UNIVERSITÉ DE MONTRÉAL

# A GOAL-ORIENTED FINITE ELEMENT METHOD AND ITS EXTENSION TO PGD REDUCED-ORDER MODELING

KENAN KERGRENE
DÉPARTEMENT DE MATHÉMATIQUES ET DE GÉNIE INDUSTRIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

THÈSE PRÉSENTÉE EN VUE DE L'OBTENTION
DU DIPLÔME DE PHILOSOPHIÆ DOCTOR
(MATHÉMATIQUES DE L'INGÉNIEUR)
NOVEMBRE 2018

© Kenan Kergrene, 2018.

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Cette thèse intitulée :

A GOAL-ORIENTED FINITE ELEMENT METHOD AND ITS EXTENSION TO PGD REDUCED-ORDER MODELING

présentée par : KERGRENE Kenan
en vue de l'obtention du diplôme de : Philosophiæ Doctor
a été dûment acceptée par le jury d'examen constitué de :

M. LAFOREST Marc, Ph. D., président
M. PRUDHOMME Serge, Ph. D., membre et directeur de recherche
M. DUFOUR Steven, Ph. D., membre et codirecteur de recherche
M. TSOGTGEREL Gantumur, Ph. D., membre
M. NAIR Prasanth, Ph. D., membre externe

# DEDICATION

*À Tadig et Mammig*

## ACKNOWLEDGEMENTS

I would like to express my sincere appreciation and gratitude to my advisor Prof. Serge Prudhomme for his constant support, his sound advice, and his precious guidance all along the doctorate. I would like to thank him for his mentorship and for encouraging me to grow as a research scientist. Special thanks also go to Prof. Ludovic Chamoin, Prof. Marc Laforest, Prof. Ivo Babuška, and Prof. Steven Dufour, for fruitful collaborations and countless discussions.

I would like to thank my committee members, Prof. Gantumur Tsogtgerel and Prof. Prasanth Nair, for the time they will dedicate to reading this manuscript and for their presence at the defense.

These last years have also been marked by numerous friendships. I would like to take this opportunity to acknowledge them as well: Antonin, Damien, Pierre, Jonathan, Quentin, Ron, Swann, Adrien, Élise, Vincent, Manon, Yoann, Simon, and the rest of the $\chi\beta\rho$.

I also want to thank my partner Alice. Your continuous understanding and unfailing patience over the past several years enabled me to complete this research work.

The last thank goes to my parents, my brother, and my family for their encouragement all along my studies.

# RÉSUMÉ

Nous proposons une méthode éléments finis formulée pour des quantités d'intérêt. L'objectif est d'accroître la précision des solutions numériques pour ces quantités, choisies par l'utilisateur, sans pour autant perdre en précision globale.

Les approches traditionnelles visant à contrôler l'erreur en quantité d'intérêt utilisent habituellement la solution d'un problème adjoint pour: (i) estimer l'erreur en quantité d'intérêt; et (ii) savoir comment adapter la discrétisation afin d'obtenir un espace éléments finis capable de mieux représenter les quantités d'intérêt de la solution. Ces approches s'inscrivent donc dans un procédé itératif de prédictions-corrections. Nous proposons d'utiliser cette même solution adjointe conjointement avec un probleme primal modifié, tel que sa solution soit ajustée à une valeur plus précise de la quantité d'intérêt. Ainsi, nous résolvons dans un espace qui est déjà adapté à la quantité d'intérêt.

L'originalité de la présente approche consiste à utiliser la solution du problème adjoint non pas en tant que substitut de la solution exacte/référence pour l'estimation d'erreur et l'adaptation, mais en extrayant de celle-ci des valeurs des quantités d'intérêt extrêmement précises. Ces valeurs sont ensuite utilisées dans une minimisation sous contrainte de l'énergie (problème primal contraint) afin d'obtenir une solution plus précise en quantité d'intérêt.

Ensuite, nous étendons cette approche en quantité d'intérêt à un contexte de réduction de modèles en utilisant la PGD. Ces méthodes reposent généralement sur des représentations spectrales, et sont de plus en plus utilisées pour simuler des problèmes en haute dimension. En ne considérant que les principaux modes propres de la solution, ces méthodes déjouent la malédiction de la dimensionnalité et rendent possibles des simulations auparavant inenvisageables.

# ABSTRACT

We present a finite element formulation of boundary-value problems that aims at constructing approximations specifically tailored for the estimation of quantities of interest of the solution, hence the name goal-oriented finite element method.

The main idea is to formulate the problem as a constrained minimization problem that includes refined information in the goal functionals, so that the resulting model is capable of delivering enhanced predictions of the quantities of interest. This paradigm constitutes a departure from classical goal-oriented approaches in which one computes first the finite element solution and subsequently adapts the mesh via a greedy approach, by controlling error estimates measured in terms of quantities of interest using a posteriori dual-based error estimates.

The formulation is then extended to the so-called Proper Generalized Decomposition method, an instance of model order reduction methods, with the aim of constructing reduced-order models tailored for the approximation of quantities of interest. Model order reduction methods aim at circumventing the curse of dimensionality arising from the high number of parameters of a given problem, by uncovering and/or exploiting lower dimensional structures present in the model or in the solution.

Numerical examples are disseminated throughout the dissertation. They appear at the end of each of the three main chapters and Chapter 5 consists of an application example, namely a parametrized electrostatic cracked composite material.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1 INTRODUCTION

*"The ancients considered mechanics in a twofold respect; as rational, which proceeds accurately by demonstration, and practical. To practical mechanics all the manual arts belong, from which mechanics took its name. But as artificers do not work with perfect accuracy, it comes to pass that mechanics is so distinguished from geometry, that what is perfectly accurate is called geometrical; what is less so is called mechanical. But the errors are not in the art, but in the artificers. He that works with less accuracy is an imperfect mechanic; and if any could work with perfect accuracy, he would be the most perfect mechanic of all; for the description of right lines and circles, upon which geometry is founded, belongs to mechanics. Geometry does not teach us to draw these lines, but requires them to be drawn; for it requires that the learner should first be taught to describe these accurately, before he enters upon geometry; then it shows how by these operations problems may be solved."*

*Sir Isaac Newton, Philosophiæ Naturalis Principia Mathematica, 1687*

## 1.1 Overview

Along with theory and experimentation, the scientific method has recently been complemented with a third pillar, that of computer simulation. The advent of the computer has helped the scientific community to build and test more sophisticated models than ever before, validate hypotheses, and contribute to new knowledge in virtually all realms of science and beyond. Advances in Computational Science and Engineering, as well as improvement of computer performances have allowed one to simulate complex multiphysics and multiscale problems. In turn, this has led computational scientists to contemplate simulations that are increasingly more time- and resource-consuming. This is the case when, for instance, one is interested in quantifying uncertainties or solving inverse and optimization problems.

The objectives of uncertainty quantification analyses usually encompass the characterization of uncertainties in input or material parameters of a given model, and the propagation of these uncertainties to model outputs. Parameter identification, model calibration, or inverse problems, which refer to the activities that consist in estimating parameters from a priori knowledge and measurement data, usually require, alike uncertainty propagation or optimization problems, that output quantities of interest be estimated for a very large number of parameter samples. These activities are sometimes collectively referred to as multi-query studies. Applications span a myriad of domains where simulation-based decision-making is

needed.

From these observations, it has now become clear that numerical simulations should focus on efficiently and accurately predicting output quantities of interest or response surfaces, that is, specific features of the solution rather than the whole solution itself. Within this context, one may think of two avenues in order to achieve this goal. First, goal-oriented error estimation and adaptive methods have been developed in order to estimate and control errors with respect to quantities of interest. These dual-based methods involve an adjoint problem whose solution provides weighted residual indicators to identify the sources of errors that influence these quantities the most. These approaches have been successful in accelerating the convergence of the approximation towards the exact quantities of interest when compared to classical approaches based on global-norm error estimates.

Secondly, one could reduce the model complexity in order to enable fast simulation of computational problems. Such a need has led to the development of model reduction techniques, whose main feature is to represent the solution in terms of modal expansions on some particular basis. Model reduction techniques usually differ by the choice of said basis. The key ingredient is that the modes are selected by order of importance, so that only a few of them are often needed to ensure a correct representation of the underlying physical phenomenon. As a result, an accurate reduced solution can be obtained at a fraction of the cost associated with that of the full solution.

The main objective of the present research work is to develop a goal-oriented model reduction technique, i.e. a model reduction method that is specifically tailored towards the approximation of quantities of interest. The originality of the proposed methodology is to simultaneously combine the concepts and methods from both the fields of model reduction and goal-oriented approaches. This is different from what is currently done when one first constructs a reduced model and subsequently corrects the model by estimating and controlling, in a post-processed fashion, the sources of errors in the reduced solution. Instead, we proposed a new paradigm in which accurate information from quantities of interest is directly incorporated into the calculation of the reduced solution. The cornerstone of the approach is to augment the original problem with constraints defined with respect to quantities of interest of the solution. The methodology will be applied to one given class of model reduction methods, namely, the Proper Generalized Decomposition (PGD) method that consists in constructing an approximate solution of an initial or boundary-value problem using the concept of separation of variables. Nonetheless, the proposed strategy is far more general and can be applied to other model reduction methods.

The main contributions of the thesis are

- the formulation of a new goal-oriented approach for the Finite Element Method: derivation, theoretical analysis, code implementation, and testing of the method through multiple numerical examples;
- the extension of numerical strategies for the imposition of constraints in the Proper Generalized Decomposition framework: review of classical methods to solve constrained problems (e.g. penalization, Lagrangian, and Uzawa methods), derivation, analysis, code implementation, and testing on several Poisson problems;
- the development of a goal-oriented methodology for reduced-order modeling: formulation, implementation to various problems with different separations of variables ($x$ and $y$ space variables, space and parameter variables);
- the application of the proposed method to the simulation of a cracked composite material: electrostatic study with extra parameters including the values of the electrical conductivity in the different plies as well as an extra parameter controlling the position of the electrode that injects the direct current into the material.

## 1.2 Literature review and state-of-the-art

### 1.2.1 Model reduction methods

Due to the so-called curse of dimensionality [21], FEM and other classical methods to approximate solutions of high-dimensional initial and boundary-value problems can rapidly become intractable when the prescribed tolerance on the solution is too small. An alternative approach to classical methods has been to develop model reduction techniques. Generally, these methods are based on the representation of the solution in terms of modes expanded on some specific basis. Such approaches significantly reduce the need for extensive computer resources. They are justified by the fact that it is often unnecessary to calculate every detail of the solution in order to obtain a good understanding of the underlying physical phenomena. The objective of using such methods is to capture the principal modes of the system. Discrete solutions of reduced models can often provide reasonable approximations in much lower dimensional spaces than those in which the fully discretized solution lives. When solving partial differential equations, using a fully discretized model, one quickly reaches the limits of most current computers, while only achieving a rather crude description of complex phenomena. On the other hand, model reduction methods lead to problems with $m \times d \times N$ unknowns, $m$ being the number of retained modes, $d$ the number of parameters or variables, and $N$ the number of degrees of freedom for each parameter. The interest is twofold: first, the modes are captured by order of importance; second, the size of the system to be solved

grows linearly with the number of degrees of freedom, which is a considerable improvement compared to classical approaches using brute force and leading to systems of exponential size $N^d$ [35]. Note that model reduction methods are also useful when the problem only contains a small number of parameters or variables as they circumvent the need to solve extremely large systems. In summary, model reduction enables one to deliver predictions using a simplified model compared to the complete one, thus allowing simulation at extremely low cost, or even in real-time, which is useful for optimization, DDDAS (Direct Dynamic Data Driven Application System), and applications involving multi-query approaches.

In this section, we briefly recall the essential features, and provide an overview, of model reduction methods. We will then focus on one of such methods, namely the PGD (Proper Generalized Decomposition) method since it will be used in the next chapters of this manuscript.

**Spectral representations**

Most approaches in model reduction methods use the technique known as POD (Proper Orthogonal Decomposition) [10, 53], also known as Karhunen-Loève decomposition [61, 71], SVD (Singular Value Decomposition), PCA (Principal Component Analysis) or PCD (Principal Component Decomposition) in other contexts [11, 86]. The method relies on the projection of the problem onto a reduced basis, predetermined during the offline stage. This first stage, also known as the learning stage (by the method of snapshots for instance; other methods can also be used, such as the CVT method (Central Voronoi Tessellation) [52]), consists in solving the full problem for some values of the parameters (in space, it comes down to solving the problem on a crude mesh, or on a sub-domain of the structure [7]; in time, one solves on a shorter time interval than that of the full simulation). These solutions then constitute what are usually called the snapshots. In the second stage of the method, one performs a Singular Value Decomposition on these snapshots in order to obtain a truncated spectral representation and to get rid of redundant information. Next is the online stage, which consists in solving the system projected onto the reduced basis composed of the first $m$ modes retained in the SVD previously performed. As a result, one finally obtains a reduced solution in the form of a truncated spectral decomposition of the fully discretized solution, which retains the modes by order of importance. The Reduced Basis (RB) methods [27, 84] constitute a similar approach to POD methods and mainly differs with respect to the choice of snapshots. In essence, the POD method aims at injecting knowledge about the solution into the model on behalf of the user in order to reduce the complexity of the underlying problem.

By construction, POD has a significant drawback, that of being extremely dependent on the

learning phase. Indeed, all the information about the system of interest is encapsulated within the reduced basis. Therefore, if critical information is missing from the learning stage, i.e. the learning phase has not been properly and carefully performed, one may obtain a solution of rather poor quality. It should be noted, however, that there exist methods allowing one to update the reduced basis when its poor efficiency is detected [29, 52], but these methods often rely on greedy algorithms. Furthermore, this happens during the online phase, which is not ideal when one is faced with time constraints.

POD methods have nevertheless demonstrated their efficiency on a wide range of complex systems, e.g. in the case of problems in solide mechanics involving large deformations, non-linearities (elasto-plastic, elasto-visco-plastic, geometric non-linearities, etc.), incompressibility, homogenization, etc. [72, 95], and in fluid mechanics [33]. Their efficiency has also been reported for uncertainty quantification [29] (in this case one only needs to incorporate an additional term corresponding to the probabilistic parameter in the spectral decomposition), and for parametrized PDEs [29, 83]. Of course, the more complicated or sophisticated the problem is, the more adaptive steps are required for the POD to work reasonably. For instance, in the incompressible or non-linear case or when uncertainty is added to the model, several authors have reported that the sensitivity to the learning phase was increased [29, 95]. Many developments have followed POD, such as the weighted POD [52], in which a higher importance is given to some snapshots than others, or the hybrid CVOD [52], cheaper than POD (it requires to solve several small eigenvalue problems instead of a large one).

In [70], the authors derive an equation-free POD, where they find the functions appearing in the spectral decomposition by explicit means, instead of solving ODEs/PDEs, as it is generally done. They assert that their POD method is thereby more stable, since a truncation of the higher modes, corresponding to shorter wavelengths, irreparably causes a drift in the solution because of the unresolved shorter scales, at which the energy is dissipated. A remedy to this problem is presented in [70]: it consists in adding a spectral viscosity to the model, which handles energy dissipation and thus improves stability.

Model reduction methods are also increasingly employed in the field of electronic engineering (automatic, optimal control, etc.) [59, 86]. The objective in these communities consists in reducing the dimension of the system while preserving input/output relations, as well as other properties like stability and passivity. Other types of model reduction thus come into play, which involve mathematical methods such as Arnoldi, Lànczos, Padé, Krylov (by the way, [86] presents a relation between Padé and Krylov), as well as diverse variants of those such as AWE (Asymptotic Waveform Equation), PVL (Padé via Lànczos) or PRIMA (Passive Reduced-order Interconnect Macromodeling Algorithm). A last category of methods aim

at preserving moments of quantities of interest [59, 86], but they seem to suffer from bad conditioning issues.

#### The Proper Generalized Decomposition

The PGD method has been proposed as an alternative to POD approaches. Its main advantage is that it does not require a learning phase [75]. For this reason, it is said to be an a priori reduction method, while POD is said to be a posteriori. The PGD method can also be understood as a generalized SVD [48] or low-rank approximation of the solution. Instead of computing the modes beforehand and projecting the problem on the subspace of snapshots at the beginning of the simulation, the modes are computed on the go as successive rank-1 updates of the current approximation. PGD is most often viewed as an iterative process, where at each iteration the solution is improved by adding a new mode, or correction, optimal with respect to a given metric.

PGD methods rely on a construction on the fly of separated representations without using a priori knowledge of the solution to a given problem. In addition to space and time, one could seek a separated representation with respect to other parameters (e.g. material coefficients, geometry, boundary conditions, initial conditions, loading) that may even involve uncertainties. The separation of variables allows for a significant reduction in computational costs, allowing simulation at extremely low cost or even in real-time, as well as optimization, and other many query applications [35].

Model reduction methods are currently in rapid expansion, and their performance in terms of reduction of computational time and memory storage can be impressive. However, these are still relatively new methods in academics and industries, and they still require further advances. Our attention will primarily focus on PGD approaches since they do not require a learning phase, and all the calculations can be done online.

### 1.2.2 Goal-oriented error estimation

Goal-oriented error estimation is the activity in computational sciences and engineering that focuses on the development of computable estimators of the error measured with respect to user-defined quantities of interest. The use of discretization methods (such as the finite element method) for solving initial- and boundary-value problems necessarily produces approximations that are in error when compared to the exact solutions. Methods to estimate discretization errors were proposed as early as the seventies [14] and initially focused on developing error estimators in terms of global (energy) norms. Typical methods are the

so-called recovery-type methods, explicit residual error estimators, subdomain residual methods, element residual methods, etc., see [4, 15, 92] and references therein. An issue in those approaches is that they provide error estimates in abstract norms, which fail to inform the users about specific quantities of engineering interest or local features of the solutions. It is only in the late nineties (except maybe the early work of Gartland [50]) that researchers started to develop error estimators with respect to user-defined quantities of interest. These error estimators and corresponding adaptive methods are based on the solution of adjoint problems associated with the quantity of interest. These methods are usually referred to as the dual-weighted residual method [18, 19, 20], a posteriori bounds for linear-functional outputs [80], or goal-oriented error estimators [78, 82]. In this case, the user is able to specify quantities of interest, written as functionals defined on the space of admissible solutions, and to assess the accuracy of the approximations in terms of these quantities. It is also noteworthy that the corresponding adaptive approaches are based on greedy strategies, which involve the following sequence of steps: 1) compute an approximate solution to the problem, 2) estimate the error in the approximation, 3) derive local refinement indicators based on the error estimate and adapt the discretization space, 4) iterate the process until convergence within some prescribed tolerance is reached. The objective of the research dissertation will be to construct approximations on a given discretization space that are more accurate in the quantity of interest than those obtained using the classical adaptive approach.

### 1.3 Objectives of the thesis

The overall objective of the manuscript is to formulate a goal-oriented model reduction method, i.e. a model reduction technique that is specifically tailored towards the approximation of quantities of interest. In the present dissertation, this objective is divided into four sub-objectives.

1. The first objective is to derive a goal-oriented finite element formulation whose solution shows improved convergence in the quantities of interest when compared to a traditional finite element solution. To achieve this specific goal, we reformulated the original minimization problem as a constrained minimization problem. The constraints essentially carry enhanced information regarding the quantities of interest, that can be obtained by solving the adjoint problems beforehand.

2. The second objective deals with the imposition of constraints in the model order reduction framework of the PGD. There exists a variety of numerical methods to enforce constraints on a boundary-value problem, e.g. Penalization, Lagrangian, Uzawa or Aug-

mented Lagrangian. To tackle the second objective, we extended these methods to the case of PGD reduced solutions, and analyzed the resulting numerical strategies. The methods were tested and compared on a two-dimensional Poisson problem with either pure Neumann, or Robin, boundary conditions.

3. The third objective answers the overarching research problematic of formulating a goal-oriented PGD method. To this particular end, we developed a "two-step progressive Galerkin approach". The first step consists in performing a rank-1 update of the adjoint solution. The second step incorporates this information in the computation of a constrained rank-1 update of the primal solution.

4. The last objective consists in applying the new methodology to a problem of engineering interest, that of the electrostatic study of a cracked composite material.

### 1.4 Outline of the dissertation

The manuscript is organized as follows. Chapters 2 through 4 are largely inspired by [66], [65], and [64], respectively. Chapter 2 is devoted to the new constrained formulation that aims at targeting specific features of the solution. It is proven that the constrained problem is well-defined and that its solution is globally near-optimal while being much more accurate in the quantities of interest. We also analyze the errors yielded by this approach both globally and in the quantities of interest. The efficiency of this methodology is demonstrated on a series of numerical examples. Next we turn to the reduced-order modeling framework. Chapter 3 describes classical computational techniques to enforce constraints and extends those to the case of PGD reduced-order problems. The main difficulty lies in the fact that the constraints should be applied globally while the reduced solution is defined using the concept of separation of variables. Numerical examples are provided to assess the performance of each of the methods we investigated. In Chapter 4 we combine both ideas and present the goal-oriented model reduction technique. Again, numerical examples illustrate that the proposed methodology is able to deliver enhanced predictions of the quantities of interest compared to a standard approach, without sacrificing too much the global accuracy of the solution. Chapter 5 consists in an application chapter where we apply the goal-oriented model reduction technique to the case of a parametrized electrostatic problem in a laminated composite. Finally, in Chapter 6 we present a synthesis of the proposed methodology, as well as guidelines for future work.

# CHAPTER 2 A NEW GOAL-ORIENTED FORMULATION OF THE FINITE ELEMENT METHOD

*In this chapter, we introduce, analyze, and numerically illustrate a method designed to take into account quantities of interest during the finite element treatment of a boundary-value problem. The objective is to derive a method whose computational cost is of the same order as that of the classical approach for goal-oriented adaptivity, which involves the solution of the primal problem and of an adjoint problem used to weigh the residual and provide indicators for mesh refinement. In the current approach, we first solve the adjoint problem, then use the adjoint information as a minimization constraint for the primal problem. As a result, the constrained finite element solution is enhanced with respect to the quantities of interest, while maintaining near-optimality in energy norm. We describe the formulation in the case of a problem defined by a symmetric continuous coercive bilinear form and demonstrate the efficiency of the new approach on several numerical examples. This chapter is largely inspired by [66].*

## 2.1 Introduction

Advances in Computational Science and Engineering have reached such a level of maturity that increasingly complex multiphysics and multiscale problems can now be simulated for decision-making and optimal design. The focus of such simulations has thus shifted towards efficiently and accurately predicting specific features of the solution rather than the whole solution itself. With that objective in mind, goal-oriented error estimation and adaptive methods [78, 82], whose predominant instance is the dual-weighted residual method [20], have been developed since the late nineties in order to estimate and control errors with respect to quantities of interest. The principle of these methods essentially relies on the solution of adjoint problems associated with quantities of interest in order to identify and refine the sources of discretization or modeling errors that influence these quantities the most [79]. So far, these approaches have been very successful in accelerating the convergence of the approximations towards the exact quantities of interest and thus at a lesser computational cost than classical a posteriori error estimation methods. However, dual-weighted residual methods are reminiscent of two-step predictor-corrector methods, in the sense that one first computes an approximate solution of a boundary-value problem and then corrects the discrete solution space in order to better approximate the quantities of interest.

The objective of the present chapter is to propose an alternative paradigm: we aim at

developing a novel finite element formulation of boundary-value problems whose approximate solutions are tailored towards the calculation of quantities of interest. The main idea is based on the reformulation of the problem as a minimization problem subjected to the additional constraint that the error in the quantities of interest be within some prescribed tolerance. Chaudhry et al. [36] have proposed a similar approach in which constraints are enforced via a penalization method. One main issue with that approach is concerned with the selection of suitable penalization parameters. We propose here to circumvent this issue by imposing the equality or inequality constraints through the use of Lagrange multipliers. The framework will be presented in the case of several quantities of interest in order to describe the method in a general setting. However, the treatment of several quantities of interest is not the primary goal of the manuscript and the reader interested in multi-objective error estimation is referred to the following literature [46, 47, 56, 89].

The present chapter is organized as follows: In Section 2.2, we present the model problem considered in this study along with some classical notations. In Section 2.3, we introduce the novel formulation of taking into account quantities of interest using a constrained minimization. We demonstrate the well-posedness of the formulation and the near-optimality of the corresponding solution. In Section 2.4, we investigate the case of inequality constraints using the Karush-Kuhn-Tucker conditions. Section 2.5 addresses the topic of error estimation and adaptivity. Numerical examples are presented in Section 2.6 and illustrate the performance of the method. In particular, we compare our approach to the classical goal-oriented adaptivity. Finally, we provide some concluding remarks in Section 2.7.

## 2.2 Preliminaries and model problem

Consider an abstract problem written in weak form as

$$\text{Find } u \in V \text{ such that } a(u,v) = f(v), \quad \forall v \in V, \tag{2.1}$$

where $(V, \|\cdot\|)$ is a Hilbert space, and bilinear form $a$ and linear form $f$ satisfy the usual regularity assumptions: $a$ is continuous and coercive and $f$ is continuous over $V$. This problem will be referred to as the primal problem and its well-posedness is ensured by the Lax-Milgram theorem. For the sake of simplicity, we require in addition that $a$ be symmetric so that the primal problem (2.1) is equivalent to minimizing the following energy functional

$$J(u) = \frac{1}{2}a(u,u) - f(u), \tag{2.2}$$

i.e.

$$\text{Find } u \in V \text{ such that } u = \underset{v \in V}{\operatorname{argmin}}\, J(v). \tag{2.3}$$

If $a$ were not symmetric, the method presented in this work could be applied by considering a Least Squares approach [23], which in effect symmetrizes the problem.

We now turn to the finite element formulation of the primal problem (2.1). Here, and in the remainder of the dissertation, we consider a general conforming finite element space $V_h = \operatorname{span}\{\varphi_i\} \subset V$, where $\varphi_i$, $i = 1, \ldots, N$ are basis functions of $V_h$. We also assume that the corresponding mesh satisfies the usual regularity properties [42, 60]. We denote by $h$ the characteristic mesh size. The classical finite element problem associated to the primal problem (2.1) is given by

$$\text{Find } u_h \in V_h \text{ such that } a(u_h, v_h) = f(v_h), \quad \forall v_h \in V_h. \tag{2.4}$$

The objective of this work is to improve the accuracy in the approximation of scalar quantities of the solution $u$ of the primal problem (2.1). Consider therefore the continuous quantities of interest $Q_i(u)$, $i = 1, \ldots, k$, with $k \in \mathbb{N}$ and assume these are linear, i.e. $Q_i \in V'$, the dual space of $V$. We will denote by $Q$ the linear map from $V$ to $\mathbb{R}^k$ whose $i$-th component is $Q_i$. Further, we assume the linear forms to be linearly independent, i.e. the map $Q$ is surjective. In other words, each functional $Q_i$ provides independent information about $u$.

Associated to these linear forms, we have the $k$ dual or adjoint problems

$$\text{For } i = 1, \ldots, k, \text{ find } p_i \in V \text{ such that } a(v, p_i) = Q_i(v), \quad \forall v \in V, \tag{2.5}$$

and the fundamental relations

$$Q_i(u) = a(u, p_i) = f(p_i), \quad \forall i = 1, \ldots, k. \tag{2.6}$$

The finite element formulation of the adjoint problems (2.5) in space $V_h$ are given by

$$\text{For } i = 1, \ldots, k, \text{ find } p_{i,h} \in V_h \text{ such that } a(v_h, p_{i,h}) = Q_i(v_h), \quad \forall v_h \in V_h. \tag{2.7}$$

The main idea is to derive a novel formulation of the problem based on the minimization of the energy functional $J$ subjected to constraints in terms of the quantities of interest $Q$.

## 2.3 Goal-oriented formulation with equality constraints

### 2.3.1 Formulation and well-posedness

Suppose for a moment that we are interested in finding a solution $w \in V$ that satisfies the constraints $Q_i(w) = \alpha_i$, where $\alpha = (\alpha_1, \ldots, \alpha_k)^T \in \mathbb{R}^k$ is given. Instead of the minimization problem (2.3), we consider the constrained minimization problem

$$\text{Find } w \in V \text{ such that } w = \underset{\substack{v \in V \\ Q(v)=\alpha}}{\operatorname{argmin}} J(v). \tag{2.8}$$

The standard way to impose constraints is by the introduction of the Lagrangian functional $\mathcal{L} : V \times \mathbb{R}^k \to \mathbb{R}$ defined as

$$\mathcal{L}(w, \lambda) = J(w) + \sum_{i=1}^{k} \lambda_i (Q_i(w) - \alpha_i), \tag{2.9}$$

where $\lambda = (\lambda_1, \ldots, \lambda_k)^T \in \mathbb{R}^k$ is the vector collecting the so-called Lagrange multipliers. This functional can be written in compact form as

$$\mathcal{L}(w, \lambda) = J(w) + \lambda \cdot (Q(w) - \alpha), \tag{2.10}$$

where $\cdot$ denotes the classical Euclidean inner-product on $\mathbb{R}^k$.

The saddle-point formulation of the Lagrangian functional $\mathcal{L}$ over $V \times \mathbb{R}^k$ yields the mixed problem

$$\text{Find } (w, \lambda) \in V \times \mathbb{R}^k \text{ such that } \begin{cases} a(w, v) + \lambda \cdot Q(v) = f(v), & \forall v \in V, \\ \tau \cdot Q(w) = \tau \cdot \alpha, & \forall \tau \in \mathbb{R}^k. \end{cases} \tag{2.11}$$

Introducing the bilinear form $b(\tau, v) = \tau \cdot Q(v)$ defined on $\mathbb{R}^k \times V$, the above problem can be recast in the classical form

$$\text{Find } (w, \lambda) \in V \times \mathbb{R}^k \text{ such that } \begin{cases} a(w, v) + b(\lambda, v) = f(v), & \forall v \in V, \\ b(\tau, w) = \tau \cdot \alpha, & \forall \tau \in \mathbb{R}^k. \end{cases} \tag{2.12}$$

**Lemma 1** (LBB condition)**.** *Let $\|\cdot\|_1$ be the 1-norm on $\mathbb{R}^k$, i.e. $\|\tau\|_1 = \sum_i |\tau_i|$. The bilinear*

*form $b$ satisfies the LBB condition*

$$\exists\, \beta > 0 \text{ such that } \forall \tau \in \mathbb{R}^k, \quad \sup_{v \in V} \frac{|b(\tau, v)|}{\|v\|} \geq \beta \|\tau\|_1. \tag{2.13}$$

*Proof.* We first consider the trivial case $k = 1$ and then the general case.

*Case $k = 1$.* Let $z \in V \backslash \text{Ker } Q$ (there exists such a $z$ since the linear form $Q$ is assumed to be surjective, i.e. non-zero in this case) and define $\beta = \frac{|Q(z)|}{\|z\|}$. Then for any $\tau \in \mathbb{R}$, $\frac{|b(\tau,z)|}{\|z\|} = \beta|\tau|$, so that $\sup_{v \in V} \frac{|b(\tau,v)|}{\|v\|} \geq \beta|\tau|$.

*General case.* Similarly, using the surjectivity of $Q$, one can find functions in $V$ such that all cases in terms of the signs of the components of $\tau \in \mathbb{R}^k$ will be accounted for. More specifically, let the "vector-valued sign function" defined over $\mathbb{R}^k$ as

$$\begin{aligned} \text{signs}: \ \mathbb{R}^k &\to \{-1, 0, 1\}^k \\ \tau &\mapsto (\text{sign}(\tau_1), \ldots, \text{sign}(\tau_k)) \end{aligned} \tag{2.14}$$

Function signs is surjective onto the set $\{-1, 1\}^k$. Indeed, for any $\sigma \in \{-1, 1\}^k$, it holds $\text{signs}(\sigma) = \sigma$. Since $Q$ is also surjective onto $\mathbb{R}^k$, for any $\sigma \in \{-1, 1\}^k$ there exists $z_\sigma \in V$ such that $\text{signs}(Q(z_\sigma)) = \sigma$. The set $\{-1, 1\}^k$ is finite, as a result this process constructs a finite set $\mathcal{Z} \subset V$. Then we define

$$\beta = \min_{\substack{z_\sigma \in \mathcal{Z} \\ i=1,\ldots,k}} \frac{|Q_i(z_\sigma)|}{\|z_\sigma\|} > 0. \tag{2.15}$$

Now, for any $\tau \in \mathbb{R}^k$, let $\sigma = \text{signs}(\tau)$. To determine $z \in \mathcal{Z}$ associated with $\sigma$ when $\sigma$ contains components of value zero, we construct $\tilde{\sigma}$ where all zero components have been replaced by one and set $z_\sigma = z_{\tilde{\sigma}}$. As a result there always exists a well-defined $z_\sigma \in \mathcal{Z}$ and it holds

$$\frac{|b(\tau, z_\sigma)|}{\|z_\sigma\|} = \frac{\left|\sum_{i=1}^k \tau_i Q_i(z_\sigma)\right|}{\|z_\sigma\|} = \frac{\sum_{i=1}^k |\tau_i||Q_i(z_\sigma)|}{\|z_\sigma\|} \geq \beta \|\tau\|_1. \tag{2.16}$$

Consequently, $\sup_{v \in V} \frac{|b(\tau,v)|}{\|v\|} \geq \beta \|\tau\|_1$. □

**Theorem 1** (Well-posedness of the constrained formulation)**.** *The constrained problem* (2.12) *has a unique solution.*

*Proof.* The proof directly follows from the LBB condition established in Lemma 1 and the Babuška-Lax-Milgram theorem [12, 28, 77]. □

In the specific case where $\alpha = Q(u)$, $u$ being the solution of the original problem (2.1), the solution of the constrained problem (2.12) is given by $w = u$ and $\lambda = 0$: the constraints are "inactive".

We now establish a key relation between the solutions to the constrained and unconstrained problems.

**Theorem 2** (Relation between constrained and unconstrained solutions)**.** *Let $(w, \lambda) \in V \times \mathbb{R}^k$ denote the solution of the constrained problem* (2.12)*, $p_i \in V$ denote the solutions of the dual problems* (2.5)*, and $u \in V$ denote the solution of the unconstrained problem* (2.1)*. Then*

$$u = w + \sum_{i=1}^{k} \lambda_i p_i. \tag{2.17}$$

*Proof.* Using the adjoint problems (2.5) and the bilinearity of $a$, it holds

$$\lambda \cdot Q(v) = \sum_{i=1}^{k} \lambda_i Q_i(v) = \sum_{i=1}^{k} \lambda_i a(v, p_i) = a\left(v, \sum_{i=1}^{k} \lambda_i p_i\right). \tag{2.18}$$

Substituting the new expression (2.18) for $\lambda \cdot Q(v)$ in the first equation of the constrained problem (2.12) yields

$$a(w, v) + a\left(v, \sum_{i=1}^{k} \lambda_i p_i\right) = f(v), \quad \forall v \in V. \tag{2.19}$$

Now, making use of the fact that $a$ is bilinear and symmetric yields

$$a\left(w + \sum_{i=1}^{k} \lambda_i p_i, v\right) = f(v), \quad \forall v \in V. \tag{2.20}$$

Finally, the Lax-Milgram theorem applied to the unconstrained problem (2.1) ensures unicity of the solution so that

$$u = w + \sum_{i=1}^{k} \lambda_i p_i, \tag{2.21}$$

which completes the proof. □

We further note that Theorem 2 holds for any choice of $\alpha \in \mathbb{R}^k$.

The mixed finite element problem on $V_h \times \mathbb{R}^k$ corresponding to the Lagrangian approach (2.11)

is given by

$$\text{Find } (w_h, \lambda_h) \in V_h \times \mathbb{R}^k \text{ such that } \begin{cases} a(w_h, v_h) + \lambda_h \cdot Q(v_h) = f(v_h), & \forall v_h \in V_h, \\ \tau_h \cdot Q(w_h) = \tau_h \cdot \alpha, & \forall \tau_h \in \mathbb{R}^k. \end{cases} \tag{2.22}$$

Note that the Lagrange multiplier is here denoted by $\lambda_h$, not because of the discretization of $\mathbb{R}^k$, but rather because it depends on $w_h \in V_h$ where $V_h$ is a finite-dimensional subspace of $V$. Furthermore, we also use this notation in order to avoid confusion with the Lagrange multiplier $\lambda$ appearing in the constrained problem (2.11).

**Remark 1.** *If $Q : V_h \to \mathbb{R}^k$ is still surjective, existence and unicity are inherited from the infinite-dimensional case; in particular, surjectivity implies here that: 1) $\dim \mathbb{R}^k \leq \dim V_h$, i.e. $k \leq N$: there are fewer constraints than degrees of freedom; 2) rank $Q = k$, i.e. the rows of the $k \times N$ constraint matrix are linearly independent: in other words it has full row-rank.*

Similarly to Theorem 2, we can establish the following relation between the solutions to the constrained and unconstrained finite element problems. This result will be used when studying convergence in Section 2.3.2 and adaptivity in Section 2.5.

