at the cost of an additional adjoint solve for each additional constraint (that depends on the PDE solution).

Fig. 13 Visualization of the flow field around wing with flapping motion corresponding to the *solution* of the optimization problem in (50). The color shows the pressure field on the wing surface as well as on an isosurface of the streamwise vorticity. Snapshots (top to bottom, left to right) taken at equally spaced time increments during the second period that correspond to times $t = 5.0, 5.83, 6.67, 7.5, 8.33, 9.17$

While this work is one step toward solving optimization problems of engineering and scientific relevance, further development is required to have an impact in practice. This work has considered a pure fluid problem and treats the structure as rigid, which is not realistic, particularly in the regime of MAVs. Additionally, as noted in [22, 42, 43, 47, 48, 65], more efficient flapping motions may be realized from a flexible structure. As such, extending the high-order discretization and corresponding adjoint method to fluid-structure interaction problems or, more generally, multiphysics problems coupled along an interface, will be the subject of future work. Furthermore, for larger scale applications the cost of repeatedly solving the conservation law becomes a challenge and calls for more efficient solvers such as those developed in [36] or a globally convergent optimization framework that incorporates fast and reliable adaptive reduced-order models [59, 60].

Acknowledgements This work was supported in part by the Luis Alvarez Postdoctoral Fellowship by the Director, Office of Science, Office of Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 (MZ), and by the Director, Office of Science, Computational and Technology Research, U.S. Department of Energy under contract number DE-AC02-05CH11231 (PP). The content of this publication does not necessarily reflect the position or policy of any of these supporters, and no official endorsement should be inferred.

References

1. Roger Alexander. Diagonally implicit Runge-Kutta methods for stiff ODEs. *SIAM J. Numer. Anal.*, 14(6):1006–1021, 1977.
2. Yves Allaneau, Matthew Culbreth, and Antony Jameson. A computational framework for low Reynolds number 3d flapping wings simulations. In *20th AIAA Computational Fluid Dynamics Conference*, Honolulu, Hawaii, June 27–30 2011.
3. Douglas N Arnold, Franco Brezzi, Bernardo Cockburn, and L Donatella Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM Journal on Numerical Analysis*, 39(5):1749–1779, 2002.
4. Sai K Banala and Sunil K Agrawal. Design and optimization of a mechanism for out-of-plane insect winglike motion with twist. *Journal of Mechanical Design*, 127(4):841–844, 2005.
5. James M Birch, William B Dickson, and Michael H Dickinson. Force production and flow structure of the leading edge vortex on flapping wings at high and low Reynolds numbers. *Journal of Experimental Biology*, 207(7):1063–1072, 2004.
6. Frank M Bos, Bas W van Oudheusden, and Hester Bijl. Wing performance and 3-d vortical structure formation in flapping flight. *Journal of Fluids and Structures*, 42:130–151, 2013.
7. Chris Chabalko, Richard D Snyder, Philip S Beran, and Gregory Parker. The physics of an optimized flapping wing micro air vehicle. In *Proceedings of the 47th AIAA Aerospace Science Meeting Including The New Horizons Forum and Aerospace Exposition, AIAA Paper*, number 2009–801, Orlando, Florida, January 5–8 2009.
8. Tony F Chan and Wing Lok Wan. Analysis of projection methods for solving linear systems with multiple right-hand sides. *SIAM Journal on Scientific Computing*, 18(6):1698–1721, 1997.
9. Jung-Sun Choi, Liangyu Zhao, Gyung-Jin Park, Sunil K Agrawal, and Raymond K Kolonay. Enhancement of a flapping wing using path and dynamic topology optimization. *AIAA Journal*, 49(12):2616–2626, December 2011.
10. Bernardo Cockburn and Chi-Wang Shu. Runge-Kutta discontinuous Galerkin methods for convection-dominated problems. *J. Sci. Comput.*, 16(3):173–261, 2001.
11. Matthew Culbreth, Yves Allaneau, and Antony Jameson. High-fidelity optimization of flapping airfoils and wings. In *29th AIAA Applied Aerodynamics Conference*, Honolulu, Hawaii, June 27–30 2011.
12. Benoît Desjardins, Emmanuel Grenier, P-L Lions, and Nader Masmoudi. Incompressible limit for solutions of the isentropic Navier–Stokes equations with Dirichlet boundary conditions. *Journal de Mathématiques Pures et Appliquées*, 78(5):461–471, 1999.
13. Michael H Dickinson, Fritz-Olaf Lehmann, and Sanjay P Sane. Wing rotation and the aerodynamic basis of insect flight. *Science*, 284(5422):1954–1960, 1999.
14. Charbel Farhat, Christoph Degand, Bruno Koobus, and Michel Lesoinne. Torsional springs for two-dimensional dynamic unstructured fluid meshes. *Computer methods in applied mechanics and engineering*, 163(1):231–245, 1998.
15. Charbel Farhat, Philippe Geuzaine, and Céline Grandmont. The discrete geometric conservation law and the nonlinear stability of ale schemes for the solution of flow problems on moving grids. *Journal of Computational Physics*, 174(2):669–694, 2001.

16. Bradley Froehle and Per-Olof Persson. Nonlinear elasticity for mesh deformation with high-order discontinuous Galerkin methods for the Navier-Stokes equations on deforming domains. In *Spectral and High Order Methods for Partial Differential Equations ICOSAHOM 2014*, pages 73–85. Springer, 2015.
17. Bradley Michael Froehle. *High-order discontinuous Galerkin fluid-structure interaction methods*. University of California, Berkeley, 2013.
18. Ryan B George, Mark B Colton, Christopher A Mattson, and Scott L Thomson. A differentially driven flapping wing mechanism for force analysis and trajectory optimization. *International Journal of Micro Air Vehicles*, 4(1):31–49, 2012.
19. Mehdi Ghommem, Nathan Collier, Antti H Niemi, and Victor M Calo. On the shape optimization of flapping wings and their performance analysis. *Aerospace Science and Technology*, 32(1):274–292, 2014.
20. Mehdi Ghommem, Muhammad R Hajj, Dean T Mook, Bret K Stanford, Philip S Beran, Richard D Snyder, and Layne T Watson. Global optimization of actively morphing flapping wings. *Journal of Fluids and Structures*, 33:210–228, 2012.
21. Peter A Gnoffo et al. CFD validation studies for hypersonic flow prediction. *AIAA paper*, 1025:2001, 2001.
22. Peter G Ifju, David A Jenkins, Scott Ettinger, Yongsheng Lian, Wei Shyy, and Martin R Waszak. Flexible-wing-based micro air vehicles. In *40th AIAA Aerospace Sciences Meeting & Exhibit*, number 2002–0705, pages 1–5, Reno, Nevada, January 14–17 2002.
23. Antony Jameson, Wolfgang Schmidt, Eli Turkel, et al. Numerical solutions of the Euler equations by finite volume methods using Runge-Kutta time-stepping schemes. *AIAA paper*, 1259:1981, 1981.
24. KD Jones and MF Platzer. Numerical computation of flapping-wing propulsion and power extraction. In *35th Aerospace Sciences Meeting and Exhibit*, volume 97, page 0826, Reno, Nevada, January 6–9 1997. AIAA.
25. Martin Jones and Nail K Yamaleev. Adjoint based shape and kinematics optimization of flapping wing propulsive efficiency. 43rd AIAA Fluid Dynamics Conference. San Diego, CA, 2013. AIAA 2013–2472, 2013.
26. Bruno Koobus and Charbel Farhat. Second-order time-accurate and geometrically conservative implicit schemes for flow computations on unstructured dynamic meshes. *Computer Methods in Applied Mechanics and Engineering*, 170(1):103–129, 1999.
27. Fritz-Olaf Lehmann. The mechanisms of lift enhancement in insect flight. *Naturwissenschaften*, 91(3):101–122, 2004.
28. Michel Lesoinne and Charbel Farhat. Geometric conservation laws for flow problems with moving boundaries and deformable meshes, and their impact on aeroelastic computations. *Computer methods in applied mechanics and engineering*, 134(1):71–90, 1996.
29. Chi-Kun Lin. On the incompressible limit of the compressible Navier-Stokes equations. *Communications in partial differential equations*, 20(3–4):677–707, 1995.
30. Karthik Mani and Dimitri J Mavriplis. Unsteady discrete adjoint formulation for two-dimensional flow problems with deforming meshes. *AIAA Journal*, 46(6):1351–1364, 2015/06/22 2008.
31. V Maple. Waterloo maple software. *University of Waterloo, Version*, 5, 1994.
32. James M. McMichael and Michael S. Francis. Micro air vehicles - towards a new dimension in flight. Technical report, DARPA, August 7 1997.
33. Siva K. Nadarajah and Antony Jameson. Optimum shape design for unsteady flows with time-accurate continuous and discrete adjoint method. *AIAA Journal*, 45(7):1478–1491, 2007.
34. Eric J Nielsen, Boris Diskin, and Nail K Yamaleev. Discrete adjoint-based design optimization of unsteady turbulent flows on dynamic unstructured grids. *AIAA Journal*, 48(6):1195–1206, 2010.
35. Kui Ou, Patrice Castonguay, and Antony Jameson. 3d flapping wing simulation with high order spectral difference method on deformable mesh. In *49th AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition*, volume 1316, page 2011, Orlando, Florida, January 4–7 2011.

36. Will Pazner and Per-Olof Persson. Stage-parallel fully implicit Runge-Kutta solvers for discontinuous Galerkin fluid simulations. *Journal of Computational Physics*, 2016.
37. Jaime Peraire and Per-Olof Persson. The Compact Discontinuous Galerkin (CDG) method for elliptic problems. *SIAM Journal on Scientific Computing*, 30(4):1806–1824, 2008.
38. Per-Olof Persson, Javier Bonet, and Jaime Peraire. Discontinuous Galerkin solution of the Navier–Stokes equations on deformable domains. *Computer Methods in Applied Mechanics and Engineering*, 198(17):1585–1595, 2009.
39. Per-Olof Persson and Jaime Peraire. Curved mesh generation and mesh refinement using Lagrangian solid mechanics. In *Proceedings of the 47th AIAA Aerospace Sciences Meeting and Exhibit*, volume 204, 2009.
40. Ravi Ramamurti and William Sandberg. Simulation of flow about flapping airfoils using finite element incompressible flow solver. *AIAA Journal*, 39(2):253–260, 2001.
41. Philip L Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *Journal of Computational Physics*, 43(2):357–372, 1981.
42. Wei Shyy, Hikaru Aono, Satish Kumar Chimakurthi, P Trizila, C-K Kang, Carlos ES Cesnik, and Hao Liu. Recent progress in flapping wing aerodynamics and aeroelasticity. *Progress in Aerospace Sciences*, 46(7):284–327, 2010.
43. Wei Shyy, Mats Berg, and Daniel Ljungqvist. Flapping and flexible wings for biological and micro air vehicles. *Progress in aerospace sciences*, 35(5):455–505, 1999.
44. Wei Shyy, Y Lian, J Tang, H Liu, P Trizila, B Stanford, L Bernal, C Cesnik, P Friedmann, and P Ifju. Computational aerodynamics of low Reynolds number plunging, pitching and flexible wings for MAV applications. *Acta Mechanica Sinica*, 24(4):351–373, 2008.
45. Valeria Simoncini and Efstratios Gallopoulos. An iterative method for nonsymmetric systems with multiple right-hand sides. *SIAM Journal on Scientific Computing*, 16(4):917–933, 1995.
46. Bret K Stanford and Philip S Beran. Analytical sensitivity analysis of an unsteady vortex-lattice method for flapping-wing optimization. *Journal of Aircraft*, 47(2):647–662, 2010.
47. WB Tay, BW Van Oudheusden, and H Bijl. Numerical simulation of X-wing type biplane flapping wings in 3D using the immersed boundary method. *Bioinspiration & Biomimetics*, 9(3), March 14 2014.
48. Alexandra H Techet. Propulsive performance of biologically inspired flapping foils at high Reynolds numbers. *Journal of Experimental Biology*, 211(2):274–279, 2008.
49. Ismail H Tuncer and Mustafa Kaya. Optimization of flapping airfoils for maximum thrust and propulsive efficiency. *AIAA Journal*, 43(11):2329–2336, 2005.
50. James R Usherwood and Charles P Ellington. The aerodynamics of revolving wings ii. propeller force coefficients from mayfly to quail. *Journal of Experimental Biology*, 205(11):1565–1576, 2002.
51. Marnix P van Schrojenstein Lantman and K Fidkowski. Adjoint-based optimization of flapping kinematics in viscous flows. In *21st AIAA Computational Fluid Dynamics Conference*, 2013.
52. Andreas Wächter and Lorenz T Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical programming*, 106(1):25–57, 2006.
53. Jingyi Wang, Matthew J Zahr, and Per-Olof Persson. Energetically optimal flapping flight via a fully discrete adjoint method with explicit treatment of flapping frequency. In *23rd AIAA Computational Fluid Dynamics Conference*, page 4412, 2017.
54. ZJ Wang, Krzysztof Fidkowski, Rémi Abgrall, Francesco Bassi, Doru Caraeni, Andrew Cary, Herman Deconinck, Ralf Hartmann, Koen Hillewaert, HT Huynh, et al. High-order CFD methods: current status and perspective. *International Journal for Numerical Methods in Fluids*, 72(8):811–845, 2013.
55. David J Willis, Emily R Israeli, Per-Olof Persson, Mark Drela, Jaime Peraire, SM Swartz, and Kenneth S Breuer. A computational framework for fluid structure interaction in biologically inspired flapping flight. In *25th AIAA Applied Aerodynamics Conference*, volume 1, pages 38–59, Miami, Florida, June 25–28 2007.
56. David J Willis, Per-Olof Persson, Emily R Israeli, Jaime Peraire, Sharon M Swartz, and Kenneth S Breuer. Multifidelity approaches for the computational analysis and design of

effective flapping wing vehicles. In *46th AIAA Aerospace Sciences Meeting and Exhibit*, page 2008, Reno, Nevada, January 7–10 2008.
57. Nail Yamaleev, Boris Diskin, and Eric Nielsen. Adjoint-based methodology for time-dependent optimization. In *12th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*. American Institute of Aeronautics and Astronautics, 2008.
58. Zhi Yang and Dimitri J Mavriplis. Unstructured dynamic meshes with higher-order time integration schemes for the unsteady Navier-Stokes equations. *AIAA paper*, 1222(2005):1, 2005.
59. Matthew J. Zahr. *Adaptive model reduction to accelerate optimization problems governed by partial differential equations*. PhD thesis, Stanford University, August 2016.
60. Matthew J Zahr and Charbel Farhat. Progressive construction of a parametric reduced-order model for PDE-constrained optimization. *International Journal for Numerical Methods in Engineering*, 102(5):1111–1135, 2015.
61. Matthew J. Zahr and Per-Olof Persson. Performance tuning of Newton-GMRES methods for discontinuous Galerkin discretizations of the Navier-Stokes equations. In *21st AIAA Computational Fluid Dynamics Conference*. American Institute of Aeronautics and Astronautics, 2013.
62. Matthew J Zahr and Per-Olof Persson. High-order, time-dependent aerodynamic optimization using a discontinuous Galerkin discretization of the Navier-Stokes equations. In *AIAA Science and Technology Forum and Exposition*, San Diego, CA, 2016.
63. Matthew J. Zahr and Per-Olof Persson. An adjoint method for a high-order discretization of deforming domain conservation laws for optimization of flow problems. *Journal of Computational Physics*, 326, 516–543, 2016.
64. Matthew J Zahr, Per-Olof Persson, and Jon Wilkening. A fully discrete adjoint method for optimization of flow problems on deforming domains with time-periodicity constraints. *Computers & Fluids*, Special Issue on USNCCM13 International Symposium on Spectral and High-Order Methods, 2016.
65. Qiang Zhu. Numerical simulation of a flapping foil with chordwise or spanwise flexibility. *AIAA Journal*, 45(10):2448–2457, October 2007.

Optimization of a Fractional Differential Equation

Enrique Otárola and Abner J. Salgado

Abstract We consider a linear quadratic optimization problem where the state is governed by a fractional ordinary differential equation. We also consider control constraints. We show existence and uniqueness of an optimal state–control pair and propose a method to approximate it. Due to the low regularity of the solution to the state equation, rates of convergence cannot be proved unless problematic assumptions are made. Instead, we appeal to the theory of Γ-convergence to show the convergence of our scheme.

1 Introduction

In recent years, a lot of attention has been paid to the study of *nonlocal* problems, of which fractional differential equations represent an instance. This is motivated by the fact that fractional derivatives are better suited to capturing long-range interactions, as well as memory effects. For instance, they have been used to describe anomalous transport phenomena [9, 10], option pricing [6], porous media flow [5], and viscoelastic materials [8], to name a few. It is only natural then, from the purely mathematical as well as the practical points of view, to try to optimize systems that are governed by these equations. In previous work [4], we dealt with a constrained optimization problem where the state is governed by a differential equation that presented nonlocal features in time as well as in space. Throughout the analysis presented in [4], the nonlocalities in time and space were intertwined and this required to develop several tools to analyze the nonlocal operator in space that are, in principle, not relevant to the nonlocality in time. It is thus our feeling that

E. Otárola
Departamento de Matemática, Universidad Técnica Federico Santa María, Valparaíso, Chile
e-mail: enrique.otarola@usm.cl

A. J. Salgado (✉)
Department of Mathematics, University of Tennessee, Knoxville, TN 37996, USA
e-mail: asalgad1@utk.edu

H. Antil et al. (eds.), *Frontiers in PDE-Constrained Optimization*, The IMA Volumes in Mathematics and its Applications 163,
https://doi.org/10.1007/978-1-4939-8636-1_8

the extensive technicalities that ensued in the analysis of [4] obscured many of the unique features that optimization of fractional differential equations contains; for instance, the lack of time regularity regardless of the smoothness of data. For this reason, our main objective in this note is to present a detailed study for the case where the state is governed by a time-fractional ordinary differential equation.

Let us be precise in our considerations. Given $m, n \geq 1$, a final time $T > 0$, a desired state $u_d \in L^2(0, T; \mathbb{R}^m)$, and a regularization parameter $\mu > 0$, we define the *cost functional* as

$$J(u, z) = \frac{1}{2} \int_0^T \left(|\mathscr{C} u - u_d|_m^2 + \mu |z|_n^2 \right) \mathrm{d}t, \tag{1}$$

where we denote the Euclidean norm in $\mathbb{R}^s$ by $|\cdot|_s$ and $\mathscr{C} \in \mathbb{M}^{m \times n}$; $\mathbb{M}^{m \times n}$ denotes the set of all m–by–n matrices. The variable u is called the *state*, while the variable z is the *control*. The control and state are related by the so-called *state equation*, which we now describe. Given an initial condition $\psi \in \mathbb{R}^n$, a forcing function $f : (0, T] \to \mathbb{R}^n$, a symmetric positive definite matrix $\mathscr{A} \in \mathbb{M}^{n \times n}$, the state equation reads

$$\mathrm{d}_t^\gamma u + \mathscr{A} u = f + z, \ t \in (0, T], \qquad u(0) = \psi. \tag{2}$$

Here, $\gamma \in (0, 1)$ and d_t^γ denotes the so-called *left-sided Caputo fractional derivative* of order γ, which is defined by [19, 28]

$$\mathrm{d}_t^\gamma v(t) = \frac{1}{\Gamma(1-\gamma)} \int_0^t \frac{1}{(t-\zeta)^\gamma} \dot{v}(\zeta) \, \mathrm{d}\zeta, \tag{3}$$

where by $\dot{v}$ we denote the usual derivative and Γ is the Gamma function. We must immediately remark that, in addition to (3), there are other, not equivalent, definitions of fractional derivatives: Riemann–Liouville, Grünwald-Letnikov, and Marchaud derivatives. In this work, we shall focus on the Caputo derivatives since they allow for a standard initial condition in (2); a highly desirable feature in applications; see, for instance, the discussion in [14, Section E.4]. For further motivation and applications, we refer the reader to [11, 14].

The problem we shall be concerned with is to find $(\breve{u}, \breve{z})$ such that

$$J(\breve{u}, \breve{z}) = \min J(u, z) \tag{4}$$

subject to the state equation (2) and the *control constraints*

$$a \preceq z \preceq b. \tag{5}$$

Here $a, b \in \mathbb{R}^n$ which we assume satisfy that $a \preceq b$. The relation $v \preceq w$ means that, for all $i = 1, \ldots, n$, we have $v_i \leq w_i$.

To our knowledge, the first work that was devoted to the study of (4) is [2] where a formal Lagrangian formulation is discussed and optimality conditions are formally derived. The author of this work also presents a numerical scheme based on shifted Legendre polynomials. However, there is no analysis of the optimality conditions or numerical scheme. Other discretization schemes using finite elements [3], rational approximations [30], spectral methods [24, 32, 33], or other techniques have been considered. Most of these works do not provide a rigorous justification or analysis of their schemes, and the ones that do obtain error estimates under rather strong regularity assumptions of the state variable; namely, they require that $\ddot{u} \in L^\infty(0,T;\mathbb{R}^n)$ which is rather problematic; see Theorem 2 below. In contrast, in this work, we carefully describe the regularity properties of the state equation and on their basis provide convergence (without rates) of the numerical scheme we propose.

Throughout our discussion, we will follow the standard notation and terminology. Nonstandard notation will be introduced in the course of our exposition. The rest of this work is organized as follows: Basic facts about fractional derivatives and integrals are presented in Section 1.1. We study the state equation in Section 2 where we construct the solution to problem (2), study its regularity, and present a somewhat new point of view for a classical scheme—the so-called L1 scheme. More importantly, we use the right regularity to obtain rates of convergence; an issue that has been largely ignored in the literature. With these ingredients at hand we proceed, in Section 3, to analyze the optimization problem (4); we show existence and uniqueness of an optimal state–control pair and propose a scheme to approximate it. We employ a piecewise linear (in time) approximation of the state and a piecewise constant approximation of the control. While not completely necessary for the analysis, we identify the discrete adjoint problem and use it to derive discrete optimality conditions. Finally, we show the strong convergence of the discrete optimal control to the continuous one. Owing to the reduced regularity of the solution to the state equation, this convergence, however, cannot have rates.

1.1 Fractional Derivatives and Integrals

We begin by recalling some fundamental facts about fractional derivatives and integrals. The left-sided Caputo fractional derivative is defined in (3). The right-sided Caputo fractional derivative of order γ is given by [19, 28]

$$\mathrm{d}^{\gamma}_{T-t} v(t) = -\frac{1}{\Gamma(1-\gamma)} \int_t^T \frac{1}{(\zeta - t)^{\gamma}} \dot{v}(\zeta)\, \mathrm{d}\zeta. \tag{6}$$

For $v \in L^1(0,T)$, the left Riemann–Liouville fractional integral of order $\sigma \in (0,1)$ is defined by

$$I_t^\sigma[v](t) = \frac{1}{\Gamma(\sigma)} \int_0^t \frac{1}{(t-\zeta)^{1-\sigma}} v(\zeta)\, \mathrm{d}\zeta; \tag{7}$$

see [28, Section 2]. Young's inequality for convolutions immediately yields that, for $p > 1$, I_t^σ is a continuous operator from $L^p(0,T)$ into itself. More importantly, a result by Flett [12] shows that

$$v \in L\log L(0,T) \implies I_t^\sigma[v] \in L^{\frac{1}{1-\sigma}}(0,T). \tag{8}$$

We refer the reader to [20] for the definition of the Orlicz space $L\log L(0,T)$. This observation will be very important in subsequent developments. Notice finally that if $v \in W_1^1(0,T)$, then we have that $\mathrm{d}_t^\gamma v(t) = I_t^{1-\gamma}[\dot{v}](t)$.

The generalized Mittag-Leffler function with parameters $\alpha > 0$ and $\beta \in \mathbb{R}$ is defined by

$$E_{\alpha,\beta}(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + \beta)}, \quad z \in \mathbb{C}. \tag{9}$$

We refer the reader to [14] for an account of the principal properties of the Mittag-Leffler function.

2 The State Equation

In this section, we construct the solution to (2), thus showing its existence and uniqueness. This shall be of uttermost importance not only when showing the existence and uniqueness of solutions to our optimization problem, but when we deal with the discretization, as we will study the smoothness of u. To shorten notation, in this section we set

$$g = f + z,$$

where f is the forcing term and z is the control in (2).

2.1 Solution Representation and Regularity

Let us now construct the solution to (2) and review its main properties. We will adapt the arguments of [26] to our setting. Since the matrix $\mathscr{A}$ is symmetric and positive definite, it is orthogonally diagonalizable; meaning that there are $\{\lambda_\ell, \xi_\ell\}_{\ell=1}^n \subset \mathbb{R}_+ \times \mathbb{R}^n$ such that

$$\mathscr{A}\xi_\ell = \lambda_\ell \xi_\ell, \qquad \xi_{\ell_1} \cdot \xi_{\ell_2} = \delta_{\ell_1,\ell_2}.$$

This, in particular, implies that the vectors $\{\xi_\ell\}_{\ell=1}^n$ form an orthonormal basis of $\mathbb{R}^n$. Moreover, for any vector $v \in \mathbb{R}^n$, we can define $|v|_{\mathscr{A}}^2 = v \cdot \mathscr{A} v$, which turns out to be a norm that satisfies

$$\lambda_1 |v|_n^2 \le |v|_{\mathscr{A}}^2 \le \lambda_n |v|_n^2, \quad \forall v \in \mathbb{R}^n. \tag{10}$$

We set

$$\|v\|_{L^2_{\mathscr{A}}(0,T;\mathbb{R}^n)}^2 = \int_0^T |v|_{\mathscr{A}}^2 \, \mathrm{d}t. \tag{11}$$

With these properties of the matrix $\mathscr{A}$ at hand, we propose the following solution ansatz:

$$u(t) = \sum_{\ell=1}^n u_\ell(t)\xi_\ell, \quad u_\ell(t) = u(t) \cdot \xi_\ell, \tag{12}$$

where the coefficients $u_\ell(t)$ satisfy

$$\mathrm{d}_t^\gamma u_\ell(t) + \lambda_\ell u_\ell(t) = g_\ell(t), \ t \in (0, T], \qquad u_\ell(0) = \psi_\ell, \tag{13}$$

for $\ell \in \{1, \cdots, n\}$. Here, $g_\ell(t) = g(t) \cdot \xi_\ell$ and $\psi_\ell = \psi \cdot \xi_\ell$. The importance of this orthogonal decomposition lies in the fact that we have reduced problem (2) to a decoupled system of equations. The theory of fractional ordinary differential equations [28] gives, for $\ell \in \{1, \cdots, n\}$, a unique function u_ℓ satisfying problem (13). In addition, standard considerations, which formally entail taking the Laplace transform of (13), yield that

$$u_\ell(t) = E_{\gamma,1}(-\lambda_\ell t^\gamma)\psi_\ell + \int_0^t (t-\zeta)^{\gamma-1} E_{\gamma,\gamma}(-\lambda_\ell (t-\zeta)^\gamma) g_\ell(\zeta) \, \mathrm{d}\zeta. \tag{14}$$

We refer the reader to [25–27] for details. This representation shall prove rather useful to describe the existence, uniqueness, and regularity of u. To concisely state it, let us define

$$\mathbb{U} = \{w \in L^2(0, T; \mathbb{R}^n) : \mathrm{d}_t^\gamma w \in L^2(0, T; \mathbb{R}^n)\}. \tag{15}$$

With this notation, a specialization of the results of [26] to the substantially simpler case when $\mathscr{A}$ is a positive definite matrix (and thus the spaces are finite dimensional) yields the following result.

Theorem 1 (Existence and Uniqueness) *Assume that $g \in L^2(0, T; \mathbb{R}^n)$. Problem (2) has a unique solution $u \in \mathbb{U}$, given by (12) and (14). Moreover, the following a*

priori estimate holds

$$I_t^{1-\gamma}\left[|u|_n^2\right](T) + \|u\|^2_{L^2_{\mathscr{A}}(0,T;\mathbb{R}^n)} \lesssim \Lambda_\gamma^2(\psi, g), \tag{16}$$

where, for $v \in \mathbb{R}^n$ *and* $h \in L^2(0,T;\mathbb{R}^n)$ *we have*

$$\Lambda_\gamma^2(v,h) = I_t^{1-\gamma}\left[|v|_n^2\right](T) + \|h\|^2_{L^2(0,T;\mathbb{R}^n)}, \tag{17}$$

where we implicitly identified v *with the constant function* $[0,T] \ni t \mapsto v \in \mathbb{R}^n$. *In this estimate, the hidden constant is independent of* ψ, g, *and* u.

Having obtained conditions that guarantee the existence and uniqueness for (2) we now study its regularity. This is important since, as it is well known, smoothness and rate of approximation go hand in hand. This is exactly the content of direct and converse theorems in approximation theory [1, 17]. Consequently, any rigorous study of an approximation scheme must be concerned with the regularity of the solution. This, we believe, is an issue that for this problem has been largely ignored in the literature since, essentially, the solution to (2) is not smooth. Let us now follow [25, 26] and elaborate on this matter. The essence of the issue is already present in the case $n = 1$ so that (14) *is* the solution. Let us, to further simplify the discussion, set $\mathscr{A} = 1$, $g \equiv 0$, and $\psi = 1$. In this case, the solution verifies the following asymptotic estimate:

$$u(t) = E_{\gamma,1}(-t^\gamma) = 1 - \frac{1}{\Gamma(1+\gamma)}t^\gamma + \mathscr{O}(t^{2\gamma}), \qquad t \downarrow 0.$$

If this is the case we then expect that, as $t \downarrow 0$, $\dot{u}(t) \approx t^{\gamma-1}$ and $\ddot{u}(t) \approx t^{\gamma-2}$. Notice that, since $\gamma \in (0,1)$, the function $\omega_1(t) = t^{\gamma-1}$ belongs to $L\log L(0,T)$ but $\omega_1 \notin L^{1+\epsilon}(0,T)$ for any $\epsilon > \gamma(1+\gamma)^{-1}$. Similarly, the function $\omega_2(t) = t^{\gamma-2}$ is not Lebesgue integrable, but

$$\int_0^T t^\sigma |\omega_2(t)|^2 \,\mathrm{d}t = \int_0^T t^{\sigma+2(\gamma-2)} \,\mathrm{d}t < \infty \Rightarrow \sigma > 3 - 2\gamma,$$

which implies that ω_2 belongs to the weighted Lebesgue space $L^2(t^\sigma;0,T)$, where $\sigma > 3 - 2\gamma > 1$. The considerations given above tell us that we should expect the following:

$$\dot{u} \in L\log L(0,T;\mathbb{R}^n) \qquad \ddot{u} \in L^2(t^\sigma;0,T;\mathbb{R}^n),\ \sigma > 3-2\gamma. \tag{18}$$

The justification of this heuristic is the content of the next result. For a proof, we refer the reader to [26, Theorem 8].

Theorem 2 (Regularity) *Assume that $g \in H^2(0,T;\mathbb{R}^n)$. Then u, the solution to (2), satisfies (18) and, for $t \in (0,T]$, we have the following asymptotic estimate:*

$$\left(\int_0^T \zeta^\sigma |\ddot{u}(\zeta)|_n^2 \,\mathrm{d}\zeta\right)^{1/2} + t^{1-\gamma}\left|\dot{u}(t) - \frac{1}{t}(u(t)-\psi)\right|_n \lesssim |\psi|_n + \|g\|_{H^2(0,T;\mathbb{R}^n)},$$

where $\sigma > 3-2\gamma$. The hidden constant is independent of t but blows up as $\gamma \downarrow 0^+$.

Remark 1 (Extensions) Under the correct framework, the conclusion of Theorem 2 can be extended to the case where $\mathscr{A}$ is an operator acting on a Hilbert space $\mathscr{H}$ and Equation (2) is understood in a Gelfand triple $\mathscr{V} \hookrightarrow \mathscr{H} \hookrightarrow \mathscr{V}'$; see [26] for details.

2.2 Discretization of the State Equation

Now that we have studied the state equation and the regularity properties of its solution u, we proceed to discretize it. To do so, we denote by $\mathscr{K} \in \mathbb{N}$ the number of time steps. We define the (uniform) time step $\tau = T/\mathscr{K} > 0$ and set $t_k = k\tau$ for $k = 0,\dots,\mathscr{K}$. We denote the time partition by $\mathscr{T} = \{t_k\}_{k=0}^{\mathscr{K}}$. We define the space of continuous and piecewise linear, over the partition $\mathscr{T}$, functions as follows:

$$\mathbb{U}(\mathscr{T}) = \left\{W \in C([0,T];\mathbb{R}^n) : W|_{(t_k,t_{k+1}]} \in \mathbb{P}_1(\mathbb{R}^n), k = 0,\dots,\mathscr{K}-1\right\}. \tag{19}$$

We also define the space of piecewise constant functions

$$\mathbb{Z}(\mathscr{T}) = \left\{W \in BV(0,T;\mathbb{R}^n) : W|_{(t_k,t_{k+1}]} \in \mathbb{P}_0(\mathbb{R}^n), k = 0,\dots,\mathscr{K}-1\right\}, \tag{20}$$

and the $L^2(0,T;\mathbb{R}^n)$-orthogonal projection onto $\mathbb{Z}(\mathscr{T})$, that is, the operator $\Pi_{\mathscr{T}} : L^2(0,T;\mathbb{R}^n) \to \mathbb{Z}(\mathscr{T})$ defined by

$$\int_0^T (r - \Pi_{\mathscr{T}} r)\cdot Z^{\boldsymbol{\tau}} \,\mathrm{d}t = 0 \quad \forall Z^{\boldsymbol{\tau}} \in \mathbb{Z}(\mathscr{T}).$$

We remark that $\Pi_{\mathscr{T}}$ satisfies

$$\|r - \Pi_{\mathscr{T}} r\|_{L^2(0,T;\mathbb{R}^n)} \lesssim \tau \|\dot{r}\|_{L^2(0,T;\mathbb{R}^n)}, \tag{21}$$

where the hidden constant is independent of r and τ.

For a function $\phi \in BV(0,T;\mathbb{R}^n)$ we set $\phi^k = \lim_{\epsilon\uparrow 0}\phi(t_k-\epsilon)$ and $\phi^{\boldsymbol{\tau}} = \{\phi^k\}_{k=0}^{\mathscr{K}}$, which can be uniquely identified with either an element of $\mathbb{U}(\mathscr{T})$ or $\mathbb{Z}(\mathscr{T})$ by the procedures we describe now. To $\phi^{\boldsymbol{\tau}}$ we associate $\bar{\phi}^{\boldsymbol{\tau}} \in \mathbb{Z}(\mathscr{T})$ defined by

$$\bar{\phi}^{\boldsymbol{\tau}}(0) = \phi^0, \quad \bar{\phi}^{\boldsymbol{\tau}}|_{(t_k,t_{k+1}]}(t) = \phi^{k+1},\ k = 0,\dots,\mathscr{K}-1. \tag{22}$$

We also associate $\hat{\phi}^{\boldsymbol{\tau}} \in \mathbb{U}(\mathscr{T})$ via

$$\hat{\phi}^{\boldsymbol{\tau}}(0) = \phi^0, \quad \hat{\phi}^{\boldsymbol{\tau}}|_{(t_k,t_{k+1}]}(t) = \frac{t_{k+1}-t}{\tau}\phi^k + \frac{t-t_k}{\tau}\phi^{k+1}, \ k = 0, \ldots, \mathscr{K}-1. \tag{23}$$

Notice that

$$\|\hat{\phi}^{\boldsymbol{\tau}}\|_{L^\infty(0,T;\mathbb{R}^n)} = \|\bar{\phi}^{\boldsymbol{\tau}}\|_{L^\infty(0,T;\mathbb{R}^n)} = \|\phi^{\boldsymbol{\tau}}\|_{\ell^\infty(\mathbb{R}^n)}$$

and that

$$\|\bar{\phi}^{\boldsymbol{\tau}}\|^2_{L^2(0,T;\mathbb{R}^n)} = \tau \sum_{k=1}^{\mathscr{K}} |\phi^k|^2_n.$$

Finally, for a sequence $\phi^{\boldsymbol{\tau}}$ we also define, for $k = 0, \ldots, \mathscr{K}-1$,

$$\mathfrak{d}\phi^{k+1} = \tau \dot{\hat{\phi}}^{\boldsymbol{\tau}}|_{(t_k,t_{k+1}]} = \phi^{k+1} - \phi^k, \tag{24}$$

which can be understood as a mapping $\mathfrak{d} : \mathbb{U}(\mathscr{T}) \to \mathbb{Z}(\mathscr{T})$.

Having introduced this notation, we propose to discretize (2) by a collocation method over $\mathbb{U}(\mathscr{T})$. In other words, we seek for $\hat{U}^{\boldsymbol{\tau}} \in \mathbb{U}(\mathscr{T})$ such that

$$\hat{U}^{\boldsymbol{\tau}}(0) = \psi, \tag{25}$$

and, for every $k = 0, \ldots \mathscr{K}-1$, it satisfies

$$\mathrm{d}_t^\gamma \hat{U}^{\boldsymbol{\tau}}(t_{k+1}) + \mathscr{A}\hat{U}^{\boldsymbol{\tau}}(t_{k+1}) = \Pi_{\mathscr{T}} g(t_{k+1}). \tag{26}$$

Remark 2 (Derivation of the Scheme) In the literature, (26) is commonly referred to as the L1-scheme [16, 21, 22, 29], even though it is not presented this way. Nevertheless, let us show that this is equivalent to the methods presented in the literature. To see the relation it is sufficient to compute, for a function $\hat{W}^{\boldsymbol{\tau}} \in \mathbb{U}(\mathscr{T})$, the value of $\mathrm{d}_t^\gamma \hat{W}^{\boldsymbol{\tau}}(t_{k+1})$. By definitions (3), (23), and (24), we obtain that

$$\begin{aligned}\mathrm{d}_t^\gamma \hat{W}^{\boldsymbol{\tau}}(t_{k+1}) &= \frac{1}{\Gamma(1-\gamma)} \int_0^{t_{k+1}} \frac{1}{(t_{k+1}-\zeta)^\gamma} \dot{\hat{W}}^{\boldsymbol{\tau}}(\zeta)\, \mathrm{d}\zeta \\ &= \frac{\tau^{-1}}{\Gamma(1-\gamma)} \sum_{j=0}^{k} \mathfrak{d}W^{j+1} \int_{t_j}^{t_{j+1}} \frac{1}{(t_{k+1}-\zeta)^\gamma}\, \mathrm{d}\zeta = \sum_{j=0}^{k} a_j^k \mathfrak{d}W^{j+1},\end{aligned} \tag{27}$$

where the coefficients a_j^k satisfy

$$
\begin{aligned}
a_j^k &= \frac{\tau^{-1}}{\Gamma(1-\gamma)} \int_{t_j}^{t_{j+1}} \frac{1}{(t_{k+1}-\zeta)^{\gamma}} \, \mathrm{d}\zeta \\
&= \frac{\tau^{-1}}{\Gamma(2-\gamma)} \left[(t_{k+1}-t_j)^{1-\gamma} - (t_{k+1}-t_{j+1})^{1-\gamma} \right] \\
&= \frac{\tau^{-\gamma}}{\Gamma(2-\gamma)} \left[(k+1-j)^{1-\gamma} - (k-j)^{1-\gamma} \right].
\end{aligned} \tag{28}
$$

Here, in the last step, we used that the time step is uniform and of size τ. The fact that the time step is uniform also implies that

$$
a_{k-j}^k = \frac{\tau^{-\gamma}}{\Gamma(2-\gamma)} \left[(j+1)^{1-\gamma} - j^{1-\gamma} \right] = a_k^{k+j},
$$

so that, after the change of indices $m = k - j$, we obtain

$$
\begin{aligned}
\mathrm{d}_t^{\gamma} \hat{W}^{\tau}(t_{k+1}) &= \frac{\tau^{-\gamma}}{\Gamma(2-\gamma)} \sum_{m=0}^{k} b_m \eth W^{k+1-m} \\
&= \frac{\tau^{-\gamma}}{\Gamma(2-\gamma)} \left(b_0 W^{k+1} + \sum_{m=1}^{k} (b_m - b_{m-1}) W^{k+1-m} - b_k W^0 \right),
\end{aligned} \tag{29}
$$

with $b_m = (m+1)^{1-\gamma} - m^{1-\gamma}$. The expression above is what is commonly referred to as the L1 scheme.

2.2.1 Stability

Let us discuss the stability of scheme (26) as originally detailed in [26, Section 3.2.2]. We begin by exploring the properties of the coefficients a_j^k.

Lemma 1 (Properties of a_j^k) *Assume that the time step is given by $\tau > 0$. For every $k = 0, \ldots, \mathscr{K} - 1$ and $j = 0, \ldots, k$, the coefficients a_j^k, defined in (28), satisfy*

$$
0 < a_j^k, \qquad a_j^k < a_{j+1}^k, \qquad a_j^{k+1} < a_j^k.
$$

Moreover $a_k^k = \tau^{-\gamma}/\Gamma(2-\gamma)$.

Proof The positivity of the coefficients follows from the fact that, for $j = 0, \ldots, k$ and $\zeta \in (t_j, t_{j+1})$, we have that $t_{k+1} - \zeta > 0$. We now show that the coefficients are increasing in the lower index. In fact, an application of the mean value theorem yields

$$a_j^k = \frac{1}{\Gamma(1-\gamma)} \fint_{t_j}^{t_{j+1}} \frac{\mathrm{d}\zeta}{(t_{k+1}-\zeta)^\gamma} = \frac{1}{\Gamma(1-\gamma)} \frac{1}{(t_{k+1}-\zeta_j)^\gamma}$$

for some $\zeta_j \in (t_j, t_{j+1})$. Since the function $\zeta \mapsto (t_{k+1}-\zeta)^{-\gamma}$ is increasing for $\zeta < t_{k+1}$, we conclude that $a_j^k < a_{j+1}^k$. To show that the coefficients are decreasing in the upper index, we note that

$$t_{k+1} > t_k \implies \frac{1}{(t_{k+1}-\zeta)^\gamma} < \frac{1}{(t_k-\zeta)^\gamma},$$

so that $a_j^{k+1} < a_j^k$. Finally, we note that

$$a_k^k = \frac{1}{\Gamma(1-\gamma)} \fint_{t_k}^{t_{k+1}} \frac{\mathrm{d}\zeta}{(t_{k+1}-\zeta)^\gamma} = \frac{\tau^{-\gamma}}{\Gamma(2-\gamma)}.$$

This concludes the proof. □

With the results of Lemma 1 at hand, we can now show stability of the scheme.

Theorem 3 (Stability) *For every $\mathscr{K} \in \mathbb{N}$, the scheme (26) is unconditionally stable and satisfies*

$$I_t^{1-\gamma}\left[|\bar{U}^{\boldsymbol{\tau}}|_n^2\right](T) + \|\bar{U}^{\boldsymbol{\tau}}\|^2_{L^2_{\mathscr{A}}(0,T;\mathbb{R}^n)} \lesssim \Lambda_\gamma^2(\psi, g),$$

where the hidden constant is independent of ψ, g, $\bar{U}^{\boldsymbol{\tau}}$ and $\mathscr{K}$; and Λ_γ is defined in (17).

Proof Multiply (26), by $2U^{k+1}$ to obtain

$$2\,\mathrm{d}_t^\gamma \hat{U}^{\boldsymbol{\tau}}(t_{k+1}) \cdot U^{k+1} + 2|U^{k+1}|_{\mathscr{A}}^2 \le 2|\Pi_{\mathscr{T}} g^{k+1}|_n |U^{k+1}|_n, \tag{30}$$

where on the right-hand side we applied the Cauchy–Schwartz inequality; $|\cdot|_{\mathscr{A}}$ is defined in Section 2.1. We thus use (10), together with Young's inequality, to say that

$$2\,\mathrm{d}_t^\gamma \hat{U}^{\boldsymbol{\tau}}(t_{k+1}) \cdot U^{k+1} + |U^{k+1}|_{\mathscr{A}}^2 \le \lambda_1^{-1} |\Pi_{\mathscr{T}} g^{k+1}|_n^2.$$

We now invoke (27) and deduce that

$$\begin{aligned}
\mathrm{d}_t^\gamma \hat{U}^{\boldsymbol{\tau}}(t_{k+1}) \cdot U^{k+1} &= a_k^k |U^{k+1}|_n^2 + \sum_{j=0}^{k-1} a_j^k U^{j+1} \cdot U^{k+1} - \sum_{j=1}^{k} a_j^k U^j \cdot U^{k+1} \\
&\quad - a_0^k U^0 \cdot U^{k+1} \\
&= a_k^k |U^{k+1}|_n^2 + \sum_{j=1}^{k} (a_{j-1}^k - a_j^k) U^j \cdot U^{k+1} - a_0^k U^0 \cdot U^{k+1}.
\end{aligned}$$

With this at hand (30) reduces to

$$2a_k^k|U^{k+1}|_n^2 + |U^{k+1}|_{\mathscr{A}}^2 \leq \lambda_1^{-1}|\Pi_{\mathscr{T}} g^{k+1}|_n^2 + 2\sum_{j=1}^{k}(a_j^k - a_{j-1}^k)U^j \cdot U^{k+1} + 2a_0^k U^0 \cdot U^{k+1}.$$

Since, as stated in Lemma 1, we have that $a_j^k - a_{j-1}^k > 0$ we estimate

$$\begin{aligned} 2\sum_{j=1}^{k}(a_j^k - a_{j-1}^k)U^j \cdot U^{k+1} &\leq \sum_{j=1}^{k}(a_j^k - a_{j-1}^k)(|U^j|_n^2 + |U^{k+1}|_n^2) \\ &= \sum_{j=1}^{k}(a_j^k - a_{j-1}^k)|U^j|_n^2 + (a_k^k - a_0^k)|U^{k+1}|_n^2, \end{aligned}$$

which can be used to obtain that

$$a_k^k|U^{k+1}|_n^2 + \sum_{j=1}^{k} a_{j-1}^k|U^j|_n^2 + |U^{k+1}|_{\mathscr{A}}^2 \leq \lambda_1^{-1}|\Pi_{\mathscr{T}} g^{k+1}|_n^2 + a_0^k|\psi|_n^2 + \sum_{j=1}^{k} a_j^k|U^j|_n^2. \tag{31}$$

Notice now that, since a_j^k are defined as in (28) and $b_m = (m+1)^{1-\gamma} - m^{1-\gamma}$, for every $j = 0, \ldots, k$ we have

$$a_j^k = \frac{\tau^{-\gamma}}{\Gamma(2-\gamma)} b_{k-j}.$$

Thus, the change of indices $m = k + 1 - j$ on the left-hand side and $l = k - j$ on the right-hand side of (31), respectively, yields

$$\begin{aligned} \frac{\tau^{-\gamma}}{\Gamma(2-\gamma)}\sum_{m=0}^{k} b_m|U^{k+1-m}|_n^2 + |U^{k+1}|_{\mathscr{A}}^2 &\leq \lambda_1^{-1}|\Pi_{\mathscr{T}} g^{k+1}|_n^2 + a_0^k|\psi|_n^2 \\ &\quad + \frac{\tau^{-\gamma}}{\Gamma(2-\gamma)}\sum_{l=0}^{k-1} b_l|U^{k-l}|_n^2, \end{aligned}$$

where the sum on the right-hand side vanishes for $k = 0$. Multiply by τ and add over k to obtain

$$\begin{aligned} \frac{\tau^{1-\gamma}}{\Gamma(2-\gamma)}\sum_{k=0}^{\mathscr{K}-1} b_k|U^{\mathscr{K}-k}|_n^2 + \|\bar{U}^{\boldsymbol{\tau}}\|_{L^2_{\mathscr{A}}(0,T;\mathbb{R}^n)}^2 &\leq \lambda_1^{-1}\|\Pi_{\mathscr{T}} g\|_{L^2(0,T;\mathbb{R}^n)}^2 \\ &\quad + \tau|\psi|_n^2 \sum_{k=0}^{\mathscr{K}-1} a_0^k, \end{aligned} \tag{32}$$

where $\|\bar{U}^{\boldsymbol{\tau}}\|_{L^2_{\mathscr{A}}(0,T;\mathbb{R}^n)}$ is defined by (11). Notice now that, since the time step is uniform,

$$\tau \sum_{k=0}^{\mathscr{K}-1} a_0^k = \frac{\tau^{1-\gamma}}{\Gamma(2-\gamma)} \sum_{k=0}^{\mathscr{K}-1} b_k = \frac{T^{1-\gamma}}{\Gamma(2-\gamma)} = I_t^{1-\gamma}[1](T). \tag{33}$$

We now analyze the first term on the left-hand side of (32): Changing indices via $l+1 = \mathscr{K} - k$ gives

$$\begin{aligned} \frac{\tau^{1-\gamma}}{\Gamma(2-\gamma)} \sum_{k=0}^{\mathscr{K}-1} b_k |U^{\mathscr{K}-k}|_n^2 &= \frac{\tau^{1-\gamma}}{\Gamma(2-\gamma)} \sum_{l=0}^{\mathscr{K}-1} b_{\mathscr{K}-l-1} |U^{l+1}|_n^2 \\ &= \sum_{l=0}^{\mathscr{K}-1} \tau a_l^{\mathscr{K}-1} |U^{l+1}|_n^2 \\ &= \frac{1}{\Gamma(1-\gamma)} \sum_{l=0}^{\mathscr{K}-1} \int_{t_l}^{t_{l+1}} \frac{1}{(t_{\mathscr{K}} - \zeta)^{\gamma}} |\bar{U}^{\boldsymbol{\tau}}(\zeta)|_n^2 \, \mathrm{d}\zeta \\ &= I_t^{1-\gamma} \left[|\bar{U}^{\boldsymbol{\tau}}|_n^2 \right](T). \end{aligned} \tag{34}$$

Inserting (33) and (34) in (32), and using $\Pi_{\mathscr{T}}$ that is a projection, yields the result. □

2.2.2 Consistency and Error Estimates

Let us now discuss the consistency of scheme (26). This will allow us to obtain error estimates. Clearly, it suffices to control the difference $\mathrm{d}_t^{\gamma}(u - \hat{u}^{\boldsymbol{\tau}})$. The following formal estimate has been shown in many references; see, for instance, [21, 22]. The proof, essentially, is a Taylor expansion argument.

Proposition 1 (Consistency for Smooth Functions) *Let $w \in C^2([0,T];\mathbb{R}^n)$, then*

$$\|\, \mathrm{d}_t^{\gamma}(w - \hat{w}^{\boldsymbol{\tau}})\|_{L^{\infty}(0,T;\mathbb{R}^n)} \lesssim \tau^{2-\gamma},$$

where the hidden constant depends on $\|w\|_{C^2([0,T];\mathbb{R}^n)}$ but is independent of τ.

We must immediately point out that this estimate *cannot* be used in the analysis of (2). The reason behind this lies in Theorem 2 which shows that, in general, the solution to the state equation is not twice continuously differentiable. For this reason, in [26] a new consistency estimate, which takes into account the correct regularity of the solution, has been developed. This is the content of the next result.

Theorem 4 (Consistency) *Let $\gamma \in (0,1)$ and u solve (2). In the setting of Theorem 2 we have that, for any $\theta < \frac{1}{2}$,*

$$\| \mathrm{d}_t^\gamma (u - \hat{u}^{\boldsymbol{\tau}}) \|_{L^2(0,T;\mathbb{R}^n)} \lesssim \tau^\theta \left(|\psi|_n + \|g\|_{H^2(0,T;\mathbb{R}^n)} \right),$$

where the hidden constant is independent of τ but blows up as $\theta \uparrow \frac{1}{2}$. Here θ is independent of γ.

For a proof of this result, we refer the reader to [26, Section 3.2.1]. We just comment that it consists of a combination of the fine regularity results of Theorem 2, weighted estimates, and the mapping properties of the fractional integral operator $I_t^{1-\gamma}$ detailed in Section 1.1. Let us, however, show how from this we obtain an error estimate.

Corollary 1 (Error Estimates) *Let u solve (2) and $U^{\boldsymbol{\tau}}$ solve (26). In the setting of Theorem 2 we have that, for any $\theta < \frac{1}{2}$,*

$$I_t^{1-\gamma} \left[|\bar{u}^{\boldsymbol{\tau}} - \bar{U}^{\boldsymbol{\tau}}|_n^2 \right](T) + \|\bar{u}^{\boldsymbol{\tau}} - \bar{U}^{\boldsymbol{\tau}}\|^2_{L^2_{\mathscr{A}}(0,T;\mathbb{R}^n)} \lesssim \tau^{2\theta} \left(|\psi|_n + \|g\|_{H^2(0,T;\mathbb{R}^n)} \right)^2,$$

where the hidden constant is independent of τ and the data but blows up as $\theta \uparrow \frac{1}{2}$.

