Politecnico di Milano

Scuola di Ingegneria dei Sistemi

Corso di Studi in Ingegneria Matematica

Tesi di Laurea Magistrale

Reduced basis method for parametrized optimal control problems governed by PDEs

Relatore: Prof. Alfio Quarteroni
Correlatore: Dr. Ing. Gianluigi Rozza

Tesi di Laurea di:
Federico Negri
Matr. 750137

Anno Accademico 2010-2011

alla mia famiglia,
Anna, Carlo e Silvia.

Abstract

This master thesis aims at the development, analysis and computer implementation of efficient numerical methods for the solution of optimal control problems based on parametrized partial differential equations. Our goal is to develop a new approach based on suitable model reduction paradigm – the reduced basis method (RB) – for the rapid and reliable solution of control problems which may occur in several engineering contexts. In particular, we develop the methodology for parametrized quadratic optimization problems with either coercive elliptic equations or Stokes equations as constraints. Firstly, we recast the optimal control problem in the framework of mixed variational problems in order to take advantage of the already developed RB theory for Stokes-type problems. Then the usual ingredients of the RB methodology are provided: a Galerkin projection onto a low-dimensional space of basis functions properly selected by an adaptive procedure; an affine parametric dependence enabling to perform competitive Offline-Online splitting in the computational procedure; an efficient *and* rigorous a posteriori error estimation on the state, control and adjoint variables as well as on the cost functional. The reduction scheme is applied to several numerical tests confirming the theoretical results and demonstrating the efficiency of the proposed technique. Moreover an application to an (idealized) inverse problem in haemodynamics is discussed, showing the versatility and potentiality of the method in tackling parametrized optimal control problems that could arise in a a broad variety of application contexts.

Sommario

L'obiettivo di questa tesi è quello di sviluppare, analizzare e implementare metodi numerici efficienti per la risoluzione di problemi di controllo ottimale per equazioni differenziali alle derivate parziali parametrizzate. L'interesse verso questo tipo di problemi nasce in diverse aree applicative, in particolar modo in tutti i casi in cui si mira non soltanto alla simulazione numerica del sistema considerato, ma anche all'ottimizzazione e al controllo di alcune sue funzionalità in corrispondenza delle possibili diverse configurazioni fisiche e/o geometriche del sistema stesso, identificate da un insieme di parametri. Poiché la risoluzione numerica di un problema di controllo ottimo richiede ingenti risorse computazionali già nel caso non-parametrico, a maggior ragione, quando dobbiamo ripetere il processo di ottimizzazione in corrispondenza di diverse configurazioni del sistema, o quando, data una certa configurazione, vogliamo ottenere la soluzione in tempi rapidi, lo sforzo computazionale richiesto può risultare incredibilmente elevato, e dunque spesso insostenibile. Si è voluto quindi sviluppare un nuovo approccio basato su di una opportuna tecnica di *riduzione d'ordine del modello* – il metodo a basi ridotte – che permetta di risolvere in maniera rapida, accurata ed affidabile questo tipo di problemi.

Il metodo proposto permette di trattare problemi di controllo ottimale governati da equazioni ellittiche coercive e dalle equazioni di Stokes. In entrambi i casi, il problema di controllo viene innanzitutto riformulato come un opportuno problema punto-sella, in modo da poter sfruttare l'analogia con la struttura delle equazioni di Stokes parametrizzate, per le quali il metodo a basi ridotte è già stato ampiamente sviluppato. Successivamente, viene proposto uno schema a basi ridotte che consta dei seguenti ingredienti fondamentali: una proiezione di Galerkin su di uno spazio di dimensione ridotta, costituito da funzioni di base opportunamente selezionate attraverso una procedura adattiva; una scomposizione offline/online delle procedure computazionali, che permette un disaccoppiamento del problema tra la fase di generazione della base (ridotta) e la procedura di proiezione; una stima a posteriori efficiente *e* rigorosa dell'errore sulle variabili di stato, controllo e stato aggiunto, così come dell'errore sul funzionale costo. L'approccio a basi ridotte sviluppato viene poi applicato ad alcuni esempi numerici che confermano i risultati teorici dimostrati e ne evidenziano l'efficienza computazionale. Infine viene presentata una possibile applicazione ad un problema inverso in ambito emodinamico, mostrando la versatilità e le potenzialità del metodo nell'affrontare problemi di controllo ottimale parametrizzati riscontrabili in numerosi contesti applicativi.

Contents

Introduction **xvii**

1 An introduction to Optimal Control Problems governed by PDEs **1**

1.1 The Lagrangian formalism 2

1.1.1 Abstract formulation in Banach spaces 2

1.1.2 Application to linear quadratic optimal control problems 4

1.2 Saddle-point formulation 9

1.2.1 Formulation and existence of solutions 9

1.2.2 Galerkin approximation, stability and convergence 11

1.2.3 Quadratic optimization with elliptic coercive equations constraints . . 12

1.2.4 Quadratic optimization with Stokes equations constraints 16

2 Numerical methods: iterative and one-shot approach **21**

2.1 Iterative methods for optimal control problems 23

2.1.1 Iterative methods for unconstrained optimization 24

2.1.2 Iterative optimization methods for optimal control of PDEs 26

2.1.3 Iterative optimization methods as linear solvers for the reduced Hessian 28

2.1.4 A numerical example: test and comparison 30

2.2 One-shot approach: multigrid preconditioning 33

2.2.1 Problems governed by elliptic equations 35

2.2.2 Block diagonal preconditioners 35

2.2.3 Comparison and numerical results 39

2.2.4 Problems governed by the Stokes system: a block diagonal preconditioner 41

2.2.5 Comparision and numerical tests 45

3 Reduced Basis Method for Parametrized Elliptic PDEs **47**

3.1 Problem formulation 49

3.1.1 Truth approximation 50

3.2 Reduced basis approximation 50

3.2.1 Formulation and main features 50

3.2.2 Offline-Online procedure 51

3.2.3 Sampling strategy: the greedy algorithm 53

3.3 A posteriori error estimation 53

3.3.1 Basic ingredients 54

3.3.2 Offline-Online procedure 55

3.3.3 SCM coercivity constant lower bounds 56

3.4 Diffusion-advection-reaction equations in parametrized domains 59

3.4.1 Affine geometrical parametrization 60
3.4.2 Parametrized formulation on a reference domain 61
3.4.3 A Graetz conduction-convection problem 61
3.5 Non-coercive problems: the Stokes equations 64
3.5.1 Problem formulation 64
3.5.2 Reduced basis approximation 67
3.5.3 Algebraic formulation 68
3.5.4 Offline-Online computational procedure and sampling strategies 69
3.5.5 A posteriori error estimation 70
3.5.6 Numerical example: Couette flow 73

4 Reduced Basis Method for Parametrized Elliptic Optimal Control Problems 77
4.1 Problem formulation 78
4.1.1 Truth approximation 81
4.2 The reduced basis approximation 82
4.2.1 Formulation 82
4.2.2 Algebraic formulation 83
4.2.3 Approximation stability: enriched spaces 85
4.2.4 Algebraic formulation with enriched spaces 86
4.2.5 Offline-Online procedure and sampling strategy 87
4.3 A posteriori error estimation 88
4.3.1 Bound for the solution: Babuška framework 88
4.3.2 Offline-Online procedure 90
4.3.3 A posteriori error bound for the cost functional 92
4.4 Numerical examples 93
4.4.1 Test 1: distributed optimal control for the Laplace equation with geometrical parametrization 93
4.4.2 Test 2: distributed optimal control for a Graetz convection-diffusion problem with physical parametrization 97
4.4.3 Test 3: boundary optimal control for a Graetz flow with physical and geometrical parametrization 100

5 Reduced Basis Method for Parametrized Stokes Optimal Control problems 105
5.1 Problem formulation 105
5.1.1 Truth approximation 108
5.2 The reduced basis approximation 109
5.2.1 Formulation 109
5.2.2 Algebraic formulation 111
5.2.3 Offline-Online procedure and sampling strategy 112
5.3 A posteriori error estimation 113
5.4 Preliminary numerical tests 115
5.4.1 Test 1: distributed optimal control for a Couette flow 115
5.4.2 Test 2: boundary optimal control for a Couette flow 118
5.5 Application to an inverse problem in haemodynamics 119
5.5.1 Mathematical modeling 119
5.5.2 Original formulation of the optimal control problem 122

5.5.3 Parametrized formulation in the reference domain 123
5.5.4 Numerical results 124

Conclusions **127**

Appendices **133**

A Classical Variational Methods **133**
A.1 Abstract variational problems 133
A.1.1 Strongly coercive problems 133
A.1.2 Weakly coercive problems 134
A.1.3 Saddle-point problems 134
A.2 Approximation of solutions of variational problems 136
A.2.1 Strongly coercive problems 136
A.2.2 Weakly coercive problems 136
A.2.3 Saddle-point problems 137
A.3 On the computation of discrete stability factors 138
A.3.1 Computation of discrete coercivity constant 138
A.3.2 Computation of discrete inf-sup constant 138

Bibliography **141**

Acknowledgements **149**

List of Tables

2.1 Size of the reduced system related to each level of refinement considered. . . . 32

2.2 Comparison of CPU time and number of iterations using the CG, GMRES and Richardson method. 32

2.3 Comparison of CPU time and number of iterations with $\alpha = 10^{-1}$ using preconditioned CG and GMRES. 33

2.4 Comparison of CPU-time and number of iterations with $\alpha = 10^{-5}$ using CG and GMRES solvers non-preconditioned and preconditioned. 33