**Theorem 3** (Relation between constrained and unconstrained solutions – Finite-dimensional case)**.** *Let $(w_h, \lambda_h) \in V_h \times \mathbb{R}^k$ denote the solution of the constrained problem* (2.22)*, $p_{i,h} \in V_h$ denote the solutions of the dual problems* (2.7) *and $u_h \in V_h$ denote the solution of the unconstrained problem* (2.4)*. Then*

$$u_h = w_h + \sum_{i=1}^{k} \lambda_{h,i} p_{i,h}. \tag{2.23}$$

*Proof.* The proof is similar to that of Theorem 2. □

**Remark 2.** *Note that the Galerkin orthogonality arising from the constrained problem* (2.22) *is slightly modified compared to the classical unconstrained approach. Indeed, subtracting the first equation of the constrained finite element problem* (2.22) *from the initial weak formulation* (2.1) *yields*

$$a(u - w_h, v_h) - b(\lambda_h, v_h) = f(v_h) - f(v_h) = 0, \quad \forall v_h \in V_h, \tag{2.24}$$

*that is $a(u - w_h, v_h) = b(\lambda_h, v_h) = \lambda_h \cdot Q(v_h),\ \forall v_h \in V_h$. In particular, $u - w_h$ is not orthogonal to the entire space $V_h$ but at least to $V_h \cap Ker\, Q$. This modified Galerkin orthogonality relation will be used when studying error estimation and adaptivity in Section 2.5.*

**Remark 3.** *In contrast with the Lagrangian approach, the penalization approach [36] seeks the minimizer of the modified energy functional*

$$J_\beta(u) = J(u) + \sum_{i=1}^{k} \frac{\beta_i}{2} \left(Q_i(u) - \alpha_i\right)^2, \tag{2.25}$$

*with a penalization parameter $\beta \in \mathbb{R}^k$ chosen to ensure convergence, efficiency, and accuracy. In that case, the relation between the penalized solution $u_\beta$ and the unconstrained solution $u$ is*

$$u = u_\beta + \sum_{i=1}^{k} \beta_i \left(Q_i(u_\beta) - \alpha_i\right) p_i, \tag{2.26}$$

*and similarly for their finite-dimensional counterparts.*

### 2.3.2 Selection of $\alpha$ and near-optimality

In this work, the goal is to obtain an approximation $w_h$ such that $Q(w_h) \approx Q(u)$, meaning that the target values $\alpha_i$ should be as close as possible to the quantities of interest $Q_i(u)$. In view of the fundamental relation (2.6), we propose to choose $\alpha$ by considering the $k$ adjoint problems (2.5). However, for most problems of practical interest, the adjoint problems cannot be solved exactly and have to be discretized, say using the finite element method. These approximate adjoint solutions $\tilde{p}_i$ are then used to derive the target values $\alpha_i$, i.e. we set $\alpha_i = f(\tilde{p}_i)$, $i = 1, \ldots, k$ and then proceed to solve the constrained finite element problem (2.22).

**Remark 4.** *We emphasize here that one needs to use a space larger than $V_h$ to compute the adjoint finite element solutions $\tilde{p}_i$. Indeed let us assume that we were to solve each discrete adjoint problem in the same space $V_h$ as the one used to solve the classical finite element problem (2.4), i.e. the adjoint problems (2.7), and then set $\alpha_i = f(p_{i,h})$. Choosing these target values $\alpha$ as constraints for the constrained primal problem (2.22) leads to $Q_i(w_h) = \alpha_i$. Repeating the computation (2.6), now in the finite element space $V_h$, for $i = 1, \ldots, k$ we find*

$$Q_i(u_h) = a(u_h, p_{i,h}) = f(p_{i,h}) = Q_i(w_h), \tag{2.27}$$

*that is, the same approximation of the quantities of interest is found whether we proceed to a constrained minimization or not: the approach would thus be useless. Indeed, the unique solution $(w_h, \lambda_h)$ of the constrained problem (2.22) would be given by $w_h = u_h$, the solution of the unconstrained problem (2.4), and $\lambda_h = 0$.*

In the remainder of the thesis, we shall use a larger finite element space for the adjoint problems, denoted by $\tilde{V}_h$, than the approximation space $V_h \subset \tilde{V}_h$ for the primal problem. In practice, $\tilde{V}_h$ consists of higher-order hierarchical elements on the same mesh. Consequently, in order to compute the target values $\alpha$, we consider the following higher-order dual problems

$$\text{Find } \tilde{p}_i \in \tilde{V}_h \text{ such that } a(\tilde{v}, \tilde{p}_i) = Q_i(\tilde{v}), \quad \forall \tilde{v} \in \tilde{V}_h, \quad \forall i = 1, \dots, k, \tag{2.28}$$

and set $\alpha_i = f(\tilde{p}_i)$ for all $i = 1, \dots, k$. In later analysis, we will nonetheless also consider the dual problems in the same space $V_h$ defined by (2.7).

We now turn our attention to the numerical method that will be used to solve the finite-dimensional mixed problem described above. The mixed formulation (2.22) yields the following system of equations

$$\begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} W \\ \lambda_h \end{bmatrix} = \begin{bmatrix} F \\ \alpha \end{bmatrix}, \tag{2.29}$$

with $A_{ij} = a(\varphi_j, \varphi_i)$, $B_{ij} = b(e_j, \varphi_i) = Q_j(\varphi_i)$, where $\{e_j\}_{j=1}^k$ denotes the canonical basis of $\mathbb{R}^k$, $F_i = f(\varphi_i)$, and the components $w_i$ of $W$ are the coefficients of the solution $w_h$ with respect to the finite element basis functions $\varphi_i$, i.e. $w_h = \sum_i w_i \varphi_i$. This system could be solved directly as given since the augmented matrix is non-singular. However, its size is larger than that yielded by the classical unconstrained finite element method (2.4), which is simply $AU = F$. If the number of constraints is large, one may consider applying either the Uzawa or Augmented Lagrangian method [85].

Applying each of the $k$ functionals $Q_1, \dots, Q_k$ to the relation (2.23) and rearranging the terms, we obtain a linear system of size $k \times k$ with vector $\lambda_h \in \mathbb{R}^k$ as unknown

$$S_h \lambda_h = Q(u_h - w_h) = Q(u_h) - \alpha, \tag{2.30}$$

with $S_{h,ij} = Q_i(p_{j,h})$ and where we highlighted the dependency of this matrix on the mesh size $h$. As a result, instead of solving the augmented and possibly ill-conditioned linear system (2.29), one only needs to compute

1. the unconstrained solution $u_h$ of (2.4),
2. the adjoint solutions $p_{i,h}$ of (2.7),
3. the Lagrange multipliers $\lambda_h$ using (2.30), then
4. form the constrained solution $w_h$ using (2.23).

In other words, one solves $k+1$ (1 primal and $k$ dual) systems of the same size as the original finite element problem as well as one $k \times k$ linear algebraic problem. In fact, $S_h$ is precisely the Schur complement arising from the augmented matrix featured in (2.29), usually defined as $B^T A^{-1} B$. Indeed, each vector counterpart to the adjoint solution $p_{j,h}$ is given by $A^{-1}B_j$, where $B_j$ denotes the $j$-th column of $B$. When concatenating these $k$ column vectors, we form $A^{-1}B$. Applying now each of the $k$ functionals $Q_i$, we get $B^T A^{-1} B$. Since $B$ is injective and $A^{-1}$ is symmetric positive-definite, $S_h$ is symmetric positive-definite and thus invertible, so that the Schur complement equation (2.30) has a unique solution.

**Lemma 2** (Technical result on the Schur complements)**.** *The Schur complements $S_h$ converge towards a symmetric positive-definite matrix $S$ as the mesh size $h$ tends to zero.*

*Proof.* The entries of $S_h$ are given by $S_{h,ij} = Q_i(p_{j,h})$. By continuity of $Q$, the matrices $S_h$ converge to the matrix $S$ defined by $S_{ij} = Q_i(p_j)$. We denote by $\mathcal{M}_k$ the space of size $k \times k$ matrices, and by $\mathcal{S}_k^+$ the subset of symmetric and positive semi-definite matrices. Since the $S_h$'s are all symmetric and positive semi-definite, by closedness of $\mathcal{S}_k^+$ in $\mathcal{M}_k$, $S$ is symmetric and positive semi-definite as well. To prove definiteness, we note that $S_{ij} = Q_i(p_j) = a(p_i, p_j)$. It follows that $S$ is the Gram matrix of the family $\{p_i \in V, i = 1, \ldots, k\}$ for the inner product $a(\cdot, \cdot)$. Since the linear forms $Q_i \in V'$ are assumed to be linearly independent, the adjoint solutions $p_i \in V$ are also linearly independent, which implies that $S$ is positive-definite (Theorem 7.2.10 in [57]). □

We now establish a theorem stating that the proposed approach yields a solution that maintains near-optimality in energy norm.

**Theorem 4** (Near-optimality of the constrained solution)**.** *Let $\|\cdot\|_{\mathcal{E}}$ denote the energy norm on $V$, i.e. the norm induced by the bilinear form $a$, and $\|\cdot\|_1$ denote the 1-norm on $\mathbb{R}^k$. Let $u \in V$ denote the solution of the primal problem (2.1), $w_h \in V_h$ the solution of the constrained problem (2.22), and $u_h \in V_h$ the solution of the unconstrained problem (2.4). Assume there exists $C > 0$, independent of the mesh size $h$, such that*

$$\|Q(u) - Q(w_h)\|_1 \leq C \|Q(u) - Q(u_h)\|_1. \tag{2.31}$$

*Then there exists $D > 0$, independent of the mesh size $h$, such that*

$$\|u - w_h\|_{\mathcal{E}} \leq D \|u - u_h\|_{\mathcal{E}}. \tag{2.32}$$

*Proof.* Using Theorem 3, it holds

$$\begin{aligned} \|u - w_h\|_{\mathcal{E}} &\leq \|u - u_h\|_{\mathcal{E}} + \|u_h - w_h\|_{\mathcal{E}}, \\ &\leq \|u - u_h\|_{\mathcal{E}} + \|\sum_{i=1}^{k} \lambda_{h,i} p_{i,h}\|_{\mathcal{E}}, \\ &\leq \|u - u_h\|_{\mathcal{E}} + \sum_{i=1}^{k} |\lambda_{h,i}| \|p_{i,h}\|_{\mathcal{E}}, \\ &\leq \|u - u_h\|_{\mathcal{E}} + C_1 \|\lambda_h\|_1, \end{aligned} \tag{2.33}$$

where $C_1 = \max_{i=1,\dots,k} \|p_i\|_{\mathcal{E}} \geq \max_{i=1,\dots,k} \|p_{i,h}\|_{\mathcal{E}}$ is independent of the mesh size $h$. Now using (2.30) and the fact that the Schur complement $S_h$ is non-singular, it follows

$$\lambda_h = S_h^{-1} Q(u_h - w_h), \tag{2.34}$$

from which we obtain

$$\|\lambda_h\|_1 = \|S_h^{-1} Q(u_h - w_h)\|_1 \leq C_{S_h^{-1}} \|Q(u_h - w_h)\|_1, \tag{2.35}$$

where $C_{S_h^{-1}}$ denotes the matrix norm of $S_h^{-1}$ induced by $\|\cdot\|_1$, which depends on the mesh size $h$. To obtain a uniform bound, we use Lemma 2 and continuity of the matrix norm, so that the sequence $C_{S_h^{-1}}$ converges to $C_{S^{-1}}$, the matrix norm of $S^{-1}$. As a convergent sequence, it is bounded so there exists $\gamma \geq C_{S_h^{-1}}$, with $\gamma$ independent of the mesh size $h$, so that

$$\|\lambda_h\|_1 \leq \gamma \|Q(u_h - w_h)\|_1. \tag{2.36}$$

Then, using assumption (2.31)

$$\begin{aligned} \|Q(u_h - w_h)\|_1 &\leq \|Q(u - w_h)\|_1 + \|Q(u - u_h)\|_1, \\ &\leq (1 + C) \|Q(u - u_h)\|_1. \end{aligned} \tag{2.37}$$

Using now the boundedness of $Q$, it holds

$$\|Q(u - u_h)\|_1 \leq C_Q \|u - u_h\|_{\mathcal{E}}, \tag{2.38}$$

where $C_Q$ denotes the operator norm of $Q$ induced by $\|\cdot\|_{\mathcal{E}}$ and $\|\cdot\|_1$. Finally, we obtain

$$\|u - w_h\|_{\mathcal{E}} \leq D \|u - u_h\|_{\mathcal{E}}, \tag{2.39}$$

where $D = 1 + C_1(1 + C)\gamma C_Q$ is independent of the mesh size $h$. □

Essentially, Theorem 4 states that if the target values $\alpha \in \mathbb{R}^k$ are consistent with the problem, then the constrained solution $w_h$ maintains near-optimality in the energy norm. In particular, we also demonstrated that the vector of the Lagrange multipliers $\lambda_h \in \mathbb{R}^k$ necessarily converges to zero as $h$ tends to zero.

### 2.4 Inequality constraints

In this section, we focus on a slightly less restrictive approach where the equality constraints $Q(w_h) = \alpha$ are replaced by the following inequality constraints

$$|Q_i(w_h) - \alpha_i| \leq \varepsilon_i, \quad \forall i = 1, \ldots, k, \tag{2.40}$$

where each $\varepsilon_i$ is a positive scalar, possibly quite small. The rationale for replacing equality constraints by inequality constraints is twofold. First, since we consider the computable approximates $\alpha$ using the discretized adjoint problems (2.28) instead of the exact quantities $Q(u)$, we introduce some error in the target values. As a result, there is no need to exactly impose those perturbed values. The quantities $\varepsilon \in \mathbb{R}^k$ could be user-specified or could represent tolerances on the errors in the quantities of interest. Second, recall that our objective is to derive a finite element formulation that adequately represents the solution globally as well as quantities of interest of that solution. Intuitively speaking, incorporating equality constraints comes down to sacrificing the energy in order to satisfy the constraints (minimization in an affine space strictly contained in the "surrounding" space). Replacing equality constraints by inequality constraints would allow one to reduce the impact of this sacrifice and better represent the solution globally while maintaining a controlled (through parameters $\varepsilon$) representation of the quantities of interest. In order to simplify the exposition, we introduce the notation

$$\alpha^- = \alpha - \varepsilon \text{ and } \alpha^+ = \alpha + \varepsilon. \tag{2.41}$$

The necessary conditions for the solution of an inequality constrained minimization problem

are given by the KKT (Karush-Kuhn-Tucker) conditions [62, 67], leading here to

$$\text{Find } (w_h, \lambda_h^+, \lambda_h^-) \in V_h \times \mathbb{R}^k \times \mathbb{R}^k \text{ such that}$$
$$\begin{cases} a(w_h, v_h) + \lambda_h^+ \cdot Q(v_h) + \lambda_h^- \cdot Q(v_h) = f(v_h), & \forall v_h \in V_h, \\ \lambda_h^+ \geq 0,\ \lambda_h^- \leq 0, \\ \alpha^+ \geq Q(w_h),\ \alpha^- \leq Q(w_h), \\ \lambda_{h,i}^+ \left(\alpha_i^+ - Q_i(w_h)\right) = 0,\ \lambda_{h,i}^- \left(Q_i(w_h) - \alpha_i^-\right) = 0, & \forall i = 1, \ldots, k, \end{cases} \tag{2.42}$$

where the notation $\tau \geq 0$ (resp. $\tau \leq 0$) for a vector $\tau \in \mathbb{R}^k$ is employed to mean that all components of the vector are positive (resp. negative). The last conditions of the inequality constrained system (2.42) are usually called the "complementary conditions" and essentially state that for each of the $k$ constraints there are three possibilities:

1. $\lambda_{h,i}^+ = \lambda_{h,i}^- = 0$ and $\alpha_i^- \leq Q_i(w_h) \leq \alpha_i^+$;
2. $\lambda_{h,i}^+ = 0$, $\lambda_{h,i}^- < 0$ and $Q_i(w_h) = \alpha_i^-$;
3. $\lambda_{h,i}^- = 0$, $\lambda_{h,i}^+ > 0$ and $Q_i(w_h) = \alpha_i^+$.

In the first case, the $i$-th constraint is usually referred to as "non-binding", in the sense that it is naturally satisfied by the unconstrained minimization and does not have to be enforced (observe in this case that the $i$-th component of the Lagrange multipliers vanishes from the weak formulation – first equation of problem (2.42)); in the other cases, it is said "binding" and equality has to be enforced on the boundary of the admissible set: either $Q_i(w_h) = \alpha_i^+$ or $Q_i(w_h) = \alpha_i^-$. As a result, in an inequality constrained minimization, each constraint is either enforced with equality or discarded. The main difficulty in such problems rests on the determination of the set of active constraints. One could use a brute force approach and solve all $3^k$ problems, but there exist more efficient approaches. We mention for instance the interior point or barrier methods, see e.g. the IPOPT package [1, 94], which can be used to solve the inequality constrained system (2.42) at the expense of an iterative scheme.

In the present study, the task of finding the set of active constraints is somewhat simplified compared to a general problem. We describe the rationale for solving the inequality constrained system (2.42) in the case where there is only one constraint, i.e. $k = 1$. We start by finding $u_h \in V_h$, the solution to the unconstrained problem (2.4). Equivalently this could be viewed as assuming that the set of active constraints is empty, i.e. all Lagrange multipliers are zero. Next we form the quantity of interest $Q(u_h) \in \mathbb{R}$ and determine whether we are in case i) $\alpha^- \leq Q(u_h) \leq \alpha^+$; or ii) $Q(u_h) < \alpha^-$; or iii) $\alpha^+ < Q(u_h)$. If we are in the first case then the constraint is non-binding, i.e. the solution $(w_h, \lambda_h^+, \lambda_h^-)$ of the inequality con-

strained problem (2.42) is given by $(u_h, 0, 0)$, which is the unconstrained solution; if we are in the second case then the lower bound is binding, i.e. the solution is of the form $(w_h, 0, \lambda_h)$ with $\lambda_h < 0$; and if we are in the third case then the upper bound is binding, i.e. the solution is of the form $(w_h, \lambda_h, 0)$ with $\lambda_h > 0$. The reason follows from the fact that the Schur complement $S_h$ is positive-definite, i.e. $S_h > 0$ since $k = 1$, and as a result equation (2.30) implies that the Lagrange multiplier and the quantity on the right-hand side have same sign. Unfortunately, the reasoning does not extend to more than one constraint ($k > 1$) since in that case being positive-definite does not yield enough information on the coefficients of the Schur complement $S_h$.

An alternative approach could be to exploit the Schur complement equation (2.30) and solve each of the $3^k$ problems of size less or equal to $k \times k$, each associated with a different set of active constraints. For each resulting vector of Lagrange multipliers, a first check is whether the KKT conditions relative to their signs are respected: a Lagrange multiplier associated to an upper (resp. lower) bound should be positive (resp. negative). For each of the remaining potential solutions, one should solve the primal unknowns and check whether the inequalities $\alpha^- \leq Q(w_h) \leq \alpha^+$ hold. Since the minimization problem is convex, the KKT conditions are necessary and sufficient so that in practice not all $3^k$ problems need to be solved.

We will not show numerical examples using inequality constraints as they do not bring new insight when compared to the results with equality constraints.

## 2.5 Error estimation and adaptivity

In this section, we derive error estimates and design an adaptive strategy for the numerical approach considered in this study. We introduce an approach that could be coined a "global implicit method". We note that implicit methods usually introduce auxiliary residual problems defined on patches of elements or single elements in an effort to spare computational effort [3]. However such a paradigm is not the primary focus of this work, and we consider a global method instead.

In the current approach, local contributions to the error are derived on elements, and the elements with the largest contributions are marked for refinement. For the sake of clarity, we first describe the method for error estimation in energy norm and with respect to quantities of interest in the case of the classical solution $u_h$ of problem (2.4). We will then turn to error estimation in energy norm and with respect to quantities of interest for the solution $w_h$ of the constrained problem (2.22).

Recall that the energy norm is defined on $V$ by $\|v\|_{\mathcal{E}} = a(v,v)^{1/2}$. Let us introduce the residual functional $R^h$, defined with respect to $u_h$, the classical unconstrained finite element solution of (2.4)

$$R^h(u_h; v) = f(v) - a(u_h, v) = a(u - u_h, v), \quad \forall v \in V. \tag{2.43}$$

Thanks to the classical Galerkin orthogonality, we have that $R^h(u_h; v_h) = 0$ for any $v_h \in V_h$. As a result, $R^h(u_h; v) = R^h(u_h; v - v_h)$ for all $v \in V$ and $v_h \in V_h$. This residual is actually used for both the error estimation in the energy norm as well as in the quantities of interest. Indeed, the error in the energy norm is defined by

$$\mathcal{E}_h = \|u - u_h\|_{\mathcal{E}} = \sqrt{a(u - u_h, u - u_h)} = \sqrt{R^h(u_h; u - u_h)}. \tag{2.44}$$

Similarly, the error in each quantity of interest $Q_i$ is defined by

$$\mathcal{E}_i = |Q_i(u) - Q_i(u_h)| = |f(p_i) - a(u_h, p_i)| = \left|R^h(u_h; p_i)\right| = \left|R^h(u_h; p_i - p_{i,h})\right|. \tag{2.45}$$

For any $v \in V$, the scalar quantity $R^h(u_h; v)$ can be decomposed into local contributions. In order to illustrate this process, consider the following example for bilinear form $a$ and linear form $f$

$$a(u,v) = \int_\Omega a \nabla u \cdot \nabla v \, dx, \text{ and } f(v) = \int_\Omega f v \, dx + \int_{\Gamma_N} g v \, ds, \tag{2.46}$$

where $\Gamma_N \subset \partial\Omega$ denotes the Neumann part of the boundary of domain $\Omega$. Then, we have

$$R^h(u_h; v) = f(v) - a(u_h, v) = \int_\Omega f v \, dx + \int_{\Gamma_N} g v \, ds - \int_\Omega a \nabla u_h \cdot \nabla v \, dx. \tag{2.47}$$

After splitting the integrals from $\Omega$ to each element $K \subset \Omega$, we can use Green's first identity and rearrange the integrals over the boundary of each element as integrals over the set of mesh edges $\Gamma$

$$R^h(u_h; v) = \sum_{K \subset \Omega} \int_K r_K v \, dx - \sum_{\gamma \subset \Gamma} \int_\gamma j_\gamma v \, ds, \tag{2.48}$$

where we have introduced the interior element residual term $r_K = (\nabla \cdot (a \nabla u_h) + f)|_K$ and

the edge element residual term $j_\gamma$ defined on $\Gamma$ by

$$j_\gamma = \begin{cases} (a\nabla u_h)|_K \cdot n_K + (a\nabla u_h)|_{K'} \cdot n_{K'} & \text{if } \gamma = \partial K \cap \partial K', \\ (a\nabla u_h)|_K \cdot n_K - g & \text{if } \gamma = \partial K \cap \Gamma_N, \\ (a\nabla u_h)|_K \cdot n_K & \text{if } \gamma = \partial K \cap (\partial\Omega \backslash \Gamma_N)\,. \end{cases} \tag{2.49}$$

For the adaptive procedure, we wish to obtain local contributions that can be computed and compared elementwise. We choose

$$R^h(u_h; v) = \sum_{K \subset \Omega} R^h_K(u_h; v), \tag{2.50}$$

where $R^h_K(u_h; v)$ is the elementary contribution to the error, defined by

$$R^h_K(u_h; v) = \int_K r_K v \, dx - \frac{1}{2} \sum_{\gamma \subset (\partial K \backslash \partial\Omega)} \int_\gamma j_\gamma v \, ds - \sum_{\gamma \subset (\partial K \cap \partial\Omega)} \int_\gamma j_\gamma v \, ds. \tag{2.51}$$

This process can be used to compute the error either in energy norm (2.44) by setting $v = u - u_h$ or in the quantities of interest (2.45) by setting $v = p_i - p_{i,h}$. Of course the exact solution $u$ (resp. $p_i$) is unavailable in practice, so that it is replaced by an approximation, denoted $\tilde{u}$ (resp. $\tilde{p}_i$) computed in the same space $\tilde{V}_h$ as the one used to get the enhanced quantities of interest values.

We note that the decomposition suggested in (2.50) and (2.51) is not unique and somewhat arbitrary. However, this is not the subject of the thesis and we will only consider the proposed approach as it is often used in the literature [81].

As refinement criterion, we choose the so-called "maximum strategy" [78, 82], i.e. we mark for refinement all elements that satisfy

$$\frac{\left|R^h_K(u_h; v)\right|}{\max_K \left|R^h_K(u_h; v)\right|} > \delta, \tag{2.52}$$

where $\delta \in (0, 1)$ is a chosen threshold. In the numerical experiments, we chose $\delta = 0.5$.

We now turn our attention to error estimates for the solution $w_h$ of the constrained problem (2.22). Again, the approach is based on the use of the residual, which is evaluated this time with respect to the computed solution $w_h$

$$R^h(w_h; v) = f(v) - a(w_h, v) = a(u - w_h, v). \tag{2.53}$$

Note that the modified Galerkin orthogonality (2.24) yields

$$\begin{aligned} R^h(w_h; v) &= a(u - w_h, v - v_h) + \lambda_h \cdot Q(v_h), \\ &= R^h(w_h; v - v_h) + \sum_{j=1}^{k} \lambda_{h,j} a(v_h, p_{j,h}). \end{aligned} \tag{2.54}$$

The error in energy norm satisfies

$$\mathcal{E}_h = \|u - w_h\|_{\mathcal{E}} = \sqrt{a(u - w_h, u - w_h)} = \sqrt{R^h(w_h; u - w_h)}, \tag{2.55}$$

and the error in each quantity of interest is given by

$$\begin{aligned} \mathcal{E}_i = |Q_i(u) - Q_i(w_h)| &= |f(p_i) - a(w_h, p_i)|, \\ &= \left|R^h(w_h; p_i)\right|, \\ &= \left| R^h(w_h; p_i - p_{i,h}) + \sum_{j=1}^{k} \lambda_{h,j} a(p_{i,h}, p_{j,h}) \right|, \end{aligned} \tag{2.56}$$

where the modified Galerkin orthogonality (2.24) was used to derive (2.56). In (2.55) (resp. (2.56)), we proceeded to a straightforward extension of the classical approach (2.44) (resp. (2.45)). Though this approach yields satisfying results, we investigated a different error representation approach aimed at separating the two sources of errors in the numerical solution $w_h$, namely the classical error due to the discretization of space $V$ into the finite-dimensional space $V_h$ and the additional error term due to the constrained minimization. Using the classical Galerkin orthogonality between $u - u_h \in V$ and $u_h - w_h \in V_h$, it holds

$$\mathcal{E}_h^2 = \|u - w_h\|_{\mathcal{E}}^2 = \|u - u_h\|_{\mathcal{E}}^2 + \|u_h - w_h\|_{\mathcal{E}}^2, \tag{2.57}$$

where the first term is the discretization error in the classical solution $u_h$. Now using (2.44) and Theorem 3, it follows that

$$\mathcal{E}_h^2 = R^h(u_h; u - u_h) + \| \sum_{j=1}^{k} \lambda_{h,j} p_{j,h} \|_{\mathcal{E}}^2. \tag{2.58}$$

The second term can be interpreted as the error due to the introduction of the constraint. We further note that it can be computed exactly since the Lagrange multipliers $\lambda_h \in \mathbb{R}^k$ and the finite element adjoint solutions $p_{i,h} \in V_h$ are known at this stage. Moreover, local contributions can be derived by decomposing the integral defined on $\Omega$ into integrals on each

element $K \subset \Omega$.

As far as the error in the quantity of interest $Q_i$ is concerned, it holds

$$\begin{aligned} \mathcal{E}_i = |Q_i(u) - Q_i(w_h)| &= |Q_i(u - u_h) + Q_i(u_h - w_h)|, \\ &= \left| R^h(u_h; p_i - p_{i,h}) + \sum_{j=1}^{k} \lambda_{h,j} Q_i(p_{j,h}) \right|, \\ &= \left| R^h(u_h; p_i - p_{i,h}) + \sum_{j=1}^{k} \lambda_{h,j} a(p_{j,h}, p_{i,h}) \right|, \end{aligned} \tag{2.59}$$

where we have used (2.45) and Theorem 3. The first term is the contribution due to the discretization error in the classical solution $u_h$. The second term can be interpreted as the error due to the introduction of the constraint. Again, the second term can be computed exactly and local contributions on each element can be derived.

The different contributions to the errors will be illustrated in the next section. However, by replacing the adjoint solution $p_i$ by the computable approximation $\tilde{p}_i \in \tilde{V}_h$ either in (2.56) or in (2.59), one obtains an estimate of the error in the quantity of interest that is zero. Indeed, $Q_i(w_h) = \alpha_i = f(\tilde{p}_i) = Q_i(\tilde{u})$. Of course, one could use an even higher-order approximation for the purpose of error estimation, but the cost of the method would then be prohibitive compared to a traditional approach. Nevertheless, the local contributions can be used to mark the elements that contribute largely to the error.

In the case of adaptation for the error in the energy norm (2.58), the element contributions are defined as

$$\begin{aligned} \mathcal{E}_h^2 &= R^h(u_h; u - u_h) + \| \sum_{j=1}^{k} \lambda_{h,j} p_{j,h} \|_{\mathcal{E}}^2, \\ &= \sum_{K \subset \Omega} R_K^h(u_h; u - u_h) + \sum_{i,j=1}^{k} \lambda_{h,i} \lambda_{h,j} a(p_{i,h}, p_{j,h}), \\ &= \sum_{K \subset \Omega} \left( R_K^h(u_h; u - u_h) + \sum_{i,j=1}^{k} \lambda_{h,i} \lambda_{h,j} a_K(p_{i,h}, p_{j,h}) \right), \end{aligned} \tag{2.60}$$

where the first term $R_K^h(u_h; u - u_h)$ was defined in (2.51) and the bilinear form $a_K$ relative to each element $K$ is given by

$$a_K(u, v) = \int_K a \nabla u \cdot \nabla v \, dx, \tag{2.61}$$

following the example for the bilinear form $a$ chosen in (2.46). To avoid eventual cancellation between the two sources during the marking process, we will consider the following refinement indicator

$$\left|R_K^h(u_h; u - u_h)\right| + \left|\sum_{i,j=1}^{k} \lambda_{h,i}\lambda_{h,j} a_K(p_{i,h}, p_{j,h})\right|. \tag{2.62}$$

The choice of introducing absolute values when computing the local indicators may lead to pessimistic results. However, it is motivated by observing that the two contributions may cancel each other while still producing large sources of errors that may need to be controlled. Again, in practice the exact solution $u \in V$ is replaced by the computable approximation $\tilde{u} \in \tilde{V}_h$.

In the case of adaptation for the error in the $i$-th quantity of interest (2.59), the element contributions are defined as

$$\begin{aligned} \mathcal{E}_i &= \left|R^h(u_h; p_i - p_{i,h}) + \sum_{j=1}^{k} \lambda_{h,j} a(p_{j,h}, p_{i,h})\right|, \\ &= \left|\sum_{K\subset\Omega} \left(R_K^h(u_h; p_i - p_{i,h}) + \sum_{j=1}^{k} \lambda_{h,j} a_K(p_{j,h}, p_{i,h})\right)\right|. \end{aligned} \tag{2.63}$$

As previously, to avoid eventual cancellation between the two sources during the marking process, we will consider the following refinement indicator

$$\left|R_K^h(u_h; p_i - p_{i,h})\right| + \left|\sum_{j=1}^{k} \lambda_{h,j} a_K(p_{j,h}, p_{i,h})\right|. \tag{2.64}$$

Again, in practice the adjoint solution $p_i \in V$ is replaced by the computable approximation $\tilde{p}_i \in \tilde{V}_h$.

## 2.6 Numerical Examples

In this section, we numerically illustrate the proposed approach on some academic boundary-value problems. For the finite element simulations, we use square elements and $V_h$ is defined as the space spanned by the bilinear Lagrange functions. For the enhanced quantities of interest and error estimation, we use $\tilde{V}_h$ the space spanned by the hierarchical integrated Legendre polynomials up to quadratic order.

We will consider three examples: the first two consist of a Poisson equation. They are used

to illustrate the efficiency of the method introduced in this work under uniform refinements. In the first example we consider a single quantity of interest that is conforming to the finite element mesh, while in the second example, we consider two quantities of interest that no longer conform to the mesh. The last example involves a diffusion equation with a piecewise constant coefficient, mimicking the so-called "L-shaped problem" in which the exact solution exhibits weak-singularities. It is used to illustrate the efficiency of the adaptive mesh refinement procedure introduced in this chapter.

**Example 1** The first model problem we consider consists of the Poisson equation with homogeneous Dirichlet conditions

$$\begin{cases} -\Delta u = 1, & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{2.65}$$

where $\Omega = (0,1)^2$. The exact solution of (2.65) can be found using Fourier series and is shown in Figure 2.1. We mention that $u \in H^3(\Omega)$ for this problem.

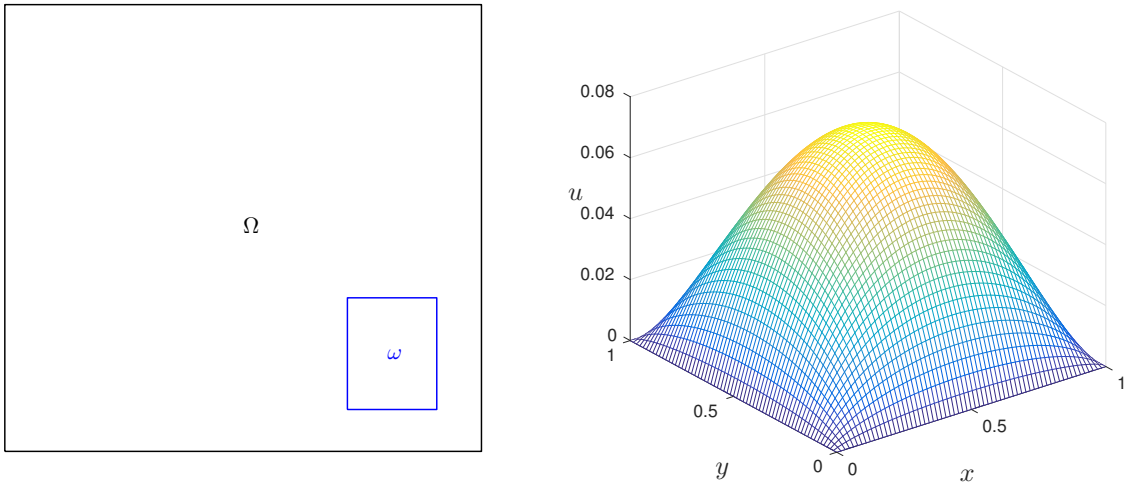


Figure 2.1 Geometry (left) and solution (right) for Example 1.

We also suppose that one is interested in the scalar quantity

$$Q(u) = \frac{1}{|\omega|} \int_{\omega} u \, dx, \tag{2.66}$$

where $\omega$ is a subdomain of $\Omega$, illustrated in Figure 2.1, and defined as

$$\omega = \{(x, y) \in \Omega;\ 23/32 \leq x \leq 29/32,\ 3/32 \leq y \leq 11/32\}\,. \tag{2.67}$$

In this first example, the region of interest $\omega$ coincides with the mesh (after a few uniform refinements).

The exact value of the quantity of interest (2.66) can be computed using the Fourier expansion of $u$. We mention that $Q$ is continuous on $H^1(\Omega)$. The adjoint solution $p \in H^3(\Omega)$ is shown in Figure 2.2.

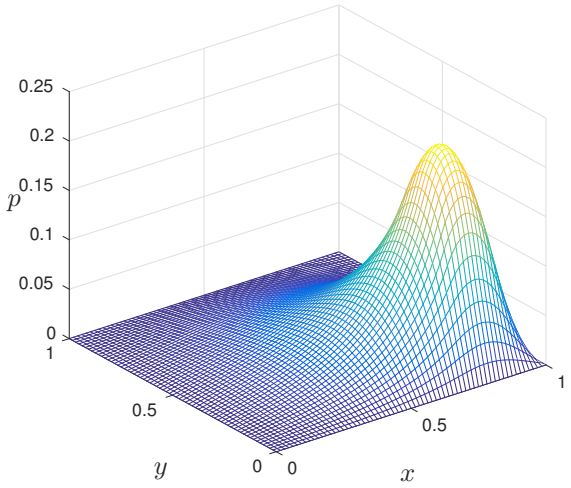


Figure 2.2 Adjoint solution for Example 1.

In order to assess the efficiency of the constrained approach introduced in this chapter as well as of the error estimation procedure, we generate a sequence of uniform refinements with inverse mesh sizes $h^{-1} = 4, 8, 16, \ldots, 256$, estimate the resulting errors, and measure the effectivity of the estimators. In Figure 2.3 (left), we show the normalized exact errors in energy norm and in the quantity of interest for both the classical unconstrained approach $u_h$ and the constrained approach $w_h$ proposed in this chapter. In Figure 2.3 (right), we also show the effectivity indices $i_{\text{eff}}$, which are classically defined as the ratio of the estimated errors over the exact errors.

As it can be seen from Figure 2.3 (left), the errors in energy norm for the classical unconstrained approach $u_h$ and for the constrained approach $w_h$ have the same convergence rate $\mathcal{O}(h)$, as predicted by the results of a priori error estimation and Theorem 4. The effec-

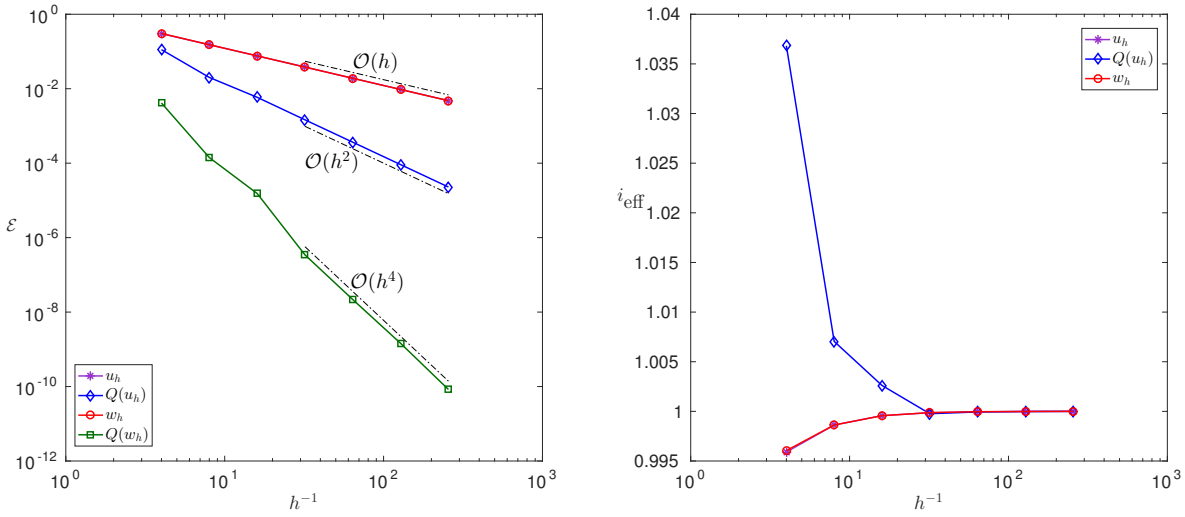


Figure 2.3 Exact errors (left) and effectivity indices (right) as functions of the inverse mesh size $h^{-1}$.

tivity indices for the errors in the energy norm are also very similar for both approaches and are in the $[0.996, 1]$ interval.