Proof Define $e^{\boldsymbol{\tau}} = u^{\boldsymbol{\tau}} - U^{\boldsymbol{\tau}}$. Subtracting (2) and (25)–(26) at $t = t_{k+1}$ yields $\hat{e}^{\boldsymbol{\tau}}(0) = 0$ and, for $k = 0, \ldots, \mathscr{K} - 1$

$$\mathrm{d}_t^\gamma \hat{e}^{\boldsymbol{\tau}}(t_{k+1}) + \mathscr{A} \hat{e}^{\boldsymbol{\tau}}(t_{k+1}) = \mathrm{d}_t^\gamma (\hat{u}^{\boldsymbol{\tau}} - u)(t_{k+1}) + (g - \Pi_{\mathscr{T}} g)(t_{k+1}).$$

Since $\bar{e}^{\boldsymbol{\tau}}(0) = 0$, the stability estimate of Theorem 3 then yields

$$I_t^{1-\gamma} \left[|\bar{e}^{\boldsymbol{\tau}}|_n^2 \right](T) + \|\bar{e}^{\boldsymbol{\tau}}\|^2_{L^2_{\mathscr{A}}(0,T;\mathbb{R}^n)} \lesssim \| \mathrm{d}_t^\gamma (u - \hat{u}^{\boldsymbol{\tau}}) \|^2_{L^2(0,T;\mathbb{R}^n)} + \|g - \Pi_{\mathscr{T}} g\|^2_{L^2(0,T;\mathbb{R}^n)}.$$

The consistency estimate of Theorem 4 gives a control of the first term. Finally, owing to the regularity of g, we have that $\|g - \Pi_{\mathscr{T}} g\|_{L^2(0,T;\mathbb{R}^n)} \lesssim \tau$; see (21). This implies the result. □

2.3 Numerical Illustration

It is natural to wonder whether the reduced rate of convergence given in Corollary 1 is nothing but a consequence of the methods of proof. Here we show, by means of some computational examples, that while the rate τ^θ might not be sharp it is not possible to obtain the rate of convergence suggested by Proposition 1.

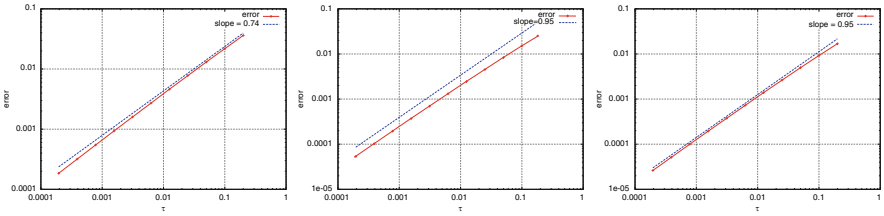

Fig. 1 Experimental rates of convergence for the solution of (2) using (25)–(26). We have set $n = 1$, $T = 1$, $\lambda_1 = \frac{1}{2}$, $\psi = 1$ and $g = 0$. The figures show the computed rates of convergence with respect to the time step for $\gamma = 0.3$ (left), $\gamma = 0.5$ (middle), and $\gamma = 0.8$ (right). We observe that the rate of convergence $\tau^{2-\gamma}$ is never attained

Let us set $n = 1$, $T = 1$, $\lambda_1 = \frac{1}{2}$, $\psi = 1$ and $g = 0$. From (14) we then obtain that the solution to the state equation (2) is given by

$$u(t) = E_{\gamma,1}\left(-\frac{1}{2}t^{\gamma}\right).$$

We implemented, in an in-house code, the scheme (25)–(26) and used it to approximate this function. We measured the $L^2(0,T)$ norm of the error, where we implemented the Mittag-Leffler function following [15]. Integration was carried out using a composite Gaussian rule with three (3) nodes; increasing the number of nodes produced no significant difference in the results.

The rates of convergence for various values of $\gamma \in (0,1)$ are presented in Figure 1. As we can see, Corollary 1 is not sharp, but consistent with the experimental orders. More importantly, the rates suggested by Proposition 1 are not obtained. In fact, the experimental rate of convergence seems to be $\mathscr{O}(\tau^{\kappa}) < \mathscr{O}(\tau^{2-\gamma})$ with $\kappa = \min\{1, \gamma + \frac{1}{2}\}$. However, the proof of such an estimate eludes us at the moment.

3 The Optimization Problem

Having studied the state equation, we can proceed with the study of the constrained optimization problem (4)–(5). We will show existence and uniqueness of a solution, along with a numerical technique to approximate it. We will also discuss the convergence properties of the proposed approximation scheme.

3.1 Existence and Uniqueness

To precisely state the constrained optimization problem, we begin by defining the set of admissible controls

$$Z_{\text{ad}} = \left\{ \zeta \in L^2(0,T;\mathbb{R}^n) : a \preceq \zeta(t) \preceq b,\ a.e.\ t \in (0,T) \right\}, \tag{35}$$

which is, under the assumption that $a \preceq b$, a nonempty, closed, convex, and bounded subset of $L^2(0,T;\mathbb{R}^n)$.

Now, as the conclusion of Theorem 1 asserts, for any $z \in L^2(0,T;\mathbb{R}^n)$ there is a unique $u = u(z) \in \mathbb{U}$ that solves (2). This uniquely defines an affine continuous mapping $\mathfrak{S} : L^2(0,T;\mathbb{R}^n) \to \mathbb{U} \subset L^2(0,T;\mathbb{R}^n)$ by the rule $u = \mathfrak{S}z$, where u solves (2). With these tools at hand, we can show the existence and uniqueness of a state–control pair, that is, a pair $(\breve{u}, \breve{z}) \in \mathbb{U} \times Z_{\text{ad}}$ such that $\breve{u} = \mathfrak{S}\breve{z}$ and satisfies (4)–(5). The proof of the following result is standard and we include it just for the sake of completeness.

Theorem 5 (Existence and Uniqueness) *The optimization problem: Find* (u, z) *such that satisfies* (4) *subject to* (2) *and* (5) *has a unique solution* $(\breve{u}, \breve{z}) \in \mathbb{U} \times Z_{ad}$.

Proof The control to state operator $\mathfrak{S}$ allows us to introduce the so-called reduced cost functional:

$$\mathscr{J}(z) := J(\mathfrak{S}z, z) = \frac{1}{2}\int_0^T \left(|\mathscr{C}\mathfrak{S}z - u_d|_m^2 + \mu |z|_n^2 \right) \mathrm{d}t,$$

and to equivalently state the problem as: minimize $\mathscr{J}$ over Z_{ad}. Since $\mu > 0$ and $\mathfrak{S}$ is affine the reduced cost $\mathscr{J}$ is strictly convex. Owing to the continuity of $\mathfrak{S}$, we have that $\mathscr{J}$ is continuous as well. Existence and uniqueness then follow from the direct method of calculus of variations [7, 23]. □

3.2 Discretization

We now proceed to discretize the optimization problem (4)–(5). We will do so by a piecewise constant approximation of the control and a piecewise linear continuous approximation of the state. We will follow the notation of Section 2.2 and, additionally, define

$$\mathbb{Z}_{\text{ad}}(\mathscr{T}) = \mathbb{Z}(\mathscr{T}) \cap Z_{\text{ad}}.$$

Once again, $\mathbb{Z}_{\text{ad}}(\mathscr{T})$ is a nonempty, convex, and closed subset of $\mathbb{Z}(\mathscr{T})$. Notice also that, since a, b are time independent $\Pi_{\mathscr{T}} Z_{\text{ad}} \subset \mathbb{Z}_{\text{ad}}(\mathscr{T})$.

We also define the discrete cost functional $J_{\mathscr{T}} : \mathbb{U}(\mathscr{T}) \times \mathbb{Z}(\mathscr{T}) \to \mathbb{R}$ by

$$J_{\mathscr{T}}(\hat{U}^{\boldsymbol{\tau}}, Z^{\boldsymbol{\tau}}) = \frac{1}{2}\int_0^T \left(|\mathscr{C}\bar{U}^{\boldsymbol{\tau}} - \bar{u}_d^{\boldsymbol{\tau}}|_m^2 + \mu |Z^{\boldsymbol{\tau}}|_n^2 \right) \mathrm{d}t,$$

where $\mathbb{U}(\mathscr{T})$ and $\mathbb{Z}(\mathscr{T})$ are defined in (19) and (20), respectively. We immediately comment that, by an abuse of notation, we defined $\bar{u}_d^{\boldsymbol{\tau}} \subset \mathbb{R}^m$ as the sequence of

values $u_d^k = \fint_{t_k}^{t_{k+1}} u_d \, \mathrm{d}t$. In other words, we are modifying the cost by replacing the desired state u_d by its piecewise constant approximation $\bar{u}_d^{\boldsymbol{\tau}}$. Additionally, we have replaced $\hat{U}^{\boldsymbol{\tau}}$ by its piecewise constant counterpart $\bar{U}^{\boldsymbol{\tau}} \in \mathbb{Z}(\mathscr{T})$. For these reasons,

$$J_{\mathscr{T}}(\hat{U}^{\boldsymbol{\tau}}, Z^{\boldsymbol{\tau}}) \neq J(\hat{U}^{\boldsymbol{\tau}}, Z^{\boldsymbol{\tau}}).$$

We propose the following discretization of the state equation (2): Given $Z^{\boldsymbol{\tau}} \in \mathbb{Z}(\mathscr{T})$, find $\hat{U}^{\boldsymbol{\tau}} \in \mathbb{U}(\mathscr{T})$ such that $\hat{U}^{\boldsymbol{\tau}}(0) = \psi$ and, for all $k = 0, \ldots, \mathscr{K} - 1$, we have

$$\mathrm{d}_t^\gamma \hat{U}^{\boldsymbol{\tau}}(t_{k+1}) + \mathscr{A}\hat{U}^{\boldsymbol{\tau}}(t_{k+1}) = \Pi_{\mathscr{T}} f(t_{k+1}) + Z^{\boldsymbol{\tau}}(t_{k+1}), \tag{36}$$

where d_t^γ is defined in (3) and $\Pi_{\mathscr{T}}$ corresponds to the $L^2(0,T;\mathbb{R}^n)$-orthogonal projection onto $\mathbb{Z}(\mathscr{T})$. We remark that (36) is nothing but discretization (25)–(26) of the state equation, where the variable z is already piecewise constant in time. Since $f + Z^{\boldsymbol{\tau}} \in L^2(0,T;\mathbb{R}^n)$, we can invoke Theorem 3 to conclude that problem (36) is stable for all $\tau > 0$.

We thus define the discrete optimization problem as follows: Find $(\breve{U}^{\boldsymbol{\tau}}, \breve{Z}^{\boldsymbol{\tau}}) \in \mathbb{U}(\mathscr{T}) \times \mathbb{Z}_{\mathrm{ad}}(\mathscr{T})$ such that

$$J_{\mathscr{T}}(\breve{U}^{\boldsymbol{\tau}}, \breve{Z}^{\boldsymbol{\tau}}) = \min J_{\mathscr{T}}(\hat{U}^{\boldsymbol{\tau}}, Z^{\boldsymbol{\tau}}) \tag{37}$$

subject to (36). Let us briefly comment on the existence and uniqueness of a minimizer, which closely follows Theorem 5. Indeed, for every $z \in L^2(0,T;\mathbb{R}^n)$ there exists a unique $\hat{U}^{\boldsymbol{\tau}} \in \mathbb{U}(\mathscr{T})$ that solves (36) with data $\Pi_{\mathscr{T}} z$. This uniquely defines a map $\mathfrak{S}_{\mathscr{T}} : L^2(0,T;\mathbb{R}^n) \to \mathbb{U}(\mathscr{T})$, which we call the discrete control to state map. We can then define the reduced cost as

$$\mathbb{Z}(\mathscr{T}) \ni Z^{\boldsymbol{\tau}} \mapsto \mathscr{J}_{\mathscr{T}}(Z^{\boldsymbol{\tau}}) = J_{\mathscr{T}}(\widehat{\mathfrak{S}_{\mathscr{T}} Z^{\boldsymbol{\tau}}}, Z^{\boldsymbol{\tau}})$$

and proceed as in Theorem 5, by using the strict convexity of $\mathscr{J}_{\mathscr{T}}$ and the continuity of the affine map $\mathfrak{S}_{\mathscr{T}}$, which follows from Theorem 3.

3.3 Discrete Optimality Conditions

Let us derive discrete optimality conditions. This is useful not only in the practical solution of the discrete optimization problem (36)–(37), but it will help us in analyzing its convergence properties. As stated before, problem (36)–(37) is equivalent to the following constrained optimization problem: Find $\breve{Z}^{\boldsymbol{\tau}} \in \mathbb{Z}_{\mathrm{ad}}(\mathscr{T})$ such that

$$\mathscr{J}_{\mathscr{T}}(\breve{Z}^{\boldsymbol{\tau}}) = \min \left\{ \mathscr{J}_{\mathscr{T}}(Z^{\boldsymbol{\tau}}) : Z^{\boldsymbol{\tau}} \in \mathbb{Z}_{\mathrm{ad}}(\mathscr{T}) \right\},$$

that is, a minimization problem over a closed, bounded, and convex set. It is standard then (since $\mathscr{J}_{\mathscr{T}}$ is convex, coercive, and differentiable) that a necessary and sufficient condition for optimality is

$$D\mathscr{J}_{\mathscr{T}}(\check{Z}^{\boldsymbol{\tau}})\left[Z^{\boldsymbol{\tau}}-\check{Z}^{\boldsymbol{\tau}}\right]\geq 0\quad \forall Z^{\boldsymbol{\tau}}\in\mathbb{Z}_{\mathrm{ad}}(\mathscr{T}), \tag{38}$$

where $D\mathscr{J}_{\mathscr{T}}(Z)[\cdot]$ is the Gâteaux derivative of $\mathscr{J}_{\mathscr{T}}$ at the point Z. Let us now rewrite and simplify the optimality condition (38) by introducing the so-called adjoint state that, as stated in [31, Section 1.4.3], is a *simple trick that is of utmost importance in optimal control theory*.

For a given $\hat{U}^{\boldsymbol{\tau}}\in\mathbb{U}(\mathscr{T})$ the adjoint is the function $\hat{P}^{\boldsymbol{\tau}}\in\mathbb{U}(\mathscr{T})$ such that $\hat{P}^{\boldsymbol{\tau}}(T)=0$ and, for all $k=\mathscr{K}-1,\ldots,0$

$$\mathrm{d}^{\gamma}_{T-t}\hat{P}^{\boldsymbol{\tau}}(t_k)+\mathscr{A}\hat{P}^{\boldsymbol{\tau}}(t_k)=\mathscr{C}^{\intercal}\left(\mathscr{C}\bar{U}^{\boldsymbol{\tau}}(t_k)-\bar{u}_d^{\boldsymbol{\tau}}(t_k)\right), \tag{39}$$

where $\mathrm{d}^{\gamma}_{T-t}$ denotes the right-sided Caputo fractional derivative of order γ defined in (6). The optimality conditions are as follows.

Theorem 6 (Optimality Conditions) *The pair* $(\breve{\hat{U}}^{\boldsymbol{\tau}},\breve{Z}^{\boldsymbol{\tau}})\in\mathbb{U}(\mathscr{T})\times\mathbb{Z}_{ad}(\mathscr{T})$ *solves* (37) *if and only if* $\breve{\hat{U}}^{\boldsymbol{\tau}}=\mathfrak{S}_{\mathscr{T}}\breve{Z}^{\boldsymbol{\tau}}$ *and*

$$\int_0^T\left(\bar{\breve{P}}^{\boldsymbol{\tau}}+\mu\breve{Z}^{\boldsymbol{\tau}}\right)\cdot\left(Z^{\boldsymbol{\tau}}-\breve{Z}^{\boldsymbol{\tau}}\right)\mathrm{d}t\geq 0\quad\forall Z^{\boldsymbol{\tau}}\in\mathbb{Z}_{ad}(\mathscr{T}), \tag{40}$$

where $\breve{P}^{\boldsymbol{\tau}}\in\mathbb{U}(\mathscr{T})$ *solves* (39) *with data* $\breve{\hat{U}}^{\boldsymbol{\tau}}$.

Proof We will obtain the result by showing that (40) is nothing but a restatement of (38). In fact, a simple calculation reveals that, for any $\Theta^{\boldsymbol{\tau}},\Psi^{\boldsymbol{\tau}}\in\mathbb{Z}(\mathscr{T})$, we have

$$D\mathscr{J}_{\mathscr{T}}(\Theta^{\boldsymbol{\tau}})[\Psi^{\boldsymbol{\tau}}]=\int_0^T\left[\left(\mathscr{C}\overline{\mathfrak{S}_{\mathscr{T}}\Theta^{\boldsymbol{\tau}}}-\bar{u}_d^{\boldsymbol{\tau}}\right)\cdot\mathscr{C}\overline{\mathfrak{S}_{\mathscr{T}}\Psi^{\boldsymbol{\tau}}}+\mu\Theta^{\boldsymbol{\tau}}\cdot\Psi^{\boldsymbol{\tau}}\right]\mathrm{d}t.$$

Consequently, (38) can be equivalently rewritten as, for every $Z^{\boldsymbol{\tau}}\in\mathbb{Z}_{\mathrm{ad}}(\mathscr{T})$,

$$\int_0^T\left[\mathscr{C}^{\intercal}\left(\mathscr{C}\overline{\mathfrak{S}_{\mathscr{T}}\breve{Z}^{\boldsymbol{\tau}}}-\bar{u}_d^{\boldsymbol{\tau}}\right)\cdot\overline{\mathfrak{S}_{\mathscr{T}}(Z^{\boldsymbol{\tau}}-\breve{Z}^{\boldsymbol{\tau}})}+\mu\breve{Z}^{\boldsymbol{\tau}}\cdot(Z^{\boldsymbol{\tau}}-\breve{Z}^{\boldsymbol{\tau}})\right]\mathrm{d}t\geq 0. \tag{41}$$

Let us focus our attention now on the first term inside the integral. Denote $U^{\boldsymbol{\tau}}=\mathfrak{S}_{\mathscr{T}}Z^{\boldsymbol{\tau}}$ and $\breve{U}^{\boldsymbol{\tau}}=\mathfrak{S}_{\mathscr{T}}\breve{Z}^{\boldsymbol{\tau}}$. Define $\Phi^{\boldsymbol{\tau}}:=U^{\boldsymbol{\tau}}-\breve{U}^{\boldsymbol{\tau}}$ and notice that $\hat{\Phi}^{\boldsymbol{\tau}}\in\mathbb{U}(\mathscr{T})$ satisfies: $\hat{\Phi}^{\boldsymbol{\tau}}(0)=0$ and, for every $k=0,\ldots,\mathscr{K}-1$,

$$\mathrm{d}^{\gamma}_{t}\hat{\Phi}^{\boldsymbol{\tau}}(t_{k+1})+\mathscr{A}\hat{\Phi}^{\boldsymbol{\tau}}(t_{k+1})=Z^{\boldsymbol{\tau}}(t_{k+1})-\breve{Z}^{\boldsymbol{\tau}}(t_{k+1}),$$

or, in view of (22), equivalently,

$$\overline{\mathrm{d}_t^\gamma \hat{\Phi}^{\boldsymbol{\tau}}} + \mathscr{A}\bar{\Phi}^{\boldsymbol{\tau}} = Z^{\boldsymbol{\tau}} - \check{Z}^{\boldsymbol{\tau}}.$$

Multiply this equation by $\bar{\check{P}}^{\boldsymbol{\tau}}$ and integrate to obtain

$$\int_0^T \left[\overline{\mathrm{d}_t^\gamma \hat{\Phi}^{\boldsymbol{\tau}}} \cdot \bar{\check{P}}^{\boldsymbol{\tau}} + \mathscr{A}\bar{\Phi}^{\boldsymbol{\tau}} \cdot \bar{\check{P}}^{\boldsymbol{\tau}} \right] \mathrm{d}t = \int_0^T \left(Z^{\boldsymbol{\tau}} - \check{Z}^{\boldsymbol{\tau}} \right) \cdot \bar{\check{P}}^{\boldsymbol{\tau}} \, \mathrm{d}t.$$

Now, multiply (39) by $\bar{\Phi}^{\boldsymbol{\tau}}$ and integrate to obtain

$$\int_0^T \left[\overline{\mathrm{d}_{T-t}^\gamma \hat{\check{P}}^{\boldsymbol{\tau}}} \cdot \bar{\Phi}^{\boldsymbol{\tau}} + \mathscr{A}\bar{\check{P}}^{\boldsymbol{\tau}} \cdot \bar{\Phi}^{\boldsymbol{\tau}} \right] \mathrm{d}t = \int_0^T \mathscr{C}^{\intercal} \left(\mathscr{C}\bar{\check{U}}^{\boldsymbol{\tau}} - \bar{u}_d^{\boldsymbol{\tau}} \right) \cdot \bar{\Phi}^{\boldsymbol{\tau}} \, \mathrm{d}t.$$

Subtract these last two identities. Upon remembering the definition of $\Phi^{\boldsymbol{\tau}}$, we thus obtain

$$\begin{aligned} &\int_0^T \left[\overline{\mathrm{d}_t^\gamma \hat{\Phi}^{\boldsymbol{\tau}}} \cdot \bar{\check{P}}^{\boldsymbol{\tau}} - \overline{\mathrm{d}_{T-t}^\gamma \hat{\check{P}}^{\boldsymbol{\tau}}} \cdot \bar{\Phi}^{\boldsymbol{\tau}} \right] \mathrm{d}t \\ &\quad = \int_0^T \left[\left(Z^{\boldsymbol{\tau}} - \check{Z}^{\boldsymbol{\tau}} \right) \cdot \bar{\check{P}}^{\boldsymbol{\tau}} - \mathscr{C}^{\intercal} \left(\mathscr{C}\bar{\check{U}}^{\boldsymbol{\tau}} - \bar{u}_d^{\boldsymbol{\tau}} \right) \cdot \overline{\mathfrak{S}_{\mathscr{T}}(Z^{\boldsymbol{\tau}} - \check{Z}^{\boldsymbol{\tau}})} \right] \mathrm{d}t, \end{aligned}$$

where we have used that the matrix $\mathscr{A}$ is symmetric. Notice that the last term in this expression is nothing but the first term on the left-hand side of (41). In other words, if we can show that

$$\int_0^T \overline{\mathrm{d}_t^\gamma \hat{\Phi}^{\boldsymbol{\tau}}} \cdot \bar{\check{P}}^{\boldsymbol{\tau}} \, \mathrm{d}t = \int_0^T \overline{\mathrm{d}_{T-t}^\gamma \hat{\check{P}}^{\boldsymbol{\tau}}} \cdot \bar{\Phi}^{\boldsymbol{\tau}} \, \mathrm{d}t \tag{42}$$

we obtain the result.

To show this we realize that, since we are dealing with piecewise constants, we can equivalently rewrite the left-hand side of this identity as

$$\begin{aligned} \int_0^T \overline{\mathrm{d}_t^\gamma \hat{\Phi}^{\boldsymbol{\tau}}} \cdot \bar{\check{P}}^{\boldsymbol{\tau}} \, \mathrm{d}t &= \tau \sum_{k=0}^{\mathscr{K}-1} \check{P}^{k+1} \cdot \mathrm{d}_t^\gamma \hat{\Phi}^{\boldsymbol{\tau}}(t_{k+1}) \\ &= \frac{\tau^{1-\gamma}}{\Gamma(2-\gamma)} \sum_{k=0}^{\mathscr{K}-1} \check{P}^{k+1} \cdot \sum_{m=0}^{k} b_m \mathfrak{d}\Phi^{k+1-m}, \end{aligned}$$

where we used (29).

In a similar manner to the computations of Remark 2, we can obtain that

$$\mathrm{d}_{T-t}^\gamma \hat{\check{P}}(t_k) = -\sum_{j=k}^{\mathscr{K}-1} a_k^j \mathfrak{d}\check{P}^{j+1} = -\frac{\tau^{-\gamma}}{\Gamma(2-\gamma)} \sum_{j=k}^{\mathscr{K}-1} b_{j-k} \mathfrak{d}\check{P}^{j+1},$$

consequently

$$\int_0^T \overline{\mathrm{d}_{T-t}^{\gamma} \hat{\check{P}}^{\boldsymbol{\tau}}} \cdot \bar{\Phi}^{\boldsymbol{\tau}} \, \mathrm{d}t = \frac{\tau^{1-\gamma}}{\Gamma(2-\gamma)} \sum_{k=1}^{\mathscr{K}} \Phi^k \cdot \sum_{j=k}^{\mathscr{K}-1} b_{j-k} \eth \check{P}^{j+1}.$$

We can invoke now the results of [4, Appendix A] to conclude that the identity (42) holds. The theorem is thus proven. □

Remark 3 (Discrete Fractional Integration by Parts) Notice that, during the course of the proof of Theorem 6 we showed that, whenever $\hat{V}^{\boldsymbol{\tau}}, \hat{W}^{\boldsymbol{\tau}} \in \mathbb{U}(\mathscr{T})$ satisfy $\hat{V}^{\boldsymbol{\tau}}(0) = 0$ and $\hat{W}^{\boldsymbol{\tau}}(T) = 0$, then they satisfy the following discrete fractional integration by parts

$$\int_0^T \overline{\mathrm{d}_t^{\gamma} \hat{V}^{\boldsymbol{\tau}}} \cdot \bar{W}^{\boldsymbol{\tau}} \, \mathrm{d}t = \int_0^T \overline{\mathrm{d}_{T-t}^{\gamma} \hat{W}^{\boldsymbol{\tau}}} \cdot \bar{V}^{\boldsymbol{\tau}} \, \mathrm{d}t.$$

This identity shall prove useful in the sequel.

Remark 4 (Projection) The solution to the variational inequality (40) can be accomplished rather easily. Indeed, since all the involved functions belong to $\mathbb{Z}(\mathscr{T})$, it suffices to consider one time interval, say $(t_{k-1}, t_k]$, where we must have

$$\left(\check{P}^k + \mu \check{Z}^k\right) \cdot \left(Z^k - \check{Z}^k\right) \geq 0.$$

From this it immediately follows that

$$\check{Z}^k = \underset{[a,b]}{\mathrm{Pr}} \left(\frac{-1}{\mu} \check{P}^k \right),$$

where, for $w \in \mathbb{R}^n$, we define $\mathrm{Pr}_{[a,b]} w$ as the projection onto the cube $[a, b] = \{x \in \mathbb{R}^n : a \preceq x \preceq b\}$, which can be easily accomplished by the formula

$$\underset{[a,b]}{\mathrm{Pr}} \, w_i = \max\left\{a_i, \min\left\{b_i, w_i\right\}\right\}, \quad i = 1, \ldots, n.$$

This is the main advantage of considering piecewise constant approximations of the control and a modified cost. Other variants might yield a better approximation, but at the cost of a more involved solution scheme.

3.4 Convergence

Let us now discuss the convergence of our approximation scheme. The main issue here is that since, even for a smooth f, the right-hand side of (36) belongs only

to $L^2(0,T;\mathbb{R}^n)$ we cannot invoke the results of Corollary 1 to establish a rate of convergence. Notice, additionally, that we modified the cost, one of the reasons being that this led us to the simplifications detailed in Remark 4. As a consequence we only show convergence without rates.

We begin by noticing that, for any $z \in L^2(0,T;\mathbb{R}^n)$ we have that $\mathfrak{S}_{\mathscr{T}} z = \mathring{\mathfrak{S}}_{\mathscr{T}} z + \hat{V}^{\boldsymbol{\tau}}$, where $\hat{V}^{\boldsymbol{\tau}} \in \mathbb{U}(\mathscr{T})$ satisfies

$$\hat{V}^{\boldsymbol{\tau}}(0) = \psi, \qquad \mathrm{d}_t^{\gamma} \hat{V}^{\boldsymbol{\tau}}(t_{k+1}) + \mathscr{A} \hat{V}^{\boldsymbol{\tau}}(t_{k+1}) = \Pi_{\mathscr{T}} f(t_{k+1}),\ k = 0, \ldots, \mathscr{K} - 1,$$

and the linear, continuous operator $\mathring{\mathfrak{S}}_{\mathscr{T}}$ is the solution operator for the scheme: Find $\hat{U}_0^{\boldsymbol{\tau}} \in \mathbb{U}(\mathscr{T})$ such that $\hat{U}_0^{\boldsymbol{\tau}}(0) = 0$ and, for $k = 0, \ldots, \mathscr{K} - 1$,

$$\mathrm{d}_t^{\gamma} \hat{U}_0^{\boldsymbol{\tau}}(t_{k+1}) + \mathscr{A} \hat{U}_0^{\boldsymbol{\tau}}(t_{k+1}) = \Pi_{\mathscr{T}} z(t_{k+1}). \tag{43}$$

Let us describe the properties of $\hat{V}^{\boldsymbol{\tau}}$.

Proposition 2 (Properties of $\hat{V}^{\boldsymbol{\tau}}$) *Assume that $f \in L^2(0,T;\mathbb{R}^n)$, then the family $\{\hat{V}^{\boldsymbol{\tau}}\}_{\mathscr{T}}$ converges, as $\mathscr{K} \to \infty$, in $L^2(0,T;\mathbb{R}^n)$ to $v \in \mathbb{U}$, which solves*

$$\mathrm{d}_t^{\gamma} v + \mathscr{A} v = f, \quad t \in (0,T], \qquad v(0) = \psi.$$

Proof The claimed result is obtained by a simple density argument, combined with stability of the continuous and discrete state equations. Let $\epsilon > 0$. Since $f \in L^2(0,T;\mathbb{R}^n)$, there is a $f_\epsilon \in H^2(0,T;\mathbb{R}^n)$ such that

$$\|f - f_\epsilon\|_{L^2(0,T;\mathbb{R}^n)} < \frac{\epsilon}{4C_1},$$

where by C_1 we denote the constant in inequality (16). Denote by v_ϵ the solution to

$$\mathrm{d}_t^{\gamma} v_\epsilon + \mathscr{A} v_\epsilon = f_\epsilon, \quad t \in (0,T], \qquad v(0) = \psi.$$

The smoothness of f_ϵ allows us to invoke Theorem 2 to assert that the regularity estimates (18), with u replaced by v_ϵ, hold. In addition, invoking Theorem 1, we get that

$$\|v - v_\epsilon\|_{L^2_{\mathscr{A}}(0,T;\mathbb{R}^n)} \leq C_1 \Lambda_\gamma(0, f - f_\epsilon) = C_1 \|f - f_\epsilon\|_{L^2(0,T;\mathbb{R}^n)} < \frac{\epsilon}{4}.$$

Let us now approximate v_ϵ via the scheme (26), over a mesh $\mathscr{T}$ where $\mathscr{K}$ remains to be chosen. In doing so we obtain a function $\hat{V}_\epsilon^{\boldsymbol{\tau}} \in \mathbb{U}(\mathscr{T})$. Moreover, since v_ϵ verifies the assumptions of Theorem 2, we invoke Corollary 1 to conclude that

$$\|\bar{v}_\epsilon - \bar{V}_\epsilon^{\boldsymbol{\tau}}\|_{L^2_{\mathscr{A}}(0,T;\mathbb{R}^n)} \leq C_2 \tau^{\theta},$$

where C_2 denotes a positive constant that depends on $\|f_\epsilon\|_{H^2(0,T;\mathbb{R}^n)}$. However, since ϵ is fixed, we can choose $\mathscr{K}$ so that

$$C_2\tau^\theta < \frac{\epsilon}{4} \implies \|\bar{v}_\epsilon - \bar{V}_\epsilon^{\boldsymbol{\tau}}\|_{L^2(0,T;\mathbb{R}^n)} < \frac{\epsilon}{4}.$$

The last ingredient is to observe that the difference $V_\epsilon^{\boldsymbol{\tau}} - V^{\boldsymbol{\tau}}$ solves (25)–(26) with zero initial condition and right-hand side $\Pi_{\mathscr{T}}(f - f_\epsilon)$. We then invoke the stability of the scheme, stated in Theorem 3, to obtain

$$\|\bar{V}_\epsilon^{\boldsymbol{\tau}} - \bar{V}^{\boldsymbol{\tau}}\|_{L^2_{\mathscr{A}}(0,T;\mathbb{R}^n)} \le C_1 \Lambda_\gamma(0, \Pi_{\mathscr{T}}(f - f_\epsilon)) \le C_1 \|f - f_\epsilon\|_{L^2(0,T;\mathbb{R}^n)} < \frac{\epsilon}{4},$$

where we used that $\Pi_{\mathscr{T}}$ is a projection.

Combine these observations to conclude that

$$\begin{aligned}\|v - \bar{V}^{\boldsymbol{\tau}}\|_{L^2_{\mathscr{A}}(0,T,\mathbb{R}^n)} &\le \|v - v_\epsilon\|_{L^2_{\mathscr{A}}(0,T;\mathbb{R}^n)} + \|v_\epsilon - \bar{v}_\epsilon\|_{L^2_{\mathscr{A}}(0,T;\mathbb{R}^n)} \\ &\quad + \|\bar{v}_\epsilon - \bar{V}_\epsilon^{\boldsymbol{\tau}}\|_{L^2_{\mathscr{A}}(0,T;\mathbb{R}^n)} + \|\bar{V}_\epsilon^{\boldsymbol{\tau}} - \bar{V}^{\boldsymbol{\tau}}\|_{L^2_{\mathscr{A}}(0,T;\mathbb{R}^n)} \\ &< \frac{3\epsilon}{4} + \|v_\epsilon - \bar{v}_\epsilon\|_{L^2_{\mathscr{A}}(0,T;\mathbb{R}^n)}.\end{aligned}$$

Conclude by noticing that, since $v_\epsilon \to v$, after possibly taking an even larger $\mathscr{K}$ we can assert

$$\|v_\epsilon - \bar{v}_\epsilon\|_{L^2_{\mathscr{A}}(0,T;\mathbb{R}^n)} < \frac{\epsilon}{4}.$$

This concludes the proof. □

The main consequence of this statement arises when we use the decomposition of $\mathfrak{S}_{\mathscr{T}}$ in the reduced cost. Namely, we get

$$\mathscr{J}_{\mathscr{T}}(Z^{\boldsymbol{\tau}}) = \frac{1}{2}\int_0^T \left[|\mathscr{C}\overline{\mathring{\mathfrak{S}}_{\mathscr{T}} Z^{\boldsymbol{\tau}}} - \bar{W}^{\boldsymbol{\tau}}|_m^2 + \mu |Z^{\boldsymbol{\tau}}|_n^2\right] \mathrm{d}t,$$

for $W^{\boldsymbol{\tau}} = u_d^{\boldsymbol{\tau}} - \mathscr{C}V^{\boldsymbol{\tau}}$, that is, the discrete desired state changes and, moreover, $W^{\boldsymbol{\tau}} \to u_d - \mathscr{C}v$ in $L^2(0,T;\mathbb{R}^m)$ as $\mathscr{K} \to \infty$. All these considerations allow us to reduce the problem to the case when $\psi = 0$ and $f \equiv 0$ so that the discrete control to state map is a linear operator.

In this setting we can assert the strong convergence of $\mathring{\mathfrak{S}}_{\mathscr{T}}$ and its adjoint, which will be a fundamental tool in proving convergence. Here and in what follows, we denote by $\mathfrak{B}(L^2(0,T;\mathbb{R}^n))$ the space of bounded linear operators on $L^2(0,T;\mathbb{R}^n)$ endowed with the operator norm.

Lemma 2 (Strong Convergence) *The family of solution operators* $\{\mathring{\mathfrak{S}}_{\mathscr{T}}\}_{\mathscr{T}}$ *and of their adjoints* $\left\{\mathring{\mathfrak{S}}^\star_{\mathscr{T}}\right\}_{\mathscr{T}}$ *is uniformly bounded in* $\mathfrak{B}(L^2(0,T;\mathbb{R}^n))$ *and strongly convergent.*

Proof We begin by realizing that the uniform boundedness, in $\mathfrak{B}(L^2(0,T;\mathbb{R}^n))$, of $\{\mathring{\mathfrak{S}}_{\mathscr{T}}\}_{\mathscr{T}}$ is a restatement of Theorem 3, see [13, 18]. Moreover, the error estimates of Corollary 1 are valid for a collection of right-hand sides that is dense in $L^2(0,T;\mathbb{R}^n)$. This means, by an argument similar to the one provided in Proposition 2, that for every $z \in L^2(0,T;\mathbb{R}^n)$ the family $\{\mathring{\mathfrak{S}}_{\mathscr{T}} z\}_{\mathscr{T}}$ converges; see [13, Proposition 5.17].

Let us now prove the same statements for the family of adjoints. To do so we must first identify it. Let $z, \eta \in L^2(0,T;\mathbb{R}^n)$ and $\hat{U}_0^{\boldsymbol{\tau}}$ solve (43). In addition, let $\hat{P}^{\boldsymbol{\tau}} \in \mathbb{U}(\mathscr{T})$ be the solution to (39) but with the right-hand side replaced by $\Pi_{\mathscr{T}}\eta$. Multiply the aforementioned equations by $\bar{P}^{\boldsymbol{\tau}}$ and $\bar{U}_0^{\boldsymbol{\tau}}$, integrate and subtract to obtain

$$\int_0^T \left[\overline{\mathrm{d}_t^{\gamma} \hat{U}_0^{\boldsymbol{\tau}}} \cdot \bar{P}^{\boldsymbol{\tau}} - \overline{\mathrm{d}_{T-t}^{\gamma} \hat{P}^{\boldsymbol{\tau}}} \cdot \bar{U}_0^{\boldsymbol{\tau}} \right] \mathrm{d}t = \int_0^T \left[\Pi_{\mathscr{T}} z \cdot \bar{P}^{\boldsymbol{\tau}} - \Pi_{\mathscr{T}} \eta \cdot \overline{\mathring{\mathfrak{S}}_{\mathscr{T}} z} \right] \mathrm{d}t$$

where we used that the matrix $\mathscr{A}$ is symmetric. We now invoke Remark 3 to conclude that the right-hand side of the previous expression vanishes, which implies that

$$\begin{aligned} \int_0^T z \cdot \overline{\mathring{\mathfrak{S}}^{\star}_{\mathscr{T}} \eta} \, \mathrm{d}t &= \int_0^T \Pi_{\mathscr{T}} z \cdot \overline{\mathring{\mathfrak{S}}^{\star}_{\mathscr{T}} \eta} \, \mathrm{d}t = \int_0^T \Pi_{\mathscr{T}} \eta \cdot \overline{\mathring{\mathfrak{S}}_{\mathscr{T}} z} \, \mathrm{d}t \\ &= \int_0^T \Pi_{\mathscr{T}} z \cdot \bar{P}^{\boldsymbol{\tau}} \, \mathrm{d}t = \int_0^T z \cdot \bar{P}^{\boldsymbol{\tau}} \, \mathrm{d}t, \end{aligned}$$

where the first and last equalities hold by the definition of $\Pi_{\mathscr{T}}$. Since the last identity holds for every $z \in L^2(0,T;\mathbb{R}^n)$, we thus have that $P^{\boldsymbol{\tau}} = \mathring{\mathfrak{S}}^{\star}_{\mathscr{T}} \eta$.

It now remains to realize that $P^{\boldsymbol{\tau}}$ is a discretization of the problem

$$\mathrm{d}_{T-t}^{\gamma} p + \mathscr{A} p = \eta, \ t \in [0,T), \quad p(T) = 0.$$

Repeating the arguments that led to Theorem 3 and Corollary 1, we get that $P^{\boldsymbol{\tau}}$ is a stable and consistent approximation, so we can, again, conclude the uniform boundedness and convergence of the family $\left\{ \mathring{\mathfrak{S}}^{\star}_{\mathscr{T}} \right\}_{\mathscr{T}}$. □

We are now ready to establish convergence of our scheme.

Theorem 7 (Convergence) *The family $\{\breve{Z}^{\boldsymbol{\tau}}\}_{\mathscr{T}}$ of optimal controls is uniformly bounded and contains a subsequence that converges strongly to $\breve{z}$, the solution to (4).*

Proof Boundedness is a consequence of optimality. Indeed, if $z_0 \in Z_{\mathrm{ad}}$ then

$$\frac{\mu}{2} \|\breve{Z}^{\boldsymbol{\tau}}\|^2_{L^2(0,T;\mathbb{R}^n)} \leq \mathscr{J}_{\mathscr{T}}(\breve{Z}^{\boldsymbol{\tau}}) \leq \mathscr{J}_{\mathscr{T}}(\Pi_{\mathscr{T}} z_0) \lesssim \|z_0\|^2_{L^2(0,T;\mathbb{R}^n)} + \|u_d\|^2_{L^2(0,T;\mathbb{R}^m)},$$

where we used the continuity of $\mathfrak{S}_{\mathscr{T}}$ and $\Pi_{\mathscr{T}}$. This implies the existence of a (not relabeled) weakly convergent subsequence.

To show convergence of this sequence to $\check{z}$, we invoke the theory of Γ-convergence [7], so that we must verify three assumptions:

1. *Lower bound*: We must show that, whenever $Z^{\boldsymbol{\tau}} \rightharpoonup z$ then $\mathscr{J}(z) \leq \liminf \mathscr{J}_{\mathscr{T}}(Z^{\boldsymbol{\tau}})$. To do so, let $\eta \in L^2(0,T;\mathbb{R}^n)$ and notice that

$$\int_0^T \left[\overline{\mathfrak{S}_{\mathscr{T}} Z^{\boldsymbol{\tau}}} - \mathfrak{S} z\right] \cdot \eta \,\mathrm{d}t = \int_0^T \left[\overline{\mathfrak{S}_{\mathscr{T}} z} - \mathfrak{S} z\right] \cdot \eta \,\mathrm{d}t + \int_0^T \overline{\mathfrak{S}_{\mathscr{T}}(Z^{\boldsymbol{\tau}} - z)} \cdot \eta \,\mathrm{d}t$$
$$= A + B.$$

   The pointwise convergence of $\{\mathring{\mathfrak{S}}_{\mathscr{T}}\}_{\mathscr{T}}$ shows that $A \to 0$, while the pointwise convergence of the adjoints shows that $B \to 0$. In conclusion, $\mathfrak{S}_{\mathscr{T}} Z^{\boldsymbol{\tau}} \rightharpoonup \mathfrak{S} z$. Now, owing to the weak lower semicontinuity of norms, and the fact that $\bar{u}_d^{\boldsymbol{\tau}} \to u_d$ in $L^2(0,T;\mathbb{R}^m)$ we conclude

$$\mathscr{J}(z) \leq \liminf \mathscr{J}_{\mathscr{T}}(Z^{\boldsymbol{\tau}}).$$

2. *Existence of a recovery sequence*: We must show that, for every $z \in Z_{\mathrm{ad}}$ there is $Z^{\boldsymbol{\tau}} \in \mathbb{Z}_{\mathrm{ad}}(\mathscr{T})$ such that $Z^{\boldsymbol{\tau}} \rightharpoonup z$ and $\mathscr{J}(z) \geq \limsup \mathscr{J}_{\mathscr{T}}(Z^{\boldsymbol{\tau}})$. To do so, it suffices to set $Z^{\boldsymbol{\tau}} = \Pi_{\mathscr{T}} z$. Indeed, we even have strong convergence so that we can say $\mathfrak{S}_{\mathscr{T}} \Pi_{\mathscr{T}} z \to \mathfrak{S} z$. Continuity of norms and the convergence of $\bar{u}_d^{\boldsymbol{\tau}}$ allow us to conclude the inequality for the costs.
3. *Equicoerciveness*: We must show that, for every $r \in \mathbb{R}$, there is a weakly closed and weakly compact $K_r \subset L^2(0,T;\mathbb{R}^n)$ such that, for all $\mathscr{T}$, the r-sublevel set of $\mathscr{J}_{\mathscr{T}}$ is contained in K_r. To do so it suffices to notice that

$$\mathscr{J}_{\mathscr{T}}(Z^{\boldsymbol{\tau}}) \geq \frac{\mu}{2} \|Z^{\boldsymbol{\tau}}\|^2_{L^2(0,T;\mathbb{R}^n)}.$$

   Thus, invoking [7, Proposition 7.7], we can immediately conclude.

With these three ingredients, we can now show convergence. Indeed, the lower bound inequality and recovery sequence property allow us to say that

$$\mathscr{J}_{\mathscr{T}} \xrightarrow{\Gamma} \mathscr{J}$$

so that minimizers of $\mathscr{J}_{\mathscr{T}}$ converge to minimizers of $\mathscr{J}$. Equicoerciveness and the uniqueness of $\check{z}$ are the conditions of the so-called fundamental lemma of Γ-convergence [7, Corollary 7.24] which allow us to conclude that $\check{Z}^{\boldsymbol{\tau}} \rightharpoonup \check{z}$.

We finalize the proof by showing strong convergence. To do so we first note that, by Dal Maso [7, equation (7.32)], we have $\mathscr{J}_{\mathscr{T}}(\check{Z}^{\boldsymbol{\tau}}) \to \mathscr{J}(\check{z})$. Therefore

$$\frac{1}{2}\int_0^T \left[\left|\overline{\mathfrak{S}_{\mathscr{T}}\check{Z}^{\boldsymbol{\tau}}} - \mathfrak{S}z\right|_n^2 + \mu\left|\bar{\check{Z}}^{\boldsymbol{\tau}} - \check{z}\right|_n^2\right]^2 \mathrm{d}t = \mathscr{J}_{\mathscr{T}}(\check{Z}^{\boldsymbol{\tau}}) + \mathscr{J}(\check{z})$$
$$- \int_0^T \overline{\mathfrak{S}_{\mathscr{T}}\check{Z}^{\boldsymbol{\tau}}} \cdot \left(\mathfrak{S}\check{z} - \bar{u}_d^{\boldsymbol{\tau}}\right) \mathrm{d}t$$
$$+ \int_0^T u_d \cdot \left(\mathfrak{S}\check{z} - \bar{u}_d^{\boldsymbol{\tau}}\right) \mathrm{d}t$$
$$- \mu \int_0^T \check{Z}^{\boldsymbol{\tau}} \cdot \check{z}\, \mathrm{d}t$$
$$\to \mathscr{J}(\check{z}) + \mathscr{J}(\check{z}) - 2\mathscr{J}(\check{z}) = 0,$$

where we, again, used the convergence of the adjoint.

This concludes the proof of convergence. □

We conclude by showing weak convergence of the state.

Corollary 2 (State Convergence) *In the setting of Theorem 7 we have that* $\check{U}^{\boldsymbol{\tau}} \rightharpoonup \check{u}$ *in* $L^2(0,T;\mathbb{R}^n)$.

Proof This follows from the strong convergence of $\check{Z}^{\boldsymbol{\tau}}$ and of the adjoints $\mathring{\mathfrak{S}}^{\star}_{\mathscr{T}}$. Indeed, let $v \in L^2(0,T;\mathbb{R}^n)$ and notice that

$$\int_0^T \mathring{\mathfrak{S}}_{\mathscr{T}}\check{Z}^{\boldsymbol{\tau}} \cdot v\, \mathrm{d}t = \int_0^T \check{Z}^{\boldsymbol{\tau}} \cdot \mathring{\mathfrak{S}}^{\star}_{\mathscr{T}} v\, \mathrm{d}t \to \int_0^T \check{z} \cdot \mathring{\mathfrak{S}}^{\star} v\, \mathrm{d}t.$$

Since $\check{U}^{\boldsymbol{\tau}} = \mathring{\mathfrak{S}}_{\mathscr{T}}\check{Z}^{\boldsymbol{\tau}} + V^{\boldsymbol{\tau}}$, we obtain the result by invoking Proposition 2. □

Acknowledgements E. Otárola was supported in part by CONICYT through FONDECYT project 3160201. A.J. Salgado was supported in part by NSF grant DMS-1418784.

References

1. N.I. Achieser. *Theory of approximation*. Dover Publications, Inc., New York, 1992. Translated from the Russian and with a preface by Charles J. Hyman, Reprint of the 1956 English translation.
2. O.P. Agrawal. A general formulation and solution scheme for fractional optimal control problems. *Nonlinear Dynam.*, 38(1–4):323–337, 2004.
3. O.P. Agrawal. A general finite element formulation for fractional variational problems. *J. Math. Anal. Appl.*, 337(1):1–12, 2008.
4. H. Antil, E. Otárola, and A.J. Salgado. A space-time fractional optimal control problem: analysis and discretization. *SIAM J. Control Optim.*, 54(3):1295–1328, 2016.
5. M. Caputo. Diffusion of fluids in porous media with memory. *Geothermics*, 28(1):113 – 130, 1999.

6. A. Cartea and D. del Castillo-Negrete. Fractional diffusion models of option prices in markets with jumps. *Physica A*, 374:749–763, 2007.
7. G. Dal Maso. *An introduction to Γ-convergence*. Progress in Nonlinear Differential Equations and their Applications, 8. Birkhäuser Boston, Inc., Boston, MA, 1993.
8. L. Debnath. Fractional integral and fractional differential equations in fluid mechanics. *Fract. Calc. Appl. Anal.*, 6(2):119–155, 2003.
9. D. del Castillo-Negrete. Fractional diffusion models of nonlocal transport. *Phys. Plasmas*, 13(8):082308, 16, 2006.
10. D. del Castillo-Negrete, B. A. Carreras, and V. E. Lynch. Nondiffusive transport in plasma turbulence: A fractional diffusion approach. *Phys. Rev. Lett.*, 94:065003, Feb 2005.
11. K. Diethelm. *The analysis of fractional differential equations*, volume 2004 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2010. An application-oriented exposition using differential operators of Caputo type.
12. T.M. Flett. A note on some inequalities. *Proc. Glasgow Math. Assoc.*, 4:7–15 (1958), 1958.
13. G.B. Folland. *Real analysis*. John Wiley & Sons, Inc., New York, 1984.
14. R. Gorenflo, A.A. Kilbas, F. Mainardi, and S.V. Rogosin. *Mittag-Leffler functions, related topics and applications*. Springer Monographs in Mathematics. Springer, Heidelberg, 2014.
15. R. Gorenflo, J. Loutchko, and Y. Luchko. Computation of the Mittag-Leffler function $E_{\alpha,\beta}(z)$ and its derivative. *Fract. Calc. Appl. Anal.*, 5(4):491–518, 2002. Dedicated to the 60th anniversary of Prof. Francesco Mainardi.
16. B. Jin, R. Lazarov, and Z. Zhou. An analysis of the L1 scheme for the subdiffusion equation with nonsmooth data. *IMA J. Numer. Anal.*, 36(1):197–221, 2016.
17. J.-P. Kahane. Teoría constructiva de funciones. *Universidad de Buenos Aires, Buenos Aires]*, 1961:111, 1961.
18. L. V. Kantorovich and G. P. Akilov. *Funktsionalnyi analiz*. "Nauka", Moscow, third edition, 1984.
19. A. A. Kilbas, H. M. Srivastava, and J. J. Trujillo. *Theory and applications of fractional differential equations*, volume 204 of *North-Holland Mathematics Studies*. Elsevier Science B.V., Amsterdam, 2006.
20. M.A. Krasnosel'skiĭ and Ja.B. Rutickiĭ. *Convex functions and Orlicz spaces*. Translated from the first Russian edition by Leo F. Boron. P. Noordhoff Ltd., Groningen, 1961.
21. Y. Lin, X. Li, and C. Xu. Finite difference/spectral approximations for the fractional cable equation. *Math. Comp.*, 80(275):1369–1396, 2011.
22. Y. Lin and C. Xu. Finite difference/spectral approximations for the time-fractional diffusion equation. *J. Comput. Phys.*, 225(2):1533–1552, 2007.
23. J.-L. Lions. *Optimal control of systems governed by partial differential equations.* Translated from the French by S. K. Mitter. Die Grundlehren der mathematischen Wissenschaften, Band 170. Springer-Verlag, New York-Berlin, 1971.
24. A. Lotfi and S. A. Yousefi. A numerical technique for solving a class of fractional variational problems. *J. Comput. Appl. Math.*, 237(1):633–643, 2013.
25. W. McLean. Regularity of solutions to a time-fractional diffusion equation. *ANZIAM J.*, 52(2):123–138, 2010.
26. R.H. Nochetto, E. Otárola, and A.J. Salgado. A PDE approach to space-time fractional parabolic problems. *SIAM J. Numer. Anal.*, 54(2):848–873, 2016.
27. K. Sakamoto and M. Yamamoto. Initial value/boundary value problems for fractional diffusion-wave equations and applications to some inverse problems. *J. Math. Anal. Appl.*, 382(1): 426–447, 2011.
28. S. G. Samko, A. A. Kilbas, and O. I. Marichev. *Fractional integrals and derivatives*. Gordon and Breach Science Publishers, Yverdon, 1993. Theory and applications, Edited and with a foreword by S. M. Nikol'skiĭ, Translated from the 1987 Russian original, Revised by the authors.
29. Z. Sun and X. Wu. A fully discrete difference scheme for a diffusion-wave system. *Appl. Numer. Math.*, 56(2):193–209, 2006.

30. C. Tricaud and Y. Chen. An approximate method for numerically solving fractional order optimal control problems of general form. *Comput. Math. Appl.*, 59(5):1644–1655, 2010.
31. F. Tröltzsch. *Optimal control of partial differential equations*, volume 112 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2010. Theory, methods and applications, Translated from the 2005 German original by Jürgen Sprekels.
32. X. Ye and C. Xu. Spectral optimization methods for the time fractional diffusion inverse problem. *Numer. Math. Theory Methods Appl.*, 6(3):499–516, 2013.
33. X. Ye and C. Xu. A spectral method for optimal control problems governed by the time fractional diffusion equation with control constraints. In *Spectral and high order methods for partial differential equations—ICOSAHOM 2012*, volume 95 of *Lect. Notes Comput. Sci. Eng.*, pages 403–414. Springer, Cham, 2014.