2.5 Computational time and number of iterations for different values of α using PCG solver. 34

2.6 Comparison of CPU time and iterations with $\alpha = 10^{-2}$ using GMRES with $\hat{\mathcal{P}}_d$ as preconditioner on the full system, varying the number of V-cycles and smoothing iterations. 39

2.7 Setup of the multigrid routines for the preconditioner $\hat{\mathcal{P}}_d$ and size of the full saddle-point system in correspondence of different levels of refinement. 40

2.8 Comparison of computational time and number of iterations using GMRES preconditioned with $\hat{\mathcal{P}}_d$ and the sparse direct solver. 40

2.9 Comparison of CPU-time and number of iterations using GMRES preconditioned with $\hat{\mathcal{P}}_{d2}$ and $\hat{\mathcal{P}}_{dr}$ and the direct solver on the reduced system. 41

2.10 Comparison of CPU-time and number of iterations using GMRES preconditioned with $\hat{\mathcal{P}}_d$ and the direct solver. 46

List of Figures

1 Parametrized optimal control problem xix

2.1 Schemes of the iterative and one-shot approach for generic optimal control problems, either linear or nonlinear. 23
2.2 Desired state and optimal solutions for $\alpha = 10^{-2}$. 31
2.3 Optimal state and control solutions for $\alpha = 10^{-5}$. 31
2.4 Comparison of residual norms and CPU time using different solvers. 32
2.5 Comparison of number of iterations for different values of α. 41
2.6 Comparison between the solvers discussed in the context of reduced Hessian methods and the one-shot approach. The regularization parameter is fixed to $\alpha = 10^{-3}$. 42
2.7 Desired velocity and optimal solution for $\alpha = 10^{-4}$. 46

3.1 Original domain $\Omega_o(\boldsymbol{\mu})$. 62
3.2 Temperature field $u(\boldsymbol{\mu})$ for different values of the parameters. 63
3.3 Bounds for the coercivity constant and a posteriori error bound for the RB error. 64
3.4 Original domain $\Omega_o(\boldsymbol{\mu})$ for the Couette flow. 74
3.5 Couette flow: representative solution for $\boldsymbol{\mu} = (2, 1.5, 0.3)$. 74
3.6 Couette flow: representative solution for $\boldsymbol{\mu} = (0.6, 1, 0.7)$. 75
3.7 Couette flow: lower bound for the Babuška inf-sup constant $\tilde{\beta}(\boldsymbol{\mu})$. 75
3.8 Couette flow: a posteriori error bound. Comparison between the average computed error and the estimator $\Delta(\boldsymbol{\mu})$, for $N = 1, \cdots, N_{max} = 9$. 75

4.1 Test 1: original domain $\Omega_o(\boldsymbol{\mu})$. 94
4.2 Test 1: on the left the computational domain, on the right representative solution for $\boldsymbol{\mu} = (0.6, 3)$. 95
4.3 Test 1: representative solutions for $\boldsymbol{\mu} = (1.5, 2)$. The value of the cost functional is $J = 0.086$. 96
4.4 Test1: lower bound for the Babuška inf-sup constant $\hat{\beta}(\boldsymbol{\mu})$ as a function of the geometrical parameter μ_1. 96
4.5 Test 1: comparison of Brezzi inf-sup constant $\beta(\boldsymbol{\mu})$ and coercivity constant of the state equation $\tilde{\alpha}(\boldsymbol{\mu})$ for the FE and RB approximations. 97
4.6 Test 1: a posteriori error bounds. Here Ξ_{train} is a sample of size $n_{train} = 1000$ and $N_{max} = 9$. 97
4.7 Test 2: domain Ω. $\hat{\Omega}_1$ and $\hat{\Omega}_2$ are the observation domains. 98
4.8 Test 2: on the left the computational domain, on the right representative solution for $\boldsymbol{\mu} = (5, 1, 2.5)$. 99

4.9 Test 2: representative solutions for $\boldsymbol{\mu} = (15, 0.6, 1.8)$. 99
4.10 Test 2: stability properties. 99
4.11 Test 2: a posteriori error bounds. Here Ξ_{train} is a sample of size $n_{train} = 1000$ and $N_{max} = 17$. 100
4.13 Original domain for Test 3. 101
4.14 Test 3: representative solution for $\boldsymbol{\mu} = (12, 2, 2.5)$. 102
4.15 Test 3: lower bound for the Babuška inf-sup constant $\hat{\beta}(\boldsymbol{\mu})$. 103
4.16 Test 3: comparison of Brezzi inf-sup constant $\beta(\boldsymbol{\mu})$ and coercivity constant of the state equation $\tilde{\alpha}(\boldsymbol{\mu})$ for the FE and RB approximations. 103
4.17 Test 3: a posteriori error bounds. Here Ξ_{train} is a sample of size $n_{train} = 2500$ and $N_{max} = 39$. 103

5.1 Original domain for Test 1. 115
5.2 Test 1: representative solutions for $\boldsymbol{\mu} = (1.7, 1.5)$. 117
5.3 Test 1: stability properties 117
5.4 Test 1: a posteriori error bounds. Here Ξ_{train} is a sample of size $n_{train} = 1000$ and $N_{max} = 15$. 118
5.5 Test 2: a posteriori error bounds. Here Ξ_{train} is a sample of size $n_{train} = 1000$ and $N_{max} = 12$. 119
5.6 An (idealized) example of inverse boundary problem in haemodynamics. . . . 120
5.7 Original domain $\Omega_o(\boldsymbol{\mu})$ of the arterial bifurcation. 120
5.8 Parametrized observation $\boldsymbol{v}_d(\boldsymbol{\mu})$. 122
5.9 Reference domain with `rbMIT` triangulation. 123
5.10 Average computed error and bound between the *truth* FE solution and the RB approximation, for $N = 1, \cdots, N_{max}$. 124
5.11 Average computed error between the *truth* FE solution and the RB approximation. 125
5.12 Representative solution for $\boldsymbol{\mu} = (0.7, \pi/6, \pi/5, 0.73, 2.1, 2.4, 0.25, 1.2)$. 125
5.13 Representative solution for $\boldsymbol{\mu} = (1, \pi/5, \pi/6, 1, 1.7, 2.2, 0.8, 1)$. 126
5.14 Representative solution for $\boldsymbol{\mu} = (1.2, \pi/6, \pi/6, 0.8, 2.5, 2.1, 0.3, 1)$. 126

Introduction

This master thesis aims at the development, analysis and computer implementation of efficient numerical methods for the solution of Optimal Control problems based on parametrized Partial Differential Equations. Our goal is to develop a new approach based on suitable model reduction paradigms for the rapid and reliable solution of complex problems which may occur in several engineering contexts. In fact, often the ultimate goal is not only the numerical simulation of the modelled system, but rather the optimization of some specific performances related to the system (as, for example, the drag forces acting on airfoils), or the optimal control of the underlying process in order to reach a desired state. Since the characterization of a system in terms of physical quantities (like source terms, boundary conditions, material properties) and/or geometrical configuration usually depends on a set of parameters, the system response will be parameter dependent as well, and so will be the optimal control. In all these cases, we are required to solve parametrized optimal control problems, where the prediction of optimal control inputs and the optimization of given output of interests is required for each different value of the parameters.
The numerical solution of an optimal control problem entails large computational costs and may be very time-consuming already in the non-parametric case. Therefore, when performing the optimization process for many different parameter values (*many-query* context) or when, for a given new configuration, we want to compute the solution in a rapid way (*real-time* context), the computational effort may be unacceptably high and, often, unaffordable. For this reason we aim at reducing the complexity of the original problem by means of suitable model order reduction techniques, yet preserving its main features and the same input-output behaviour. The ultimate goal of this thesis is to build a general framework based on the reduced basis method for the efficient numerical solution of high dimension/-complexity optimal control problems.

In order to identify the key features of the problem at hand and the strategy we chose for the resolution, let us clarify what is meant by Optimal Control problem for Parametrized PDEs. In general, an optimal control problem (OCP) consists of:

1. a controlled system, i.e. an input-output process (or relationship),
2. an observation of the output of the controlled system,
3. a control function, which can be seen as an input for the system,
4. an objective to be achieved.

The goal is to find the best input such that the observation of the output satisfies the objective to be achieved. In our abstract framework the controlled system is indeed a physical system modelled by PDEs, on which we have to act in some way, in order to obtain some desired/prescribed functionality. To be more specific, we are interested in situations where [52, 33]:

1. the controlled system is given by a (or a system of) PDE(s), that we denote as $\mathcal{E}(y, u) = 0$, also called the state equation; y is the state variable (*output*) and u is the control variable (*input*). The control variable may represent a forcing term, a boundary condition, an initial condition or even a coefficient in the equation. Given an input u, the input-output relationship $u \to y(u)$ requires the solution of the state equation, which thus represents a constraint of the OCP;

2. the observation is typically a linear mapping of the state variable, say $z(y) = \mathcal{C}y$, being C a suitable (linear or nonlinear) operator; $z(y)$ could be the state variable itself, the restriction of the state variable on a portion of the boundary, or even quantities depending on the derivatives of the state variable;

3. the objective is given by an objective or cost functional $\mathcal{J}(y, u)$, expressed as a function of the observation $z(y)$ and possibly of some target state y_d.

Then the optimal control problem can be expressed as:

find the optimal control u^* and the state $y(u^*)$ such that the cost functional $\mathcal{J}(y, u)$ is minimized subject to $\mathcal{E}(y, u) = 0$. (OCP)

The abstract formulation (OCP) is again very general and comprehensive of many different kinds of problems. In this work we limit to consider to the most typical linear/quadratic case, i.e. optimal control problems where the cost functional is a quadratic functional of the observation while the state equation is a linear (scalar or vectorial, coercive or noncoercive) elliptic PDE.