Concerning the error in the quantity of interest, the results are striking on this simple example: the rate of convergence for the constrained approach is twice as large as that obtained by the classical approach: $\mathcal{O}(h^4)$ vs $\mathcal{O}(h^2)$, again in agreement with the results of a priori error estimation [13], based on the fact that $Q(w_h) = Q(\tilde{u})$ and that both $u$ and $p$ are sufficiently smooth. The error estimator for the quantities of interest is only available for the unconstrained approach (recall the discussion about the error estimator for $Q_i(w_h)$ in Section 2.5) with an effectivity index ranging in the $[0.9998, 1.037]$ interval.

In Figure 2.4, we compare the contributions to the error as defined in (2.58) and in (2.59). In Figure 2.4 (left), we observe that the two sources of error in the energy norm do not have the same convergence rate. The error term due to the constraint decreases much more rapidly than the classical discretization error. As a result, the total error is similar to the discretization error: indeed $w_h$ is near-optimal in the energy norm. The situation is completely different for the two terms of the error in the quantity of interest: in Figure 2.4 (right), we observe that the two terms are almost equal. In fact, they have opposite signs so that they mostly cancel when added. As a result the convergence rate of the error in the quantity of interest for $w_h$ is increased.

**Example 2** The second model problem consists of the same boundary-value problem (2.65)

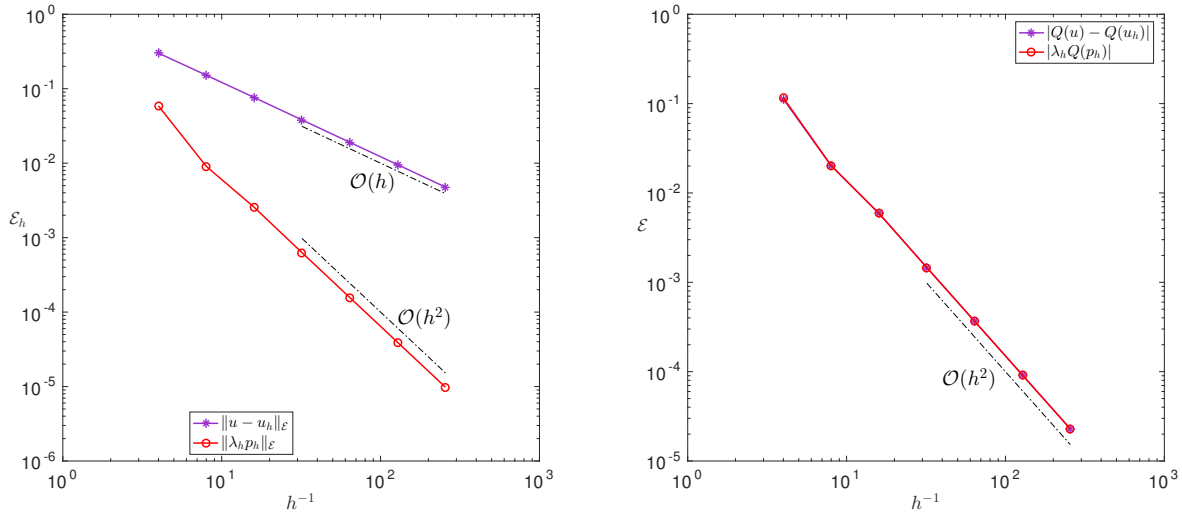


Figure 2.4 Detail of the contributions as functions of the inverse mesh size $h^{-1}$: error in the energy norm (left); error in the quantity of interest (right).

introduced in Example 1. However, this time we suppose that we are interested in the two quantities

$$Q_1(u) = \frac{1}{|\omega_1|}\int_{\omega_1} u\,dx, \text{ and } Q_2(u) = \frac{1}{|\omega_2|}\int_{\omega_2} \boldsymbol{i}\cdot\nabla u\,dx, \tag{2.68}$$

where $\boldsymbol{i}$ denotes the horizontal unit vector, and $\omega_1$, $\omega_2$ are two subdomains of $\Omega$, illustrated in Figure 2.5 (left), and defined as

$$\begin{aligned}\omega_1 &= \left\{(x,y)\in\Omega;\ 1/\sqrt{2}\le x\le 1/\sqrt{2}+1/\sqrt{30},\ 1/\sqrt{18}\le y\le 1/\sqrt{18}+1/\sqrt{17}\right\},\\ \omega_2 &= \left\{(x,y)\in\Omega;\ 1/\sqrt{40}\le x\le 1/\sqrt{40}+1/\sqrt{20},\ 1/\sqrt{3}\le y\le 1/\sqrt{3}+1/\sqrt{13}\right\},\end{aligned} \tag{2.69}$$

where the irrational coordinates were chosen so that the regions of interest $\omega_1$, $\omega_2$ never coincide with the meshes. In Figure 2.5 (right), we present the adjoint solution $p_2$. The adjoint solution $p_1$ is similar to that shown in Figure 2.2 and is not shown here.

Again, the exact values of the quantities of interest (2.68) can be computed using the Fourier expansion of $u$. We mention that $Q_1$ and $Q_2$ are continuous on $H^1(\Omega)$. Furthermore, we have the following regularity for the adjoint solutions: $p_1 \in H^3(\Omega)$ while $p_2 \in H^2(\Omega)$, only.

Once more, we perform a sequence of uniform refinements, estimate the resulting errors, and measure the effectivity indices of the estimators. In Figure 2.6 (left), we show the normalized exact errors in energy norm and in the two quantities of interest for both the classical unconstrained solution $u_h$ and the constrained solution $w_h$. The effectivity indices $i_{\text{eff}}$

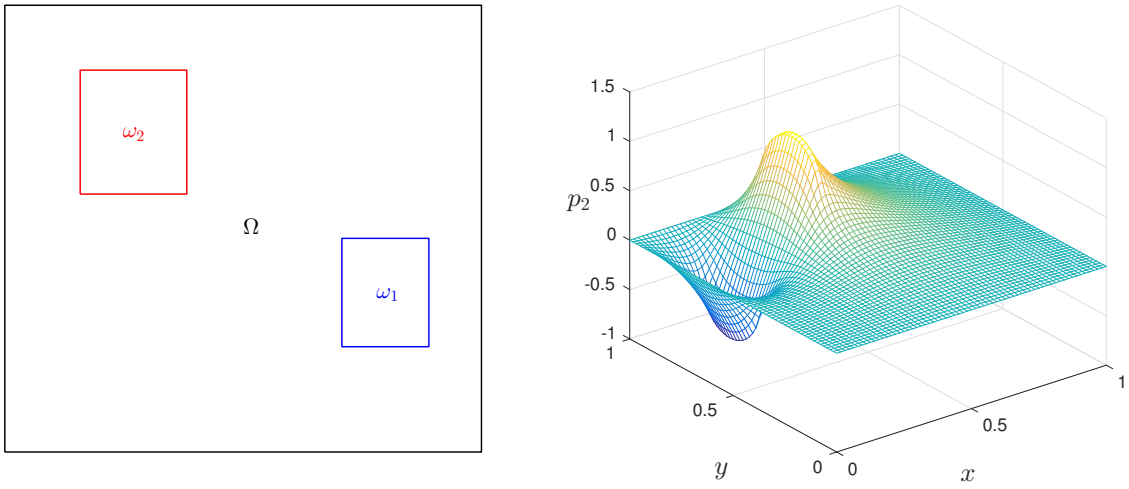


Figure 2.5 Geometry (left) and adjoint solution $p_2$ (right) for Example 2.

are shown for this case in Figure 2.6 (right).

We observe from Figure 2.6 (left) that the errors in energy norm in $u_h$ and $w_h$ are again almost equal: the error for the constrained solution is 2% larger than for the unconstrained solution on the coarsest mesh considered, and only 0.0002% larger for the finest mesh considered. This behavior is very similar to that observed in the first example. The effectivity indices for the errors in the energy norm are also very similar for both approaches and remain in the $[0.996, 1]$ interval.

Concerning the error in the quantity of interest $Q_1$, the rate of convergence for the constrained approach is again twice as large as that obtained by the classical approach: $\mathcal{O}(h^4)$ vs $\mathcal{O}(h^2)$. For the quantity of interest $Q_2$, the rate of convergence only increases by one order: $\mathcal{O}(h^3)$ vs $\mathcal{O}(h^2)$. This difference is due to the limited regularity of the adjoint solution for the second quantity of interest: recall $p_1 \in H^3(\Omega)$ while $p_2 \in H^2(\Omega)$ only. Indeed, with such regularity but no more, we have $\|p_2 - \tilde{p}_2\|_{\mathcal{E}} = \mathcal{O}(h)$ as $h \to 0$ while $\|u - \tilde{u}\|_{\mathcal{E}} = \mathcal{O}(h^2)$. Hence the quantity $Q_2(u)$ is approximated with order $\mathcal{O}(h^3)$. The error estimator for the quantities of interest shows an effectivity index ranging in the $[0.969, 1.110]$ interval.

**Example 3** Again we consider $\Omega = (0,1)^2$, and choose a point $(x_c, y_c) \in \Omega$ so that $\Omega$ is split into two regions: $\Omega_1 = \{(x,y) \in \Omega; x > x_c \text{ and } y > y_c\}$, and the complementary region $\Omega_0 = \Omega \backslash \Omega_1$. We choose $x_c = y_c = 1/2$. We introduce on $\Omega$ the piecewise constant coefficient $a$ such that $a_{|\Omega_i} = a_i, i = 0, 1$, with $a_0 = 1$ and $a_1 = 100$. The third problem consists of a diffusion equation with diffusivity coefficient $a$ so that it features a weak-singularity (i.e.

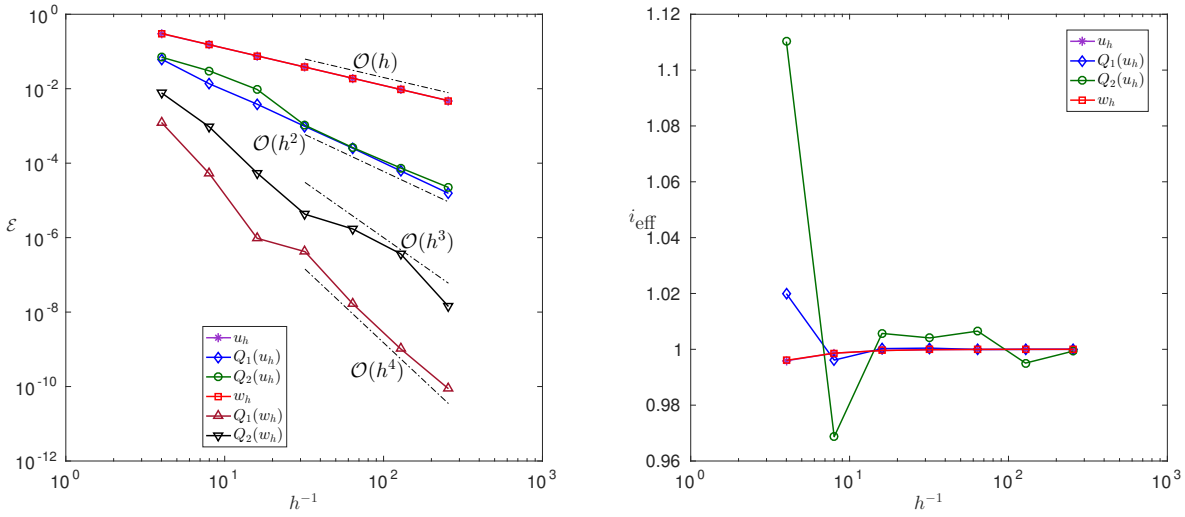


Figure 2.6 Exact errors (left) and effectivity indices (right) as functions of the inverse mesh size $h^{-1}$.

the gradient of the solution is singular), whose solution $u$ is subjected to Robin boundary conditions on $\partial\Omega$

$$\begin{cases} -\nabla \cdot a\nabla u = f, & \text{in } \Omega, \\ \boldsymbol{n} \cdot a\nabla u + u = g, & \text{on } \partial\Omega. \end{cases} \tag{2.70}$$

The exact solution $u$ is constructed using the so-called manufactured solution method and is chosen to be harmonic of the form

$$u = u(r,\theta) = \begin{cases} A_0 r^\mu \cos(\mu\theta) + B_0 r^\mu \sin(\mu\theta), & \text{in } \Omega_0, \\ A_1 r^\mu \cos(\mu\theta) + B_1 r^\mu \sin(\mu\theta), & \text{in } \Omega_1, \end{cases} \tag{2.71}$$

where $(r,\theta)$ are the polar coordinates centered at $(x_c, y_c)$. The constants $\mu, A_0, B_0, A_1$ and $B_1$ are chosen such that $u$ is continuous in $\Omega$ and $\boldsymbol{n} \cdot a\nabla u$ is continuous across the interface between $\Omega_0$ and $\Omega_1$. The source term $f$ and boundary datum $g$ are derived by injecting (2.71) into (2.70). We mention that $f = 0$ because $u$ is taken to be harmonic in $\Omega$. We report the values of the constant parameters $\mu, A_0$, and $B_0$ in Table 2.1, while $A_1 = A_0$ and $B_1 = (a_0/a_1)B_0$.

Note that by construction, we have $u \in H^{1+\mu-\epsilon}(\Omega)$, where $\epsilon > 0$ is arbitrarily small. The manufactured problem resembles the so-called "L-shaped problem" constructed here with a finite contrast $a_1/a_0$. As a result, the solution exhibits a weak-singularity at the corner $(x_c, y_c)$ and its gradient is discontinuous along the interface $\partial\Omega_1 \backslash \partial\Omega$. In order to simplify the pre-

Table 2.1 Values of the parameters $\mu$, $A_0$, and $B_0$ used for Example 3.

| $\mu$ | $A_0$ | $B_0$ |
|---|---|---|
| 0.6739 | 0.0171 | 0.9998 |

sentation, the initial mesh is chosen to be conforming to the interface by taking $h = 1/2$.

We show in Figure 2.7 the geometry and the manufactured solution for this example. The quantities of interest are the same as in the second example, see (2.68)–(2.69). The adjoint solutions associated with these quantities of interest are illustrated in Figure 2.8.

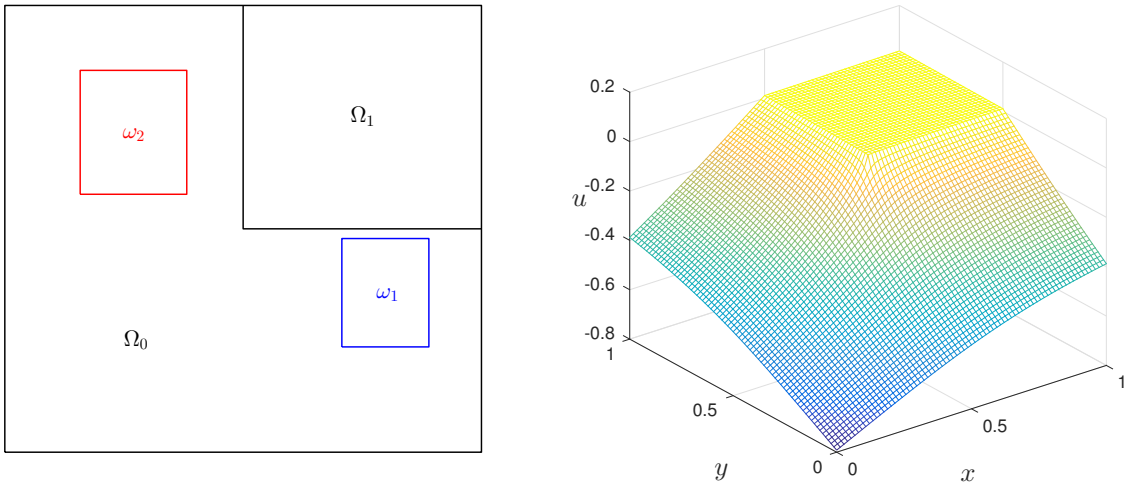


Figure 2.7 Geometry (left) and solution (right) for Example 3.

We now turn to the adaptive procedure for the above problem. When an element is marked for refinement, it is divided into four squares of equal areas, which introduces hanging nodes [31]. We will compare four types of refinement based on the following criteria:

(a) adaptation in norm for the unconstrained solution $u_h$,

(b) adaptation in the quantities of interest for the unconstrained solution $u_h$,

(c) adaptation in norm for the constrained solution $w_h$,

(d) adaptation in the quantities of interest for the constrained solution $w_h$.

We mention that for the adaptation based on the two quantities of interest, elements are marked for refinement if any of the two error indicators associated with $Q_1$ and $Q_2$ exceeds

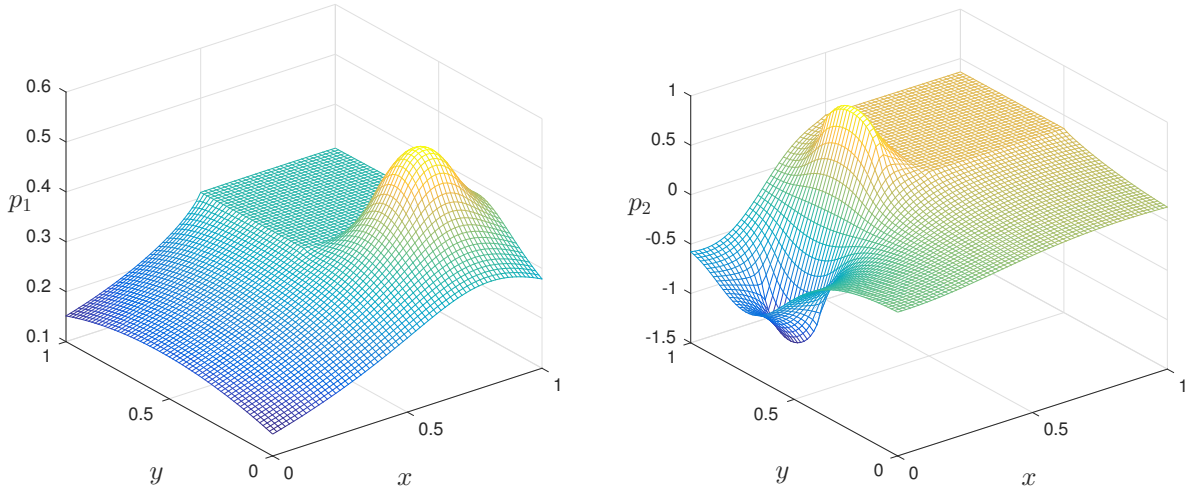


Figure 2.8 Adjoint solutions for Example 3.

the prescribed threshold.

We show in Figure 2.9 the resulting sequences of adapted meshes. All four methods manage to capture the singularity at the corner of the interface. In addition, we note that the approaches based on the quantities of interest, (b) and (d), accentuate the refinement in the two regions of interest $\omega_1$, $\omega_2$. Furthermore, the refinements in energy norm (a) and (c) are very similar, which is due to the relatively small contribution of the term related to the constraint, recall Figure 2.4 and Eq. (2.58). Conversely, the adapted meshes obtained for the refinement based on the quantities of interest (b) and (d) are less similar, recall Figure 2.4 and Eq. (2.59).

The convergence plots for the four methods are shown in Figure 2.10. We mention that 25 levels (15, 412 dofs) were considered for approaches (a) and (c), 26 levels (16, 020 dofs) for approach (d), and 30 levels (16, 041 dofs) for approach (b). The two approaches based on the constrained solution yield the best results in terms of convergence of the quantities of interest. In particular, all results asymptotically converge at optimal rates, that is, denoting the number of degrees of freedom as $N_{\text{dof}}$,

$$\|u - u_h\|_{\mathcal{E}} \leq C \left(N_{\text{dof}}\right)^{-p/d} \text{ and } \|u - w_h\|_{\mathcal{E}} \leq C \left(N_{\text{dof}}\right)^{-p/d}, \tag{2.72}$$

with $p = 1$ for both $u_h$ and $w_h$, and

$$|Q_i(u) - Q_i(u_h)| \leq C \left(N_{\text{dof}}\right)^{-2p/d} \text{ and } |Q_i(u) - Q_i(w_h)| \leq C \left(N_{\text{dof}}\right)^{-2p/d}, \tag{2.73}$$

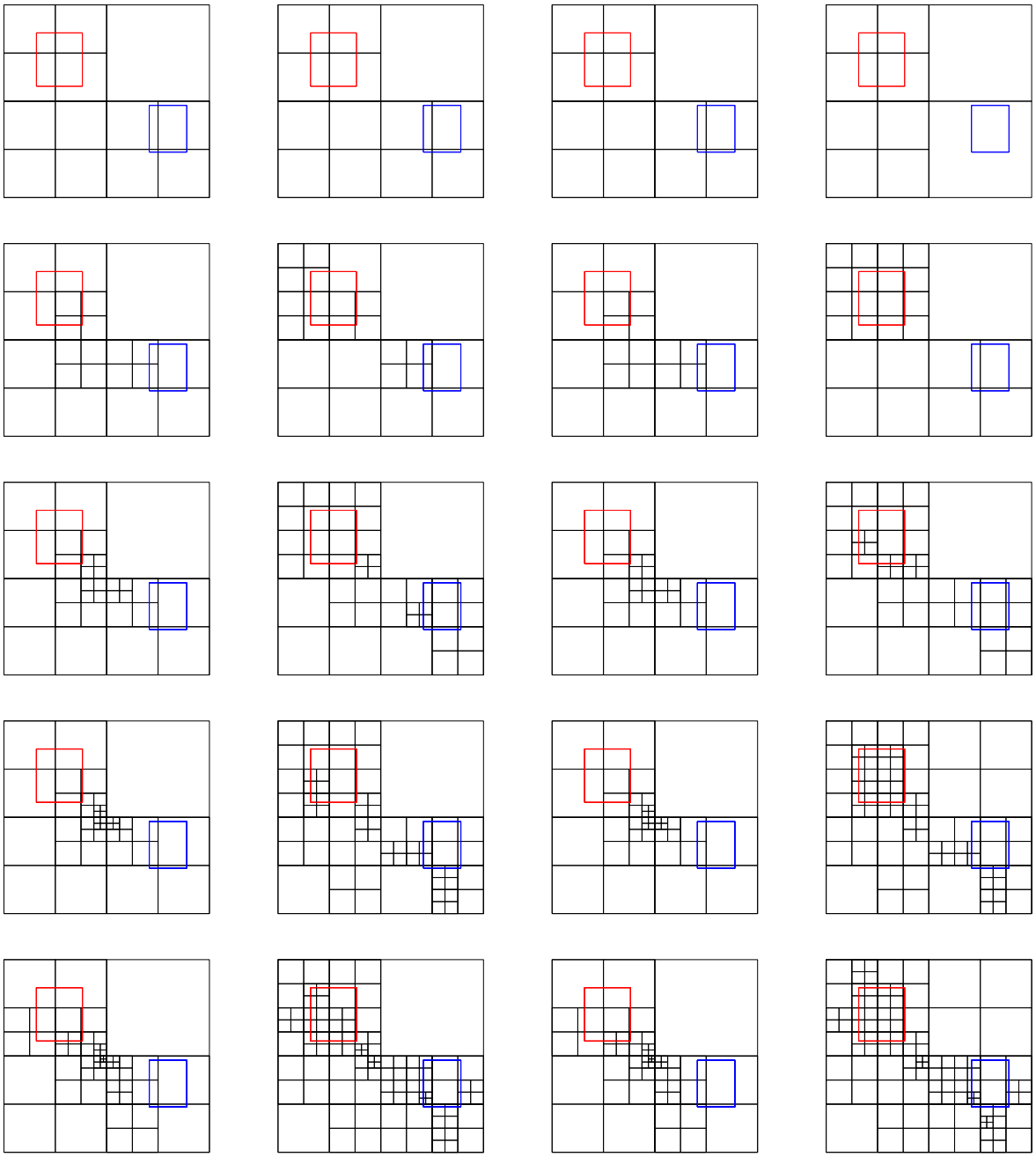

Figure 2.9 Sequences of adapted meshes for Example 3. (a): refinement in energy norm for $u_h$; (b): refinement in the quantities of interest for $u_h$; (c): refinement in energy norm for $w_h$; (d): refinement in the quantities of interest for $w_h$.

with $p = 1$ for $u_h$ and $p = 2$ for $w_h$, for both $i = 1, 2$.

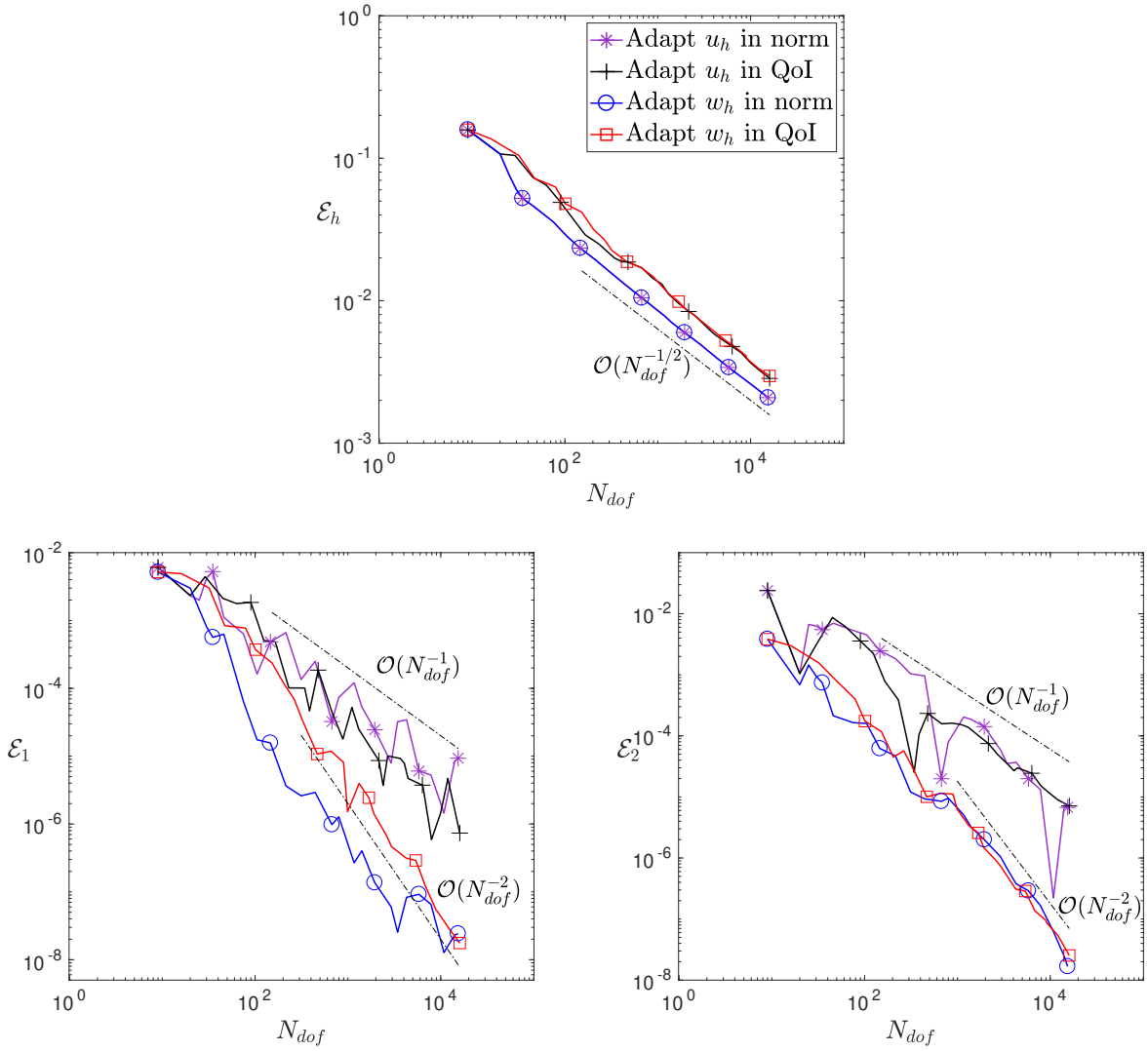


Figure 2.10 Convergence results for the four considered methods. top: convergence in energy norm; bottom-left and bottom-right: convergence in the two quantities of interest.

## 2.7 Conclusion

We have introduced a novel formulation designed to take into account quantities of interest in finite element approximations. The approach is very different from classical procedures involving goal-oriented adaptivity. In the latter the adjoint problems are solved after computing the primal solution in order to assess and control the errors in the quantity of interest. In the proposed approach, the adjoint problems are solved beforehand in order to obtain en-

hanced values for the quantities of interest, which are then introduced in the formulation of the primal problem using constraints. In this chapter, we have proved that the corresponding mixed formulation was well-posed, and that the constrained finite element solution retained near-optimality in the energy norm while being much more accurate in the quantities of interest. Error estimators were derived for the proposed approach, with an emphasis on explicitly identifying the two contributions to the error, namely the classical discretization error and the error due to the introduction of a constraint. The efficiency of the novel formulation and of the corresponding mesh refinement procedure was demonstrated on a several numerical examples.

Future work will focus on the extension of the present work to non-linear problems and non-linear quantities of interest. We also note that the methodology can be straightforwardly extended to a worst-case multi-objective formulation [89] by considering one dual problem using an approximate supporting functional of the objective set rather than solving a dual problem for each quantity of interest.

In the following chapters, we will extend the proposed method to reduced-order modeling, exemplified using the Proper Generalized Decomposition [39, 75] method, for which we first need to develop the framework needed to enforce constraints. We anticipate that the construction of reduced models using such an approach could help provide accurate estimates of quantities of interest at very low computational cost. This would be particularly useful for the treatment of uncertainty quantification problems, in which case one has to estimate output quantities of interest for a very large number of parameter samples.

# CHAPTER 3 APPROXIMATION OF CONSTRAINED PROBLEMS USING THE PGD METHOD WITH APPLICATION TO PURE NEUMANN PROBLEMS

*In this chapter we introduce, analyze, and compare several approaches designed to incorporate a linear (or affine) constraint within the Proper Generalized Decomposition framework. We apply the considered methods and numerical strategies to two classes of problems: the pure Neumann case where the role of the constraint is to recover unicity of the solution; and the Robin case, where the constraint forces the solution to move away from the already existing unique global minimizer of the energy functional. This chapter is largely inspired by [65].*

## 3.1 Introduction

The need for fast evaluation of surface responses in parametric analyses has spurred the development of novel model reduction methods to construct, in an effective manner, solutions to boundary-value problems. One such method is the Proper Generalized Decomposition (PGD) framework [39, 40], in which the solution is sought numerically using the concept of separation of variables. The PGD approximation scheme allows one to simplify a complex problem into a set of coupled problems, defined with respect to each spatial and/or parametric variable, which can be further decoupled using the so-called Alternating Directions scheme [8, 40, 45]. There exist to date a variety of PGD methods [75], which have been adapted to the nature of the problem at hand and which have been successfully tested on a wide range of applications and model problems, see e.g. [9, 25, 26, 34, 38, 41, 58, 76, 93, 97]. Yet, and to the best of the authors' knowledge, none of these applications include problems subjected to constraints defined on the solution space, except, maybe, the case of the incompressible Navier-Stokes equations, for which the divergence-free constraint is treated using a fractional-step or projection method [43, 44]. We also mention the works presented in [2, 51] and the references therein where a penalization formulation is used to circumvent the mixed formulation arising from the constrained problem.

The objective of the chapter is therefore to study how a general boundary-value problem involving a linear or affine constraint can be treated within the PGD setting. For the sake of simplicity in the exposition, but without loss of generality, the model problem that we have chosen to focus on consists of a two-dimensional Poisson equation with pure Neumann boundary conditions prescribed on the whole boundary of the domain. It is well-known that the solution to such a problem is given only within a constant and that one needs to

prescribe an additional constraint on the solution in order to fix the constant [24]. The main challenge in applying a constraint functional within the PGD framework arises from the fact that the original problem is decoupled into subproblems with respect to each spatial and/or parametric variable while the constraint should be applied to the solution globally. Classical methods used to enforce constraints are the Penalization, Lagrange Multiplier, and Augmented Lagrangian methods [85]. Our goal here is to see if and how these methods and their numerical implementations (direct, Uzawa, iterative Uzawa) can be extended to the case of PGD formulations.

The chapter is organized as follows: In Section 3.2, we first describe the model problem, namely a pure Neumann boundary-value problem in terms of the Poisson equation. We then review different approaches, namely Penalization, Lagrange Multiplier, and Augmented Lagrangian methods, to impose a constraint in order to recover the unicity of the solution. We also introduce a Robin boundary-value problem as a perturbation of the Neumann problem. The main difference with the latter is that it already admits a unique solution without resorting to any constraint on the solution. We will nevertheless consider a constrained Robin problem in order to compare the influence of the methods on the behavior of the solution with the case of the pure Neumann problem. In Section 3.3, we briefly describe the finite element discretization of the constrained problems and review some classical numerical strategies for solving these problems. In Section 3.4, we present a classical PGD formulation and extend above methods and strategies to the PGD formulation of the constrained problems. Numerical examples are presented in Section 3.5 to analyze the performance of each of the methods to the Neumann and Robin problems. We finally provide some concluding remarks in Section 3.6.

## 3.2 Model problem

Let $d \in \mathbb{N}$ be such that $d \geq 2$ and let $\Omega_i$ be open intervals $(a_i, b_i) \subset \mathbb{R}$, $i = 1, \ldots, d$ such that the domain $\Omega = \Pi_{i=1}^{d} \Omega_i$ forms an open, hyper-rectangular, bounded subset of $\mathbb{R}^d$ with boundary $\partial\Omega$. We shall denote by $\boldsymbol{n}$ the outward normal unit vector to $\Omega$ and by $|\Omega|$ a measure of $\Omega$.

We consider in this chapter the so-called pure Neumann boundary-value problem

$$\text{Find } u \text{ such that } \begin{cases} -\nabla \cdot (a \nabla u) = f, & \text{in } \Omega, \\ \boldsymbol{n} \cdot a \nabla u = g, & \text{on } \partial\Omega, \end{cases} \tag{3.1}$$

where $a = a(x) \in L^\infty(\Omega)$ is strictly positive and the data $f \in L^2(\Omega)$ and $g \in H^{1/2}(\partial\Omega)$ are

given such that the so-called compatibility condition

$$\int_\Omega f\,dx + \int_{\partial\Omega} g\,ds = 0, \tag{3.2}$$

is satisfied. In that case, the Fredholm alternative implies that above problem admits solutions up to an additive constant [24].

A weak formulation associated with Problem (3.1) reads

$$\text{Find } u \in H^1(\Omega) \text{ such that } a(u,v) = f(v), \quad \forall v \in H^1(\Omega), \tag{3.3}$$

where the bilinear form $a$ and linear form $f$ defined on $H^1(\Omega)$ are given by

$$\begin{aligned} a(u,v) &= \int_\Omega a\nabla u \cdot \nabla v\,dx, \\ f(v) &= \int_\Omega fv\,dx + \int_{\partial\Omega} gv\,ds. \end{aligned} \tag{3.4}$$

Alternatively, Problem (3.3) can be recast as a minimization problem by introducing the energy functional

$$J(u) = \frac{1}{2}a(u,u) - f(u), \tag{3.5}$$

and by minimizing $J$ over $H^1(\Omega)$.

Solutions to Problems (3.3) or (3.5) are not unique in $H^1(\Omega)$ since the bilinear form $a$ fails to be coercive in that space. In practice, unicity of the solution is often recovered by imposing the value of the solution at a given point in $\Omega$ or on $\partial\Omega$. Unfortunately, this approach yields an ill-posed problem as the point-value functional is not well defined for functions of $H^1(\Omega)$ when $d \geq 2$. A proper way to proceed is to search solutions in the subspace $V$ of $H^1(\Omega)$ of zero-mean functions [24]

$$V = \left\{ v \in H^1(\Omega); \frac{1}{|\Omega|}\int_\Omega v\,dx = 0 \right\}, \tag{3.6}$$

often referred to as the quotient space and denoted by $V = H^1(\Omega)/\mathbb{R}$. Since the bilinear form $a$ is coercive over $V$, the problem

$$\text{Find } u \in V \text{ such that } a(u,v) = f(v), \quad \forall v \in V, \tag{3.7}$$

is now well-posed. However, when considering discretization methods such as the Finite Element Method, Problem (3.7) is never solved as is, as it is difficult to construct trial

and test functions with zero-mean. Instead, one reformulates the problem as a constrained problem by minimizing $J$ over $V$, that is, by minimizing $J$ over $H^1(\Omega)$ subjected to the constraint that the solution has zero-mean. Let $Q$ denote the functional in $H^{-1}(\Omega)$ such that

$$Q(v) = \frac{1}{|\Omega|}\int_\Omega v\,dx. \tag{3.8}$$

The zero-mean constraint on function $u \in H^1(\Omega)$ now reads $Q(u) = 0$.

In this chapter we shall consider a class of problems that is slightly larger in two respects. First, the linear constraint $Q(u) = 0$ will be replaced by the affine constraint $Q(u) = \gamma$, where $\gamma \in \mathbb{R}$ is a prescribed mean. Secondly, the constraint will be further extended to the case where the solution has a prescribed mean on a subset $\omega \subset \Omega$, which will be denoted as $Q_\omega(u) = \gamma$, where

$$Q_\omega(v) = \frac{1}{|\omega|}\int_\omega v\,dx. \tag{3.9}$$

*For simplicity, we will drop the $Q_\omega$ notation and simply refer to this linear functional as $Q$.*

In this setting, the strong form of the constrained pure Neumann problem reads

$$\text{Find } u \text{ such that } \begin{cases} -\nabla\cdot(a\nabla u) = f, & \text{in } \Omega, \\ \boldsymbol{n}\cdot a\nabla u = g, & \text{on } \partial\Omega, \\ Q(u) = \gamma. \end{cases} \tag{3.10}$$

The standard way to impose constraints is by the introduction of the Lagrangian functional. For $(u,\lambda) \in H^1(\Omega)\times\mathbb{R}$ consider the functional

$$\mathcal{L}(u,\lambda) = J(u) + \lambda(Q(u)-\gamma), \tag{3.11}$$

where $\lambda \in \mathbb{R}$ is the so-called Lagrange multiplier.