Sensitivity-Based Topology and Shape Optimization with Application to Electric Motors

Peter Gangl

Abstract In many industrial applications, one is interested in finding an optimal layout of an object, which often leads to PDE-constrained shape optimization problems. Such problems can be approached by shape optimization methods, where a domain is altered by smooth deformation of its boundary, or by means of topology optimization methods, which in addition can alter the connectivity of the initial design. We give an overview over established topology optimization methods and focus on an approach based on the sensitivity of the cost function with respect to a topological perturbation of the domain, called the topological derivative. We illustrate a way to derive this sensitivity and discuss the additional difficulties arising in the case of a nonlinear PDE constraint. We show numerical results for the optimization of an electric motor which are obtained by a combination of two methods: a level set algorithm which is based on the topological derivative, and a shape optimization method together with a special treatment of the evolving material interface which assures accurate approximate solutions to the underlying PDE constraint as well as a smooth final design.

1 Introduction

This chapter deals with PDE-constrained topology and shape optimization and is motivated by a concrete application from electrical engineering, namely the design optimization of an electric motor. The goal is to identify an admissible subset Ω of the design region Ω^d of the motor which yields the best possible performance of the motor. The performance is measured by a functional J which is related to the smoothness of the rotation or to the torque of the motor. In shape optimization,

P. Gangl (✉)
Graz University of Technology, Institute of Applied Mathematics, Graz, Austria
e-mail: gangl@math.tugraz.at

H. Antil et al. (eds.), *Frontiers in PDE-Constrained Optimization*, The IMA Volumes in Mathematics and its Applications 163,
https://doi.org/10.1007/978-1-4939-8636-1_9

the domain can be modified by a smooth deformation of the boundary, whereas the topology optimization methods can also alter the connectivity of the domain by, e.g., introducing new holes.

In contrast to optimal control problems, here the set of admissible controls is a set of subsets of $\mathbb{R}^d$, which does not admit a vector space structure. Nevertheless, we will use the notions of derivatives of the objective functional J with respect to the control variable Ω: On the one hand, the *shape derivative* represents the sensitivity of a domain-dependent functional with respect to a smooth variation of the boundary of this domain whereas, on the other hand, the *topological derivative* is the sensitivity of the functional with respect to a topological perturbation of the set Ω, i.e., with respect to the introduction of a hole in its interior. Starting out from an initial design, these sensitivities can be used to successively update the shape and topology of the control Ω in order to reach an optimal design. In our case, the set Ω is a subset of the computational domain D representing the motor, and its boundary $\partial\Omega$ represents a material interface, e.g., the interface between a ferromagnetic region and an air region of the motor. Both the shape derivative and the topological derivative of this PDE-constrained optimization problem involve the solution to the state equation u and the solution to the adjoint equation p. In a numerical algorithm, these quantities must be computed approximately in each iteration, which is often done by a finite element method. In order to obtain accurate approximations u_h, p_h to the state and adjoint variable, one has to take care of the material interface between the different subdomains. This interface evolves over the iterations of the algorithm and is, in general, not aligned with the underlying finite element mesh.

This book chapter is meant to give an overview over various aspects of topology and shape optimization approaches and many details and proofs are omitted. For more details and more mathematical rigor, we refer the interested reader to [20]. The rest of this chapter is organized as follows: The design optimization problem for the electric motor, which serves as a model problem throughout this chapter, is introduced in Section 2. In Section 3, we give an overview over established topology optimization methods and demonstrate the main steps in the derivation of the topological derivative for the optimization problem at hand, which is constrained by a quasilinear PDE constraint. Section 4 deals with shape optimization and we will derive the shape derivative for our problem. In Section 5, we give an overview over possible ways to treat moving interfaces in the context of finite elements, before combining all of these techniques to an efficient design tool in Section 6, where we will also give numerical optimization results.

2 Problem Description

We consider an interior permanent magnet electric motor as depicted in Figure 1 which consists of a fixed outer part (called the stator) and a rotating inner part (the rotor). The stator contains coils where alternating electric current is induced and

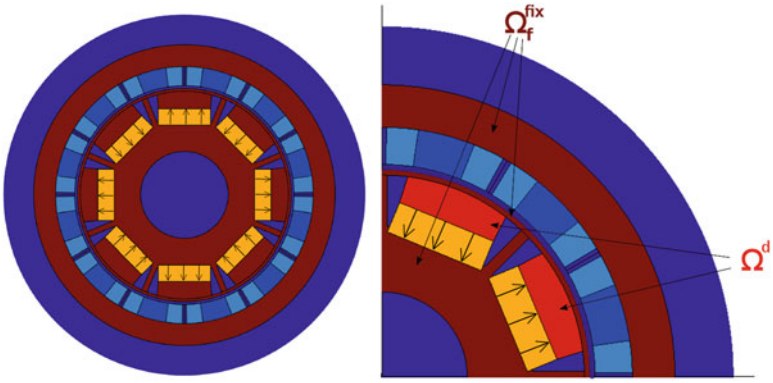

Fig. 1 Left: computational domain D representing electric motor with different subdomains. Right: zoom on upper left quarter (for a different rotor-to-stator constellation) with design region Ω^d and fixed ferromagnetic set Ω_f^{fix}. For a given design $\Omega \subset \Omega^d$, we have $\overline{\Omega_f} = \overline{\Omega_f^{fix}} \cup \overline{\Omega}$

the rotor holds permanent magnets which are magnetized in the directions indicated in Figure 1. Both parts contain a ferromagnetic subdomain and they are separated by a thin air gap. A rotation of the rotor occurs due to the interaction between the magnetic fields produced by magnets and the electric currents in the coils. As it is common for the simulation of electric motors at constant rotation speed, we consider the setting of two-dimensional magnetostatics. There, the magnetic flux density $\mathbf{B}$ is given as $\mathbf{B} = \operatorname{curl}\left((0,0,u)^\top\right)$ where u solves the quasilinear boundary value problem

$$-\operatorname{div}\left(\nu_\Omega(x,|\nabla u|)\nabla u\right) = J_3 - \nu_0 \operatorname{div} M^\perp, \qquad x \in D, \tag{1a}$$

$$u = 0, \qquad x \in \partial D, \tag{1b}$$

on a circular hold-all domain D where J_3 denotes the currents impressed in the coil areas and $M^\perp$ is the perpendicular of the permanent magnetization in the magnets. Here, $\Omega \subset \Omega^d$ denotes the unknown subset of the design region that is occupied with ferromagnetic material, and the magnetic reluctivity ν_Ω is a nonlinear function $\hat{\nu}$ in the ferromagnetic subdomain and a constant ν_0 in the rest of the motor,

$$\nu_\Omega(x,s) = \begin{cases} \hat{\nu}(s), & x \in \Omega_f, \\ \nu_0, & x \in D \setminus \Omega_f. \end{cases} \tag{2}$$

Here, Ω_f consists of the fixed ferromagnetic domain Ω_f^{fix} outside the design region Ω^d and the variable ferromagnetic subset of the design region $\Omega \subset \Omega^d$, i.e., $\overline{\Omega_f} = \overline{\Omega_f^{fix}} \cup \overline{\Omega}$, see Figure 1. Note that, in general, the nonlinear function $\hat{\nu}$ is not known in a closed form but is usually approximated from measured values, see [28].

The magnetic reluctivity ν_Ω, which is the reciprocal of the magnetic permeability μ_Ω, is much larger in the air subdomains compared to the ferromagnetic subdomains of the motor. Therefore, the distribution of ferromagnetic material inside the design area Ω^d (i.e., the shape and topology of the unknown set Ω) influences the magnetic flux density $\mathbf{B}$ in Ω^d and also in the rest of the motor. The magnetic flux density $\mathbf{B}$ inside the air gap between rotor and stator has a big impact on the behavior of the rotation of the motor. The goal of the optimization problem is to identify a subset Ω which minimizes the objective function J that is related to the smoothness of the rotation of the motor. The functional J depends on the magnetic flux density $\mathbf{B}$ in the air gap and will be introduced later in Section 6. The PDE-constrained design optimization problem reads

$$\min_{\Omega} J(u, \Omega) \tag{3a}$$

$$\text{subject to } u \in H_0^1(D) : \int_D \nu_\Omega(x, |\nabla u|)\nabla u \cdot \nabla \eta \, \mathrm{d}x = \langle F, \eta \rangle \quad \forall \eta \in H_0^1(D), \tag{3b}$$

where (3b) is the weak form of boundary value problem (1) with $F \in H^{-1}(D)$ given by

$$\langle F, \eta \rangle = \int_D J_3 \, \eta + \nu_0 \, M^\perp \cdot \nabla \eta \, \mathrm{d}x, \ \eta \in H_0^1(D).$$

Remark 1 In applications of electric motors, the functional J is usually supported only in the air gap between the rotor and the stator. Since the design areas are part of the rotor, we assume that J does not depend on Ω directly, but only via the state variable u, $J = J(u) \neq J(u, \Omega)$. Furthermore, we introduce the reduced functional $\mathscr{J}(\Omega) = J(u(\Omega))$ where $u(\Omega)$ is the solution to (3b) for given Ω.

3 Topology Optimization

In this chapter, we employ a topology optimization algorithm which is based on the topological derivative. Beside this approach, there exist a number of other approaches to topology optimization. We give an overview over the most widely used methods in Section 3.1 before coming to the derivation of the topological derivative for the problem at hand in Section 3.2.

3.1 Overview of Topology Optimization Methods

The concept of topology optimization originates from applications in mechanical engineering but has been applied to a large variety of other applications such as fluid dynamics, acoustics, or electromagnetics. This section is meant to give a brief

overview over the most common methods of topology optimization. For a more detailed discussion of the single approaches, we refer the reader to the review articles [14, 29, 34] and the references therein.

The starting point of numerical topology optimization is widely considered to be the seminal paper by Bendsøe and Kikuchi [9] introducing the homogenization method for topology optimization, followed by the paper [8], where Bendsøe introduced what is now known as the Solid Isotropic Material with Penalization (SIMP) method, giving rise to the large class of density-based methods.

3.1.1 Homogenization Method

The idea of the homogenization method is to represent a domain as a periodic microstructure (usually consisting of rectangular cells like a regular quadrilateral finite element grid) and then to find the optimal layout for each cell. Each of these cells is considered to consist of material and void regions (often a rectangular hole surrounded by solid material) and the dimensions and orientations of these holes are the design variables with respect to which the optimization is performed. Finally, one ends up with a perforated design which can be interpreted as a microstructure. A black-and-white structure can be obtained by setting those cells which are mostly occupied with material to solid, and the other cells to void. The method uses several degrees of freedom for each of the cells, resulting in a large number of degrees of freedom, which is considered a significant drawback of this method. For more details on the homogenization method, we refer the reader to the monograph [1] and the references therein.

3.1.2 Density Methods

In density-based approaches to topology optimization, a design can be represented by a function ρ which takes the value 1 in areas of material and the value 0 in void areas. We remark that, in applications of mechanical engineering, if ρ is 0, the elasticity tensor vanishes and the global stiffness matrix becomes singular. Therefore, it is common practice in density-based topology optimization of mechanical structures to replace the value of 0 by a small, but positive number $\rho_{min} > 0$. The idea of density-based topology optimization approaches is to relax this strict 0–1 nature of the problem by allowing the function ρ to attain any value between 0 and 1. The function ρ is called a density variable. In order to enforce a 0–1 structure of the final design, the idea of [8] is to combine this idea with a penalization of intermediate density values, i.e., to replace the density function ρ in the state equation (and only there) by a penalized version of the density, $\tilde{\rho}(\rho) = \rho^p$ for some $p > 1$. In combination with a constraint on the volume of the arising structure, the algorithm favors the use of "black" and "white" regions, i.e., regions where $\rho = 1$ and $\rho = 0$, respectively, because intermediate values "give very little stiffness at an unreasonable cost" [8]. As remarked in [29], a constraint which limits the volume is

important for this penalizing effect to appear. The method described in [8] together with the choice $\tilde{\rho}(\rho) = \rho^p$ for some $p > 1$ became well known as the SIMP method. We remark that the method is sensitive with respect to the value of p and that good results are usually obtained by using $p = 3$ or by gradually increasing the parameter from $p = 1$ to higher values in the course of the optimization procedure [29].

While the penalization of intermediate density values yields designs with a 0–1 structure, these problems usually lack existence of a solution, a fact which often results in a mesh dependence of the optimized designs. For a detailed survey on the numerical problems resulting from the ill-posedness of such problems, we refer the reader to [30]. The most widely used approach to regularizing these ill-posed problems is by applying a filter to the density variable ρ. This means that one replaces the actual density at a point by an average over the density values in a neighborhood of a certain radius R, called the filter radius. Other approaches include a filtering of the sensitivities, adding a bound on the perimeter of the arising structure or on the gradient of the density variable ρ, see [29, 30].

A more detailed overview of density-based topology optimization methods can be found in, e.g., [10, 29].

3.1.3 Phase-Field Method

The phase-field method for topology optimization is a density-based method using a linear material interpolation, $\tilde{\rho}(\rho) = \rho$. A regularization is achieved by adding a term to the cost functional which approximates the total variation of the density variable. This term is a Cahn-Hilliard type functional, which itself is a weighted sum of two terms. One of these two terms causes a regularizing effect whereas the other term penalizes intermediate density values. We mention that the choices of the weighting factor between these two parts, as well as the weight of the Cahn-Hilliard type functional relative to the objective function, are often crucial for obtaining good results. The phase-field method has been applied to many topology optimization problems, see, e.g., [12, 22].

3.1.4 Level Set Methods

In the level set method [27], a material interface is represented by the zero level set of an evolving function $\psi = \psi(x, t)$ which attains positive values in one subdomain and negative values in the other. The evolution of ψ is given by the solution to the Hamilton-Jacobi equation

$$\frac{\partial}{\partial t}\psi + V \cdot \nabla\psi = 0,$$

where t is a pseudo-time variable and V determines the direction of the evolution. In applications from shape optimization, this vector field is given according to shape sensitivities. For a thorough overview over level set methods, we refer the reader to the review papers [29, 34].

3.1.5 Topological Derivative

The concept of the topological derivative was introduced in [17] as a means to allow for changes of the topology in the course of a classical shape optimization method. The topological derivative of a domain-dependent functional at an interior point of the domain describes its sensitivity with respect to the introduction of a hole around that point. We will deal with the topological derivative in detail in Section 3.2.

3.2 Topological Derivative for Nonlinear Magnetostatics

The topological derivative of a domain-dependent functional $\mathscr{J} = \mathscr{J}(\Omega)$ was introduced in a mathematically rigorous way in [31]. Given a domain Ω, an interior point $x_0 \in \Omega$, and a bounded domain ω which contains the origin, let $\omega_\varepsilon = x_0 + \varepsilon\,\omega$ represent the hole of radius ε around x_0 and let $\Omega_\varepsilon := \Omega \setminus \overline{\omega}_\varepsilon$ denote the perturbed domain. Then, the topological derivative of $\mathscr{J}$ at the point x_0 is defined as the quantity $G(x_0)$ satisfying a topological asymptotic expansion of the form

$$\mathscr{J}(\Omega_\varepsilon) - \mathscr{J}(\Omega) = f(\varepsilon)\, G(x_0) + o(f(\varepsilon)),$$

where f is a positive function which tends to zero with ε, most often $f(\varepsilon) = \varepsilon^d$ with d the space dimension. In many situations such as in the context of electromagnetics, one is not interested in a perturbation of the domain where a hole is excluded from the computational domain, but rather in a local perturbation of a material coefficient. In fact, in the context of magnetostatics, introducing a "hole" in the ferromagnetic subdomain corresponds to the introduction of a different material, namely air. Then, one is interested in an expansion of the form

$$J_\varepsilon(u_\varepsilon) - J_0(u_0) = f(\varepsilon)\, G(x_0) + o(f(\varepsilon)). \tag{4}$$

Here, u_ε is the solution to the state equation where the material coefficient is perturbed within a radius ε around x_0, and u_0 is the solution to the unperturbed state equation. Likewise, J_ε and J_0 denote the objective functional in the perturbed and unperturbed configuration, respectively. We remark that this interpretation is possible in the context of electromagnetics where air is just a different material with a different, positive material coefficient, whereas in mechanical engineering an inclusion of void would lead to a loss of the ellipticity of the perturbed bilinear form.

It can be seen from the expansion (4) that, at points x_0 where $G(x_0) < 0$, for $\varepsilon > 0$ small enough, the objective value for the perturbed configuration is smaller than that in the unperturbed configuration, $J_\varepsilon(u_\varepsilon) - J_0(u_0) < 0$. Thus, in order to minimize a given functional J, it is beneficial to change the material in areas where the topological derivative is negative. Using this information for the iterative introduction of holes at the most favorable positions is one possible topology optimization algorithm using the topological derivative. A different algorithm that is based on the topological derivative is the level set algorithm introduced in [5]. As opposed to the classical level set method for shape optimization, the updates are based on the topological derivative rather than on the shape derivative. Therefore, this algorithm can also nucleate new holes in the interior. More details on the algorithm can be found in [4].

3.2.1 Preliminaries

We show the main steps in the derivation of the topological derivative according to (4) in the context of two-dimensional magnetostatics. We consider a simplified version of the PDE constraint (3b) where, in the unperturbed configuration, the entire computational domain is occupied with ferromagnetic material. Let $F \in H^{-1}(D)$ denote the sources on the right-hand side of the PDE constraint and let Ω^d denote the design subdomain which we assume to be compactly contained in $D \setminus \operatorname{supp}(F)$, i.e., $\Omega^d \subset\subset D \setminus \operatorname{supp}(F)$. Let $x_0 \in \Omega^d$ denote a fixed interior point around which the material coefficient is perturbed. Given a smooth bounded domain ω containing the origin, which represents the shape of the material perturbation, let $\omega_\varepsilon = x_0 + \varepsilon\,\omega$ for small $\varepsilon > 0$. Then, the ferromagnetic subdomain in the perturbed configuration is given by $\Omega_\varepsilon = \Omega \setminus \overline{\omega}_\varepsilon$. For $\varepsilon > 0$ and $W \in \mathbb{R}^2$, we define

$$T(W) = \hat{\nu}(|W|)W \quad \text{and} \quad T_\varepsilon(x, W) = \nu_{\Omega_\varepsilon}(x, |W|)W,$$

where ν_{Ω_ε} is defined according to (2). For the rest of this chapter, we will use $\omega = B(0, 1)$ the unit disk in $\mathbb{R}^2$. For more details about a possible extension of the results to ellipse-shaped inclusions, see [20].

Let $\varepsilon > 0$ small enough such that $\omega_\varepsilon \subset \Omega^d$. Using the notation introduced above, the state equation in the unperturbed and in the perturbed setting read

$$\text{Find } u_0 \in H_0^1(D): \int_D T(\nabla u_0) \cdot \nabla \eta \, \mathrm{d}x = \langle F, \eta \rangle \; \forall \eta \in H_0^1(D), \tag{5}$$

$$\text{Find } u_\varepsilon \in H_0^1(D): \int_D T_\varepsilon(x, \nabla u_\varepsilon) \cdot \nabla \eta \, \mathrm{d}x = \langle F, \eta \rangle \; \forall \eta \in H_0^1(D), \tag{6}$$

respectively. Note that the right-hand sides coincide since we assumed that $x_0 \in \Omega^d \subset\subset D \setminus \operatorname{supp}(F)$. We will be interested in the behavior of the difference between

the solution to these two boundary value problems in terms of ε. By subtracting (5) from (6), we get the boundary value problem defining the variation of the direct state $\tilde{u}_\varepsilon := u_\varepsilon - u_0$,

$$\begin{aligned}\text{Find } \tilde{u}_\varepsilon \in H_0^1(D) : \int_D (T_\varepsilon(x, \nabla u_0 + \nabla \tilde{u}_\varepsilon) - T_\varepsilon(x, \nabla u_0)) \cdot \nabla \eta \, \mathrm{d}x \\ = -\int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|\nabla u_0|)) \nabla u_0 \cdot \nabla \eta \, \mathrm{d}x \quad \forall \eta \in H_0^1(D).\end{aligned} \tag{7}$$

Furthermore, for simplicity, we assume that the objective functional is the same in the perturbed and in the unperturbed configuration, $J_\varepsilon = J_0 = J$. Note that this is satisfied for functionals which are supported only outside the design area like in the case of electric motors, cf. Remark 1. For deriving the topological derivative, we make the following assumption on the objective function:

Assumption 1 *For $\varepsilon > 0$, there exist $\tilde{G} \in H^{-1}(D)$ and $\delta_J \in \mathbb{R}$ such that*

$$J(u_\varepsilon) - J(u_0) = \langle \tilde{G}, u_\varepsilon - u_0 \rangle + \delta_J \, \varepsilon^2 + o(\varepsilon^2). \tag{8}$$

Note that this assumption is satisfied, e.g., for quadratic functionals which are supported only outside the design region.

Moreover, we introduce the following adjoint equations in the unperturbed and perturbed configurations:

$$\begin{aligned}\text{Find } p_0 \in H_0^1(D) : \int_D \mathrm{D}T(\nabla u_0) \nabla p_0 \cdot \nabla \eta \mathrm{d}x = -\langle \tilde{G}, \eta \rangle \ \forall \eta \in H_0^1(D), \\ \text{Find } p_\varepsilon \in H_0^1(D) : \int_D \mathrm{D}T_\varepsilon(x, \nabla u_0) \nabla p_\varepsilon \cdot \nabla \eta \mathrm{d}x = -\langle \tilde{G}, \eta \rangle \ \forall \eta \in H_0^1(D).\end{aligned} \tag{9}$$

Here, $\tilde{G}$ is according to Assumption 1 and $\mathrm{D}T$, $\mathrm{D}T_\varepsilon$ denote the Jacobians of the operators T, T_ε, respectively. Also here, we introduce the difference between the solutions to the two problems above, called the variation of the adjoint state $\tilde{p}_\varepsilon := p_\varepsilon - p_0$, which is the solution to

$$\begin{aligned}\text{Find } \tilde{p}_\varepsilon \in H_0^1(D) : \int_D \mathrm{D}T_\varepsilon(x, \nabla u_0) \nabla \tilde{p}_\varepsilon \cdot \nabla \eta \, \mathrm{d}x \\ = -\int_{\omega_\varepsilon} (\nu_0 \, I - \mathrm{D}T(\nabla u_0)) \nabla p_0 \cdot \nabla \eta \, \mathrm{d}x \quad \forall \eta \in H_0^1(D).\end{aligned} \tag{10}$$

For the rest of this section, we will drop the differential $\mathrm{d}x$ in the volume integrals as there is no danger of confusion.

3.2.2 Derivation of Topological Derivative

By virtue of Assumption 1, choosing $f(\varepsilon) = \varepsilon^2$, it remains to show that there exists $G_0 \in \mathbb{R}$ such that $\langle \tilde{G}, \tilde{u}_\varepsilon \rangle = \varepsilon^2 G_0 + o(\varepsilon^2)$. Testing the perturbed adjoint equation (9) with $\eta = \tilde{u}_\varepsilon$ and exploiting the symmetry of $\mathrm{D}T_\varepsilon$, we get

$$\begin{aligned}
\langle \tilde{G}, \tilde{u}_\varepsilon \rangle &= -\int_D \mathrm{D}T_\varepsilon(x, \nabla u_0)\nabla \tilde{u}_\varepsilon \cdot \nabla p_\varepsilon \\
&= -\int_D \mathrm{D}T_\varepsilon(x, \nabla u_0)\nabla \tilde{u}_\varepsilon \cdot \nabla p_\varepsilon \\
&\quad + \int_D (T_\varepsilon(x, \nabla u_0 + \nabla \tilde{u}_\varepsilon) - T_\varepsilon(x, \nabla u_0)) \cdot \nabla p_\varepsilon \\
&\quad + \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|\nabla u_0|))\nabla u_0 \cdot \nabla p_\varepsilon,
\end{aligned}$$

where we added the left- and right-hand side of (7) tested with $\eta = p_\varepsilon$. For $\varepsilon > 0$, $V, W \in \mathbb{R}^2$, we introduce the operator

$$S_V^\varepsilon(x, W) := T_\varepsilon(x, V + W) - T_\varepsilon(x, V) - \mathrm{D}T_\varepsilon(x, V)W, \tag{11}$$

which characterizes the nonlinearity of the operator T_ε. Then, we get

$$\langle \tilde{G}, \tilde{u}_\varepsilon \rangle = \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|\nabla u_0|))\nabla u_0 \cdot \nabla p_\varepsilon + \int_D S^\varepsilon_{\nabla u_0}(x, \nabla \tilde{u}_\varepsilon) \cdot \nabla p_\varepsilon.$$

Noting that $p_\varepsilon = p_0 + \tilde{p}_\varepsilon$, and defining

$$\begin{aligned}
j_1(\varepsilon) &:= \int_{\omega_\varepsilon} (\nu_0 - \hat{\nu}(|\nabla u_0|))\nabla u_0 \cdot (\nabla p_0 + \nabla \tilde{p}_\varepsilon), \\
j_2(\varepsilon) &:= \int_D S^\varepsilon_{\nabla u_0}(x, \nabla \tilde{u}_\varepsilon) \cdot (\nabla p_0 + \nabla \tilde{p}_\varepsilon),
\end{aligned}$$

we get from Assumption 1 that

$$J_\varepsilon(u_\varepsilon) - J_0(u_0) = j_1(\varepsilon) + j_2(\varepsilon) + \delta_J \varepsilon^2 + o(\varepsilon^2).$$

In view of (4), it remains to show that there exist numbers J_1, J_2 such that

$$j_1(\varepsilon) = \varepsilon^2 J_1 + o(\varepsilon^2), \quad \text{and} \tag{12}$$

$$j_2(\varepsilon) = \varepsilon^2 J_2 + o(\varepsilon^2). \tag{13}$$

Then, the topological derivative is given by $G(x_0) = J_1 + J_2 + \delta_J$.

In what follows, we will first sketch the procedure to obtain the topological derivative in the case of a linear model, i.e., in the case where the nonlinear function $\hat{\nu}$ introduced in (2) is replaced by a constant $\nu_1 < \nu_0$, before discussing the additional difficulties in the case of nonlinear material behavior in the ferromagnetic subdomain.

3.2.3 Linear Case

It can be seen from the definition of the operator S^ε in (11) that, in the linear case, the second term $j_2(\varepsilon)$ vanishes. Thus, we only have to consider the term

$$\begin{aligned} j_1(\varepsilon) &= (\nu_0 - \nu_1) \int_{\omega_\varepsilon} \nabla u_0 \cdot (\nabla p_0 + \nabla \tilde{p}_\varepsilon) \\ &= (\nu_0 - \nu_1) \int_{\omega_\varepsilon} \nabla u_0 \cdot \nabla p_0 - \int_D \nu_\varepsilon \nabla \tilde{u}_\varepsilon \cdot \nabla \tilde{p}_\varepsilon, \end{aligned} \tag{14}$$

where we used (7) with $\eta = \tilde{p}_\varepsilon$ and introduced $\nu_\varepsilon(x) = \chi_{D\setminus\omega_\varepsilon}(x)\,\nu_1 + \chi_{\omega_\varepsilon}(x)\,\nu_0$ with χ_S denoting the characteristic function of a set S.

Assuming enough regularity for the unperturbed direct and adjoint state, it can be seen that, for the first term in (14), we have

$$(\nu_0 - \nu_1) \int_{\omega_\varepsilon} \nabla u_0 \cdot \nabla p_0 = |\omega| \varepsilon^2 \,(\nu_0 - \nu_1)\, \nabla u_0(x_0) \cdot \nabla p_0(x_0) + o(\varepsilon^2) \tag{15}$$

as ε approaches zero.

In order to treat the second term in (14), we define $\tilde{\nu}(x) = \chi_{\mathbb{R}^2\setminus\omega}(x)\nu_1 + \chi_\omega(x)\nu_0$ for $x \in \mathbb{R}^2$, and introduce ε-independent approximations to boundary value problems (7) and (10). After a change of scale, we get the transmission problem defining the variation of the direct state at scale 1,

$$\begin{aligned} &\text{Find } H \in \mathscr{H} \text{ such that} \\ &\int_{\mathbb{R}^2} \tilde{\nu}\, \nabla H \cdot \nabla \eta + (\nu_0 - \nu_1) \int_\omega \nabla u_0(x_0) \cdot \nabla \eta = 0 \quad \forall \eta \in \mathscr{H}, \end{aligned} \tag{16}$$

approximating (7), and the problem defining the variation of the adjoint state at scale 1,

$$\begin{aligned} &\text{Find } K \in \mathscr{H} \text{ such that} \\ &\int_{\mathbb{R}^2} \tilde{\nu}\, \nabla \eta \cdot \nabla K + (\nu_0 - \nu_1) \int_\omega \nabla p_0(x_0) \cdot \nabla \eta = 0 \quad \forall \eta \in \mathscr{H}, \end{aligned} \tag{17}$$

as an approximation of (10), where $\mathscr{H}$ is a suitable weighted Hilbert space over $\mathbb{R}^2$. The solutions H, K are approximations to $\tilde{u}_\varepsilon$ and $\tilde{p}_\varepsilon$, respectively, at scale 1 and it holds

$$\tilde{u}_\varepsilon(x) \approx \varepsilon H(\varepsilon^{-1}x) \quad \text{and} \quad \tilde{p}_\varepsilon(x) \approx \varepsilon K(\varepsilon^{-1}x),$$

for almost every $x \in D$. An important ingredient for deriving an expansion of the form (12) is to show that these ε-independent approximations of $\tilde{u}_\varepsilon$ and $\tilde{p}_\varepsilon$ have a sufficiently fast decay as $|x|$ approaches infinity. This would imply that the impact of the local variation of the material is small "far away" from the inclusion. In the case of a linear state equation, this sufficiently fast decay can be established by convolution of the right-hand side of problems (16) and (17) with the fundamental solution of the Laplace equation. Exploiting these sufficiently fast decays allows us to show that

$$\int_D \nu_\varepsilon \nabla\tilde{u}_\varepsilon \cdot \nabla\tilde{p}_\varepsilon = \varepsilon^2 \int_{\mathbb{R}^2} \tilde{\nu}\, \nabla H \cdot \nabla K + o(\varepsilon^2),$$

which, by means of (16) tested with $\eta = K$, together with the term (15), yields (12) with

$$J_1 = (\nu_0 - \nu_1) \int_\omega \nabla u_0(x_0) \cdot (\nabla K + \nabla p_0(x_0)).$$

It can be seen from (17) that K depends linearly on $\nabla p_0(x_0)$ and, therefore, J_1 can be represented by means of a matrix $\mathscr{M}$. Finally, in the linear case we get

$$\begin{aligned} J_1 &= \nabla u_0(x_0)^\top \mathscr{M}\, \nabla p_0(x_0), \\ J_2 &= 0. \end{aligned}$$

Here, $\mathscr{M} = \nu_1 \mathscr{P}(\omega, \nu_0/\nu_1)$ where the matrix $\mathscr{P}(\omega, \nu_0/\nu_1)$ only depends on the shape of the inclusion ω and the contrast ν_0/ν_1 and is called a polarization matrix, see, e.g., [2]. Explicit formulas for these matrices are available if ω is a disk or ellipse in two space dimensions, or a ball or ellipsoid in three space dimensions, see also [2, 3]. We mention that in the case where ω is the unit disk in $\mathbb{R}^2$, the polarization matrix in the linear setting reads

$$\mathscr{P}_{\omega,\nu_0/\nu_1} = 2\pi \frac{\nu_0/\nu_1 - 1}{\nu_0/\nu_1 + 1} I,$$

where I is the identity matrix. A more detailed derivation of the topological derivative in the linear setting can be found in [3].

3.2.4 Nonlinear Case

In the nonlinear case, the procedure to treat the term $j_1(\varepsilon)$ is similar. However, one big difference is that in the nonlinear setting a sufficiently fast decay of the variation of the direct state H cannot be shown by convolution, but other, more technical tools must be used, see [6, 20]. Furthermore, the term $j_2(\varepsilon)$ does not vanish here and an estimate of the type (13) has to be established. Under certain assumptions on the nonlinearity of the function $\hat{\nu}$, this was done in [20]. If those assumptions on the ferromagnetic material are fulfilled, the result is the following:

Theorem 1 ([20]) *Let $\omega = B(0,1)$. Assume that the functional J fulfills Assumption 1 and that, for the unperturbed direct and adjoint state, it holds $u_0, p_0 \in C^{1,\beta}(D)$ for some $\beta > 0$. For $V, W \in \mathbb{R}^2$, let $\tilde{S}_V(x,W) = \chi_{\mathbb{R}^2\setminus\omega}(x)\,(T(V+W) - T(V) - \mathrm{D}T(V)W)$. Then, the topological derivative of the PDE-constrained optimization problem (3) according to (4) at point $x_0 \in \Omega^d$ reads*

$$G(x_0) = \nabla u_0(x_0)^\top \mathscr{M} \nabla p_0(x_0) + \int_{\mathbb{R}^2} \tilde{S}_{\nabla u_0(x_0)}(x, \nabla H) \cdot (\nabla p_0(x_0) + \nabla K) + \delta_J \tag{18}$$

with $\mathscr{M} = \mathscr{M}(\omega, \mathrm{D}T(\nabla u_0(x_0))) \in \mathbb{R}^{2\times 2}$ and H and K being the variations of the direct and adjoint state at scale 1, respectively.

Remark 2 In order to make use of this formula in numerical optimization algorithms, the following aspects are treated in [20]:

- An explicit formula for the matrix $\mathscr{M}$ is computed. This matrix is related to the concept of polarization matrices.
- The term

$$J_2 = \int_{\mathbb{R}^2} \tilde{S}_{\nabla u_0(x_0)}(x, \nabla H) \cdot (\nabla p_0(x_0) + \nabla K)$$

  seems to be computationally extremely costly since H depends on $\nabla u_0(x_0)$ via (16) and, thus, the (nonlinear) transmission problem (16) defining H would have to be solved for every point x_0 in order to evaluate the term J_2. This problem was overcome by exploiting a rotational invariance property of J_2 with respect to a simultaneous rotation of the quantities $\nabla u_0(x_0)$ and $\nabla p_0(x_0)$. This property allows to precompute a range of typical values of J_2 in a computationally expensive offline stage and to look up the precomputed values during the optimization procedure.
- The formula of Theorem 1 represents the sensitivity of the objective function with respect to the introduction of an inclusion of air around a point x_0. In order to be able to employ bidirectional optimization algorithms which are capable of both removing and reintroducing material at the most favorable positions such as the algorithm introduced in [5], also the topological derivative for the reverse scenario must be computed. We refer to these two topological derivatives as $G^{f\to air}$ and $G^{air\to f}$.

4 Shape Optimization

In contrast to topology optimization, in shape optimization the connectivity of a domain is assumed to be fixed. Here, one is interested in finding the shape of a domain or subdomain which is optimal with respect to a given criterion by means of smooth variations of the boundary or of a material interface. In this section, we are concerned with finding the optimal shape of the ferromagnetic part Ω within the design area Ω^d of the electric motor introduced in Section 2. An essential tool for gradient-based shape optimization is the notion of the sensitivity of a shape functional $\mathscr{J} = \mathscr{J}(\Omega)$ with respect to a smooth perturbation of the boundary of the shape Ω, called the shape derivative. A shape functional $\mathscr{J}$ is said to be shape differentiable if the limit

$$d\mathscr{J}(\Omega; V) = \lim_{t \searrow 0} \frac{\mathscr{J}(\Omega_t) - \mathscr{J}(\Omega)}{t}$$

exists and the mapping $V \mapsto d\mathscr{J}(\Omega; V)$ is linear and continuous with respect to the topology of $C_c^\infty(D, \mathbb{R}^2)$. Here, $\Omega_t = T_t(\Omega)$ denotes the transformed domain under the flow T_t generated by a smooth vector field V.

We mention that there are two ways to define this flow given a smooth vector field V. In the perturbation of identity method, the transformation is given by $T_t(X) = X + t\,V(X)$ for all $X \in \mathbb{R}^d$ and $t \geq 0$, whereas in the velocity or speed method, it is given as $T_t(X) = x(t, X)$ with $x(t, X)$ the solution to the initial value problem

$$\begin{aligned} \frac{d}{dt} x(t, X) &= V(x(t, X)), \quad 0 < t < \tau, \\ x(0, X) &= X, \end{aligned}$$

which, for small $\tau > 0$, has a unique solution, see [16, 32]. Note that, for simplicity, we assumed the vector field V to be autonomous. We remark that both approaches are equivalent for the derivation of first-order shape derivatives but differ by an acceleration term in the case of second-order shape derivatives [16].

4.1 Representation of Shape Derivative

There are basically two ways how one can represent the shape derivative of a functional depending on a domain Ω: either as a distribution on the boundary $\partial\Omega$ which only depends on the normal component of the perturbation, called the Hadamard form, or in a more general volume form, also called the distributed shape derivative. If the shape Ω is regular enough, the Hadamard form can be rewritten as an integral over the boundary,

$$d\mathscr{J}(\Omega; V) = \int_{\partial\Omega} g_\Gamma\, V \cdot n \, \mathrm{d}s, \tag{19}$$

with an integrable function g_Γ. The volume form can be written as

$$d\mathscr{J}(\Omega; V) = \int_\Omega g(V, \mathrm{D}V)\,\mathrm{d}x, \tag{20}$$

for some function g.

One obvious advantage of the boundary-based form (19) is that a descent direction $V = -g_\Gamma n$ is readily available. However, in many situations this choice of V might not be regular enough and has to be regularized. Furthermore, in many numerical procedures for shape optimization, it is not enough to have a descent direction that is only defined on the material interface and it has to be extended to a neighborhood or to the entire computational domain.

On the other hand, in the case where the shape derivative is given in the distributed form (20), the extraction of a descent direction V such that $d\mathscr{J}(\Omega; V) < 0$ can also be achieved easily but requires the solution of an auxiliary boundary value problem of the form

$$\text{Find } V : b(V, W) = -d\mathscr{J}(\Omega; W)\ \forall\, W, \tag{21}$$

where V, W are elements of a suitable function space and $b(\cdot, \cdot)$ is a positive definite bilinear form on the same space. Obviously, a solution V to (21) is a descent direction since $d\mathscr{J}(\Omega; V) = -b(V, V) < 0$. One benefit of the volume form is that it is more general, meaning that for shapes with lower regularity the distributed shape derivative (20) may be well defined whereas the Hadamard form (19) is not. A different aspect favoring the volume-based form (20) is concerned with numerical accuracy of the approximation of the shape derivative when the underlying state and adjoint equations are solved by the finite element method. In [24], the authors show that the finite element approximation to the volume-based form converges quadratically to the "true" shape derivative on the continuous level as the mesh size tends to zero, whereas the boundary-based form converges only linearly.

We mention that, in the case of the Hadamard form of the shape derivative (19), the auxiliary boundary value problem (21) with $b(\cdot, \cdot)$ defined on $\partial\Omega$ can be used to compute a regularized gradient descent velocity for the case where the choice $V = -g_\Gamma n$ is not smooth enough.

A more detailed comparison between these two possible representations can be found in [25].

4.2 Shape Derivative for Nonlinear Magnetostatics

For the reasons mentioned above, we restrict ourselves to the shape derivative in its volume-based representation (20). The rigorous derivation of the shape derivative for the model problem involving the quasilinear PDE of two-dimensional magnetostatics, which was introduced in Section 2, can be found in [21]. There, the shape derivative was computed using the averaged adjoint method introduced in [33].

The shape derivative of the model problem (3) reads

$$\begin{aligned} d\mathscr{J}(\Omega; V) = &- \int_D (J_3 \operatorname{div}(V) + \nabla J_3 \cdot V) p \, \mathrm{d}x - \int_{\Omega_{mag}} \nu_0 \mathbb{P}'(0) \nabla p \cdot M^{\perp} \, \mathrm{d}x \\ &+ \int_D \nu_\Omega(x, |\nabla u|) \mathbb{Q}'(0) \nabla u \cdot \nabla p \, \mathrm{d}x \\ &- \int_{\Omega_f} \frac{\hat{\nu}'(x, |\nabla u|)}{|\nabla u|} (DV^\top \nabla u \cdot \nabla u)(\nabla u \cdot \nabla p) \, \mathrm{d}x, \end{aligned} \tag{22}$$

where $\mathbb{P}'(0) = (\operatorname{div} V) I - \mathrm{D}V^\top$, $\mathbb{Q}'(0) = (\operatorname{div} V) I - \mathrm{D}V^\top - \mathrm{D}V$, $I \in \mathbb{R}^{2\times 2}$ is the identity matrix, and u, $p \in H_0^1(D)$ are the state and adjoint state, respectively.

5 Interface Handling

Both the topological derivative (18) and the shape derivative (22) involve the solution to the state equation (3b) and to the adjoint equation in the current configuration. These two quantities are usually computed approximately by means of the finite element method. In the course of the numerical optimization algorithm, the interface between the ferromagnetic and the air subdomain evolves. In order to get accurate solutions using standard finite element methods, this material interface must be resolved by the underlying mesh. We give an overview over the possible approaches to deal with evolving material interfaces in Section 5.1, before introducing our method in Section 5.2.

5.1 Finite Element Methods for Interface Problems

One way to deal with evolving interfaces in the context of finite elements is to create a new triangulation in each step of the algorithm, which is computationally very costly. Another approach, which is often used in shape optimization, is to start with a mesh that resolves the interface and to advect all nodes of the mesh in the direction of the descent vector field V – provided that V is defined on the whole computational domain. This procedure has the limitation that it does not allow for topological changes and can become problematic when more complex geometries with geometric constraints are involved, as it is the case for our model problem. Here, fixed parts of the electric motor like the circular air gap should not be altered under any circumstances.

The idea of the extended finite element method (XFEM) is to enrich the finite element basis by additional basis functions which are modified versions of the standard basis functions. The solution is seeked in the enriched space $V_h^\Gamma = V_h \oplus V_h^x$

where V_h is a standard finite element space, and V_h^x the space of standard finite element functions which are supported at the interface, multiplied with a so-called enrichment function, see, e.g., [7, 19].

The idea of the immersed finite element method [26] is similar to that of the XFEM. However, rather than adding basis functions to the basis, existing basis functions of the finite element space which are supported across the interface are modified in such a way that the interface jump conditions are satisfied.

In the unfitted Nitsche method introduced in [23], a discontinuity or kink of the solution across an interface is enforced in a weak sense. This way of treating the interface conditions is often used in combination with XFEM, called the Nitsche-XFEM. In this method, just like in all other methods mentioned above, a crucial task is to establish stability of the method with respect to the location of the interface relative to the mesh. Generally, if an element of the underlying unfitted background mesh is cut by the interface very close to one of the vertices, the condition of the system becomes very bad. This issue is treated in the CutFEM [13], which is a stabilized version of the Nitsche-XFEM.

An alternative to these fixed mesh approaches is to modify the mesh and always work with a fitted discretization while still guaranteeing a certain quality of the mesh. We mention the deformable simplicial complex (DSC) method [15].

In [18], an interface finite element method on a fixed mesh is introduced where the interface is resolved by locally modifying the finite element basis functions. Optimal order of convergence and also, when choosing a special hierarchical basis, optimal conditioning of the system matrix are shown. We note that this parametric approach can be equivalently interpreted as a fitted finite element method where some of the mesh nodes close to the interface are moved in such a way that the interface is resolved by the mesh. In the next section, we will follow the approach of [18] and translate it to the case of triangular finite elements.

5.2 A Local Mesh Modification Strategy

We adapt the method presented in [18] for quadrilateral meshes to the case of piecewise linear finite elements on a triangular grid. Our method is based on the assumption that the mesh has a one-level hierarchy, i.e., that always four triangles of the mesh $\mathscr{T}_h$ can be combined to one triangle of a coarser mesh $\mathscr{T}_{2h}$. We will refer to this bigger triangle as a macro triangle and call $\mathscr{T}_{2h}$ the macro mesh. Furthermore, we assume that each element of the macro mesh which is cut by the interface is intersected either in two distinct edges or in one vertex and the opposite edge. Note that this assumption can be enforced by choosing a fine enough macro mesh $\mathscr{T}_{2h}$.

The idea of the method is the following: If a macro element is not cut by the material interface, it is left unchanged. For those macro elements which are cut by the interface in two distinct edges, two of the three vertices lying on the edges of the macro element are moved along these edges to the intersection points of the interface and the macro edge, see Figure 2. If necessary, the vertex lying on the third edge

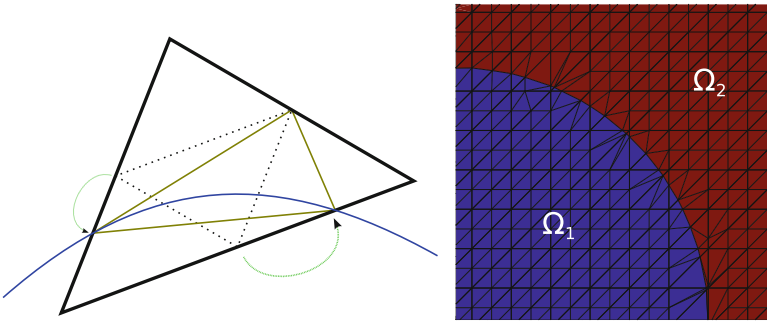

Fig. 2 Left: modification of one macro element that is cut by the material interface (blue). Right: mesh modification strategy for circular material interface

can be moved along that edge in order to avoid some angle to become too close to 180°. The vertices of the macro element remain unchanged. Similarly, if the macro element is cut in one vertex and the opposite edge, the vertex lying on the intersected macro edge is moved to meet the intersection point and the other two vertices may be moved such that a maximum angle condition is satisfied. More details on this procedure can be found in [20] where also the optimal order of convergence of the finite element solution to the true solution in the $L^2(D)$ and in the $H^1(D)$ norm is shown.

6 Numerical Optimization Results

In this section, we combine the results of Sections 3 and 4 and the method introduced in Section 5.2 to one efficient design tool. We describe the procedure in Section 6.1 before applying it to the model problem introduced in Section 2 in Section 6.2.

6.1 Combined Topology and Shape Optimization with Interface Handling

We present a two-stage algorithm where topology optimization is performed in the first stage in order to find the optimal connectivity of the design domain, followed by shape optimization in combination with the interface resolution method of Section 5.2 as a post-processing in order to obtain smoother designs.

In the first stage, topology optimization is performed using the level set algorithm [5]. In order to apply this algorithm, it is important to have the topological

derivative for both directions, i.e., the sensitivity of the objective function with respect to the nucleation of a hole of air inside ferromagnetic material, $G^{f\to air}$, and the sensitivity for the creation of ferromagnetic material in an air region, $G^{air\to f}$.

The shape optimization is done by means of a gradient descent algorithm. Starting out from an initial design, the interface between the ferromagnetic and the air subdomain of the design area Ω^d is moved a certain distance in a descent direction V which was obtained from (21). The step size is chosen in such a way that a decrease of the objective functional is achieved. Note that, for the evaluation of the shape derivative on the right-hand side of (21), the state and adjoint equations have to be solved, which is done by the finite element method. In order to obtain accurate finite element solutions, the mesh modification strategy of Section 5.2 should be applied whenever the interface is updated.

The proposed optimization procedure is summarized in the following algorithm:

Algorithm 1 *(Combined topology and shape optimization with interface handling)*

Stage I: *Apply the algorithm [5] to find an optimal topology.*
Stage II: *Use the final design of Stage I as an initial design and perform gradient-based shape optimization where for each solve of the state and adjoint equations, the local mesh adaptation strategy of Section 5.2 is applied.*

A more detailed description of the algorithm can be found in [20].

6.2 Minimizing Total Harmonic Distortion

The goal of the model problem introduced in Section 2 was to achieve a smooth rotation of the rotor. This can be achieved by ensuring a smooth radial component of the magnetic flux density $B_r = \mathbf{B}\cdot n = \nabla u \cdot \tau$ in the air gap between the rotor and the stator when the electric current is switched off ($J_3 = 0$). Here, n and τ denote the unit normal and tangential vectors on a circular path in the air gap, respectively. For that purpose, we consider B_r along this circular curve inside the air gap as a periodic signal and decompose it into its Fourier coefficients,

$$B_r(u)(\varphi) = \sum_{k=1}^{\infty} A_k \sin(\omega\, k\, \varphi) + B_k \cos(\omega\, k\, \varphi), \tag{23}$$

where $A_k, B_k \in \mathbb{R}$, $\varphi \in [0, 2\pi]$, and ω denotes the number of pole pairs of the motor. In the motor introduced in Section 2, we have eight magnetic poles, thus $\omega = 4$. Due to the geometry of the motor, the coefficients A_k are approximately zero and will be neglected. The total harmonic distortion (THD) measures the contributions of

higher harmonics (i.e., $k > 1$) to the total signal, see [11]. For practical purposes, we only consider the first $N = 20$ harmonics. Then, the total harmonic distortion of B_r reads

$$THD(B_r) = \sqrt{\frac{\sum_{k=2}^{N} B_k^2}{\sum_{k=1}^{N} B_k^2}},$$

where the coefficients B_k are according to (23). The minimization of the THD filters out all higher harmonics. In order to make sure that the first harmonic does not become too small, we minimize the functional

$$J(u) = \frac{THD(B_r(u))^2}{B_1(B_r(u))},$$

where $B_1(B_r(u))$ denotes the coefficient B_1 in (23). In our implementation, we computed the Fourier coefficients by a least square approach.

Figure 3 shows the evolution of the design by using Algorithm 1 starting from an initial design. The final design of Stage I obtained after a total of 47 iterations is approximated by an explicit polygonal interface, which serves as an initial guess for the shape optimization. The final design after the shape optimization procedure together with the local mesh modification strategy introduced in Section 5.2 can be seen in the bottom row of Figure 3. Figure 4 shows the curve B_r for the initial and the final design of both stages of the optimization procedure, and Figure 5 the final design together with the magnetic field.

7 Conclusion

This book chapter was motivated by a concrete application from electrical engineering, the design optimization of an electric motor. We addressed the problem by a two-stage algorithm. In the first stage, we used a topology optimization approach which is based on the mathematical concept of the topological derivative. Here, the derivation and efficient implementation of the topological derivative for the optimization problem at hand, which is constrained by a nonlinear PDE, turned out to be particularly challenging. The second stage of the global algorithm is a shape optimization algorithm where we additionally took care to accurately resolve the evolving material interfaces by means of a mesh modification strategy. Finally, we showed numerical results obtained by applying the introduced algorithm to find a motor design which exhibits very smooth rotation properties.