In the parametrized case, the state equation has a more general form $\mathcal{E}(y(\boldsymbol{\mu}), u(\boldsymbol{\mu}); \boldsymbol{\mu}) = 0$, where $\boldsymbol{\mu} \in \mathcal{D} \subset \mathbb{R}^p$ denotes a p-vector of parameters representing physical or geometrical parameters of interest. The parametrized optimal control problem can be expressed as follows:

given $\boldsymbol{\mu} \in \mathcal{D}$, find the optimal control $u^*(\boldsymbol{\mu})$ and the state $y^*(\boldsymbol{\mu})$ such that the cost functional $\mathcal{J}(y(\boldsymbol{\mu}), u(\boldsymbol{\mu}); \boldsymbol{\mu})$ is minimized subject to $\mathcal{E}(y(\boldsymbol{\mu}), u(\boldsymbol{\mu}); \boldsymbol{\mu}) = 0$. $(OCP_{\boldsymbol{\mu}})$

The repeated solution of the $(OCP_{\boldsymbol{\mu}})$ for many different parameters values is an unaffordable task from a computational point of view. This is why a suitable model order reduction is necessary. From an abstract point of view, the mapping $\boldsymbol{\mu} \mapsto (y(\boldsymbol{\mu}), u(\boldsymbol{\mu}))$ defines a *smooth* and rather *low-dimensional* parametrically induced manifold

$$\mathcal{M} = \{(y(\boldsymbol{\mu}), u(\boldsymbol{\mu})) \in X : \boldsymbol{\mu} \in \mathcal{D}\},$$

where $y(\boldsymbol{\mu})$ and $u(\boldsymbol{\mu})$ are the state and control solutions of $(OCP_{\boldsymbol{\mu}})$ and X is a suitable functional space such that $(y, u) \in X$. Attempting to solve numerically the problem $(OCP_{\boldsymbol{\mu}})$ can be seen as trials to approximate the manifold $\mathcal{M}$. In a classical discretization approach, after introducing an approximation space $X^{\mathcal{N}}$ of (typically very large) dimension $\mathcal{N}$ – e.g. a finite element (FE) space – for every value of the parameters $\boldsymbol{\mu}$ we are supposed to solve the whole optimal control problem in order to compute the solution $(y^{\mathcal{N}}(\boldsymbol{\mu}), u^{\mathcal{N}}(\boldsymbol{\mu}))$. This amounts to constructing a pointwise approximation $\mathcal{M}^{\mathcal{N}}$ of the manifold $\mathcal{M}$, thus ignoring

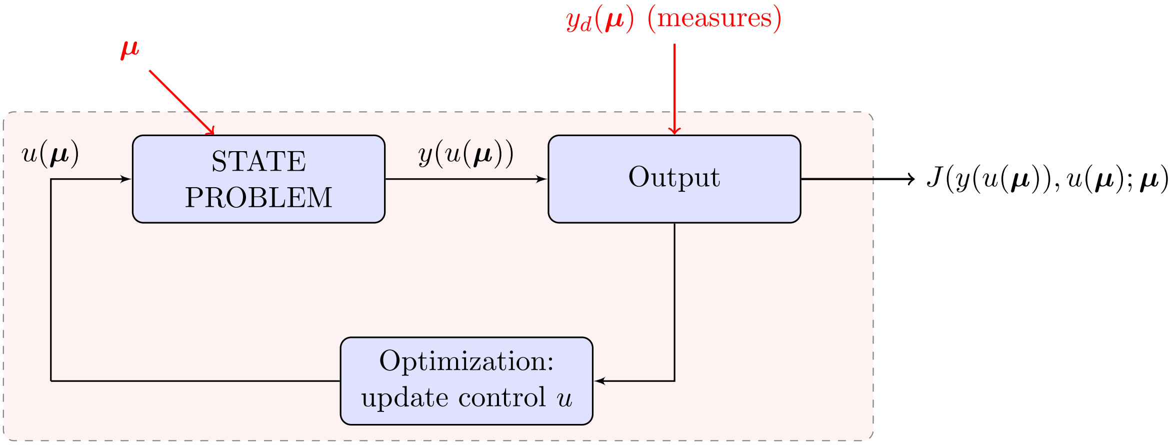

$y_d(\boldsymbol{\mu})$ (measures)
$\boldsymbol{\mu}$
$u(\boldsymbol{\mu})$
STATE PROBLEM
$y(u(\boldsymbol{\mu}))$
Output
$J(y(u(\boldsymbol{\mu})), u(\boldsymbol{\mu}); \boldsymbol{\mu})$
Optimization: update control u

Figure 1: Parametrized optimal control problem

the possibly *smooth* relation between parameters and solutions. In other words, this kind of generic approximation spaces are unnecessarily rich and hence unnecessarily expensive within the parametric framework.

A reduced (basis) approach is premised upon a classical finite element method (for example) and consists in a low-order approximation of the *truth* manifold $\mathcal{M}^{\mathcal{N}}$, based on two stages: in the former we sample some parameters values in the space $\mathcal{D}$ and compute the corresponding FE solutions, which can be seen as snapshots of the *truth* manifold $\mathcal{M}^{\mathcal{N}}$; in the latter we build a lower-dimensional approximation $\mathcal{M}^N$ of the *truth* manifold, by means of a suitable interpolation procedure (a Galerkin projection) of the precomputed snapshots. In particular, the main ingredients of the reduced basis (RB) methods [61, 66, 77] are the following ones:

(i) a rapidly convergent global approximation (Galerkin projection) onto a space spanned by solution of the original problem at some selected parameters value;

(ii) a rigorous a posteriori error estimation procedures which provides inexpensive yet sharp bounds for the error between the RB and the *truth* solution. The a posteriori error estimation is crucial for the certification of the method as well as for the design of the sampling (Greedy) procedure used for the construction of the reduced basis;

(iii) an Offline/Online computational procedures, i.e. a splitting between a time-consuming and parameter independent Offline stage and an inexpensive Online calculation for each new input/output evaluation.

Some reduced order strategies have already been used for the efficient solution of parametrized optimal control problems; beyond reduced basis methods, also proper orthogonal decomposition [41] applied to OCPs has been widely analyzed. However, several fundamental aspects have still to be analyzed and extended to more general cases. In particular:

1. the solution of OCPs within the RB framework was faced firstly by Ito and Ravindran [48, 50, 49]. More recently, the RB method has been applied to parametrized linear-quadratic advection-diffusion optimal control with environmental applications in [67], in the acoustic context [88], in the parametrized linear-quadratic parabolic optimal

control problem, in both the unconstrained [17] and constrained case [18]. However, in all these works the control variable is low-dimensional (e.g. a set of scalars, that could be treated themselves as parameters). We aim at developing a reduced framework that enables to handle with general control functions, i.e. infinite dimensional distributed and/or boundary control functions.

2. An efficient and rigorous a posteriori error estimation, necessary both for constructing the reduced order model and measuring its accuracy, is still missing for a large class of optimal control problems. The main difficulty stands in the construction of rigorous error bounds not just for the reduced state variable, but also for the reduced cost functional and the reduced control. For example, the a posteriori estimators for the error in the cost functional and in the control variable proposed in [17, 18] show to be efficient in practice but unfortunately lack of rigorousness, whereas the estimator proposed in [88] is proved to be rigorous but not efficient. Only recently in [32] an efficient and rigorous estimator has been proposed in the case of a scalar constant control function. In this thesis we aim at developing both efficient and rigorous a posteriori error bounds in order to estimate, simultaneously, the errors in the optimal control, the state variable and the cost functional and by considering a very general control function.

3. Concerning the numerical solution of optimal control problems, all the previous works have exploited an iterative algorithm for tackling the optimization stage. This may often require very large computational costs as it involves the repeated evaluation of the state solution and the gradient of the cost functional. We propose to apply an alternative approach, based on the so-called *one-shot method* [7, 23, 71, 86], to pursue a computational speedup during both the very expensive *offline* construction of the reduced basis and the many *online* evaluations of the solution of the optimal control problem for different configurations of the system.

The goal of this thesis is to apply the RB method to the class of parametrized optimal control problems featuring quadratic cost functionals and linear constraints with high-dimensional control variable, and to develop rigorous and efficiently evaluable a posteriori error bounds for the errors in the optimal control, the state variable and the cost functional. In particular, with reference to the basic ingredients of the RB method previously introduced, we point out that:

(i) in our approach, the reduced basis is made of optimal solutions of the original problem, hence the computation of each basis function requires to solve the FE *truth* model, for which we need efficient methods. Therefore a preliminary part of the thesis will be devoted to the analysis of numerical methods for the solution of optimal control problems, focusing on the so-called one-shot approach (in contrast to the iterative or reduced-space approaches);

(ii) in order to provide the a posteriori error estimations for the optimal control problem, we take advantage of the theory developed for Stokes-type problems [76, 74, 79, 75], by recasting it in the framework of mixed variational (also called saddle-point) problems;

(iii) we rely on the the affine parameter dependence assumption, which provides the possibility to extract the parameter dependent components from our operators and thus exploit an Offline/Online computational procedure, but results are easily extendible to non-affine parametric dependency.

With respect to the previously introduced background, we have organized the work with the following structure.

Chapter 1 In this first part we briefly introduce the theory of optimal control problems governed by PDEs. We first present the Lagrangian approach for general nonlinear problems in Banach spaces, focusing on the case of linear-quadratic problems in Hilbert spaces. Then we study linear-quadratic problems in the framework of mixed variational problems, proving some results on the well-posedness of the problem, as well as for its Galerkin approximation, in the case of elliptic equations and Stokes system as constraints.