The saddle-point formulation of $\mathcal{L}$ over $H^1(\Omega)\times\mathbb{R}$ yields the mixed problem

$$\text{Find } (u,\lambda) \in H^1(\Omega)\times\mathbb{R} \text{ such that } \begin{cases} a(u,v) + \lambda Q(v) = f(v), & \forall v \in H^1(\Omega), \\ \tau Q(u) = \tau\gamma, & \forall \tau \in \mathbb{R}. \end{cases} \tag{3.12}$$

**Remark 5.** *An alternative approach to take the constraint $Q(u) = \gamma$ into account, although not exactly, is to consider a penalized formulation where the goal is to minimize $J(u) + \frac{\beta}{2}(Q(u)-\gamma)^2$ over $H^1(\Omega)$, with $\beta > 0$ a fixed penalization parameter. In that case, the*

*penalization problem reads*

$$\text{Find } u_\beta \in H^1(\Omega) \text{ such that } a(u_\beta, v) + \beta Q(u_\beta)Q(v) = f(v) + \beta\gamma Q(v), \quad \forall v \in H^1(\Omega), \tag{3.13}$$

*where the bilinear form on the left-hand side is coercive over $H^1(\Omega)$ due to the addition of the "mass-term" governed by parameter $\beta$. The penalization problem* (3.13) *is thus well-posed.*

*The corresponding strong form of the problem reads in that case*

$$\text{Find } u_\beta \text{ such that } \begin{cases} -\nabla \cdot (a\nabla u_\beta) + \dfrac{\beta}{|\omega|} Q(u_\beta) = f + \dfrac{\beta}{|\omega|}\gamma, & \text{in } \Omega, \\ \boldsymbol{n} \cdot a\nabla u_\beta = g, & \text{on } \partial\Omega. \end{cases} \tag{3.14}$$

**Remark 6.** *The so-called Augmented Lagrangian method is yet another way of taking the constraint $Q(u) = \gamma$ into account and can be seen as a combination of the Lagrangian and penalization methods. In this method, the mixed problem to be solved is*

$$\text{Find } (u, \lambda) \in H^1(\Omega) \times \mathbb{R} \text{ such that} \\ \begin{cases} a(u, v) + \lambda Q(v) + \beta Q(u)Q(v) = f(v) + \beta\gamma Q(v), & \forall v \in H^1(\Omega), \\ \tau Q(u) = \tau\gamma, & \forall \tau \in \mathbb{R}. \end{cases} \tag{3.15}$$

In order to highlight the performances of the different methods and numerical strategies for solving (3.12), e.g. the Lagrangian or Uzawa methods, we also introduce a class of perturbed problems where the pure Neumann boundary condition in (3.10) is replaced by a Robin boundary condition with an impedance coefficient controlled by a parameter $\varepsilon > 0$, i.e. the weak form of the Robin problem reads

$$\text{Find } (u, \lambda) \in H^1(\Omega) \times \mathbb{R} \text{ such that } \begin{cases} a_\varepsilon(u, v) + \lambda Q(v) = f(v), & \forall v \in H^1(\Omega), \\ \tau Q(u) = \tau\gamma, & \forall \tau \in \mathbb{R}, \end{cases} \tag{3.16}$$

where $a_\varepsilon(u, v) = a(u, v) + \varepsilon \int_{\partial\Omega} uv \, ds$. *For simplicity, we will drop the $a_\varepsilon$ notation and simply refer to this bilinear form by a when the context is clear.*

In this Robin problem, the role of the constraint is not to enforce unicity. Indeed, the mass term on the boundary controlled by $\varepsilon$ provides a coercive bilinear form on $H^1(\Omega)$ so the solution of the unconstrained Robin problem is unique for any fixed $\varepsilon$. The role of the constraint is rather to force the unconstrained solution to move away from the global minimum of the energy functional $J$. For any fixed $\varepsilon$, the solution of the constrained Robin problem (3.16) is unique.

The perturbed problem (3.16) provides a solution $u_\varepsilon$ that converges to the solution $u$ of (3.10) as $\varepsilon$ goes to zero. The fact that we introduce this problem here will become clear when we consider the Uzawa method, which is introduced in Section 3.3.3.

## 3.3 Finite element formulations of constrained problems

In this section, we derive the finite element formulations of the above constrained problems. Whenever relevant, we also highlight the differences between the Neumann and Robin problems. Here, and in the remainder of the chapter, we consider a general conforming finite element space $V_h = \text{span}\,\{\varphi_i\} \subset H^1(\Omega)$, where $\varphi_i$, $i = 1, \ldots, N$ are basis functions of $V_h$. We also assume that the corresponding mesh satisfies the usual regularity properties, see [42, 60].

### 3.3.1 Penalization method

The finite element problem corresponding to the penalization approach (3.13) is given by

$$\text{Find } u_h \in V_h \text{ such that } a(u_h, v_h) + \beta Q(u_h)Q(v_h) = f(v_h) + \beta\gamma Q(v_h), \quad \forall v_h \in V_h. \tag{3.17}$$

The linear system associated with this finite-dimensional problem is of the form

$$(A + \beta BB^T)U = F + \beta\gamma B, \tag{3.18}$$

where $A_{ij} = a(\varphi_j, \varphi_i)$, $B_i = Q(\varphi_i)$, $F_i = f(\varphi_i)$ and the solution vector $U$ collects the degrees of freedom of $u_h$, i.e. $u_h = \sum_{i=1}^{n} U_i\varphi_i$. The rank-one matrix $\beta BB^T$ can be viewed as a correction to the original (unconstrained) stiffness matrix $A$. In the pure Neumann case, the original stiffness matrix $A$ is positive semi-definite, with a rank deficiency of one, while the matrix $A + \beta BB^T$ is positive definite.

Drawbacks of the penalization approach are now briefly recalled. First, the choice of $\beta$ has a strong influence on the quality of the numerical solution. Second, depending on the value of $\beta$, the condition number of the matrix can become very high and adversely affect the accuracy of the approach. Asymptotically, we observed in the considered numerical experiments that $\kappa = \mathcal{O}(\beta)$, where $\kappa$ denotes the scaled condition number of the stiffness matrix based on the $\|\cdot\|_2$ vector norm. Third, due to the addition of the term $BB^T$ in the stiffness matrix, the sparsity of the matrix is lost and the cost of the method increases. Finally, we mention that one can recover an approximation of the Lagrange multiplier $\lambda$ by computing $\beta(B^TU - \gamma)$.

### 3.3.2 Lagrangian method

The mixed finite element problem on $V_h \times \mathbb{R}$ corresponding to the Lagrangian approach (3.12) or (3.16) is given by

$$\text{Find } (u_h, \lambda) \in V_h \times \mathbb{R} \text{ such that } \begin{cases} a(u_h, v_h) + \lambda Q(v_h) = f(v_h), & \forall v_h \in V_h, \\ \tau Q(u_h) = \tau\gamma, & \forall \tau \in \mathbb{R}. \end{cases} \tag{3.19}$$

The linear system associated with this finite-dimensional problem is in this case of the form

$$\begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} U \\ \lambda \end{bmatrix} = \begin{bmatrix} F \\ \gamma \end{bmatrix}. \tag{3.20}$$

The system could be directly solved as given since the augmented matrix is indeed non-singular. However, its size is also larger, which results in higher computational cost, as the constraint is globally applied to the solution. Our goal is nevertheless to decouple the system in order to preserve the efficiency of the PGD approximation solution process. This issue will be addressed in Section 3.4.3.

The presence of the entry zero on the diagonal of the augmented matrix prevents one from uncoupling the solution $U$ from the Lagrange multiplier $\lambda$. The next method aims at circumventing this issue.

### 3.3.3 Uzawa method

The Uzawa method [85, 88] is a numerical strategy aiming at decoupling the constraint from the original problem in (3.20). Two versions of the method are available: the first one, referred to as direct Uzawa, relies on the evaluation of the Schur complement to compute the Lagrange multiplier; the second one, the so-called iterative Uzawa, computes a sequence approximating the Lagrange multiplier within an iterative scheme. However, both methods need for $A$ to be invertible, which is the case for the Robin problem, but not for the Neumann problem. *In the rest of this section, we will thus consider only the Robin problem.*

**Direct Uzawa** Let us develop the linear system of equations for (3.20) as

$$\begin{cases} AU + B\lambda = F, \\ B^T U = \gamma. \end{cases} \tag{3.21}$$

Since $A$ is invertible, one can manipulate the first equation to get $U = A^{-1}(F - B\lambda)$. Then,

using this result in the second equation yields

$$\gamma = B^T\Big(A^{-1}(F - B\lambda)\Big) = B^T A^{-1}F - B^T A^{-1}B\lambda. \tag{3.22}$$

Denoting the Schur complement by $S = B^T A^{-1}B$, one gets

$$S\lambda = B^T A^{-1}F - \gamma. \tag{3.23}$$

In other words, the Lagrangian formulation (3.20) has been recast as

$$\begin{bmatrix} A & B \\ 0 & S \end{bmatrix} \begin{bmatrix} U \\ \lambda \end{bmatrix} = \begin{bmatrix} F \\ B^T A^{-1}F - \gamma \end{bmatrix}, \tag{3.24}$$

where the matrix is now upper triangular: direct Uzawa performs a triangularization by blocks, as a result, the constraint is indeed decoupled from the rest of the problem and the system can be solved by a backward substitution by blocks. However, it still requires one to explicitly invert the stiffness matrix $A$. Iterative Uzawa provides a means to avoid explicitly calculating the inverse $A^{-1}$.

**Iterative Uzawa** In the iterative Uzawa method, the system (3.24), and most particularly the constraint equation, is solved in an iterative manner. The corresponding algorithm, given here in its most simple form using for example the linear descent, is described in Algorithm 1. In this algorithm, the residual circumvents the use of $A^{-1}$ and $S$, indeed

$$r^{(k)} = \gamma - B^T A^{-1}(F - B\lambda^{(k)}) = \gamma - B^T U^{(k)}, \tag{3.25}$$

which corresponds to the constraint residual of the original Lagrangian system (3.21).

**Algorithm 1:** Iterative Uzawa method.

```
Initialize λ^(0), k = 0
while convergence not reached do
    Solve for U^(k): A U^(k) = F − B λ^(k)
    Compute the residual r^(k) = γ − B^T U^(k)
    Compute the step length α^(k)
    Update λ^(k+1) = λ^(k) − α^(k) r^(k)
    k ← k + 1
end
```

The step length $\alpha^{(k)} \in \mathbb{R}$ can be taken as a constant or be evaluated using a gradient approach

to improve the performance of the method. In [85], bounds for the step length are provided in order for the method to converge and optimal step lengths are given, for the case where $A$ is symmetric positive-definite and $B$ is full rank. These bounds and the optimal step length are

$$0 < \alpha^{(k)} < \frac{2}{\lambda_{\max}(S)} \quad \text{and} \quad \alpha_{\text{opt}} = \frac{2}{\lambda_{\min}(S) + \lambda_{\max}(S)}, \tag{3.26}$$

where $\lambda_{\max}(S)$ (resp. $\lambda_{\min}(S)$) denotes the largest (resp. smallest) eigenvalue of the Schur complement $S$. In the present case, since the constraint is scalar, we have that $B$ is a (non-zero) column-vector, and so it has full-rank (equal to one). Moreover, $A$ is positive definite, so that $S$ is a strictly positive scalar and $\lambda_{\max}(S) = \lambda_{\min}(S) = S > 0$. The conditions (3.26) thus reduce to

$$0 < \alpha^{(k)} < \frac{2}{S} \text{ and } \alpha_{\text{opt}} = \frac{1}{S}. \tag{3.27}$$

Note that the optimal step length is not used in practice since it requires the Schur complement $S$. A classical refinement concerning the step length $\alpha^{(k)}$ is to use a gradient descent on the constraint equation (3.23), in which case the step length would be given by

$$\alpha^{(k)} = \frac{r^{(k)} \cdot r^{(k)}}{r^{(k)} \cdot Sr^{(k)}}. \tag{3.28}$$

To avoid the use of $S$, one can write

$$Sr^{(k)} = B^T A^{-1} B r^{(k)} = B^T w^{(k)}, \tag{3.29}$$

where $w^{(k)}$ is the solution of the auxiliary problem $Aw^{(k)} = Br^{(k)}$. Finally, one can use this auxiliary solution $w^{(k)}$ and the step length $\alpha^{(k)}$ to update all variables in the Uzawa algorithm. Indeed, updating $\lambda^{(k+1)} = \lambda^{(k)} - \alpha^{(k)} r^{(k)}$ results in an update of the constraint residual as

$$r^{(k+1)} = \gamma - B^T A^{-1}(F - B\lambda^{(k+1)}) = r^{(k)} - \alpha^{(k)} S r^{(k)} = r^{(k)} - \alpha^{(k)} B^T w^{(k)}, \tag{3.30}$$

and similarly for the solution vector

$$U^{(k+1)} = A^{-1}(F - B\lambda^{(k+1)}) = U^{(k)} + \alpha^{(k)} A^{-1} B r^{(k)} = U^{(k)} + \alpha^{(k)} w^{(k)}. \tag{3.31}$$

In the end, the iterative Uzawa algorithm with gradient descent is described by Algorithm 2. This algorithm has thus eliminated all uses of $A^{-1}$ and $S$.

**Uzawa Adjoint** The constraint considered in this chapter is scalar and so is $r^{(k)}$, as a

**Algorithm 2:** Uzawa method with gradient descent.

Initialize $\lambda^{(0)}$, $k = 0$
Solve for $U^{(0)}$: $AU^{(0)} = F - B\lambda^{(0)}$
Compute the constraint residual $r^{(0)} = \gamma - B^T U^{(0)}$
**while** *convergence not reached* **do**
    Solve for $w^{(k)}$: $Aw^{(k)} = Br^{(k)}$
    Compute the step length $\alpha^{(k)} = \frac{r^{(k)} \cdot r^{(k)}}{r^{(k)} \cdot B^T w^{(k)}}$
    Update $\lambda^{(k+1)} = \lambda^{(k)} - \alpha^{(k)} r^{(k)}$
    Update $r^{(k+1)} = r^{(k)} - \alpha^{(k)} B^T w^{(k)}$
    Update $U^{(k+1)} = U^{(k)} + \alpha^{(k)} w^{(k)}$
    $k \leftarrow k + 1$
**end**

result Algorithm 2 can be further simplified introducing the adjoint problem associated to the constraint functional

$$\text{Find } p \in H^1(\Omega) \text{ such that } a(v, p) = Q(v), \quad \forall v \in H^1(\Omega). \tag{3.32}$$

This problem is well-posed since bilinear form $a$ is coercive (recall we are only considering the Robin problem in this section) and $Q$ is continuous, so that there is a unique solution $p \in H^1(\Omega)$. Note that this approach cannot be applied to the pure Neumann problem since the loading of the adjoint problem $Q$ does not satisfy the compatibility condition (3.2). Now, going back to the mixed-weak formulation arising from the Lagrangian method (3.12) and denoting its solution by $(u_\lambda, \lambda)$, we have

$$\begin{cases} a(u_\lambda, v) + \lambda Q(v) = f(v), & \forall v \in H^1(\Omega), \\ \tau Q(u_\lambda) = \tau\gamma, & \forall \tau \in \mathbb{R}. \end{cases} \tag{3.33}$$

Then, using the adjoint problem $a(v, p) = Q(v)$ we obtain

$$a(u_\lambda, v) + \lambda a(v, p) = f(v), \quad \forall v \in H^1(\Omega). \tag{3.34}$$

Now, making use of the fact that $a$ is bilinear and symmetric yields

$$a(u_\lambda + \lambda p, v) = f(v), \quad \forall v \in H^1(\Omega). \tag{3.35}$$

Finally, the Lax-Milgram theorem applied to the unconstrained Robin problem ensures unic-

ity of the solution so that

$$u_\lambda + \lambda p = u_0, \tag{3.36}$$

where $u_0 \in H^1(\Omega)$ denotes the unconstrained solution. We see that the scalar constraint allows one to simplify the problem, since the Lagrange multiplier $\lambda$ can readily be obtained by applying the functional $Q$ to (3.36) and rearranging the terms, that is

$$\lambda = \frac{Q(u_0) - Q(u_\lambda)}{Q(p)} = \frac{Q(u_0) - \gamma}{Q(p)}, \tag{3.37}$$

where $Q(p) \neq 0$ since $Q(p) = a(p,p)$. As a result, one only needs to compute the unconstrained solution $u_0$, the adjoint solution $p$ and the Lagrange multiplier $\lambda$ to solve the constrained problem. We will subsequently refer to this approach as "Uzawa Adjoint".

### 3.3.4 Augmented Lagrangian method

The mixed finite element problem on $V_h \times \mathbb{R}$ corresponding to the Augmented Lagrangian approach (3.15) is given by

Find $(u_h, \lambda) \in V_h \times \mathbb{R}$ such that

$$\begin{cases} a(u_h, v_h) + \lambda Q(v_h) + \beta Q(u_h)Q(v_h) = f(v_h) + \beta\gamma q(v_h), & \forall v_h \in V_h, \\ \tau Q(u_h) = \tau\gamma, & \forall \tau \in \mathbb{R}. \end{cases} \tag{3.38}$$

The linear system associated with this finite-dimensional problem is

$$\begin{bmatrix} A + \beta BB^T & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} U \\ \lambda \end{bmatrix} = \begin{bmatrix} F + \beta\gamma B \\ \gamma \end{bmatrix}. \tag{3.39}$$

Since the matrix $A + \beta BB^T$ is positive-definite, and thus invertible, one can apply the Uzawa method here to both the Neumann and Robin problems. The direct Uzawa method yields the block triangular system

$$\begin{bmatrix} A + \beta BB^T & B \\ 0 & S_\beta \end{bmatrix} \begin{bmatrix} U \\ \lambda \end{bmatrix} = \begin{bmatrix} F + \beta\gamma B \\ B^T(A + \beta BB^T)^{-1}(F + \beta\gamma B) - \gamma \end{bmatrix}, \tag{3.40}$$

where $S_\beta$ denotes the Schur complement of the perturbed matrix, i.e. $S_\beta = B^T(A + \beta BB^T)^{-1}B$.

The iterative Uzawa scheme is derived mutatis mutandis as the earlier one. However, some simplifications can be made to compute the step length $\alpha^{(k)}$ (see steps 5 and 6 in Algorithm 2) assuming, for instance, that $A + \beta BB^T \approx \beta BB^T$. The auxiliary problem in step 5

then reduces to finding $w^{(k)}$ such that $\beta BB^T w^{(k)} = Br^{(k)}$. Viewing now $B \in \mathbb{R}^N$ as a linear application from $\mathbb{R}$ to $\mathbb{R}^N$, by the rank-nullity theorem, dim (Ker $B$) = dim $(\mathbb{R})$ − rk $(B)$ = $1 - \text{rk}\ (B^T) = 1 - 1 = 0$. As a result, $B$ is injective and $\beta BB^T w^{(k)} = Br^{(k)}$ implies that $\beta B^T w^{(k)} = r^{(k)}$. Then, in step 6 of Algorithm 2, the step length $\alpha^{(k)}$ is approximately $\beta$; in other words, the auxiliary problem has been avoided. As a result, the Augmented Lagrangian method reduces to an Uzawa method on the penalized bilinear form with a constant step length $\beta$, see Algorithm 1.

**Remark 7.** *For the Robin case, in which the matrix $A$ is invertible, one can rely on the Sherman-Morrison-Woodbury matrix identity [55] instead of the approximation $A+\beta BB^T \approx \beta BB^T$. In the present context, the identity states*

$$\left(A + \beta BB^T\right)^{-1} = A^{-1} - \frac{\beta A^{-1} BB^T A^{-1}}{1 + \beta B^T A^{-1} B}, \tag{3.41}$$

*which allows one to write the Schur complement of the perturbed matrix $S_\beta$ in terms of $S$ as*

$$S_\beta = B^T \left( A^{-1} - \frac{\beta A^{-1} BB^T A^{-1}}{1 + \beta B^T A^{-1} B} \right) B = S - \frac{\beta S^2}{1 + \beta S} = \frac{S}{1 + \beta S}. \tag{3.42}$$

*According to Saad [85], the optimal step length $\alpha_{opt}$ is then given by the inverse of the Schur complement: $\frac{1}{S_\beta} = \beta + \frac{1}{S}$. Taking $\beta$ large enough, we obtain $\alpha_{opt} \approx \beta$.*

Similarly to the penalization approach discussed in Section 3.3.1, the choice of the penalization parameter $\beta$ has some influence on the performance of the algorithm. First, the stiffness matrix looses its sparsity pattern resulting in higher computational costs. Secondly, the condition number of the matrix may increase significantly. Finally, choosing $\beta$ too large may introduce round-off errors, which could affect the accuracy of the method, while choosing $\beta$ too small may result in an incorrect step length $\alpha$, leading to an increased number of iterations needed to reach convergence. Note however that the Augmented Lagrangian approach is consistent with the Lagrangian method so that the solution of the former coincides with that of the latter.

## 3.4 PGD formulations of constrained problems

The objective of this section is to apply the above formulations to the Proper Generalized Decomposition (PGD) framework. In order to grasp the essential ingredients of the PGD method, we first introduce the concepts of tensor product of Hilbert spaces, and rank of a tensor, before presenting the formulation arising from the PGD framework. The reader is

referred to [39, 48, 75] for more in-depth analyses of the method and its variants.

### 3.4.1 Introduction and main concepts

**Tensor product of Hilbert spaces.** An alternative approach to approximate the solution $u \in V$ of (2.1) is to construct a reduced order solution $u_m \in V$ using a separated representation involving $m$ terms, or modes. To this end, we make specific assumptions on the Hilbert space $V$. We will assume that the Hilbert space $V$ is a tensor product of Hilbert spaces

$$V = V^{(1)} \otimes V^{(2)} \otimes \cdots \otimes V^{(d)} = \bigotimes_{i=1}^{d} V^{(i)}, \tag{3.43}$$

where $V^{(i)}$ are separable Hilbert spaces and $d$ is the dimension of the problem. Associated to each $V^{(i)}$ we have an inner product $(\cdot,\cdot)_i$ and an associated norm $\|\cdot\|_i$. Using these, we can build a new inner product $(\cdot,\cdot)_V$ and a new norm $\|\cdot\|_V$ on $V$ by tensorization

$$\begin{aligned} (\cdot,\cdot)_V &= \prod_{i=1}^{d} (\cdot,\cdot)_i \,, \\ \|\cdot\|_V &= \prod_{i=1}^{d} \|\cdot\|_i. \end{aligned} \tag{3.44}$$

The tensor product Hilbert space $(V, \|\cdot\|_V)$ is in fact constructed by completion under this inner product; indeed let

$$\underset{a}{\bigotimes_{i=1}^{d}} V^{(i)} = \operatorname{span}\left\{ \otimes_{i=1}^{d} v_i,\ v_i \in V^{(i)},\ \ 1 \le i \le d \right\}. \tag{3.45}$$

This is the so-called algebraic tensor product space. It contains all linear combinations of the $v_i$'s involving a finite number of terms. Then, the tensor product Hilbert space $\bigotimes_{i=1}^{d} V^{(i)}$ is obtained by completion of the algebraic tensor product space $\underset{a}{\bigotimes_{i=1}^{d}} V^{(i)}$ with respect to the inner-product $(\cdot,\cdot)_V$.

For instance, the function $f : (x, y) \mapsto \cos(xy)$ does not belong to the algebraic tensor product space as it does not admit a tensor representation in a finite number of terms. However, it does belong to the tensor product Hilbert space as the limit of a sequence of the algebraic tensor product space, consider e.g. the Fourier series or Taylor expansion of $f$.

**Low-rank tensor subsets.** Let us now introduce a subset of the tensor product Hilbert

space $V$, noted $\mathcal{S}_1$, composed of all rank-1 tensors of $V$

$$\mathcal{S}_1 = \left\{ \otimes_{i=1}^{d} z_i \in V; z_i \in V^{(i)}, i = 1, \ldots, d \right\}. \tag{3.46}$$

This subset is of paramount importance to build low-rank tensor approximations of $u \in V$. We can then define inductively the sets of rank-$m$ tensors $\mathcal{S}_m$ by

$$\mathcal{S}_m = \mathcal{S}_{m-1} + \mathcal{S}_1, \quad \forall m \geq 2. \tag{3.47}$$

Note that $\mathcal{S}_{m-1} \subset \mathcal{S}_m$. However, sets of low-rank tensors do not verify what we may call "nice" properties in linear algebra. For instance, $\mathcal{S}_m$ is not a linear space (the sum of two rank-$m$ tensors may be rank-$2m$). It is not even convex. However, if $z_m \in \mathcal{S}_m$ and $c \in \mathbb{R}$, then $cz_m \in \mathcal{S}_m$, which we may write $c\mathcal{S}_m \subset \mathcal{S}_m$, meaning $\mathcal{S}_m$ is a cone. Also, it can be proved that $\mathcal{S}_1$ is weakly closed in $(V, \|\cdot\|_V)$ [48].

**Tensor rank-1 projection.** The fact that $\mathcal{S}_1$ is weakly closed in $(V, \|\cdot\|_V)$ allows us to define rank-1 projection of any tensor $u \in V$ (which may not be unique) by

$$\Pi(u) = \underset{v \in \mathcal{S}_1}{\operatorname{argmin}} \|u - v\|_V. \tag{3.48}$$

We can thus consider a rank-1 projection of $u$, noted $u_1 \in \Pi(u) \subset \mathcal{S}_1$. Then, we consider again a rank-1 projection of the residual $u - u_1$ given by $e_2 \in \Pi(u - u_1) \subset \mathcal{S}_1$, and form $u_2 = u_1 + e_2 \in \mathcal{S}_2$. By repeating this progressive process we can construct iteratively a sequence $u_m \in \mathcal{S}_m$ that can be shown to converge strongly towards $u$ [48].

A possible way for constructing those rank-1 projections is now detailed. For a given rank-1 tensor $\otimes_{i=1}^{d} z_i \in \mathcal{S}_1$ and a given direction $j \in \{1, \ldots, d\}$, we denote the vector space of "admissible variations" around $\otimes_{i=1}^{d} z_i$ in the $j$-th direction by

$$\begin{aligned} \mathcal{T}^{(j)}(\otimes_{i=1}^{d} z_i) &= \left\{ \otimes_{i=1}^{d} v_i \in \mathcal{S}_1; v_i = z_i, i = 1, \ldots, d, i \neq j \right\}, \\ &= z_1 \otimes z_2 \otimes \cdots \otimes V^{(j)} \otimes \cdots \otimes z_d. \end{aligned} \tag{3.49}$$

Considering now all directions, the tangent space $\mathcal{T}(\otimes_{i=1}^{d} z_i) \subset V$ to the elementary tensor $\otimes_{i=1}^{d} z_i \in \mathcal{S}_1$ is given by

$$\mathcal{T}(\otimes_{i=1}^{d} z_i) = \mathcal{T}^{(1)}(\otimes_{i=1}^{d} z_i) + \mathcal{T}^{(2)}(\otimes_{i=1}^{d} z_i) + \cdots + \mathcal{T}^{(d)}(\otimes_{i=1}^{d} z_i). \tag{3.50}$$

The functions $v \in \mathcal{T}(\otimes_{i=1}^{d} z_i)$ are thus of the form

$$v = \underbrace{v_1 \otimes z_2 \otimes \cdots \otimes z_d}_{\psi_1 \in \mathcal{T}^{(1)}(\otimes_{i=1}^{d} z_i)} + \underbrace{z_1 \otimes v_2 \otimes \cdots \otimes z_d}_{\psi_2 \in \mathcal{T}^{(2)}(\otimes_{i=1}^{d} z_i)} + \cdots + \underbrace{z_1 \otimes z_2 \otimes \cdots \otimes v_d}_{\psi_d \in \mathcal{T}^{(d)}(\otimes_{i=1}^{d} z_i)}, \tag{3.51}$$

where the $v_i$'s are arbitrary functions in $V^{(i)}$.

Starting from a given approximation $u_{m-1} \in \mathcal{S}_{m-1}$, the goal of the progressive PGD method is to compute a rank-one update, or correction, $\delta u = \otimes_{i=1}^{d} z_i \in \mathcal{S}_1$, such that the updated solution

$$u_m = u_{m-1} + \delta u \in \mathcal{S}_m, \tag{3.52}$$

minimizes the updated energy functional $J(u_{m-1} + v)$ over the set of rank-1 tensors. This minimization problem is formulated as

$$\text{Find } \delta u \in \mathcal{S}_1 \text{ such that } J(u_{m-1} + \delta u) = \min_{v \in \mathcal{S}_1} J(u_{m-1} + v). \tag{3.53}$$

The weak formulation associated to this non-linear problem reads

$$\text{Find } \delta u \in \mathcal{S}_1 \text{ such that } a(u_{m-1} + \delta u, v) = f(v), \quad \forall v \in \mathcal{T}(\delta u). \tag{3.54}$$

We note that (3.54) is a necessary condition for (3.53), but it is not sufficient. Indeed, the weak formulation (3.54) arises from the stationarity of the energy functional $J$, but since the set $\mathcal{S}_1$ is not convex, stationarity does not constitute a sufficient condition. In fact, one can even design problems to highlight this non-equivalence, e.g. by considering a problem with cleverly arranged symmetries. In particular, solutions of (3.54) are not unique and include local minima of the energy functional $J$. Nevertheless, since the construction of $u_m$ is by design an iterative process, this possible non-optimality does not constitute a clear-cut flaw of the method.

**Elementary convergence results.** We provide below some straightforward theoretical results about the solution $\delta u$ to (3.53). For a complete proof of the strong convergence of $\{u_m\}$ towards $u$ in terms of the energy norm $\|\cdot\|_{\mathcal{E}}$, the reader is referred to [48]. By (3.53), we immediately obtain

$$J(u_m) = J(u_{m-1} + \delta u) = \min_{v \in \mathcal{S}_1} J(u_{m-1} + v) \leq J(u_{m-1}). \tag{3.55}$$

As a result, the sequence $\{J(u_m)\}$ is decreasing. Being bounded from below by $J(u)$ (the

global minimizer of $J$ over $V$, see (2.3)), the sequence $\{J(u_m)\}$ is convergent. We also have

$$\begin{aligned} J(u_m) &= \frac{1}{2}a(u_m, u_m) - f(u_m), \\ &= \frac{1}{2}a(u_{m-1} + \delta u, u_{m-1} + \delta u) - f(u_{m-1} + \delta u), \\ &= \frac{1}{2}a(u_{m-1}, u_{m-1}) - f(u_{m-1}) + \frac{1}{2}a(\delta u, \delta u) + a(u_{m-1}, \delta u) - f(\delta u), \\ &= J(u_{m-1}) + \frac{1}{2}a(\delta u, \delta u) - a(\delta u, \delta u), \text{ since } \delta u \in \mathcal{T}(\delta u), \\ &= J(u_{m-1}) - \frac{1}{2}a(\delta u, \delta u). \end{aligned} \tag{3.56}$$

Consequently

$$J(u_{m-1}) - J(u_m) = \frac{1}{2}\|\delta u\|_{\mathcal{E}}^2. \tag{3.57}$$

Concerning Galerkin orthogonality, we have the following result

$$a(u - u_m, v) = a(u - (u_{m-1} + \delta u), v) = 0, \quad \forall v \in \mathcal{T}(\delta u). \tag{3.58}$$

Concerning Céa's lemma, we have

$$\|u - (u_{m-1} + \delta u)\|_{\mathcal{E}} \leq \inf_{v \in \mathcal{T}(\delta u)} \|u - (u_{m-1} + v)\|_{\mathcal{E}}. \tag{3.59}$$

In particular, we have

$$\|u - u_m\|_{\mathcal{E}} \leq \|u - u_{m-1}\|_{\mathcal{E}}, \tag{3.60}$$

i.e. the sequence $\{\|u - u_m\|_{\mathcal{E}}\}$ is decreasing. Being bounded from below by zero, the sequence $\{\|u - u_m\|_{\mathcal{E}}\}$ is convergent.

**Alternating directions.** Denoting $\delta u = \otimes_{i=1}^d z_i$, Problem (3.54) is non-linear in the unknowns $z_i$. At the expense of an iterative scheme, we can lift the non-linearity.

Problem (3.54) naturally leads to the set of coupled one-dimensional problems

$$\text{Find } \otimes_{i=1}^d z_i \in \mathcal{S}_1 \text{ such that } \begin{cases} a(\otimes_{i=1}^d z_i, \psi_1) = R_{m-1}(\psi_1), & \forall \psi_1 \in \mathcal{T}^{(1)}(\otimes_{i=1}^d z_i), \\ a(\otimes_{i=1}^d z_i, \psi_2) = R_{m-1}(\psi_2), & \forall \psi_2 \in \mathcal{T}^{(2)}(\otimes_{i=1}^d z_i), \\ \qquad\qquad \vdots & \\ a(\otimes_{i=1}^d z_i, \psi_d) = R_{m-1}(\psi_d), & \forall \psi_d \in \mathcal{T}^{(d)}(\otimes_{i=1}^d z_i), \end{cases} \tag{3.61}$$

where $R_{m-1}$ denotes the residual $R_{m-1}(v) = f(v) - a(u_{m-1}, v)$. Problem (3.61) is still non-

linear.

An approach for solving (3.61) is the so-called Alternating Directions scheme, a fixed-point algorithm in which one successively solves each of the previous equations. To be more precise, each iteration of the Alternating Directions scheme is as follows: from the current iterate $z_1^{(k)}, \ldots, z_d^{(k)}$, compute the new $z_1^{(k+1)}$ using $z_2^{(k)}, \ldots, z_d^{(k)}$ by solving the first equation of (3.61). Then, compute the new $z_2^{(k+1)}$ using the just computed $z_1^{(k+1)}$ and $z_3^{(k)}, \ldots, z_d^{(k)}$ by solving the second equation of (3.61). All the $d$ problems are thus solved until the last one, where we compute $z_d^{(k+1)}$ using the already computed $z_1^{(k+1)}, \ldots, z_{d-1}^{(k+1)}$ by solving the last equation of (3.61). This process repeats $k^\star$ times until the fixed point $\otimes_{i=1}^d z_i^{(k^\star)}$ is (approximately) reached, then we set $u_m = u_{m-1} + \otimes_{i=1}^d z_i^{(k^\star)}$ and repeat the search for the next rank-1 update until convergence.

The name alternating directions scheme stems from the fact that the $V^{(i)}$ are explored successively. Also, it can be viewed as a block Gauss-Seidel method. In practice, for each $1 \leq i \leq d$, we obtain an Ordinary Differential Equation if derivatives with respect to the $i$-th variable are involved in bilinear form $a$ (such as space and time). If no derivatives with respect to the $i$-th variable are involved in bilinear form $a$ (e.g. material parameter, boundary condition), we obtain an algebraic equation. The alternating directions scheme can also be viewed as a pseudo-eigenvalue problem, for which a classical solution is to use a power iteration to capture the dominant singular value.

There exist many different ways to date for constructing a rank-1 correction [39]. The one presented above could be viewed as the Galerkin approach. Others include minimizing the residual [75], where the optimal rank-1 correction is defined as that achieving the minimum of the residual. In practice, it leads to symmetric least-squares problems, but they seem to suffer from bad convergence properties with respect to usual norms. Another way to construct a rank-1 correction is through Minimax PGD, which can be viewed as a Petrov-Galerkin PGD [75]. In this approach, two rank-1 corrections are sought in each iteration: one for the test space, one for the trial space. Finally, instead of computing the modes successively, it is also possible to compute several modes simultaneously at the expense of higher computational costs, leading to the so-called simultaneous version of the PGD method [75]. In this case, the decomposition is usually of greater quality, since the minimization involved when computing one mode after the other is a greedy process. The analogy with the power iteration is in that case replaced by a subspace iteration process.

Before investigating the constrained approaches, we propose to examine some properties of the linear systems arising from discretization and alternating directions when applied to the unconstrained pure Neumann problem, as detailed below. For simplicity, we consider $d = 2$,

that is $\Omega = \Omega_1 \times \Omega_2 \subset \mathbb{R}^2$, with a diffusivity constant $a$ equal to unity throughout $\Omega$, and $m = 1$, i.e. we want to compute a rank-one solution $u_1 = z_1 \otimes z_2$. In this case, the problem of finding a rank-1 approximation of the solution to problem (3.3) simplifies to

$$\text{Find } (z_1, z_2) \in V_h^{(1)} \times V_h^{(2)} \text{ such that } \begin{cases} a(z_1 \otimes z_2, v_1 \otimes z_2) = f(v_1 \otimes z_2), & \forall v_1 \in V_h^{(1)}, \\ a(z_1 \otimes z_2, z_1 \otimes v_2) = f(z_1 \otimes v_2), & \forall v_2 \in V_h^{(2)}. \end{cases} \tag{3.62}$$

Recalling the definition of the bilinear form $a$, at a given iteration of the Alternating Directions scheme, the first equation of (3.62) is found to be

$$\begin{aligned} &\text{Given } z_2 \in V_h^{(2)}, \text{ find } z_1 \in V_h^{(1)} \text{ such that} \\ &\|z_2\|_{L^2(\Omega_2)}^2 \left( \int_{\Omega_1} z_1' v_1' \, dx \right) + |z_2|_{H^1(\Omega_2)}^2 \left( \int_{\Omega_1} z_1 v_1 \, dx \right) \quad = f(v_1 \otimes z_2), \quad \forall v_1 \in V_h^{(1)}, \end{aligned} \tag{3.63}$$

where $\|z_2\|_{L^2(\Omega_2)}^2$ and $|z_2|_{H^1(\Omega_2)}^2$ in the left-hand side appear as known constant coefficients in front of what yield essentially a stiffness matrix and a mass matrix, respectively.

From (3.63), we can already make some simple observations about system (3.62). First, in the case where $|z_2|_{H^1(\Omega_2)} \neq 0$, i.e. $z_2$ is not constant over $\Omega_2$, then the matrix arising from the FE formulation of problem (3.63) over $\Omega_1$ will be positive-definite, even though the solution of the original Neumann problem was only defined up to a constant. As a result, the PGD sets for itself an additive constant during the process. However, this value depends, among other things, on the initialization of the fixed-point algorithm and it is not clear how or even if it can be controlled.