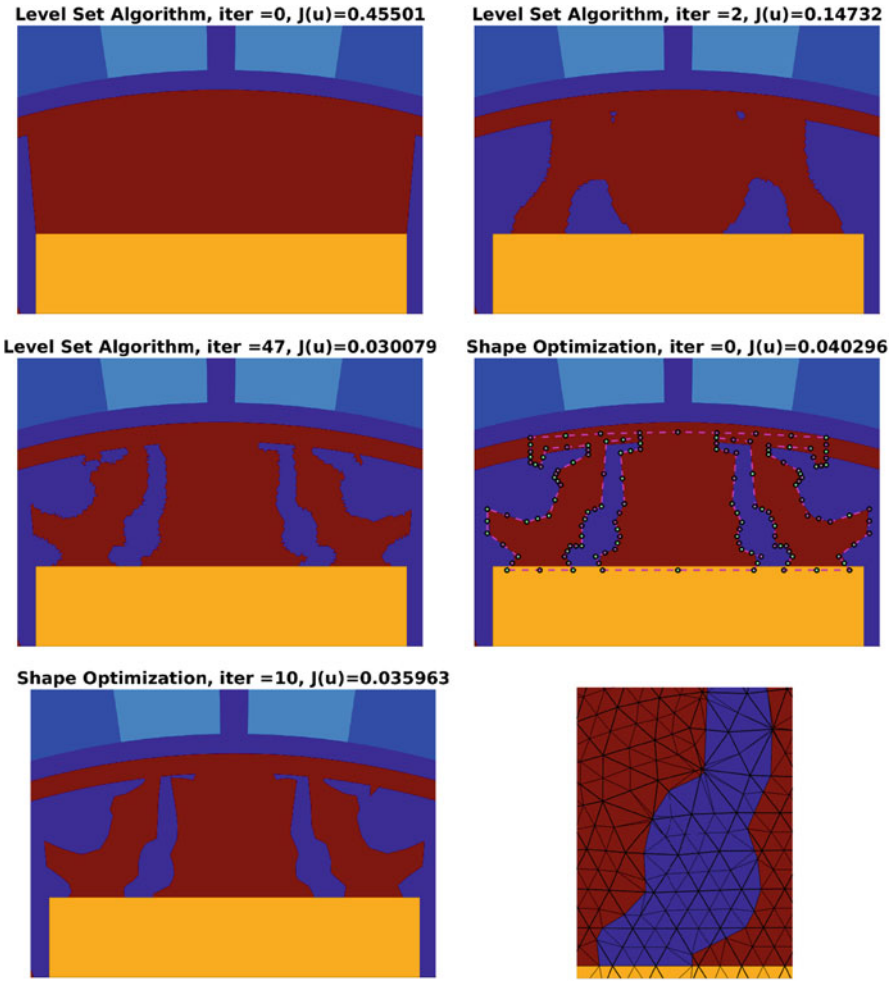

Fig. 3 Top left: initial design. Top right: design after two iterations of topology optimization by algorithm [5]. Center left: final design of topology optimization after 47 iterations. Center right: initial design for shape optimization by approximation of topology optimization result. Bottom left: final design of shape optimization with mesh adaptation strategy after 10 iterations. Bottom right: zoom on modified mesh

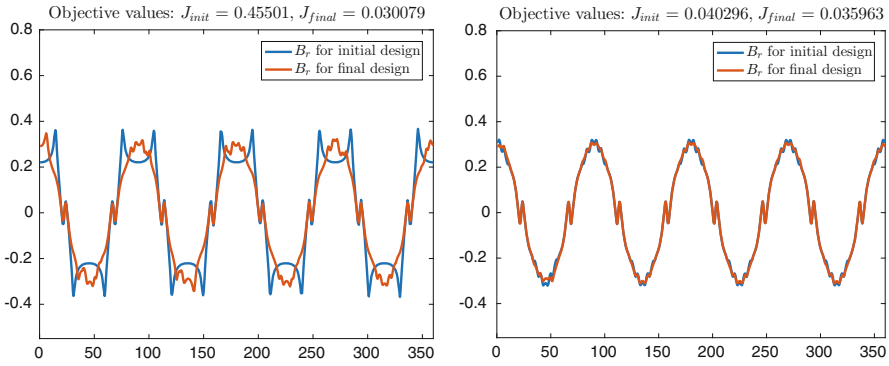

Fig. 4 Radial component of magnetic flux density along the air gap for initial and final designs. Left: Stage I (topology optimization). Right: Stage II (shape optimization)

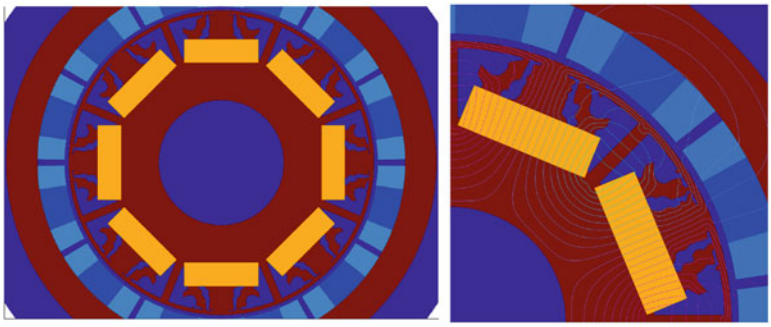

Fig. 5 Final designs after Stage II together with magnetic field lines

References

1. G. Allaire. *Shape optimization by the homogenization method*. Applied mathematical sciences. Springer, New York, 2002.
2. H. Ammari and H. Kang. *Polarization and Moment Tensors*. Springer-Verlag New York, 2007.
3. S. Amstutz. Sensitivity analysis with respect to a local perturbation of the material property. *Asymptotic analysis*, 49(1), 2006.
4. S. Amstutz. Analysis of a level set method for topology optimization. *Optimization Methods and Software - Advances in Shape an Topology Optimization: Theory, Numerics and New Application Areas*, 26(4–5):555–573, 2011.
5. S. Amstutz and H. Andrä. A new algorithm for topology optimization using a level-set method. *Journal of Computational Physics*, 216(2):573–588, 2006.
6. S. Amstutz and A. Bonnafé. Topological derivatives for a class of quasilinear elliptic equations. *Journal de mathématiques pures et appliquées*.
7. T. Belytschko, R. Gracie, and G. Ventura. A review of extended/generalized finite element methods for material modeling. *Model. Simul. Mater. Sci. Eng.*, 17(4), 2009.
8. M. P. Bendsøe. Optimal shape design as a material distribution problem. *Structural Optimization*, 1(4):193–202, 1989.

9. M. P. Bendsoe and N. Kikuchi. Generating optimal topologies in structural design using a homogenization method. *Comput. Methods Appl. Mech. Eng.*, 71(2):197–224, Nov. 1988.
10. M. P. Bendsøe and O. Sigmund. *Topology Optimization: Theory, Methods and Applications*. Springer, Berlin, 2003.
11. A. Binder. *Elektrische Maschinen und Antriebe: Grundlagen, Betriebsverhalten*. Springer-Lehrbuch. Springer, 2012.
12. M. Burger and R. Stainko. Phase-field relaxation of topology optimization with local stress constraints. *SIAM J. Control Optim.*, 45(4):1447–1466, 2006.
13. E. Burman, S. Claus, P. Hansbo, M. G. Larson, and A. Massing. CutFEM: Discretizing geometry and partial differential equations. *International Journal for Numerical Methods in Engineering*, 104(7):472–501, 2015.
14. F. Campelo, J. Ramırez, and H. Igarashi. A survey of topology optimization in electromagnetics: considerations and current trends. 2010.
15. A. N. Christiansen, M. Nobel-Jørgensen, N. Aage, O. Sigmund, and J. A. Bærentzen. Topology optimization using an explicit interface representation. *Structural and Multidisciplinary Optimization*, 49(3):387–399, 2014.
16. M. C. Delfour and J.-P. Zolésio. *Shapes and geometries*, volume 22 of *Advances in Design and Control*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 2011. Metrics, analysis, differential calculus, and optimization.
17. H. A. Eschenauer, V. V. Kobelev, and A. Schumacher. Bubble method for topology and shape optimization of structures. *Structural optimization*, 8(1):42–51, 1994.
18. S. Frei and T. Richter. A locally modified parametric finite element method for interface problems. *SIAM J. Numer. Anal.*, 52(5):2315–2334, 2014.
19. T.-P. Fries and T. Belytschko. The extended/generalized finite element method: An overview of the method and its applications. *Int. J. Numer. Meth. Eng.*, 84(3):253–304, 2010.
20. P. Gangl. *Sensitivity-based topology and shape optimization with application to electrical machines*. PhD thesis, Johannes Kepler University Linz, 2016.
21. P. Gangl, U. Langer, A. Laurain, H. Meftahi, and K. Sturm. Shape optimization of an electric motor subject to nonlinear magnetostatics. *SIAM Journal on Scientific Computing*, 37(6):B1002–B1025, 2015.
22. H. Garcke, C. Hecht, M. Hinze, and C. Kahle. Numerical approximation of phase field based shape and topology optimization for fluids. *SIAM Journal on Scientific Computing*, 37(4):A1846–A1871, 2015.
23. A. Hansbo and P. Hansbo. An unfitted finite element method, based on Nitsche's method, for elliptic interface problems. *Computer Methods in Applied Mechanics and Engineering*, 191(47–48):5537 – 5552, 2002.
24. R. Hiptmair, A. Paganini, and S. Sargheini. Comparison of approximate shape gradients. *BIT*, 55(2):459–485, 2015.
25. A. Laurain and K. Sturm. Distributed shape derivative via averaged adjoint method and applications. *ESAIM: M2AN*, 50(4):1241–1267, 2016.
26. Z. Li. The immersed interface method using a finite element formulation. *Appl. Num. Math.*, 27:253–267, 1998.
27. S. Osher and J. A. Sethian. Fronts propagating with curvature dependent speed: Algorithms based on Hamilton-Jacobi formulations. *Journal of Computational Physics*, 79(1):12–49, 1988.
28. C. Pechstein and B. Jüttler. Monotonicity-preserving interproximation of B-H-curves. *J. Comp. App. Math.*, 196:45–57, 2006.
29. O. Sigmund and K. Maute. Topology optimization approaches: A comparative review. *Structural and Multidisciplinary Optimization*, 48(6):1031–1055, 2013.
30. O. Sigmund and J. Petersson. Numerical instabilities in topology optimization: A survey on procedures dealing with checkerboards, mesh-dependencies and local minima. *Structural Optimization*, 16(1):68–75, 1998.
31. J. Sokołowski and A. Zochowski. On the topological derivative in shape optimization. *SIAM Journal on Control and Optimization*, 37(4):1251–1272, 1999.

32. J. Sokołowski and J.-P. Zolésio. *Introduction to shape optimization*, volume 16 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1992. Shape sensitivity analysis.
33. K. Sturm. Minimax Lagrangian approach to the differentiability of nonlinear PDE constrained shape functions without saddle point assumption. *SIAM Journal on Control and Optimization*, 53(4):2017–2039, 2015.
34. N. P. van Dijk, K. Maute, M. Langelaar, and F. van Keulen. Level-set methods for structural topology optimization: a review. *Structural and Multidisciplinary Optimization*, 48(3): 437–472, 2013.

Distributed Parameter Estimation for the Time-Dependent Radiative Transfer Equation

Oliver Dorn

Abstract The time-dependent radiative transfer equation (RTE), also called a linear transport equation, is an integro-differential equation that is frequently used for modeling transport processes of "particles" (or "wave packages") through a medium where the trajectories of the particles are affected by random absorption and scattering processes at any location inside the medium. The probabilities that such absorption and scattering events take place at a given location are quantified by two parameters of the RTE, namely the absorption and scattering cross section, which are usually unknown in many practical applications and need to be estimated as space-dependent functions from the same set of measurements or observations. This is the so-called *inverse problem* or *distributed parameter estimation problem* for the RTE, which has close links to the field of PDE-constrained optimization. Due to the complex structure and high dimensionality of the RTE, often PDE-approximations to the RTE are employed for obtaining these estimates where however the two distributed parameters of the RTE appear in transformed form. In this contribution we describe some typical practical approaches for solving such distributed parameter estimation problems for the time-dependent RTE including some of its approximations. We highlight some difficulties encountered in the simultaneous reconstruction of two independent distributed parameters from the same data set.

1 Introduction

In 1905, Schuster [72] proposed a radiative transfer model for describing the propagation of light through a foggy atmosphere. Since then the Radiative Transfer Equation (RTE), often also called a linear transport equation or radiative transport

O. Dorn (✉)
School of Mathematics, The University of Manchester, Alan Turing Building, Oxford Rd, Manchester M13 9PL, UK
e-mail: oliver.dorn@manchester.ac.uk

H. Antil et al. (eds.), *Frontiers in PDE-Constrained Optimization*, The IMA Volumes in Mathematics and its Applications 163,
https://doi.org/10.1007/978-1-4939-8636-1_10

equation, has been applied to a wide range of situations where energy is propagating through a scattering environment. Examples are the propagation of light through the atmospheres of remote stars [17, 74], the propagation of neutrons inside a nuclear reactor [9, 24, 32, 79, 80], the coda-analysis in seismology [1, 69], the simulation of special light effects in computer graphics [41], the radiation therapy planning problem [11], and the imaging of optical properties inside the human body with near-infrared light in Diffuse Optical Tomography (DOT) [5, 6]. The RTE model is extremely versatile and powerful and is capable of describing a large variety of physical effects. Accordingly, it is also quite complex and difficult to implement numerically for studying such physical effects. It contains a certain number of physical parameters related to absorption and scattering of energy packages when propagating in the environment that need to be known a priori when doing this modeling. Often, these physical parameters, defined over the entire domain, can only be inferred indirectly from few observations at discrete locations that are accessible by measurement equipment. This process of inferring medium properties in the entire domain from few indirect observations constitutes a so-called *inverse problem* or *distributed parameter estimation problem* which is in many cases under-determined and *ill-posed*. Adding any form of a priori information into this estimation process helps to stabilize the solution of the inverse problem and is called *regularization*. We will concentrate here mainly on one particular inverse problem which arises in DOT, with a special emphasis on the simultaneous estimation of the absorption and the scattering cross section from the same data set. We will, however, try to emphasize throughout the text that many of the techniques and strategies developed in this particular application will directly apply also to many of the other above mentioned fields where the RTE is applicable.

The remainder of this chapter is organized as follows. In Section 2, we outline some general properties of the RTE and formulate the corresponding inverse problem considered in this text. In Section 3, we formulate then the particular RTE problem encountered in the specific application of DOT, and derive several approximations to that model that can be used (and in practice often *are* used) for obtaining estimates of the unknown model parameters. In particular, we outline in this section how the parameters transform into derived parameters governing the inverse problems of these approximations. Section 4 presents some 2-D Monte Carlo simulations visualizing the propagation of particle populations in some typical regimes of the RTE and thereby illustrating the validity of some of these approximations. In Section 5, we briefly outline the derivation of the gradients corresponding to the various models introduced in Section 3. Section 6 then briefly mentions some alternative regularization and nonstandard optimization techniques for the DOT inverse problem that have been addressed in the literature using this RTE model. Section 7 finally summarizes and provides some outlook on possible future research and open problems. We conclude by providing a list of references for further reading, which is by far not intended to be complete. However, this list should contain sufficient material to provide the reader with a useful starting point for further investigations.

2 General Results for the RTE

2.1 The Cauchy Problem of the RTE

We present first the RTE in a sufficiently general form which applies to a large number of applications outlined above. In the time-dependent case, the RTE can be studied conveniently using concepts of semi-group theory [22, 62]. We will outline some general results in the following.

Let $u(x, \zeta, t)$ denote the density of particles (photons, phonons, neutrons ...) at location $x \in \Omega \subset \mathbb{R}^3$ propagating with velocity $\zeta \in V \subset \mathbb{R}^3$ at time $t \in [0, \tau[$ in a compact domain Ω. The *Cauchy Problem for the RTE* is described by the evolution equation

$$\frac{\partial u}{\partial t}(x, \zeta, t) + \zeta \cdot \nabla u(x, \zeta, t) + a(x, \zeta) u(x, \zeta, t) \tag{1}$$

$$- \int_V f(x, \zeta', \zeta) u(x, \zeta', t) d\zeta' = q(x, \zeta, t) \quad \text{in} \quad \Omega \times V \times]0, \tau[$$

with initial condition

$$u(x, \zeta, 0) = u_0(x, \zeta) \quad \text{in} \quad \Omega \times V \tag{2}$$

and boundary condition

$$u(x, \zeta, t) = g(x, \zeta, t) \quad \text{at} \quad \Gamma_-. \tag{3}$$

The set $\Gamma_\pm$ is defined as

$$\Gamma_\pm := \{ (x, \zeta, t) \in \partial\Omega \times V \times]0, \tau[\,, \quad \pm \nu(x) \cdot \zeta > 0 \}, \tag{4}$$

where $\nu(x)$ is the outer normal to $\partial\Omega$ at boundary point x. Similarly we obtain Γ_0 by replacing '$>$' by '$=$'. Therefore we have

$$\partial\Omega \times V \times]0, \tau[=: \Gamma = \Gamma_- \cup \Gamma_+ \cup \Gamma_0.$$

Often we write $\zeta = \upsilon\theta$ with $\upsilon = |\zeta|$ and $\theta = \zeta/|\zeta| \in S^2$ with S^2 being the unit sphere in $\mathbb{R}^3$. This has the interpretation that the particle propagates with speed υ in direction θ. In the special case (usually assumed in DOT where scattering events do not change the energy of the photons) that the speed is given and fixed, we also write $|\zeta| =: c =$const instead of υ. In this case

$$V = S_c := \{ \zeta \in \mathbb{R}^3 \,, \, |\zeta| = c = \text{const} > 0 \},$$

which is the sphere of radius c centered in the origin. However, when energy is lost in scattering events the velocity is usually not constant such that we can use instead

$$V = B_\rho := \{\, \zeta \in \mathbb{R}^3 \,,\ |\zeta| \leq \rho \,\}$$

where ρ indicates a maximally permitted velocity (energy) in the system. In some applications also, the situation can occur that energy or velocity is increased in a scattering event. We will not consider such situations here.

The above defined Cauchy problem can be analyzed theoretically using L^p spaces, $p \in [1, \infty[$, for the photon density u [23]. More precisely, we introduce the Sobolev spaces

$$\tilde{W}_p := \Big\{ u \in L^p(\Omega \times V \times]0,\tau[),\ \frac{\partial u}{\partial t} + \zeta \cdot \nabla u \in L^p(\Omega \times V \times]0,\tau[),$$
$$u(.\,,.\,,0) \in L^p(\Omega \times V), \quad u|_{\Gamma_-} \in L^p(\Gamma_-,\ |\zeta \cdot \nu|\, d\gamma d\mu dt) \Big\}$$

where $d\gamma$ is a surface measure on $\partial\Omega$. Using these spaces, existence and uniqueness of the solution of the Cauchy problem can be shown using semi-group theory under certain assumptions on the physical parameters involved. For example, in the classical theory the following assumptions are listed [23]:

$$a(x,\zeta) \in L^\infty(\Omega \times V) \quad \text{with} \quad a(x,\zeta) \geq 0,$$

and $f(x,\zeta',\zeta)$ in (1) is a positive function with

$$\int_V f(x,\zeta',\zeta)\, d\mu(\zeta) \leq M_a \quad \forall (x,\zeta') \in \Omega \times V,$$
$$\int_V f(x,\zeta',\zeta)\, d\mu(\zeta') \leq M_b \quad \forall (x,\zeta) \in \Omega \times V,$$

with positive constants M_a, M_b. The source q and initial particle distribution u_0 are assumed to satisfy

$$q \in L^p(\Omega \times V \times]0,\tau[) \ \text{ and } \ u_0 \in L^p(\Omega \times V)$$

for $p \in [1, \infty[$. Existence and uniqueness of the solution (in the weak sense) of the forward problem are then shown in [23]. Furthermore it is shown that under certain conditions a strong solution exists. Choosing $p = 1$ is the most natural choice since the 1-norm of u physically represents the total number of photons contained in the domain at a given time. However, when solving the underlying inverse problem often $p = 2$ is preferred since it imposes a convenient Hilbert space structure on the functions spaces. For more details, we refer to [23].

2.2 Physical Interpretation of the RTE in Applications

The time-dependent RTE stated above models the *propagation* of particles along straight paths ("rays") where also *scattering* is involved. A slight modification of it can also capture propagation along curved paths [46, 54, 66, 69], which however will not be considered here. The RTE is a balance equation which looks at an infinitesimal volume and describes the density of particles $u(x, \zeta, t)$ as they pass at time t through such a volume centered at position x. "Particles" can be any countable quantity such as photons, phonons, neutrons, energy wave packages, or vector descriptions of those, where any phase information is neglected. In its continuous setting, though, it describes real-valued (usually positive) densities of particles instead of an integer number of particles. The RTE model is used in a large variety of applications such as seismology, biomedical optics, radar, neutron physics, computer graphics, and others. Particles can have an "energy" which is represented by the magnitude υ (sometimes called "speed") of the "velocity" vector ζ. By introducing the normalized direction vector θ, we can then write $\zeta = \upsilon\theta$ as already mentioned above. In a 3-D environment, both the position x and the velocity ζ are three-dimensional, whereas time is one-dimensional, such that particle propagation is modeled in a seven-dimensional phase space.

Scattering of particles can take place at any location in the domain with position-dependent probabilities. Scattering is only considered as interaction with the internal material of the volume but not with other particles, which renders the RTE being linear. Therefore, the RTE is inappropriate for modeling the propagation of charged particles or the modeling in gas dynamics, where the interactions with other particles introduce nonlinearities in the model [16, 39, 40, 76]. The specific rules of what happens in a given scattering event are application-dependent, but in principle a particle will change the direction of propagation θ as well as energy (speed υ) in such an event. At every location in the domain, particles are entering in an infinitesimal volume centered at that location from all possible directions according to the incoming flux density, can be scattered inside the volume or alternatively pass through the volume without being affected at all by the internal structure, and then leave the volume after an infinitesimal time step into directions specified by these interaction laws. Particles can usually also be "absorbed" inside the infinitesimal volume, which means they disappear completely, such that particle conservation might not be granted. On the other hand, source terms can be contained inside this volume which produce new particles also contributing to the outflux from this volume. By combining the contribution of all these physical events, a balance equation can be formulated which essentially is the time-dependent RTE. We have described its mathematical form above in (1)–(3) and will discuss specific examples further below.

Notice that some applications actually use time-harmonic source terms in time-independent background media, which transforms the time-domain RTE into a frequency-domain or time-harmonic RTE. Also, in some applications, a continuously working source term is employed, such that after some short transition period

the time-dependence might disappear and the process becomes stationary. This is then described better by the stationary RTE. In the following, we will focus mainly on time-dependent phenomena where the full-time-dependent RTE is necessary for the modeling. Nevertheless, historically the stationary and time-harmonic RTE have often been the methods of choice in many applications since the dimensionality of the RTE in these situations is reduced. Some background on these alternative models can be found for example in [5, 7, 57].

2.3 Integral Form of the RTE and Neumann Series Solution

Existence and uniqueness of the solutions of the RTE can also be addressed by using its integral formulation instead of resorting to semi-group theory. This formulation also leads directly to a number of numerical approaches for solving the RTE, and provides valuable physical insight. We follow here [15].

By integrating (1)–(3) along characteristics we arrive at the following integral representation of the RTE

$$u(x,\zeta,t) = Q(x,\zeta,t) + \int_0^t \int_V f[x-\zeta(t-t'),\zeta',\zeta]\cdot \tag{5}$$

$$\cdot\, u(x-\zeta(t-t'),\zeta,t')\,\exp\left\{-\int_{t'}^{t} a[x-\zeta(t-t''),\zeta]\,dt''\right\} d\mu(\zeta')dt'$$

with

$$Q(x,\zeta,t) = u(x-\zeta t,\zeta,0)\,\exp\left\{-\int_0^t a[x-\zeta(t-t'),\zeta]\,dt'\right\} \tag{6}$$

$$+\int_0^t q(x-\zeta(t-t'),\zeta,t')\,\exp\left\{-\int_{t'}^{t} a[x-\zeta(t-t''),\zeta]\,dt''\right\} dt'.$$

Any boundary condition (3) is represented here by an equivalent surface source [15]

$$q_s(x_s,\zeta,t) := |\zeta\cdot\nu(x_s)|\,g(x_s,\zeta,t)\,\delta(z_\nu(x_s)), \tag{7}$$

where z_ν is a local coordinate which is perpendicular to $\partial\Omega$ in $x_s \in \partial\Omega$ and δ denotes the one-dimensional Dirac delta distribution. We assume that any such $q_s(x_s,\zeta,t)$ is already contained in $q(x,\zeta,t)$ of (6). Equation (5) can be written in operator form as

$$u(x,\zeta,t) = Q(x,\zeta,t) + Ku(x,\zeta,t), \tag{8}$$

where K is given by the integral expression in (5). Mathematically, (5) or (8) are examples for Volterra integral equations of the second kind. Such equations are well known and can be solved formally by expansion in a Neumann series [15, 32]

$$u(x,\zeta,t) = \sum_{k=0}^{\infty} u_k(x,\zeta,t), \tag{9}$$

$$\text{with} \quad u_0(x,\zeta,t) := Q(x,\zeta,t), \tag{10}$$

$$\text{and} \quad u_k(x,\zeta,t) := K u_{k-1}(x,\zeta,t) \quad \text{for } k \geq 1. \tag{11}$$

Certainly, criteria need to be established for when this series converges [15, 26, 63]. Physically, $u_k(x,\zeta,t)$ represents the part of the flux which has been "scattered" exactly k times.

2.4 Distributed Parameter Estimation and the Inverse Problem

The above RTE model contains two distributed parameters $m_1 = a$ and $m_2 = f$ which are indicators of the medium properties in which the particles propagate. In many applications, these medium properties are unknown and are the main focus of interest. In the *distributed parameter estimation problem* or *inverse problem* of the RTE data $\mathbf{G}$ (consisting for example of influx–outflux pairs related by the *albedo* operator $\mathscr{A}$ to be defined further below) are gathered at locations accessible to measurement equipment and then the mathematical goal is to find m_1 and m_2 as space-dependent functions that honor these data. Depending on the physical and mathematical setup, it might be possible or not to reconstruct uniquely both parameters simultaneously from the measured data. We first treat in this section the case where uniqueness results for the inverse problem are available that come with a constructive procedure for estimating those parameters. In later sections, we will then outline some situations where explicit formulas for the unique recovery of two distributed parameters are not known so far (or might not exist at all) and where instead optimization approaches are employed for estimating those parameters (possibly under some uncertainty) from the given data.

The theoretical analysis of inverse problems often starts with defining some kind of "scattering operator" following general concepts introduced in [65]. This operator maps incoming to outgoing radiation for a given bounded volume Ω and depends on the parameter distribution inside this volume. In the framework of linear transport theory usually, the related "albedo operator" $\mathscr{A}$ is used instead which is given as

$$\mathscr{A}[a,f] : u_{|\Gamma_-} \mapsto u_{|\Gamma_+} \tag{12}$$

with $\Gamma_\pm$ defined in (4). This operator is mapping incoming flux to outgoing flux through $\partial\Omega$, and its dependence on the internal parameters a and f is indicated

by the notation $\mathscr{A}[a, f]$. Using appropriate function spaces for $u_{|\Gamma_-}$, we can "probe" the medium by applying a variety of incoming fluxes and measuring the corresponding outgoing fluxes $u_{|\Gamma_+}$.

The *albedo-operator* can be defined in different settings, whose technical differences are not of much importance for our general introduction. We refer the interested reader to [3, 7, 19, 33] and follow in the following mainly [3, 33] neglecting some of the technical details in our brief exposition.

Some physical restrictions on the variables and parameters are necessary for obtaining a well-defined albedo operator. Let V be an open set in $\mathbb{R}^3$ and let Ω be a convex set in $\mathbb{R}^3$ with $\mathscr{C}^1$-regular boundary. Adopting terminology from [65] we call a pair (a, f) *admissible and regular* if the following four conditions are satisfied:

$$
\begin{aligned}
&(i)\ \ 0 \le a \in L^\infty(\mathbb{R}^3 \times V);\\
&(ii)\ \ 0 \le f(x, \zeta^{'}, .) \in L^1(V) \quad \text{for a.e. } (x, \zeta^{'}) \in \mathbb{R}^3 \times V;\\
&(iii)\ \sigma_p(x, \zeta^{'}) := \int_V f(x, \zeta^{'}, \zeta)\, d\zeta \in L^\infty(\mathbb{R}^3 \times V).\\
&(iv)\ \ a \text{ and } f \text{ vanish for } x \notin \Omega
\end{aligned}
$$

Let similar to before $\tilde{\Gamma}_\pm$ be

$$
\tilde{\Gamma}_\pm = \{\, (x, \zeta) \in \partial\Omega \times V\ , \quad \pm\nu(x) \cdot \zeta > 0 \,\}. \tag{13}
$$

On $\tilde{\Gamma}_\pm$ we have the measure $d\xi = |\nu(x) \cdot \zeta| d\mu(x) d\zeta$ where $d\mu(x)$ is the corresponding measure on $\partial\Omega$. Let $u(x, \zeta, t)$ be the solution of the evolution problem

$$
\Big(\frac{\partial}{\partial t} - T\Big)u = 0 \quad \text{in } [0, \infty[\times\Omega \times V, \tag{14}
$$

$$
u|_{]0,\infty[\times\tilde{\Gamma}_-} = g, \tag{15}
$$

$$
u|_{t=0} = 0, \tag{16}
$$

where $T := -\zeta \cdot \nabla - a(x, \zeta) + \int_V f(x, \zeta^{'}, \zeta)\, .\, d\zeta^{'}$ is treated as an operator on $\Omega \times V$. Then the albedo operator can be defined by

$$
\mathscr{A} g = u_{|[0,\infty[\times\tilde{\Gamma}_+}. \tag{17}
$$

Depending on assumptions on g (amongst others), $\mathscr{A}$ from (17) can be considered as an operator from $\mathscr{C}^0([0, \infty[; L^1(\tilde{\Gamma}_-, d\xi))$ to $\mathscr{C}^0([0, \infty[; L^1(\tilde{\Gamma}_+, d\xi))$ (or suitable L^1 spaces, for details we refer to [3, 7, 12, 19, 23, 33]).

Whereas the (linear) albedo operator is mapping influxes to outfluxes at $\partial\Omega$, its dependence on the parameters $(m_1, m_2) = (a, f)$ gives rise to the *idealistic*

nonlinear forward mapping

$$\tilde{A}_{RTE} : (a, f) \mapsto \mathscr{A}[a, f] \tag{18}$$

which maps a given parameter pair (a, f) to all possible influx–outflux pairs belonging to the above described function spaces that are produced by the corresponding albedo operator for this particular parameter choice (a, f).

In practice not the idealistic forward map $\mathscr{A}[a, f]$ is accessible via physical measurements but only a small part of it, corresponding to a finite set of influx–outflux pairs, which we call $\mathbf{G}(a, f)$. This means, in practice, we want to determine $(m_1, m_2) = (a, f)$ from the *realistic nonlinear forward mapping*

$$A_{RTE} : (a, f) \mapsto \mathbf{G}(a, f) \tag{19}$$

where $\mathbf{G}$ defines a restriction of the albedo operator to certain experimentally feasible subsets of all possible influxes where possibly also the corresponding outfluxes might be observed only partially depending on the available measurement equipment. These two mappings $\tilde{A}_{RTE}$ and A_{RTE} define *inverse problems for the linear transport equation*, the former in an idealistic and the latter in a more practical setting.

We will demonstrate in this contribution that practically very often the RTE is approximated by a simpler model which has the consequence that also the idealistic and realistic forward mappings $\tilde{A}_{RTE}$ and A_{RTE} will have to be modified in order to accommodate the assumptions inherent in those approximations.

In [19] the following uniqueness result for the inverse problem of the RTE is stated and proven for suitable function spaces.

Let (a, f) and $(\hat{a}, \hat{f})$ be two admissible pairs with ζ-independent a,$\hat{a}$. Let further Ω be an open bounded set with $\mathscr{C}^1$-regular boundary such that $a, \hat{a}, f, \hat{f}$ vanish outside $\bar{\Omega}$. If in this case the corresponding albedo operators coincide on $\partial\Omega$ then it follows $a = \hat{a}$, $f = \hat{f}$.

Notice that this statement refers to the idealistic albedo operator (18) rather than its realistic counterpart (19). In [19] (as well as in [12, 73]), the singular structure of the fundamental solution of the linear transport equation is used for obtaining constructive expressions for determining (a, f) from complete knowledge of the albedo operator. Thus, a practical algorithm can be derived from complete observations for uniquely calculating the unknown scattering and absorption parameters from an idealized data set. In the terminology of the previous section, it amounts to using the lowest-order terms of the Neumann series for the inversion making use of the singular structure of the fundamental solution of the RTE. Physically these singular terms correspond to unscattered and single-scattered particles only. Higher-order terms in the Neumann series tend to be increasingly regular.

In some applications of imaging with the RTE, it is possible to extract some of these singular parts of the fundamental solution, at least approximately, from the measured data. This is, for example, the case in X-ray tomography where only few scattering events take place and singularities in the source terms can survive with

sufficient strength to be separated in the measurements from the more regular parts. Also collimators or energy filters can be used in order to isolate certain parts of the radiation. The standard approach is then to use the inversion of the Radon Transform (or X-ray transform in 3-D) for uniquely estimating the attenuation distribution $a(x)$ inside the domain of interest from the unscattered data [55]. Information on the scattering parameter f can then, if desired, be obtained from the additional analysis of once or multiple scattered radiation. See, for example, [7, 12, 55] for more details and additional information on practical aspects of this approach.

3 Forward Modeling of Transport Processes in DOT

3.1 The RTE in Biomedical Imaging

In most practical scenarios of imaging in highly scattering media, the unscattered and single-scattered contributions highlighted in the previous section are hidden in the noise and are dominated by the multiple scattered more regular terms. Therefore, separating data by the number of scattering events is impossible and different techniques need to be employed for the inversion. In particular, it is not clear in those cases whether a unique reconstruction of two independent distributed parameters is still possible. To make things worse, in most practical applications, only a finite set of measurements can be taken such that the complete albedo operator cannot be assumed known. The standard approach is here the use of optimization techniques where parameters of the RTE or of a related forward model are estimated as space-dependent functions from the measured data. This distributed parameter estimation problem with two different parameters interfering with each other during the inversion is a challenging problem where standard techniques of optimization often fail to yield satisfactory results due to the high degree of ill-posedness of the inverse problem and due to a strong cross talk between these two parameters. We will outline in the following a few of the approaches that have been followed so far in the literature and describe some of the challenges encountered in these approaches.

In DOT [4–6], the domain of interest Ω is irradiated by low-energy laser light in the near-infrared regime which travels through this domain following the model of an RTE. The outgoing light around the boundary of the domain of interest is measured and defines the data of the inverse problem. It is usually assumed that particles (photons) are propagating with constant speed c and that scattering events will not change the speed. Moreover, it is assumed that the scattering phase function f separates into a product of a spatial scattering cross section b and an angular component η which only depends on the cosine of the scattering angle. We will follow this approach and divide the RTE considered above by c to obtain the *RTE for DOT* in the form

$$\frac{1}{c}\frac{\partial u}{\partial t}+\theta\cdot\nabla u(x,\theta,t)+a(x)u(x,\theta,t)=b(x)\int_{S^2}\eta(\theta\cdot\theta')u(x,\theta',t)d\theta'+q(x,\theta,t), \tag{20}$$

with initial condition

$$u(x,\theta,0)=0 \qquad \text{in} \quad \Omega\times S^2, \tag{21}$$

and boundary condition

$$u(x,\theta,t)=0 \qquad \text{on} \quad \Gamma_-. \tag{22}$$

Here again S^2 denotes the unit sphere in $\mathbb{R}^3$ and we choose to model the incoming boundary flux in (3)

$$g(x_s,t_s):=\int_{S^2_-}u(x_s,\theta,t_s)\nu(x_s)\cdot\theta\,d\theta \quad \text{on } \partial\Omega\times[0,T] \tag{23}$$

as an equivalent surface source term q in (20) instead of an incoming boundary condition in (22) using expression (7). Here $S^2_\pm$ denotes the subset of direction vectors $\theta\in S^2$ for which $\pm\nu(x_r)\cdot\theta>0$. Furthermore, $(x,\theta,t)\in\mathbb{R}^3\times S^2\times\mathbb{R}$ and

$$a(x),b(x),\eta(\theta\cdot\theta')\geq 0, \quad a(x)\geq b(x), \quad c>0, \tag{24}$$

as well as

$$\int_{S^2}\eta(\theta\cdot\theta')d\theta'=1. \tag{25}$$

In many biomedical applications, the scattering phase function $\eta(\theta\cdot\theta')$ is highly forward peaked. A popular model for these applications is the Henyey–Greenstein (HG) scattering function in 3-D

$$\eta(\theta\cdot\theta')=\frac{1-g^2}{4\pi(1+g^2-2g\theta\cdot\theta')^{\frac{3}{2}}}=\sum_{n=0}^{\infty}\frac{2n+1}{4\pi}g^nP_n(\theta\cdot\theta'), \tag{26}$$

where P_n is the n-th order Legendre polynomial. The anisotropy factor $-1\leq g\leq 1$ in this function has the meaning of a mean scattering cosine. $g=0$ indicates isotropic scattering, $g>0$ primarily forward scattering, and $g<0$ primarily backward scattering. In biomedical applications, we have approximately $0.9<g<0.95$ or higher which indicates highly peaked forward scattering.

Motivated by (24) we denote further

$$\sigma_a(x)=a(x)-b(x)\geq 0 \tag{27}$$

which measures the probability at a given location that a particle is absorbed and disappears completely. Obviously knowledge of a and b entails that of σ_a and b and vice versa, such that we can search for any of these two parameter pairs in the inverse problem. By introducing the scattering operator $\mathscr{L}$ as

$$\mathscr{L}u \;=\; -u + \int_{S^2} \eta(\theta \cdot \theta')u(x, \theta', t)d\theta' \tag{28}$$

we can write (20) in shorter form as

$$\frac{1}{c}\frac{\partial u}{\partial t} + \theta \cdot \nabla u + \sigma_a(x)u - b(x)\mathscr{L}u = q. \tag{29}$$

The outgoing flux across the boundary $\partial\Omega$ at receiver position x_r and receiving time t_r for given parameters (σ_a, b) is given by

$$G[\sigma_a, b](x_r, t_r) := \int_{S^2_+} u(x_r, \theta, t_r)\nu(x_r) \cdot \theta \, d\theta \quad \text{on } \partial\Omega \times [0, T], \tag{30}$$

where u is the solution of the RTE (20)–(22) for the source q and parameters (σ_a, b). Expression (30) typically corresponds to the expected measurements given a parameter pair (σ_a, b). We obtain therefore the following variant of (19) for the application in DOT

$$A_{RTE} \;:\; (a, b) \mapsto \mathbf{G}(a, b) \tag{31}$$

which due to (27) can also be formulated as (using the same symbol as before)

$$A_{RTE} \;:\; (\sigma_a, b) \mapsto \mathbf{G}(\sigma_a, b)\,. \tag{32}$$

When physically measured (or independently simulated) data $\tilde{G}$ are available for a given source q, then we can define the corresponding residuals $R[\sigma_a, b]$ as

$$R[\sigma_a, b](x_r, t_r) = G[\sigma_a, b](x_r, t_r) - \tilde{G}(x_r, t_r), \tag{33}$$

describing the mismatch between predicted and measured data. One standard way to proceed now is to write down the least-squares data misfit functional

$$\mathscr{J}(m_1, m_2) \;=\; \mathscr{J}(\sigma_a, b) := \frac{1}{2}\,\|R[\sigma_a, b]\|_2^2\,, \tag{34}$$

and adding suitable regularization terms in order to obtain the so-called "cost functional" of the underlying optimization problem.

Similar expressions to (34) and equivalent regularization terms are obtained for the approximations to the RTE discussed below where $(m_1, m_2) \;=\; (\sigma_a, b)$ are replaced by the two parameters that appear in the corresponding approximation.

Gradient-based optimization techniques then require to calculate $\nabla_{m_1,m_2}\mathscr{J}$ in each step of an iteration. The practical calculation of these gradients clearly has an impact on the convergence properties of each method but also on the way how different distributed parameters are separated from each other in the inversion. We will discuss further below some of these gradient calculation techniques and highlight some specific features of them. Before doing so, we will outline the most common approximations applied to this RTE-constrained optimization problem using partial differential equations (PDEs) or systems of those instead.

We notice that those derivations as presented here are mainly based on formal calculations, and validity of each of these approximate models, including the choice of appropriate function spaces, needs to be justified by doing additional quantitative analysis. In fact, for example, the diffusion approximation stated further below comes in different flavors whose details depend on details of its formal derivation technique. Starting point for this particular model can be either the above introduced RTE or more fundamental models such as Maxwell's equations, linear elasticity or general wave equations. These aspects have been studied intensively in the literature and we refer the reader to the list of references, for example [5, 7, 40, 61, 66, 70, 76]. We will restrict ourselves to outline just one of these different derivations, which is based on an expansion in spherical harmonics.

3.2 The $\mathscr{P}_1$ Approximation

In some applications, the solution of the RTE can be approximated by systems of PDEs considering certain lower-order angular moments of the flux density. The $\mathscr{P}_N$-approximation is such a system where the positive integer $N > 0$ describes the level of approximation. It can be derived from the RTE by using a spherical harmonics expansion of the angularly dependent quantities. In this approach, we expand

$$u(x,\theta,t) = \sum_{l=0}^{\infty}\sum_{m=-l}^{+l}\left(\frac{2l+1}{4\pi}\right)^{\frac{1}{2}}\psi_{lm}(x,t)Y_{lm}(\theta), \tag{35}$$

$$q(x,\theta,t) = \sum_{l=0}^{\infty}\sum_{m=-l}^{+l}\left(\frac{2l+1}{4\pi}\right)^{\frac{1}{2}}q_{lm}(x,t)Y_{lm}(\theta), \tag{36}$$

$$\eta(\theta\cdot\theta') = \sum_{l=0}^{\infty}\frac{2l+1}{4\pi}f_l P_l(\theta\cdot\theta') = \sum_{l=0}^{\infty}\sum_{m=-l}^{+l}f_l Y_{lm}^*(\theta')Y_{lm}(\theta) \tag{37}$$

where $Y_{lm}(\theta)$ are the (complex-valued) spherical harmonics, the P_l are Legendre polynomials, and * means "complex conjugate." By inserting (35)–(37) into (20) we obtain an infinite hierarchical sequence of coupled PDEs for the expansion coefficients $\psi_{lm}(x,t)$, $q_{lm}(x,t)$ and f_l which can be truncated at an arbitrary level

$N > 0$ in order to obtain computable approximations to the RTE. This yields the $\mathscr{P}_N$-approximation to the RTE. The expressions obtained in the general $\mathscr{P}_N$-approximation are lengthy such that we omit their presentation here. Instead we refer to the detailed discussions in [5, 15, 27, 43, 81]. The lowest-order truncation terms with $N = 1$ yield the $\mathscr{P}_1$-approximation which then can be used for deriving the telegraph approximation and the diffusion approximation by applying further simplifications. This will be outlined in the following.

We introduce the combined parameter

$$\sigma_{tr}(x) := a(x) - \bar{\mu}_0 b(x) \tag{38}$$

with the "mean scattering cosine"

$$\bar{\mu}_0 = \int_{S^2} \theta' \cdot \theta \eta(\theta' \cdot \theta) d\theta' . \tag{39}$$

We will consider in the following the three angular density moments

$$\Phi(x,t) := \int_{S^2} u(x,\theta,t) d\theta , \tag{40}$$

$$J(x,t) := \int_{S^2} \theta\, u(x,\theta,t) d\theta , \tag{41}$$

$$T(x,t) := \int_{S^2} \theta \otimes \theta\, u(x,\theta,t) d\theta \tag{42}$$

and also will make use of the angular source moments

$$q_0(x,t) := \int q(x,\theta,t)\, d\theta, \tag{43}$$

$$q_1(x,t) := \int_{S^2} \theta q(x,\theta,t) d\theta . \tag{44}$$

Notice that $\Phi(x,t)$ is a scalar quantity which describes the (direction-independent) photon density at location x and time t, whereas the vectorial term $J(x,t)$ physically describes a photon flux density. The symbol $\otimes$ denotes the tensor product of two vectors. There is no direct physical meaning associated with the tensor T. It can be described component-wise by

$$T = (T_{ij})_{i,j=1,\ldots,n} = (\int_{S^2} \theta_i \theta_j u d\theta)_{i,j=1,\ldots,n} ,$$

where θ_i, θ_j are the cartesian components of θ, see [43]. Truncating the above described infinite expansion at level $N = 1$ and using the expressions (27) and (38) for σ_a and σ_{tr}, we obtain the lowest-order terms

$$\frac{1}{c}\frac{\partial \Phi}{\partial t} + \nabla \cdot J + \sigma_a \phi = q_0, \tag{45}$$

$$\frac{1}{c}\frac{\partial J}{\partial t} + \nabla \cdot T + \sigma_{tr} J = q_1. \tag{46}$$

In (45), (46) we have identified the lowest-order coefficients of (35) as $\psi_{0,0} = \Phi$ and $J = (J_x, J_y, J_z)$ with

$$J_x = \frac{1}{\sqrt{2}}\left[\psi_{1,-1} - \psi_{1,1}\right], \quad J_y = -\frac{i}{\sqrt{2}}\left[\psi_{1,-1} + \psi_{1,1}\right], \quad J_z = \psi_{1,0},$$

where i is the imaginary unit. A similar relationship holds for (36) with respect to q_0, q_1. The system (45), (46) so far does not contain any approximation and needs to be satisfied by any solution of (20). However, it is not closed and would refer to coefficients of degree higher than $N = 1$ in order to uniquely determine the tensor T. Therefore it cannot be solved for all unknowns in a unique way. We can formally close the system by expressing the higher order moment $T(x, t)$ as a linear combination of the two lower-order moments Φ, J. *This can be achieved by assuming that $u(x, \theta, t)$ depends only linearly on the angle θ.* Using a projection approach, this assumption gets the form

$$u(x, \theta, t) \doteq \frac{1}{4\pi}\Phi(x, t) + \frac{3}{4\pi} J(x, t) \cdot \theta \tag{47}$$

where we introduced the notation $\doteq$ (also used in the following) to indicate that a given quantity is replaced by a model-dependent approximation (or representation) of it. Plugging this into the definition of $T(x, t)$ yields then $\nabla \cdot T \doteq \frac{1}{3}\nabla\Phi$ which eliminates the T-dependence in (45), (46). In a similar way, we can approximate

$$q(x, \theta, t) \doteq \frac{1}{4\pi} q_0 + \frac{3}{4\pi} q_1 \cdot \theta$$

assuming that the source term only linearly depends on the direction θ. Also here q_1 is a vectorial quantity analogous to J in the expansion of u. Combining these approximations, we arrive at the (real-valued) *$\mathscr{P}_1$-approximation of the RTE*

$$\frac{1}{c}\frac{\partial \Phi}{\partial t} + \nabla \cdot J + \sigma_a \Phi = q_0, \tag{48}$$

$$\frac{1}{c}\frac{\partial J}{\partial t} + \frac{1}{3}\nabla\Phi + \sigma_{tr} J = q_1. \tag{49}$$

(48) is a scalar equation, whereas (49) is vectorial with three components. Therefore, the $\mathscr{P}_1$-approximation is a system of four PDEs in four unknowns which formally is closed and can be solved analytically or numerically. Notice that the inverse problem of the $\mathscr{P}_1$-approximation would look for $\sigma_a(x)$ and $\sigma_{tr}(x)$ instead of $\sigma_a(x)$ and $b(x)$

(or alternatively $a(x)$ and $b(x)$) of the RTE using P_1-interpretations of the applied sources and measured data:

$$A_{P1}(m_1, m_2) = A_{P1}(\sigma_a, \sigma_{tr}).$$

An optimization scheme would require gradient directions with respect to these two parameters, which are actually combinations of $a(x)$ and $b(x)$ as seen in (27) and (38). This will affect in particular the simultaneous reconstruction of $a(x)$ and $b(x)$ from a given data set.

3.3 Approximation by a Telegraph Equation

From the $\mathscr{P}_1$-system (48), (49) for Φ and J a scalar second-order PDE model can be extracted which describes the behavior of $\Phi(x, t)$ only. This PDE has the form of a telegraph approximation. For this purpose, we define the diffusion coefficient D by

$$D(x) = \frac{1}{3\sigma_{tr}(x)} = \frac{1}{3(a - \bar{\mu}_0 b)}. \tag{50}$$

Using this definition (48) and (49) write as

$$\frac{\partial \Phi}{\partial t} + c\nabla \cdot J + c\sigma_a \Phi = cq_0, \tag{51}$$

$$\frac{3D}{c}\frac{\partial J}{\partial t} + D\nabla\Phi + J = 3Dq_1. \tag{52}$$

Formal differentiation of (51) with respect to time yields

$$\frac{\partial^2 \Phi}{\partial t^2} + c\nabla \cdot \frac{\partial J}{\partial t} + c\sigma_a \frac{\partial \Phi}{\partial t} = c\frac{\partial q_0}{\partial t}. \tag{53}$$

Inserting now (52) into (53) gives

$$\frac{\partial^2 \Phi}{\partial t^2} + \frac{c^2}{3D}\nabla \cdot [3Dq_1 - D\nabla\Phi - J] + c\sigma_a \frac{\partial \Phi}{\partial t} = c\frac{\partial q_0}{\partial t}, \tag{54}$$

such that

$$\frac{3D}{c^2}\frac{\partial^2 \Phi}{\partial t^2} + 3D\nabla \cdot q_1 - \nabla \cdot D\nabla\Phi - \nabla \cdot J + \frac{3D\sigma_a}{c}\frac{\partial \Phi}{\partial t} = \frac{3D}{c}\frac{\partial q_0}{\partial t}.$$

By using again (51) we can furthermore eliminate $\nabla \cdot J$ obtaining the *telegraph approximation to the RTE*

$$\frac{3D}{c^2}\frac{\partial^2\Phi}{\partial t^2} - \nabla\cdot D\nabla\Phi + \frac{1}{c}(1+3D\sigma_a)\frac{\partial\Phi}{\partial t} + \sigma_a\Phi = q_{TA} \tag{55}$$

with the effective source term

$$q_{TA} = \frac{3D}{c}\frac{\partial q_0}{\partial t} - 3D\nabla\cdot q_1 + q_0. \tag{56}$$

It describes some form of damped wave which at the same time shows diffusive behavior (see [22] for some general aspects of such telegraph equation models). The front of this damped wave propagates with finite speed c in the scattering medium. As pointed out in [79], the group velocity of the wave is however smaller than the speed of individual particles, namely $\frac{c}{\sqrt{3}}$, due to the added diffusive component.

Notice that the inverse problem of the telegraph equation would attempt to reconstruct $\sigma_a(x)$ and $D(x)$ instead of $\sigma_a(x)$ and $b(x)$ (or alternatively $a(x)$ and $b(x)$) of the RTE:

$$A_{TE}(m_1, m_2) = A_{TE}(\sigma_a, D).$$

An optimization scheme would require gradient directions with respect to these two parameters, which are actually combinations of $a(x)$ and $b(x)$ as seen in (27) and (50). Observe also that the combination $(1 + 3D\sigma_a)$ in front of the first-order time derivative term introduces a complicated structure into the inverse problem. This will affect in particular the simultaneous reconstruction of $a(x)$ and $b(x)$ from a given data set.

The telegraph equation in its complexity is not necessary in DOT once it is accepted that only an equation for the scalar angular-independent quantity $\Phi(x,t)$ is desired. Diffusion becomes dominant in DOT due to the physically considered range of parameters (see the numerical simulations shown in Section 4). The wave propagation feature is hardly observable in this application such that a much simpler approximation is usually employed, namely a diffusion approximation, which is described next.

3.4 Approximation by a Diffusion Equation

When putting in (52)

$$\frac{3D}{c}\frac{\partial J}{\partial t} \equiv 0 \quad , \quad q_1 \equiv 0 \tag{57}$$

then a simpler PDE can be derived for modeling photon propagation in tissue. (57) essentially amounts to the assumptions that the source term is fully isotropic and that the speed of particle propagation is much larger than the diffusion coefficient

(or, alternatively, that J varies very slowly over time). Combining the assumptions (57) with (52) yields directly *Fick's diffusion law*

$$J(x,t) = -D(x)\nabla\Phi(x,t). \tag{58}$$

Plugging this into (51) we obtain the *diffusion equation*

$$\frac{1}{c}\frac{\partial\Phi}{\partial t} - \nabla\cdot D(x)\nabla\Phi(x,t) + \sigma_a(x)\Phi(x,t) = q_0(x,t). \tag{59}$$

We still need to equip this new model with appropriate initial and boundary conditions, and take care of a suitable model for the measurements, in order to finally obtain the *diffusion approximation (DA) of the RTE*. This will be done in the following section.

As already mentioned, alternative derivations of the diffusion approximation might yield slightly different representations for the diffusion coefficient $D(x)$. See, for example, [5, 7, 40, 61, 66, 70, 76].

According to our assumptions, the diffusion model uses an infinite speed of signal propagation, and furthermore is unable to model correctly the propagation of singularities which might be contained in the applied source terms or which might occur close to interfaces or the boundary. Therefore, sometimes a Neumann series expansion is employed for separating unscattered and single-scattered particles (carrying any singular behavior of the sources) from the more regular higher-order terms, and only the higher-order terms are then represented by a diffusion model. Also boundary regions and regions of low scattering coefficient (if present) can be modeled by adjusting some terms in the diffusion approximation up to a certain degree of accuracy (which is not always satisfactory, though). We refer for details to the general references given in the introduction and the next section.

Finally we mention that, similar to the telegraph model, the inverse problem of the diffusion approximation attempts to reconstruct $\sigma_a(x)$ and $D(x)$ from the given data:

$$A_{DA}(m_1, m_2) = A_{DA}(\sigma_a, D).$$

Once these are obtained, in principle the corresponding parameters $\sigma_a(x)$ and $b(x)$ of the RTE model can be deduced using (27) and (50). Practically difficulties might arise though due to the ill-posedness and non-uniqueness of the inverse problem and the slightly more complicated structure of (50). Moreover, standard regularization strategies focus on $\sigma_a(x)$ and $D(x)$ rather than $\sigma_a(x)$ and $b(x)$ which might have a negative impact on the separation of the latter parameters from each other.

3.5 Measurements and Boundary Conditions in the Diffusion Approximation

Any approximation used for the forward modeling of radiative transport phenomena still needs to be able to approximate the measured data (30) reasonably well if we want to have a chance of reconstructing the distributed parameters from measured data in a reliable way. We will outline in the following briefly how this can be achieved in the diffusion approximation (DA) and refer the reader for more details to [10, 36, 40, 44, 45, 52, 53, 67].

In (47) we have seen that in the $\mathscr{P}_1$-approximation the flux $u(x,\theta,t)$ is linearly approximated by

$$u_{diff}(x,\theta,t) := \frac{1}{4\pi}\Phi(x,t) + \frac{3}{4\pi}J(x,t)\cdot\theta\,. \tag{60}$$

The incoming radiation $I_-(x,t)$ at position $x \in \partial\Omega$ and time t is then described in the DA by

$$I_-(x,t) \;=\; \int_{\nu(x)\cdot\theta<0} \nu(x)\cdot\theta\left[\frac{1}{4\pi}\Phi(x,t) + \frac{3}{4\pi}J(x,t)\cdot\theta\right]d\theta. \tag{61}$$

Replacing $J(x,t)$ in this expression by using (58) and performing the integration the free boundary condition (22) obtains in the DA the form

$$I_-(x,t) \;=\; -\frac{1}{4}\left[\Phi(x,t) + 2D\nu(x)\cdot\nabla\Phi\right] \;=\; 0 \tag{62}$$

or, equivalently,

$$\Phi(x,t) + 2D\frac{\partial\Phi}{\partial\nu} = 0 \tag{63}$$

with ($\frac{\partial}{\partial\nu}$ = outer normal derivative). This is a *Marshak*- or *Robin* boundary condition. Other approximations are possible as well, see [15].