Chapter 2 In this chapter we discuss some (finite element) numerical methods to solve linear-quadratic optimal control problems. In particular we introduce the two most popular strategies to solve numerically this kind of problems: iterative (or reduced Hessian) methods and one-shot (or full-system) methods.

Chapter 3 In this chapter we provide an overview of reduced basis approximation and a posteriori error estimation methods for parametrized elliptic coercive equations and Stokes system. In particular we describe the main ingredients of the method that will be extended to parametrized optimal control problems in the following chapters.

Chapter 4 We provide here a reduced basis framework for the efficient solution of parametrized linear-quadratic optimal control problems governed by coercive second-order elliptic PDEs. We prove the well-posedness of the RB approximation and we develop an efficient and rigorous a posteriori error estimation. The reduction scheme is applied to some numerical examples that confirm the theoretical results and demonstrate the efficiency of the method.

Chapter 5 We extend the methodology developed in Chapter 4 to the case of Stokes constraints, providing some numerical tests and an application to an inverse problem in haemodynamics.

The simulations reported in this work have been carried out by means of MATLAB®[57] software using the `MLife` (finite element) library [82] and an enhanced version (co-developed at CMCS, EPFL) of the `rbMIT`© (reduced basis) library [44, 61]. A considerable part of the work has been devoted to extend these two libraries in order to handle optimal control problems.

This work has been carried out in the Chair of Modeling and Scientific Computing (CMCS) of the École Polytechnique Fédérale de Lausanne (CH), in the framework of the international exchange European Union program Erasmus.

Milano and Lausanne, December 2011.

Chapter 1

An introduction to Optimal Control Problems governed by PDEs

The classical approach to optimal control for PDEs is based on the theory developed by J.L. Lions [52, 53], which provides existence and uniqueness results for optimal control problems described by elliptic, parabolic, hyperbolic and mixed PDEs. An alternative, more general approach, is based on the Lagrangian formalism (see for instance [40, 89, 33, 47]): the control problem is recast in a constrained minimization problem, for which a Lagrangian functional is defined. The optimum, if it does exist, is a stationary 'point' of the Lagrangian functional. The Lagrangian also allows to easily provide a posteriori error estimates for approximated optimal control problems as, e.g., discussed in [5]. The solution of the Optimal Control Problem can be characterized by the Karush-Kuhn-Tucker conditions, which in our case yield a system of PDEs – state equation, adjoint equation and optimality condition. In the case of optimal control problems here considered, featuring quadratic cost functionals and linear constraints, this coupled system, also called optimality conditions system, features a saddle-point structure. In view of the application of the reduced basis method, we are interested in highlighting this structure, hence we study linear-quadratic optimal control problems in the framework of mixed variational problems [14], as already done in [84, 38, 63, 34, 35].

In Section 1.1 we introduce the Lagrangian formalism, we first present existence results and optimality conditions for general nonlinear problems in Banach spaces, then we focus on the case of linear-quadratic problems in Hilbert spaces, with applications to the optimal control of coercive elliptic equations and Stokes system. In Section 1.2 we study linear-quadratic problems in Hilbert spaces in the framework of mixed variational problems, proving some results on the well-posedness of the problem, as well as for its Galerkin approximation, in the case of elliptic equations and Stokes system as constraints.

To avoid misunderstandings we clarify that analysing linear-quadratic optimal control problems using either the Lagrangian or the mixed variational framework leads exactly to the same results. In fact, being the two methods equivalent and representing simply different formalisms, usually the choice of the one to use is simply a question of taste. However, in some contexts, using the saddle-point formalism can be more convenient since it permits to highlight some analogies with different problems sharing the same abstract structure. In our case we are interested in highlighting the saddle-point structure and its properties in view of the application of the reduced-basis method, other contexts in which the mixed variational framework is used are the design of efficient preconditioner for the discrete problem, as we will discuss in Chapter 2, and the development of a particular class of numerical methods known as penalty methods (see [34, 35]).

1.1 The Lagrangian formalism

The following issues must be addressed when we study an optimal control problem: (i) existence and uniqueness of a solution, (ii) optimality conditions, (iii) numerical approximation and optimization algorithms. While the last step will be addressed in the next Chapter, here we discuss the first two points. In particular we first present existence results and optimality conditions for general nonlinear problems in Banach spaces, then we focus on the case of linear-quadratic problems in Hilbert spaces, with applications to the optimal control of elliptic equations and Stokes system. For the main results and theorems we refer to the monographs [40, 33], while the basics on Banach and Hilbert spaces can be found in any book on linear functional analysis, e.g. [91, 80].

As regards the notation, given a Banach space X, X^* will denote its dual, i.e. the space of linear functionals on X, while $\langle \cdot, \cdot \rangle_{X,X^*}$ will denote the duality pairing of X^* and X.

1.1.1 Abstract formulation in Banach spaces

Let U, Y be reflexive Banach spaces and Z be a Banach space, $U_{ad} \subset U$ and $Y_{ad} \subset Y$, consider the following nonlinear optimization problem

$$\min_{(y,u)\in Y\times U} J(y,u) \quad \text{subject to} \quad \mathcal{E}(y,u) = 0, \quad u \in U_{ad},\, y \in Y_{ad}, \tag{1.1.1}$$

where the functional $J : Y \times U \to \mathbb{R}$ and the state equation $\mathcal{E} : Y \times U \to Z$. Note that when $U_{ad} \subsetneq U$ the problem is said to be control-constrained, similarly when $Y_{ad} \subsetneq Y$ the problem is said to be state-constrained. Denote the feasible set by

$$F_{ad} = \{(y,u) \in Y \times U : (y,u) \in Y_{ad} \times U_{ad},\, \mathcal{E}(y,u) = 0\}.$$

The following assumptions are required to prove existence results:

1. U_{ad} is convex, bounded and closed;
2. Y_{ad} is convex and closed, such that (1.1.1) has a feasible point;
3. the state equation $\mathcal{E}(y,u) = 0$ has a bounded solution operator $u \in U_{ad} \mapsto y(u) \in Y$;
4. $(y,u) \in Y \times U \mapsto \mathcal{E}(y,u) \in Z$ is continuous under weak convergence;
5. J is weakly lower semicontinuous.

We can state the following existence theorem (see [40, Sec. 1.5.2] for the proof)

Theorem 1.1. *Let assumptions (1)-(5) hold. Then problem (1.1.1) has an optimal solution* $(\bar{y}, \bar{u})$.

We now give a result about first order necessary condition in the simpler case of control-constrained problem,

$$\min_{(y,u)\in Y\times U} J(y,u) \quad \text{subject to} \quad \mathcal{E}(y,u)=0, \quad u\in U_{ad}. \tag{1.1.2}$$

We make the following assumptions

(i) $U_{ad}\subseteq U$ is nonempty, convex and closed;

(ii) $J : Y\times U\to\mathbb{R}$ and $\mathcal{E} : Y\times U\to Z$ are continuously Fréchet differentiable and U, Y, Z are Banach spaces;

(iii) for all $u\in V$ in a neighborhood $V\subset U$ of U_{ad}, the state equation $\mathcal{E}(y,u)=0$ has a unique solution $y=y(u)\in Y$;

(iv) $\mathcal{E}_y(y(u),u)\in\mathcal{L}(Y,Z)$ has a bounded inverse for all $u\in V\supset U_{ad}$.

Remark 1.1. Assumption (iv) and the implicit function theorem ensure that the solution operator of the state equation, i.e. $V\ni u\mapsto y(u)\in Y$, is continuously differentiable. We denote with $y'(u)$ the corresponding derivative.

Now let us introduce the reduced problem

$$\min_{u\in U}\hat{J}(u) \quad \text{subject to} \;\; u\in U_{ad}, \tag{1.1.3}$$

with the reduced objective functional $\hat{J}(u)=J(y(u),u)$. Holding the assumptions (i)-(iv) the formulations (1.1.2) and (1.1.3) are equivalent. We have the following result [40, Th. 1.48]

Theorem 1.2. *Let assumptions (i)-(iv) hold. If $\bar{u}$ is a local solution of the reduced problem (1.1.3) then $\bar{u}\in U_{ad}$ satisfies the variational inequality (called optimality condition or minimum principle)*

$$\langle\hat{J}'(\bar{u}), v-\bar{u}\rangle_{U^*,U}\geq 0, \qquad \forall v\in U_{ad}. \tag{1.1.4}$$

Note that to make operative the optimality condition (1.1.4) we need an explicit expression for the derivative of the reduced cost functional $\hat{J}(u)$. There are at least two ways: the sensitivity approach and the adjoint approach. We discuss here the adjoint approach with Lagrangian functional. Let us define the Lagrangian functional $\mathcal{L} : Y\times U\times Z^*\to\mathbb{R}$,

$$\mathcal{L}(y,u,p)=J(y,u)+\langle p,\mathcal{E}(y,u)\rangle_{Z^*,Z}, \tag{1.1.5}$$

where the variable p is called Lagrange multiplier or adjoint variable. We want to express the reduced cost functional in terms of $\mathcal{L}$: for arbitrary $p\in Z^*$

$$\hat{J}(u)=J(y(u),u)=J(y(u),u)+\langle p,\mathcal{E}(y(u),u)\rangle_{Z^*,Z}=\mathcal{L}(y(u),u,p). \tag{1.1.6}$$

Differentiating we obtain

$$\langle\hat{J}'(u),v\rangle_{U^*,U}=\langle\mathcal{L}_y(y(u),u,p),y'(u)v\rangle_{Y^*,Y}+\langle\mathcal{L}_u(y(u),u,p),v\rangle_{U^*,U},$$

and, to get an explicit expression for the derivative of $\hat{J}'(u)$, we choose $p=p(u)$ such that $\mathcal{L}_y(y(u),u,p)=0$, this way

$$\hat{J}'(u)=\mathcal{L}_u(y(u),u,p(u))=J_u(y(u),u)+\mathcal{E}_u(y(u),u)^*p(u). \tag{1.1.7}$$

Of course $p(u)$ is still unknown, however from (1.1.6) we see that the condition $\mathcal{L}_y(y(u),u,p) = 0$ implies that $p(u)$ is the solution of

$$\begin{aligned}\langle \mathcal{L}_y(y,u,p), z\rangle_{Y^*,Y} &= \langle J_y(y,u), z\rangle_{Y^*,Y} + \langle p, \mathcal{E}_y(y,u)z\rangle_{Z^*,Z}\\ &= \langle J_y(y,u) + \mathcal{E}_y(y,u)^* p, z\rangle_{Y^*,Y} = 0, \qquad \forall z \in Y.\end{aligned}$$

Therefore $p \in Z^*$ is the solution of

$$\mathcal{E}_y(y(u),u)^* p = -J_y(y(u),u), \tag{1.1.8}$$

the so-called adjoint equation. Having an explicit representation of $\hat{J}'(\bar{u})$ we can state the following corollary to Theorem 1.2, which provides necessary first order optimality conditions.