Second, if $|z_2|_{H^1(\Omega_2)} \approx 0$, i.e. $z_2$ is almost constant over $\Omega_2$ or the variations of $z_2$ are small compared to its magnitude, then the matrix arising from the FE formulation of the problem over $\Omega_1$ will be ill-conditioned, or sometimes non-invertible with a rank deficiency of one (just like the matrix $A$ from Section 3.3 was). This is a degenerate case of PGD we came across, hence the motivation for the methodology developed in this chapter. Nevertheless, it is worth mentioning that even in that case, the compatibility condition is inherited from the original problem so that solutions do exist.

We now assume that the input data admit affine representations [27], meaning that the diffusivity coefficient $a$, as well as the loadings $f$ and $g$, admit exact separated representations. As a consequence, the bilinear form $a$ and linear form $f$ can be separated accordingly. In addition, we also require the subset $\omega$ (and, consequently, the linear functional $Q$ as well) to admit a separated representation. For $\omega$, this means that the domain can be written as a (possibly non-disjoint) union of $d$-dimensional hyper-rectangles, while for $Q$ this means it

can be written in tensor form. The reader is referred to [96] for the case where the input data is not separable.

### 3.4.2 Penalization method

Since the penalization approach is nothing but an ad hoc stabilization of the bilinear form together with a consistent correction of the right-hand side, the PGD formulation of this problem can readily be established: given a previously computed numerical approximation $u_{m-1}$,

$$\begin{aligned} &\text{Find } \delta u \in \mathcal{S}_1 \text{ such that} \\ &\qquad a(u_{m-1}+\delta u, v) + \beta Q(u_{m-1}+\delta u)Q(v) = f(v) + \beta\gamma Q(v), \quad \forall v \in \mathcal{T}(\delta u). \end{aligned} \tag{3.64}$$

Rearranging the terms in the equation, the problem can be recast as

$$\begin{aligned} &\text{Find } \delta u \in \mathcal{S}_1 \text{ such that} \\ &\qquad a(\delta u, v) + \beta Q(\delta u)Q(v) = R_{m-1}(v) + \beta\Big(\gamma - Q(u_{m-1})\Big)Q(v), \quad \forall v \in \mathcal{T}(\delta u). \end{aligned} \tag{3.65}$$

With the assumed separation of the input data and $Q$, this leads to a problem that possesses the same structure as problem (3.61), namely

$$\begin{aligned} &\text{Find } \delta u \in \mathcal{S}_1 \text{ such that} \\ &\begin{cases} a(\delta u, \psi_1) + \beta Q(\delta u)Q(\psi_1) = R_{m-1}(\psi_1) + \beta\Big(\gamma - Q(u_{m-1})\Big)Q(\psi_1), & \forall \psi_1 \in \mathcal{T}^{(1)}(\delta u), \\ a(\delta u, \psi_2) + \beta Q(\delta u)Q(\psi_2) = R_{m-1}(\psi_2) + \beta\Big(\gamma - Q(u_{m-1})\Big)Q(\psi_2), & \forall \psi_2 \in \mathcal{T}^{(2)}(\delta u), \\ \qquad\qquad\qquad\qquad \vdots & \\ a(\delta u, \psi_d) + \beta Q(\delta u)Q(\psi_d) = R_{m-1}(\psi_d) + \beta\Big(\gamma - Q(u_{m-1})\Big)Q(\psi_d), & \forall \psi_d \in \mathcal{T}^{(d)}(\delta u). \end{cases} \end{aligned} \tag{3.66}$$

This system is solved in an Alternating Directions manner until convergence of the new mode $\delta u = \otimes_{i=1}^{d} z_i$, after which one can set $u_m = u_{m-1} + \otimes_{i=1}^{d} z_i$.

Following [48], and under the assumption of weak closedness therein, the penalized PGD converges towards the penalized FEM, at least in the norm induced by the penalized bilinear form.

### 3.4.3 Lagrangian method

Once again, using a progressive approach, we assume that $u_{m-1}$ is given and we seek for a next mode $\delta u$ and Lagrange multiplier $\lambda$ satisfying

$$\text{Find } (\delta u, \lambda) \in \mathcal{S}_1 \times \mathbb{R} \text{ such that } \begin{cases} a(u_{m-1} + \delta u, v) + \lambda Q(v) = f(v), & \forall v \in \mathcal{T}(\delta u), \\ \tau Q(u_{m-1} + \delta u) = \tau\gamma, & \forall \tau \in \mathbb{R}. \end{cases} \tag{3.67}$$

Note that in this work we have not studied the existence nor the unicity of the solution $(\delta u, \lambda)$ and are only concerned with finding critical points of the Lagrangian functional.

Rearranging the terms, we have

$$\text{Find } (\delta u, \lambda) \in \mathcal{S}_1 \times \mathbb{R} \text{ such that } \begin{cases} a(\delta u, v) + \lambda Q(v) = R_{m-1}(v), & \forall v \in \mathcal{T}(\delta u), \\ \tau Q(\delta u) = \tau\Big(\gamma - Q(u_{m-1})\Big), & \forall \tau \in \mathbb{R}, \end{cases} \tag{3.68}$$

leading to the following problem

$$\text{Find } (\delta u, \lambda) \in \mathcal{S}_1 \times \mathbb{R} \text{ such that } \begin{cases} a(\delta u, \psi_1) + \lambda Q(\psi_1) = R_{m-1}(\psi_1), & \forall \psi_1 \in \mathcal{T}^{(1)}(\delta u), \\ a(\delta u, \psi_2) + \lambda Q(\psi_2) = R_{m-1}(\psi_2), & \forall \psi_2 \in \mathcal{T}^{(2)}(\delta u), \\ \vdots & \\ a(\delta u, \psi_d) + \lambda Q(\psi_d) = R_{m-1}(\psi_d), & \forall \psi_d \in \mathcal{T}^{(d)}(\delta u), \\ \tau Q(\delta u) = \tau\Big(\gamma - Q(u_{m-1})\Big), & \forall \tau \in \mathbb{R}. \end{cases} \tag{3.69}$$

It is interesting to observe that problem (3.69) has a structure that is clearly different from that of (3.61) or (3.66) due to the constraint equation and the Lagrange multiplier $\lambda$. In the spirit of the Alternating Directions scheme, one would be tempted to associate the constraint with one (or more) of the $d$ other equations and perform the Alternating Directions as usual until convergence. However, this approach raises several questions: does the method converge? Does the choice of the coupling have an influence on convergence?

From our preliminary experiments, it turns out that none of these approaches yield satisfactory results. To be more specific, the constraint $Q(u_{m-1} + \delta u) = \gamma$ is satisfied, but as we increase the number of modes $m$, the PGD solution does not converge towards the FE solution of the original Lagrangian problem (3.19). Furthermore, depending on the choice adopted when coupling the constraint with one of the $d$ problems, we obtain different results.

Finally, the Lagrange multiplier $\lambda$ does not converge either.

In order to solve (3.69) efficiently, it appears there are two options: either associate the constraint with all $d$ problems, or fully decouple the constraint from the other equations, as the Uzawa and Augmented Lagrangian methods do. In the rest of this section, we will investigate the first option: the goal is to solve (3.69) in such a way that the constraint equation is evenly associated to the other problems.

The approach considered in this section differs from the classical Alternating Directions scheme in that we update simultaneously all the $z_i^{(k+1)}, i = 1, \ldots, d$ using the previous iterates $z_j^{(k)}, j = 1, \ldots, d, j \neq i$. If one were to make an analogy between the classical Alternating Directions scheme and the block Gauss-Seidel method, then the present approach could be viewed as a block Jacobi method. Since we update each function $z_i^{(k+1)}$ using only information that is already available from the previous iterate $k$, one could do this process in parallel, but this is not the approach we take. Instead, we assemble a global block-diagonal system whose unknown is the concatenation of all the $z_i^{(k+1)}, i = 1, \ldots, d$. This way, we have a system where all the $d$ functions are updated simultaneously. Finally, we incorporate the constraint into this global system, after having linearized it around the current iterate $\delta u^{(k)} = \otimes_{i=1}^{d} z_i^{(k)}$, e.g. by the Newton method. Therefore, instead of

$$Q\Big( \otimes_{i=1}^{d} z_i^{(k+1)} \Big) = \gamma - Q(u_{m-1}), \tag{3.70}$$

we consider

$$\begin{aligned} Q\Big( z_1^{(k+1)} \otimes z_2^{(k)} \otimes \cdots \otimes z_d^{(k)} + z_1^{(k)} \otimes z_2^{(k+1)} \otimes \cdots \otimes z_d^{(k)} + \cdots + z_1^{(k)} \otimes z_2^{(k)} \otimes \cdots \otimes z_d^{(k+1)} \Big) \\ = \gamma - Q\Big( u_{m-1} + (1-d) \otimes_{i=1}^{d} z_i^{(k)} \Big). \end{aligned} \tag{3.71}$$

In the end, each fixed-point iteration consists in solving a linear system where all $d$ functions are updated simultaneously and in which the constraint couples the one-dimensional problems.

The finite element counterpart yields the following system of equations, where $z_i^{(k+1)}$ defines

the vectors of unknown in each direction $i$

$$\begin{bmatrix} A^{(1,k)} & 0 & \dots & 0 & B^{(1,k)} \\ 0 & A^{(2,k)} & \dots & 0 & B^{(2,k)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & A^{(d,k)} & B^{(d,k)} \\ B^{(1,k)^T} & B^{(2,k)^T} & \dots & B^{(d,k)^T} & 0 \end{bmatrix} \begin{bmatrix} z_1^{(k+1)} \\ z_2^{(k+1)} \\ \vdots \\ z_d^{(k+1)} \\ \lambda \end{bmatrix} = \begin{bmatrix} F^{(1,k)} \\ F^{(2,k)} \\ \vdots \\ F^{(d,k)} \\ \tilde{\gamma} \end{bmatrix}, \tag{3.72}$$

with, for each $i = 1, \dots, d$

$$\begin{cases} A_{ln}^{(i,k)} = a(z_1^{(k)} \otimes z_2^{(k)} \otimes \cdots \otimes \varphi_n^{(i)} \otimes \cdots \otimes z_d^{(k)}, z_1^{(k)} \otimes z_2^{(k)} \otimes \cdots \otimes \varphi_l^{(i)} \otimes \cdots \otimes z_d^{(k)}), \\ B_l^{(i,k)} = Q(z_1^{(k)} \otimes z_2^{(k)} \otimes \cdots \otimes \varphi_l^{(i)} \otimes \cdots \otimes z_d^{(k)}), \\ F_l^{(i,k)} = R_{m-1}(z_1^{(k)} \otimes z_2^{(k)} \otimes \cdots \otimes \varphi_l^{(i)} \otimes \cdots \otimes z_d^{(k)}), \end{cases} \tag{3.73}$$

where $V_h^{(i)} = \text{ span} \left\{ \varphi_j^{(i)} \right\}_{j=1}^{\dim V_h^{(i)}}$, and $\tilde{\gamma} = \gamma - Q\Big(u_{m-1} + (1-d) \otimes_{i=1}^{d} z_i^{(k)}\Big)$.

It is worth mentioning here that the method needs help with convergence. We noticed in our numerical experiments that starting each fixed-point problem with a few iterations using the penalized version PGD (or the forthcoming Uzawa or Augmented Lagrangian methods) provides a remedy to this issue. It is likely because these few iterations bring the iterates closer to the attraction basin of the solution.

### 3.4.4 Uzawa method

In this section and the next, we show how system (3.69) can be solved by decoupling the constraint from the other equations. In Section 3.3.3, we described three versions for the Uzawa method: Algorithm 1, where the step length $\alpha$ was not elaborated upon (one can take $\alpha$ constant for a first grab at the method, note however that it cannot be too large, otherwise the method will not converge, according to [85]); Algorithm 2, where the step length $\alpha$ was computed using a gradient algorithm, which required the solution of an auxiliary problem (step 5); and the so-called Uzawa Adjoint method. We will now adapt those to the PGD setting.

We start by mentioning that the linearization of the constraint equation in (3.71) is consistent with (3.25).

As far as Algorithm 1 is concerned, step 3 has to be adapted to the PGD framework. For simplicity, this step can be written in weak form, leading to, for an already computed ap-

proximate value of $\lambda$

$$\text{Find } \delta u \in \mathcal{S}_1 \text{ such that } a(\delta u, v) = R_{m-1}(v) - \lambda Q(v), \quad \forall v \in \mathcal{T}(\delta u). \tag{3.74}$$

Since $\lambda$ is (approximately) known at this stage, this problem is of the same form as the problem arising from classical PGD, and is solved using the Alternating Directions scheme.

Then, several choices are available in order to update the Lagrange multiplier $\lambda$ (step 6): it could be updated after reaching the fixed point satisfying (3.74), i.e. for the current approximation of $\lambda$, or more frequently, for example after one full Alternating Directions iteration, or even more frequently. that is after each problem in the Alternating Directions scheme. In the present work, we did not study in detail the influence of this feature and simply chose to update after one full Alternating Directions iteration.

Concerning Algorithm 2, some more in-depth modifications need to be made since we cannot afford to solve the auxiliary problem $Aw^{(k)} = Br^{(k)}$, which lives in the fully discretized space (in fact, this approach would be the so-called Uzawa Adjoint and will be investigated at the end of this subsection). Instead, we proceed as follows: for the current approximate value of the Lagrange multiplier $\lambda$ we perform one full Alternating Directions iteration on (3.74) and we compute the constraint residual $r^{(k)}$. For the step length $\alpha$, the analysis of [85] is again required because we are no longer working with system (3.19), but with (3.72) instead, having uncoupled the $d$ problems by the Alternating Directions scheme and linearized the constraint equation by Newton's method. As a consequence, the Schur complement is now the sum of $d$ "unidimensional Schur complements"

$$S = \sum_{i=1}^{d} S_i, \text{ where } S_i = B^{(i,k)^T} \left(A^{(i,k)}\right)^{-1} B^{(i,k)}, \tag{3.75}$$

where these vectors and matrices were defined in (3.73). Note that each matrix $A^{(i,k)}$ is associated with a one-dimensional problem in $V_h^{(i)}$, $i = 1, \ldots, d$.

Then, the optimal step length is given by $\alpha_{\text{opt}} = \frac{1}{S} = \frac{1}{\sum_{i=1}^{d} S_i}$, and to approximate it, one can use a gradient descent on the constraint equation, in which the step length would be given by

$$\alpha = \frac{r^{(k)} \cdot r^{(k)}}{r^{(k)} \cdot S r^{(k)}}. \tag{3.76}$$

To avoid the use of $S$, one can write

$$Sr^{(k)} = \sum_{i=1}^{d} B^{(i,k)^T} \left(A^{(i,k)}\right)^{-1} B^{(i,k)} r^{(k)} = \sum_{i=1}^{d} B^{(i,k)^T} w_i, \tag{3.77}$$

where each $w_i$ is the solution of an auxiliary one-dimensional problem $A^{(i,k)} w_i = B^{(i,k)} r^{(k)}$.

Finally, the extension of the Uzawa Adjoint method to the PGD setting is relatively straightforward, since one only needs to compute the unconstrained and adjoint solutions as well as the Lagrange multiplier using (3.37). One iteration of the PGD version of this algorithm is as follows: compute one mode $\delta u$ for the unconstrained solution, one mode $\delta p$ for the adjoint solution, and update the approximate Lagrange multiplier

$$\lambda = \frac{Q(u_{m-1} + \delta u) - \gamma}{Q(p_{m-1} + \delta p)}. \tag{3.78}$$

### 3.4.5 Augmented Lagrangian method

The Augmented Lagrangian method is essentially the same as the Uzawa method but with the bilinear form replaced by its penalized version. We only state in this subsection the simplifications associated with the step length $\alpha$. Similarly to what was derived in Section 3.3.4, when $\beta$ is large enough, the auxiliary problems can be circumvented assuming $\beta B^{(i,k)^T} w_i = r^{(k)}$. The step length is then computed as

$$\alpha = \frac{r^{(k)} \cdot r^{(k)}}{r^{(k)} \cdot \sum_{i=1}^{d} \frac{r^{(k)}}{\beta}} = \frac{\beta}{d}, \tag{3.79}$$

instead of (3.76)–(3.77). Here again, the auxiliary problems have been avoided. Note that the Sherman-Morrison-Woodbury matrix identity applied to each (penalized) term in the sum (3.75) yields the same result.

## 3.5 Numerical Examples

In this section, we apply above methods to the two model problems considered in this chapter, namely the constrained pure Neumann problem (3.12) and the constrained Robin problem (3.16).

For the numerical simulations, we consider $d = 2$ and $\Omega = \omega = (0,1)^2$ and choose a point $(x_c, y_c) \in \Omega$ so that $\Omega$ is split into two regions: $\Omega_1 = \{(x,y) \in \Omega; x > x_c \text{ and } y > y_c\}$, and the complementary region $\Omega_0 = \Omega \backslash \Omega_1$. Then $a$ is chosen piecewise constant in each $\Omega_i, i =$

$0, 1$. We choose $x_c = 7/32$, $y_c = 19/32$, $a_0 = 1$ and $a_1 = 10$. Finally, we take $\gamma = 0$. The exact solution $u$ of the constrained pure Neumann problem (3.10) is constructed using the so-called manufactured solution method, and is chosen to be harmonic of the form

$$u(r,\theta) = \begin{cases} A_0 r^\mu \cos(\mu\theta) + B_0 r^\mu \sin(\mu\theta) + C, & \text{in } \Omega_0, \\ A_1 r^\mu \cos(\mu\theta) + B_1 r^\mu \sin(\mu\theta) + C, & \text{in } \Omega_1, \end{cases} \tag{3.80}$$

where $(r, \theta)$ is the polar coordinate centered at $(x_c, y_c)$. The constants $\mu, A_0, B_0, A_1$, and $B_1$ are chosen such that $u$ is continuous in $\Omega$ and $\boldsymbol{n} \cdot (a\nabla u)$ is continuous across the interface between $\Omega_0$ and $\Omega_1$. Finally, $C$ is chosen so that $u$ satisfies the constraint $Q(u) = 0$. We mention that $\mu$ is taken greater than the degree of the shape functions considered in the numerical experiments so that the manufactured solution has sufficient regularity. The loadings $f$ and $g$ are derived using (3.1). In fact $f = 0$ since $u$ is taken to be harmonic in $\Omega$. Table 3.1 collects the values of the constant parameters $\mu, A_0, B_0$, and $C$ while we have $A_1 = A_0$ and $B_1 = (a_0/a_1)B_0$.

Table 3.1 Values of the parameters $\mu, A_0, B_0$, and $C$ used in the numerical experiments.

| $\mu$ | $A_0$ | $B_0$ | $C$ |
|---|---|---|---|
| 2.7317 | 0.1526 | 0.9883 | 0.0534 |

In order to compare the numerical solutions, we will use the semi-norm induced by $a$, denoted by $|\cdot|_{\mathcal{E}}$, and the mean-functional $Q$. Concerning the Robin problem (3.16), we consider the same loadings $f$ and $g$ as for the pure Neumann problem (3.12), independently of $\varepsilon$. As a result, the exact solution $u_\varepsilon$ of the Robin problem (3.16) is unknown, but this is not the focus of the present study (it was verified though that for every value of $\varepsilon$, the unconstrained finite element solution of the Robin problem did not already satisfy the constraint). Furthermore, we will also use the semi-norm induced by $a$ and the mean-functional $Q$ for the Robin problem. Finally, a regular mesh of square elements with associated mesh size $h = 1/32$ is used, and the bilinear Lagrange polynomials are chosen as basis functions.

First, we start by illustrating some properties of the Robin problem when the impedance coefficient $\varepsilon$ goes to zero within the FE framework described in Section 3.3. Afterwards, we present some results for the penalized FEM and compare the Neumann and Robin problems. Finally, we present results for the constrained PGD approaches introduced in Section 3.4.

### 3.5.1 Constrained FEM solutions

As stated in Section 3.2, for any $\varepsilon > 0$, the Schur complement $S$ for the Robin problem exists and is finite. However, as Figure 3.1 shows, when $\varepsilon$ goes to zero, $S^{-1}$ goes to zero as well, with a slope of one. On the same figure, we also show:

- the algebraic error $|u_{\varepsilon,h} - u_h|_{\mathcal{E}}$ between the FE solutions of the Robin problem and the Neumann problem, both obtained by the Lagrangian method (3.20); as $\varepsilon$ goes to zero, Figure 3.1 shows that the constrained Robin solution converges towards the constrained Neumann solution, at least in terms of the semi-norm $|\cdot|_{\mathcal{E}}$;
- the absolute value of the mean-value of the Robin solution $|Q(u_{\varepsilon,h})|$; Figure 3.1 shows that the constraint is numerically enforced for all values of $\varepsilon$, so that, together with the previous point, $u_{\varepsilon,h}$ does converge towards $u_h$;
- the absolute value of the Lagrange multiplier $|\lambda|$; Figure 3.1 shows that it has the same behavior as $\varepsilon$. From our numerical experiments, there could be two reasons (or a combination of both): as $\varepsilon$ goes to zero, either the unconstrained solution progressively satisfies the zero-mean condition, or the matrix $A$ becomes more and more numerically singular;
- the inverse of the scaled condition number of matrix $A$, denoted by $\kappa^{-1}$; Figure 3.1 shows that, as $\varepsilon$ goes to zero, the matrix becomes numerically singular, reflecting that the underlying bilinear form progressively looses coercivity;
- the absolute value of the mean-value of the unconstrained Robin solution $|Q(\tilde{u}_{\varepsilon,h})|$; Figure 3.1 shows that it is bounded away from zero for any $\varepsilon$, and so we can conclude that $|\lambda|$ goes to zero because of the lost coercivity.

Following (3.27), in order for the Uzawa method to converge, the upper bound for the step length $\alpha$ is given by $\frac{2}{S}$, which is of the order of $\varepsilon$. Therefore, when $\varepsilon$ goes to zero, the convergence of the Uzawa method deteriorates. At the limit $\varepsilon = 0$, which is the pure Neumann case, $S$ does not exist and one cannot use the Uzawa method.

We now consider the penalized FE methods and compare the Robin problem (with fixed impedance parameter $\varepsilon = 1$) and the Neumann problem when the penalization coefficient $\beta$ varies. The results are collected in Figure 3.2, where we show:

- the algebraic error $|u_{\beta,h} - u_h|_{\mathcal{E}}$ between the penalized and Lagrangian solutions;
- the absolute value of the mean-value $|Q(u_{\beta,h})|$;
- the error in the post-processed Lagrange multiplier $|\beta Q(u_{\beta,h}) - \lambda|$;

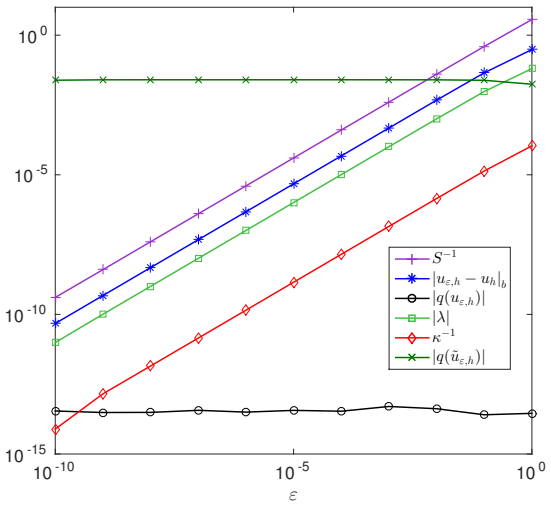


Figure 3.1 Evolution of some outputs of the Robin problem solved by the Lagrangian method with respect to the impedance coefficient $\varepsilon$.

- the inverse of the scaled condition number of the penalized matrix $\kappa^{-1}$;
- for the Robin problem only, the algebraic error between the penalized solution and the unconstrained solution $|u_{\beta,h} - \tilde{u}_h|_{\mathcal{E}}$.

We observe strikingly different results between the two problems. For the Robin case, the choice of $\beta$ has evidently a strong influence on the solution: it has to be sufficiently large to enforce the constraint. This is because the unconstrained Robin problem already possesses a unique solution, and the penalization method is nothing but a trade-off between this unconstrained solution and the solution of the Lagrangian method. This is in contrast with the Neumann problem, for which the unconstrained solution is not unique. As a result, there is no trade-off where the energy would have to be sacrificed in favor of the constraint. To some extent, as $A$ is singular and thus not coercive, any $\beta > 0$ is large enough to impose the constraint so that the penalized solution coincides with the Lagrangian solution. However, for small values of the parameter $\beta$, the constraint fails to be enforced, due to the fact that the penalized matrix $A$ becomes more and more singular as $\beta$ goes to zero, so that the numerical solutions get polluted by round-off errors.

Based on the results collected in Figure 3.2, we will now set $\beta = 10^2$ for the penalization and Augmented Lagrangian approaches. We purposely take $\beta$ not too large in order to observe the limitations of the penalization approach. We emphasize here that the penalization parameter $\beta$ in the Augmented Lagrangian approach has to be chosen large enough to ensure

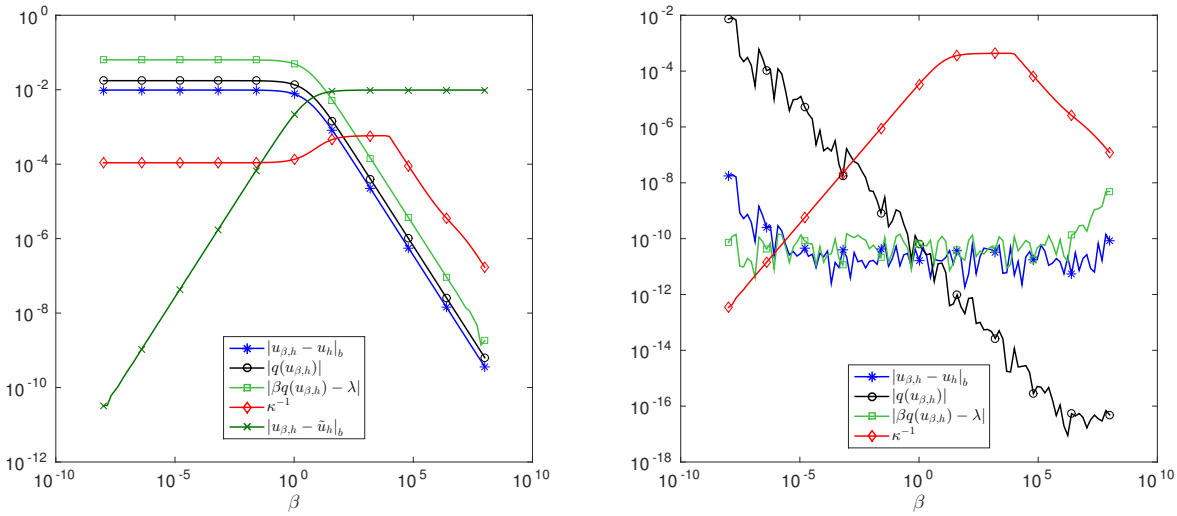


Figure 3.2 Evolution of some outputs of the penalization approach with respect to the penalization parameter $\beta$. Left: Robin problem. Right: Neumann problem.

the precision of the step length $\alpha$ and, thus, the fast convergence of the solution, but not too large to avoid that round-off errors pollute the numerical solution.

### 3.5.2 Constrained PGD solutions

We now turn our attention to the constrained PGD solutions. In Section 3.4, we introduced five different methods, namely the penalization method; the Lagrangian method with simultaneous update of the functions by a block Jacobi method; the iterative Uzawa and the Uzawa Adjoint methods (recall that they can only be applied to the Robin case where the Schur complement exists); and the Augmented Lagrangian method. For completeness, we also consider the classical (unconstrained) PGD method, which can equivalently be seen as a penalization method with $\beta = 0$. We mention that in these experiments, the step length $\alpha$ for the iterative Uzawa method was set to unity. We will subsequently analyze the influence of this parameter on the convergence of the method. We recall that for the iterative Uzawa and Augmented Lagrangian methods, the Lagrange multiplier $\lambda$ is updated after each Alternating Direction iteration. Finally, the PGD algorithms were initialized with $\lambda = 1$ and with random modes.

For each method and each problem, we measure the truncation error in the semi-norm induced by $a$ between the FE solution of the Lagrangian problem and the PGD solutions, as displayed in Figure 3.3, as well as the absolute value of the mean-value, as displayed in Figure 3.4, and

the error in the Lagrange multiplier, as displayed in Figure 3.5.

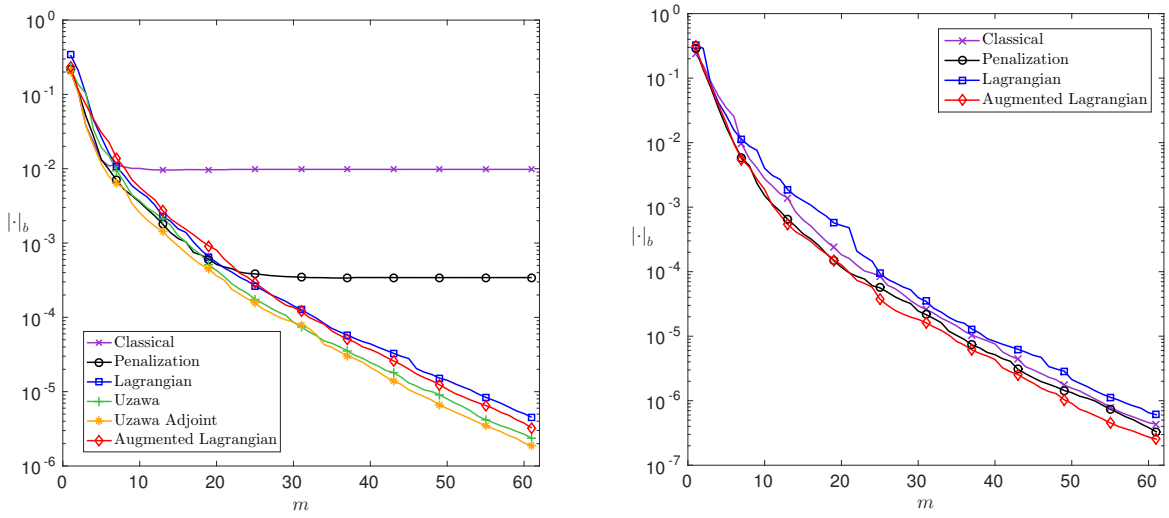


Figure 3.3 Truncation error between the FE Lagrangian solution and the constrained PGD methods. Left: Robin problem. Right: Neumann problem.

In terms of the semi-norm induced by $a$, all PGD solutions converge towards the FE solution of the Lagrangian problem, except for the unconstrained and the penalized approaches for the Robin case, as expected. Concerning the mean-value, all constrained PGD solutions tend to satisfy the constraint as the number of modes increases, except for the penalized Robin case, because of the aforementioned trade-off between energy and constraint satisfaction. We note that for the Uzawa Adjoint method, the constraint is satisfied up to machine precision, but this is only a consequence of the way the Lagrange multiplier is computed for this approach, i.e. following (3.78). Finally, the Lagrange multiplier also converges with the number of modes except for the penalized Robin case.

To summarize, the PGD methods based on the Lagrangian formulation (including the Uzawa and Augmented Lagrangian approaches) converge towards the FE Lagrangian solution. The PGD method based on the penalization formulation converges towards its penalized FE solution counterpart (not directly shown in these figures).

We now investigate the influence of the step length $\alpha$ and of the impedance coefficient $\varepsilon$ on the performance of the constrained PGD solved by the Uzawa method, again applied to the Robin problem. The results are collected in Table 3.2, where we measure the number of modes required to achieve a truncation error in the energy norm between the constrained PGD solution and the FE solution of the full Lagrangian problem (3.19) smaller than $10^{-2}$.

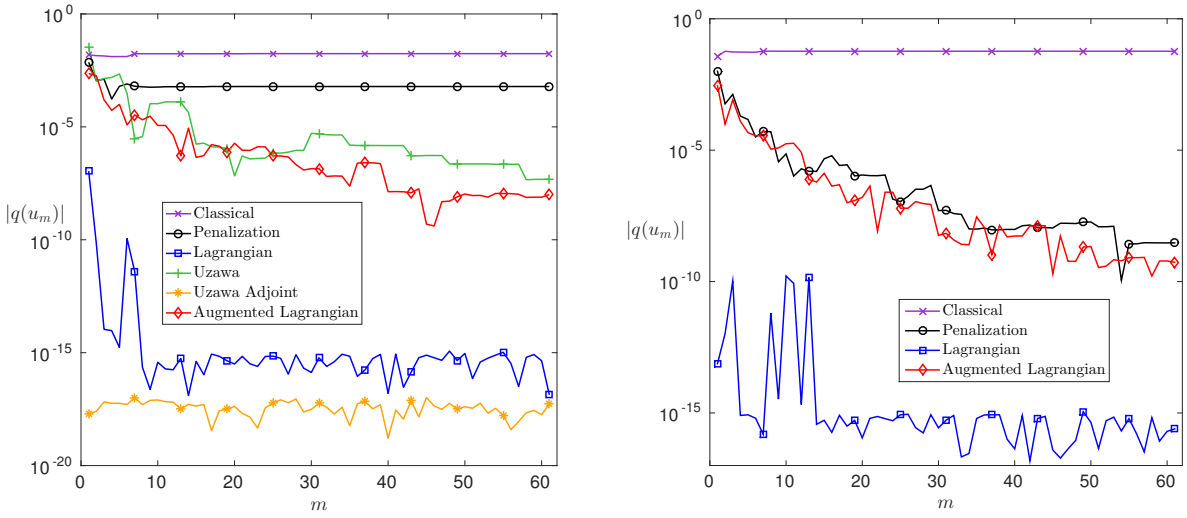


Figure 3.4 Absolute value of the mean-value. Left: Robin problem. Right: Neumann problem.

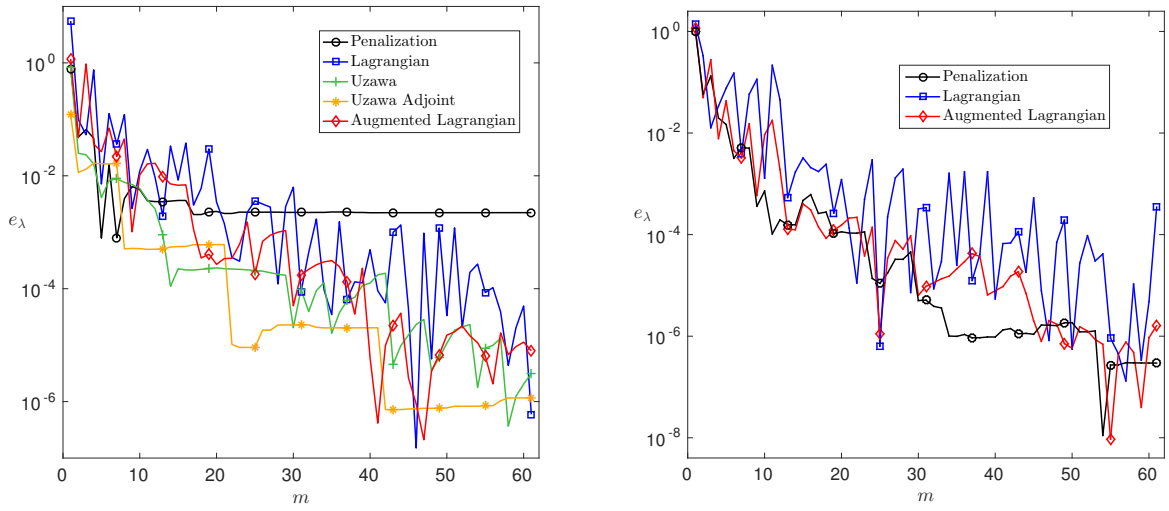


Figure 3.5 Error in the Lagrange multiplier. Left: Robin problem. Right: Neumann problem.

For several values of the impedance coefficient $\varepsilon$ (and thus several values of the Schur complement $S$) we used different constant values of $\alpha$ as well as the step length $\alpha_{\text{opt}}$ computed solving the auxiliary problems resulting from (3.76)–(3.77).

Table 3.2 Number of modes needed to achieve a truncation error in the energy norm smaller than $10^{-2}$ as a function of the step length $\alpha$ and impedance coefficient $\varepsilon$. The "$\times$" notation is employed to mean the numerical solutions failed to converge (unbounded energy norm). The "$\circ$" notation is employed to mean the numerical solution had bounded energy norm but did not converge to the FE solution of the full Lagrangian problem (3.19).

| $\alpha$ \ $\varepsilon$ | $10^{-8}$ | $10^{-7}$ | $10^{-6}$ | $10^{-5}$ | $10^{-4}$ | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | 1 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | $\times$ | $\times$ | $\times$ | $\times$ | $\times$ | $\times$ | $\times$ | $\times$ | 7 |
| $10^{-1}$ | $\times$ | $\times$ | $\times$ | $\times$ | $\times$ | $\times$ | $\times$ | 8 | 9 |
| $10^{-2}$ | $\times$ | $\times$ | $\times$ | $\times$ | $\times$ | $\times$ | 8 | 9 | 42 |
| $10^{-3}$ | $\times$ | $\times$ | $\times$ | $\times$ | $\times$ | 7 | 7 | 43 | 335 |
| $\alpha_{\text{opt}}$ | $\circ$ | $\circ$ | 10 | 10 | 7 | 8 | 11 | 9 | 7 |

We observe that choosing a constant step length $\alpha$ is effective as long as the impedance coefficient $\varepsilon$ is neither too small, which may lead to numerical solutions with unbounded energy norm, nor too large, which may require more iterations to observe convergence. This is in agreement with Saad's analysis [85]. Conversely, opting for the step length $\alpha_{\text{opt}}$ is a much more robust choice. However we note that when the impedance coefficient $\varepsilon$ gets very close to zero, even this choice yields unsatisfactory results. The reason is that for such values of the impedance coefficient $\varepsilon$, the Robin problem becomes closer and closer to the pure Neumann problem, for which the Uzawa approach is not applicable. We mention that the Uzawa Adjoint approach was stable for values of the impedance coefficient $\varepsilon$ as small as $10^{-13}$.