Similarly, recall that the detector measurements are given as

$$g(x,t) \;=\; \int_{S^2_+} \nu(x)\cdot\theta\, u(x,\theta,t)d\theta \tag{64}$$

for $(x,t) \in \partial\Omega \times \mathbb{R}^+$. With the same approach as before and using (22) we obtain here

$$g(x,t) \doteq \int_{S^2} \nu(x)\cdot\theta\,[\frac{1}{4\pi}\Phi(x,t) + \frac{3}{4\pi}J(x,t)\cdot\theta]\,d\theta.$$

Applying again Fick's law, we obtain after integration

$$g(x,t) \doteq -D\nu(x)\cdot\nabla\Phi(x,t) \;=\; -D\frac{\partial\Phi}{\partial\nu}(x,t). \tag{65}$$

For more details see [5, 27, 57]

3.6 Fokker–Planck and δ-Eddington Approximations

We follow here closely [35]. We mentioned that in biomedical applications the mean scattering cosine of the Henyey–Greenstein phase function (26) takes values of approximately $0.9 < g < 0.95$ or higher, which indicates highly peaked forward scattering. These situations are difficult to model *numerically* with the RTE [48] such that specific approximations have been developed in the literature for simplifying such situations of highly peaked forward scattering. Assuming in the following the HG scattering function in 3-D, the Fokker-Planck (FP) approximation replaces the scattering operator $\mathscr{L}$ given in (28) by the easier to evaluate FP-scattering operator

$$\mathscr{L}u \approx \mathscr{L}_{FP}u \;=\; \frac{1}{2}(1-g)\Delta_\theta u \tag{66}$$

where Δ_θ denotes the spherical Laplacian. Likewise, in the δ-Eddington (δE) approximation $\mathscr{L}$ is replaced by

$$\mathscr{L}u \approx \mathscr{L}_{\delta E}u = -(1-g^2)u + \frac{1-g^2}{4\pi}\int_{S^2}\left[P_0(\theta\cdot\theta') + \frac{3g}{1+g}P_1(\theta\cdot\theta')\right]u(x,\theta',t)d\theta'. \tag{67}$$

In both approximations, the cumbersome scattering integral of the RTE (20) is replaced by simpler expressions, which simplifies the numerical modeling. Notice that the unknowns of the inverse problems are still $\sigma_a(x)$ and $b(x)$:

$$A_{FP}(m_1, m_2) = A_{FP}(\sigma_a, b).$$

No transformation is necessary after solving the inverse problem of these approximations to obtain the original parameters of the RTE.

3.7 Monte Carlo Methods

Monte Carlo (MC) methods for modeling particle propagation in scattering media have a long history. Using statistical sampling, the goal in these methods is to track a large number of particles in an equivalent reference medium in order to

find statistical estimates of the quantities of interest, for example of measured data. The unknown parameters in the inverse problem are here probability densities of absorption and scattering events which can be estimated using statistical estimation theory. The MC method is extremely versatile and not limited by regularity assumptions compared to the other models discussed so far. Moreover, it is possible to make a direct link between quantities in the RTE and the probabilities needed in a MC simulation such that several concepts of parameter estimation from the theory of RTE can be applied also to the parameter estimation problem arising in MC strategies. For more details, see [8, 18, 37, 50, 51, 71, 75, 78].

4 Numerical Simulations

We mentioned in the introduction that the RTE can model a large variety of phenomena and we have introduced different approximations to the RTE describing particle propagation by quite different-looking PDEs. The tool of MC simulations (see Section 3.7) allows us to demonstrate that indeed different combinations of parameters lead to different-looking propagation patterns of particles in the domain. In the following figures (Figures 1, 2, 3, 4, and 5) we present some 2-D snapshots of time evolutions that have been generated by an in-house 2-D MC simulator using in total 5×10^7 particles doing a random walk through a rectangular domain Ω of size $4\,\mathrm{cm} \times 6\,\mathrm{cm}$. For more details on the setup of this simulator, we refer to [27].

All simulations have in common that both the absorption cross section σ_a and the scattering cross section b have been chosen constant over Ω. The absorption cross section is chosen to be the same in all figures, namely $\sigma_a = 0.001\,\mathrm{cm}^{-1}$, but both the scattering cross section b as well as the scattering parameter g in a 2-D adaptation of the Henyey–Greenstein phase function (26) are varied between different simulations. The choice $g = 0$ amounts here to isotropic scattering, and $g = 0.9$ indicates highly forward peaked scattering.

The particle speed c has been normalized to $c = 1\,\mathrm{cm\,s}^{-1}$. The ten snapshots show ten different equidistant time steps $t_1 < t_2 < \ldots < t_{10}$ of the particle density $\Phi(x, t)$ calculated by a MC version of (40). The source is an ultrashort pulse of particles injected at time $t = 0$ at the center of the upper straight boundary in the direction perpendicular to $\partial\Omega$ into this rectangular domain Ω. In the bottom row images of the figures, different vertical cross sections of particle densities are shown that are indicated by (coloured) lines in the corresponding snapshots at times $t_2, t_4, t_6, t_8, t_{10}$, respectively. The left-hand cross sections are taken through the line defined by the incoming source direction which contains (amongst others) "ballistic" (i.e., unscattered) components of the particle density. The right-hand side images show cross sections taken slightly off this "ballistic" line. Further details are given in the caption of each figure.

Finally, Figure 6 shows a data set obtained by running the MC simulation code for a 2-D version of a typical DOT application. The source here is similar to before but now located at one of the longer sides of the domain Ω. The bottom

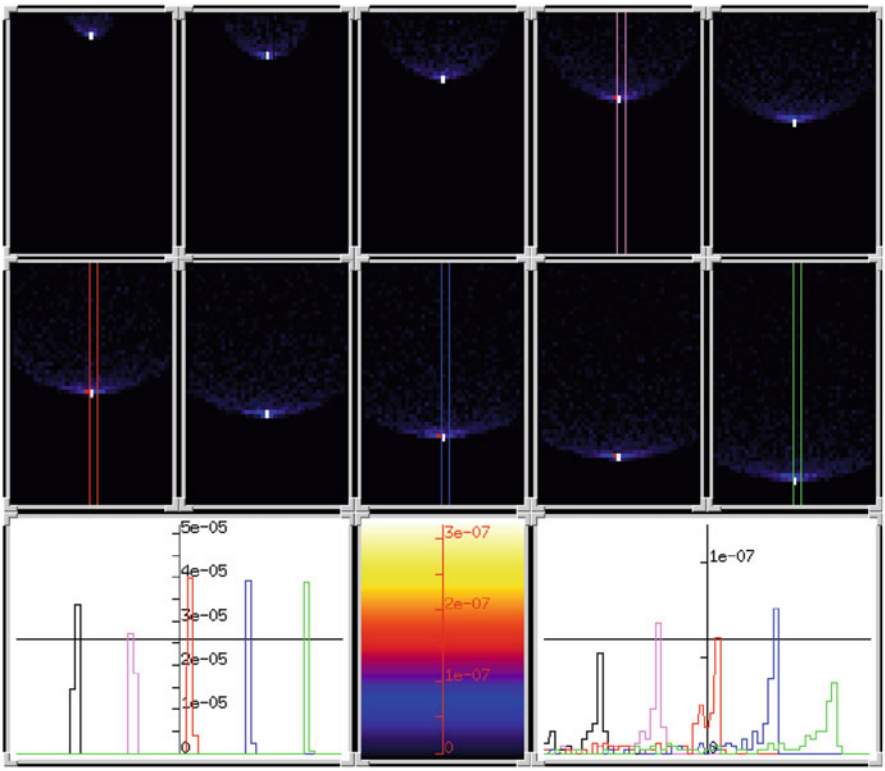

Fig. 1 $b = 0.05\,\mathrm{cm}^{-1}$, $g = 0.0$. Here the ballistic (unscattered) contribution is dominant (bottom left image) and only a small scattered side-lobe appears (bottom right image)

left figure shows the MC simulated data (30) arranged such that the vertical axis corresponds to the boundary position (clearly showing one main lobe and three side lobes corresponding to the four sides of Ω) and the horizontal axis corresponds to time (increasing from left to right). The boundary position where the source is located can be easily identified as the position where data are available immediately at time $t \approx 0$. Two cross sections are indicated in this data set by (green) lines, and the corresponding data curves are shown on the two right-hand side plots. The upper plot corresponds to a receiver position close to the source where a high number of photons is detected due to its proximity to the source position. The lower plot corresponds to a receiver position at one of the adjacent (shorter) sides of the domain Ω which shows a relatively low number of detected photons due to the increased distance from the source position.

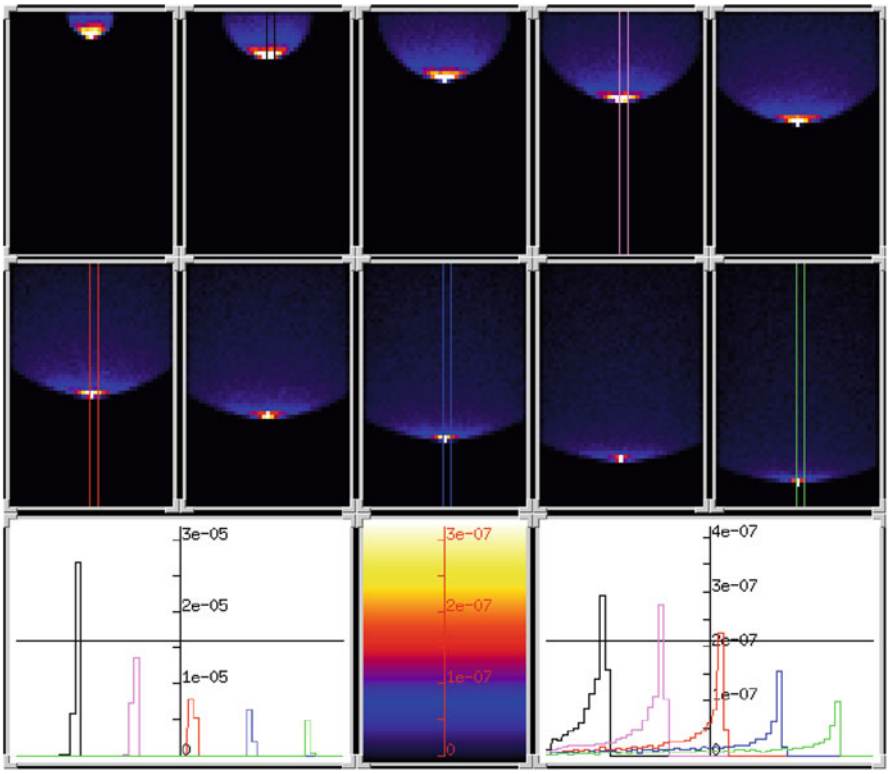

Fig. 2 $b = 0.5\,\mathrm{cm}^{-1}$, $g = 0.0$. The scattered part becomes stronger but the ballistic contribution is still is visible. A more spread-out "wave front" becomes visible and increased isotropic scattering contributions give rise to more significant particle densities "behind" this wave front

5 Gradient Calculation

When solving inverse problems using gradient-based techniques, an important question to be answered is whether we want to follow the rule "first optimize then discretize" or "first discretize then optimize." In the first case, some form of Fréchet derivative or Gateaux derivative of the forward model needs to be calculated first, usually based on so-called adjoint techniques, and then both the forward and the adjoint continuous model (both represented by RTEs or PDEs) are discretized and combined in order to arrive at a discretization of a "continuous descent direction." These obtained descent directions are closely related to so-called "sensitivities" of the parameter-to-data maps, which play a major role in many applications [29, 47, 49, 56]. In the second approach, first the forward problem is discretized

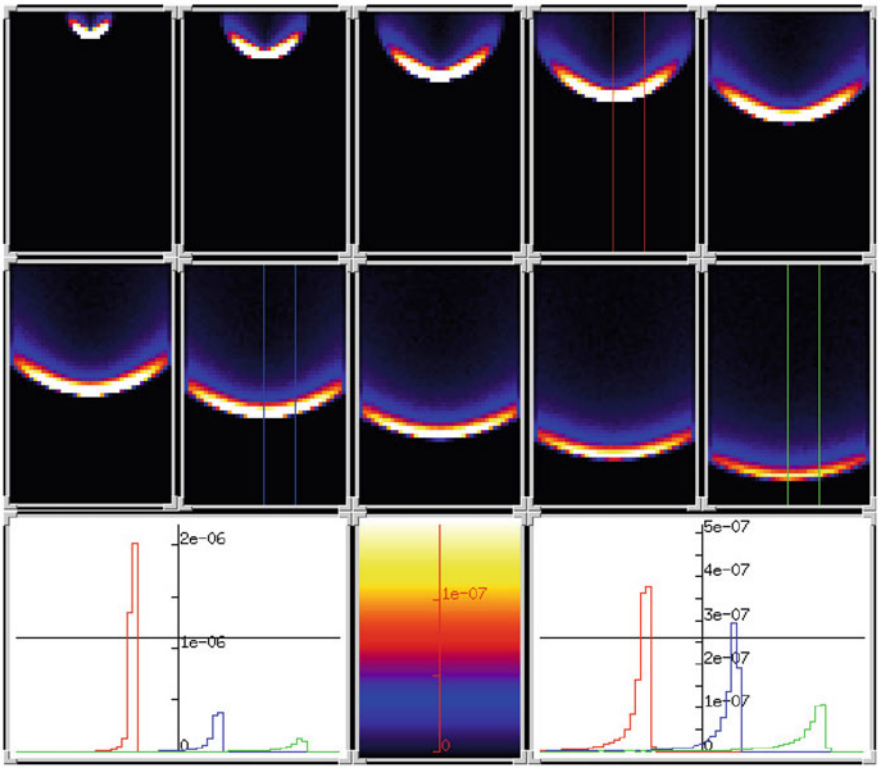

Fig. 3 $b = 2.5\,\mathrm{cm}^{-1}$, $g = 0.9$. Now scattering becomes more dominant with a highly forward peaked scattering phase function. The corresponding photon density distribution resembles now more a damped wave propagating inside Ω which might be well approximated by a telegraph equation

arriving at a discrete optimization problem. Then, a gradient of the resulting finite-dimensional problem is calculated using standard rules from optimization. Also in this second approach, adjoint techniques are often employed, which usually lead to matrix-based adjoints. Unfortunately, there is no guarantee that these matrix-based adjoints coincide with the discretized version of the continuous adjoints obtained in the first approach. In this contribution, we will only follow the first approach ("first optimize then discretize") and refer to general explanations regarding the second approach to [58]. Moreover, we will only state practical expressions for derivatives or gradients without actually proving differentiability and without (except for the RTE itself) stating specific function spaces. Many of those details are discussed in the cited literature.

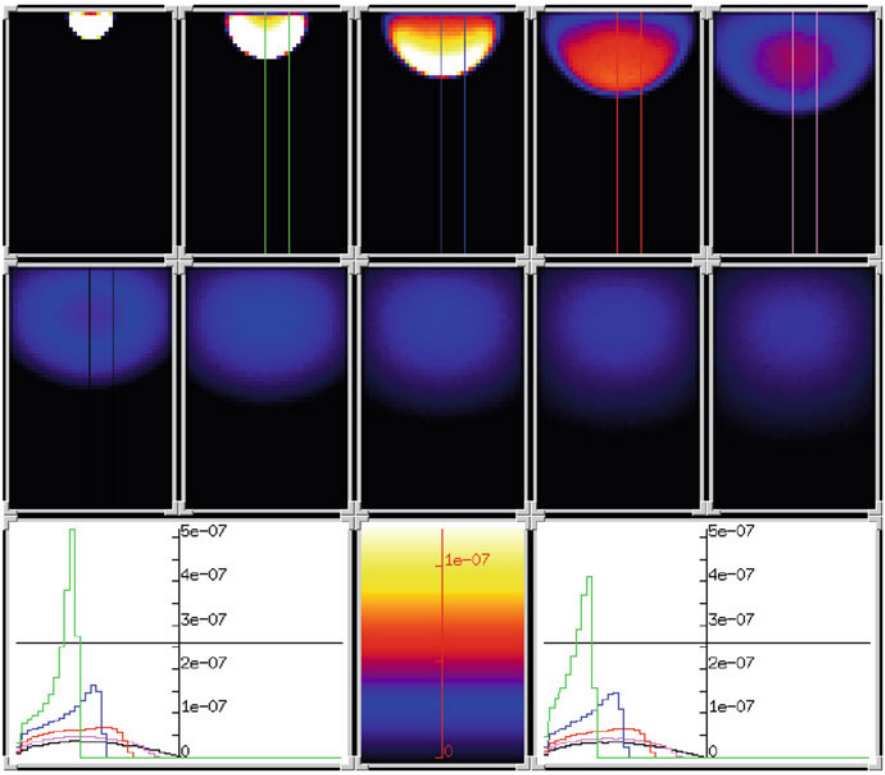

Fig. 4 $b = 25.0\,\mathrm{cm}^{-1}$, $g = 0.9$. Now scattering dominates the particle propagation but due to the highly forward peaked scattering phase function the penetration depth of the particles is still good. The diffusion approximation seems a good model for this situation

5.1 Gradient for the RTE

Following [27, 28], we denote the function space

$$W_\infty := \Big\{ z \in L^\infty(\Omega \times S^2 \times [0,T]),\ (\partial_t + \theta\cdot\nabla)z \in L^\infty(\Omega \times S^2 \times [0,T]),$$
$$z|_{\Gamma_+} \in L^\infty(\Gamma_+, \nu\cdot\theta d\sigma d\theta dt),\ z(x,\theta,T) = 0 \text{ at } \Omega \times S^2 \Big\}.$$

Let $z \in W_\infty$ be the solution of the *adjoint RTE*

$$-\frac{\partial z}{\partial t} - \theta\cdot\nabla z(x,\theta,t) + (\sigma_a(x)+b(x))\, z(x,\theta,t) - b(x)\int_{S^2} \eta(\theta\cdot\theta')\, z(x,\theta',t)d\theta'$$
$$= 0 \qquad \text{in} \quad \Omega \times S^2 \times [0,T], \tag{68}$$

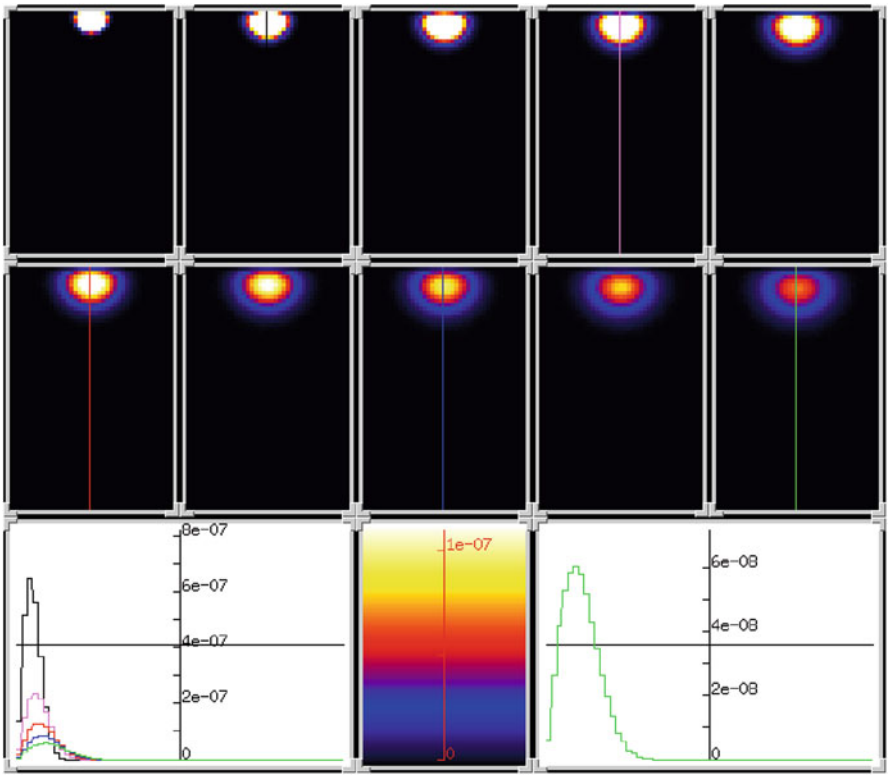

Fig. 5 $b = 25.0\,\mathrm{cm}^{-1}$, $g = 0.0$. This situation is similar to the one in Figure 4 but now, due to the isotropic scattering, the penetration depth of the particles is smaller. Also here the diffusion approximation seems a good model for this situation

with the final condition

$$z(x, \theta, T) = 0 \qquad \text{on} \quad \overset{\circ}{\Omega} \times S^2 \tag{69}$$

and boundary condition

$$z|_{\Gamma_+} = \tilde{z}(x, t) := \lambda(x, t) \in L^\infty(\partial\Omega \times [0, T]) \tag{70}$$

with $\lambda(x, t) = R[\sigma_a, b](x, t)$. In (69) the symbol $\overset{\circ}{\Omega}$ denotes the open interior of the region Ω and in (70) the residuals $\lambda(x, t)$ are assumed to be applied uniformly into all directions with $\theta \cdot \nu > 0$. This adjoint RTE models some form of back-transport where virtual photon densities proportional to the residual values are applied at receiver positions and then propagated backward in time and direction into the medium. This concept is well known in many fields and has a long history in neutron transport, see, for example, [47].

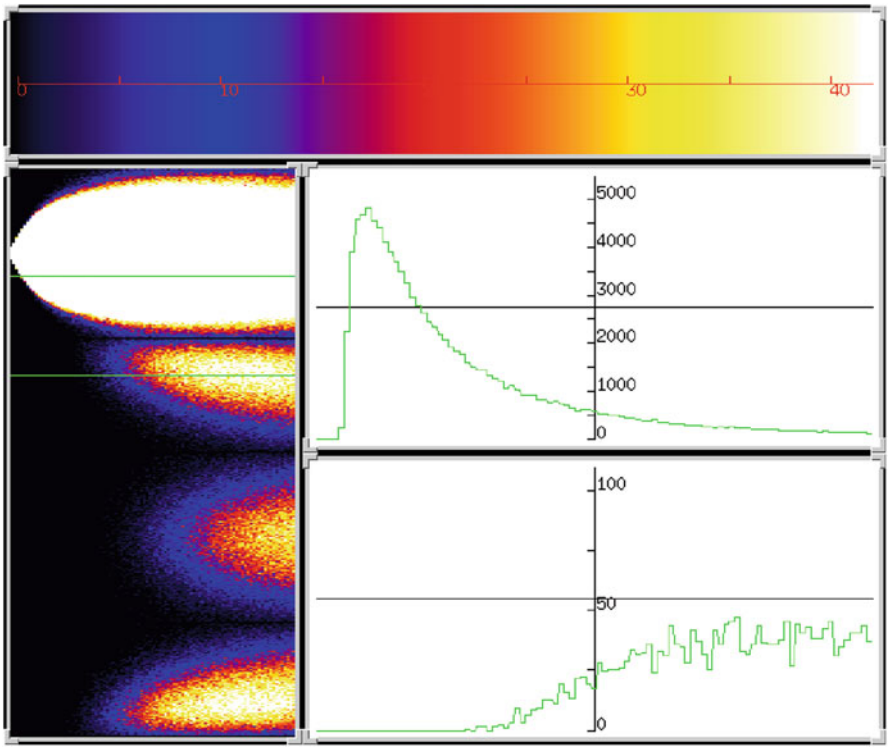

Fig. 6 $\sigma_a = 0.025\,\mathrm{cm}^{-1}$, $b = 100.0\,\mathrm{cm}^{-1}$, $g = 0.9$. Shown are the MC-simulated data as explained in the text

The gradient $\nabla_{\sigma_a,b}\mathscr{J}$ is now given as

$$\nabla_{\sigma_a,b}\mathscr{J}(x) = \left[R'\begin{pmatrix}\sigma_a\\ b\end{pmatrix}^*\lambda\right](x) := \begin{pmatrix}-I_1(u,z;x)\\ I_2(u,z;x) - I_1(u,z;x)\end{pmatrix}, \tag{71}$$

$$I_1(u,z;x) := \int_{[0,T]}\int_{S^2} u(x,\theta,t)z(x,\theta,t)\,d\theta dt, \tag{72}$$

$$I_2(u,z;x) := \int_{[0,T]}\int_{S^2}\left(\int_{S^2}\eta(\theta\cdot\theta')u(x,\theta',t)d\theta'\right)z(x,\theta,t)\,d\theta dt \tag{73}$$

where u is the solution of the forward problem (20)–(22). Using the operator $\mathscr{L}$ defined in (28) we can also write

$$\nabla_{\sigma_a,b}\mathscr{J}(x) = \left[R'\begin{pmatrix}\sigma_a\\ b\end{pmatrix}^*\lambda\right](x) := \begin{pmatrix}-\int_{[0,T]}\int_{S^2} uz\,d\theta dt\\ \int_{[0,T]}\int_{S^2}(\mathscr{L}u)z\,d\theta dt\end{pmatrix} \tag{74}$$

where the adjoint equation (68) can be written as

$$-\frac{1}{c}\frac{\partial z}{\partial t} - \theta \cdot \nabla z + \sigma_a(x) z - b(x) \mathscr{L} z = 0. \tag{75}$$

The adjoint linearized residual operator $R'\left(\begin{smallmatrix}\sigma_a\\ b\end{smallmatrix}\right)^*$ defined by (71) or (74) is linear.

Notice that the forward and adjoint photon densities are only integrated in (71) and (74) but not differentiated such that no strong regularity assumptions are made on u and z. These expressions can be evaluated even if u or z are not differentiable but contain certain singularities. For more details and numerical results see [5, 26–29, 47, 56, 57].

5.2 Gradient for the $\mathscr{P}_1$ Approximation

Gradient expressions for the general $\mathscr{P}_N$ approximation to the RTE can be obtained either directly from the infinite system or alternatively from (71) by plugging the $\mathscr{P}_N$ expansions (35) of u and z into (72), (73). For more details, we refer to [81] where numerical reconstructions are presented for $N = 1, 3, 5$ from data simulated with the $N = 7$ model (using a frequency-domain RTE). The telegraph approximation can be dealt with directly as well but as such it shows quite a complex structure with respect to the unknown parameters as mentioned in Section 3.3. Due to its close relationship to the $\mathscr{P}_1$ approximation, it seems easier to address the inverse problem by the $\mathscr{P}_1$ approximation, or alternatively to resort to the diffusion approximation as described next.

5.3 Gradient for the Diffusion Approximation

We consider the diffusion approximation to the RTE

$$\frac{\partial \Phi}{\partial t} + \sigma_a \Phi - \nabla \cdot (D \nabla \Phi) = Q, \tag{76}$$

$$\Phi(x, 0) = 0 \qquad \text{in } \Omega, \tag{77}$$

$$\Phi(x, t) + 2D \frac{\partial \Phi}{\partial \nu}(x, t) = 0 \quad \text{at } \partial\Omega \times [0, T] \tag{78}$$

with Robin boundary condition (63)

$$I_-(\Phi)(x, t) = 0 \quad \text{at } \partial\Omega\,. \tag{79}$$

Data $G(x,t)$ are given by (65) as

$$G(x,t) \;=\; -D\frac{\partial\tilde{\Phi}}{\partial\nu}(x,t) \quad \text{at } \partial\Omega\times[0,T]. \tag{80}$$

We define the residual operator by

$$R(\sigma_a,D)(x,t) \;=\; -D\frac{\partial\Phi}{\partial\nu}(x,t) - G(x,t), \tag{81}$$

with Fréchet derivative denoted by $R^{'}(\sigma_a,D)$. Let furthermore $g(x,t) = R(\sigma_a,D)(x,t)$. Then the gradient $\nabla_{\sigma_a,D}\mathscr{J}$ is given by

$$\nabla_{\sigma_a,D}\mathscr{J}(x) \;=\; \left([R^{'}(\sigma_a,D)]^{*}g\right)(x) \;=\; -\int_0^T \begin{pmatrix} \Phi z \\ \nabla\Phi\cdot\nabla z \end{pmatrix} dt. \tag{82}$$

Here $[R^{'}(\sigma_a,D)]^{*}$ denotes the adjoint of $[R^{'}(\sigma_a,D)]$ and z is solution of

$$-\frac{\partial z}{\partial t} + \sigma_a z - \nabla\cdot(D\nabla z) \;=\; 0, \tag{83}$$

with the final condition

$$z(x,T) \;=\; 0 \qquad \text{in } \Omega, \tag{84}$$

and mixed boundary condition

$$z(x,t) + 2D\frac{\partial z}{\partial\nu}(x,t) \;=\; g \quad \text{at } \partial\Omega\times[0,T]. \tag{85}$$

Notice that the adjoint diffusion equation mimics "back-diffusion" as an approximation to "back-transport" described by the adjoint RTE in (68)–(70). It looks backward in time the same way as the adjoint RTE does. The mixed boundary condition (85) is similar to (70) in this approximation. Notice also that (82) contains derivatives of u and z which require these quantities to be sufficiently smooth. These assumptions are usually well-approximated in large parts of the computational domain for situations with small mean free path. However, close to singular sources or initial conditions as well as close to the boundaries of the domain, the "true" photon distribution might not be that smooth such that sensitivities are estimated incorrectly in those regions. Also, the presence of clear (low-scattering) regions in the domain of interest renders these updates incorrect. For more details and some numerical results, see [5, 6, 27, 29, 56, 57].

5.4 Gradient for the Fokker-Planck Approximation

The gradient of the Fokker–Planck or δ-Eddington model is given as in (74) by replacing the scattering operator $\mathscr{L}$ by the corresponding approximation $\mathscr{L}_{FP}$ and $\mathscr{L}_{\delta E}$, respectively. For more details and numerical results, see [35].

5.5 Gradient by the Monte Carlo Method

As mentioned above, a MC-based approach would suggest directly the use of statistical estimation techniques for obtaining the corresponding parameters (probabilities) of absorption and scattering events. However, due to the direct link to concepts of the RTE, it is as well possible to proceed in a similar way as in the optimization problem of the RTE model. The adjoint RTE can be modeled by a so-called adjoint MC simulation which is well known from importance sampling in nuclear reactor theory [47]. When following this approach, advantages and disadvantages are similar as those seen when using the RTE approach. The computational cost of MC modeling might however become quickly prohibitive if no sophisticated computing techniques are employed (such as parallel computing or the use of graphics processors [2]). This approach has been taken for example in [18, 38].

6 Alternative Optimization Strategies

6.1 Shape and Topology Optimization

In some applications, additional prior knowledge about the parameter distributions of the RTE inside a medium is available in form of structural information. For example, in biomedical imaging, it might be known that regions of different tissue types (indicating organs, blood vessels, muscles, etc.) are separated by sharp interfaces and that both parameters of the RTE (or their approximations) change values at those interfaces and might be either slowly varying or constant inside each of these regions. In geophysical applications, these regions might indicate geological structures such as salt domes, ancient rivers, rock structures, etc. In those applications, it is possible to apply a region-based model to the inverse problem where the interfaces as well as some simple internal structures need to be reconstructed from the given data set. Such situations have been studied intensively during the last twenty years in the inverse problems community, see, for example, [13, 30, 31]. In these approaches, the forward model can be either one of the above described RTE approximations or the RTE itself. The gradient however will be replaced by a *shape derivative* (or *shape sensitivity*) that moves

an initial shape during the optimization process into a descent direction with respect to the chosen cost functional [25]. We refer for more details to the above mentioned references, and in particular for applications to the time-dependent RTE to [64]. Since additional prior information is implicitly used in those approaches, the reconstruction task will be better posed due to the corresponding regularization effect.

Practically, in these approaches the topology of the final shapes is unknown. When using only shape derivatives or shape sensitivities, it is still difficult to model the necessary changes in topology theoretically and computationally. Here the recently developed concept of a topological derivative or topological sensitivity provides a way of circumventing some of the related difficulties [14, 34]. This can be combined with the shape sensitivity analysis for improving results.

A practically convenient tool for modeling the shape evolution as part of these optimization approaches is the *level set technique* which has been originally proposed in [59] in the framework of computational physics and image processing and then was introduced into the inverse problems community in [68]. This technique allows for a straightforward implementation of both, shape evolution and topological changes, following the theory provided by the above mentioned concepts of shape and topology optimization. For more details and further references, see [13, 30, 31].

6.2 Sparsity-Promoting Reconstruction and Non-Differentiable Optimization

During the last ten years or so also sparsity-promoting regularization techniques have been developed in the literature that differ from the more classical gradient-based scheme in that the cost functional is not differentiable [21, 42]. Also the well-known total-variation regularization technique has been applied with good success to a variety of inverse problems [60, 77]. Often concepts from convex analysis are involved when solving those problems practically. Sparsity-promoting and total variation-based regularization have been applied to the time-dependent RTE for example in [63, 64]. We refer the reader to those publications for more details. So far it is not clear how well these novel regularization approaches will help to distinguish different distributed parameters from each other in the reconstructions. However, first numerical results look promising [20, 64].

7 Summary and Suggestions for Further Research

We have outlined several approaches for estimating two distributed parameters of the time-dependent RTE simultaneously from the same data set. Most current approaches make use of certain approximations of the RTE when practically solving

this task. However, the two parameters of the RTE, the absorption and the scattering cross section, occur in transformed form in those approximations, where also some specific characteristics (e.g., the mean scattering cosine) of the scattering phase function might be involved, which often is inaccurately known. The general structure of the gradients or sensitivities depends strongly on the forward model used. This has an impact on the task of separating the corresponding two parameters from each other in the reconstructions. So far it is not clear which approach is best suitable for practical applications. A trade-off between practicality, computational cost, physical or biological significance of individual parameters and mathematical complexity of the forward problem as well as the corresponding inversion strategies needs to be found here.

We also mention that extensions of the RTE model as outlined here to more general situations have as well been discussed recently in the literature which provide interesting paths for further research. For example, some structures in the domain of interest (such as muscle fibers in the human body or fractures and sedimentary layering in the earth) might suggest the presence of anisotropic effects in the individual parameters of the RTE such that they cannot be modeled as scalars anymore. Tensors need to be used and reconstructed instead. Also, the propagation of particles might not follow straight lines but curved paths, which renders the computational modeling more difficult. Sometimes more complicated matching conditions at interfaces between different physical regions in the domain of interest have to be taken into account. This includes mode conversion at interfaces in geophysical applications where a vector-version of the RTE is employed for modeling the propagation of elastic waves in the scattering earth, or a mismatch of refractive index between different optical regions in DOT. Also the scattering cross section η might be unknown or might follow more complicated physical laws than used here. A large variety of local and global optimization techniques can be applied to the resulting inverse problems, including statistical approaches that include the estimation of uncertainties. Finally, there are also applications where the unknown physical parameters might change at time scales which are in the order of magnitude of the measurement time, which renders the inverse problem time-dependent. All the above make this topic an interesting and broadly still open area of current and future research.

References

1. Aki, K. and Richards, P.G.: *Quantitative Seismology* 2nd edition, University Science Books (2009)
2. Alerstam E., Svensson T. and Andersson-Engels S.: Parallel computing with graphics processing units for high-speed Monte Carlo simulation of photon migration *J. Biomed. Opt.* **13** 060504 (2008).
3. Arianfar P. and Emamirad H.: Relation between scattering and albedo operators in linear transport theory, Transport Theory and Statistical Physics, 23:4, 517–531 (1994).
4. Arridge S.R., Optical tomography in medical imaging, Inverse Problems **15**, R41 (1999)

5. Arridge S.R. and Schotland J.C.: Optical Tomography: forward and inverse problems, *Inverse Problems* **25** (12) 123010 (59pp) (2009)
6. Arridge, S.: Methods in Diffuse Optical Imaging, *Phil. Trans. R. Soc.* A (2011) 369, 4558–4576
7. Bal G.: Inverse Transport Theory and Applications, *Inverse Problems* **25** (5) (2009)
8. Bal G. and Moscoso M.: Polarization Effects of Seismic Waves on the Basis of Radiative Transport Theory, *Geophys. J. Int.* **142** (2), pp 1639–1666 (2000).
9. Bell, G.I. - Glasstone, S.: *Nuclear Reactor Theory*, Van Nostrand-Reinhold, Raleigh, North Carolina, (1970).
10. Börgers, C., Larsen, E.W. and Adams, M.L.: The Asymptotic Diffusion Limit of a Discontinuous Linear Transport Equation, *J. Comp. Phys.* **98**, (1992), pp 285ff.
11. Börgers, C.: The radiation therapy planning problem, in *Computational Radiology and Imaging: Therapy and Diagnostics*, IMA Volumes in Mathematics and its Applications 110, C. Börgers and F. Natterer (eds.), 1–15 (1999)
12. Bondarenko, A.N.: Structure of singularities of the fundamental solution of the transport equation, *Dokl. Akad. Nauk SSSR* **322**, (1992), pp 274–276.
13. Burger M.: A level set method for inverse problems, *Inverse Problems* **17** (5) pp 1327–55 (2001)
14. Carpio, A. and Rapún, M.-L.: Solving inhomogeneous inverse problems by topological derivative methods, *Inverse Problems* **24** 045014 (32pp) (2008)
15. Case, K.M. and Zweifel, P.F.: *Linear Transport Theory*, Plenum Press, New York, (1967).
16. Cercignani, C.: *Theory and Applications of the Boltzmann Equation*, Elsevier, New York, (1975).
17. Chandrasekhar, S.: *Radiative Transfer*, Oxford Univ. Press, London, (1950). Also: Dover, New York, (1960).
18. Chen J. and Intes X.: Comparison of Monte Carlo methods for fluorescence molecular tomography-computational efficiency, *Med. Phys.* **38** (10) pp 5788–98 (2011)
19. Choulli, M. and Stefanov, P.: Inverse Scattering and Inverse Boundary Value Problems for the Linear Boltzmann Equation, *Comm. Part. Diff. Equ.* **21**(5&6), (1996), pp 763–785.
20. Cooper J.: Sparsity Regularization in Diffuse Optical Tomography, *PhD Thesis* Clemson University (2012)
21. Daubechies I., Defrise M. and De Mol C.: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, *Communications on Pure and Applied Mathematics* **57** (11) pp 1413–1457 (2004).
22. Dautray, R. and Lions, J.L.: *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol.5, Springer, Berlin, (1993).
23. Dautray, R. and Lions, J.L.: *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol.6, Springer, Berlin, (1993).
24. Davison, B.: *Neutron Transport Theory*, Oxford Univ. Press, London, (1957).
25. Delfour M.C. and Zolésio J. P.: *Shapes and Geometries: Metrics, Analysis, Differential Calculus, and Optimization* (2nd ed.) SIAM (2011)
26. Dierkes T., Dorn O., Natterer F., Palamodov V. and Sielschott H.: Fréchet Derivatives for some Bilinear Inverse Problems, *SIAM J. Appl. Math.* **62** (6) pp 2092–2113 (2002).
27. Dorn O.: Das inverse Transportproblem in der Lasertomographie, *PhD-thesis*, Westfälische Wilhelms Universität Münster, Germany (1997)
28. Dorn O.: A transport-backtransport method for optical tomography, *Inverse Problems* **14** pp 1107–1130 (1998)
29. Dorn O.: Scattering and absorption transport sensitivity functions for optical tomography *Optics Express* **7** (13) pp 492–506 (2000)
30. Dorn O. and Lesselier D.: Level set methods for inverse scattering, *Inverse Problems* **22** pp R67–R131 (2006)
31. Dorn O. and Lesselier D.: Level set methods for inverse scattering - some recent developments, *Inverse Problems* **25** 125001 (11pp) (2009)
32. Duderstadt, J.J. and Martin, W.R.: *Transport Theory*, Wiley, New York, (1979).

33. H. Emamirad and V. Protopopescu: Relationship between the albedo and scattering operators for the Boltzmann equation with semi-transparent boundary conditions, *Mathematical Methods in the Applied Sciences*, Vol 19, 1–13 (1996)
34. Feijóo, G.: A new method in inverse scattering based on the topological derivative, *Inverse Problems* **20** pp 1819–1840 (2004).
35. González-Rodríguez P. and Kim A.D., Comparison of light scattering models for diffuse otpical tomography, *Optics Express* **17** (11) pp 8756–8774 (2009)
36. Habetler, G.J. and Matkowsky, B.J.: Uniform asymptotic expansions in transport theory with small mean free paths, and the diffusion approximation, *Journ. Math. Phys.* **16** (4), (1975), pp 846ff.
37. Hammersley, J.M. and Handscomb, D.C.: *Monte Carlo Methods*, Methuen & Co LTD, London, (1965).
38. Hayakawa C.K., Spanier J. and Venugopalan, V.: Coupled Forward-adjoint Monte-Carlo Simulations of Radiative Transport for the Study of Optical Probe Design in Heterogeneous Tissues, *SIAM J. Appl. Math.* **68** (1) pp 253–270, (2007)
39. Hejtmanek, J.: Time-Dependent Linear Transport Theory, in: *Kinetic Theories and the Boltzmann Equation*, Lecture Notes in Mathematics **1048**, Eds. A.Dold, B. Eckmann, (1981).
40. Ishimaru, A.: *Wave Propagation and Scattering in Random Media*, 2 Vol., Academic Press, New York, (1978).
41. Jensen, H.W.: *Realistic Image Synthesis Using Photon Mapping*, AK Peters Publisher, ISBN 1568811470 (2001).
42. Jin B. and Maass P.: Sparsity regularization for parameter identification problems, *Inverse Problems* **28** (12) 123001 (70pp) (2012)
43. Kaltenbach, J.M. and Kaschke, M.: Frequency- and Time-Domain Modelling of Light transport in Random Media, in: *Medical Optical Tomography*, ed. Potter, SPIE Optical Engineering Press, Vol. IS11, (1993), pp 65–86.
44. Larsen, E.W. and Keller, J.B.: Asymptotic solution of neutron transport problems for small mean free paths, *J. Math. Phys.* **15** (1), (1974), pp 75ff.
45. Larsen, E.W.: Asymptotic Theory of the Linear Transport Equation for small mean free Paths, II, *Siam J. Appl. Math.* **33**, (1976), pp 427ff.
46. Lau C.W. and Watson K.M.: Radiation Transport along Curved Ray Paths, *J Math Phys* **11** pp 3125–37 (1970)
47. Lewins, J.: *Importance - The Adjoint Function*, Pergamon Press, Oxford, (1965).
48. Lewis, E.E. and Miller Jr., W.F.: *Computational Methods of Neutron Transport*, Wiley, New York, (1984).
49. Lions, J.L.: *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, Berlin Heidelberg, (1971).
50. Lux, I. and Koblinger, L.: *Monte Carlo Particle Transport Methods: Neutron and Photon Calculations*, CRC Press, (1991).
51. Margerin L., Campilo M. and Van Tiggelen B., Monte Carlo simulation of multiple scattering of elastic waves, *Journal of Geophyisical Research* **105** No B4, pp 7873–7892 (2000)
52. McCormick, N.J.: Recent Developments in Inverse Scattering Transport Methods, *Transp. Theory and Statist. Physics* **13** (1&2), (1984), pp 15–28.
53. McCormick, N.J.: Methods for solving Inverse Problems for Radiation Transport - An update, *Transp. Theory and Stat. Physics* **15** (6&7), (1986), pp 759–772.
54. McDowall, S.R.: Optical Tomography on Simple Riemannian Surfaces, *Communications in Partial Differential Equations* **30**, pp 1379–1400 (2005).
55. Natterer, F.: *The Mathematics of Computerized Tomography*, B.G. Teubner, Stuttgart, (1986).
56. Natterer, F.: Numerical Solution of Bilinear Inverse Problems, *Preprint* Westfälische Wilhelms Universität Münster, Germany (1995)
57. Natterer, F. and Wübbeling F.: *Mathematical Methods in Image Reconstruction*, SIAM, Philadelphia, (2001).
58. Nocedal J. and Wright S.: *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering (2006)

59. Osher S. and Sethian J.A.: Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations, *J. Comput. Phys.* **79** pp 12–49 (1988)
60. Osher, S., Burger, M., Goldfarb, D., Xu, J., and Yin, W.: An iterative regularization method for total variation-based image restoration *Multiscale Modeling and Simulation: A SIAM Interdisciplinary Journal* **4** (2) pp 460–489 (2005)
61. Papanicolaou, G.: Asymptotic analysis of transport processes, *Bulletin of the American Mathematical Society*, Vol 81 (2), (1975)
62. Pazy, A.: *Semigroups of Linear Operators and Applications to Partial Differential Equations* Springer Series 'Applied Mathematical Sciences' Vol 44 (1983).
63. Prieto Moreno K.E.: Novel mathematical techniques for structural inversion and image reconstruction in medical imaging governed by a transport equation, *PhD thesis*, The University of Manchester, UK (2015).
64. Prieto K. and Dorn O.: Sparsity and Level Set Regularization For Diffuse Optical Tomography Using a Transport Model in 2D, *Inverse Problems* **33** 014001 (28pp) (2017).
65. Reed, M. and Simon, B.: *Methods of Modern Mathematical Physics*, Vol. 3, Academic Press, New York, (1979).
66. Ryzhik L., Papanicolaou G. and Keller J.B.: Transport Equations for Elastic and Other Waves in Random Media, *Wave Motion* **24**, pp 327–370.
67. Sanchez, R. and McCormick, N.J.: A Review of Neutron Transport Approximations, *Nucl. Sci. Eng.***80**, (1981), pp 481–535.
68. Santosa F.: A level-set approach for inverse problems involving obstacles, *ESAIM: Control, Optimization and Calculus of Variations* **1** pp 17–33 (1996).
69. Sato, H. and Fehler, M.C. and Maeda, T.: *Seismic Wave Propagation and Scattering in the Heterogeneous Earth: Second Edition*, Springer, (2012).
70. Sheng P., *Introduction to wave scattering, localization and mesoscopic phenomena*, Springer, Berlin-Heidelberg, (2006)
71. Shreider, Yu.A.: *Method of Statistical Testing: Monte Carlo Method*, Elsevier Publishing Company, Amsterdam, (1964).
72. Schuster, A.: Radiation through a foggy atmosphere, *Astrophys. Journal* **21**, (1905).
73. Siewert, C.E.: *On the Singular Components of the Solution to the Searchlight Problem in Radiative Transfer* J. Quant. Radiat. Transfer, **33** (6), (1985), pp 551–554.
74. Sobolev, V.V.: *A Treatise on Radiative Transfer* (translated from Russian), Princeton, New Jersey, (1963).
75. Spanier, J. and Gelbard, E.M.: *Monte Carlo Principles and Neutron Transport Problems*, Addison-Wesley, Reading, (1969).
76. Van de Hulst, H.C.: *Multiple Light Scattering*, 2 Vol., Academic Press, New York, (1980).
77. Vogel, C. and Oman, M.: Fast, robust total variation-based reconstruction of noisy blurred images, *IEEE Transactions on Image Processing* **7**(6) pp 813–824 (1998)
78. Wang L. and Jacques S.L.: Hybrid model of Monte Carlo simulation diffusion theory for light reflectance by turbid media *J. Opt. Soc. Am. A* **10** pp 1746–52 (1993).
79. Weinberg A.M. and Wigner E.P.: *The Physical Theory of Neutron Chain Reactors*, University Chicago Press, (1958).
80. Wing, G.M.: *An Introduction to Transport Theory*, Wiley, New York, (1962).
81. Wright, S., Schweiger, M. and Arridge, S.R.: Reconstruction in optical tomography using the P_N approximations, *Meas. Sci. Technol.* **18** pp 79–86 (2007).

On the Use of Optimal Transport Distances for a PDE-Constrained Optimization Problem in Seismic Imaging

L. Métivier, A. Allain, R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux

Abstract Full waveform inversion is a PDE-constrained nonlinear least-squares problem dedicated to the estimation of the mechanical subsurface properties with high resolution. Since its introduction in the early 80s, a limitation of this method is related to the non-convexity of the misfit function which is minimized when the method is applied to the estimation of the subsurface wave velocities. Recently, the definition of an alternative misfit function based on an optimal transport distance has been proposed to mitigate this difficulty. In this study, we review the difficulties for exploiting standard optimal transport techniques for the comparison of seismic data. The main difficulty is related to the oscillatory nature of the seismic data, which requires to extend optimal transport to the transport of signed measures. We review three standard possible extensions relying on a decomposition of the data into its positive and negative part. We show how the two first might not be adapted for full waveform inversion and focus on the third one. We present a numerical strategy based on the dual formulation of a particular optimal transport distance yielding an efficient implementation. The interest of this approach is illustrated on the 2D benchmark Marmousi model.

L. Métivier (✉)
ISTerre/LJK, CNRS, Univ. Grenoble Alpes, Saint-Martin-d'Hères, France
e-mail: ludovic.metivier@univ-grenoble-alpes.fr

A. Allain · E. Oudet
LJK, Univ. Grenoble Alpes, Saint-Martin-d'Hères, France
e-mail: aude.allain@imag.fr; edouard.oudet@imag.fr

R. Brossier · J. Virieux
Univ. Grenoble Alpes, ISTerre, Grenoble, France
e-mail: romain.brossier@univ-grenoble-alpes.fr; jean.virieux@univ-grenoble-alpes.fr

Q. Mérigot
LMO, Univ. Paris Sud, Orsay, France
e-mail: quentin@mgrt.fr

H. Antil et al. (eds.), *Frontiers in PDE-Constrained Optimization*, The IMA Volumes in Mathematics and its Applications 163,
https://doi.org/10.1007/978-1-4939-8636-1_11

1 Full Waveform Inversion as a PDE-Constrained Nonlinear Optimization Problem

Full waveform inversion (FWI) is a high resolution seismic imaging technique which aims at reconstructing subsurface mechanical properties such as wave velocities, density, attenuation, or anisotropy parameters, from the recording of seismic waves at the surface. Compared to conventional tomography strategies, based on the interpretation of arrival times only, FWI should exploit the totality of the seismic signal, which is expected to provide higher resolution estimates of the subsurface parameters, in the limit of half the shortest wavelength of the propagated signal following the theory of diffraction tomography [12]. A recent review of FWI is proposed by Virieux et al. [43]. FWI is usually formulated as the minimization over the space of the subsurface parameters of the misfit between predicted and observed data. The predicted data is computed through the solution of partial differential equations (PDE) describing the seismic waves propagation. In the simplest settings, which we consider in this study, the acoustic approximation is adopted. Using this formalism, the problem is cast as the following PDE-constrained nonlinear optimization problem [18, 40]

$$\begin{cases} \min\limits_{v_P} J(v_P) = g(d_{cal}, d_{obs}) + \alpha R(v_P), \;\; v_P(x) \in \mathscr{C}^p(\Omega), \quad \Omega \subset \mathscr{R}^d \\ \dfrac{1}{\rho v_P(x)^2} \partial_{tt} u(x,t) - \text{ div}\left(\dfrac{1}{\rho(x)} \nabla u(x,t)\right) = s(x,t), \;\; (x,t) \in \Omega \times [0,T], \\ d_{cal}(x_r,t) = H(u)(x_r,t), \;\; (x_r,t) \in \Gamma \times [0,T]. \end{cases} \tag{1}$$

In the system (1), the spatial domain Ω is a subset of $\mathscr{R}^d$, where $d = 2$ or $d = 3$, while Γ denotes a subset of the border $\partial\Omega$. The time interval is defined by $[0, T]$, where $T > 0$. The control variable is denoted by $v_P(x)$: this is the pressure wave (P-wave) velocity, which is supposed to be smooth up to a certain level of regularity $p \in \mathscr{N}$. The P-wave velocity is generally the main parameter to be reconstructed, even if the density $\rho(x)$ can also be included in the inverse problem yielding a so-called multiparameter problem (see [32] for a review on multiparameter FWI). The functional $J(v_P)$ measures the misfit between predicted data $d_{cal}(x_r, t)$ and observed data $d_{obs}(x_r, t)$ through a misfit measurement function g which is often taken as the least-squares norm

$$g(d_{cal}, d_{obs}) = \frac{1}{2} \| d_{cal} - d_{obs} \|^2_{L^2}. \tag{2}$$

It shall be noted that this least-squares distance measure is local: each sample of the observed data is compared with its synthetic counterpart at the same position in the data space, neglecting any information which could come from the neighboring samples. As a result, the least-squares distance is unable to detect shifted patterns between two datasets.

A regularization term $R(v_p)$, weighted by a positive coefficient α, is also generally added to the misfit measurement to reduce the null space of the underlying inverse problem. Usual choices for $R(v_P)$ include prior information regularization, or penalization of the first-order spatial derivatives (Tikhonov regularization)

$$R(v_P) = \frac{1}{2}\|v_P - v_{P,0}\|_{L^2}^2, \quad R(v_P) = \sum_{i=1}^{d} \frac{1}{2}\|\partial x_i v_P\|_{L^2}^2. \tag{3}$$

The calculated data $d_{cal}(x_r, t)$ is computed from the solution $u(x, t)$ of the acoustic wave equation through the observation operator $H(u)$. In practice, this observation operator simply extracts the value of the wavefield $u(x, t)$ at the receivers' locations.