Corollary 1.1. *Let $(\bar{y},\bar{u})$ be an optimal solution of the problem (1.1.2) and let assumptions (i)-(iv) hold. Then there exists an adjoint state (or Lagrange multiplier) $\bar{p} \in Z^*$ such that the following optimality conditions hold*

$$\begin{cases} \mathcal{E}(\bar{y},\bar{u}) = 0,\\ \mathcal{E}_y(\bar{y},\bar{u})^*\bar{p} = -J_y(\bar{y},\bar{u}),\\ \langle J_u(\bar{y},\bar{u}) + \mathcal{E}_u(\bar{y},\bar{u})^*\bar{p}, v-\bar{u}\rangle_{U^*,U} \geq 0, \quad \forall v \in U_{ad}.\end{cases} \tag{1.1.9}$$

Equivalently using the Lagrangian functional

$$\begin{cases} \langle \mathcal{L}_p(\bar{y},\bar{u},\bar{p}), q\rangle_{Z,Z^*} = 0 & \forall q \in Z^*\\ \langle \mathcal{L}_y(\bar{y},\bar{u},\bar{p}), z\rangle_{Y^*,Y} = 0 & \forall z \in Y\\ \langle \mathcal{L}_u(\bar{y},\bar{u},\bar{p}), v-\bar{u}\rangle_{U^*,U} \geq 0 & \forall v \in U_{ad}.\end{cases} \tag{1.1.10}$$

With respect to the functional J and the state equation $\mathcal{E}$ the variational formulation (1.1.10) becomes

$$\begin{cases} \langle \mathcal{E}(\bar{y},\bar{u}), q\rangle_{Z,Z^*} = 0 & \forall q \in Z^*\\ \langle \mathcal{E}_y(\bar{y},\bar{u})^*\bar{p} + J_y(\bar{y},\bar{u}), z\rangle_{Y^*,Y} = 0 & \forall z \in Y\\ \langle J_u(\bar{y},\bar{u}) + \mathcal{E}_u(\bar{y},\bar{u})^*\bar{p}, v-\bar{u}\rangle_{U^*,U} \geq 0 & \forall v \in U_{ad}.\end{cases} \tag{1.1.11}$$

Note that when $U_{ad} \equiv U$ the variational inequality reduces to an equation,

$$\langle J_u(\bar{y},\bar{u}) + \mathcal{E}_u(\bar{y},\bar{u})^*\bar{p}, v\rangle_{U^*,U} = 0, \qquad \forall v \in U.$$

In this case the system of equations (1.1.9) can be viewed as the Euler-Lagrange system of the Lagrangian functional. In fact the solutions of equations (1.1.10) are the stationary points of $\mathcal{L}(\cdot,\cdot,\cdot)$, i.e.

$$\nabla\mathcal{L}(\bar{y},\bar{u},\bar{p})[z,v,q] = 0 \qquad \forall (z,v,q) \in Y \times U \times Z^*.$$

1.1.2 Application to linear quadratic optimal control problems

We consider the following unconstrained linear-quadratic optimal control problem

$$\begin{aligned}\min_{(y,u)\in Y\times U} \quad & J(y,u) = \frac{1}{2}\|\mathcal{Q}y - y_d\|^2_{\mathcal{Z}} + \frac{\alpha}{2}\|u\|^2_U\\ \text{subject to} \quad & Ay = Cu + f\end{aligned} \tag{1.1.12}$$

where $Y, U, \mathcal{Z}$ are Hilbert spaces, $Z = Y^*$, $y_d \in \mathcal{Z}$, $f \in Y^*$, we assume that $A \in \mathcal{L}(Y, Y^*)$ has a bounded inverse (i.e. $A^{-1} \in \mathcal{L}(Y^*, Y)$), $C \in \mathcal{L}(U, Y^*)$ and $\mathcal{Q} \in \mathcal{L}(Y, \mathcal{Z})$. The problem is quadratic in the functional, linear in the state equation and unconstrained because both $U_{ad} \equiv U$ and $Y_{ad} \equiv Y$. With these assumptions Theorem 1.1 holds, moreover if $\alpha > 0$ the solution is unique [40]. Since Hilbert spaces are reflexive we have the identification $Y = Y^{**}$, whence $Z^* = Y$, also $U = U^*$ and $\mathcal{Z}^* = \mathcal{Z}$. Then

$$\mathcal{E}(y,u) = Ay - Cu - f, \qquad \mathcal{E}_y(y,u) = A, \qquad \mathcal{E}_u(y,u) = C,$$

the Lagrangian functional reads

$$\mathcal{L}(y,u,p) = \frac{1}{2}(\mathcal{Q}y - y_d, \mathcal{Q}y - y_d)_{\mathcal{Z}} + \frac{\alpha}{2}(u,u)_U + \langle p, Ay - Cu - f\rangle_{Y,Y^*}$$

and assumptions (i)-(iv) hold, in particular the hypothesis that $A \in \mathcal{L}(Y, Y^*)$ has a bounded inverse ensures the fulfilment of assumptions (iii)-(iv). To derive the optimality system we have to compute the derivative of $\mathcal{L}(\cdot,\cdot,\cdot)$ with respect to (y,u,p):

$$\begin{aligned}
\langle \mathcal{L}_y(y,u,p), z\rangle_{Y^*,Y} &= (\mathcal{Q}y - y_d, \mathcal{Q}z)_{\mathcal{Z}} + \langle p, Az\rangle_{Y,Y^*} \\
&= \langle \mathcal{Q}^*(\mathcal{Q}y - y_d) + A^*p, z\rangle_{Y^*,Y} && \forall z \in Y, \\
\langle \mathcal{L}_u(y,u,p), v\rangle_{U^*,U} &= (\mathcal{L}_u(y,u,p), v)_U = \alpha(u,v)_U - \langle p, Cv\rangle_{Y,Y^*} \\
&= (\alpha u - C^*p, v)_U && \forall v \in U, \\
\langle \mathcal{L}_p(y,u,p), q\rangle_{Y^*,Y} &= \langle Ay - Cu - f, q\rangle_{Y^*,Y} && \forall q \in Y,
\end{aligned}$$

where $A^* \in \mathcal{L}(Y^*, Y)$ is the adjoint operator of A, $C^* \in \mathcal{L}(Y, U)$ is the adjoint operator of C and $\mathcal{Q}^* \in \mathcal{L}(\mathcal{Z}, Y^*)$ is the adjoint operator of $\mathcal{Q}$. Thus the optimality system takes the form (omitting the bar over the optimal variable)

$$\begin{cases} A^*p = -\mathcal{Q}^*(\mathcal{Q}y - y_d) \\ \alpha u - C^*p = 0 \\ Ay = Cu + f, \end{cases} \tag{1.1.13}$$

or, equivalently, using the variational formulation

$$\begin{cases} \langle \mathcal{Q}^*(\mathcal{Q}y - y_d) + A^*p, z\rangle_{Y^*,Y} = 0 & \forall z \in Y, \\ (\alpha u - C^*p, v)_U = 0 & \forall v \in U, \\ \langle Ay - Cu - f, q\rangle_{Y^*,Y} = 0 & \forall q \in Y. \end{cases} \tag{1.1.14}$$

Thanks to the Riesz representation theorem we can also obtain an explicit expression for the derivative of the cost functional: by imposing

$$(\nabla J(u), v)_U = \langle \hat{J}'(u), v\rangle_{U^*,U} = \langle \mathcal{L}_u(y,u,p), v\rangle_{U^*,U}, \qquad \forall v \in U,$$

we get $\nabla J(u) = -C^*p + \alpha u$, denoting with $\nabla J(u) \in U$ the Riesz representation of $J'(u) \in U^*$.

We now provide some concrete examples of linear quadratic optimal control problems, specifying for each case the operator A, C and $\mathcal{Q}$ introduced in the general formulation (1.1.12). In order to prove the well-posedness of these problems, the crucial point is to check the fulfilment of the hypotheses made above, in particular to check that $A \in \mathcal{L}(Y, Y^*)$ has a bounded inverse (i.e. $A^{-1} \in \mathcal{L}(Y^*, Y)$), $C \in \mathcal{L}(U, Y^*)$ and $\mathcal{Q} \in \mathcal{L}(Y, \mathcal{Z})$.