Finally, computing times for the constrained PGD approaches were recorded and compared to the unconstrained (classical) PGD. We found that the extra cost associated to the constraint was in the 20% range of the total time, except for the Lagrangian approach, in which case the extra time was about 75%. This difference is likely due to the "Jacobi" nature of the fixed-point algorithm for the Lagrangian approach, which is known to require more iterations to converge than its "Gauss-Seidel" counterpart.

## 3.6 Conclusion

In this chapter, we have introduced and analyzed several methods to incorporate a constraint within the PGD framework. We have considered the Lagrangian formulation and some

classical numerical strategies such as the Uzawa and Augmented Lagrangian approaches, and the penalization approach. Using two problems, namely a pure Neumann problem and a Robin problem, we were able to show from the numerical examples that the constrained PGD based on the Lagrangian formulation converges towards the FE solution of the Lagrangian problem (except for the Uzawa approach when the Schur complement was too large), while the penalized PGD solution converges towards the penalized FE solution.

As a conclusion of the chapter, we recommend the use of the Uzawa method, if the Schur complement is moderately small, and the Lagrangian or Augmented Lagrangian methods otherwise, which offer satisfactory results. As far as the penalization parameter $\beta$ in the Augmented Lagrangian approach is concerned, it should be chosen in such a manner that the step length remains accurate while avoiding introducing round-off errors in the solution. The optimal value of $\beta$ for a given problem actually depends on many parameters, including the spectrum of the Schur complement. The penalization approach is much simpler to implement but provides only an approximation of the constraint, except in the particular case of the pure Neumann problem (unicity recovered through the constraint).

Future developments will focus on when to update the Lagrange multiplier in the Uzawa and Augmented Lagrangian approaches, on a possible proof of convergence for the PGD solutions when using the Lagrangian formulation, on higher dimensional constraints, e.g. for the Stokes problem or quasi-incompressible solid mechanics (the methods proposed in this paper could be extended as alternative approaches to the method presented in [58]), and on the incorporation of inequality constraints using the KKT (Karush-Kuhn-Tucker) conditions [62, 67].

In the following chapters, we extend the proposed method to the goal-oriented formulation for the PGD framework, which involves a minimization problem under several constraints in terms of quantities of interest of the solution.

# CHAPTER 4 A GOAL-ORIENTED VERSION OF THE PROPER GENERALIZED DECOMPOSITION METHOD

*In this chapter, we introduce, analyze, and numerically illustrate a goal-oriented version of the Proper Generalized Decomposition method. The objective is to derive a reduced-order formulation such that the accuracy in given quantities of interest is increased when compared to a standard Proper Generalized Decomposition method. Traditional goal-oriented methods usually compute the solution of an adjoint problem following the calculation of the primal solution for error estimation and adaptation. In the present work, we propose to solve the adjoint problem first, based on a reduced approach, in order to extract estimates of the quantities of interest and use this information to constrain the reduced primal problem. The resulting reduced-order constrained solution is thus capable of delivering more accurate estimates of the quantities of interest. The performance of the proposed approach is illustrated on several numerical examples. This chapter is largely inspired by [64].*

## 4.1 Introduction

Uncertainty quantification for computational science and engineering applications requires fast methods to efficiently estimate response surfaces in sampling methods or multi-query optimization. Such a need has led to the development of model reduction approaches, for instance the Proper Orthogonal Decomposition (POD) [10, 53] and the Proper Generalized Decomposition (PGD) [39, 75] methods, the Reduced-Basis methods (RB) [27, 84], etc. These methods propose to approximate the solutions of initial and boundary-value problems as modal expansions involving a finite number of modes. Moreover, one can resort to adaptive methods based on a posteriori error estimation for the construction of the reduced models in hope of reducing the computational cost for a given accuracy. Adaptive methods are usually implemented in a greedy fashion where the discrete solution space is iteratively and a posteriori corrected based on refinement indicators provided by error estimates on the computed solution. Adjoint-based methods, often referred to as goal-oriented methods [78, 82], allow one to estimate the numerical error with respect to quantities of interest (QoI) and, thus, to adapt the solution towards the specific goals of the computer simulation.

The objective of the present work is to propose a different paradigm in which information from quantities of interest is directly incorporated into the calculation of the modes in model reduction methods. The proposed approach will be exemplified here using the Proper Generalized Decomposition method, the main objective being to approximate given outputs of interest

at a lesser cost than the classical PGD approach. It is different from the goal-oriented error estimation greedy approach for PGD solutions [5, 6, 8, 35, 68, 69] or the natural approach that consists in computing a global PGD, then localizing at the QoI level, and compacting to get fewer modes.

The proposed approach presents some similarity with the work described in [22], where the authors define specific norms with additional weighting terms taking into account the error in the quantity of interest. The main idea in [22] is to minimize a norm weighted by a functional involving the adjoint solution, via a penalization approach, in order to obtain a goal-oriented PGD using the so-called ideal minimal residual approach. The choice of this specific norm allows one to enhance accuracy with respect to some quantity of interest. The method can also be seen as a perturbation of a minimal residual method with a measure of the residual corresponding to the error in specified solution norm. A variant of the RB method, referred to as the weighted RB method, was introduced in [37] and the approach was extended to the POD framework in [90]. A comparison of the weighted RB method with the weighted POD method is presented in [91]. The weighted RB method applies specific weights, defined according to the underlying probability distribution, to the snapshots in order to properly account for the stochastic nature of the problem. In contrast, the weighted POD method introduces weights in the correlation matrix. In yet another work [32], a goal-oriented POD method was derived by considering additional snapshots based on the derivatives with respect to the parameter variables and referred to as sensibility factors. The authors showed that the resulting reduced-order solution provided more accurate results than those obtained from the classical POD due to the fact that the reduced-order model was able to better account for parameter changes. The motivation and rationale for using weights or sensibility factors have in fact some similarities with the proposed adjoint-based method of this work. Earlier works also include [30] where a goal-oriented reduced-order model is derived using concepts borrowed from the POD and the optimal control communities. There, the authors define an optimized reduced-order basis as the one minimizing the errors in the quantities of interest between the full-space solution and the reduced-order solution, subject to the constraint that the underlying governing equations are enforced.

The current approach actually builds on the a priori goal-oriented methods for least squares finite element formulations [36] and for finite element approximations of symmetric boundary-value problems [66], and is essentially an extension of these methods to reduced-order modeling. The method in [36] incorporates the error in QoI into the least squares functional via a penalization approach and inherits the global approximation properties of the standard formulation as well as increased accuracy in the quantity of interest. In practice, the value of the QoI is not known a priori; this issue can be addressed by finding a Riesz representation of

the QoI using the original least squares functional. However, the value of the QoI is replaced by an approximation of the QoI computed on a finer mesh than that used for the solution. This method represents a departure from classical goal-oriented a posteriori error estimation methods, which solve an adjoint problem to determine the sensitivity of the QoI to perturbations in the PDE solution [20]. Using the approximate Riesz representer can actually be seen as an approximation of the adjoint problem with refined discretization parameters.

A critical issue in the methods presented in [22] for PGD and in [36] for FEM is the choice of the penalization parameter, which strongly influences the quality of the numerical solution. Indeed, the penalization parameter needs to be properly chosen to ensure convergence, efficiency, and accuracy. A large value puts more weight on the QoI in the functional and, as a result, the approximation yields a more accurate QoI. However, there is a limit to the benefit of overweighting the QoI as the problem becomes ill-conditioned when the penalization parameter becomes too large. Conversely, when the penalization parameter is too small, the method fails to improve the accuracy in the QoI since the additional term in the functional is too small to account for the QoI. Moreover, the penalization parameter ought to depend on the discretization parameter in order to maintain balance between the least squares residual term and the QoI error term, which makes the selection of an optimal value of the parameter non trivial.

In [66], the issue of selecting the penalization parameter was circumvented by enforcing the constraint exactly through the use of Lagrange multipliers. In that work, it was shown that the constrained problem was well-posed and that the corresponding constrained finite element solution retained near-optimality in energy norm while being much more accurate in the desired quantities of interest. The objective of the present chapter is thus to extend this framework to the case of PGD approximations. This actually gives rise to a new issue, that of enforcing constraints in the PGD setting. Several methods were investigated and compared in [65], namely the penalization approach, the Lagrangian approach using Uzawa-type techniques, and the Augmented Lagrangian method. The conclusions of [65] were that the Lagrangian-based approaches resulted in better solutions than the penalization approach.

The chapter is organized as follows: In Section 4.2, we describe the model problem. In Section 4.3, we present the novel goal-oriented PGD method. The main idea is to use the reduced-order adjoint problems in order to obtain more accurate information about the quantities of interest, and to define a constrained PGD primal problem whose solution is tailored towards the approximation of the quantities of interest. Numerical experiments are presented in Section 4.4 to illustrate the performance of the proposed approach. We finally provide some concluding remarks in Section 4.5.

## 4.2 Model problem

Let $d \in \mathbb{N}$ and $\Omega \subset \mathbb{R}^d$ denote a hyper-rectangular domain of interest. The $d$ variables could represent space, time, and/or parameters. In general, the space variables are not separated from each other, meaning that the spatial domain need not be tensorizable. However, for simplicity in the exposition, we consider here the case where the spatial variables can also be separated. We consider a $d$-dimensional problem, defined over the tensor space $V = \bigotimes_{i=1}^{d} V^{(i)}$, where $V^{(i)}$ denotes the functional space associated to the $i$-th variable, written in global weak form as

$$\text{Find } u \in V \text{ such that } a(u, v) = f(v), \quad \forall v \in V, \tag{4.1}$$

for $a$ a continuous coercive bilinear form on $V \times V$ and $f$ a continuous linear form on $V$. For simplicity, we shall assume that above weak form can be derived from the minimization of a potential energy. If it were not the case, the method presented in this chapter could be applied by resorting to a least squares formulation [23] and the corresponding Minimal Residual PGD [75].

A classical approach for approximating the solution of (4.1) consists in discretizing space $V$ using for instance the finite element method. The mesh associated with the finite-dimensional space $V_h \subset V$ is obtained by the tensor product of one-dimensional meshes along each of the $d$ directions of domain $\Omega \subset \mathbb{R}^d$. In other words, $V_h$ is the tensor product of $d$ one-dimensional FE spaces: $V_h = \bigotimes_{i=1}^{d} V_h^{(i)}$. If $V_h^{(i)}$ introduces $N$ degrees of freedom, the dimension of the discretized space $V_h$ is thus $N^d$. This exponential growth is the so-called "curse of dimensionality" and quickly results in prohibitive costs as $N$ grows or when $d$ becomes large. A possible remedy consists in adopting a reduced-order modeling framework, whose objective is to represent the solution in terms of modes expanded on some specific basis. As a result, the number of variables used to characterize the reduced-order solution scales with $m \times d \times N$, where $m$ is the number of retained modes. Reduced-order methods differ from one another by the choice of said basis. In the present work, we will focus our attention on the PGD method.

## 4.3 Goal-oriented PGD reduced model

We assume in the present chapter that the input data admit affine representations [27]. As a consequence, the bilinear form $a$ and linear form $f$ can be separated accordingly. In addition, we also require that the linear map $Q$ representing the quantities of interest be separable with

respect to the decomposition of $V$. The reader is referred to [96] for guidelines on how to handle non-separable input data.

Let us consider $k \in \mathbb{N}$ quantities of interest $Q(u) = (Q_1(u), Q_2(u), \dots, Q_k(u)) \in \mathbb{R}^k$ where $Q$ is a linear map from $V$ to $\mathbb{R}^k$. As in [66], it is assumed that the map $Q$ is surjective onto $\mathbb{R}^k$ (if it were not it would mean that the quantities of interest $Q_i$ are redundant). The constrained optimization problem we consider is

$$\min_{\substack{v \in V \\ Q(v)=\gamma}} J(v), \tag{4.2}$$

where $\gamma \in \mathbb{R}^k$ is chosen such that the accuracy in the quantities of interest is increased when compared to the classical approach. The values of $\gamma_i$, $i = 1, \dots, k$, can be obtained through the solution $p_i$ of each adjoint problem

$$\text{Find } p_i \in V \text{ such that } a(v, p_i) = Q_i(v), \quad \forall v \in V, \tag{4.3}$$

since

$$\gamma_i = Q_i(u) = f(p_i), \quad i = 1, \dots, k. \tag{4.4}$$

In [36, 66], the target values $\gamma$ were estimated by approximating the adjoint problems on refined or enriched spaces. We propose here to use the PGD method for each of the adjoint problems, yielding approximations $\tilde{p}_{m,i}$, where $m$ is the same as in $u_m$. In that case, the values of $\gamma_i$ are approximated by

$$\gamma_{m,i} = f(\tilde{p}_{m,i}), \quad i = 1, \dots, k. \tag{4.5}$$

We note that the tilde symbol will be employed throughout the present chapter to refer to an approximation in a refined or enriched space/set. Hence, $\tilde{V}_h$ denotes a larger space than $V_h$ (piecewise quadratic and piecewise linear basis functions, respectively, in our numerical examples). Similarly for $\tilde{\mathcal{S}}_{1,h}$ and $\tilde{\mathcal{T}}_h(\delta u)$, the set of rank-1 tensors and the tangent space to the rank-1 tensor $\delta u$.

**Remark 8.** *As detailed in [66], one needs to use a larger space for the computation of the target values. If one were to use the same space for the primal and the adjoint problems, then the accuracy in the quantities of interest of the constrained solution would be the same as that of the unconstrained solution: the whole approach would be useless.*

Two strategies, introduced in [65] for constrained PGD formulations, will be used in the following to take into account the constraint in the minimization problem (4.2): the penalization

method, and the Uzawa method.

### 4.3.1 Penalization approach

The penalization method amounts to minimizing the functional

$$J_\beta(v) = J(v) + \sum_{i=1}^{k} \frac{\beta_i}{2}(Q_i(v) - \gamma_i)^2, \tag{4.6}$$

where the $\beta_i$ are penalization parameters that need to be provided. The weak formulation associated to this minimization problem reads

$$\text{Find } u \in V \text{ such that } a(u,v) + \sum_{i=1}^{k} \beta_i Q_i(u) Q_i(v) = f(v) + \sum_{i=1}^{k} \beta_i \gamma_i Q_i(v), \quad \forall v \in V. \tag{4.7}$$

Using a "two-step progressive Galerkin PGD approach" (a step for the adjoint problem followed by a step for the primal problem) leads to the following sequence of discrete problems:

**Adjoint problems:** For each $i = 1, \ldots, k$,

1) Find $\delta\tilde{p}_i \in \tilde{\mathcal{S}}_{1,h}$ such that $a(\tilde{v}, \tilde{p}_{m-1,i} + \delta\tilde{p}_i) = Q_i(\tilde{v}), \quad \forall \tilde{v} \in \tilde{\mathcal{T}}_h(\delta\tilde{p}_i)$.

2) Compute $\gamma_{m,i} = f(\tilde{p}_{m,i})$.

**Penalized primal problem:**

3) Find $\delta u \in \mathcal{S}_{1,h}$ such that

$$a(u_{m-1} + \delta u, v) + \sum_{i=1}^{k} \beta_i Q_i(u_{m-1} + \delta u) Q_i(v) = f(v) + \sum_{i=1}^{k} \beta_i \gamma_{m,i} Q_i(v), \quad \forall v \in \mathcal{T}_h(\delta u). \tag{4.8}$$

In the sequence of problems (4.8), each adjoint problem is solved using the classical Alter-

nating Directions approach. The penalized primal problem can be recast as

$$\text{Find } \delta u \in \mathcal{S}_{1,h} \text{ such that}$$
$$\begin{cases} a(\delta u, \psi_1) + \sum_{i=1}^{k} \beta_i Q_i(\delta u) Q_i(\psi_1) = R_{m-1}(\psi_1) + \sum_{i=1}^{k} \beta_i r_{m-1,i} Q_i(\psi_1), & \forall \psi_1 \in \mathcal{T}_h^{(1)}(\delta u), \\ a(\delta u, \psi_2) + \sum_{i=1}^{k} \beta_i Q_i(\delta u) Q_i(\psi_2) = R_{m-1}(\psi_2) + \sum_{i=1}^{k} \beta_i r_{m-1,i} Q_i(\psi_2), & \forall \psi_2 \in \mathcal{T}_h^{(2)}(\delta u), \\ \qquad\qquad\qquad\qquad \vdots & \\ a(\delta u, \psi_d) + \sum_{i=1}^{k} \beta_i Q_i(\delta u) Q_i(\psi_d) = R_{m-1}(\psi_d) + \sum_{i=1}^{k} \beta_i r_{m-1,i} Q_i(\psi_d), & \forall \psi_d \in \mathcal{T}_h^{(d)}(\delta u), \end{cases} \tag{4.9}$$

where we have introduced the constraint residual

$$r_{m-1} = \gamma_m - Q(u_{m-1}). \tag{4.10}$$

The penalized primal PGD (4.9) presents the advantage of being very simple to analyze and implement, since it has the same structure as that of the classical PGD problem (3.61). As a result, it can also be solved using the Alternating Directions approach. However, we mention two disadvantages of the method, as already indicated in Chapter 3: first, the choice of the penalization parameters $\beta_i$ has a strong influence on the quality of the numerical solution $u_m$; second, the constraints $Q(u_m) = \gamma_m$ are not satisfied exactly. In order to circumvent those two issues, we consider below a Lagrangian approach.

### 4.3.2 Lagrangian approach

Denoting by $\cdot$ the Euclidean inner product on $\mathbb{R}^k$, the Lagrangian approach consists in finding the saddle-point of the following Lagrangian functional

$$\mathcal{L}(v, \lambda) = J(v) + \lambda \cdot (Q(v) - \gamma), \tag{4.11}$$

so that one obtains the mixed problem

$$\text{Find } (u, \lambda) \in V \times \mathbb{R}^k \text{ such that } \begin{cases} a(u, v) + \lambda \cdot Q(v) = f(v), & \forall v \in V, \\ \tau \cdot Q(u) = \tau \cdot \gamma, & \forall \tau \in \mathbb{R}^k. \end{cases} \tag{4.12}$$

In [66] and in Chapter 2 it was shown that above problem, under the assumption of surjec-

tivity of $Q$, was well-posed.

Using a "two-step progressive Galerkin PGD approach" leads to the following sequence of problems:

**Adjoint problems:** For each $i = 1, \dots, k$,

1) Find $\delta\tilde{p}_i \in \tilde{\mathcal{S}}_{1,h}$ such that $a(\tilde{v}, \tilde{p}_{m-1,i} + \delta\tilde{p}_i) = Q_i(\tilde{v}), \quad \forall \tilde{v} \in \tilde{\mathcal{T}}_h(\delta\tilde{p}_i)$.

2) Compute $\gamma_{m,i} = f(\tilde{p}_{m,i})$.

**Constrained primal problem:** (4.13)

3) Find $(\delta u, \lambda) \in \mathcal{S}_{1,h} \times \mathbb{R}^k$ such that

$$\begin{cases} a(u_{m-1} + \delta u, v) + \lambda \cdot Q(v) = f(v), & \forall v \in \mathcal{T}_h(\delta u), \\ \tau \cdot Q(u_{m-1} + \delta u) = \tau \cdot \gamma_m, & \forall \tau \in \mathbb{R}^k. \end{cases}$$

The adjoint problems are obviously identical to those considered in the penalization approach. Only the constrained (mixed non-linear) primal problem differs from the previous approach, which can be recast as

$$\text{Find } (\delta u, \lambda) \in \mathcal{S}_{1,h} \times \mathbb{R}^k \text{ such that}$$
$$\begin{cases} a(\delta u, \psi_1) + \lambda \cdot Q(\psi_1) = R_{m-1}(\psi_1), & \forall \psi_1 \in \mathcal{T}_h^{(1)}(\delta u), \\ a(\delta u, \psi_2) + \lambda \cdot Q(\psi_2) = R_{m-1}(\psi_2), & \forall \psi_2 \in \mathcal{T}_h^{(2)}(\delta u), \\ \vdots & \\ a(\delta u, \psi_d) + \lambda \cdot Q(\psi_d) = R_{m-1}(\psi_d), & \forall \psi_d \in \mathcal{T}_h^{(d)}(\delta u), \\ \tau \cdot Q(\delta u) = \tau \cdot r_{m-1}, & \forall \tau \in \mathbb{R}^k, \end{cases} \tag{4.14}$$

where $r_{m-1}$ again denotes the constraint residual (4.10).

Problem (4.14) has a different structure from the classical PGD (3.61) due to the constraint equation. Following [65], one could solve it by considering a direct Lagrangian method, the iterative Uzawa scheme, or the Augmented Lagrangian approach. Due to its relative simplicity and its performance, as demonstrated in [65], we choose in the present work the Uzawa method. Its main feature is to decouple the constraint from the rest of the system and to approximate the Lagrange multipliers using an iterative method. The extension of the Uzawa method for finding a constrained PGD mode and the Lagrange multipliers satisfying (4.14) is described in Algorithm 3. We note that the problem in step 2 of the algorithm can be solved using the classical Alternating Directions strategy.

The step length $\alpha$, which appears in step 4 of Algorithm 3, can be taken as a constant or can

**Algorithm 3:** Algorithm for finding a constrained PGD mode (Uzawa).

1 **while** *convergence not reached* **do**
2 | Solve $a(\delta u, v) = R_{m-1}(v) - \lambda \cdot Q(v),\ \forall v \in \mathcal{T}_h(\delta u)$
3 | Compute constraint residual $r = \gamma_m - Q(u_{m-1} + \delta u)$
4 | Compute step length $\alpha$
5 | Update Lagrange multiplier $\lambda \leftarrow \lambda - \alpha r$
6 **end**
7 Update solution $u_m = u_{m-1} + \delta u$

be computed using a gradient method [65, 85]. In order to do so, we introduce $d$ (uncoupled) one-dimensional auxiliary problems. Let $\delta u = \otimes_{i=1}^{d} z_i$ be the current primal solution and $r$ the current constraint residual (see step 3 of Algorithm 3 and (4.10)), the $d$ auxiliary problems are given by

$$\begin{aligned} &\text{For each } j = 1, \dots, d, \text{ find } w_j \in V_h^{(j)} \text{ such that} \\ &\qquad a(z_1 \otimes \cdots \otimes w_j \otimes \cdots \otimes z_d, \psi_j) = Q(\psi_j) \cdot r, \quad \forall \psi_j \in \mathcal{T}_h^{(j)}(\otimes_{i=1}^{d} z_i). \end{aligned} \tag{4.15}$$

Then the gradient method leads to computing the step length $\alpha$ as

$$\alpha = \frac{r \cdot r}{r \cdot Q(w)}, \tag{4.16}$$

where $w \in \mathcal{T}_h(\otimes_{i=1}^{d} z_i)$ is constructed from the solutions $w_j$ of the auxiliary problems (4.15) as

$$w = \sum_{j=1}^{d} z_1 \otimes \cdots \otimes w_j \otimes \cdots \otimes z_d. \tag{4.17}$$

Instead of the gradient method, one could consider the conjugate gradient method. In that case, one would only need to carry out the modifications associated to the constraint residual $r$.

**Remark 9.** *As an alternative to the Uzawa approach, one could consider an augmented Lagrangian method. This consists in finding the saddle-point of* $\mathcal{L}(v, \lambda) + \sum_{i=1}^{k} \frac{\beta_i}{2}(Q_i(v) - \gamma_i)^2$, *rather than simply* $\mathcal{L}(v, \lambda)$ *in order to enhance the convergence properties of the method, see e.g. [65].*

**Remark 10.** *To get more flexibility, it may be useful to consider the relaxed constraints* $(Q_i(v) - \gamma_i)^2 \leq \epsilon$, *where* $\epsilon$ *represents a user-defined tolerance on the errors in the quantities of interest. This requires, when enforcing the constraint, to use KKT (Karush-Kuhn-Tucker) conditions [62, 67].*

## 4.4 Numerical examples

In this section we illustrate the goal-oriented PGD method on two numerical examples.

The first problem consists of an elastic beam in traction and composed of two different materials. The PGD method is used to separate the space variable from each of the two parameter variables (Young's modulus of each material). In this first example, we did not consider any approximation of the adjoint problems (4.3) and simply used the exact value of $\gamma$ as target value. This first problem is used to illustrate the method in the case where there are only two sources of numerical error with respect to the exact solution: namely (i) the truncation error arising from the finite number of modes used in the PGD expansion, and (ii) the error arising from the treatment of the constraint. I would like to thank Prof. Ludovic Chamoin for having provided the initial code, results, and report of this first example.

The second problem consists of a diffusion equation in 2D and the PGD is used to separate the two space variables. In that second example, we compute the target values $\gamma_m$ using an enriched PGD approximation of the adjoint solution.

**Example 1: A parametrized beam problem** We consider an elastic beam of length $L$ and constant cross section area $A = 1$, made of two different materials. The Young modulus $E = E(x)$ is chosen piecewise constant

$$E(x) = \begin{cases} E_1, & x \in (0, L/2), \\ E_2, & x \in (L/2, L), \end{cases} \tag{4.18}$$

where $E_1 \in \Omega_1 = [E_1^{\min}, E_1^{\max}]$ and $E_2 \in \Omega_2 = [E_2^{\min}, E_2^{\max}]$, see Figure 4.1. The beam is fixed at $x = 0$ and subjected to a unit traction force at $x = L$. Forms $a$ and $f$ are given by

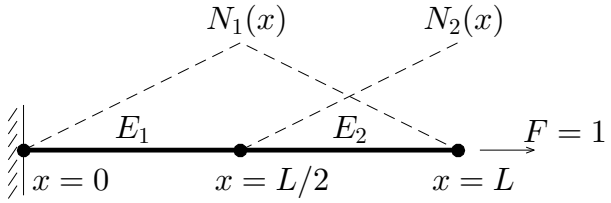


Figure 4.1 Schematic of the elastic beam in traction.

$$a(u, v) = \int_{\Omega_1} \int_{\Omega_2} \int_0^L E \frac{\partial u}{\partial x} \frac{\partial v}{\partial x}; \qquad f(v) = \int_{\Omega_1} \int_{\Omega_2} v(L). \tag{4.19}$$

For this problem, the analytical solution reads

$$u(x, E_1, E_2) = \begin{cases} \dfrac{1}{E_1} x, & x \in [0, L/2], \\ \dfrac{1}{E_1}\dfrac{L}{2} + \dfrac{1}{E_2}\left(x - \dfrac{L}{2}\right), & x \in [L/2, L], \end{cases} \tag{4.20}$$

which can be expressed in compact form as

$$u(x, E_1, E_2) = \frac{L}{2E_1}\left(N_1(x) + N_2(x)\right) + \frac{L}{2E_2} N_2(x), \tag{4.21}$$

where $N_1$ and $N_2$ are the piecewise linear Lagrange basis functions associated with nodes $x = L/2$ and $x = L$, respectively. In this example, the solution space is thus given by $V = \text{span}\{N_1, N_2\} \otimes L^2(\Omega_1) \otimes L^2(\Omega_2)$, and $u = u(x, E_1, E_2) \in V$.

The considered quantities of interest are the mean value, over the parameter domain $\Omega_1 \times \Omega_2$, of the displacement at mid-point $x = L/2$ and end-point $x = L$, respectively

$$\begin{aligned} Q_1(u) &= \frac{1}{|\Omega_1| \times |\Omega_2|} \int_{\Omega_1} \int_{\Omega_2} u(L/2), \\ Q_2(u) &= \frac{1}{|\Omega_1| \times |\Omega_2|} \int_{\Omega_1} \int_{\Omega_2} u(L), \end{aligned} \tag{4.22}$$

so that $Q = (Q_1, Q_2) : V \to \mathbb{R}^2$ denotes the functional of interest.

The exact value of the quantities of interest are

$$\begin{aligned} Q_1(u) &= \frac{L}{2}\left[\frac{1}{|\Omega_2|} \ln\left(\frac{E_1^{\max}}{E_1^{\min}}\right)\right], \\ Q_2(u) &= \frac{L}{2}\left[\frac{1}{|\Omega_2|} \ln\left(\frac{E_1^{\max}}{E_1^{\min}}\right) + \frac{1}{|\Omega_1|} \ln\left(\frac{E_2^{\max}}{E_2^{\min}}\right)\right]. \end{aligned} \tag{4.23}$$

These exact values are those used to determine $\gamma$, i.e. we do not consider an approximation of the adjoint solution for this first example. We search a PGD solution under the form

$$u_m(x, E_1, E_2) = \sum_{i=1}^{m} \varphi_i(x) \phi_{1i}(E_1) \phi_{2i}(E_2) \tag{4.24}$$

We choose $L = 1$, $E_1^{\min} = E_2^{\min} = 1$, and $E_1^{\max} = E_2^{\max} = 10$. Furthermore, we mention that the spaces $L^2(\Omega_1)$ and $L^2(\Omega_2)$ are each discretized using 500 uniformly distributed points. It was verified that the resulting discretization error was negligible with respect to the truncation errors observed in this first numerical example.

*Penalization method.* We consider the minimization of (4.6) leading to the non-linear problem (4.9). For simplicity we take the same penalization parameter $\beta$ for both constraints. We consider a progressive approach, i.e. for $u_{m-1}$ given, we compute the next iterate $u_m = u_{m-1} + \varphi \otimes \phi_1 \otimes \phi_2$, where $\varphi, \phi_1$, and $\phi_2$ are solutions to

$$a(\varphi\phi_1\phi_2, v) + \beta Q(\varphi\phi_1\phi_2) \cdot Q(v) = R_{m-1}(v) + \beta r_{m-1} \cdot Q(v), \quad \forall v \in \mathcal{T}(\varphi\phi_1\phi_2), \tag{4.25}$$

which is solved using an Alternating Directions strategy.

We consider several values of the penalization parameter, i.e. $\beta \in \{0, 10^2, 10^5\}$, in order to illustrate its influence on the results. Of course, the value $\beta = 0$ corresponds to the classical PGD decomposition (3.61), i.e. without the additional weighting term.

*Uzawa method.* We consider the saddle-point problem (4.11) leading to the constrained non-linear problem (4.14), which is solved by a progressive approach. Assuming $u_{m-1}$ is known, we compute the next iterate $u_m = u_{m-1} + \varphi \otimes \phi_1 \otimes \phi_2$, where $\varphi, \phi_1, \phi_2$, and $\lambda$ are solutions to

$$\begin{cases} a(\varphi\phi_1\phi_2, v) + \lambda \cdot Q(v) = R_{m-1}(v), & \forall v \in \mathcal{T}(\varphi\phi_1\phi_2), \\ \tau \cdot Q(\varphi\phi_1\phi_2) = \tau \cdot r_{m-1}, & \forall \tau \in \mathbb{R}^k. \end{cases} \tag{4.26}$$

This mixed non-linear problem is solved with the Uzawa method described in Algorithm 3.

We collect in Figure 4.2 the first four PGD modes obtained using

- Classical PGD (i.e. penalization with $\beta = 0$);
- Penalized PGD with $\beta = 10^2$;
- Penalized PGD with $\beta = 10^5$;
- Constrained PGD using Uzawa method.

We also provide the numerical values of the global potential energy $J(u_m) = \frac{1}{2}a(u_m, u_m) - f(u_m)$ in Table 4.1 and of the normalized quantities of interest $Q_i(u_m)/Q_i(u), i = 1, 2$ in Table 4.2. For each method, the potential energy converges to the value $-10.3618$.

The classical PGD, the penalized PGD with $\beta = 10^2$, and the Uzawa-based PGD all reach the exact value of the potential energy $J(u)$ to four digits within 7–8 modes while the penalized PGD with $\beta = 10^5$ needs 12 modes (not shown here) to reach the same accuracy. Meanwhile, concerning the accuracy in the quantities of interest, the best results are obtained using the penalized PGD with $\beta = 10^5$ and the Uzawa-based PGD.

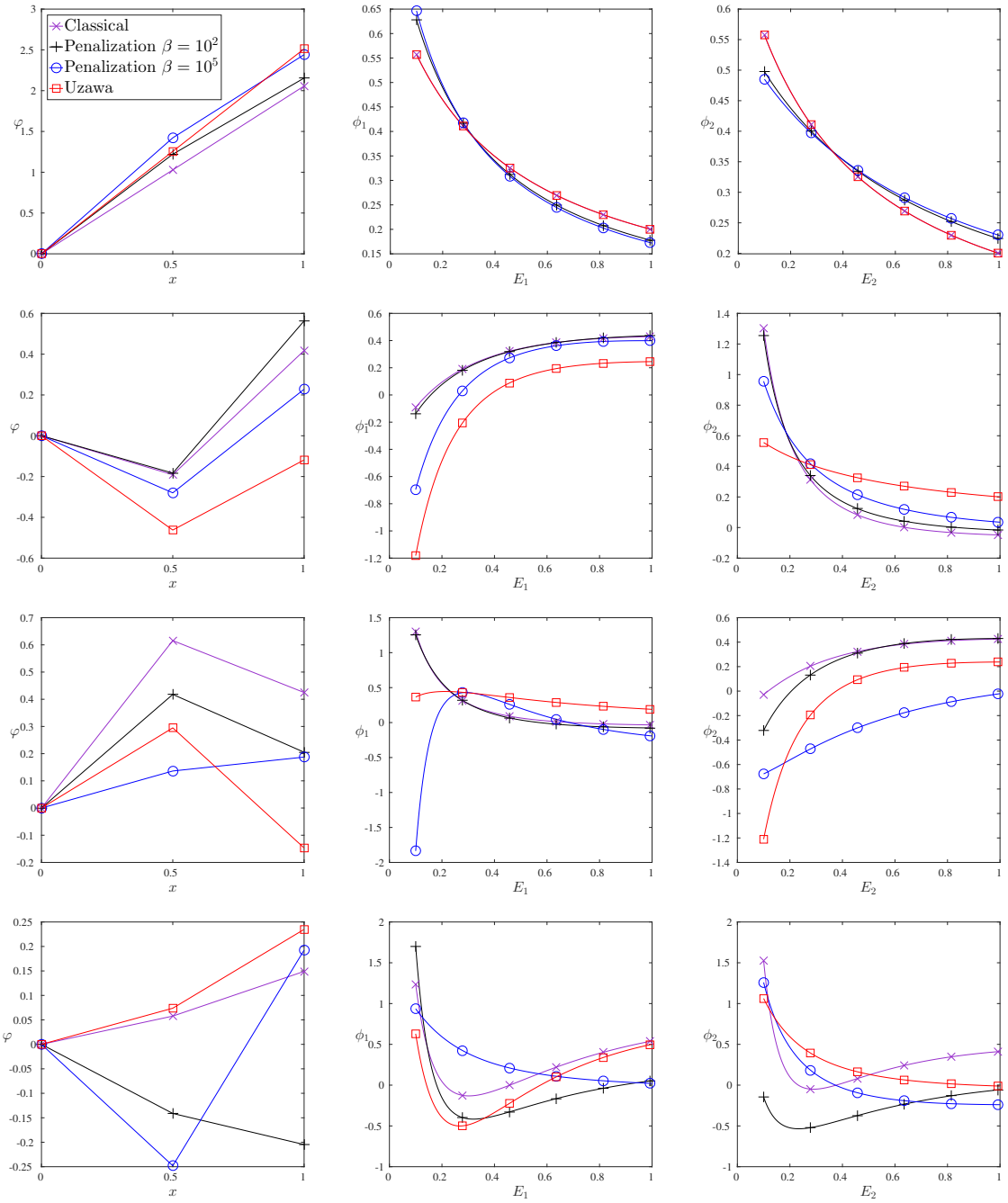


Figure 4.2 First four PGD modes (top row $m = 1$; bottom row $m = 4$) for the different methods: functions $\varphi$ (left), $\phi_1$ (center), and $\phi_2$ (right).

Table 4.1 Evolution of the potential energy $J(u_m)$ with respect to $m$.

| $m$ | $J(u_m)$ | | | |
|---|---|---|---|---|
| | $\beta = 0$ | $\beta = 10^2$ | $\beta = 10^5$ | Uzawa |
| 1 | -8.4810 | -8.3982 | -8.0812 | -8.0646 |
| 2 | -9.3514 | -9.6009 | -9.3856 | -9.1776 |
| 3 | -10.2313 | -10.2219 | -9.4526 | -10.1169 |
| 4 | -10.2894 | -10.2905 | -10.1508 | -10.1971 |
| 5 | -10.3278 | -10.3301 | -10.1888 | -10.2747 |
| 6 | -10.3525 | -10.3523 | -10.2796 | -10.3267 |
| 7 | -10.3557 | -10.3574 | -10.2967 | -10.3528 |
| 8 | -10.3578 | -10.3594 | -10.3257 | -10.3581 |
| 9 | -10.3606 | -10.3606 | -10.3293 | -10.3590 |
| 10 | -10.3612 | -10.3610 | -10.3338 | -10.3603 |

Table 4.2 Evolution of the normalized quantities of interest $Q_i(u_m)/Q_i(u)$ with respect to $m$.

| $m$ | $Q_1(u_m)/Q_1(u)$ | | | | $Q_2(u_m)/Q_2(u)$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $\beta = 0$ | $\beta = 10^2$ | $\beta = 10^5$ | Uzawa | $\beta = 0$ | $\beta = 10^2$ | $\beta = 10^5$ | Uzawa |
| 1 | 0.8185 | 0.9666 | 1.1226 | 1.0000 | 0.8185 | 0.8527 | 0.9637 | 0.9998 |
| 2 | 0.7447 | 0.8846 | 1.0125 | 1.0005 | 0.8991 | 0.9789 | 1.0085 | 0.9999 |
| 3 | 1.0000 | 1.0157 | 0.9998 | 1.0001 | 0.9872 | 1.0111 | 0.9998 | 1.0000 |
| 4 | 1.0228 | 0.9784 | 1.0003 | 1.0002 | 1.0165 | 0.9841 | 0.9996 | 1.0001 |
| 5 | 0.9898 | 0.9925 | 0.9999 | 1.0001 | 0.9929 | 0.9953 | 0.9997 | 1.0000 |
| 6 | 0.9897 | 0.9923 | 1.0000 | 1.0000 | 0.9912 | 0.9947 | 0.9998 | 1.0000 |
| 7 | 0.9961 | 1.0010 | 0.9997 | 1.0000 | 0.9948 | 0.9979 | 1.0000 | 1.0000 |
| 8 | 1.0005 | 1.0003 | 1.0000 | 1.0000 | 0.9964 | 1.0022 | 1.0002 | 1.0000 |
| 9 | 0.9980 | 1.0021 | 1.0002 | 1.0000 | 0.9993 | 1.0026 | 1.0000 | 1.0000 |
| 10 | 0.9979 | 1.0000 | 1.0002 | 1.0000 | 0.9987 | 1.0014 | 1.0000 | 1.0000 |

The numerical results show that a large penalization parameter $\beta$ improves the accuracy in the quantities of interest but worsens the convergence in the potential energy. This is because the penalization approach is a trade-off between minimizing the energy and minimizing the additional term related to the quantity of interest. Conversely, the Lagrangian/Uzawa approach circumvents the issue of selecting "a good $\beta$" by imposing the constraints through Lagrange multipliers. It is worth mentioning that the Uzawa method provides the correct values of the quantities of interest using only a couple of modes without sacrificing much the value of the potential energy.