A Lagrangian function associated with the PDE-constrained problem (1) is

$$\begin{aligned} L(v_P, d_{cal}, u, \lambda_1, \lambda_2) = {} & g(d_{cal}, d_{obs}) + \alpha R(v_P) \\ & + \int_{x_r \in \Gamma} \int_0^T (d_{cal}(x_r, t) - Hu(x_r, t))\, \lambda_2(x_r, t) dx_r dt \\ & + \int_{x \in \Omega} \int_0^T \left(\frac{1}{\rho v_P^2} \partial_{tt} u(x, t) \right. \\ & \left. - \operatorname{div}\left(\frac{1}{\rho} \nabla u(x, t) \right) - s(x, t) \right) \lambda_1(x, t) dx dt \end{aligned} \tag{4}$$

First-order Karush-Kuhn-Tucker conditions give necessary conditions to characterize the solution of (1). They are obtained by canceling the first-order partial derivatives of the Lagrangian operator.

$$\begin{cases} -\dfrac{2}{\rho v_P^3} \displaystyle\int_0^T \partial_{tt} u(x, t) \lambda_1(x, t) dt + \alpha \nabla R(v_P) = 0 & (5) \\ d_{cal} = H(u) & (6) \\ \dfrac{1}{\rho v_P^2} \partial_{tt} u - \operatorname{div}\left(\dfrac{1}{\rho} \nabla u \right) = s & (7) \\ \lambda_2 = -\partial_{d_{cal}} g(d_{cal}, d_{obs}) & (8) \\ \partial_{tt} \lambda_1 - \rho v_P^2 \operatorname{div}\left(\dfrac{1}{\rho} \nabla \lambda \right) = -\partial_u H(u) \lambda_2 & (9) \end{cases}$$

Instead of solving the Karush-Kuhn-Tucker system iteratively through a Newton algorithm, a "reduced space" method is preferred [31] for efficiency. The misfit function $J(v_P)$ is minimized following iterative local optimization methods for smooth nonlinear functions, which rely on the ability to compute its gradient $\nabla J(v_P)$. This gradient is computed from the equation

$$\nabla J(v_P) = -\frac{2}{\rho v_P^3} \int_0^T \partial_{tt} \overline{u}(x,t) \overline{\lambda}_1(x,t) dt + \alpha \nabla R(v_P), \tag{10}$$

where fields $\overline{u}(x,t)$ and $\overline{\lambda}_1(x,t)$ are obtained through the solution of the Equations from (6) to (9). In particular, using the L^2 norm for the definition of the misfit measurement function g yields the simple expression

$$\lambda_2 = -(d_{cal} - d_{obs}). \tag{11}$$

The reduced space method thus yields an efficient strategy to compute the gradient $\nabla J(v_P)$. This technique, also introduced as the adjoint-state method within the optimal control theory [21], has been known for a long time in seismic imaging [9] and in weather forecasting [19]. A review of the adjoint-state method and its application in seismic imaging has been proposed by Plessix [34].

Among different minimization strategies, the nonlinear conjugate gradient method, the quasi-Newton l-BFGS [30], or the truncated Newton approach [29] are used to solve the FWI problem (see [25] for a review of standard minimization algorithms used in FWI).

Since its introduction in the 80's, one of the main challenges for FWI is related to the non-convexity of the P-wave velocity reconstruction problem. For practical applications, the size of the discrete problem prevents the use of global or semi-global optimization strategies (Monte-Carlo or genetic algorithms, for instance): in 2D, the number of unknowns easily reaches $O(10^6)$, in 3D this number grows up to $O(10^9)$. The use of local optimization strategies thus requires to start the iterative process from a v_P model close enough from the solution, otherwise the method converges to a local minimum. Wave physics analysis provides useful information to better assess what are the requirements that an initial model should satisfy to ensure the convergence toward the global minimum.

The non-convexity of the misfit function with respect to the P-wave velocity is related to the choice of the function $g(d_{cal}, d_{obs})$ to measure the discrepancy between observed and calculated data. Seismic observations are in essence oscillatory signals. Macroscale P-wave velocity perturbation mainly affects the seismic data by modifying the propagation time rather than the amplitude of the seismic events [16]. As a result, observed and calculated data mainly differ through time-shifts of the different seismic arrivals. The function $g(d_{cal}, d_{obs})$ should thus be convex with respect to these time-shifts. This is not the case for the L^2 distance which is used in practice. This is illustrated in Figure 1 where the seismic data is schematically represented as a periodic sinusoidal signal. When the signals are shifted by a multiple of one period of the signal, the L^2 differences between the signals reach a local minimum: this is what is referred to a cycle skipping, or phase ambiguity problem, in the FWI community. Avoiding these local minima thus requires to start the minimization from less than half-a-phase shift. In other words, the initial velocity model should be sufficiently accurate to predict the kinematic of the wave propagation up to half-a-phase shift.

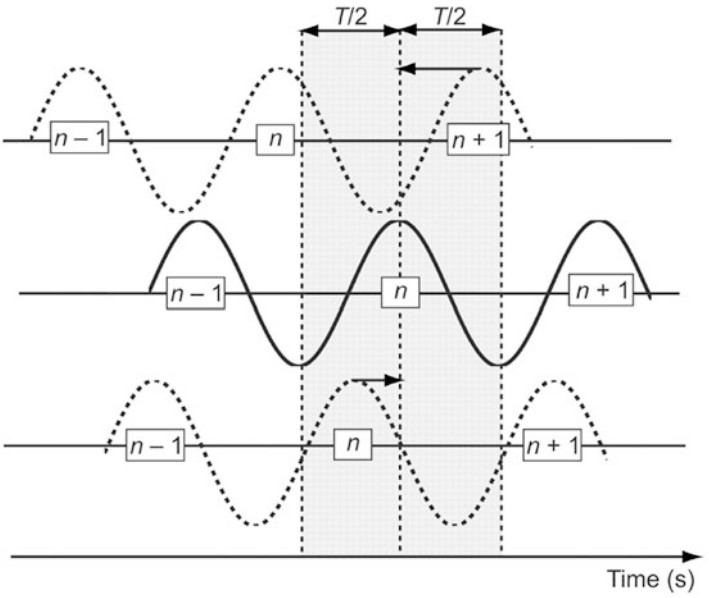

Fig. 1 Schematic example of the cycle skipping/phase ambiguity issue on sinusoidal signals. As soon as the initial shift is larger than half a period of the signal, the fit of the signal using a least-squares distance is performed up to one or several phase shifts. One may try to fit the $n+1$ dashed wriggle of the top signal with the n continuous wriggle of the middle signal moving to the wrong direction. The bottom dashed signal predicts the n wriggle in less than half-period leading to a correct updating direction (figure from [44])

Mitigating this non-convexity has been the aim of numerous methods proposed during the past decades. Three main lines of investigation have been followed. The first one relies on the design of hierarchical schemes. The data is interpreted through a sequence of FWI problems, the estimation obtained from the problem i being used as an initial guess for the problem $i+1$. For each FWI problem, only a subset of the data is interpreted. The usual data decomposition is performed in the frequency domain: the data is interpreted from low-to-high frequencies. Low frequency components of the signal have a larger period, therefore the requirement on the initial model to fit the observed data within half-a-period of the signal is partially relaxed. Additional level of hierarchy can also be applied (time-windowing and offset selection, for instance) following layer stripping approaches [8, 35, 37]. The second line of investigation is based on the modification of the misfit measurement function $g(d_{cal}, d_{obs})$. Cross-correlation functions have been first investigated [23], and later on warping techniques [15], deconvolution approaches [22, 45] as well as envelope and phase separation [6, 14]. The third line of investigation relies on probing the consistency of the velocity model by building reflectivity images using different subset of the data. The velocity is updated such that the different reflectivity images become similar (see [39] and references therein for a review). These methods are known as (extended) image-domain techniques.

None of these approaches has completely overcome the cycle skipping or phase ambiguity problem. Hierarchical approaches relax the constraint on the accuracy of the initial velocity model by working first at low frequencies; however, this strategy is limited by the lowest available frequency, which is most of the time not low enough to sufficiently constrain the model. The different modifications of the misfit function proposed so far also enables to start from an initial velocity model further away from the solution; however, this is often at the expense of the resolution of the final estimation. Image-domain techniques also exhibit interesting properties in terms of convexity of the misfit function; however, the computation cost associated with the repeated computation of reflectivity images seems to preclude their use to large-scale datasets, especially in 3D configuration.

In this study, we discuss how optimal transport distances could be used to define an alternative misfit function measurement g in the framework of FWI. In particular, these distances provide natural tools to go beyond the point-to-point comparison underlaid by the least-squares distance by performing global comparison. The field of optimal transport has been very active in the last years, as testified by the number of textbooks published on this topic recently [2, 36, 41, 42]. Recent applications in image processing demonstrate the interest of optimal transport distance to compare images, notably for its ability to detect shifted patterns from one image to another [20]. We discuss what are the main difficulties when applying optimal transport distance for the comparison of seismic data. In particular, we show that the oscillatory nature of the seismic data requires to extend optimal transport to the comparison of signed measures, which is a nontrivial problem. We review three different propositions found in the literature relying on the decomposition of the data in its positive and negative part. We show how the two first options might not be adapted for full waveform inversion. We thus focus on the third possibility and show how an efficient implementation can be obtained, as we have presented it in previous studies [26, 28]. We present numerical results obtained on the 2D Marmousi case study, a benchmark in the seismic imaging community, which illustrate the interest of this approach.

In Section 2, we discuss the optimal transport problem formulation for positive measures and present a state-of-the-art for its extension to the comparison of signed measures. In Section 3, we present the alternative strategy we have promoted in previous studies and its application to the 2D Marmousi case study. Conclusion and perspectives are given in Section 4.

2 Optimal Transport for Full Waveform Inversion

2.1 Basics on Optimal Transport

Optimal transport has its roots in the work of a French scientist named Gaspard Monge, in an attempt to devise the best strategy to move piles of sand to a building

site. The aim was to minimize the volume of the sand to be displaced as well as the distance on which it had to be displaced. In modern mathematics, an expression of this problem is the following. Consider two probability measures $\mu \in \mathscr{P}(X)$ and $\nu \in \mathscr{P}(Y)$ (μ would represent the initial configuration of sand and ν the targeted one). We consider the mapping $T(x)$ from X to Y such that

$$\begin{cases} X & \longrightarrow & Y \\ T : x & \longrightarrow & T(x), \end{cases} \tag{12}$$

The push forward distribution of μ through the mapping T is denoted by $T_{\#}\mu$, such that for any measurable subset $A \subset Y$, we have

$$(T_{\#}\mu)(A) \equiv \mu\left(T^{-1}(A)\right) = \nu(A). \tag{13}$$

In this framework, the original Monge problem is formulated as

$$\inf_{T} \left\{ \int_X ||x - T(x)|| d\mu(x), \quad T_{\#}\mu = \nu \right\}. \tag{14}$$

This problem has not necessarily a solution, and when the solution exists, it is difficult to compute because of the nonlinear constraint $T_{\#}\mu = \nu$.

A relaxation of this problem has been proposed by Kantorovich [17], under the form

$$\inf_{\gamma} \left\{ \int_{X\times Y} c(x, y) d\gamma(x, y), \quad \gamma \in \Pi(\mu, \nu) \right\}, \tag{15}$$

where the ensemble of transport plans $\Pi(\mu, \nu)$ is defined by

$$\Pi(\mu, \nu) = \{\gamma \in \mathscr{P}(X \times Y), \quad (\pi_X)_{\#}\gamma = \mu, \quad (\pi_Y)_{\#}\gamma = \nu\}. \tag{16}$$

The operators π_X and π_Y are the projectors on X and Y, respectively. This relaxation is the cornerstone of modern application of optimal transport as the problem (15) has always a solution which coincides with the one of the original Monge problems when this one exists. The problem (15) generalizes (14) in the sense that, instead of considering a mapping T transporting each particle of the distribution μ to the distribution ν, it considers all pairs (x, y) of the space $X \times Y$ and for each pair defines how many particles of μ go from x to y.

In discrete form, the Kantorovich problem becomes a linear programming problem of the form

$$\min_{\gamma_{ij}} \sum_{ij} \gamma_{ij} c_{ij}, \quad \gamma \in \Pi(\mu, \nu) \tag{17}$$

where

$$\Pi(\mu,\nu) = \{\gamma \geq 0, \sum_{j=1} \gamma_{ij} = \mu_i, \quad \sum_{i=1} \gamma_{ij} = \nu_j\} \tag{18}$$

The entry γ_{ij} represents how much mass should be moved from x_i to y_j while c_{ij} measures the distance between x_i to y_j. The constraint ensures that the initial distribution is equal to μ while the transported distribution through the transport plan γ is equal to ν.

Of particular interest, optimal transport induces distances between distribution, named as Wasserstein distances or earth mover's distances (EMD). They are defined by

$$W_p(\mu,\nu) = \left(\min_{\gamma_{ij}} \sum_{ij} \gamma_{ij} \|x_i - y_j\|^p, \quad \gamma \in \Pi(\mu,\nu) \right)^{1/p} \tag{19}$$

One interest for using such distance for signal processing applications is their ability to detect shifted pattern from one signal to another. This property is also referred to in the literature as the fact that W_p distances should be seen as "horizontal distances" while L^p distances should be seen as "vertical distances" [36]. The W_p distance between two shifted probability distributions is convex with respect to this shift, while the L^p distance is insensitive to this shift.

2.2 *Applying Optimal Transport for the Comparison of Seismic Data: The Difficulty of Transporting Signed Measures*

The existence of a solution to the optimal transport problem (16) depends on two assumptions that shall be satisfied by the measures μ and ν

1. μ and ν shall be positive
2. μ and ν shall have the same total mass

$$\int_X d\mu(x) = \int_X d\nu(x). \tag{20}$$

In this section, for the sake of simplicity, we assume that the two measures μ and ν are defined on the same space X . This is the case when μ and ν represent seismic data. Seismic data do not satisfy the positivity requirement due to its oscillatory nature. However, the zero frequency component of each seismic trace is zero

$$\forall x_r, \quad \int_0^T d_{cal}(x_r,t)dt = \int_0^T d_{obs}(x_r,t)dt = 0. \tag{21}$$

Therefore, we have

$$\int_{x_r}\int_0^T d_{cal}(x_r,t)dtdx_r = \int_{x_r}\int_0^T d_{obs}(x_r,t)dtdx_r = 0. \tag{22}$$

Thus, interpreting seismic data as density functions, Equation (22) shows that the seismic data satisfy the second assumption: observed and calculated data have the same total mass, which is zero.

The main difficulty to apply optimal transport to the comparison of seismic data thus relies on the non-positivity of the seismic data. This is a well-identified issue in the optimal transport community. The question how to extend optimal transport to signed measures is investigated in particular by Ambrosio et al. [2] and Mainini [24]. Mainini makes use of the following Jordan-Hahn decomposition,

$$\mu = \mu^+ - \mu^-, \tag{23}$$

where μ^+ (respectively, μ^-) is the positive part of μ (respectively, the negative part of μ). Three strategies are reviewed in [24] to extend optimal transport to signed measures. The corresponding extension of the W_p distances to signed measures is introduced as $W_{p,i}(\mu,\nu)$, $i = 1,2,3$ in the following. The three strategies proposed by Mainini are

1. Transport separately the positive and negative part of the measures

$$W_{p,1}(\mu,\nu) = W_p(\mu^+,\nu^+) + W_p(\mu^-,\nu^-) \tag{24}$$

2. Transport the absolute value of the measures

$$W_{p,2}(\mu,\nu) = W_p(|\mu|,|\nu|) \tag{25}$$

3. Perform the decomposition

$$W_{p,3}(\mu,\nu) = W_p(\mu^+ + \nu^-, \nu^+ + \mu^-) \tag{26}$$

The first strategy, which might appear as the more intuitive, is the one proposed originally by Engquist and Froese [13]. It is successfully applied to the comparison of two time-shifted Ricker functions. The function $W_{2,1}^2(\mu,\nu)$ exhibits a quadratic convexity with respect to the time-shift between the two Ricker functions (Figure 2). Two drawbacks can nonetheless be identified. First, the mass conservation between positive and negative parts of the measure is not ensured. Second, there is no obvious reason that the positive and negative parts of the seismic data should be uncorrelated. For realistic application, the source wavelet $s(x,t)$ is not known, and a prior source estimation is required to perform FWI. Hence, we can expect this decomposition to be strongly sensitive to errors in this source wavelet estimation.

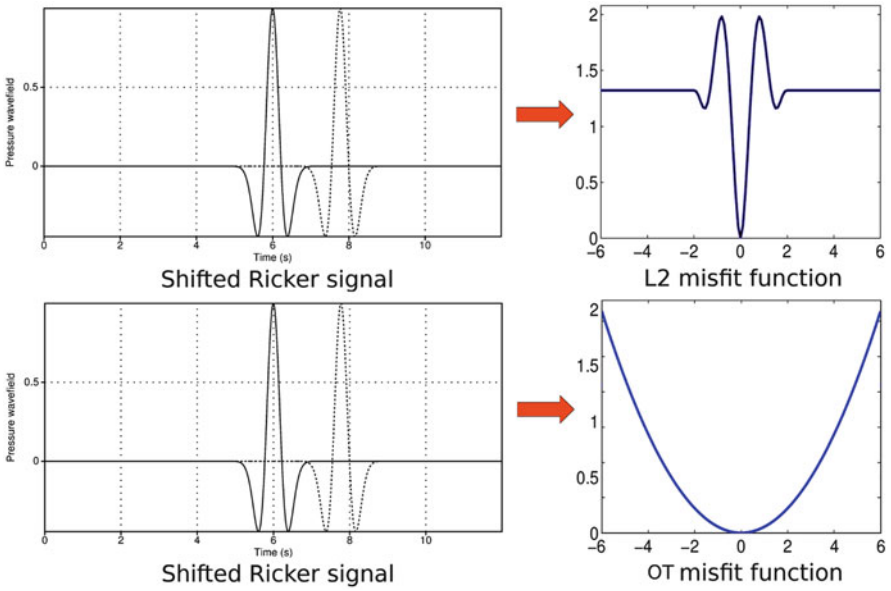

Fig. 2 Computation of the misfit function between two time-shifted Ricker signals depending on the time-shift, using a least-squares distance and an optimal transport distance. While the least-squares distance yields a non-convex misfit function with two local minima aside the global minimum at zero time-shift, the optimal transport distance yields a perfectly convex misfit function [13]

The second strategy is straightforward to apply; however, the mass conservation between $|\mu|$ and $|\nu|$ is also not ensured. In addition, FWI misfit functions relying on the absolute value of the data lose the sensitivity to the polarity of the signal. As a result, positive or negative impedance contrasts cannot be distinguished. This prevents from the correct interpretation of reflected waves.

The third strategy comes from the decomposition

$$\mu - \nu = \left(\mu^+ + \nu^-\right) - \left(\nu^+ + \mu^-\right). \tag{27}$$

This method seems appealing as, for any μ and ν satisfying the mass conservation assumption, one has

$$\int_X d\mu^+ - d\mu^-(x) = \int_X d\nu^+(x) - d\nu^-(x), \tag{28}$$

therefore

$$\int_X d\mu^+ + d\nu^-(x) = \int_X d\nu^+(x) + d\mu^-(x), \tag{29}$$

and the mass conservation is ensured for the distance $W_{p,3}$.

We thus see that the mass conservation assumption is not satisfied in the definition of $W_{p,1}$, $W_{p,2}$. This might not be a shortcoming as severe as the one associated with the transport of signed measures as several possibilities exist to extend optimal transport to situation where the mass conservation is not ensured, known as partial optimal transport. However, the correlation between the negative and positive part of the seismic data is not accounted for using $W_{p,1}$. The sensitivity to the polarity of the seismic data is lost using $W_{p,2}$. These two drawbacks are severe. On the other hand, $W_{p,3}$ is based on a formulation for which the mass conservation is ensured and only positive measures are compared. For this reason, we are interested in investigating the use of this strategy for FWI.

2.3 A Strategy Using the W_1 Distance in Its Dual Form

2.3.1 Link Between the Dual W_1 Distance and the Mainini Decomposition

As the size of seismic data easily reaches several millions of discrete parameters for realistic FWI applications, we need to design a numerical strategy for large-scale optimal transport problem with at most quasi-linear complexity.

Standard approaches for fast optimal transport computation encompass

- the direct solution of the Monge-Ampère equations [33]
- the solution of a fluid dynamic problem following the Benamou-Brenier formulation [3]
- the solution of a regularized optimal transport problem following an entropic regularization strategy [4, 11]

The last of this strategy can be applied for the computation of general W_p distances, while the two first strategies are dedicated to the computation of the W_2 distance.

Instead of relying on these developments, we rather propose another fast optimal transport computation technique, dedicated to a particular instance of the W_1 distance. The reason we focus on the W_1 distance is related to the Mainini technique, described in the previous paragraph, we want to apply. We explain it in the following.

The very important duality result due to [17] states that the Kantorovich optimal transport problem (16) is equivalent to the maximization problem

$$\max_{\varphi,\psi} \int_X \varphi(x) d\mu(x) + \int_X \psi(x) d\nu(x), \quad \varphi(x) + \psi(x') \leq c(x, x'). \tag{30}$$

In the particular case of the W_1 distance, the dual problem (30) can be expressed using a single potential function $\varphi(x)$ as

$$\max_{\varphi \in \mathrm{Lip}_1(X)} \int_X \varphi(x) d(\mu - \nu)(x), \tag{31}$$

where the space of 1-Lipschitz function over X is denoted by $\mathrm{Lip}_1(X)$. This simplification comes from the fact that for W_1, we have

$$c(x, y) = |x - y| \tag{32}$$

which is itself a distance over $X \times X$ (see [36] for a complete proof). Note that this is not the case for W_p distances with $p > 1$.

Interestingly, using this duality result, we see that

$$\begin{aligned} W_{1,3}(\mu, \nu) &= W_1(\mu^+ + \nu^-, \nu^+ + \mu^-) \\ &= \max_{\varphi \in \mathrm{Lip}_1(X)} \int_X \varphi(x) d(\mu^+ + \nu^- - \nu^+ + \mu^-)(x), \\ &= \max_{\varphi \in \mathrm{Lip}_1(X)} \int_X \varphi(x) d(\mu - \nu)(x) \\ &= W_1(\mu, \nu) \end{aligned} \tag{33}$$

This equality is important, as it reveals that through its particular dual formulation, the distance W_1 (31) can be computed for signed measures satisfying the mass conservation assumption (22). Indeed, as it is mentioned in [20] and [5, 8.10.viii], the problem

$$\max_{\varphi \in Lip(X)} \int_X \varphi_x d\mu(x), \tag{34}$$

defines the norm $\|\mu\|^*_{KR}$ on the space of signed measures with first-order moment equal to zero

$$\int_X d\mu(x) = 0. \tag{35}$$

We have mentioned that for seismic data, the measure $\mu - \nu$ satisfies (35), therefore we have

$$\left\{ \max_{\varphi \in Lip_1(X)} \int_X \varphi(x) d(\mu - \nu)(x), \right\} = \|\mu - \nu\|^*_{KR} \tag{36}$$

In addition, this shows that the Mainini decomposition is directly embedded in the dual formulation of W_1 as soon as signed measures are involved.

This has the following important advantage for our application: there is no need to numerically perform the Jordan-Han decomposition into positive and negative part of the data to compute our misfit function. This could be problematic as we minimize this misfit function through local optimization strategies for differentiable functions, relying on the gradient and the Hessian of this function. As the Jordan-Han decomposition is not differentiable (by definition), the resulting misfit function would not be differentiable, and we would need to use optimization strategies for non-smooth misfit functions.

Note that in the case the mass conservation assumption is not satisfied, the norm $\|.\|^*_{KR}$ can be easily extended to the Kantorovich-Rubinstein norm, defined by

$$\|\mu-\nu\|_{KR} = \left\{ \max_{\varphi} \int_X \varphi(x) d(\mu-\nu)(x), \quad \varphi(x) \in Lip_1(X), \|\varphi\|_\infty < 1 \right\} \tag{37}$$

This problem admits a solution even in the case $\mu - \nu$ does not satisfy (35). It might be more flexible to use for realistic application as the mass conservation is satisfied only at machine precision, which might occur instabilities when using the formulation (31).

In a series of articles [26–28], we have investigated the use of this Kantorovich-Rubinstein norm for realistic FWI applications. In the following, we summarize the numerical method developed in these studies to compute this norm.

2.3.2 Numerical Method

We consider in the following the computation of the Kantorovich-Rubinstein norm for $d_{obs}(x_r, t) - d_{cal}(x_r, t)$. In discrete form, this is equivalent to the solution of the problem

$$\begin{aligned} &\max_{\varphi_{rn}} \sum_{r=1}^{N_r} \sum_{n=1}^{N_t} \varphi_{rn} \left((d_{obs})_{rn} - (d_{cal})_{rn} \right), \\ &\forall r, n, r', n' \quad |\varphi_{rn} - \varphi_{r'n'}| \leq \|(x_r, t_n) - (x'_r, t'_n)\|, \\ &\forall r, n, \quad |\varphi_{rn}| \leq 1 \end{aligned} \tag{38}$$

where the integer r is the index associated with the receiver coordinate x_r and the integer n is the index associated with the time coordinate t.

We denote by $N = N_r \times N_t$ the total number of discrete samples associated with one dataset. In this framework, the computation of the Kantorovich-Rubinstein norm is a linear programming problem with $O(N)$ unknowns and $O(N^2)$ constraints. For realistic application, N easily reaches $O(10^6)$, already for 2D problems. It is therefore important to reduce the number of constraints of the problem to reach feasible complexity algorithms.

With this purpose, we focus on the particular case where, instead of the Euclidean distance $\|.\|$, we use the ℓ_1 distance we denote by $|.|$ to measure the distance between (x_r, t_n) and (x'_r, t'_n). In [28], we show that satisfying the N^2 constraints

$$\forall r, n, r', n' \quad |\varphi_{rn} - \varphi_{r'n'}| \leq |(x_r, t_n) - (x'_r, t'_n)| = |x_r - x'_r| + |t_n - t'_n| \tag{39}$$

is equivalent to satisfying the $2N$ constraints

$$\forall r, n \quad |\varphi_{rn} - \varphi_{r+1,n}| \leq |x_r - x_{r+1}| \quad |\varphi_{rn} - \varphi_{r,n+1}| \leq |t_n - t_{n+1}| \tag{40}$$

This is simply due to the "Manhattan" property of the ℓ_1 norm. This yields the following ℓ_1 Kantorovich-Rubinstein problem

$$\begin{aligned}
&\max_{\varphi_{rn}} \sum_{r=1}^{N_r} \sum_{n=1}^{N_t} \varphi_{rn} \left((d_{obs})_{rn} - (d_{cal})_{rn} \right), \quad \forall r, n \\
&|\varphi_{rn} - \varphi_{r+1,n}| \leq |x_r - x_{r+1}| \\
&|\varphi_{rn} - \varphi_{r,n+1}| \leq |t_n - t_{n+1}| \\
&|\varphi_{rn}| \leq 1
\end{aligned} \tag{41}$$

which is a linear programming problem with $O(N)$ unknowns and $O(N)$ constraints.

In [28], we have detailed how this problem can be recast as the convex non-smooth optimization problem

$$\max_{\varphi} f_1(\varphi) + f_2(A\varphi), \tag{42}$$

where

$$f_1(\varphi) = \sum_{r=1}^{N_r} \sum_{n=1}^{N_t} \varphi_{rn} \left((d_{obs})_{rn} - (d_{cal})_{rn} \right), \quad f_2(\psi) = i_K(\psi). \tag{43}$$

The function i_K is the indicator function on the unit hypercube K such that

$$i_K(x) = \begin{vmatrix} 0 & \text{if } x \in K \\ +\infty & \text{if } x \notin K, \end{vmatrix} \tag{44}$$

The operator A is the rectangular real matrix

$$A = \begin{bmatrix} D_{x_r} & D_t & I_N \end{bmatrix}^T, \tag{45}$$

where I_N is the real identity matrix of size N and $D_{x_r}, \quad D_t$ are the forward finite-difference operators

$$\begin{cases} \left(D_{x_r} \varphi \right)_{rn} = \dfrac{\varphi_{r+1,n} - \varphi_{rn}}{\Delta x_r}, \\ \left(D_t \varphi \right)_{rn} = \dfrac{\varphi_{r,n+1} - \varphi_{rn}}{\Delta t}, \end{cases} \tag{46}$$

Efficient strategies based on proximal splitting can be used to solve problems such as (42), where the functions f_i might not be differentiable. Among several

$\gamma > 0$, $y_1^0 = 0$, $y_2^0 = 0$, $z_1^0 = 0$, $z_2^0 = 0$;
for $n = 0, 1, \ldots$ **do**
  $\varphi^k = \left(I_N + A^T A\right)^{-1} \left[\left(y_1^k - z_1^k\right) + A^T \left(y_2^k - z_2^k\right)\right]$;
  $y_1^{k+1} = \text{prox}_{\gamma f_1}\left(\varphi^k + z_1^k\right)$;
  $z_1^{k+1} = z_1^k + \varphi^k - y_1^{k+1}$;
  $y_2^{k+1} = \text{prox}_{\gamma i_K}\left(A\varphi^k + z_2^k\right)$;
  $z_2^{k+1} = z_2^k + A\varphi^k - y_2^{k+1}$;
end

Algorithm 1: SDMM method for the solution of the problem (42)

possibilities, we choose the simultaneous direction method of multipliers (SDMM), which is well described in [10], for its good convergence properties. The method can be summarized as the Algorithm 1. The proximity operator can be seen as the generalization of the convex projection operator. For a given function f, it is defined as

$$\text{prox}_f(x) = \arg\min_y f(y) + \frac{1}{2}\|x - y\|_2^2, \tag{47}$$

For the particular case of the function f_1 and f_2, closed-form formulations exist

$$\text{prox}_{\gamma f_1}(\varphi) = \varphi - \gamma(d_{obs} - d_{cal}), \tag{48}$$

$$\forall i = 1, \ldots, P, \quad \left(\text{prox}_{\gamma f_2}(x)\right)_i = \left(\text{prox}_{i_K}(x)\right)_i = \begin{cases} x_i & \text{if } -1 \leq x_i \leq 1 \\ 1 & \text{if } x_i > 1 \\ -1 & \text{if } x_i < -1. \end{cases} \tag{49}$$

The closed-form formulations (48) and (49) are inexpensive to compute with an overall complexity in $O(N)$ operations. However, the SDMM algorithm requires the solution of a linear system involving the matrix $I + A^T A$. In [28], we show that the matrix $A^T A$ is a second-order finite-difference discretization of the Laplacian operator with homogeneous Neumann boundary conditions. Therefore, these linear systems can be solved in $O(NlogN)$ complexity using fast Fourier transform-based algorithms [38], or in $O(N)$ complexity using multigrid strategies [1, 7]. The combination of the reduction of the number of constraints using the property of the ℓ_1 distance and the observation that the matrix $I + A^T A$ appearing in the SDMM strategy actually corresponds to the discretization of the Poisson's equation offers the possibility to design an efficient numerical method to compute the ℓ_1 Kantorovich-Rubinstein norm for large-scale problems.

Following the notations used in Section 1, the use of the ℓ_1 Kantorovich-Rubinstein as the misfit measurement function for FWI implies that

$$g(d_{obs}, d_{cal}) = \|d_{cal} - d_{obs}\|_{KR} \tag{50}$$

The computation of the gradient of the resulting misfit function only requires the definition of the source of the adjoint field $\lambda_1(x,t)$ through

$$\frac{\partial \|d_{cal} - d_{obs}\|_{KR}}{\partial d_{cal}} \tag{51}$$

Interestingly, following the definition of $\|d_{cal} - d_{obs}\|_{KR}$, if we denote by $\overline{\varphi}$ the solution of the maximization problem (42), we have

$$\frac{\partial \|d_{cal} - d_{obs}\|_{KR}}{\partial d_{cal}} = \overline{\varphi} \tag{52}$$

As a consequence, the computation of the solution to the problem (42) yields simultaneously the value of the misfit function, through the value of the criterion at the maximum, as well as the quantity $\overline{\varphi}$ required to compute the gradient of the misfit function through the adjoint-state approach. The solution of a single optimal transport problem per seismic source is thus required at each iteration of FWI.

3 Example of Application of the Kantorovich-Rubinstein Norm to FWI

In order to illustrate the property of the Kantorovich-Rubinstein norm for the interpretation of seismic data, we first reproduce the experiment proposed in [13] where the distance between shifted in time Ricker signal is computed using the L^2 distance and the W_2 distance applied to the positive and negative part of the Ricker separately. Here, instead of the W_2 distance, we compute directly the Kantorovich-Rubinstein distance without separating positive and negative parts of the signal. The results are presented in Figure 3. Compared to the least-squares distance, a single minimum is recovered. However, the convexity of the misfit function with respect to the time-shift is lost. The loss of convexity is due to the signed nature of the Ricker signal (presence of negative values). One could expect optimal transport to be able to detect that the same pattern is shifted when comparing the Ricker, and that the W_1 distance would be proportional to this shift. This is not the case, which results from the presence of negative values. However, an important feature is preserved, with respect to the L^2 distance: a single minimum is obtained, while the L^2 distance displays two local minima aside the global minimum. This prompts us to test the use of the Kantorovich-Rubinstein norm to a more realistic case study.

To this purpose, we consider the Marmousi model presented in Figure 4(a). A synthetic dataset is computed in the 2D acoustic constant-density approximation. A fixed-spread surface acquisition is used, with 128 sources each 125 m and 168 receivers each 100 m, at 50 m depth. A Ricker source function centered on 5 Hz is used to generate the synthetic dataset. The frequency content of the source is high-pass filtered below 3 Hz to mimic realistic seismic data. In practical application,

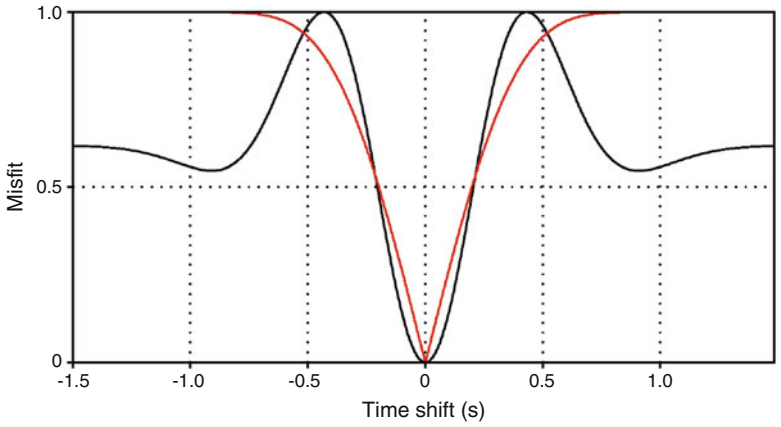

Fig. 3 Computation of the misfit function between two time-shifted Ricker signals depending on the time-shift, using a least-squares distance (black) and the Kantorovich-Rubinstein distance (red). We recover a single minimum; however, compared to the optimal transport distance used by Engquist and Froese [13], the convexity of the misfit function is lost

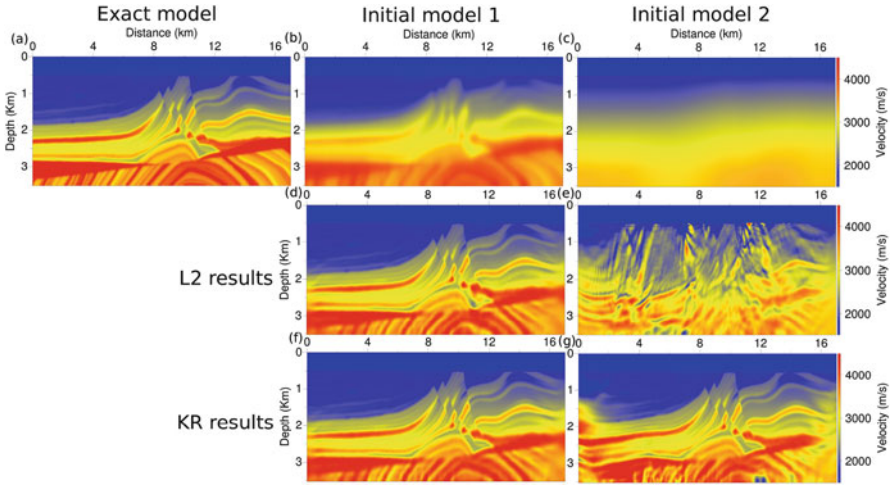

Fig. 4 Marmousi model case study. Exact model (a), initial model 1 (b), initial model 2 (c), results obtained with the L^2 distance starting from model 1 (d), from model 2 (e), results obtained with the ℓ_1 Kantorovich-Rubinstein distance starting from model 1 (f), from model 2 (g)

this frequency band is contaminated by noise, and therefore filtered out before inversion. Two initial P-wave velocity models are considered: the first contains the main features of the exact model, only with smoother interfaces (Figure 4(b)). The second is a strongly smoothed version of the exact model with very weak lateral variations and underestimated growth of the velocity in depth (Figure 4(c)). Starting from these two initial models, we compare the FWI results obtained

using a least-squares distance and the ℓ_1 Kantorovich-Rubinstein distance. The minimization is performed using the l-BFGS algorithm [30] implemented in the SEISCOPE optimization toolbox [25].

These results are presented in Figure 4(d–g). Starting from the first initial model, a correct estimation of the P-wave velocity model is obtained, using both the L^2 distance (Figure 4(d)) and the ℓ_1 Kantorovich-Rubinstein distance (Figure 4(f)). The estimation of the low velocity zone at $x = 11\,\text{km}$, $z = 2.5\,\text{km}$ is slightly improved using the ℓ_1 Kantorovich-Rubinstein distance, as a high velocity artifact located in this zone is computed using the L^2 estimation. Starting from the second initial model, only the results obtained using ℓ_1 Kantorovich-Rubinstein distance are meaningful (Figure 4(g)). The poor initial approximation of the P-wave velocity is responsible for the cycle skipping effect and the L^2 estimation corresponds to a local minimum of the misfit function (Figure 4(f)). The estimation obtained with the ℓ_1 Kantorovich-Rubinstein distance is significantly closer from the true model, despite low velocity artifacts in the shallow part at $x = 1.5\,\text{km}$, $z = 1\,\text{km}$ and in depth at $x = 12\,\text{km}$, $z = 3.4\,\text{km}$. This example illustrates the potential of optimal transport for FWI: starting from a very crude approximation of the P-wave velocity, a meaningful estimation is computed. In the same configuration, FWI based on the least-squares distance fails and produces a heavily cycle skipped estimation.

4 Conclusion and Perspectives

The use of optimal transport distances for seismic imaging is promising. Comparing seismic data through these distances should yield more convex misfit functions with respect to the P-wave velocity parameter. However, the application of optimal transport to the comparison of seismic data requires the extension of the standard optimal transport problem to the transport of signed measures, which is not straightforward. Standard decomposition techniques proposed in [24], which are based on the negative and the positive part of the data, either are not adapted to FWI (separate transport of the positive and negative part, transport of the absolute value of the data) or lose the convexity property with respect to time-shifts which is one of the key properties one would like to satisfy for FWI.

Nonetheless, in the particular case of the dual formulation of the W_1 distance, the optimal transport distance can be related to a norm in the space of signed measure, the Kantorovich-Rubinstein norm. Hence, it can be directly use to compare seismic data. This is the strategy we have followed in previous works and which is summarized in this study. The results are encouraging. The resulting misfit function is not convex with respect to time-shifts, however, it already allows to start the FWI process from more crude initial velocity model, which denotes a wider valley of attraction of the misfit function. This method has been already successfully applied to 2D synthetic datasets in the context of deep water salt structures imaging (BP 2004 case study) and reflection dominated data (Chevron 2014 case study) [27] as

well as to a 3D synthetic dataset (SEG/EAGE overthrust model) [28]. The method should now be applied to 2D and 3D real datasets to further investigate the interest of this strategy for FWI.

Despite the interesting results provided by the Kantorovich-Rubinstein norm, the convexity property of the optimal transport distance with respect to shifted patterns on the data one could expect is lost. Further investigations are thus required to assess the feasibility of the design of a misfit function, based on optimal transport, adapted to the comparison of seismic data, which would benefit from this convexity property. Among different possibilities, one could think of the construction of positive observable from the seismic data, such as its envelope, which could thus be compared through W_p distances.

Acknowledgements This study was partially funded by the SEISCOPE consortium (*http://seiscope2.osug.fr*), sponsored by CGG, CHEVRON, EXXON-MOBIL, JGI, SHELL, SINOPEC, STATOIL, TOTAL, and WOODSIDE. This study was granted access to the HPC resources of the Froggy platform of the CIMENT infrastructure (https://ciment.ujf-grenoble.fr), which is supported by the Rhône-Alpes region (GRANT CPER07_13 CIRA), the OSUG@2020 labex (reference ANR10 LABX56), and the Equip@Meso project (reference ANR-10-EQPX-29-01) of the programme Investissements d'Avenir supervised by the Agence Nationale pour la Recherche, and the HPC resources of CINES/IDRIS/TGCC under the allocation 046091 made by GENCI.

References

1. Adams, J. C. (1989). MUDPACK: Multigrid portable FORTRAN software for the efficient solution of linear elliptic partial differential equations. *Applied Mathematics and Computation*, 34(2):113–146.
2. Ambrosio, L., Gigli, N., and Savaré, G. (2008). *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.
3. Benamou, J.-D. and Brenier, Y. (2000). A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*.
4. Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. (2015). Iterative Bregman Projections for Regularized Transportation Problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138.
5. Bogachev, V. I. (2007). *Measure Theory*. Number vol. I,II in Measure Theory. Springer Berlin Heidelberg.
6. Bozdağ, E., Trampert, J., and Tromp, J. (2011). Misfit functions for full waveform inversion based on instantaneous phase and envelope measurements. *Geophysical Journal International*, 185(2):845–870.
7. Brandt, A. (1977). Multi-level adaptive solutions to boundary-value problems. *Mathematics of Computation*, 31:333–390.
8. Bunks, C., Salek, F. M., Zaleski, S., and Chavent, G. (1995). Multiscale seismic waveform inversion. *Geophysics*, 60(5):1457–1473.
9. Chavent, G. (1971). *Analyse fonctionnelle et identification de coefficients répartis dans les équations aux dérivées partielles*. PhD thesis, Université de Paris.
10. Combettes, P. L. and Pesquet, J.-C. (2011). Proximal splitting methods in signal processing. In Bauschke, H. H., Burachik, R. S., Combettes, P. L., Elser, V., Luke, D. R., and Wolkowicz, H., editors, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, volume 49 of *Springer Optimization and Its Applications*, pages 185–212. Springer New York.

11. Cuturi, M. (2013). Sinkhorn distances: lightspeed computation of optimal transportation distances. *Advances in Neural Information Processing Systems*.
12. Devaney, A. (1984). Geophysical diffraction tomography. *Geoscience and Remote Sensing, IEEE Transactions on*, GE-22(1):3–13.
13. Engquist, B. and Froese, B. D. (2014). Application of the Wasserstein metric to seismic signals. *Communications in Mathematical Science*, 12(5):979–988.
14. Fichtner, A., Kennett, B. L. N., Igel, H., and Bunge, H. P. (2008). Theoretical background for continental- and global-scale full-waveform inversion in the time-frequency domain. *Geophysical Journal International*, 175:665–685.
15. Hale, D. (2013). Dynamic warping of seismic images. *Geophysics*, 78(2):S105–S115.
16. Jannane, M., Beydoun, W., Crase, E., Cao, D., Koren, Z., Landa, E., Mendes, M., Pica, A., Noble, M., Roeth, G., Singh, S., Snieder, R., Tarantola, A., and Trezeguet, D. (1989). Wavelengths of Earth structures that can be resolved from seismic reflection data. *Geophysics*, 54(7):906–910.
17. Kantorovich, L. (1942). On the transfer of masses. *Dokl. Acad. Nauk. USSR*, 37:7–8.
18. Lailly, P. (1983). The seismic inverse problem as a sequence of before stack migrations. In Bednar, R. and Weglein, editors, *Conference on Inverse Scattering, Theory and application, Society for Industrial and Applied Mathematics, Philadelphia*, pages 206–220.
19. Le Dimet, F. and Talagrand, O. (1986). Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A*, 38A(2):97–110.
20. Lellmann, J., Lorenz, D., Schönlieb, C., and Valkonen, T. (2014). Imaging with Kantorovich–Rubinstein discrepancy. *SIAM Journal on Imaging Sciences*, 7(4):2833–2859.
21. Lions, J. L. (1968). *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*. Dunod, Paris.
22. Luo, S. and Sava, P. (2011). A deconvolution-based objective function for wave-equation inversion. *SEG Technical Program Expanded Abstracts*, 30(1):2788–2792.
23. Luo, Y. and Schuster, G. T. (1991). Wave-equation traveltime inversion. *Geophysics*, 56(5):645–653.
24. Mainini, E. (2012). A description of transport cost for signed measures. *Journal of Mathematical Sciences*, 181(6):837–855.
25. Métivier, L. and Brossier, R. (2016). The SEISCOPE optimization toolbox: A large-scale nonlinear optimization library based on reverse communication. *Geophysics*, 81(2):F11–F25.
26. Métivier, L., Brossier, R., Mérigot, Q., Oudet, E., and Virieux, J. (2016). Increasing the robustness and applicability of full waveform inversion: an optimal transport distance strategy. *The Leading Edge*, 35(12):1060–1067.
27. Métivier, L., Brossier, R., Mérigot, Q., Oudet, E., and Virieux, J. (2016). Measuring the misfit between seismograms using an optimal transport distance: Application to full waveform inversion. *Geophysical Journal International*, 205:345–377.
28. Métivier, L., Brossier, R., Mérigot, Q., Oudet, E., and Virieux, J. (2016c). An optimal transport approach for seismic tomography: Application to 3D full waveform inversion. *Inverse Problems*, 32(11):115008.
29. Nash, S. G. (2000). A survey of truncated Newton methods. *Journal of Computational and Applied Mathematics*, 124:45–59.
30. Nocedal, J. (1980). Updating Quasi-Newton Matrices With Limited Storage. *Mathematics of Computation*, 35(151):773–782.
31. Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer, 2nd edition.
32. Operto, S., Brossier, R., Gholami, Y., Métivier, L., Prieux, V., Ribodetti, A., and Virieux, J. (2013). A guided tour of multiparameter full waveform inversion for multicomponent data: from theory to practice. *The Leading Edge*, Special section Full Waveform Inversion(September):1040–1054.
33. Philippis, G. D. and Figalli, A. (2014). The Monge-Ampère equation and its link to optimal transportation. *BULLETIN (New Series) OF THE AMERICAN MATHEMATICAL SOCIETY*.
34. Plessix, R. E. (2006). A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Geophysical Journal International*, 167(2):495–503.

35. Pratt, R. G. (1999). Seismic waveform inversion in the frequency domain, part I : theory and verification in a physical scale model. *Geophysics*, 64:888–901.
36. Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing.
37. Shipp, R. M. and Singh, S. C. (2002). Two-dimensional full wavefield inversion of wide-aperture marine seismic streamer data. *Geophysical Journal International*, 151:325–344.
38. Swarztrauber, P. N. (1974). A Direct Method for the Discrete Solution of Separable Elliptic Equations. *SIAM Journal on Numerical Analysis*, 11(6):1136–1150.
39. Symes, W. W. (2008). Migration velocity analysis and waveform inversion. *Geophysical Prospecting*, 56:765–790.
40. Tarantola, A. (1984). Inversion of seismic reflection data in the acoustic approximation. *Geophysics*, 49(8):1259–1266.
41. Villani, C. (2003). *Topics in optimal transportation*. Graduate Studies In Mathematics, Vol. 50, AMS.
42. Villani, C. (2008). *Optimal transport : old and new*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin.
43. Virieux, J., Asnaashari, A., Brossier, R., Métivier, L., Ribodetti, A., and Zhou, W. (2017). An introduction to Full Waveform Inversion. In Grechka, V. and Wapenaar, K., editors, *Encyclopedia of Exploration Geophysics*, page R1–1–R1–40. Society of Exploration Geophysics.
44. Virieux, J. and Operto, S. (2009). An overview of full waveform inversion in exploration geophysics. *Geophysics*, 74(6):WCC1–WCC26.
45. Warner, M. and Guasch, L. (2014). Adaptative waveform inversion - FWI without cycle skipping - theory. In *76th EAGE Conference and Exhibition 2014*, page We E106 13.

Exploiting Sparsity in Solving PDE-Constrained Inverse Problems: Application to Subsurface Flow Model Calibration

Azarang Golmohammadi, M-Reza M. Khaninezhad, and Behnam Jafarpour

Abstract Inverse problems are frequently encountered in many areas of science and engineering where observations are used to estimate the parameters of a system. In several practical applications, the dynamic processes that take place in a physical system are described using a set of partial differential equations (PDEs), which are typically nonlinear and coupled. The inverse problems that arise in those systems ought to be constrained to honour the governing PDEs. In this chapter, we consider high-dimensional PDE-constrained inverse problems in which, because of spatial patterns and correlations in the distribution of physical properties of a system, the underlying parameters tend to reside in (usually unknown) low-dimensional manifolds, thus have sparse (low-rank) representations. The sparsity of the parameters is amenable to an effective and flexible regularization form that can be exploited to improve the solution of such inverse problems. In applications where prior training data are available, sparse manifold learning methods can be adopted to tailor parameter representations to the specific requirements of the prior data. However, a major risk in employing prior training data is the significant uncertainty about the underlying conceptual models and assumptions used to develop the prior. A group-sparsity formulation is discussed for addressing the uncertainty in the prior training data when multiple distinct, but plausible, prior scenarios are encountered. Examples from geosciences application are presented where images of rock material properties are reconstructed from limited nonlinear fluid flow measurements.

A. Golmohammadi · M.-R. M. Khaninezhad
Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA
e-mail: agolmoha@usc.edu; m.khaninezhad@usc.edu

B. Jafarpour (✉)
Mork Family Department of Chemical Engineering and Material Science, University of Southern California, Los Angeles, CA, USA

Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA
e-mail: behnam.jafarpour@usc.edu

H. Antil et al. (eds.), *Frontiers in PDE-Constrained Optimization*, The IMA Volumes in Mathematics and its Applications 163,
https://doi.org/10.1007/978-1-4939-8636-1_12

1 Introduction

The spatiotemporal evolution of dynamic state variables in many physical systems is governed by coupled partial differential equations (PDEs) that are typically derived from the balance laws of physics (mass, momentum, and energy conservation). The observable responses of these dynamical systems can usually be described as a function of their state variables, which in turn depend on model inputs, including controls, initial/boundary conditions, and parameters. In general, the functional relation between model input parameters and observable responses can be expressed as a (typically nonlinear) mapping that involves the solution of the underlying PDEs. Examples of these physical systems include fluid flow and heat transfer processes [64], electromagnetic systems [65], motion of planets in solar system [60], human's neural mechanism [42]. The exponential increase in computing power has enabled considerable advances in numerical simulation of complex processes in large-scale physical systems that have high-dimensional PDEs as governing equations. Advances in computing power have also led to development of computationally demanding inverse modelling algorithms with potentially thousands of forward model simulations, which was once considered infeasible.

The parameters that appear in the governing PDEs of physical systems are either directly observable or they need to be inferred from indirect and often limited observable quantities (outputs) of the system [52, 62, 79, 85, 89]. In some cases, a spatially distributed physical property may only be directly observable at finite points in space, requiring spatial interpolation techniques to predict unobserved parameter values. In general, estimation of model parameters from limited output measurements of the system leads to an inference or inverse problem [59, 78]. In many cases, the inverse modelling formulations involve a minimization problem where the objective function represents the mismatch between model predicted and observed data as well as other terms that penalize departure from prior (explicit or implicit) knowledge about the solution. When the system outputs depend on the solution of the PDEs that establish physical laws (e.g. mass/momentum/energy balance), the resulting inverse problem formulation must ensure that the PDE constraints are honoured, thus leading to a PDE-constrained inverse problem. Including the PDE constraints ensures that the solution of the resulting inverse problem honours the underlying governing equations (i.e. well-established physical laws such as mass/momentum conservation).

Inverse problems that arise in many practical applications are ill-posed, as the measured data are not sufficient to find a unique solution [35, 63]. When there are fewer measurements than unknown model parameters in a system, a situation that is commonly encountered in practice, the problem is underdetermined and cannot have a unique solution. Additional (a priori) information are needed to constrain the solution and eliminate implausible outcomes. A common approach to address solution non-uniqueness is to adopt a probabilistic (Bayesian) inverse modelling framework [1, 26, 32, 51, 53, 78], where the elements of the inverse problem (parameters, data, and forward model) are represented with their respective

uncertainties, typically using probability density functions (PDFs). In this chapter, we focus on deterministic inverse problems. First, an overview of inverse modelling formulation is presented, followed by general strategies for solving ill-posed inverse problems that are constrained by complex PDEs. In numerical solution techniques, the PDEs are solved by first discretizing the domain and assigning input parameters to the discrete cells. This approach leads to a discrete ill-posed inverse problem in which vector representations (as opposed to continuous functions) are used to describe the unknown parameters. The main focus of this chapter is on formulation and solution of such discrete inverse problems in which the parameters are either inherently sparse or can have a sparse approximation.