Example 1.1 (Distributed control of the Laplace equation)**.** We consider the distributed optimal control of the Laplace equation

$$\begin{aligned} &\text{minimize} \quad J(y,u) = \frac{1}{2}\int_\Omega (y-y_d)^2\, d\Omega + \frac{\alpha}{2}\int_\Omega u^2\, d\Omega, \\ &\text{s.t.} \quad \begin{cases} -\Delta y = u + f & \text{in } \Omega, \\ \quad y = 0 & \text{on } \partial\Omega. \end{cases} \end{aligned} \tag{1.1.15}$$

Let $\Omega \subset \mathbb{R}^d$ $(d = 1,2,3)$ be an open bounded domain with Lipschitz boundary $\partial\Omega$, $y_d \in L^2(\Omega)$ and $f \in L^2(\Omega)$ given, $Y = H^1_0(\Omega)$, $U = L^2(\Omega)$, $\mathcal{Z} = L^2(\Omega)$, the dual space of Y is $Y^* = H^{-1}(\Omega)$. The weak formulation of the state equation reads:

$$\text{find } y \in Y \text{ s.t.} \quad a(y,q) = c(u,q) + (f,q)_{L^2}, \qquad \forall q \in Y,$$

where the bilinear forms $a : Y \times Y \to \mathbb{R}$ and $c : U \times Y \to \mathbb{R}$ are given by

$$a(y,q) = \int_\Omega \nabla y \cdot \nabla q\, d\Omega, \qquad c(u,q) = \int_\Omega uq\, d\Omega.$$

Fixed $u \in U$, applying Lax-Milgram lemma (Lemma A.1) follows the well-posedness of the state equation. We can identify the operators introduced in the formulation (1.1.12): $A \in \mathcal{L}(Y,Y^*)$ is the operator induced by the bilinear form a, i.e. $\langle Ay,q\rangle_{Y^*,Y} = a(y,q)$, $C \in \mathcal{L}(U,Y^*)$ is the operator induced by the bilinear form c, i.e. $\langle Cu,q\rangle_{Y^*,Y} = c(u,q)$, the observation operator $\mathcal{Q} \in \mathcal{L}(Y,\mathcal{Z})$ is the identity operator, i.e. $\mathcal{Q}y = y$. Then, instead of computing the adjoint operators and substituting in the optimality conditions (1.1.13), we express the Lagrangian functional using the bilinear form defined above:

$$\begin{aligned} \mathcal{L}(y,u,p) &= \frac{1}{2}(\mathcal{Q}y - y_d, \mathcal{Q}y - y_d)_{\mathcal{Z}} + \frac{\alpha}{2}(u,u)_U + \langle p, Ay - Cu - f\rangle_{Y,Y^*} \\ &= \frac{1}{2}(y-y_d, y-y_d)_{L^2} + \frac{\alpha}{2}(u,u)_{L^2} + a(y,p) - c(u,p) - (f,q)_{L^2}. \end{aligned}$$

Setting equal to zero the derivatives of $\mathcal{L}$ with respect to (y,u,p) we obtain the optimality conditions system

$$\begin{cases} a(z,p) = -(y-y_d, z)_{L^2} & \forall z \in Y, \\ (\alpha u, v)_{L^2} = c(v,p) & \forall v \in U, \\ a(y,q) = c(u,q) + (f,q)_{L^2} & \forall q \in Y. \end{cases} \tag{1.1.16}$$

To recover the formulations (1.1.13) or (1.1.14) it is sufficient to compute the adjoint operators; for A^*, noting that the bilinear form a is symmetric, we obtain

$$\langle A^*p, z\rangle_{Y^*,Y} = \langle Az, p\rangle_{Y^*,Y} = a(z,p) = a(p,z) = \langle Ap, z\rangle_{Y^*,Y}, \quad \forall p,z \in Y,$$

which implies $A^* = A$. With the same argument we obtain also $C^* = C$ and $\mathcal{Q}^* = \mathcal{Q}$. Moreover the adjoint equation in (1.1.16) is just the weak formulation of

$$\begin{cases} -\Delta p = y_d - y & \text{in } \Omega, \\ \quad p = 0 & \text{on } \partial\Omega, \end{cases}$$

while the gradient of the cost functional is given by $\nabla J(u) = -p + \alpha u$.

Remark 1.2. Rewriting the optimality system (1.1.16) in the following form

$$\begin{cases} (y,z)_{L^2} & & +\,a(z,p) & = (y_d,z)_{L^2} & \forall z \in Y, \\ & \alpha(u,v)_{L^2} & -\,c(v,p) & = 0 & \forall v \in U, \\ a(y,q) & -\,c(u,q) & & = (f,q)_{L^2} & \forall q \in Y. \end{cases} \tag{1.1.17}$$

the saddle-point structure of this coupled system of PDEs becomes quite evident.

Example 1.2 (Boundary control of an advection-diffusion-reaction equation)**.** Let us now consider a more involved elliptic optimal control with boundary control and observation on a portion of the domain,

$$\begin{aligned} &\text{minimize} \quad J(y,u) = \frac{1}{2}\int_{\Omega_o} (y-y_d)^2\, d\Omega + \frac{\alpha}{2}\int_{\Gamma_n} u^2\, d\Gamma, \\ &\text{s.t.} \quad \begin{cases} -\mathrm{div}(\varepsilon \nabla y) + \beta \cdot \nabla y + \sigma y = f & \text{in } \Omega, \\ y = 0 & \text{on } \Gamma_d, \\ -\varepsilon \nabla y \cdot \boldsymbol{n} = u & \text{on } \Gamma_n. \end{cases} \end{aligned} \tag{1.1.18}$$

Let Ω be a bounded Lipschitz domain with boundary $\partial\Omega$ such that $\Gamma_d \cap \Gamma_n = \emptyset$ and $\Gamma_d \cup \Gamma_n = \partial\Omega$, while the source term $f \in L^2(\Omega)$, the diffusivity $\varepsilon(x) \in L^\infty(\Omega)$ with $\varepsilon(x) \geq \varepsilon_0 > 0$ a.e. in Ω, the reaction $\sigma \in L^\infty(\Omega)$ and the advective field $\beta \in [L^\infty(\Omega)]^2$ with $\mathrm{div}\,\beta \in L^\infty(\Omega)$ are assigned functions. Moreover let us assume that $-\frac{1}{2}\mathrm{div}\beta + \sigma \geq r_0 \geq 0$ a.e. in Ω for a suitable r_0 and $\Gamma_n \subset \{x \in \partial\Omega : \beta(x) \cdot \mathbf{n}(x) \geq 0\}$, being $\boldsymbol{n}$ the unit outward normal vector on $\partial\Omega$. The observation domain Ω_o is an open subset of Ω, $y_d \in L^2(\Omega)$ is given, $Y = H^1_{\Gamma_d}(\Omega)$, $U = L^2(\Gamma_n)$ and $\mathcal{Z} = L^2(\Omega)$, the dual space of Y is $Y^* = H^{-1}(\Omega)$. The weak formulation of the state equation reads

$$\text{find } y \in Y \;\text{ s.t.} \qquad a(y,q) = c(u,q) + (f,q)_{L^2}, \qquad \forall q \in Y,$$

where the bilinear forms $a : Y \times Y \to \mathbb{R}$ and $c : U \times Y \to \mathbb{R}$ are given by

$$a(y,q) = \int_\Omega \big\{\varepsilon \nabla y \cdot \nabla q + \beta \cdot \nabla y q + \sigma y q\big\}\, d\Omega, \qquad c(u,q) = \int_{\Gamma_n} uq\, d\Gamma.$$

Fixed $u \in U$, applying Lax-Milgram Lemma follows the well-posedness of the state equation (e.g [65]). The Lagrangian functional reads

$$\mathcal{L}(y,u,p) = \frac{1}{2}(y-y_d, y-y_d)_{L^2(\Omega_o)} + \frac{\alpha}{2}(u,u)_{\Gamma_n} + a(y,p) - c(u,p) - (f,p)_{L^2}.$$

Setting equal to zero the derivatives of $\mathcal{L}$ with respect to (y,u,p) we obtain the optimality conditions system

$$\begin{cases} a(z,p) = -(\chi_{\Omega_o} y - y_d, z)_{L^2} & \forall z \in Y, \\ (\alpha u, v)_{\Gamma_n} = c(v,p) & \forall v \in U, \\ a(y,q) = c(u,q) + (f,q)_{L^2} & \forall q \in Y. \end{cases} \tag{1.1.19}$$

Note that this time the bilinear form $a(\cdot,\cdot)$ is not symmetric, hence $A^* \neq A$, and the strong formulation of the adjoint equation reads

$$\begin{cases} -\mathrm{div}(\varepsilon \nabla p) - \mathrm{div}(\beta p) + \sigma p = \chi_{\Omega_o}(y_d - y) & \text{in } \Omega, \\ p = 0 & \text{on } \Gamma_d, \\ \varepsilon \nabla p \cdot \mathbf{n} + \beta \cdot \mathbf{n} p = 0 & \text{on } \Gamma_n. \end{cases} \tag{1.1.20}$$

The derivative of the cost functional is given by $\nabla J(u) = -p_{|\Gamma_n} + \alpha u$.