**Example 2: A 2D diffusion problem** The second model problem consists of the Poisson equation with homogeneous Dirichlet conditions

$$\begin{cases} -\Delta u = 1, & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{4.27}$$

where $\Omega = (0,1)^2$. The exact solution of (4.27) can be found in terms of Fourier series and is shown in Figure 4.3 left-hand side. We mention that $u \in H^3(\Omega)$ for this problem.

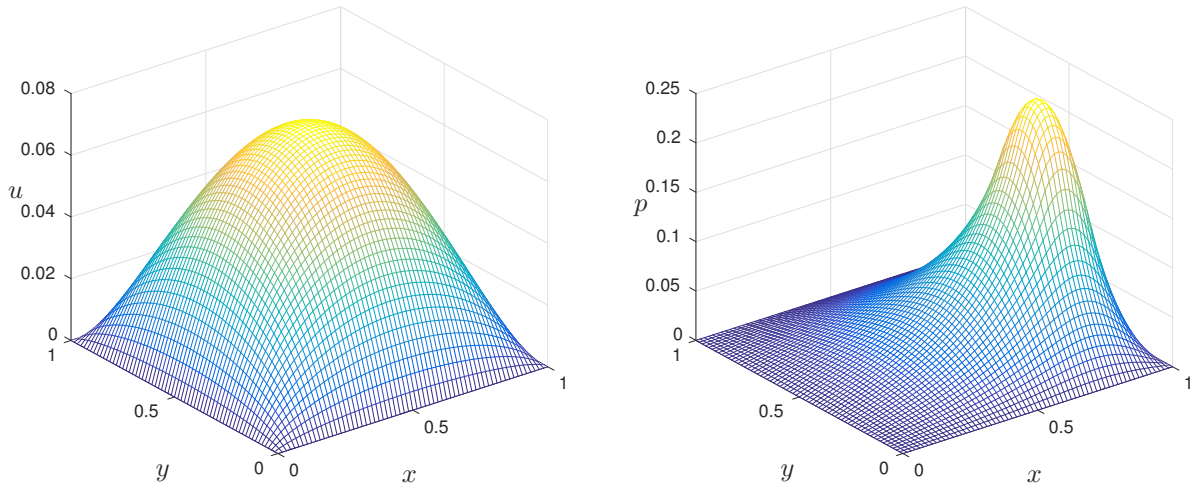


Figure 4.3 Primal $u$ (left) and adjoint $p$ (right) solutions for the diffusion example.

We also suppose that one is interested in the scalar quantity

$$Q(u) = \frac{1}{|\omega|} \int_{\omega} u, \tag{4.28}$$

where $\omega = (x_{\min}, x_{\max}) \times (y_{\min}, y_{\max})$ is a rectangular subdomain of $\Omega$ defined with $x_{\min} =$

$1/\sqrt{2}$, $x_{\max} = x_{\min} + 1/\sqrt{30}$, $y_{\min} = 1/\sqrt{18}$, $y_{\max} = y_{\min} + 1/\sqrt{17}$. The irrational bounds were intentionally chosen so that the region of interest $\omega$ does not coincide with the mesh.

The exact value of the quantity of interest (4.28) can be computed using the Fourier expansion of $u$. The adjoint solution $p \in H^3(\Omega)$ is shown in Figure 4.3 right-hand side.

Forms $a$ and $f$ are given by

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v; \qquad f(v) = \int_\Omega v. \tag{4.29}$$

The domain $\Omega$ is discretized into a uniform mesh made of $128 \times 128$ quadrilateral elements. The finite element space $V_h \subset V = H_0^1(\Omega)$ for the approximation of the primal problem is the span of piecewise continuous bilinear Lagrange basis functions. For the adjoint problem, we consider here the enriched space $\tilde{V}_h \subset V$ constructed from biquadratic hierarchical basis functions on the same mesh.

In this second example, we use the PGD to separate the $x$ and $y$ variables. We apply the "two-step progressive Galerkin PGD" method described in Section 4.3.2 combined with the Uzawa scheme to compute the constrained PGD modes.

In order to analyze the convergence results for this example, we introduce $(w_h, \lambda) \in V_h \times \mathbb{R}$, the fully discretized solution of the constrained problem

$$\text{Find } (w_h,\lambda) \in V_h \times \mathbb{R} \text{ such that } \begin{cases} a(w_h, v_h) + \lambda \cdot Q(v_h) = f(v_h), & \forall v_h \in V_h, \\ \tau \cdot Q(w_h) = \tau \cdot f(\tilde{p}), & \forall \tau \in \mathbb{R}, \end{cases} \tag{4.30}$$

where $\tilde{p} \in \tilde{V}_h$ is the fully discretized solution of the enriched adjoint problem

$$\text{Find } \tilde{p} \in \tilde{V}_h \text{ such that } a(\tilde{v}, \tilde{p}) = Q(\tilde{v}), \quad \forall \tilde{v} \in \tilde{V}_h. \tag{4.31}$$

The numerical results are collected in Figure 4.4 where we show

- in Figure 4.4 (left) the errors in energy norm, in the Lagrange multiplier, and in the quantity of interest, with respect to the fully discretized solution $(w_h, \lambda) \in V_h \times \mathbb{R}$ of the constrained problem (4.30);
- in Figure 4.4 (right), the errors in the quantity of interest with respect to the exact solution $u \in V$ of problem (4.27) for both the Classical PGD and the Goal-Oriented PGD. The dash (resp. dash-dot) line is used to show the error in the quantity of interest for the fully discretized solution in $V_h$ (resp. in $\tilde{V}_h$).

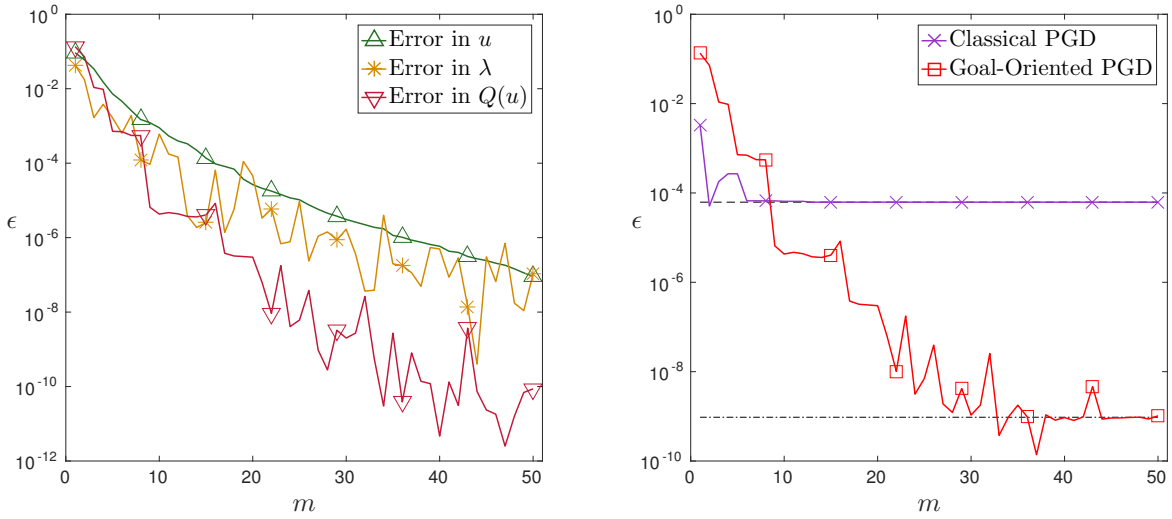


Figure 4.4 Errors with respect to the constrained solution $(w_h, \lambda) \in V_h \times \mathbb{R}$ of the fully discretized problem (4.30) (left). Errors in the quantity of interest with respect to the exact solution $u \in V$ of problem (4.27) (right).

The numerical results show that the constrained PGD solution converges to the solution $(w_h, \lambda) \in V_h \times \mathbb{R}$ of the constrained problem (4.30). Moreover, they also indicate that the error in the quantity of interest is roughly squared when considering the constrained approach. This is in agreement with the results of a priori error estimation [13] since both $u \in H^3(\Omega)$ and $p \in H^3(\Omega)$ are sufficiently regular for this problem.

## 4.5 Conclusion

In this chapter, we have extended the goal-oriented method proposed in [66] for the finite element method to the Proper Generalized Decomposition framework. The enriched adjoint approximation is computed using a PGD scheme to provide an enhanced estimate of the quantity of interest. The knowledge on the adjoint solution is then included in the primal problem as a constraint on the error in the quantity of interest and the constrained PGD problem is solved following the methodology developed in [65]. The proposed goal-oriented PGD method allows one to construct separated representations of the solution that deliver approximations of the quantity of interest with much better accuracy than the classical PGD. It is worth noting that the methodology does also handle the case where several quantities of interest are simultaneously considered. The performance of the proposed approach is illustrated on several numerical examples, namely a simple parametrized beam problem, and

a 2D Poisson problem. We observe for these linear problems a significant improvement in the accuracy of the quantities of interest obtained from the constructed PGD solutions. These preliminary results are very promising and future works will focus on: 1) the application of the method to more complex initial and boundary-value problems; 2) its extension to the case of non-linear problems and non-linear quantities of interest; 3) its extension to other reduced-order methods such as the Proper Orthogonal Decomposition approach; 4) the development of an error estimator and an adaptive strategy to enhance its performance; 5) the assessment of its efficiency for the treatment of uncertainty quantification, inverse or optimization problems, in which one has to extensively evaluate surface responses.

In the next chapter, we will apply the proposed Goal-oriented PGD strategy to a problem of engineering interest. We will model the electrostatic potential equation in a composite material, possibly featuring a delamination between two plies.

# CHAPTER 5 AN APPLICATION EXAMPLE: A PARAMETRIZED ELECTROSTATIC STUDY OF A CRACKED COMPOSITE PROBLEM

*This last chapter consists of an application example involving the electrostatic potential equation in a simplified 2D model of a composite material, featuring a possible delamination.*

## 5.1 Introduction

Composite laminates have become materials of choice in many engineering applications due to their high strength to weight ratio. However, they are prone to delamination, a mode of failure that causes layers to separate, among other types of failures. Delamination is not visually observable and its detection usually requires nondestructive testing techniques. One possible approach is electrical impedance tomography (EIT) whose main objective is to reconstruct the conductivity field of a medium by injecting current through electrodes and measuring resulting voltages. One could hypothetically use EIT data and solve inverse problems in order to identify the position and size of delamination regions in composite materials. However, the cost of solving inverse problems under the presence of uncertainties may become prohibitive for three-dimensional composite materials as the computer model should accurately predict differences of potential between electrodes. The goal of this chapter is to show that a goal-oriented PGD model tailored to the calculation of these output quantities of interest could provide a very effective surrogate computer model of the EIT experiments. In this chapter, we propose to build an accurate and cheap PGD solution to a simplified 2D electrostatic potential model of a composite material featuring a possible delamination.

## 5.2 Modelisation and problem formulation

Let $\Omega$ be the rectangle $\Omega = (0, L_x) \times (0, L_y) \subset \mathbb{R}^2$ modeling the spatial domain occupied by the composite material. As shown in Figure 5.1, we consider a 5-ply composite whose plies have thickness $L_y/5$. In this work, and for simplicity, we have chosen to model the "matrix/fiber" arrangement in the composite by a piecewise homogeneous orthotropic material. The electrical conductivity $\sigma$ corresponds to a $2 \times 2$ diagonal matrix whose values are strictly positive and piecewise constant in domain $\Omega$, modeling the directional conductivities in the horizontal and vertical directions of the different plies. Explicitely, the conductivity tensor

at each point of the domain is given by

$$\sigma = \begin{bmatrix} \sigma_x & 0 \\ 0 & \sigma_y \end{bmatrix}, \tag{5.1}$$

with $\sigma_y = 1$ and

$$\sigma_x = \sigma_x(y) = \begin{cases} \sigma_a & \text{if} \quad \frac{L_y}{5} < y < \frac{2L_y}{5}, \\ \sigma_b & \text{if} \quad \frac{3L_y}{5} < y < \frac{4L_y}{5}, \\ 10 & \text{otherwise.} \end{cases} \tag{5.2}$$

For the parametrized study, we have introduced two additional variables, $\sigma_a$ and $\sigma_b$, where $\sigma_a \in \Omega_a = \left[\sigma_a^{\min}, \sigma_a^{\max}\right]$ and $\sigma_b \in \Omega_b = \left[\sigma_b^{\min}, \sigma_b^{\max}\right]$, with $\sigma_a^{\min} > 1$ and $\sigma_b^{\min} > 1$. These additional variables could be thought of as control parameters of the stacking sequence in the composite. A high value of $\sigma_a$ in the second ply describes for example a ply where the fibers conduct extremely well the electrical current, in other words a ply where the fibers are almost perfectly aligned with the $x$-direction. A smaller conductivity indicates that the fibers become orthogonal to that direction. Finally, the lowest value of the electrical conductivity is $\sigma_y = 1$, the same in all plies, and describes the poor contact between fibers, as well as the fact that the matrix material has poor conductivity properties.

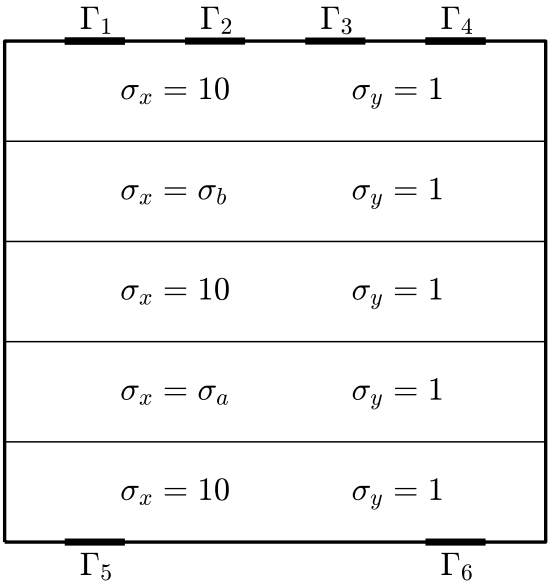


Figure 5.1 Schematic of the composite material and layout of the different electrodes.

We also equip the composite with six electrodes of same size $\ell$, $\Gamma_i$, $i = 1, \ldots, 6$, on its top and bottom boundaries, as shown in Figure 5.1. Here, we introduce the additional parame-

ter $\theta$ that describes an offset in the position of the first electrode $\Gamma_1$, that is $\Gamma_1 = \Gamma_1(\theta) = (\theta, \theta + \ell) \times \{L_y\}$. Similarly to the other parameters, we consider that $\theta$ takes values in the bounded interval $\Omega_\theta = \left[\theta^{\min}, \theta^{\max}\right]$. The other electrodes $\Gamma_2$ to $\Gamma_6$ are fixed (i.e. no additional parameter). The electrodes $\Gamma_1$ and $\Gamma_4$ are used to inject a current into the domain (non-homogeneous Neumann boundary condition). The remainder of the boundary $\partial\Omega \backslash (\Gamma_1 \cup \Gamma_4)$ is electrically insulated (homogeneous Neumann boundary condition). The other electrodes, $\Gamma_2$, $\Gamma_3$, $\Gamma_5$ and $\Gamma_6$, are used to measure differences of potential, which actually define the quantities of interest in this problem.

In the absence of an internal electrical loading force, the electrostatic equations consist of the diffusion equation and Neumann boundary conditions

$$\begin{cases} -\nabla \cdot (\sigma \nabla u) = 0, & \text{in } \Omega, \\ \quad \boldsymbol{n} \cdot \sigma \nabla u = g, & \text{on } \partial\Omega, \end{cases} \tag{5.3}$$

where $g = \chi_{\Gamma_1} - \chi_{\Gamma_4}$ (with $\chi_\Gamma$ denoting the characteristic function with respect to $\Gamma$). There exist models that better describe the behavior of the current passing through the electrodes than the Neumann boundary conditions (see e.g. [87]) but those would not change the conclusions of the present study.

Let us note that model problem (5.3) is a pure Neumann problem (with Neumann loading $g$ satisfying the compatibility condition). Consequently, the solution $u$ is defined up to an additive constant. In order to recover unicity of the solution, we look for the zero-mean weak solution to problem (5.3). As a result, the weak form reads

$$\text{Find } u \in H^1(\Omega)/\mathbb{R} \text{ such that } \int_\Omega \sigma \nabla u \cdot \nabla v = \int_{\partial\Omega} g v, \quad \forall v \in H^1(\Omega)/\mathbb{R}, \tag{5.4}$$

where $H^1(\Omega)/\mathbb{R}$ denotes the quotient space

$$H^1(\Omega)/\mathbb{R} = \left\{ v \in H^1(\Omega); \frac{1}{|\Omega|} \int_\Omega v = 0 \right\}. \tag{5.5}$$

Finally, we consider a possible delamination between two plies in the domain. The motivation for introducing the shift $\theta$ is to consider the case where the position of the delamination is not known exactly. In this study, rather than moving the delamination, we choose to simply move the electrodes. Since the delamination is fixed in the spatial domain $\Omega$, and when the context is clear, we choose not to emphasize the presence or not of a delamination in the exposition below in order to simplify the presentation. We show in Figure 5.2 some examples of the solution $u$ to (5.4) in the $(x, y)$ domain for various values of the position $\theta$ of the

electrode $\Gamma_1$ on the top boundary.

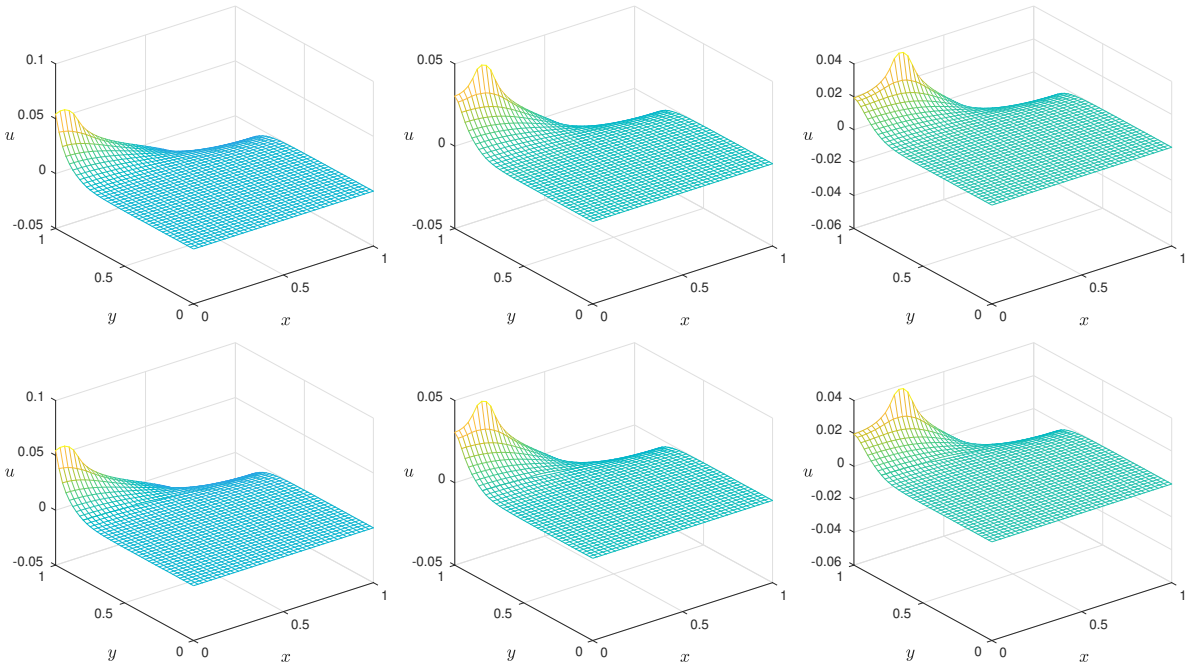

Figure 5.2 Evolution of $u(x, y)$ for fixed values of $(\sigma_a, \sigma_b)$ and $\theta = \frac{1}{90}$ (left), $\theta = \frac{10}{90}$ (center), and $\theta = \frac{19}{90}$ (right). Top row: case without delamination; bottom row: case with delamination.

We now consider the parametrized boundary-value problem of finding the weak solution to Problem (5.4) with $(\sigma_a, \sigma_b, \theta) \in \Omega_a \times \Omega_b \times \Omega_\theta$. To this end, we introduce the global weak form of the problem

$$\text{Find } u \in V \text{ such that } \int_{\Omega_a}\int_{\Omega_b}\int_{\Omega_\theta}\int_{\Omega} \sigma \nabla u \cdot \nabla v = \int_{\Omega_a}\int_{\Omega_b}\int_{\Omega_\theta}\int_{\partial\Omega} gv, \quad \forall v \in V, \tag{5.6}$$

which, in compact form, reads

$$\text{Find } u \in V \text{ such that } a(u, v) = f(v), \quad \forall v \in V, \tag{5.7}$$

where $u = u(x, y, \sigma_a, \sigma_b, \theta)$ and the solution space $V$ is defined as

$$V = \left(H^1(\Omega)/\mathbb{R}\right) \otimes L^2(\Omega_a) \otimes L^2(\Omega_b) \otimes L^2(\Omega_\theta). \tag{5.8}$$

The three scalar quantities of interest for this problem, $Q(u) \in \mathbb{R}^3$, are defined as the differ-

ences of potential averaged over the parameter space

$$\begin{cases} Q_1(u) = \dfrac{1}{\ell \times |\Omega_a| \times |\Omega_b| \times |\Omega_\theta|} \displaystyle\int_{\Omega_a}\int_{\Omega_b}\int_{\Omega_\theta} \left( \int_{\Gamma_2} u - \int_{\Gamma_3} u \right), \\ Q_2(u) = \dfrac{1}{\ell \times |\Omega_a| \times |\Omega_b| \times |\Omega_\theta|} \displaystyle\int_{\Omega_a}\int_{\Omega_b}\int_{\Omega_\theta} \left( \int_{\Gamma_2} u - \int_{\Gamma_5} u \right), \\ Q_3(u) = \dfrac{1}{\ell \times |\Omega_a| \times |\Omega_b| \times |\Omega_\theta|} \displaystyle\int_{\Omega_a}\int_{\Omega_b}\int_{\Omega_\theta} \left( \int_{\Gamma_2} u - \int_{\Gamma_6} u \right). \end{cases} \tag{5.9}$$

These quantities of interest define three adjoint problems, which are also defined on the solution space $V = (H^1(\Omega)/\mathbb{R}) \otimes L^2(\Omega_a) \otimes L^2(\Omega_b) \otimes L^2(\Omega_\theta)$

$$\text{For } i = 1, 2, 3, \text{ find } p_i \in V \text{ such that } a(v, p_i) = Q_i(v), \quad \forall v \in V. \tag{5.10}$$

We show in Figure 5.3 several examples of the adjoint solutions $p_1$, $p_2$, and $p_3$ to (5.10). It is interesting to note that $p_2$ and $p_3$ clearly show the presence of the delamination while $p_1$ does not. In other words, the quantity $Q_1$ is relatively insensitive to the delamination, as expected.

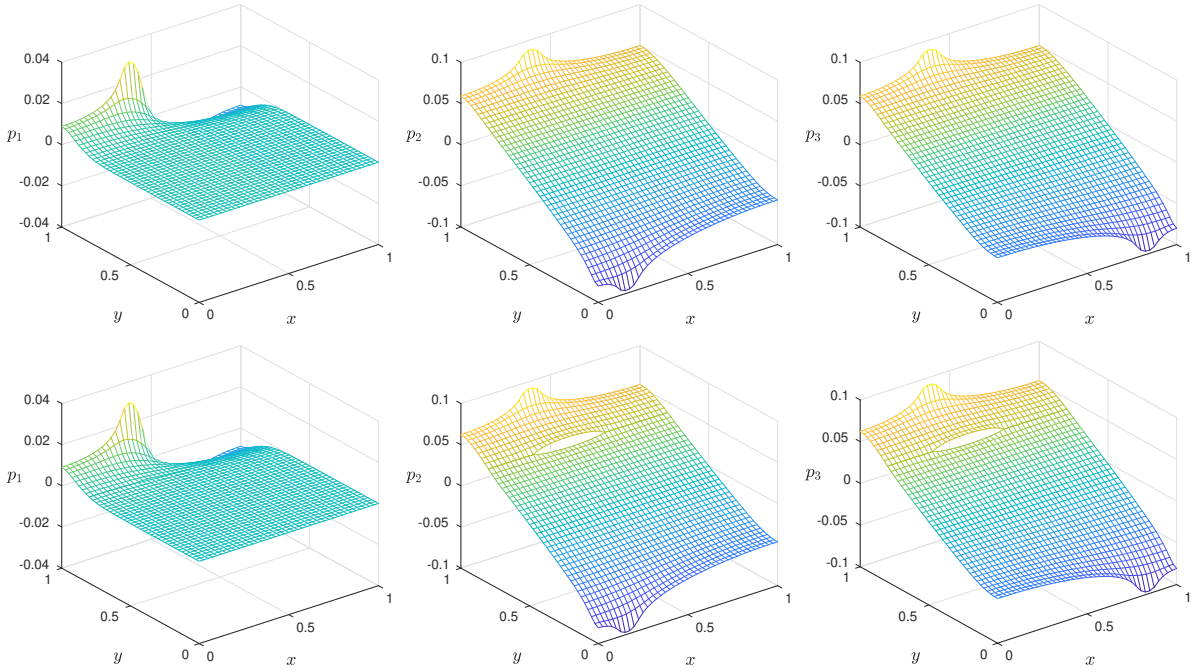


Figure 5.3 Typical instances of the adjoint solutions $p_1$ (left), $p_2$ (center), and $p_3$ (right). Top row: case without delamination; bottom row: case with delamination.

We use the Proper Generalized Decomposition method to approximate the solutions to the primal problem (5.7) and dual problems (5.10), where we separate the parameter variables from the space variables. Hence we look for a low-rank solution of the form

$$u_m(x, y, \sigma_a, \sigma_b, \theta) = \sum_{i=1}^{m} f_i(x, y) g_i(\sigma_a) h_i(\sigma_b) k_i(\theta). \tag{5.11}$$

We note that the spatial zero-mean condition is imposed globally in the 2D spatial domain. As a result, this constraint does not require a special treatment within the PGD framework since the spatial variables are not separated from each other. This constraint is merely applied using the classical Lagrangian approach when solving each 2D spatial problem during the fixed-point strategy.

Inspection of (5.6) reveals that there is a separability issue in the loading term due to the integrals over $\Omega_\theta$ and $\partial\Omega$ and the integrand term $g = g(x, \theta)$. In order to circumvent this issue in the PGD process, we construct a separated representation of the Neumann boundary term $g(x, \theta) \approx \sum_{i=1}^{M} w_i(x) z_i(\theta)$, using its Singular Value Decomposition (SVD) on a sufficiently fine grid to avoid oscillation errors [96].

### 5.3 Numerical results and analysis

The spatial domain $\Omega$ is discretized into a uniform mesh made of $40 \times 40$ quadrilateral elements. We mention that the mesh is conforming to the discontinuities in the electrical conductivity $\sigma$. The numerical treatment of the delamination is handled using double nodes. Obviously, more sophisticated approaches could have been chosen, typically by enriching the patches crossed by the delamination and those around its extremities, e.g. using PUM (Partition of Unity Methods) [73], eXtended/Generalized FEM [49] or the Stabilized GFEM [16, 17, 54, 63].

The parameter domains $\Omega_a$ and $\Omega_b$ are each discretized using a grid of 51 logarithmically-spaced points, while the parameter domain $\Omega_\theta$ is discretized using a grid of 51 linearly spaced points. The finite element space $V_h \subset V$ used for the approximation of the primal problem is the span of continuous piecewise bilinear Lagrange basis functions. For the adjoint problems, we consider the enriched space $\tilde{V}_h \subset V$ constructed from hierarchical basis functions of degree two on the same mesh. Finally, for the reference solutions, we enrich one more time and consider the space $\tilde{\tilde{V}}_h \subset V$ constructed from hierarchical basis functions of degree three on the same mesh.

For the numerical simulations, we choose $L_x = L_y = 1$, $\sigma_a^{\min} = 2$, $\sigma_a^{\max} = 4$, $\sigma_b^{\min} = 20$,

$\sigma_b^{\max} = 40$, $\theta^{\min} = 1/90$, $\theta^{\max} = 19/90$, and $\ell = 1/9$. When domain $\Omega$ features a delamination, it is located at $(0.25, 0.65) \times \{0.8\}$, i.e. between the fourth and fifth plies.

We show in Figure 5.4 (resp. Figure 5.5) the first few modes function for the PGD solutions of the primal problem (resp. of the adjoint problem corresponding to the third quantity of interest). Notice that some of the functions displayed in those figures are, or appear, constant. The specific value of the constant is due to normalization. Indeed, in each mode the parameter functions $g_i$, $h_i$ and $k_i$ are normalized and the amplitude of mode $i$ is carried in the space function $f_i$. For instance, the functions $k_i$ are constant up to machine precision for the adjoint solution as can be seen in the last column of Figure 5.5. This is due to the fact that the adjoint problem (5.10) does not depend on parameter $\theta$. One could have anticipated that the adjoint solutions were 4D and not 5D, but this was not our focus. Concerning Figure 5.4, we observe that in seven occurences, function $g_i$ or $h_i$ appears constant. In those cases, there are in fact small variations that reflect the limited influence of the parameters $\sigma_a$ or $\sigma_b$ on the respective modes. This is especially striking for $\sigma_a$, which has little influence on the first four modes, as expected. Indeed, recall that $\sigma_a$ controls the conductivity in the second ply from the bottom, whereas in the primal problem, the active electrodes, $\Gamma_1$ and $\Gamma_4$, are on the top side.

We now turn our attention to the convergence results. First, for the case without delamination, we examine the discretization error in the solutions computed in spaces $V_h$ and $\tilde{V}_h$ with respect to the reference solution obtained in $\tilde{\tilde{V}}_h$ and using $m = 41$ modes. Table 5.1 reports the relative error in energy norm and in the quantities of interest.

Table 5.1 Relative error in energy for each problem and each QoI for the case without delamination.

| Relative discretization error | in space $V_h$ | in space $\tilde{V}_h$ |
|---|---|---|
| Primal problem | 16.71% | 4.77% |
| First adjoint problem | 22.92% | 6.07% |
| Second adjoint problem | 12.17% | 3.43% |
| Third adjoint problem | 12.17% | 3.43% |
| First QoI | 0.92% | 0.0097% |
| Second QoI | 1.09% | 0.0024% |
| Third QoI | 1.09% | 0.0023% |

We now look at to the total PGD error, i.e. both discretization and truncation errors are considered, for the case without delamination. In Figure 5.6, we compare the classical PGD in $V_h$ and the goal-oriented PGD in $V_h$ with quantities of interest computed in $\tilde{V}_h$. The errors in energy norm, and in each quantity of interest, are collected as the number of modes $m$ is

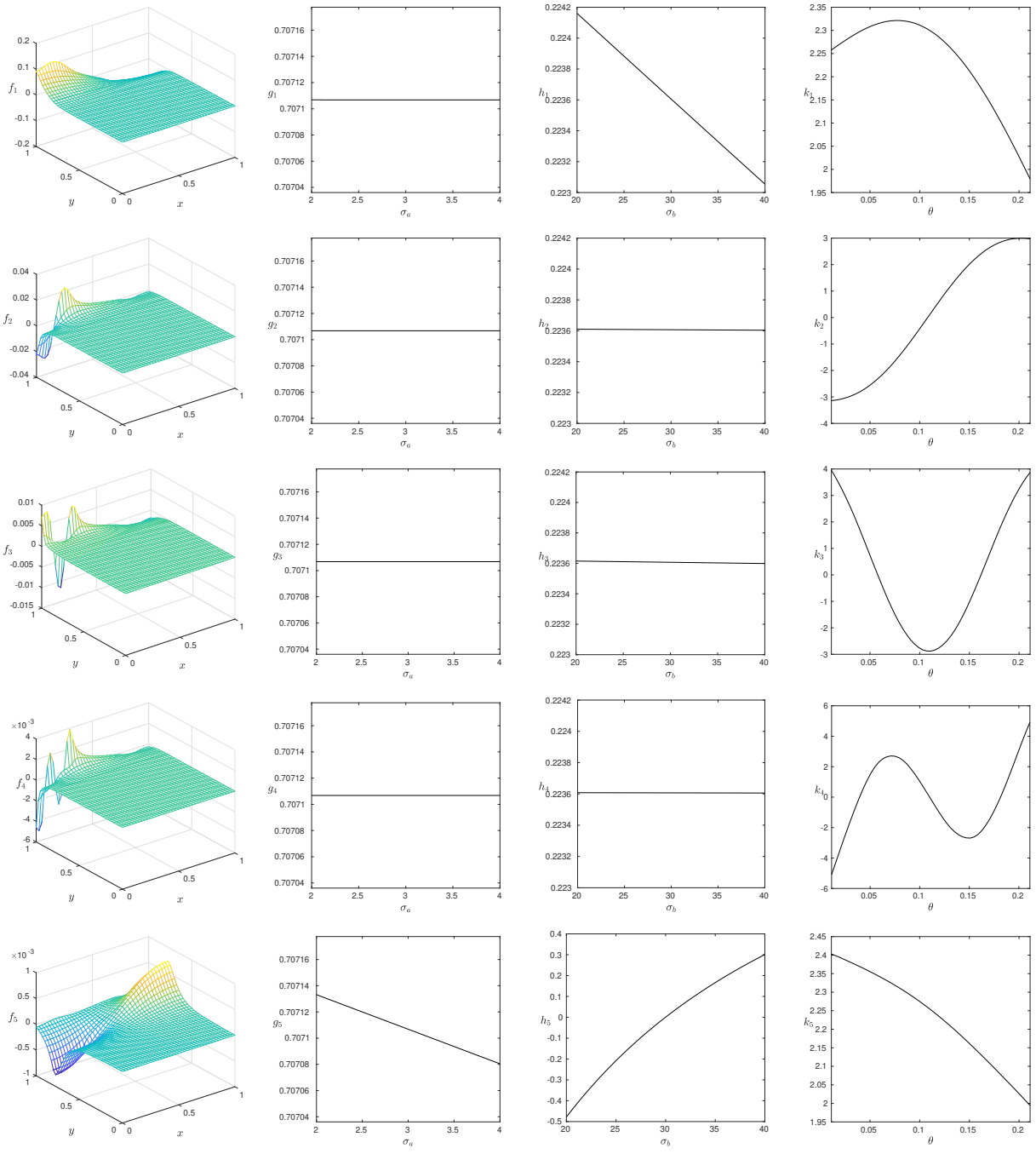


Figure 5.4 First five modes for the classical PGD solution of the primal problem.

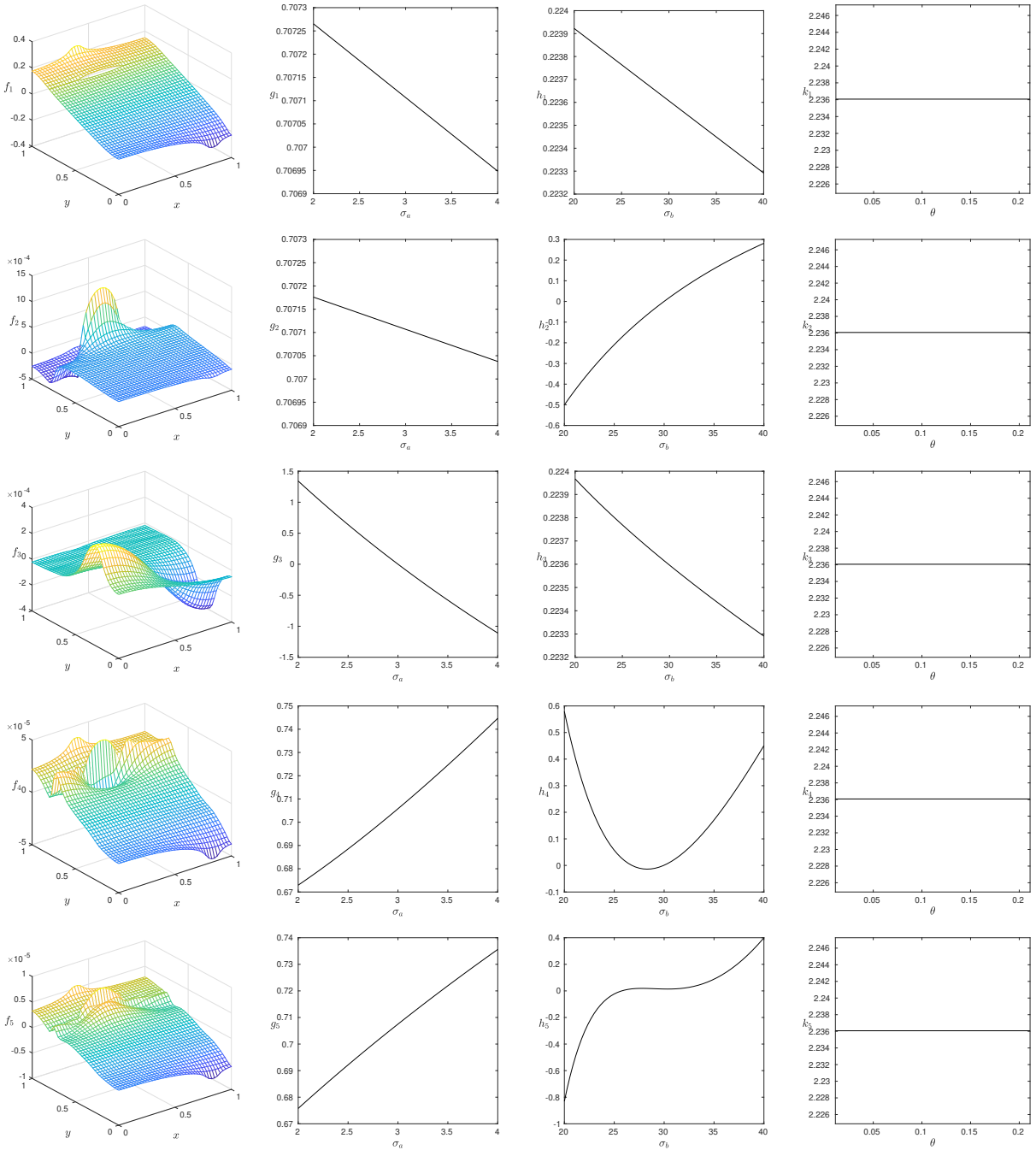


Figure 5.5 First five modes for the classical PGD solution of the third adjoint problem.

increased. As in the previous Chapter, dash (resp. dash-dot) lines are used to show the error in the quantity of interest for the fully discretized solution in $V_h$ (resp. in $\tilde{V}_h$).