2 Inverse Problem Formulation

To formulate a general inverse problem, consider collecting the observations of a physical system in a vector $\mathbf{d}$. These observations are related to the parameters of the system through a (generally nonlinear) mapping, i.e. $\mathbf{d} = \mathbf{g}(\mathbf{u})$. Here, $\mathbf{u}$ contains the parameters of the system, and $\mathbf{g}(.)$ is the nonlinear function that maps the parameter space onto the observation space. We assume that the observations $\mathbf{d}$ and the parameters $\mathbf{u}$ are vectors in $\mathbb{R}^{m\times 1}$ and $\mathbb{R}^{n\times 1}$, respectively.

Definition (General Inverse Problem) Consider the *Banach* spaces $\mathscr{U}$ and $\mathscr{D}$, and a mapping $\mathbf{G} : \mathscr{U} \rightarrow \mathscr{D}$. The inverse problem consists of the solution to the equation [66]:

$$\mathbf{g}(\mathbf{u}) = \mathbf{d} \qquad \mathbf{u} \in \mathscr{U} \quad \& \quad \mathbf{d} \in \mathscr{D} \tag{1}$$

If an exact solution is not expected (e.g. due to observation errors), the inverse problem in (1) is expressed as a minimization of the form:

$$\min_{\mathbf{u}} \quad J(\mathbf{u}) = \|\mathbf{g}(\mathbf{u}) - \mathbf{d}\|_2^2 \qquad \mathbf{u} \in \mathscr{U} \tag{2}$$

When the *Banach* space $\mathscr{D}$ is some ℓ^2-space, then this becomes a classical least-squares problem [57].

The simplest form of an inverse problem is obtained when observations and model parameters are related linearly [13, 78], i.e. $\mathbf{d} = \mathbf{G}\mathbf{u} + \boldsymbol{\epsilon}$. Here, $\mathbf{u}$ is the parameter of interest, $\mathbf{G}$ is the linear mapping from parameter space to the observation space, and $\boldsymbol{\epsilon}$ is the observation noise, which is usually considered to be independent of the parameters $\mathbf{u}$. In the linear case, the inverse problem in Equation (2) is expressed as:

$$\min_{\mathbf{u}} \quad \|\mathbf{G}\mathbf{u} - \mathbf{d}\|_2^2 \qquad \text{s.t.,} \qquad \mathbf{u} \in \mathscr{U} \tag{3}$$

with a simple quadratic objective function. In practical applications, when data is noisy, the least-square term in Equation (3) is generalized to $\|\mathbf{C}_{\boldsymbol{\epsilon}}^{-\frac{1}{2}}(\mathbf{Gu}-\mathbf{d})\|_2^2$, where $\mathbf{C}_{\boldsymbol{\epsilon}}$ is the (usually diagonal) noise covariance matrix $\boldsymbol{\epsilon}$. For ill-posed linear inverse problems, the formulation often takes the form:

$$\min_{\mathbf{u}} \quad J(\mathbf{u}) \quad \text{s.t.,} \quad \|\mathbf{d}-\mathbf{Gu}\|_2^2 \leq \sigma^2 \tag{4a}$$

$$\min_{\mathbf{u}} \quad J(\mathbf{u}) + \frac{1}{\lambda^2}(\|\mathbf{d}-\mathbf{Gu}\|_2^2 - \sigma^2) \tag{4b}$$

$$\min_{\mathbf{u}} \quad \|\mathbf{d}-\mathbf{Gu}\|_2^2 + \lambda^2 J(\mathbf{u}) \tag{4c}$$

In Equation (4a), the constraint, i.e. $\|\mathbf{d}-\mathbf{Gu}\|_2^2 \leq \sigma^2$, is added to the objective function by the penalty method [8], and the resulting equation in (4b) is reshaped into Equation (4c) by multiplying the objective function by λ^2. In Equation (4), $J(\mathbf{u})$ is a function that restricts (regularizes) the behaviour/structure of $\mathbf{u}$, and σ^2 is a bound on the observation error. For example, if $\mathbf{u}_0$ is a prior belief about the parameter $\mathbf{u}$, minimization of $J(\mathbf{u}) = \|\mathbf{u}-\mathbf{u}_0\|_2^2$ results in a solution with minimum departure from $\mathbf{u}_0$ [78]. A classical example of regularization functions are the Tikhonov regularization forms [81], for which $J(\mathbf{u})$ is defined as the second norm of the first or second derivatives of the parameters (to promote solution smoothness or flatness, respectively). It is important to note that the regularization parameter λ has a significant impact on the solution by balancing the importance of data misfit and regularization terms. For linear problems, cross validation [31] and L-curve [34] methods have been proposed for finding an optimal value for the regularization parameter.

In many practical problems, the relationship between the observed data and model parameters is nonlinear, i.e. $\mathbf{d} = \mathbf{g}(\mathbf{u}) + \boldsymbol{\epsilon}$ [74, 79]. The corresponding nonlinear inverse problem can be expressed as:

$$\min_{\mathbf{u}} \quad J(\mathbf{u}) \quad \text{s.t.,} \quad \|\mathbf{d}-\mathbf{g}(\mathbf{u})\|_2^2 \leq \sigma^2 \tag{5a}$$

$$\min_{\mathbf{u}} \quad \|\mathbf{d}-\mathbf{g}(\mathbf{u})\|_2^2 + \lambda^2 J(\mathbf{u}) \tag{5b}$$

For physical systems in which the evolution of the state variables is determined by solving PDE systems, the resulting inverse problems include the PDEs as constraints, that is,

$$\min_{\mathbf{u}} \quad \|\mathbf{d}-\mathbf{g}(\mathbf{u})\|_2^2 + \lambda^2 J(\mathbf{u}) \qquad \text{s.t.,} \qquad f(\mathbf{u}, \mathbf{x}(\mathbf{u})) = 0 \tag{6}$$

where $f(\mathbf{u}, \mathbf{x}(\mathbf{u})) = 0$ represents the PDE system. We note that the measurement operator $\mathbf{g}(\mathbf{u})$ is usually a function of the state vector $\mathbf{x}(\mathbf{u})$, which, for compactness, is not explicitly expressed in Equation (6). It is common to enforce the constraints by first solving the PDE system to obtain the state variables and then using them

to predict the measurements. In other words, the PDE system is solved to derive the nonlinear measurements, resulting in predicted measurements that honour the constraints.

In practice, nonlinear inverse problems do not lend themselves to analytical solutions, and iterative numerical optimization techniques must be employed to find the solution. In iterative solution schemes, given the current iterate $\mathbf{u}^{(k)}$, an updated solution is sought by expanding the nonlinear function $\mathbf{g}(\mathbf{u})$ around the current iterate, using either first- or second-order Taylor expansions. For example, when a linear approximation is used, the resulting objective function takes the form:

$$\mathbf{u}^{(k+1)} = \underset{\mathbf{u}}{\operatorname{argmin}} \quad \|\mathbf{d} - (\mathbf{g}(\mathbf{u}^{(k)}) + \mathbf{G}_{\mathbf{u}}(\mathbf{u} - \mathbf{u}^{(k)}))\|_2^2 + \lambda^2 J(\mathbf{u}) \tag{7}$$

where $\mathbf{G}_{\mathbf{u}}$ is the Jacobian matrix that contains the first-order derivative of multivariate vector function $\mathbf{g}(\mathbf{u})$ with respect to entries of $\mathbf{u} = \mathbf{u}^{(k)}$. The linear objective function in Equation (7) can be readily minimized to find $\mathbf{u}^{(k+1)}$, and the process is continued until the algorithm converges to a solution [78].

3 Parametrization and Regularization Techniques

Inverse problems often involve high-dimensional parameters with complex relations that need to be estimated from low-resolution nonlinear data. In addition to numerical stability issues (due to high-dimensional and low-rank nature of the matrices involved) in solving such ill-posed inverse problems, several non-unique solutions can be found that reproduce the (limited) available data, but fail to predict the future response of the system. In some physical systems, the parameters may represent spatially distributed material properties with specific architecture or patterns. In such cases, in addition to dealing with high parameter dimensionality, it is important to preserve the expected spatial structure of the parameters [6, 12, 18, 37, 47, 90]. Parametrization and regularization are two common approaches that aim to achieve these two goals by reducing parameter dimensionality and imparting pre-specified attributes on the solution. Techniques for regularizing the solution of ill-posed inverse problems have been extensively studied in the literature (e.g. see [24, 78, 81, 86]). Regularization is usually implemented by minimizing a penalty function (e.g. $J(\mathbf{u})$ in Equations (4)–(7)) that promotes an attribute of interest in the solution, e.g. using a roughness penalty function to obtain smooth solutions. By imposing certain patterns/attributes on the solution, regularization creates correlation structures that, in effect, implicitly reduce the dimension of the parameter space.

Inverse problem formulations are directly influenced by the choice of parameters (i.e. parameterization or re-parameterization) [24, 39]. Parameterization refers to changing the original parameters of an inverse problem to a (typically much smaller) set of new parameters that facilitate the search for a solution. It is often

used to explicitly reduce the number of unknown parameters, while capturing their main characteristics, with the purpose of alleviating problem ill-posedness. Parameterization can also provide more compact descriptions of complex parameter structures and facilitate their reconstruction. In solving inverse problems, choosing an appropriate domain that affords an effective description of the parameters is complicated by the lack of complete knowledge about the solution. However, a reasonable choice for the parameter domain may be deduced from the knowledge about the physics of the system under analysis and/or based on the past experience. Parameterization can be performed either in the original domain (space/time), in which the PDEs are solved, or they can be implemented by transforming the parameters into a different (often abstract) domain with certain desirable properties.

A linear parameterization [87] can generally be expressed as:

$$\mathbf{u} = \mathbf{\Phi}\mathbf{v} = \sum_{i=1}^{k} \boldsymbol{\phi}_i v_i \tag{8}$$

where $\mathbf{u}$ and $\mathbf{v}$ are vectors of the original and transformed model parameters, respectively; and $\mathbf{\Phi}$ is the linear transformation matrix with columns corresponding to the basis functions (i.e. $\phi_{i:i=1,\ldots,k}$), which are linearly combined, using the entries of $\mathbf{v}$ as coefficients, to yield $\mathbf{u}$. Matrix $\mathbf{\Phi}$ can be viewed as a linear mapping of the transformed parameters $\mathbf{v}$ onto the original parameters $\mathbf{u}$. Different choices of $\mathbf{\Phi}$ lead to alternative parameterization bases (domains) with distinct properties that can be exploited in formulating the inverse problem.

Using the linear relation $\mathbf{u} = \mathbf{\Phi}\mathbf{v}$, it is straightforward to rewrite the inverse problem objective function in Equation (7) in terms of $\mathbf{v}$ as follows:

$$\mathbf{v}^{(k+1)} = \underset{\mathbf{v}}{\operatorname{argmin}} \quad \|\mathbf{d} - (\mathbf{g}(\mathbf{v}^{(k)}) + \mathbf{G}_{\mathbf{v}}(\mathbf{v} - \mathbf{v}^{(k)}))\|_2^2 + \lambda^2 J(\mathbf{v}) \tag{9}$$

where $J(\mathbf{v})$ defines a regularization constraint on the new parameters $\mathbf{v}$ in the transform domain (more details in subsequent sections). Note that the transformation matrix $\mathbf{\Phi}$ is assumed to be constant and dropped for brevity. Furthermore, $\mathbf{G}_{\mathbf{v}}$ in Equation (9) presents the Jacobian matrix of the observations with respect to the transformed coefficients and can be simply calculated through the chain rule of differentiation as:

$$\mathbf{G}_{\mathbf{v}} = \frac{\partial}{\partial \mathbf{v}} \mathbf{g}(\mathbf{v})|_{\mathbf{v}=\mathbf{v}^k} = \mathbf{G}_{\mathbf{u}}\mathbf{\Phi} \tag{10}$$

A nonlinear parameterization can be expressed as $\mathbf{u} = \phi(\mathbf{v})$, where the mapping $\phi(.)$ represents a general nonlinear transformation. For instance, kernel functions provide mappings that can be used to reduce parameter nonlinearity prior to applying a linear parameterization [70, 71, 84]. Kernel-based methods use kernel functions to operate in high-dimensional feature spaces without computing the coordinates of the feature space. Instead, they compute the inner products of

(training) data pairs in the feature space. Using this approach, the inner product of the vectors in the nonlinear space is calculated by kernel functions, $k(\mathbf{v}, \mathbf{v}') =< \phi(\mathbf{v}), \phi(\mathbf{v}') >$, where $\phi(.)$ is a feature map (e.g. a polynomial). The kernel $k(\mathbf{v}, \mathbf{v}')$ is a function of $\mathbf{v}$ and $\mathbf{v}'$, and it eliminates the need for nonlinear expansion of the parameters. A major difficulty that arises in implementing nonlinear transformations is the lack of unique back transformation due to the nonlinear form of the transform function $\phi(.)$. In this chapter, linear transforms are discussed.

3.1 Parameterization/Regularization in Space

Zonation Zonation [37] is the simplest spatial parameterization technique in which subsets of the parameter vector $\mathbf{u}$ are assumed to have (approximately) identical values and can be aggregated into a single parameter. In imaging applications where $\mathbf{u}$ is a spatial image (of an unknown property distribution), subsets of entries of $\mathbf{u}$ that correspond to a local neighbourhood in the image form a segment or a zone with identical parameter values. By aggregating such multiple entries into a single parameter, zonation can significantly reduce the number of parameters. Figure 1(a) depicts a sample parameter distribution (shown in x-y plane) that consists of k regions or zones ($\boldsymbol{R}_1, \ldots, \boldsymbol{R}_k$). If the parameter values in each region are similar,

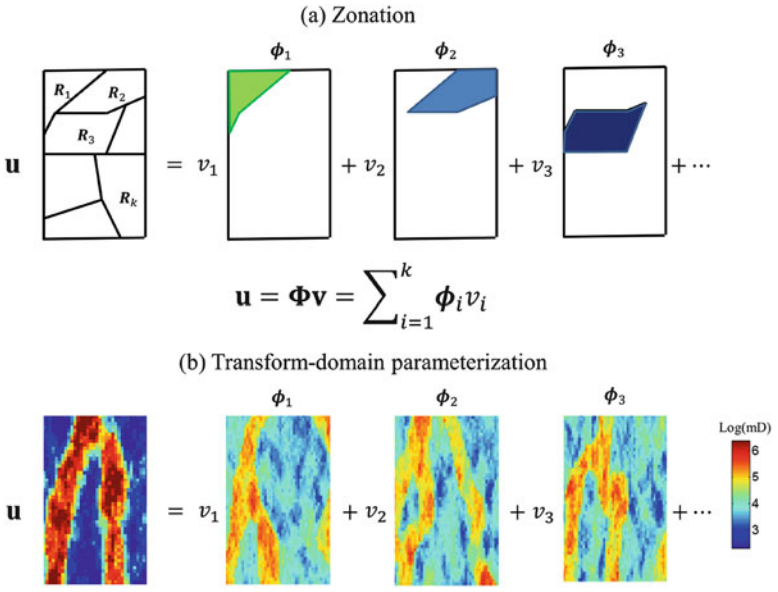

Fig. 1 Schematic of parameter representation via linear expansion: (a) spatial zonation with predefined regions with similar parameter values; (b) expansion with functions derived from compressive transform

the number of parameters can be reduced to $k << n$. This parameterization can be effectively expressed using a general linear expansion representation, consisting of basis vectors $\boldsymbol{\phi}_{s:1\leq s\leq k}$ in which only the entries corresponding to region $\boldsymbol{R}_s$ are non-zero (ones) and the remaining entries are zero (see Figure 1(a)). Using zonation, the formulation of the inverse problem is reduced to:

$$\min_{\mathbf{v}} \quad \|\mathbf{d}-\mathbf{g}(\mathbf{u})\|_2^2+\lambda^2 J(\mathbf{u}) \qquad \text{s.t.,} \qquad \mathbf{u}=\sum_{i=1}^{k}\boldsymbol{\phi}_i v_i \tag{11a}$$

where with the new parameters, i.e. $[v_1 v_2 \dots v_k]$, the problem is better posed (only k unknowns). In many cases, zonation leads to very few zones, eliminating the need for the regularization term, i.e. $J(\mathbf{u})$ in Equation (11a). Therefore, a simpler version of the problem can be expressed as:

$$\min_{\mathbf{v}} \quad \|\mathbf{d}-\mathbf{g}(\mathbf{u})\|_2^2 \qquad \text{s.t.,} \qquad \mathbf{u}=\sum_{i=1}^{k}\boldsymbol{\phi}_i v_i \tag{11b}$$

Although zonation is a simple and intuitive parameterization approach, it suffers from a number of shortcomings. First, it is not trivial to define the zones for an unknown map a-priori. Adaptive multi-resolution zonation techniques [33] have been developed that allow the zones to be redefined (updated) during inversion. Second, the sharp boundaries that separate the zones may not be realistic or plausible. Finally, eliminating the variability (heterogeneity) within each region can result in unintended elimination of local, but important, features and introduce undesired solution bias. Several other parameterization methods have been developed to improve the ill-posedness of inverse problems. Examples of these methods include transform-domain techniques such as the principal component analysis (PCA), the Fourier-based discrete cosine transform (DCT), and the discrete wavelet transform (DWT) (Section 3.2).

Tikhonov Regularization Tikhonov regularization [81] is achieved by minimizing the zeroth-/first-/second-order derivative of the solution to promote minimum-length/smooth/flat solutions, respectively. Tikhonov regularization has been widely applied to inverse problems in several imaging applications, where the parameters are expected to show some degree of continuity. The reason for this attribute is that images that represent the parameters are often related to physical properties that naturally follow certain continuity in their formation. To illustrate how Tikhonov regularization works, consider the local operator that approximates the first-order directional derivative for entry $u_{i,j}$ of the parameter vector $\mathbf{u}$ (defined on a two-dimensional x-y coordinates), i.e. $(\nabla\mathbf{u})_{i,j} \approx \begin{bmatrix} u_{i,j}-\frac{1}{2}(u_{i-1,j}+u_{i+1,j}) \\ u_{i,j}-\frac{1}{2}(u_{i,j-1}+u_{i,j+1}) \end{bmatrix}$. This notation is used to demonstrate the central point finite difference approximation to the first-order directional derivative. Minimizing $\int \|\nabla\mathbf{u}\|_2^2 \mathrm{d}\mathbf{u} \approx \Delta \times$

$\sum_{i,j} \|(\nabla \mathbf{u})_{i,j}\|_2^2$, where Δ denotes a small spatial perturbation, corresponds to solutions that exhibit smooth transition (in parameter values) from $u_{i,j}$ to its neighbouring grid cells. With the first-order Tikhonov regularization, the inverse problem objective function takes the form:

$$\min_{\mathbf{u}} \quad \|\mathbf{d} - \mathbf{g}(\mathbf{u})\|_2^2 + \lambda^2 \int \|\nabla \mathbf{u}\|_2^2 \mathrm{d}\mathbf{u} \tag{12}$$

For discrete problems, the spatial derivatives and the regularization function can be written as a linear operator $\mathbf{W}$; that is, the regularization term can be simplified to $\int \|\nabla \mathbf{u}\|_2^2 \mathrm{d}\mathbf{u} = \|\mathbf{W}\mathbf{u}\|_2^2$.

Total Variation Total variation [27, 50, 69] is a regularization technique that is used to promote piecewise smooth solutions. Hence, the regularization penalty is lenient to solutions that are generally smooth but can have discontinuity in certain parts. This form of regularization is implemented by applying a milder penalty to spatial derivatives of the parameters. In Total Variation, the ℓ_1-norm (instead of the ℓ_2-norm) of the first-order derivative of the solution is minimized. The ℓ_1-norm is less sensitive to larger entries and tends to tolerate discontinuity, which is often exhibited through large directional derivatives. In implementing the total variation, one seeks to minimize the following regularized least-squares form:

$$\min_{\mathbf{u}} \quad \|\mathbf{d} - \mathbf{g}(\mathbf{u})\|_2^2 + \lambda^2 \int \sqrt{\sum_j (\nabla_j \mathbf{u})^2} \mathrm{d}\mathbf{u} \tag{13}$$

where the index j is the number of directional derivatives, and $\nabla_j \mathbf{u}$ calculates the derivative of $\mathbf{u}$ at a direction specified by index j. The total variation regularization can be implemented for any specified direction. In its standard implementation, the directions j are the three Cartesian coordinates.

3.2 Transform-Domain Parameter Representations

Compressive transforms are used to compactly represent/approximate the most salient features of images and signals. In inverse problems, it may be possible to apply a transformation to the original parameters to achieve an effective low-rank representation. Examples of transform-domain representation techniques are those that are used in image compression, e.g. Wavelet [55] or Fourier [9] transforms. These transforms use predefined basis functions with strong compression property to provide compact (low-rank) description of natural images. The compression property of a basis directly corresponds to the decay rate of the transformed coefficients. The main steps in transform-domain low-rank representation include (i) choosing an appropriate transformation basis (expansion functions), (ii) performing

the forward transformation to obtain the transformed representation of the original parameters, (iii) identifying and retaining only significant coefficients of the transformed representation, and (iv) back transformation to the original domain using only the retained coefficients. The compressive nature of the transforms implies that the transformed representation is sparse, that is, very few of the transformed coefficients are significant. In this section, we present some of the important compressive transforms that have been used for parameterization. The discussion on identifying and retaining the significant elements in the transformed representation is presented in Section 4.

The choice of an appropriate basis to compactly represent model parameters is intimately related to the prior knowledge about the characteristics of the underlying properties of the model, e.g. existing correlation/connectivity structures or possible discontinuous features. In fact, when specific prior models are available, one could construct a specialized transformation that is learned from those models and training data. Examples of specialized transform basis functions that are learned from prior information include the principal component analysis (PCA) [41] and the k-SVD [2] for sparse dictionary learning, which are discussed in this section. In many situations, however, explicit prior models or training data may not be available. In those cases, generic transforms that are used in image compression provide an attractive option for parameterization. We briefly discuss two popular generic transformation methods, namely Fourier transform [9] and its practical and efficient variation known as the discrete cosine transform (DCT) [3, 39] and the wavelet transform [55, 77].

3.2.1 Generic Compressive Transforms

Generic compressive transforms consist of n linearly independent basis vectors in $\mathbb{R}^n$ that can be used to span any length-n vector (or vectorized image). While a complete representation of a length-n parameter vector is possible in a compressive basis, the objective is to obtain an approximate representation by only retaining $k << n$ significant basis elements. Suppose that the set $\{\boldsymbol{\phi}_{i:i=1,\ldots,n}\}$ contains all the basis vectors that are needed for perfect representation in $\mathbb{R}^n$, and a subset $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_{i:i=1,\ldots,k}\}$, with no particular order, provides an acceptable approximation for a vector of interest $\mathbf{u}$. Selection of the subset with k elements depends on the original vector to be approximated and the basis used.

Fourier and Wavelet Transforms Fourier basis functions describe a signal in terms of its frequency content. In this case, if an $n = \prod n_i$ dimensional signal $\mathbf{u}$ is defined in $\mathbb{R}^{n_1 \times n_2 \times \ldots \times n_d}$, the FT at frequency $(f_1, \ldots, f_{n_d})$ will be calculated as:

$$\mathbf{v}(f_1, \ldots, f_{n_d}) = \sum_{i_1=0}^{n_1-1} \cdots \sum_{i_{n_d}=0}^{n_d-1} \mathbf{u}(i_1, \ldots, i_{n_d}) e^{-i2\pi(\sum_{t=1}^{n_d} \frac{f_t i_t}{n_t})} \tag{14}$$

The back transformation that returns **u** can be expressed as:

$$\mathbf{u}(i_1,\ldots,i_{n_d}) = \frac{1}{n}\sum_{f_1=0}^{n_1-1}\cdots\sum_{f_{n_d}=0}^{n_d-1}\mathbf{v}(f_1,\ldots,f_{n_d})e^{i2\pi(\sum_{t=1}^{n_d}\frac{f_t i_t}{n_t})} \tag{15}$$

If the main features in **u** are captured by low-frequency elements, which is especially true for smooth and correlated vectors, one could approximate **u** by truncating the basis elements with frequencies exceeding a certain threshold. The $(n\text{-}k)$ coefficients corresponding to frequencies higher than the specified threshold are then set to zero.

The DCT is a special case of the Fourier transform that only considers the real part of $e^{-i2\pi(\sum_{t=1}^{n_d}\frac{f_t i_t}{n_t})}$, which is $\cos\{2\pi(\sum_{t=1}^{n_d}\frac{f_t i_t}{n_t})\}$. Hence, the transformation takes the form:

$$\mathbf{v}(f_1,\ldots,f_{n_d}) = \sum_{i_1=0}^{n_1-1}\cdots\sum_{i_{n_d}=0}^{n_d-1}\mathbf{u}(i_1,\ldots,i_{n_d})\cos\{2\pi(\sum_{t=1}^{n_d}\frac{f_t i_t}{n_t})\} \tag{16}$$

Similar to Fourier transform, an approximation of the original signal **u** is obtained by truncating the frequencies above a certain threshold. Fourier-based transforms can only represent information either in space or frequency domains. This means that once a signal is transformed to Fourier domain, it loses the spatial information and vice versa. Hence, the Fourier basis elements are global and do not encode local information.

Unlike the Fourier transform, the basis elements in Wavelet transform contain both space and frequency information. This implies that each basis vector is localized in space and represents a certain frequency content. Therefore, for any spatial location, one can retain (truncate) specific frequency components that are significant (insignificant). Figure 2(a) and (b) shows 64 sample basis elements for the DCT and Haar wavelet transforms in $\mathbb{R}^{64\times 64}$, respectively. As can be verified, the basis images for the DCT transform are not localized in space while those for the discrete Haar wavelet clearly exhibit localized patterns. While generic compressive transforms have useful properties that make them very desirable when explicit prior knowledge is not available, in applications where prior knowledge about the solution (e.g. a training dataset) is available, one may be able to construct more specialized transforms with better performance.

3.2.2 Learned Compressive Transforms

Pre-constructed compressive bases achieve good compression performance in representing smooth and piecewise smooth images when specific knowledge about the image to be compressed is not available. For most natural images only a small subset of the transformed coefficients is sufficient to capture the main features of an image. This implies that most natural images have sparse approximations

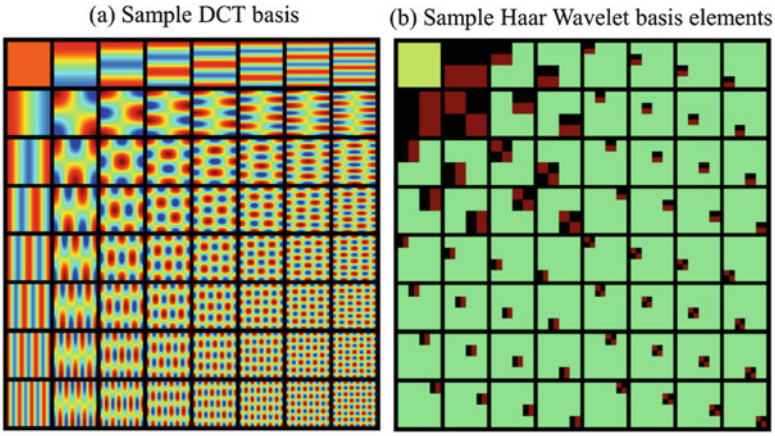

Fig. 2 Examples of generic (pre-computed) compressive transform bases: (a) sample low-frequency basis elements from the DCT basis; (b) sample basis elements from the discrete Haar wavelet. Example is shown for a 64×64 two-dimensional image. The basis elements are separated with black boxes

in these compressive transform domains. However, compressed representation of complex image features with generic transforms may require too many coefficients, which is not desirable for parameterization. Hence, a more sophisticated approach is needed to capture complex features in certain applications. In general, when a specific type of image (e.g. human face) is to be compressed, transforms that are specialized to represent the underlying features are more efficient. For example, in subsurface modelling, where extensive efforts go into data collection and site surveys to construct prior models, specialized transform-domain representations that are tailored to the information in the prior knowledge are more suitable.

Principal Component Analysis (PCA) The PCA [41] is widely used for dimensionality reduction in a wide range of applications. The PCA basis functions capture the main variability and structures in multivariate datasets, which can be exploited in compactly representing/approximating them with minimum loss of information. When the PCA is applied to the covariance matrix of a stochastic process, it diagonalizes the covariance and can be used to define a new (often more desirable) uncorrelated random process. In this case, the PCA provides an orthogonal transformation matrix with decorrelating power that contains, in its columns, the eigenvectors of the covariance matrix. The strong decorrelating property of the PCA basis is advantageous in eliminating parameter correlations (redundancies) to reduce dimensionality. In fact, the PCA sets the standard for dimension reduction with linear transforms as it gives the minimum error (in least-squares sense) in approximating an n-dimensional signal with $S << n$ basis elements (for a fixed S).

The parameterization with PCA follows the same format as in Equation (8), i.e. $\mathbf{u} = \mathbf{\Phi v} = \sum_{i=1}^{k} \boldsymbol{\phi}_i v_i$, where the basis functions ϕ_i's are the eigenvectors of the

covariance matrix of $\mathbf{u}$. Denoting an $n \times 1$-dimensional random variable as $\mathbf{u}$ and its covariance matrix as $\mathbf{C_u}$, the eigenvalue decomposition of the covariance matrix provides the following diagonalization form:

$$\mathbf{C_u} = \mathbf{\Phi \Lambda \Phi}^T \tag{17}$$

where $\mathbf{\Lambda}$ is a diagonal matrix (with eigenvalues of $\mathbf{C_u}$ in its diagonal entries) and $\mathbf{\Phi}$ is an orthonormal (transformation) matrix that has the eigenvectors of $\mathbf{C_u}$ in its columns. If sample realizations of $\mathbf{u}$ are collected into a data matrix $\mathbf{U}_{n\times L} = [\mathbf{u}_1 \dots \mathbf{u}_i \dots \mathbf{u}_L]$, the sample covariance matrix $\mathbf{C_u}$ can be computed as:

$$\mathbf{C_u} = \frac{1}{L-1}(\mathbf{U} - \bar{\mathbf{u}}\mathbf{1}_{1\times L})(\mathbf{U} - \bar{\mathbf{u}}\mathbf{1}_{1\times L})^T \tag{18}$$

where $\bar{\mathbf{u}}$ denotes the mean of $\mathbf{U}$, that is $\bar{\mathbf{u}} = \frac{1}{L}\sum_{i=1}^{L} \mathbf{u}_i$. The term $\frac{1}{\sqrt{L-1}}(\mathbf{U} - \bar{\mathbf{u}}\mathbf{1}_{1\times L})$ can be expressed in terms of its singular value decomposition (SVD) as [48, 72]:

$$\frac{1}{\sqrt{L-1}}(\mathbf{U} - \bar{\mathbf{u}}\mathbf{1}_{1\times L}) = \mathbf{\Psi \Sigma V}^T \tag{19}$$

where $\mathbf{\Psi}$ and $\mathbf{V}$ are orthonormal matrices that contain the left- and right-singular vectors of $\frac{1}{\sqrt{L-1}}(\mathbf{U} - \bar{\mathbf{u}}\mathbf{1}_{1\times L})$, respectively. Combining (18) and (19) yields

$$\mathbf{C_u} = (\mathbf{\Psi \Sigma V}^T)(\mathbf{\Psi \Sigma V}^T)^T = \mathbf{\Psi \Sigma V}^T\mathbf{V \Sigma \Psi}^T = \mathbf{\Psi \Sigma}^2\mathbf{\Psi}^T \tag{20}$$

which reveals $\mathbf{\Psi} = \mathbf{\Phi}$; that is, the left-singular vectors of $\frac{1}{\sqrt{L-1}}(\mathbf{U} - \bar{\mathbf{u}}\mathbf{1}_{1\times L})$ are identical to the eigenvectors of the sample covariance matrix $\mathbf{C_u}$. This relation shows that for high-dimensional variables the PCA transformation matrix can be more efficiently computed by obtaining the left-singular vectors of $\frac{1}{\sqrt{L-1}}(\mathbf{U} - \bar{\mathbf{u}}\mathbf{1}_{1\times L})$, which is computationally more efficient for large n. One could therefore see the correspondence between the left-singular vectors of the sample data matrix and the eigenvectors of the data covariance [30]. It can be shown that amongst all S-term (rank-S) linear approximations of $\mathbf{U}$ the expansion using its S leading left-singular vectors (denoted as $\mathbf{\Phi}_{n\times S}$) gives the smallest root-mean-square error (RMSE).

Sparse Dictionary Learning (k-SVD) While PCA offers a very efficient decorrelating basis for compact representations, it is a linear transform in which the significant basis elements are predetermined and fixed. Recent developments in sparse signal processing have led to growing interest in sparse dictionary learning algorithms. A major distinction between PCA and sparse dictionaries is in the way the significant elements are selected. In sparse reconstruction, the significant elements are neither predetermined (ranked) nor fixed; rather, they must be identified independently for each instance of the parameter vector.

For construction of sparse dictionaries from a training dataset with L elements, $\mathbf{U}_{n\times L} = [\mathbf{u}_1 \dots \mathbf{u}_i \dots \mathbf{u}_L]$, one can solve either of the following optimization problems [2, 44, 82]:

$$\min_{[\mathbf{v}_1\mathbf{v}_2\dots\mathbf{v}_L],\boldsymbol{\Phi}} \quad \|\mathbf{v}_i\|_0 \qquad \text{s.t.,} \qquad \sum_{i=1}^{L} \|\mathbf{u}_i - \boldsymbol{\Phi}\mathbf{v}_i\|_2^2 \le \epsilon \quad \text{for} \quad i \in 1:L \tag{21a}$$

$$\min_{[\mathbf{v}_1\mathbf{v}_2\dots\mathbf{v}_L],\boldsymbol{\Phi}} \quad \sum_{i=1}^{L} \|\mathbf{u}_i - \boldsymbol{\Phi}\mathbf{v}_i\|_2^2 \qquad \text{s.t.,} \qquad \|\mathbf{v}_i\|_0 \le S \quad \text{for} \quad i \in 1:L \tag{21b}$$

where $\|\mathbf{v}_i\|_0$ refers to the number of non-zero entries in $\mathbf{v}_i$ (i.e. S). Equations (21a) and (21b) are alternative formulations for sparse dictionary learning. In Equation (21a), a maximum allowable representation error is used as a constraint while the level of sparsity for each realization of the prior model is minimized. In Equation (21b), the level of sparsity is constrained while minimizing the approximation error to represent each realization. Finding the exact solution to the problems in (21) is intractable. However, heuristic methods, such as the k-SVD algorithm, provide practical approximate solutions. We note that in our notation S refers to the sparsity level (number of significant elements retained in the approximation), and k is the dictionary size (total number of dictionary elements), with $S << k$. We briefly describe the k-SVD algorithm as one approach to learn sparse geologic dictionaries from a set of prior training models (more details can be found in the original publications [2]). The k-SVD algorithm takes its name from the k-means clustering algorithm. While the latter computes k mean values at each iteration, the former applies k SVD operations at each iteration. The k-SVD algorithm constructs a dictionary $\boldsymbol{\Phi}$ with size $n \times k$ from L samples of $\mathbf{u}_i$, while ensuring that the projection of each $\mathbf{u}_i$ on $\boldsymbol{\Phi}$ is S-sparse, a problem that is formalized in Equation (21). We also note that for model reduction and approximation purposes, we consider under-complete dictionaries, where exact representation may not be achievable. However, the resulting representation can provide close approximations in a very low-dimensional space.

To construct $\boldsymbol{\Phi}$ and $\mathbf{V}$ from $\mathbf{U}$, the k-SVD algorithm iteratively solves the problem specified in (21). Each iteration of the algorithm consists of two steps: Step 1, sparse coding, used to find the sparse representations (i.e. $\mathbf{V}$) for the entire prior library by fixing $\boldsymbol{\Phi}$; and Step 2, dictionary updating, which finds a new $\boldsymbol{\Phi}$ after fixing the sparse representation $\mathbf{V}$ from Step 1. These two basic steps in the k-SVD algorithm are summarized in Appendix 1. While no formal convergence proof has been given for this algorithm, numerical experiments show that it is generally robust [2, 44, 45]. It is important to note that the k-SVD algorithm is computationally demanding, especially when the dimension of the dictionary increases. Each iteration of the k-SVD algorithm requires k orthogonal matching pursuit (OMP) [83] sparse coding and k rank-one SVD operations, both are computationally expensive operations. However, the computations related to construction of a sparse dictionary are

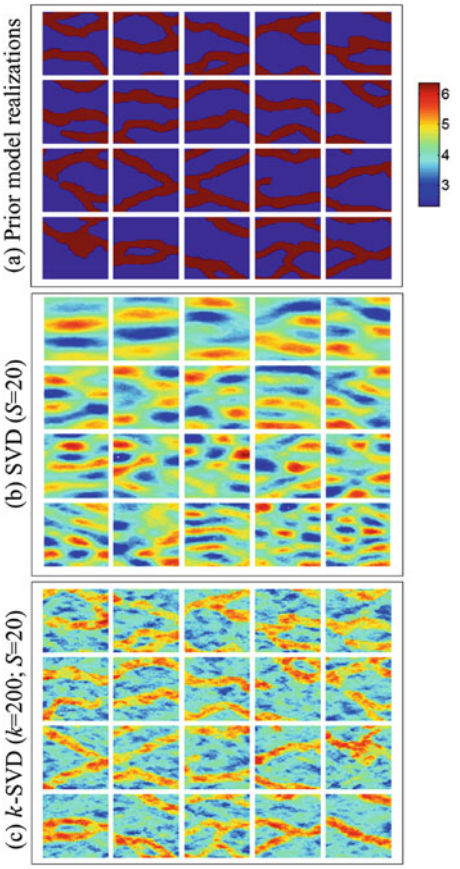

Fig. 3 Examples of learned expansion images using prior training data: (a) prior (training) models used for constructing linear expansion images; (b) $S = 20$ leading PCA basis elements; and (c) sample k-SVD dictionary elements with $S = 20$ and $k = 200$. Examples are shown for $n_x \times n_y = 100 \times 100$ two-dimensional model. The images are separated using white borders

performed offline and can be considered as part of the training step. In addition, the original k-SVD algorithm is typically applied to obtain over-complete dictionaries for small image segments [2]. For large-scale inverse problems, the method has been used to obtain under-complete dictionaries [44, 45].

Figure 3 shows an example of dictionary learning in geosciences applications. Figure 3(a) depicts samples from the training data that represent two-dimensional fluvial channel configurations (generated using SNESIM conditional simulation algorithm [75]). In this figure, the red regions represent fluvial channels that are composed of sandstone with very high rock permeability (to fluid flow) values while the blue regions describe shale or mudstone with very low-permeability values. The high-permeability values manifest their importance in fluid flow and displacement patterns by creating preferential flow patterns within the channel regions. Figure 3(b) presents the first $S = 20$ PCA basis (Eigen) images that

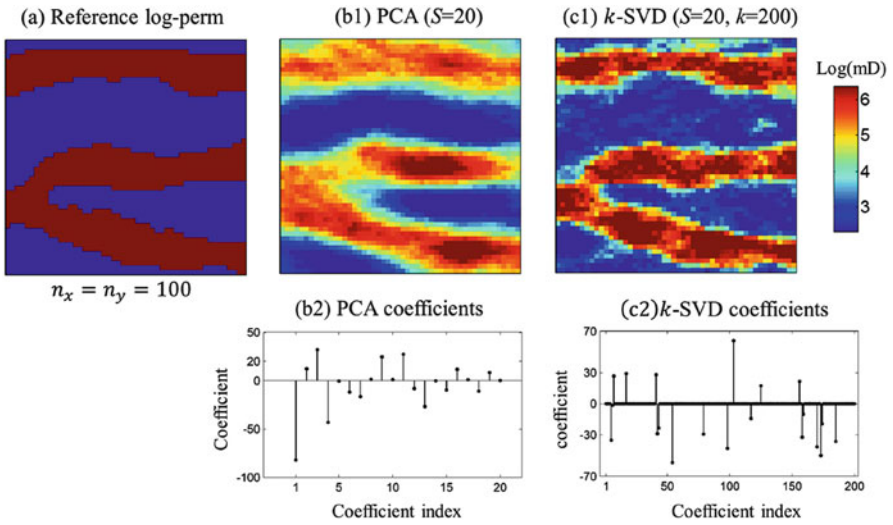

Fig. 4 Compression performance of the PCA and k-SVD: (a) a sample image similar to the training data; (b1)–(b2) results of compressed representation with $S = 20$ leading PCA basis images and the corresponding PCA coefficients, respectively; and (c1)–(c2) results of compressed representation using the k-SVD with $S = 20$ and $k = 200$ and the corresponding k-SVD coefficients

correspond to this training data, and Figure 3(c) shows the corresponding sample elements from the k-SVD dictionary, using $S = 20$ and $k = 200$. To illustrate the approximation performance of the PCA and k-SVD, Figure 4(a) depicts a model that is structurally similar to those in the training data, along with its PCA and k-SVD approximations in Figure 4(b1) and (c1), respectively, using $S = 20$. The corresponding transform coefficients for each case are shown in the second row (Figure 4(b2) and (c2)). Figure 4(b2) and (c2) shows a major difference between the PCA and k-SVD representations, which is the sorting of the PCA elements that leads to identification of $S = 20$ fixed elements. In case of k-SVD, the significant elements are not predetermined. Instead, the significant elements and their corresponding coefficients are identified by searching through a larger set ($k = 200$) of dictionary elements. However, the selection of significant dictionary elements is not trivial and is usually accomplished through a sparsity-promoting optimization algorithm, generally known as sparse reconstruction.

4 Sparse Reconstruction

Selecting a small subset of dictionary elements out of a large set is posed as a sparse reconstruction problem. A signal $\mathbf{v} \in \mathbb{R}^k$ is considered sparse if a large fraction of its entries are (approximately) zero [5]. A signal is S-sparse if it has at most S non-

zero entries. A signal that may not appear as sparse (in space or time) may have a sparse representation in a different (transform) domain. For instance, in many cases a parameter vector $\mathbf{u}$ may not be sparse but can have a sparse representation $\mathbf{v}$ after transformation through $\mathbf{\Phi}$, that is $\mathbf{u} = \mathbf{\Phi v}$.

Depending on the application, identification of significant dictionary elements can be based either on complete (e.g. image compression [80]) or incomplete knowledge (inverse problem) about the unknown parameters. In inverse problems, often limited measurements are available for identification of the significant dictionary elements, and estimation of their corresponding expansion coefficients. Compressed sensing (also called compressive sensing or compressive sampling) [4, 11, 21] is a relatively new paradigm that provides an alternative to the well-known Shannon sampling theory. Compressed sensing adopts sparsity as prior knowledge about signals, while Shannon theory was designed for frequency band-limited signals. The widespread application of compressed sensing is, in part, due to the universality of the sparsity property that is encountered in a wide range of natural phenomena (especially images). In many cases, sparsity may not be immediately apparent and certain manipulation (e.g. transformations) of the original parameters may be necessary for their sparsity to emerge. For instance, natural images that have various elements with spatial correlations in them do not exhibit sparsity in the space domain but are highly compressible and are well-known to have sparse representation in the Wavelet or DCT domains. One of the main contributors to the widespread application of compressed sensing is its direct application to solving underdetermined inverse problems, such as tomographic image reconstruction [15].

Compressed sensing gives a strong theoretical support and an efficient solution algorithm (under appropriate conditions) for solving otherwise intractable (NP-hard) inverse problems that have sufficiently sparse solutions. To recover a sparse solution $\mathbf{v}$ from a set of linear measurements $\mathbf{d} = \mathbf{Gv}$, one can solve the following minimization problem:

$$\min_{\mathbf{v}} \quad \|\mathbf{v}\|_0 \qquad \text{s.t.,} \qquad \mathbf{d} = \mathbf{Gv} \tag{22}$$

where $\|\mathbf{v}\|_0$ is the ℓ_0-norm (note that ℓ_0 does not conform to norm definition and is often loosely referred to as a norm) of vector $\mathbf{v}$ and represents its cardinality. In this formulation, the optimization problem searches for a solution that reproduces the observed data (constraint) while having a minimum number of non-zero entries (support). The ℓ_0-norm is not a differentiable function and does not lend itself to solution with standard gradient-based optimization methods. In practice, two types of approximate algorithms have been developed to solve (22): (i) greedy pursuit algorithms, such as OMP [83], COSAMP [61], IHT [7], or IMAT [56], and (ii) convex approximations, in which the non-convex ℓ_0-norm is replaced with its convex relaxations, e.g. ℓ_1-norm in basis pursuit [17] or a heuristically defined exponential norm in [58].

Compressed sensing derives the solution by replacing the ℓ_0-norm with ℓ_1-norm and offers conditions under which an exact solution to the original problem is

guaranteed (see [10, 21] for details). In this case, the optimization problem takes the form:

$$\min_{\mathbf{v}} \quad \|\mathbf{v}\|_1 \qquad \text{s.t.,} \qquad \mathbf{d} = \mathbf{G}\mathbf{v} \tag{23}$$

The fundamental importance of this formulation is that it converts the problem from an NP-hard problem to a linear programming problem, which can be solved efficiently. In practice, it can be demonstrated that the ℓ_p-norm, for $0 \leq p \leq 1$, while non-convex, has a similar sparsity-promoting property; however, in addition to solution complexity, the mathematical proof and the required conditions for this case are not well understood.

In many applications, the conditions required to guarantee exact solution may not be met. A particular example of departure from those conditions, which is often encountered in physical systems, is when the measurements are not adequate or the measurement operator is nonlinear. In those cases, it may still be possible to exploit the sparsity-promoting property of the ℓ_1-norm to formulate and solve an inverse problem. The selection property of the ℓ_1-norm penalty offers an important regularization form that can be used to enhance the solution of nonlinear inverse problems when applicable. When the measurement equations are nonlinear, the resulting sparse reconstruction problem takes the form:

$$\min_{\mathbf{v}} \quad \|\mathbf{v}\|_1 \qquad \text{s.t.,} \qquad \|\mathbf{d} - \mathbf{g}(\mathbf{v})\|_2^2 \leq \sigma^2 \tag{24}$$

where $\mathbf{g}(\mathbf{v})$ is a nonlinear operator. Appendix 2 discusses an iteratively reweighted least-squares (IRLS) algorithm for solving the ℓ_1-norm regularized minimization problem. In the next section, we discuss the application of sparsity regularization under nonlinear measurements in subsurface flow and transport inverse problems. In addition to ℓ_1-norm regularization, we will also present the use of a mixed ℓ_1/ℓ_2-norm [23, 40], which is known as group sparsity. When the signal of interest $\mathbf{v}$ has block-sparse behaviour, the ℓ_1/ℓ_2-norm can have a superior reconstruction performance compared to the standard ℓ_1-norm. In block-sparse signals, the entries are collected in predefined groups and the sparsity penalty is applied across the groups. In this case, the ℓ_2-norm is applied to the elements inside each group to quantify the group contribution, and the ℓ_1-norm operates on the computed ℓ_2-norm of the groups to impart sparsity. Mathematically, if $\mathbf{v}_i$'s are subsets of $\mathbf{v}$ and $\bigcup_i \mathbf{v}_i = \mathbf{v}$, then the ℓ_1/ℓ_2-norm is defined as $\|\mathbf{v}\|_{1,2} = \sum_i \|\mathbf{v}_i\|_2$. In this case, the inverse problem formulation minimizes the ℓ_1/ℓ_2-norm of the solution while honouring the measurement constraint, that is:

$$\min_{\mathbf{v}} \quad \|\mathbf{v}\|_{1,2} \qquad \text{s.t.,} \qquad \mathbf{d} = \mathbf{G}\mathbf{v} \qquad \text{linear} \tag{25a}$$

$$\min_{\mathbf{v}} \quad \|\mathbf{v}\|_{1,2} \qquad \text{s.t.,} \qquad \mathbf{d} = \mathbf{g}(\mathbf{v}) \qquad \text{nonlinear} \tag{25b}$$

An example application of group sparsity is presented in the next section. In this case, the objective is to select a small set of the groups within $\mathbf{v}$ that have significant contribution to the solution. In other words, the sparsity is applied to the groups and not individual entries.

5 Subsurface Flow Inverse Modelling

Fluid flow and transport in underground porous rock formations plays a key role in developing the related energy and water resources in these systems. Mathematical modelling of the underlying physical processes is commonly used to predict the response of these systems to perturbations (forcing) introduced during resource development (extraction or injection of fluids). The description of the physical processes that take place in these systems leads to high-dimensional and coupled nonlinear PDEs, which include various rock properties as spatially distributed unknown parameters. It is common to formulate inverse problems to estimate the unknown parameters of these PDEs from observations of the dynamical response of these systems. In this section, we describe the formulation of the related PDE-constrained inverse problem and provide examples to demonstrate their practical application.

5.1 Subsurface Flow Forward Modelling

An important example of PDE-constrained inverse problems is the multi-phase flow equations in the subsurface environments. The spatiotemporal evolution of multi-phase fluid flow can be expressed as a special form of the Navier-Stokes equations [19, 20]. Conservation of mass, momentum, and energy are three fundamental principles in the Navier-Stokes equations, which yield the following PDEs, respectively:

$$\frac{\partial \rho}{\partial t} + \nabla.(\rho \mathbf{v}) = 0 \tag{26a}$$

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}.\nabla)\mathbf{v} = -\frac{1}{\rho}\nabla P + \mathbf{F} + \frac{\mu}{\rho}\nabla^2 \mathbf{v} \tag{26b}$$

$$\rho\left(\frac{\partial E}{\partial t} + \mathbf{v}.\nabla E\right) - \nabla.(K_H \nabla T) + \rho P \nabla.\mathbf{v} = 0 \tag{26c}$$

where $\mathbf{v}$, E, P, T, ρ, μ, K_H, and $\mathbf{F}$ correspond to velocity, internal thermodynamic energy, pressure, temperature, density, viscosity, heat conduction coefficient, and external forces per unit mass. Here, we consider a special case involving two-phase incompressible and immiscible fluid flow system, for which the governing PDE

Table 1 A summary of physical properties and their definition

Property	Definition
Phase mobility	The ratio of effective permeability to phase viscosity
Phase density	The density of fluids, i.e. oil or water
Formation volume factor	Volume of the phase at the in-situ pressure to its volume at standard surface condition
Permeability	Ability for fluids (gas or liquid) to flow through porous rocks
Porosity	Ratio of void space to total rock volume
Phase saturation	Ratio of pore volume occupied by specific fluid phase
Flux	Flow rate per unit area
Pore volume	Total void volume of reservoir
Wetting phase	The phase with more tendency to maintain contact with the solid surface

equations are expressed by combining mass balance and Darcy's law (representing the momentum balance) [16, 22] as:

$$\nabla.(\frac{\lambda_{\mathrm{w}}}{B_{\mathrm{w}}}\mathbf{u}(\nabla P_{\mathrm{w}} - \gamma_{\mathrm{w}}\nabla Z)) = \frac{\partial}{\partial t}(\phi\frac{S_{\mathrm{w}}}{B_{\mathrm{w}}}) + q_{\mathrm{w}} \tag{27a}$$

$$\nabla.(\frac{\lambda_{\mathrm{n}}}{B_{\mathrm{n}}}\mathbf{u}(\nabla P_{\mathrm{n}} - \gamma_{\mathrm{n}}\nabla Z)) = \frac{\partial}{\partial t}(\phi\frac{S_{\mathrm{n}}}{B_{\mathrm{n}}}) + q_{\mathrm{n}} \tag{27b}$$

In the above equations, w and n represent the wetting and non-wetting phases, and λ, γ, B, $\mathbf{u}$, ϕ, Z, S, and q correspond to the phase mobility, phase density, formation volume factor, intrinsic rock permeability, rock porosity, gravity potential, phase saturation, and flux, respectively (see Table 1 for definitions). The governing equations in Equation (27) involve four unknown dynamic state variables: P_{n}, S_{n}, P_{w}, and S_{w}. Two additional equations are needed to close the PDE system. These two equations are the constitutive equations on the pressures and saturations and are typically expressed as:

$$P_{\mathrm{n}} - P_{\mathrm{w}} = P_c(S_{\mathrm{w}}) \tag{28}$$

$$S_{\mathrm{w}} + S_{\mathrm{n}} = 1 \qquad 0 \leq S_{\mathrm{w}}, S_{\mathrm{n}} \leq 1 \tag{29}$$

The first equation describes the capillary pressure (difference between non-wetting and wetting phase pressures) as a function of the wetting phase saturation (see Table 1 for definition) [49], while the second equation imposes a physical constraint on the saturation of two phases in a fully saturated porous medium.