Example 1.3 (Distributed control of the Stokes equations)**.** Let us now consider a distributed optimal control problem for the Stokes equations,

$$\begin{aligned}&\text{minimize} \quad J(\boldsymbol{v},\boldsymbol{u}) = \frac{1}{2}\int_\Omega |\boldsymbol{v}-\boldsymbol{v}_d|^2\,dx + \frac{\alpha}{2}\int_\Omega |\boldsymbol{u}|^2\,dx,\\ &\text{s.t.} \quad \begin{cases} -\nu\Delta\boldsymbol{v} + \nabla p = \boldsymbol{u} & \text{in } \Omega,\\ \operatorname{div}\boldsymbol{v} = 0 & \text{in } \Omega,\\ \boldsymbol{v} = 0 & \text{on } \partial\Omega,\end{cases}\end{aligned} \tag{1.1.21}$$

where $\Omega \subset \mathbb{R}^2$ is an open bounded domain, ν is the kinematic viscosity (a given constant), $\boldsymbol{v} \in [H_0^1(\Omega)]^2$ is the velocity and $p \in L_0^2(\Omega) = \{r \in L^2(\Omega) : \int_\Omega r = 0\}$ the pressure. The functional settings is as follows: $Y = [H_0^1(\Omega)]^2 \times L_0^2(\Omega)$ is the space of the state variables $(\boldsymbol{v}, p)$, while $U = [L^2(\Omega)]^2$ is the space of the control variable $\boldsymbol{u}$; the dual space of Y is given by $Y^* = [H^{-1}(\Omega)]^2 \times L^2(\Omega)$. In order to give the weak formulation of the state equation let us define the following bilinear forms:

$$a(\boldsymbol{v},\boldsymbol{\xi}) = \nu\int_\Omega \nabla\boldsymbol{v}\cdot\nabla\boldsymbol{\xi}\,d\Omega, \qquad b(\boldsymbol{v},p) = -\int_\Omega p\nabla\cdot\boldsymbol{v}\,d\Omega, \qquad c(\boldsymbol{u},\boldsymbol{\xi}) = \int_\Omega \boldsymbol{u}\cdot\boldsymbol{\xi}\,d\Omega.$$

The weak formulation reads: find $(\boldsymbol{v}, p) \in Y$ such that

$$\begin{cases} a(\boldsymbol{v},\boldsymbol{\xi}) + b(\boldsymbol{\xi},p) = c(\boldsymbol{u},\boldsymbol{\xi}) & \forall\boldsymbol{\xi} \in [H_0^1(\Omega)]^2,\\ b(\boldsymbol{v},\tau) = 0 & \forall\tau \in L^2(\Omega).\end{cases} \tag{1.1.22}$$

Given $\boldsymbol{u} \in U$, it is well known that the saddle-point problem (1.1.22) satisfies the assumptions of (Brezzi) Theorem A.2 (see, e.g., [14, 69]) and hence admits a unique solution that depends continosly with respect to the data. This means that the state operator is continuous from Y to Y^* and has a bounded inverse. We can proceed formally writing the Lagrangian functional

$$\mathcal{L}(\boldsymbol{v},p,\boldsymbol{u},\boldsymbol{w},q) = \frac{1}{2}(\boldsymbol{v}-\boldsymbol{v}_d,\boldsymbol{v}-\boldsymbol{v}_d)_{L^2} + \frac{\alpha}{2}(\boldsymbol{u},\boldsymbol{u})_{L^2} + a(\boldsymbol{v},\boldsymbol{w}) + b(\boldsymbol{w},p) + b(\boldsymbol{v},q) - c(\boldsymbol{u},\boldsymbol{w}),$$

requiring the derivatives of $\mathcal{L}$ to vanish we obtain the optimality conditions system

$$\begin{cases} a(\boldsymbol{v},\boldsymbol{\xi}) + b(\boldsymbol{\xi},p) = c(\boldsymbol{u},\boldsymbol{\xi}) & \forall\boldsymbol{\xi} \in [H_0^1(\Omega)]^2,\\ b(\boldsymbol{v},\tau) = 0 & \forall\tau \in L_0^2(\Omega),\\ a(\boldsymbol{\varphi},\boldsymbol{w}) + b(\boldsymbol{\varphi},q) = (\boldsymbol{v}-\boldsymbol{v}_d,\boldsymbol{\varphi})_{L^2} & \forall\boldsymbol{\varphi} \in [H_0^1(\Omega)]^2,\\ b(\boldsymbol{w},\pi) = 0 & \forall\pi \in L_0^2(\Omega),\\ \alpha(\boldsymbol{u},\boldsymbol{\lambda})_{L^2} = c(\boldsymbol{\lambda},\boldsymbol{w}) & \forall\boldsymbol{\lambda} \in [L^2(\Omega)]^2.\end{cases} \tag{1.1.23}$$

By counter-integrating by parts the third and fourth equations in (1.1.23) we obtain the strong formulation of the adjoint problem

$$\begin{cases} -\nu\Delta\boldsymbol{w} + \nabla q = \boldsymbol{v}_d - \boldsymbol{v} & \text{in } \Omega,\\ \operatorname{div}\boldsymbol{w} = 0 & \text{in } \Omega,\\ \boldsymbol{w} = 0 & \text{on } \partial\Omega,\end{cases}$$

while the derivative of the cost functional is given by $\nabla J(\boldsymbol{u}) = -\boldsymbol{w} + \alpha\boldsymbol{u}$.

Remark 1.3. As in Example 1.1 let us highlight the saddle-point structure of the coupled system (1.1.23)

$$\left\{\begin{array}{lllllll}
(\boldsymbol{v},\boldsymbol{\varphi})_{L^2} & & & +a(\boldsymbol{\varphi},\boldsymbol{w}) & +\,b(\boldsymbol{\varphi},q) & =(\boldsymbol{v}_d,\boldsymbol{\varphi})_{L^2} & \forall\boldsymbol{\varphi}\in[H^1_0(\Omega)]^2,\\
 & & & b(\boldsymbol{w},\pi) & & =0 & \forall\pi\in L^2_0(\Omega),\\
 & & \alpha(\boldsymbol{u},\boldsymbol{\lambda})_{L^2} & -c(\boldsymbol{\lambda},\boldsymbol{w}) & & =0 & \forall\boldsymbol{\lambda}\in[H^1_0(\Omega)]^2,\\
a(\boldsymbol{v},\boldsymbol{\xi}) & +\,b(\boldsymbol{\xi},p) & -c(\boldsymbol{u},\boldsymbol{\xi}) & & & =0 & \forall\boldsymbol{\xi}\in[H^1_0(\Omega)]^2,\\
b(\boldsymbol{v},\tau) & & & & & =0 & \forall\tau\in L^2_0(\Omega).
\end{array}\right.$$

1.2 Saddle-point formulation

In the previous section, discussing some examples, we noted that, at least for the problems considered, the optimality conditions system features a saddle-point structure. Here, we want to investigate more in-depth this property, in particular we want to show that this structure is not a particular feature of the examples considered, but it is common to every linear-quadratic optimal control problem. The idea is to formulate in a slightly different way the problem, more precisely in the form

$$\begin{cases}
\min\ \mathcal{J}(\underline{x})=\dfrac{1}{2}\mathcal{A}(\underline{x},\underline{x})-\langle\underline{F},\underline{x}\rangle, & \text{subject to}\\
B\underline{x}=G \qquad \text{in } Q^*, &
\end{cases}$$

where we denote $\underline{x}=(y,u)$, being y the state variable and u the control variable. Then, using the theory of saddle-point problems, we shall show the equivalence between the minimization problem stated above and the following: find $(\underline{x},p)\in X\times Q$ such that

$$\begin{cases}
\mathcal{A}(\underline{x},\underline{w})+\mathcal{B}(\underline{w},p)=\langle\underline{F},\underline{w}\rangle & \forall\underline{w}\in X,\\
\mathcal{B}(\underline{x},q)=\langle G,q\rangle & \forall q\in Q,
\end{cases}$$

the last one being simply the optimality conditions system (1.1.13), with p the Lagrange multiplier associated to the constraint $B\underline{x}=G$. To prove the equivalence between the two problems we will have to guarantee the fulfilment of the well known hypotheses of Brezzi theorem [14], hypotheses that could be reinterpreted in terms of the assumptions made in the previous section.

This formulation has been already used with different objectives in [84, 38, 63, 93, 83], a unified presentation can be found in [34, 35], where it has been developed in view of the application of penalty methods.

1.2.1 Formulation and existence of solutions

Let X and Q be Hilbert spaces and X^*, Q^* their dual spaces respectively, we consider the continuous bilinear form $\mathcal{A}(\cdot,\cdot)$ on $X\times X$, the continuous bilinear form $\mathcal{B}(\cdot,\cdot)$ on $X\times Q$, the functional $F\in X^*$ and $G\in Q^*$. Note that the bilinear form $\mathcal{A}(\cdot,\cdot)$ defines a linear continuous operator $A: X\to X^*$ by

$$\langle A\underline{x},\underline{w}\rangle_{X^*,X}=\mathcal{A}(\underline{x},\underline{w}),\qquad \forall\underline{x},\underline{w}\in X,$$

while the bilinear form $\mathcal{B}(\cdot,\cdot)$ defines a linear continuous operator $B : X \to Q^*$ and its transpose $B^t : Q \to X^*$ by

$$\langle B\underline{w}, q\rangle_{Q^*,Q} = \langle \underline{w}, B^t q\rangle_{X,X^*} = \mathcal{B}(\underline{w}, q), \qquad \forall \underline{w} \in X, q \in Q.$$

We consider the following saddle-point problem (also called mixed variational problem): find $(\underline{x}, p) \in X \times Q$ such that

$$\begin{cases} \mathcal{A}(\underline{x}, \underline{w}) + \mathcal{B}(\underline{w}, p) = \langle \underline{F}, \underline{w}\rangle & \forall \underline{w} \in X, \\ \mathcal{B}(\underline{x}, q) = \langle G, q\rangle & \forall q \in Q. \end{cases} \tag{1.2.1}$$

The problem (1.2.1) can also be written as

$$\begin{cases} A\underline{x} + B^t p = \underline{F} & \text{in } X^*, \\ B\underline{x} = G & \text{in } Q^*. \end{cases}$$