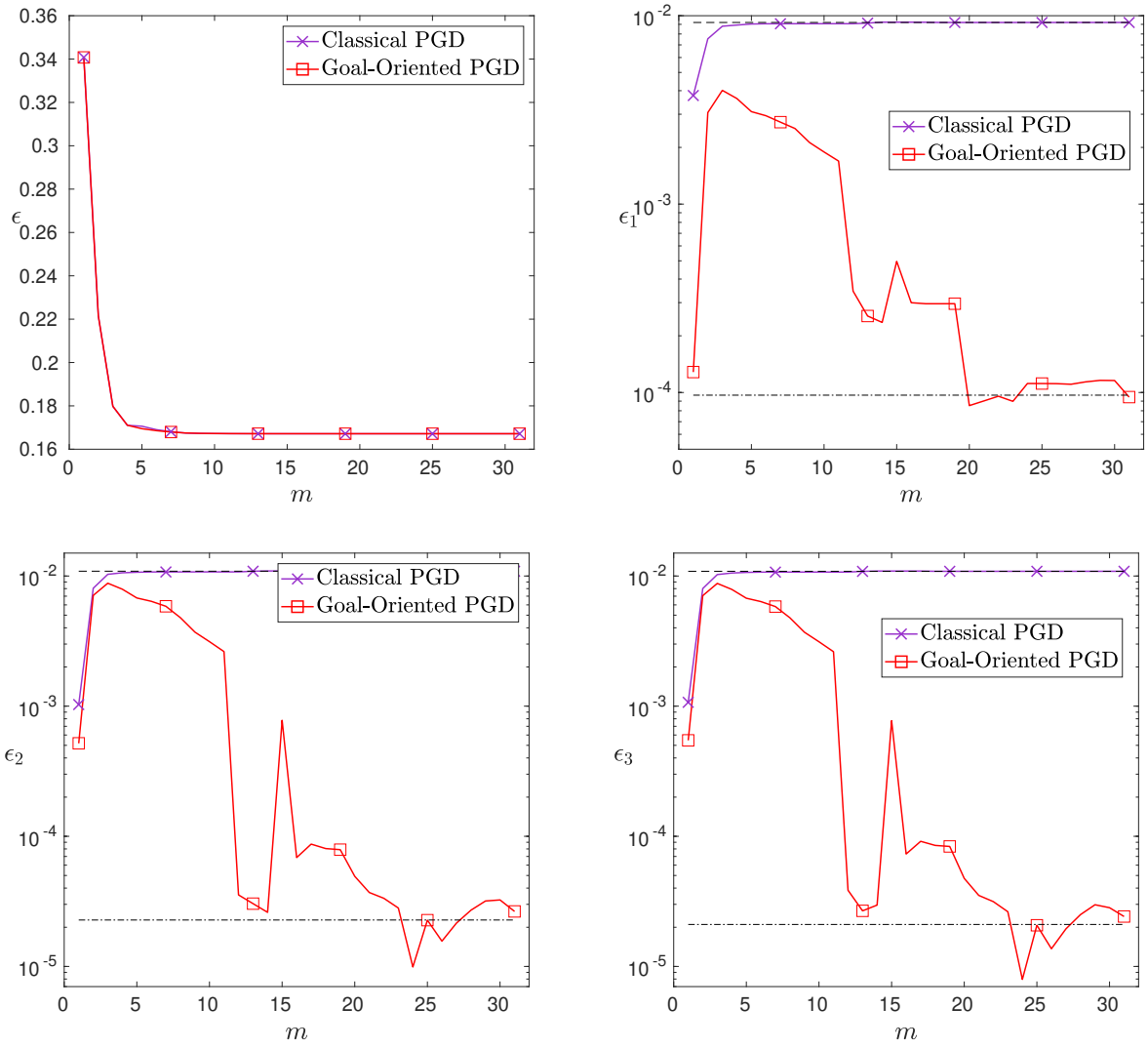


Figure 5.6 Case without delamination. Error in the energy norm (top-left). Error in each quantity of interest: error in $Q_1$ (top-right); error in $Q_2$ (bottom-left); error in $Q_3$ (bottom-right).

In Table 5.2 and Figure 5.7, we show the same results in the case where we consider a delamination between the fourth and fifth plies.

In all cases, the numerical results indicate once more that the constrained PGD is able to deliver enhanced predictions of the quantities of interest, compared to the unconstrained classical PGD. Similarly to the other examples, this is achieved without sacrificing too much

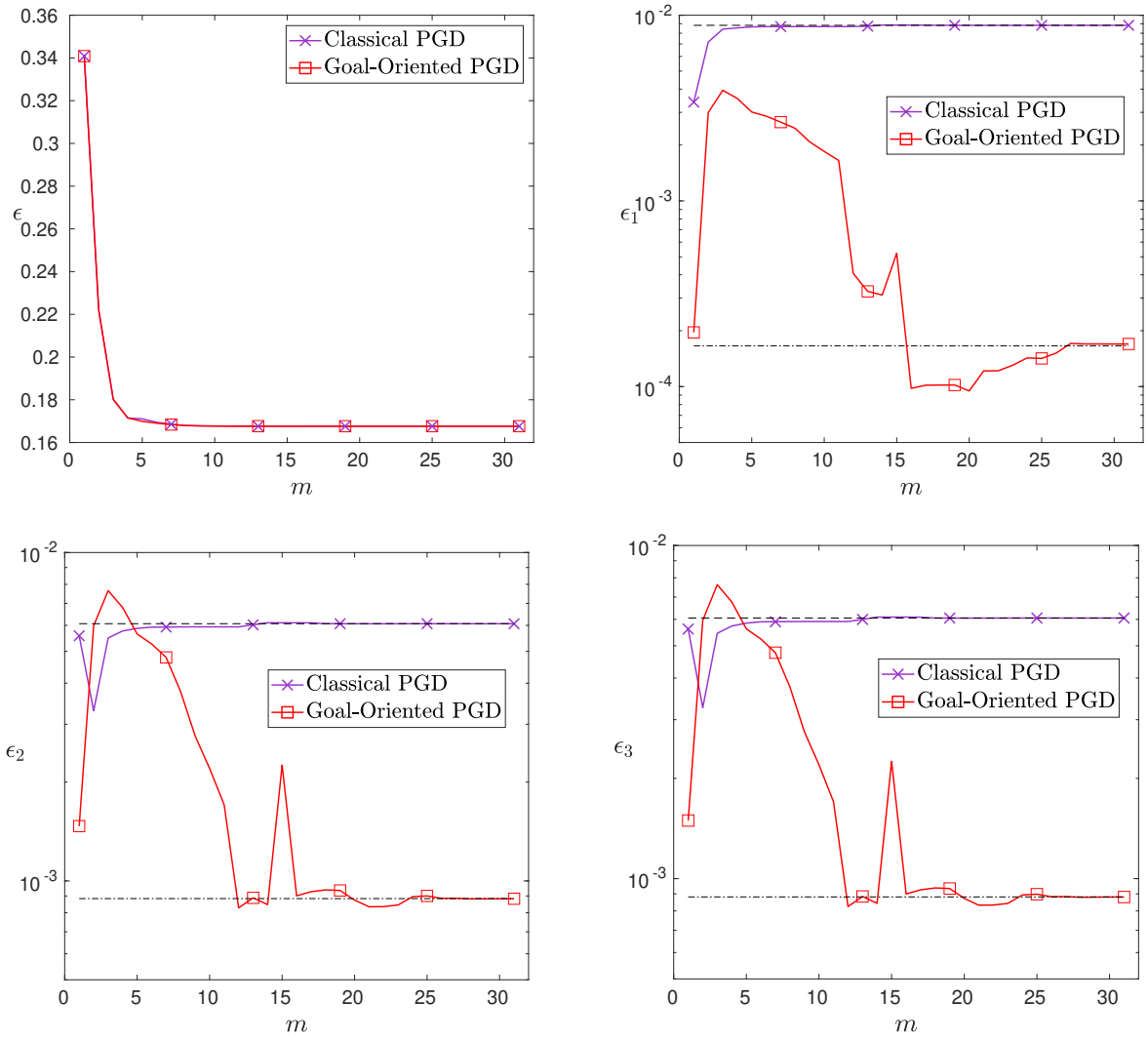


Figure 5.7 Case with delamination. Error in the energy norm (top-left). Error in each quantity of interest: error in $Q_1$ (top-right); error in $Q_2$ (bottom-left); error in $Q_3$ (bottom-right).

Table 5.2 Relative error in energy for each problem and each QoI for the case with delamination.

| Relative discretization error | in space $V_h$ | in space $\tilde{V}_h$ |
|---|---|---|
| Primal problem | 16.76% | 4.80% |
| First adjoint problem | 22.94% | 6.07% |
| Second adjoint problem | 13.25% | 4.21% |
| Third adjoint problem | 13.25% | 4.21% |
| First QoI | 0.88% | 0.017% |
| Second QoI | 0.61% | 0.088% |
| Third QoI | 0.61% | 0.088% |

the global convergence in energy norm.

Table 5.3 gathers the number of degrees of freedom associated to the full space solutions, for linear up to cubic polynomials. Recall that those spaces suffer the curse of dimensionality, as can be seen from the large number of degrees of freedom arising from the rather crude discretization chosen here. Table 5.3 also shows the dimension of the reduced spaces associated with the PGD solutions with $m = 31$ modes, along with an extra column containing the dimension reduction factors computed as the dimension of full space to dimension of reduced space ratio.

Table 5.3 Number of degrees of freedom of the full and reduced spaces, and dimension reduction factor.

| Space | Dimension | Reduced dimension | Reduction factor |
|---|---|---|---|
| $V_h$ | $0.22 \times 10^9$ | 56 854 | $4 \times 10^3$ |
| $\tilde{V}_h$ | $6.8 \times 10^9$ | 212 784 | $32 \times 10^3$ |
| $\tilde{\tilde{V}}_h$ | $50 \times 10^9$ | 467 914 | $108 \times 10^3$ |

The reduction factors are quite impressive, but they have to be mitigated because they do not reflect the number of iterations during the Alternating Direction scheme. For fair comparison, we gather in Table 5.4 the wall clock time for both the classical PGD in $V_h$ and in $\tilde{V}_h$, and for the goal-oriented PGD in space $V_h$ with dual problems solved in $\tilde{V}_h$. For reference, this is compared to the time required to assemble and solve the 2D $(x, y)$-problem for one point of the parameter grid for $(\sigma_a, \sigma_b, \theta)$, with linear (resp. quadratic) polynomials. This was recorded on a computer with 64 bits architecture, a 2.2 GHz processor and 8 GB of RAM. On this computer and using a simple incremental loop, we measured that Matlab was able to "count" up to $220 \times 10^6$ in 1 s. We mention that for simplicity, we consider the case with

delamination and only the first quantity of interest, i.e. $Q = Q_1$ in this case. Table 5.4 also includes the relative errors in the quantity of interest $Q_1(u)$.

Table 5.4 Wall clock time for each method.

| Method | Wall clock time | Error in $Q(u)$ |
|---|---|---|
| Classical PGD in $V_h$ | 21 s | 0.88% |
| Classical PGD in $\tilde{V}_h$ | 65 s | 0.017% |
| Goal-oriented PGD | 53 s | 0.017% |
| 1 simulation linear | 0.71s | |
| 1 simulation quadratic | 1.05 s | |

Table 5.4 reveals that the computational times are similar between the Classical PGD in $\tilde{V}_h$ and the Goal-oriented PGD. We even note that the Goal-oriented PGD is slighthly ahead. One reason could be that the dual problem is cheaper to solve than the primal problem.

Finally, we record the reference values (i.e. using space $\tilde{\tilde{V}}_h$) of the quantities of interest $Q_1(u)$, $Q_2(u)$, and $Q_3(u)$, and compare the cases with and without delamination in Table 5.5. The table also includes the relative difference between the two cases. The results in Table 5.5 indicate that the smallest difference is accounted for the first quantity of interest $Q_1(u)$: this corresponds to the difference of potentials between two electrodes on the same side of the composite. Conversely, the relative differences for the two other quantities of interest $Q_2(u)$, $Q_3(u)$ are larger: those are the quantities of interest for which electrodes were placed on each side of the composite.

Table 5.5 Comparison of the cases with and without delamination.

| | $Q_1(u)$ | $Q_2(u)$ | $Q_3(u)$ |
|---|---|---|---|
| Without delamination | $19.241 \times 10^{-3}$ | $9.9951 \times 10^{-3}$ | $9.9986 \times 10^{-3}$ |
| With delamination | $19.276 \times 10^{-3}$ | $10.346 \times 10^{-3}$ | $10.350 \times 10^{-3}$ |
| Relative difference | 0.18% | 3.5% | 3.5% |

This can be interepreted as follows: if the two electrodes are on the same side of the composite, then the delamination, being inside the domain, has a very small influence on the measured difference of potentials. Indeed, the path of least resistance from one electrode to the other is located between the delamination and said side. Conversely, if the two electrodes are on each side of the composite, then the streamlines of the electrical current have to adjust and go around the delamination to reach the other side of the domain. Essentially since the delamination is placed on its pathway, it has a stronger influence on the measured difference of potentials.

Finally we note from Tables 5.4 and 5.5 that the relative difference between the cases with/without delamination for the first quantity of interest is smaller than the relative error due to discretization in $V_h$. The space $V_h$ is not rich enough to allow detecting the delamination using the two electrodes that are on the same side. Conversely, the space $\tilde{V}_h$ produces a lower error in $Q_1(u)$, which is smaller than the relative difference between the cases with/without delamination. This means the enriched space $\tilde{V}_h$ is rich enough to detect the delamination. If one uses electrodes on opposite sides, which is the case for $Q_2(u)$ and $Q_3(u)$, then both space $V_h$ and space $\tilde{V}_h$ are rich enough to allow detecting the delamination.

## 5.4 Conclusion

In this chapter we have shown that a goal-oriented PGD model tailored to the calculation of targeted output quantities of interest can provide a very effective, accurate, and cheap, surrogate computer model of the simplified 2D electrostatic model of a composite material featuring a possible delamination. The parametrized solution map could also be used in an experimental design campaign, in order to help decide where best to place the electrodes to detect a possible delamination.

# CHAPTER 6 CONCLUSION AND RECOMMENDATIONS

We have presented a goal-oriented finite element methodology and its extension to the Proper Generalized Decomposition framework. Goal-oriented approaches are used when one is interested in efficiently and accurately predicting specific outputs of the solution, rather than the whole solution itself. This research topic has grown a lot since its inception, yet many challenges remain. In a multi-query study, e.g. when evaluating high-dimensional surface responses, one has to deal with problems involving a large number of parameters. Classical approaches based on discretization of the whole space then meet the so-called curse of dimensionality, which prohibits direct numerical simulation. Model-order reduction methods, of which PGD is an instance, provide a class of methods that aim at circumventing the said curse, by exploiting lower dimensional structures in the problem. A series of numerical examples, from 2D to 5D (including the parameters), demonstrated the efficiency, and the limits, of the proposed approach.

In the following two paragraphs, we provide some concluding remarks regarding the goal-oriented FEM developed in this research work. Traditional strategies in goal-oriented communities consider the primal solution $u_h \in V_h$, and then use the adjoint solution in an enriched space $\tilde{p} \in \tilde{V}_h$, for error estimation and adaptivity. The approach proposed here requires the knowledge of the adjoint solution $\tilde{p} \in \tilde{V}_h$ before computing the goal-oriented solution $(w_h, \lambda_h) \in V_h \times \mathbb{R}$ by a constrained minimization of the energy. We have shown in Chapter 2: (i) the well-posedness of the problem, (ii) the near-optimality of the corresponding solution in energy norm, and (iii) the enhanced accuracy of the solution in the quantity of interest, namely, the accuracy that one would achieve by computing the approximate solution in the space $\tilde{V}_h$.

If one were only interested in the values of the quantities of interest, one would obtain the same accuracy by considering the primal solution $\tilde{u}$ in the enriched space $\tilde{V}_h$. However, suppose now one wants to adapt the mesh: one has to solve the dual problems in an even richer space, $\tilde{\tilde{V}}_h$. Conversely, in our approach we can adapt the mesh without additional problems to be solved. The downside is that if we want an a posteriori error estimator for the quantities of interest, we also need to consider $\tilde{\tilde{V}}_h$. In short, for a similar CPU spending (corresponding to one primal solve in $V_h$ and one dual solve in $\tilde{V}_h$), the traditional approach yields: (i) an approximation of the quantity of interest, (ii) an error estimate, and (iii) adaptivity. The approach proposed here yields: (i) a more accurate quantity of interest, and (ii) adaptivity. The methodology developed here could easily be included in a black-box fashion into commercial codes that

can already handle adjoint problems and constraints. Finally, the constrained goal-oriented approach developed in this thesis could also be applied to the problem of data-assimilation, viewed from a frequentist perspective. Indeed, the methodology developed here can be readily applied in situations where experimental measurements of a set of quantities of interest are available and it is sought to incorporate these measurements as constraints while solving the primal problem. Essentially, the experimental measurements would serve as the target values $\alpha$.

Concerning the model order reduction framework, the proposed goal-oriented PGD method also allows one to obtain a more accurate estimate of the quantity of interest than the classical PGD, as demonstrated by the numerical examples from Chapters 3 through 5. The goal-oriented reduced-order approach developed here can be used to provide accurate estimates of quantities of interest at very low computational cost. This can be particularly useful for the treatment of uncertainty quantification, inverse, and optimization problems in which cases one has to evaluate surface responses a very large number of times.

Although presented for linear symmetric coercive problems, the approach is much more general. Future work will focus on the extension of the goal-oriented FEM to non-linear problems and non-linear quantities of interest. These problems are commonly solved by linearizing the non-linear equations and by using iterative schemes. In order to extend the goal-oriented methodology to these problems, one would need to find an approximation of the primal solution first in order to be able to solve for the adjoint problem. Then, one would have to iteratively consider an enriched dual problem followed by a constrained primal problem. Dual problems are linear by construction, even when the primal problem is non-linear. As a result, the full potential of the proposed goal-oriented method would appear even more clearly in that situation. Indeed the linear dual problem will be solved in the enriched space, providing enhanced approximation of the quantities of interest, while the non-linear primal problem will only be solved in the coarser space.

Some future developments for the goal-oriented PGD method include:

- the development of an adaptive goal-oriented PGD (an example of adaptive strategy is presented in [35]). The objective would be to design a goal-oriented adaptive reduced-order methodology;
- a proof of convergence for the constrained PGD solutions (Lagrangian formulation). The idea would be to adapt the proof in the case of PGD without constraints from [48];
- the extension of the goal-oriented method to problems with higher dimensional constraints, such as for the Stokes problem or quasi-incompressible solid mechanics. Due

to the increase in the number of constraints, the Augmented Lagrangian method would be the most appropriate;

- the extension of the proposed methodology to stochastic problems or problems with uncertain data. The Proper Generalized Decomposition already has a stochastic equivalent in the Stochastic Galerkin framework, the so-called Generalized Spectral Decomposition [74]. The extension of our goal-oriented approach to GSD should be straightforward.

# REFERENCES


[1] Interior Point OPTimizer (IPOPT) software package. `https://projects.coin-or.org/Ipopt/`.

[2] M. S. Aghighi, A. Ammar, C. Metivier, M. Normandin, and F. Chinesta. Non-incremental transient solution of the Rayleigh–Bénard convection model by using the PGD. *Journal of Non-Newtonian Fluid Mechanics*, 200:65–78, 2013.

[3] M. Ainsworth and J. T. Oden. A posteriori error estimation in finite element analysis. *Computer Methods in Applied Mechanics and Engineering*, 142(1–2):1–88, 1997.

[4] M. Ainsworth and J. T. Oden. *A Posteriori Error Estimation in Finite Element Analysis*. John Wiley Sons, 2000.

[5] I. Alfaro, D. González, S. Zlotnik, P. Díez, E. Cueto, and F. Chinesta. An error estimator for real-time simulators based on model order reduction. *Advanced Modeling and Simulation in Engineering Sciences*, 2(1):30, 2015.

[6] J. P. Almeida. A basis for bounding the errors of proper generalised decomposition solutions in solid mechanics. *International Journal for Numerical Methods in Engineering*, 94(10):961–984, 2013.

[7] D. Alonso, A. Velazquez, and J. Vega. A method to generate computationally efficient reduced order models. *Computer Methods in Applied Mechanics and Engineering*, 198(33–36):2683–2691, 2009.

[8] A. Ammar, F. Chinesta, P. Díez, and A. Huerta. An error estimator for separated representations of highly multidimensional models. *Computer Methods in Applied Mechanics and Engineering*, 199(25–28):1872–1880, 2010.

[9] A. Ammar, A. Huerta, F. Chinesta, E. Cueto, and A. Leygue. Parametric solutions involving geometry: a step towards efficient shape optimization. *Computer Methods in Applied Mechanics and Engineering*, 268:178–193, 2014.

[10] D. Amsallem and C. Farhat. Interpolation method for adapting reduced-order models and application to aeroelasticity. *AIAA Journal*, 46(7):1803–1813, 2008.

[11] 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.

[12] I. Babuška. Error-bounds for finite element method. *Numerische Mathematik*, 16(4):322–333, 1971.

[13] I. Babuška and A. Miller. The post-processing approach in the finite element method – Part 1: Calculation of displacements, stresses and other higher derivatives of the displacements. *International Journal for Numerical Methods in Engineering*, 20(6):1085–1109, 1984.

[14] I. Babuška and W. C. Rheinboldt. Error estimates for adaptive finite element computations. *SIAM Journal on Numerical Analysis*, 15(4):736–754, 1978.

[15] I. Babuška and T. Strouboulis. *The finite element method and its reliability*. Oxford university press, 2001.

[16] I. Babuška and U. Banerjee. Stable generalized finite element method (sgfem). *Computer Methods in Applied Mechanics and Engineering*, 201-204:91–111, 2012.

[17] I. Babuška, U. Banerjee, and K. Kergrene. Strongly stable generalized finite element method: Application to interface problems. *Computer Methods in Applied Mechanics and Engineering*, 327:58–92, 2017. Advances in Computational Mechanics and Scientific Computation—the Cutting Edge.

[18] W. Bangerth and R. Rannacher. Adaptive finite element methods for differential equations. *Lectures in Mathematics, ETH Zürich, Birkhäuser*, 2003.

[19] R. Becker and R. Rannacher. A feed-back approach to error control in finite element methods: Basic analysis and examples. *Journal of Numerical Mathematics*, 4:237–264, 1996.

[20] R. Becker and R. Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numerica*, 10:1–102, 2001.

[21] R. E. Bellman. *Adaptive control processes: a guided tour*, volume 2045. Princeton university press, 2015.

[22] M. Billaud-Friess, A. Nouy, and O. Zahm. A tensor approximation method based on ideal minimal residual formulations for the solution of high-dimensional problems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 48(6):1777–1806, 2014.

[23] P. Bochev and M. D. Gunzburger. *Least-Squares Finite Element Methods*, volume 166. 2009.

[24] P. Bochev and R. B. Lehoucq. On the finite element solution of the pure Neumann problem. *SIAM Review*, 47(1):50–66, 2005.

[25] B. Bognet, F. Bordeu, F. Chinesta, A. Leygue, and A. Poitou. Advanced simulation of models defined in plate geometries: 3D solutions with 2D computational complexity. *Computer Methods in Applied Mechanics and Engineering*, 201–204:1–12, 2012.

[26] R. Bouclier, F. Louf, and L. Chamoin. Real-time validation of mechanical models coupling PGD and constitutive relation error. *Computational Mechanics*, 52(4):861–883, 2013.

[27] S. Boyaval, C. Le Bris, T. Lelièvre, Y. Maday, N. C. Nguyen, and A. T. Patera. Reduced basis techniques for stochastic problems. *Archives of Computational methods in Engineering*, 17(4):435–454, 2010.

[28] F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. *Revue Française d'Automatique, Informatique, Recherche Opérationnelle. Analyse Numérique*, 8(2):129–151, 1974.

[29] T. Bui-Thanh, K. Willcox, and O. Ghattas. Parametric reduced-order models for probabilistic analysis of unsteady aerodynamic applications. *AIAA Journal*, 46:2520–2529, 2008.

[30] T. Bui-Thanh, K. Willcox, O. Ghattas, and B. van Bloemen Waanders. Goal-oriented, model-constrained optimization for reduction of large-scale systems. *Journal of Computational Physics*, 224(2):880–896, 2007.

[31] G. F. Carey and J. T. Oden. *Finite Elements: Computational Aspects*, volume 3. 1984.

[32] K. Carlberg and C. Farhat. A low-cost, goal-oriented "compact proper orthogonal decomposition" basis for model reduction of static systems. *International Journal for Numerical Methods in Engineering*, 86:381–402, 04 2011.

[33] K. Carlberg, C. Farhat, J. Cortial, and D. Amsallem. The gnat method for nonlinear model reduction: Effective implementation and application to computational fluid dynamics and turbulent flows. *Journal of Computational Physics*, 242:623 – 647, 2013.

[34] L. Chamoin, P.-E. Allier, and B. Marchand. Synergies between the constitutive relation error concept and PGD model reduction for simplified V&V procedures. *Advanced Modeling and Simulation in Engineering Sciences*, 3(1):1–26, 2016.

[35] L. Chamoin, F. Pled, P.-E. Allier, and P. Ladevèze. A posteriori error estimation and adaptive strategy for PGD model reduction applied to parametrized linear parabolic problems. *Computer Methods in Applied Mechanics and Engineering*, 327:118–146, 2017.

[36] J. H. Chaudhry, E. C. Cyr, K. Liu, T. A. Manteuffel, L. N. Olson, and L. Tang. Enhancing least-squares finite element methods through a quantity-of-interest. *SIAM Journal on Numerical Analysis*, 52(6):3085–3105, 2014.

[37] P. Chen, A. Quarteroni, and G. Rozza. A weighted reduced basis method for elliptic partial differential equations with random input data. *SIAM Journal on Numerical Analysis*, 51(6):3163–3185, 2013.

[38] F. Chinesta, A. Ammar, and E. Cueto. Recent advances and new challenges in the use of the proper generalized decomposition for solving multidimensional models. *Archives of Computational methods in Engineering*, 17(4):327–350, 2010.

[39] F. Chinesta, R. Keunings, and A. Leygue. *The Proper Generalized Decomposition for Advanced Numerical Simulations*. 2014.

[40] F. Chinesta, P. Ladevèze, and E. Cueto. A short review on Model Order Reduction based on Proper Generalized Decomposition. *Archives of Computational Methods in Engineering*, 18(4):395–404, 2011.

[41] F. Chinesta, A. Leygue, F. Bordeu, J. V. Aguado, E. Cueto, D. González, I. Alfaro, A. Ammar, and A. Huerta. PGD-based computational vademecum for efficient design, optimization and control. *Archives of Computational Methods in Engineering*, 20(1):31–59, 2013.

[42] P. G. Ciarlet. *The finite element method for elliptic problems*. 1978.

[43] A. Dumon, C. Allery, and A. Ammar. Proper general decomposition (PGD) for the resolution of Navier–Stokes equations. *Journal of Computational Physics*, 230(4):1387–1407, 2011.

[44] A. Dumon, C. Allery, and A. Ammar. Proper Generalized Decomposition method for incompressible Navier–Stokes equations with a spectral discretization. *Applied Mathematics and Computation*, 219(15):8145–8162, 2013.

[45] A. El Hamidi, H. Ossman, and M. Jazar. On the convergence of alternating minimization methods in variational PGD. *Computational Optimization and Applications*, 68(2):455–472, 2017.

[46] B. Endtmayer and T. Wick. A partition-of-unity dual-weighted residual approach for multi-objective goal functional error estimation applied to elliptic problems. *Computational Methods in Applied Mathematics*, 2017.

[47] D. Estep, M. Holst, and M. Larson. Generalized Green's functions and the effective domain of influence. *SIAM Journal on Scientific Computing*, 26(4):1314–1339, 2005.

[48] A. Falcó and A. Nouy. A proper generalized decomposition for the solution of elliptic problems in abstract form by using a functional Eckart–Young approach. *Journal of Mathematical Analysis and Applications*, 376(2):469–480, 2011.

[49] T. P. Fries and T. Belytschko. The extended/generalized finite element method: An overview of the method and its applications. *International Journal for Numerical Methods in Engineering*, 84(3):253–304, 2010.

[50] E. C. Gartland. Computable pointwise error bounds and the Ritz method in one dimension. *SIAM Journal on Numerical Analysis*, 21(1):84–100, 1984.

[51] C. Ghnatios, E. Abisset-Chavanne, C. Binetruy, F. Chinesta, and S. Advani. 3D modeling of squeeze flow of multiaxial laminates. *Journal of Non-Newtonian Fluid Mechanics*, 234:188–200, 2016.

[52] M. D. Gunzburger. Reduced-order modeling, data compression and the design of experiments, 2004. Second DOE Workshop on Multiscale Mathematics, July 20–22, Broomfield, Colorado.

[53] M. D. Gunzburger, J. S. Peterson, and J. N. Shadid. Reduced-order modeling of time-dependent PDEs with multiple parameters in the boundary data. *Computer Methods in Applied Mechanics and Engineering*, 196(4):1030–1047, 2007.

[54] V. Gupta, C. A. Duarte, I. Babuška, and U. Banerjee. A stable and optimally convergent generalized fem (sgfem) for linear elastic fracture mechanics. *Computer Methods in Applied Mechanics and Engineering*, 266:23–39, 2013.

[55] W. W. Hager. Updating the inverse of a matrix. *SIAM review*, 31(2):221–239, 1989.

[56] R. Hartmann. Multitarget error estimation and adaptivity in aerodynamic flow simulations. *SIAM Journal on Scientific Computing*, 31(1):708–731, 2008.

[57] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, 1990.

[58] R. Ibáñez, E. Abisset-Chavanne, F. Chinesta, and A. Huerta. Simulating squeeze flows in multiaxial laminates: towards fully 3d mixed formulations. *International Journal of Material Forming*, pages 1–17, 2016.

[59] S. Janardhanan. Model order reduction and controller design techniques, 2005.

[60] C. Johnson. *Numerical solution of partial differential equations by the finite element method*. 2012.

[61] K. Karhunen. *Zur spektraltheorie stochastischer prozesse*, volume 34. 1946.

[62] W. Karush. Minima of functions of several variables with inequalities as side constraints. Master's thesis, Department of Mathematics, University of Chicago, Chicago, Illinois, 1939.

[63] K. Kergrene, I. Babuška, and U. Banerjee. Stable generalized finite element method and associated iterative schemes; application to interface problems. *Computer Methods in Applied Mechanics and Engineering*, 305:1–36, 2016.

[64] K. Kergrene, L. Chamoin, M. Laforest, and S. Prudhomme. On a Goal-Oriented version of the Proper Generalized Decomposition. *Accepted in Journal of Scientific Computing*, 2018.

[65] K. Kergrene, S. Prudhomme, L. Chamoin, and M. Laforest. Approximation of constrained problems using the PGD method with application to pure Neumann problems. *Computer Methods in Applied Mechanics and Engineering*, 317:507–525, 2017.

[66] K. Kergrene, S. Prudhomme, L. Chamoin, and M. Laforest. A new goal-oriented formulation of the finite element method. *Computer Methods in Applied Mechanics and Engineering*, 327:256–276, 2017.

[67] H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492, Berkeley, California, 1951.

[68] P. Ladevèze and L. Chamoin. Toward guaranteed PGD-reduced models. *Bytes and Science. CIMNE: Barcelona*, pages 143–154, 2013.

[69] P. Ladevèze and L. Chamoin. On the verification of model reduction methods based on the Proper Generalized Decomposition. *Computer Methods in Applied Mechanics and Engineering*, 200(23–24):2032–2047, 2011.

[70] O. P. Le Maître and L. Mathelin. Equation-free model reduction for complex dynamical systems. *International Journal for Numerical Methods in Fluids*, 63(2):163–184, 2010.

[71] M. Loève. *Fonctions aléatoires du second ordre*, volume 220. 1945.

[72] D. J. Lucia, P. S. Beran, and W. A. Silva. Reduced-order modeling: new approaches for computational physics. *Progress in Aerospace Sciences*, 40(1–2):51–117, 2004.

[73] J. M. Melenk and I. Babuška. The partition of unity finite element method: Basic theory and applications. *Computer Methods in Applied Mechanics and Engineering*, 139(1):289–314, 1996.

[74] A. Nouy. A generalized spectral decomposition technique to solve a class of linear stochastic partial differential equations. *Computer Methods in Applied Mechanics and Engineering*, 196(45):4521–4537, 2007.

[75] A. Nouy. A priori model reduction through proper generalized decomposition for solving time-dependent partial differential equations. *Computer Methods in Applied Mechanics and Engineering*, 199(23–24):1603–1626, 2010.

[76] A. Nouy. Proper generalized decompositions and separated representations for the numerical solution of high dimensional stochastic problems. *Archives of Computational Methods in Engineering*, 17(4):403–434, 2010.

[77] J. T. Oden and L. F. Demkowicz. *Applied Functional Analysis*. 1996.

[78] J. T. Oden and S. Prudhomme. Goal-oriented error estimation and adaptivity for the finite element method. *Computers & Mathematics with Applications*, 41(5–6):735–756, 2001.

[79] J. T. Oden and S. Prudhomme. Estimation of modeling error in computational mechanics. *Journal of Computational Physics*, 182(2):496–515, 2002.

[80] M. Paraschivoiu, J. Peraire, and A. T. Patera. A posteriori finite element bounds for linear-functional outputs of elliptic partial differential equations. *Computer Methods in Applied Mechanics and Engineering*, 150(1-4):289–312, 1997.

[81] S. Prudhomme, K. Kergrene, D. Guignard, D. Pardo, and V. Darrigrand. Refinement indicators and adaptive schemes for goal-oriented error estimation. In *International Conference on Adaptive Modeling and Simulation ADMOS 2017*, Verbania, Italy, June 2017.

[82] S. Prudhomme and J. T. Oden. On goal-oriented error estimation for elliptic problems: application to the control of pointwise errors. *Computer Methods in Applied Mechanics and Engineering*, 176(1–4):313–331, 1999.

[83] G. Rozza. An introduction to reduced basis method for parametrized PDEs. In *Applied and Industrial Mathematics in Italy III*, volume 82 of *Series on Advances in Mathematics for Applied Sciences*, pages 508–519, 2010.

[84] G. Rozza, D. Huynh, and A. T. Patera. Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations. *Archives of Computational Methods in Engineering*, 15(3):229–275, 2008.

[85] Y. Saad. *Iterative Methods for Sparse Linear Systems*. 2003.

[86] W. Schilders. Introduction to model order reduction. In *Model Order Reduction: Theory, Research Aspects and Applications*, volume 13 of *Mathematics in Industry*, pages 3–32. 2008.

[87] E. Somersalo, M. Cheney, and D. Isaacson. Existence and uniqueness for electrode models for electric current computed tomography. *SIAM Journal on Applied Mathematics*, 52(4):1023–1040, 1992.

[88] H. Uzawa. Iterative methods for concave programming. *Studies in linear and nonlinear programming*, 6, 1958.

[89] E. H. van Brummelen, S. Zhuk, and G. J. van Zwieten. Worst-case multi-objective error estimation and adaptivity. *Computer Methods in Applied Mechanics and Engineering*, 313:723–743, 2017.

[90] L. Venturi, D. Torlo, F. Ballarin, and G. Rozza. A weighted POD method for elliptic PDEs with random inputs. *ArXiv e-prints*, February 2018.

[91] L. Venturi, D. Torlo, F. Ballarin, and G. Rozza. Weighted reduced order methods for parametrized partial differential equations with random inputs. *ArXiv e-prints*, May 2018.

[92] R. Verfürth. *A posteriori error estimation techniques for finite element methods*. Oxford university press, 2013.

[93] P. Vidal, L. Gallimard, and O. Polit. Proper generalized decomposition and layer-wise approach for the modeling of composite plate structures. *International Journal of Solids and Structures*, 50(14–15):2239–2250, 2013.

[94] A. Wächter and L. 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.

[95] J. Yvonnet and Q.-C. He. The reduced model multiscale method (r3m) for the non-linear homogenization of hyperelastic media at finite strains. *Journal of Computational Physics*, 223(1):341–368, 2007.

[96] S. Zlotnik, P. Díez, D. Gonzalez, E. Cueto, and A. Huerta. Effect of the separated approximation of input data in the accuracy of the resulting PGD solution. *Advanced Modeling and Simulation in Engineering Sciences*, 2(1):1–14, 2015.

[97] S. Zlotnik, P. Díez, D. Modesto, and A. Huerta. Proper generalized decomposition of a geometrically parametrized heat problem with geophysical applications. *International Journal for Numerical Methods in Engineering*, 103(10):737–758, 2015.