With specified rock and fluid properties, initial and boundary conditions, and other input parameters and control forcing, the coupled PDE system can be discretized and solved numerically. In practice, the resulting discretized system can be high-dimensional ($\sim 10^{6-7}$ unknowns) and computationally demanding to solve. A simple example of immiscible two-phase flow, in which water is injected to displace oil, is depicted in Figure 5. Figure 5(a) shows a two-dimensional

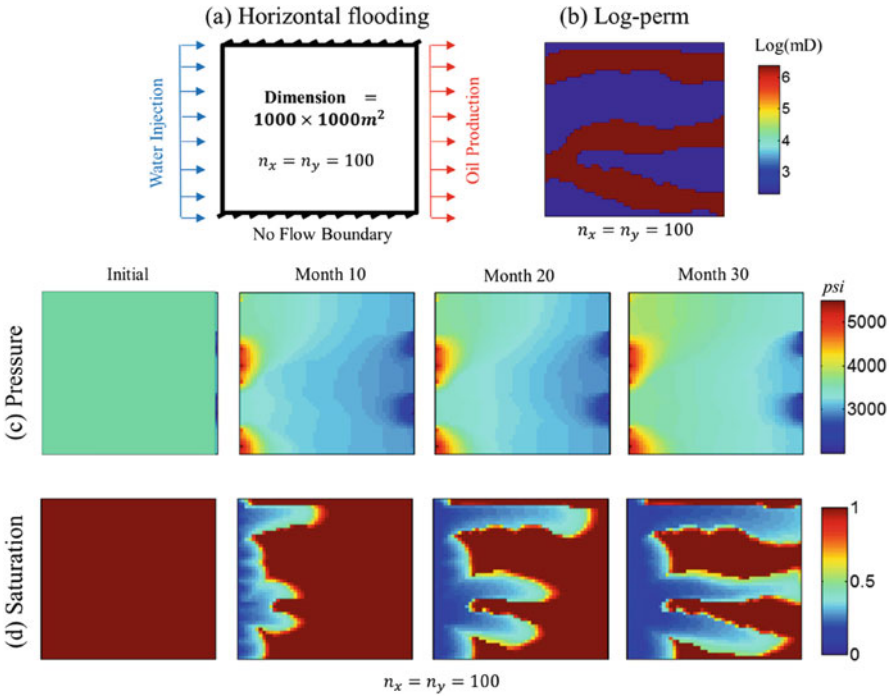

Fig. 5 The forward simulation model used in Example 1: (a) schematic of a reservoir with injection (production) wells on the left (right) side of the domain; (b) the intrinsic permeability distribution in the reference model consisting of high-permeability fluvial channels (red) and low-permeability background shale (blue); and snapshots of pressure (c) and saturation (d) profiles after 10, 20, and 30 months

($1000 \times 1000\,\mathrm{m}^2$) reservoir, which is discretized into 100×100 cells of the same size. A series of water injection wells are placed on the left side of the domain to displace the hydrocarbons toward a similar array of production wells placed on the right side. In this example, the capillary pressure is set to zero everywhere, that is $P_{\mathrm{n}}(x,t) = P_{\mathrm{w}}(x,t)$. Figure 5(b) depicts the intrinsic permeability distribution for this model, which shows a fluvial channel system with high-permeability (red) channels embedded in low-permeability (blue) background shale. As shown in the saturation plots of Figure 5(c), fluids move faster in the high-permeability channel sections. Figure 5(c) and (d) displays the solution of the PDE system as snapshots of pressure and saturation (S_{n}) fields at different times within the first 30 months of the simulation. In our example, the configuration includes the production wells (on the right) that produce water and oil, and injection wells (on the left) that inject water into the reservoir. Initially, the reservoir is fully saturated with the non-wetting phase (oil). Water injection into the reservoir displaces the oil from the left side toward the production wells on the right side, where the mixture of oil and water is extracted.

The forward simulation described above is used to predict the spatiotemporal evolution of the dynamical states (pressure and saturation distributions) of the

system for a given set of input parameters and controls. The state variables of the system are only observable through indirect measurements (e.g. flowrates and pressures) at scattered well locations. The related inverse problem can then be posed to find the unknown parameters (e.g. rock flow properties) from their limited, indirect, and nonlinear measurements.

5.2 Subsurface Flow Inverse Problem

Calibration of subsurface flow forward models against nonlinear dynamic data, i.e. data that are measured at different times and are nonlinearly related to parameters of interest, is commonly used to update model parameters and improve future model predictions [36]. Examples of dynamic data include time series of pressure or fluid flowrate measurements made at the well locations and differential images of fluid saturations typically obtained from seismic surveillance. In particular, dynamic data carry important information about heterogeneous rock flow properties, such as permeability distribution. The difficulty and cost associated with direct sampling from deep geologic formations necessitate the use of subjective assumptions and interpolations, which introduce significant uncertainty in the constructed rock properties. Calibration against dynamic data is a routine task performed to improve the description of these models (e.g. [44, 46]) and the related future forecasts. Dynamic flow data from scattered wells often contain spatially averaged information and offer limited resolution. Therefore, using high-resolution detailed models for unknown parameters can lead to discrepancy between data and model resolutions, a major contributor to the problem ill-posedness.

Prior models of parameters play a significant role in subsurface flow inverse modelling and are commonly used to constrain the inverse modelling solution. Of particular prominence in describing rock flow properties is the type and connectivity of geologic patterns that are expected in a given formation [75, 88]. Even qualitative knowledge about the depositional environment and the type of geologic features can be useful in eliminating implausible solutions. However, in solving the related inverse problems, it is important to acknowledge and reflect the uncertainty in the conceptual models of geologic continuity [25, 28, 43]. In this section, we present subsurface flow inverse modelling formulations that are developed by exploiting the selection property of the sparsity-promoting formulations that were discussed above.

We first consider the same setup in the forward model of Figure 5 and use the PCA and k-SVD representations to solve the corresponding inverse problem. A total of 2000 prior model realizations are generated using geostatistical simulation (Figure 3(a) shows 20 samples). The corresponding PCA and k-SVD basis images are shown in Figure 3(b) and (c), respectively. The ℓ_1-norm regularized formulation is applied to the k-SVD representation while a traditional parameterization using 20 leading basis elements is used for the PCA solution. More specifically, in the case of k-SVD the regularization term $J(\mathbf{v}) = \|\mathbf{v}\|_1$ is minimized along with the data

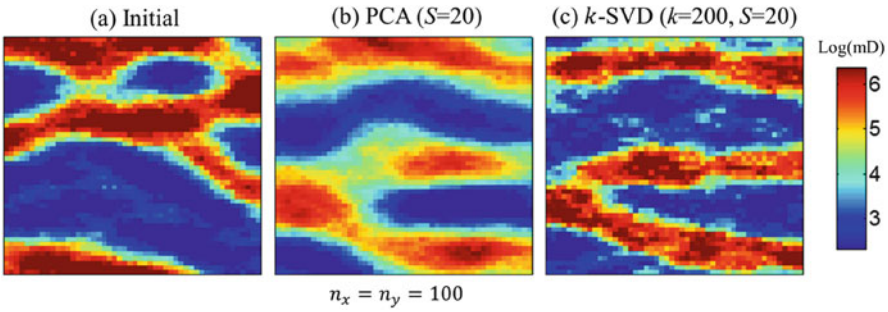

Fig. 6 Solution of the inverse problem in Example 1: (a) initial log-permeability distribution before data integration; (b) estimated log-permeability distribution using the PCA parameterization with $S = 20$; and (c) reconstructed log-permeability distribution using k-SVD with $S = 20$ and $k = 200$. The reference model is shown in Figure 5(b)

mismatch norm $\|\mathbf{d} - \mathbf{g}(\mathbf{u})\|_2^2$. In the case of PCA, the leading $S = 20$ basis elements are selected a-priori and used as parameterization basis vectors; hence, during inversion the coefficients corresponding to these elements are estimated without using the ℓ_1-norm regularization term. The initial model for the inversion is shown in Figure 6(a). The reconstruction results using the PCA and k-SVD descriptions are shown in Figure 6(b) and (c), respectively. The results show better estimation quality with the k-SVD representation and sparse reconstruction algorithm. The improved performance can be attributed to several factors, including flexibility in identifying the low-rank subspace during inversion (PCA provides a predetermined subspace), and better representation of geologic patterns that are not amenable to covariance-based description used in the PCA parameterization. Figure 7 depicts the data match and predictions obtained from the two methods, which seem to be comparable. It is important to note that while the two methods produce similar data matches, the solution from the k-SVD algorithm is visibly superior. This can be understood by recognizing the ill-posed nature of the problem, which implies that many solutions can be found to match the observed data. In this case, the k-SVD representation is better able to capture the connectivity structure in the prior model and the sparse reconstruction algorithm can recover the correct structure more accurately.

5.3 Uncertainty in Initial Geologic Scenario

Prior models of geologic continuity describe the type of geologic patterns and their connectivity. When used as prior model, geologic connectivity carries important weight in finding a solution to subsurface flow inverse problems. However, describing the exact form of connectivity from limited available data is a subjective process and depends on the geologist's interpretation. In generating a subsurface flow model, the connectivity patterns are typically constructed by integrating quantitative (well

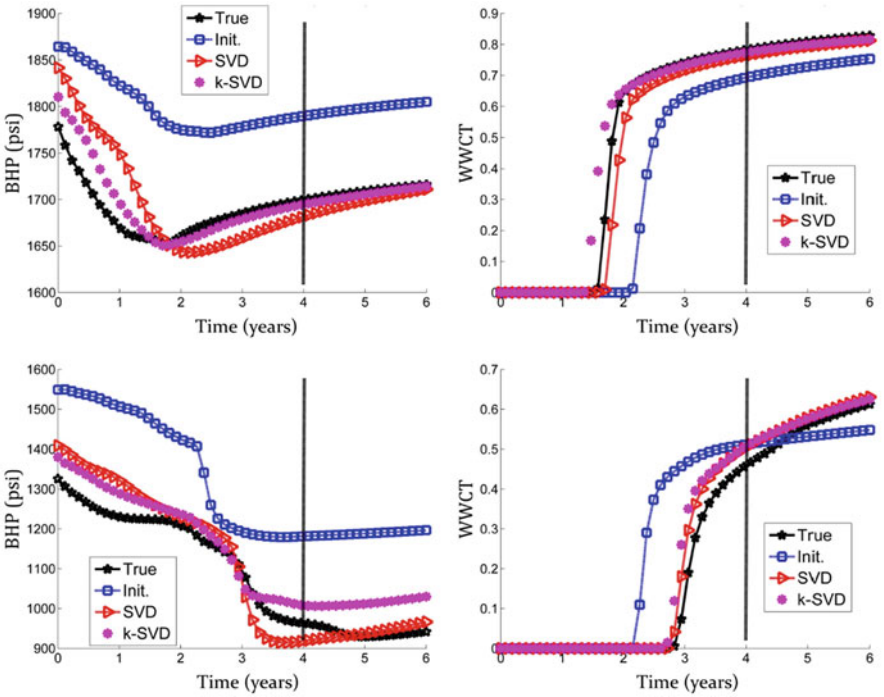

Fig. 7 Sample well data match (first four years) and prediction (last two years) results for Example 1. The BHP and watercut observations at two sample production wells are shown (the producers are under total production rate control)

log, core analysis, and seismic data) and qualitative information (e.g. outcrops) with expert knowledge and interpretation as well as process-based geologic modelling of the depositional environment. Traditionally, a single conceptual model of continuity (e.g. variogram model) is constructed and used to constrain the solution of the inverse problem, assuming perfect knowledge about the continuity model. However, a major source of uncertainty is related to the adopted conceptual model of geologic continuity. Adhering to a single conceptual geologic scenario can lead to underestimation of the initial uncertainty in the prior models and result in solutions that depend heavily on the quality of the adopted prior model (which can be questionable) [25, 28, 38, 43, 46, 67, 68, 73, 76].

Another important implication of adopting a single geologic scenario is eliminating the opportunity to confirm, reject, or correct a proposed geologic scenario based on dynamic data. In generating or selecting the prior geologic scenario, dynamic data is typically not included (usually dynamic data are obtained at later stages). A simple way to address this issue is to include multiple plausible geologic scenarios as possible prior models [25, 28, 38, 43]. These alternative geologic scenarios could be developed as independent interpretation of existing data by different geologists (experts), or they could be derived from a stochastic process-

based geologic modelling framework. Inverse modelling can then be applied to evaluate the plausibility of the proposed geologic scenarios based on available dynamic data. Inversion methods that can incorporate multiple geologic scenarios are not widely studied in the literature. In this section, we present one such inversion method by exploiting the selection property of sparsity-promoting regularization techniques, or more specifically the group-sparsity regularization.

The group-sparsity regularization is implemented by minimizing the ℓ_1/ℓ_2-norm to identify relevant geologic scenarios (from a list of proposed scenarios) based on dynamic flow-related data. Consider p alternative geologic scenarios, each used to generate L different realizations as prior models; that is, $\mathbf{U}_1 = [\mathbf{u}_{11}\mathbf{u}_{12}\dots\mathbf{u}_{1L}], \dots, \mathbf{U}_p = [\mathbf{u}_{p1}\mathbf{u}_{p2}\dots\mathbf{u}_{pL}]$ are p sets of prior model realizations in which the columns of $\mathbf{U}_i = [\mathbf{u}_{i1}\mathbf{u}_{i2}\dots\mathbf{u}_{iL}]$ represent L realizations from the i^{th} geologic scenarios. If the prior model realizations for each scenario are used to generate p different PCA bases, then a hybrid dictionary can be constructed to include all the bases $\mathbf{\Phi} = [\mathbf{\Phi}_1\mathbf{\Phi}_2\dots\mathbf{\Phi}_p]$. Here, the realizations $\mathbf{U}_i$ for each geologic scenario are used to generate a corresponding PCA basis $\mathbf{\Phi}_i = [\boldsymbol{\phi}_{i1}\dots\boldsymbol{\phi}_{is_i}]$, where s_i is the size of low-rank representation. Using this hybrid dictionary, the parameter of interest $\mathbf{u}$ is approximated through a linear expansion of the form $\mathbf{u} = \mathbf{\Phi}\mathbf{v} = [\mathbf{\Phi}_1\mathbf{\Phi}_2\dots\mathbf{\Phi}_p][\mathbf{v}_1;\mathbf{v}_2;\dots;\mathbf{v}_p]$. This formulation implies that all prior geologic scenarios have a chance to represent the solution. However, the underlying assumption is that many of the included prior scenarios are not relevant and should not contribute to reconstruction of the solution. Hence, only very few (if not just one) of the groups are expected to have non-zero weights.

Using a mixed ℓ_1/ℓ_2-norm for group sparsity [29, 54], the regularized objective function of the inverse problem can be expressed as:

$$\begin{aligned} &\min_{\mathbf{v}} \quad J(\mathbf{v}) = \sum_{i=1}^{p} \|\mathbf{v}_i\|_2 \quad \text{s.t.,} \quad \|\mathbf{d} - \mathbf{g}(\mathbf{u})\|_2^2 \le \sigma^2 \\ &\text{and} \qquad \mathbf{u} = \mathbf{\Phi}\mathbf{v} = [\mathbf{\Phi}_1\mathbf{\Phi}_2\dots\mathbf{\Phi}_p][\mathbf{v}_1;\mathbf{v}_2;\dots;\mathbf{v}_p] \end{aligned} \tag{30}$$

After solving this inverse problem, the solution $\mathbf{u}$ and the geologic scenario(s) that significantly contribute to constructing it are identified simultaneously (for more details, see [28]). Appendix 3 presents the details of solving the optimization problem in Equation (30).

The example in this section consists of a numerical two-phase flow in a heterogeneous reservoir for which the intrinsic permeability values in the entire field are unknown. The reservoir model has a dimension of $1000 \times 1000 \times 10\,\mathrm{m}^3$, which is discretized into a $100 \times 100 \times 1$ uniform grid system. Figure 8(a) depicts the configuration of this water-flooding example. An injection well is placed in the middle of the field and eight producers are located along the edges of the reservoir to build a traditional 9-spot water-flooding scheme. A total of 0.8 pore volume (PV) (see Table 1 for definition) of water is injected into the formation during the first 4 years of water flooding. Also, 0.4 PV of water is injected in the following two years,

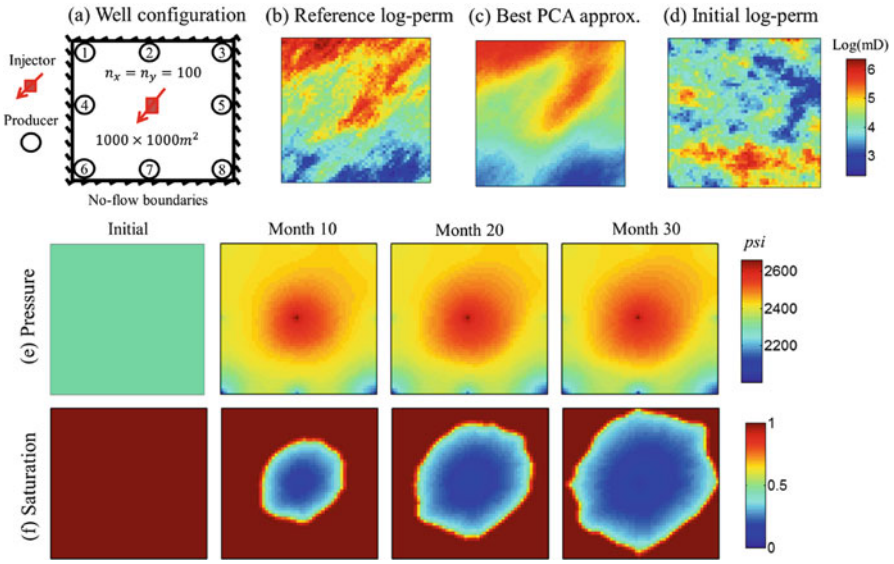

Fig. 8 The forward simulation model used in Example 2: (a) schematic of a reservoir with one injection well in the centre and eight production wells symmetrically distributed along the edges of the domain; (b) the intrinsic log-permeability distribution in the reference model; (c) the best achievable log-permeability estimate when the correct prior geologic scenario is known (Scenario 6); (d) the initial log-permeability distribution, assuming equal contributions from each prior variogram model; and snapshots of pressure (e) and saturation (f) profiles after 10, 20, and 30 months

during the prediction phase. The porosity of the field is assumed to be 0.25 for the entire field, and oil/water viscosity ratio is set to 1. The pressure at the injection well and the total (water and oil) flowrates at the production wells are measured every 40 days and used for inversion. The reference log-permeability map along with its best achievable PCA approximation, and the initial log-permeability map before inversion are shown in Figure 8(b)–(d), respectively. The initial log-permeability map considers equal weight given to all 12 groups. Figure 8(e) and (f) displays the pressure and saturation profiles, respectively, after 10, 20 and 30 months.

To reflect the uncertainty in the prior variogram model, the direction of maximum continuity and the minimum and maximum variogram ranges are assumed to be uncertain. The variogram parameters for these prior models and four samples from the corresponding realizations are shown in Figure 9. Comparing the reference model with the realizations generated using these 12 different prior geologic models reveals that the consistent model belongs to Scenario 6. The projection of the reference map onto the PCA basis corresponding to Scenario 6 is shown in Figure 8(c). Other models either present different directions of global continuity or inaccurate ranges. The realizations from these 12 variogram models are used to build 12 different PCA bases, which are combined to form a hybrid dictionary.

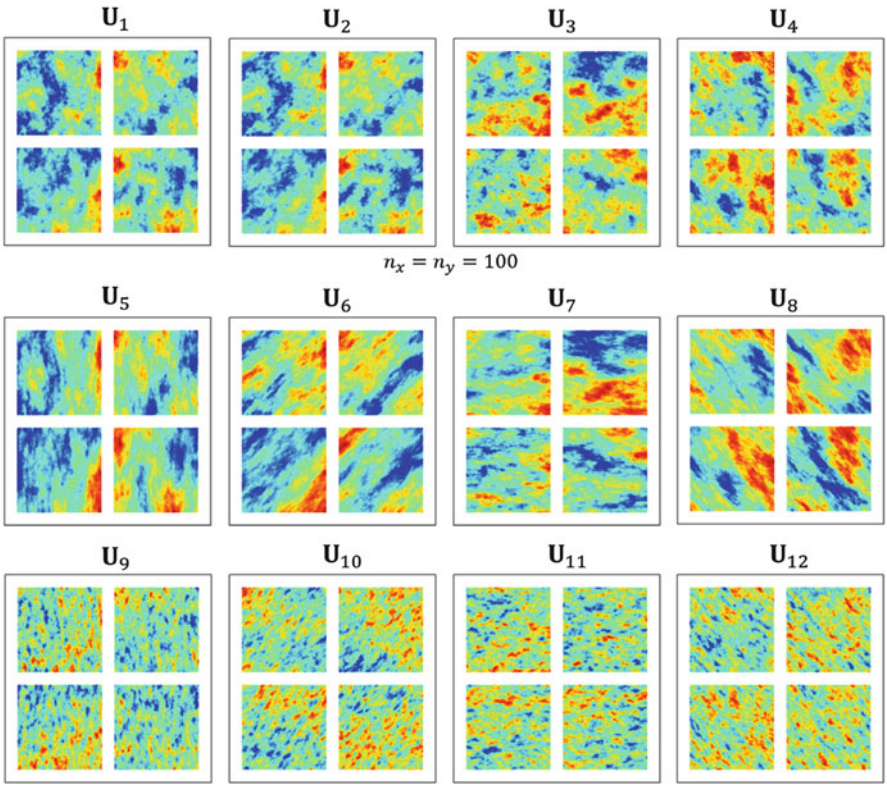

Fig. 9 Alternative prior training data derived from 12 different variogram models; each box contains four sample realization from the training data corresponding to a variogram model; the alternative variogram models are obtained by using three variogram range combinations ($a_{max} = 300m, a_{min} = 240m$), ($a_{max} = 600m, a_{min} = 300m$), and ($a_{max} = 100m, a_{min} = 60m$) and four different azimuth values . The reference model is consistent with Scenario 6 with training data $\mathbf{U}_6$

Figure 10 depicts the inversion solution at different iterations. The initial model (Figure 8(d) and the top row of Figure 10) is projected onto all elements of the hybrid dictionary, and all the prior geologic scenarios equally contribute to the representation of the initial model. The global continuity in the permeability field is captured within the first few iterations. At later iterations, the regularization term fine-tunes the solution by selecting the geologic scenarios that best represent the estimated parameter. Group 6, which has the correct variogram model, has been identified as the most significant prior model (with largest ℓ_2-norm) after convergence of the group-sparsity inversion algorithm. Figure 11 shows a summary of the data match and forecast performance of the solution compared to the initial and reference models in two production wells. The data match and prediction results clearly show the improvements achieved after model calibration.

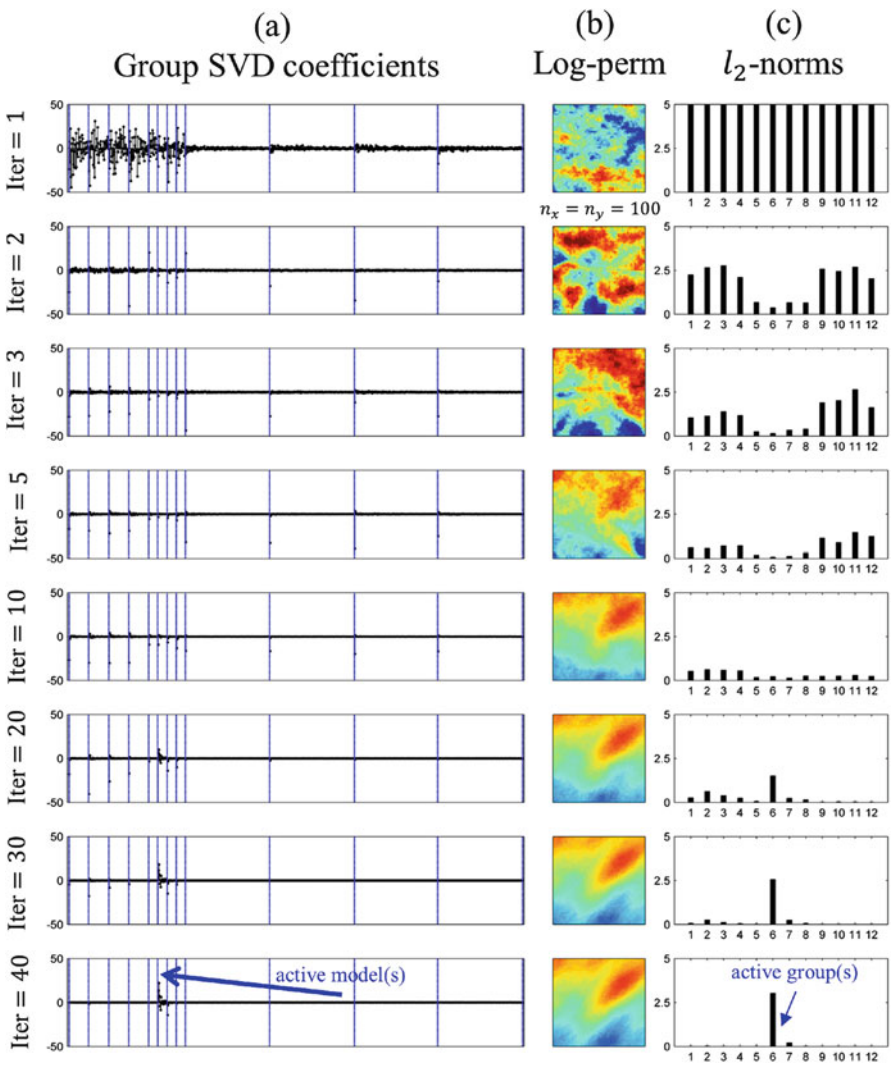

Fig. 10 Results of group-sparsity inversion iterations for Example 2: (a) the coefficients of the expansion using 12 different groups (PCA bases); (b) reconstructed log-permeability map; and (c) the ℓ_2-norm of the coefficients of the PCA representation in each group. Groups with larger ℓ_2-norm have greater contributions to the solution and persist during the iterations, whereas irrelevant groups are assigned insignificant group ℓ_2-norm. Group 6 in this example stay active with a large ℓ_2-norm while other groups disappear during inversion iterations

Two alternative ways may also be used to formulate and solve the above inverse problem: (i) by using the same parameterization, i.e. $\mathbf{\Phi} = [\mathbf{\Phi}_1 \mathbf{\Phi}_2 \dots \mathbf{\Phi}_p]$, with ℓ_1-norm regularization (without group sparsity), and (ii) by combining all the prior models and generating a single PCA parameterization. However, both of

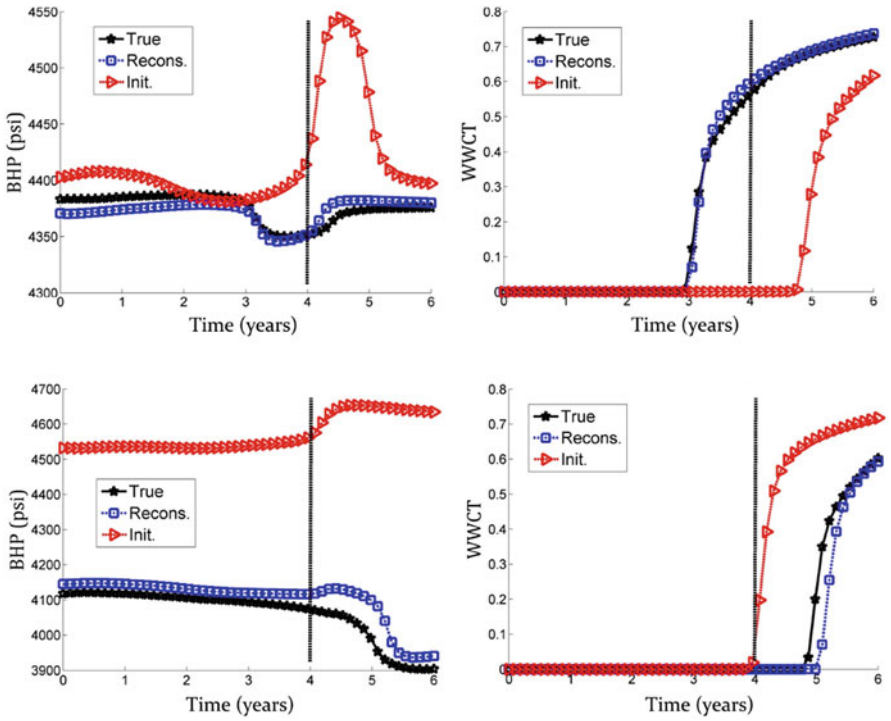

Fig. 11 The pressure and watercut data match (first four years) and forecasts (last two years) for sample production wells. The group-sparsity regularization not only identifies the correct variogram model, it also provides a calibrated model at convergence

these approaches provide inferior solutions. In the first case, the group sparsity, by formulation, has been shown to be more effective in reconstructing the solution as it imposes a stronger constraint on the problem. In the second case, a simple least-square formulation is solved to search for the PCA coefficients in the low-rank subspace defined by the leading PCs, which are not representative of any particular prior (as they represent an aggregate of all prior models).

6 Conclusion

In this chapter, we discussed a general formulation for solving sparse PDE-constrained inverse problems. In particular, we presented sparse inverse problem formulations that use sparsity to regularize ill-posed problems that can arise in various applications. Sparsity is an inherent property of many types of natural images and can be used to improve the solution of ill-posed inverse problems in which the solutions have sparse representations. Examples from multiphase fluid flow in subsurface rock formations, which involve the solution of coupled

PDEs to describe the underling physical processes, were used to demonstrate the effectiveness of the method. To calibrate heterogeneous subsurface flow models against dynamic data, scattered nonlinear measurements of flowrate and pressure are often used. Spatially distributed rock flow properties are known to have a sparse representation in a properly designed basis. High-resolution grid-based description of these properties leads to over-parameterization. When combined with data scarcity, over-parameterized descriptions often lead to problem ill-posedness, introducing great difficulty in solving these inverse problems. Furthermore, prior geologic scenarios that are typically used to regularize these ill-posed inverse problems often involve significant uncertainty that should be taken into account in formulating and solving these problems. We propose the use of learned sparse geologic dictionaries and sparsity-promoting regularization functions as powerful and robust approaches to address these issues. Specifically, we present a formulation in which prior models are used as training data to learn sparse representations of rock flow properties. We show that by promoting sparsity through minimization of regular ℓ_1-norm of the solution in the learned k-SVD dictionary (along with minimization of the predicted and observed data mismatch term) a better-posed inverse problem can be obtained to reconstruct complex geologic patterns. In addition, group-sparsity regularization that minimizes a mixed ℓ_1/ℓ_2-norm was used to discriminate against multiple prior geologic scenarios using flow data. An important implication of the latter is that it allows the use of dynamic flow data in selecting, rejecting, and correcting prior geologic scenarios, a novel concept that can improve traditional subsurface flow model calibration workflows.

Acknowledgements The content of this chapter is based on research partially funded by the US Department of Energy, Foundation CMG, and American Chemical Society.

Appendix 1: k-SVD Dictionary Learning

The k-SVD algorithm is used to construct learned sparse dictionaries from a training dataset. The algorithm is similar to the k-means clustering method and is designed to find a dictionary $\mathbf{\Phi} \in \mathbb{R}^{n\times k}$ containing k elements that sparsely represent each of the training samples in $\mathbf{U}_{n\times L} = [\mathbf{u}_1 \dots \mathbf{u}_i \dots \mathbf{u}_L]$. To achieve this goal, the algorithm attempts to solve the following minimization problem:

$$\hat{\mathbf{V}}, \hat{\mathbf{\Phi}} = \text{argmin}_{\mathbf{V},\mathbf{\Phi}} \quad \sum_{i=1}^{L} \|\mathbf{u}_i - \mathbf{\Phi}\mathbf{v}_i\|_2^2 \qquad \text{s.t.,} \qquad \|\mathbf{v}_i\|_0 \leq S \quad \text{for} \quad i \in 1:L \tag{31}$$

where $\mathbf{V}_{k\times L} = [\mathbf{v}_1 \dots \mathbf{v}_i \dots \mathbf{v}_L]$ are the expansion coefficients corresponding to the training data. Given the NP-hard nature of the problem, the k-SVD algorithm uses a heuristic greedy solution technique by dividing the above optimization problem into

Table 2 k-SVD algorithm

Initialization: Initialize dictionary with $\mathbf{\Phi}^{(0)} \in \mathbb{R}^{n\times k}$. Set $j = 1$.
REPEAT until stopping criteria is met
a. Sparse Coding Step:
-Using a pursuit algorithm (e.g. OMP) compute $\mathbf{V}_{k\times L}^{(j)} = [\mathbf{v}_1\mathbf{v}_2 \dots \mathbf{v}_L]$ as the solution of
$\mathbf{V}^{(j)} = \text{argmin}_{\mathbf{v}_i} \quad \|\mathbf{u}_i - \mathbf{\Phi}^{(j-1)}\mathbf{v}_i\|_2^2 \qquad \text{s.t.,} \qquad \|\mathbf{v}_i\|_0 \le S \quad \text{for} \quad i \in 1:L$
b. Dictionary Update Step:
For each column $c = 1, 2, \dots, k$ in $\mathbf{\Phi}^{(j-1)}$
-Define the group of prior model instances that use this element
$\omega_c = \{i \vert 1 \le i \le L, \mathbf{V}^{(j)}(c, i) \ne 0\}$
-Compute the residual matrix $\mathbf{E}_c = \mathbf{U} - \sum_{i\ne c} \boldsymbol{\phi}_i {\mathbf{v}_c}^T$, where ${\mathbf{v}_c}^T$ is the c^{th} row of $\mathbf{V}^{(j)}$
-Restrict $\mathbf{E}_c$ by choosing columns corresponding to ω_c , i.e. find $\mathbf{E}_c^{\omega}$
-Apply rank-1 SVD decomposition $\mathbf{E}_c^{\omega} = \mathbf{A\Delta B}$
-Update the dictionary element $\boldsymbol{\phi}_c = \mathbf{a}_1$ and the sparse representation $\mathbf{v}_c$ by $\mathbf{v}_c^{\omega} = \mathbf{\Delta} \mathbf{b}_1$
-END

two subproblems: (i) sparse coding and (ii) dictionary update. In the sparse coding step, for the current dictionary, a basis pursuit algorithm is used to find the sparse representation for each member of the training dataset. In the dictionary update step, the sparse representation obtained in the first step is fixed and the dictionary elements are updated to reduce the sparse approximation error. These two steps are repeated until convergence. Table 2 summarizes the k-SVD algorithm. Further details about the k-SVD algorithm may be found in [2]. We note that for high-dimensional training data the k-SVD dictionary learning can be computationally expensive. The computational complexity of each iteration of k-SVD is $O(L(2nk + S^2k + 7Sk + S^3 + 4Sn) + 5nk^2)$, where S is the sparsity level. One strategy to improve the computational efficiency of the algorithm includes using segmentation or approximate low-rank representations of the training data (to reduce n).

Appendix 2: IRLS Algorithm

We use the IRLS algorithm [14] to solve the ℓ_1-norm regularized least-square minimization problem, that is:

$$\min_{\mathbf{v}} \quad J(\mathbf{v}) = \|\mathbf{v}\|_1 + \lambda^2 \|\mathbf{d} - \mathbf{g}(\mathbf{\Phi v})\|_2^2 \tag{32}$$

At iteration n of the IRLS algorithm, the ℓ_1-norm is approximated using a weighted ℓ_2-norm as follows:

$$\min_{\mathbf{v}^{(n)}} \quad J(\mathbf{v}^{(n)}) = \sum_i w_i^{(n)} {v_i^{(n)}}^2 + \lambda^2 \|\mathbf{d} - \mathbf{g}(\mathbf{\Phi v}^{(n)})\|_2^2 \tag{33}$$

where $w_i^{(n)} = \frac{1}{({v_i^{(n-1)}}^2 + \epsilon^{(n)})^{0.5}}$, (n) stands for the iteration n, and $\epsilon^{(n)}$ is a sequence of small numbers (that converge to zero with increasing n). Using this approximation of the objective function, and a first-order Taylor expansion for $\mathbf{g}(\mathbf{\Phi v}^{(n)})$, the objective function in (33) takes the form:

$$\min_{\mathbf{v}^{(n)}} \quad J(\mathbf{v}^{(n)}) = \sum_i w_i^{(n)} {v_i^{(n)}}^2 + \lambda^2 \|\mathbf{d} - \mathbf{g}(\mathbf{\Phi v}^{(n-1)}) - {\mathbf{G_v}}^{(n)}(\mathbf{v}^{(n)} - \mathbf{v}^{(n-1)})\|_2^2 \tag{34}$$

Here, ${\mathbf{G_v}}^{(n)}$ is the Jacobian matrix of $\mathbf{g}(.)$ with respect to $\mathbf{v}$ at $\mathbf{v} = \mathbf{v}^{(n-1)}$. The updated solution at iteration n can be easily found by taking the derivative of the above convex function w.r.t. $\mathbf{v}^{(n)}$ and setting it to zero.

Appendix 3: Group-Sparsity Inversion

The objective function for group-sparsity regularization can be expressed as:

$$\min_{\mathbf{v}} \quad J(\mathbf{v}) = \sum_{i=1}^{p} \|\mathbf{v}_i\|_2 + \lambda^2 \|\mathbf{d} - \mathbf{g}(\mathbf{\Phi v})\|_2^2 \tag{35}$$

where the notations are discussed in the text. At iteration n, using the Gauss-Newton method and the first-order Taylor series for $\mathbf{g}(\mathbf{\Phi v})$, the linearized version of the above function takes the form:

$$\min_{\mathbf{v}^{(n)}} \quad J(\mathbf{v}^{(n)}) = \sum_{i=1}^{p} (\sum_{j=1}^{s_i} ({v_i^j}^{(n)})^2)^{\frac{1}{2}} + \lambda^2 \|\mathbf{d} - \mathbf{g}(\mathbf{\Phi v}^{(n-1)}) - {\mathbf{G_v}}^{(n)}(\mathbf{v}^{(n)} - \mathbf{v}^{(n-1)})\|_2^2 \tag{36}$$

where ${\mathbf{G_v}}^{(n)}$ is the Jacobian matrix of $\mathbf{g}(\mathbf{v})$, and v_i^j is the jth basis in the ith group. Denoting $\mathbf{\Delta d}^{(n)} = \mathbf{d} - \mathbf{g}(\mathbf{\Phi v}^{(n-1)}) + {\mathbf{G_v}}^{(n)}\mathbf{v}^{(n-1)}$, (36) can be simplified to:

$$\min_{\mathbf{v}^{(n)}} \quad J(\mathbf{v}^{(n)}) = \sum_{i=1}^{p} (\sum_{j=1}^{s_i} ({v_i^j}^{(n)})^2)^{\frac{1}{2}} + \lambda^2 \|\mathbf{\Delta d}^{(n)} - {\mathbf{G_v}}^{(n)}\mathbf{v}^{(n)}\|_2^2 \tag{37}$$

The derivative of the regularization term with respect to ${v_i^j}^{(n)}$ can be approximated as:

$$\frac{{v_i^j}^{(n)}}{(\sum_{k=1}^{s_i} ({v_i^k}^{(n)})^2)^{\frac{1}{2}}} \approx \frac{{v_i^j}^{(n)}}{(\sum_{k=1}^{s_i} ({v_i^k}^{(n-1)})^2 + {\epsilon_i}^{(n)})^{\frac{1}{2}}} \tag{38}$$

where ${\epsilon_i}^{(n)}$ is a small positive number that is used to avoid zero denominators. Note that ${v_i^k}^{(n)}$ in the denominator is approximated as ${v_i^k}^{(n-1)}$. Choosing ϵ such that $0 < {\epsilon_i}^{(n)} < {\epsilon_i}^{(n-1)}$ and $\lim_{n\to\infty} {\epsilon_i}^{(n)} = 0$, it can be shown that this approximation does

not change the solution of the original minimization problem. The iterative solution of (37) can now be derived as:

$$(\mathbf{\Lambda}^{(n)} + \alpha {\mathbf{G_v}^{(n)}}^T \mathbf{G_v}^{(n)})\mathbf{v}^{(n)} = \alpha {\mathbf{G_v}^{(n)}}^T \mathbf{\Delta d}^{(n)} \tag{39}$$

where $\alpha = 2\lambda^2$, and $\mathbf{\Lambda}^{(n)}$ is a diagonal matrix with diagonal entries $\frac{1}{(\sum_{k=1}^{s_i}(v_i^{k(n-1)})^2+\epsilon_i{}^{(n)})^{\frac{1}{2}}}$.

References

1. Aanonsen SI, Nævdal G, Oliver DS, Reynolds AC, Vallès B, et al (2009) The ensemble Kalman filter in reservoir engineering–a review. Spe Journal 14(03):393–412
2. Aharon M, Elad M, Bruckstein A (2006) *rmk*-svd: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on signal processing 54(11): 4311–4322
3. Ahmed N, Natarajan T, Rao KR (1974) Discrete cosine transform. IEEE transactions on Computers 100(1):90–93
4. Baraniuk RG (2007) Compressive sensing [lecture notes]. IEEE signal processing magazine 24(4):118–121
5. Berinde R, Gilbert AC, Indyk P, Karloff H, Strauss MJ (2008) Combining geometry and combinatorics: A unified approach to sparse signal recovery. In: Communication, Control, and Computing, 2008 46th Annual Allerton Conference on, IEEE, pp 798–805
6. Bhark EW, Jafarpour B, Datta-Gupta A (2011) A generalized grid connectivity–based parameterization for subsurface flow model calibration. Water Resources Research 47(6)
7. Blumensath T, Davies ME (2009) Iterative hard thresholding for compressed sensing. Applied and computational harmonic analysis 27(3):265–274
8. Boyd S, Vandenberghe L (2004) Convex optimization. Cambridge university press
9. Bracewell RN, Bracewell RN (1986) The Fourier transform and its applications, vol 31999. McGraw-Hill New York
10. Candes EJ (2008) The restricted isometry property and its implications for compressed sensing. Comptes Rendus Mathematique 346(9–10):589–592
11. Candès EJ, Wakin MB (2008) An introduction to compressive sampling. IEEE signal processing magazine 25(2):21–30
12. Carrera J, Neuman SP (1986) Estimation of aquifer parameters under transient and steady state conditions: 1. maximum likelihood method incorporating prior information. Water Resources Research 22(2):199–210
13. Chandrasekaran V, Recht B, Parrilo PA, Willsky AS (2012) The convex geometry of linear inverse problems. Foundations of Computational mathematics 12(6):805–849
14. Chartrand R, Yin W (2008) Iteratively reweighted algorithms for compressive sensing. In: Acoustics, speech and signal processing, 2008. ICASSP 2008. IEEE international conference on, IEEE, pp 3869–3872
15. Chen GH, Tang J, Leng S (2008) Prior image constrained compressed sensing (PICCS): a method to accurately reconstruct dynamic CT images from highly undersampled projection data sets. Medical physics 35(2):660–663
16. Chen S, Doolen GD (1998) Lattice Boltzmann method for fluid flows. Annual review of fluid mechanics 30(1):329–364
17. Chen SS, Donoho DL, Saunders MA (2001) Atomic decomposition by basis pursuit. SIAM review 43(1):129–159
18. Chen Y, Oliver DS (2012) Multiscale parameterization with adaptive regularization for improved assimilation of nonlocal observation. Water resources research 48(4)

19. Chorin AJ (1968) Numerical solution of the Navier-Stokes equations. Mathematics of computation 22(104):745–762
20. Constantin P, Foias C (1988) Navier-stokes equations. University of Chicago Press
21. Donoho DL (2006) Compressed sensing. IEEE Transactions on information theory 52(4):1289–1306
22. Efendiev Y, Durlofsky L, Lee S (2000) Modeling of subgrid effects in coarse-scale simulations of transport in heterogeneous porous media. Water Resources Research 36(8):2031–2041
23. Eldar YC, Kuppinger P, Bolcskei H (2010) Block-sparse signals: Uncertainty relations and efficient recovery. IEEE Transactions on Signal Processing 58(6):3042–3054
24. Engl HW, Hanke M, Neubauer A (1996) Regularization of inverse problems, vol 375. Springer Science & Business Media
25. Feyen L, Caers J (2006) Quantifying geological uncertainty for flow and transport modeling in multi-modal heterogeneous formations. Advances in Water Resources 29(6):912–929
26. Gavalas G, Shah P, Seinfeld JH, et al (1976) Reservoir history matching by Bayesian estimation. Society of Petroleum Engineers Journal 16(06):337–350
27. Gholami A (2015) Nonlinear multichannel impedance inversion by total-variation regularization. Geophysics 80(5):R217–R224
28. Golmohammadi A, Jafarpour B (2016) Simultaneous geologic scenario identification and flow model calibration with group-sparsity formulations. Advances in Water Resources 92:208–227
29. Golmohammadi A, Khaninezhad MRM, Jafarpour B (2015) Group-sparsity regularization for ill-posed subsurface flow inverse problems. Water Resources Research 51(10):8607–8626
30. Golub G, Kahan W (1965) Calculating the singular values and pseudo-inverse of a matrix. Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis 2(2):205–224
31. Golub GH, Heath M, Wahba G (1979) Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21(2):215–223
32. Gómez-Hernánez JJ, Sahuquillo A, Capilla J (1997) Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data-i. theory. Journal of Hydrology 203(1–4):162–174
33. Grimstad AA, Mannseth T, Nævdal G, Urkedal H (2003) Adaptive multiscale permeability estimation. Computational Geosciences 7(1):1–25
34. Hansen PC (1992) Analysis of discrete ill-posed problems by means of the l-curve. SIAM review 34(4):561–580
35. Hansen PC (1998) Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion. SIAM
36. Hill MC, Tiedeman CR (2006) Effective groundwater model calibration: with analysis of data, sensitivities, predictions, and uncertainty. John Wiley & Sons
37. Jacquard P, et al (1965) Permeability distribution from field pressure data. Society of Petroleum Engineers Journal 5(04):281–294
38. Jafarpour B, Tarrahi M (2011) Assessing the performance of the ensemble Kalman filter for subsurface flow data integration under variogram uncertainty. Water Resources Research 47(5)
39. Jafarpour B, McLaughlin DB, et al (2009) Reservoir characterization with the discrete cosine transform. SPE Journal 14(01):182–201
40. Jenatton R, Obozinski G, Bach F (2010) Structured sparse principal component analysis. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp 366–373
41. Jolliffe IT (1986) Principal component analysis and factor analysis. In: Principal component analysis, Springer, pp 115–128
42. Kandel ER, Schwartz JH, Jessell TM, Siegelbaum SA, Hudspeth AJ, et al (2000) Principles of neural science, vol 4. McGraw-Hill New York
43. Khaninezhad MM, Jafarpour B (2014) Prior model identification during subsurface flow data integration with adaptive sparse representation techniques. Computational Geosciences 18(1):3–16

44. Khaninezhad MM, Jafarpour B, Li L (2012) Sparse geologic dictionaries for subsurface flow model calibration: Part i. inversion formulation. Advances in Water Resources 39:106–121
45. Khaninezhad MM, Jafarpour B, Li L (2012) Sparse geologic dictionaries for subsurface flow model calibration: Part ii. robustness to uncertainty. Advances in water resources 39:122–136
46. Khodabakhshi M, Jafarpour B (2013) A Bayesian mixture-modeling approach for flow-conditioned multiple-point statistical facies simulation from uncertain training images. Water Resources Research 49(1):328–342
47. Kitanidis PK (1997) Introduction to geostatistics: applications in hydrogeology. Cambridge University Press
48. Klema V, Laub A (1980) The singular value decomposition: Its computation and some applications. IEEE Transactions on automatic control 25(2):164–176
49. Landis EM (1934) Capillary pressure and capillary permeability. Physiological Reviews 14(3):404–481
50. Lee J, Kitanidis P (2013) Bayesian inversion with total variation prior for discrete geologic structure identification. Water Resources Research 49(11):7658–7669
51. Li L, Jafarpour B (2010) A sparse Bayesian framework for conditioning uncertain geologic models to nonlinear flow measurements. Advances in Water Resources 33(9):1024–1042
52. Liu X, Kitanidis P (2011) Large-scale inverse modeling with an application in hydraulic tomography. Water Resources Research 47(2)
53. Lochbühler T, Vrugt JA, Sadegh M, Linde N (2015) Summary statistics from training images as prior information in probabilistic inversion. Geophysical Journal International 201(1):157–171
54. Luo J, Wang W, Qi H (2013) Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1809–1816
55. Mallat SG (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE transactions on pattern analysis and machine intelligence 11(7):674–693
56. Marvasti F, Azghani M, Imani P, Pakrouh P, Heydari SJ, Golmohammadi A, Kazerouni A, Khalili M (2012) Sparse signal processing using iterative method with adaptive thresholding (IMAT). In: Telecommunications (ICT), 2012 19th International Conference on, IEEE, pp 1–6
57. Miller K (1970) Least squares methods for ill-posed problems with a prescribed bound. SIAM Journal on Mathematical Analysis 1(1):52–74
58. Mohimani H, Babaie-Zadeh M, Jutten C (2009) A fast approach for overcomplete sparse decomposition based on smoothed l^0 norm. IEEE Transactions on Signal Processing 57(1):289–301
59. Mueller JL, Siltanen S (2012) Linear and nonlinear inverse problems with practical applications. SIAM
60. Murray CD, Dermott SF (1999) Solar system dynamics. Cambridge university press
61. Needell D, Tropp JA (2009) CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. Applied and Computational Harmonic Analysis 26(3):301–321
62. Oliver DS, Chen Y (2011) Recent progress on reservoir history matching: a review. Computational Geosciences 15(1):185–221
63. Oliver DS, Reynolds AC, Liu N (2008) Inverse theory for petroleum reservoir characterization and history matching. Cambridge University Press
64. Patankar S (1980) Numerical heat transfer and fluid flow. CRC press
65. Peterson AF, Ray SL, Mittra R, of Electrical I, Engineers E (1998) Computational methods for electromagnetics. IEEE press New York
66. Resmerita E (2005) Regularization of ill-posed problems in Banach spaces: convergence rates. Inverse Problems 21(4):1303
67. Riva M, Panzeri M, Guadagnini A, Neuman SP (2011) Role of model selection criteria in geostatistical inverse estimation of statistical data-and model-parameters. Water Resources Research 47(7)
68. Rousset M, Durlofsky L (2014) Optimization-based framework for geological scenario determination using parameterized training images. In: ECMOR XIV-14th European Conference on the Mathematics of Oil Recovery

69. Rudin LI, Osher S, Fatemi E (1992) Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena 60(1–4):259–268
70. Sarma P, Durlofsky LJ, Aziz K (2008) Kernel principal component analysis for efficient, differentiable parameterization of multipoint geostatistics. Mathematical Geosciences 40(1): 3–32
71. Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge university press
72. Shirangi MG (2014) History matching production data and uncertainty assessment with an efficient TSVD parameterization algorithm. Journal of Petroleum Science and Engineering 113:54–71
73. Shirangi MG, Durlofsky LJ (2016) A general method to select representative models for decision making and optimization under uncertainty. Computers & Geosciences 96:109–123
74. Snieder R (1998) The role of nonlinearity in inverse problems. Inverse Problems 14(3):387
75. Strebelle S (2002) Conditional simulation of complex geological structures using multiple-point statistics. Mathematical Geology 34(1):1–21
76. Suzuki S, Caers JK, et al (2006) History matching with an uncertain geological scenario. In: SPE Annual Technical Conference and Exhibition, Society of Petroleum Engineers
77. Talukder KH, Harada K (2010) Haar wavelet based approach for image compression and quality assessment of compressed image. arXiv preprint arXiv:10104084
78. Tarantola A (2005) Inverse problem theory and methods for model parameter estimation. SIAM
79. Tarantola A, Valette B (1982) Generalized nonlinear inverse problems solved using the least squares criterion. Reviews of Geophysics 20(2):219–232
80. Taubman D, Marcellin M (2012) JPEG2000 image compression fundamentals, standards and practice: image compression fundamentals, standards and practice, vol 642. Springer Science & Business Media
81. Tikhonov A, Arsenin VY (1979) Methods of solving incorrect problems
82. Tosic I, Frossard P (2011) Dictionary learning. IEEE Signal Processing Magazine 28(2):27–38
83. Tropp JA, Gilbert AC (2007) Signal recovery from random measurements via orthogonal matching pursuit. IEEE Transactions on information theory 53(12):4655–4666
84. Vo HX, Durlofsky LJ (2014) A new differentiable parameterization based on principal component analysis for the low-dimensional representation of complex geological models. Mathematical Geosciences 46(7):775–813
85. Vogel CR (2002) Computational methods for inverse problems. SIAM
86. Vrugt JA, Stauffer PH, Wöhling T, Robinson BA, Vesselinov VV (2008) Inverse modeling of subsurface flow and transport properties: A review with new developments. Vadose Zone Journal 7(2):843–864
87. Yeh WWG (1986) Review of parameter identification procedures in groundwater hydrology: The inverse problem. Water Resources Research 22(2):95–108
88. Zhou H, Gómez-Hernández JJ, Li L (2012) A pattern-search-based inverse method. Water Resources Research 48(3)
89. Zhou H, Gómez-Hernández JJ, Li L (2014) Inverse methods in hydrogeology: Evolution and recent trends. Advances in Water Resources 63:22–37
90. Zimmerman D, Marsily Gd, Gotway CA, Marietta MG, Axness CL, Beauheim RL, Bras RL, Carrera J, Dagan G, Davies PB, et al (1998) A comparison of seven geostatistically based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow. Water Resources Research 34(6):1373–1413