Let us define the subspace of X

$$X_0 = \{\underline{w} \in X \colon \mathcal{B}(\underline{w}, q) = 0 \quad \forall q \in Q\} \subset X, \tag{1.2.2}$$

consisting of those elements $\underline{w} \in X$ that belong to the null space of the operator induced by the bilinear form $\mathcal{B}(\cdot,\cdot)$. The existence, uniqueness and stability of a solution to this saddle-point problem is well-established by the well known Brezzi theorem, see Theorem A.2 in Appendix A. We recall here the two main assumptions of the theorem: we require the bilinear form $\mathcal{A}(\cdot,\cdot)$ to be weakly coercive on X_0, i.e. there exists a constant $\alpha_0 > 0$ such that

$$\inf_{\underline{x}\in X_0} \sup_{\underline{w}\in X_0} \frac{\mathcal{A}(\underline{x}, \underline{w})}{\|\underline{x}\|_X \|\underline{w}\|_X} \geq \alpha_0 \qquad \text{and} \qquad \inf_{\underline{w}\in X_0} \sup_{\underline{x}\in X_0} \frac{\mathcal{A}(\underline{x}, \underline{w})}{\|\underline{x}\|_X \|\underline{w}\|_X} > 0,$$

and the bilinear form $\mathcal{B}(\cdot,\cdot)$ to satisfy the inf-sup condition, i.e. there exists a constant $\beta_0 > 0$ such that

$$\beta = \inf_{q\in Q} \sup_{\underline{w}\in X} \frac{\mathcal{B}(\underline{w}, q)}{\|\underline{w}\|_X \|q\|_Q} \geq \beta_0. \tag{1.2.3}$$

Holding these assumptions the solution to problem (1.2.1) is unique, moreover there exists a constant $C > 0$ such that the solution satisfies the following a priori estimate:

$$\|\underline{x}\|_X + \|p\|_Q \leq C\big(\|F\|_{X^*} + \|G\|_{Q^*}\big). \tag{1.2.4}$$

The following proposition (see [14, Remark 1.3] and [35, Prop. 1.7]) clarifies the relation between mixed variational principles and constrained optimization problems.

Proposition 1.1. *Assume that the hypotheses of Theorem A.2 hold. Assume further that the bilinear form $\mathcal{A}(\cdot,\cdot)$ is symmetric, nonnegative, and strongly coercive on X_0, i.e., that*

$$\mathcal{A}(\underline{x}, \underline{w}) = \mathcal{A}(\underline{w}, \underline{x}), \qquad \mathcal{A}(\underline{x}, \underline{x}) \geq 0 \qquad \forall \underline{x}, \underline{w} \in X \tag{1.2.5}$$

and

$$\mathcal{A}(\underline{x}, \underline{x}) \geq \alpha_0 \|\underline{x}\|_X \qquad \forall \underline{x} \in X_0, \tag{1.2.6}$$

for a suitable $\alpha_0 > 0$. Then, the problem (1.2.1) is equivalent to the constrained optimization problem

$$\begin{cases} \min \mathcal{J}(\underline{x}) = \dfrac{1}{2}\mathcal{A}(\underline{x}, \underline{x}) - \langle \underline{F}, \underline{x}\rangle, & \textit{subject to} \\ B\underline{x} = G \quad \textit{in } Q^*. \end{cases} \tag{1.2.7}$$

Remark 1.1. If we introduce the Lagrangian functional

$$\mathcal{L}(\underline{w}, q) = \mathcal{J}(\underline{w}) + \mathcal{B}(\underline{w}, q) - \langle F, \underline{w} \rangle_{X^*,X},$$

then the constrained minimization problem (1.2.7) is equivalent to the unconstrained optimization problem of finding saddle points $(\underline{x}, p)$ in $X \times Q$ of the Lagrangian functional. These saddle points may be found by solving the optimality system (1.2.1).

Remark 1.2. As we aim to exploit the saddle-point structure of quadratic optimization problems with linear constraints, our plan for the next sections will be: (i) formulate the problem in the form (1.2.7) and (ii) prove that the assumptions of Proposition 1.1 hold in order to (iii) gain the well-posedness of the saddle-point system (1.2.1), i.e. the optimality conditions sistem.

However, before applying these results to concrete optimization problems, let us introduce the Galerkin approximation of (1.2.1) and discuss its analysis.

1.2.2 Galerkin approximation, stability and convergence

To introduce the Galerkin approximation of the saddle-point problem (1.2.1), we consider two families of finite dimensional subspaces $X^{\mathcal{N}}$ and $Q^{\mathcal{N}}$ of the space X and Q. In this work these discrete spaces will be finite element piecewise polynomial spaces, however they could be either polynomial spaces (in spectral methods) or spectral element spaces, see [65]. The Galerkin-FE approximation of problem (1.2.1) has the following form: find $(\underline{x}^{\mathcal{N}}, p^{\mathcal{N}}) \in X^{\mathcal{N}} \times Q^{\mathcal{N}}$ such that

$$\begin{cases} \mathcal{A}(\underline{x}^{\mathcal{N}}, \underline{w}) + \mathcal{B}(\underline{w}, p^{\mathcal{N}}) = \langle \underline{F}, \underline{w} \rangle & \forall \underline{w} \in X^{\mathcal{N}}, \\ \mathcal{B}(\underline{x}^{\mathcal{N}}, q) = \langle G, q \rangle & \forall q \in Q^{\mathcal{N}}. \end{cases} \tag{1.2.8}$$

Note that this is equivalent to the optimality system for the minimization of the functional $\mathcal{J}(\cdot)$ over $X^{\mathcal{N}}$ subject to $\mathcal{B}(\underline{x}^{\mathcal{N}}, q) = \langle G, q \rangle$, $\forall q \in Q^{\mathcal{N}}$. Let

$$X_0^{\mathcal{N}} = \{\underline{w} \in X^{\mathcal{N}} : \mathcal{B}(\underline{w}, q) = 0 \quad \forall q \in Q^{\mathcal{N}}\} \subset X^{\mathcal{N}}, \tag{1.2.9}$$

in general $X_0^{\mathcal{N}} \not\subset X_0$ even though $X^{\mathcal{N}} \subset$ and $Q^{\mathcal{N}} \subset Q$ so that the assumption of strong coercivity of $\mathcal{A}(\cdot,\cdot)$ and the inf-sup condition on $\mathcal{B}(\cdot,\cdot)$ may not be satisfied. If $X^{\mathcal{N}}$ and $Q^{\mathcal{N}}$ are such that the latter conditions hold, then Theorem A.4 provides the discrete counterpart of Theorem A.2, ensuring existence, uniqueness and stability for problem (1.2.8). In particular, in the case considered in Proposition 1.1, we assume that the bilinear form $\mathcal{A}(\cdot,\cdot)$ be strongly coercive on $X_0^{\mathcal{N}}$, i.e. there exists a constant $\alpha^{\mathcal{N}} > 0$ such that

$$\mathcal{A}(\underline{w}, \underline{w}) \geq \alpha^{\mathcal{N}} \|\underline{w}\|_X^2 \qquad \forall \underline{w} \in X_0^{\mathcal{N}}. \tag{1.2.10}$$

Moreover, we suppose that the bilinear form $\mathcal{B}(\cdot,\cdot)$ satisfies the following inf-sup condition: there exists $\beta_0^{\mathcal{N}} > 0$ such that

$$\beta^{\mathcal{N}} = \inf_{q \in Q^{\mathcal{N}}} \sup_{\underline{w} \in X^{\mathcal{N}}} \frac{\mathcal{B}(\underline{w}, q)}{\|\underline{w}\|_X \|q\|_Q} \geq \beta_0^{\mathcal{N}}. \tag{1.2.11}$$

Holding these assumptions the solution to problem (1.2.8) is unique and there exists a constant $C > 0$ such that the solution satisfies the following a priori estimate:

$$\|\underline{x}^{\mathcal{N}}\|_X + \|p^{\mathcal{N}}\|_Q \leq C\big(\|F\|_{X^*} + \|G\|_{Q^*}\big). \tag{1.2.12}$$

If we can prove that the discrete counterparts of the assumptions of Proposition 1.1 are verified, we obtain the well-posedness of the discrete problem (1.2.8). Moreover, holding these assumptions, if $(\underline{x}, p) \in X \times Q$ denotes the unique solution of (1.2.1), the following optimal error estimate hold

$$\|\underline{x} - \underline{x}^{\mathcal{N}}\|_X + \|p - p^{\mathcal{N}}\|_Q \leq C \Big(\inf_{\underline{w}^{\mathcal{N}} \in X^{\mathcal{N}}} \|\underline{x} - \underline{w}^{\mathcal{N}}\|_X + \inf_{q^{\mathcal{N}} \in Q^{\mathcal{N}}} \|p - q^{\mathcal{N}}\|_Q \Big), \tag{1.2.13}$$

where C is independent of $\mathcal{N}$.

Let us now investigate the structure of the algebraic system associated to the Galerkin approximation (1.2.8). Let $\mathcal{N}_X = \dim X^{\mathcal{N}}$, $\mathcal{N}_Q = \dim Q^{\mathcal{N}}$, denote with

$$\{\underline{\varphi}_j \in X^{\mathcal{N}}\}_{j=1}^{\mathcal{N}_X}, \qquad \{\phi_k \in Q^{\mathcal{N}}\}_{k=1}^{\mathcal{N}_Q},$$

the basis functions of the spaces $X^{\mathcal{N}}$, $Q^{\mathcal{N}}$, respectively. Let us expand the discrete solutions $\underline{x}^{\mathcal{N}}$ and $p^{\mathcal{N}}$ with respect to such bases,

$$\underline{x}^{\mathcal{N}} = \sum_{j=1}^{\mathcal{N}_X} x_j \underline{\varphi}_j(x), \qquad p^{\mathcal{N}} = \sum_{k=1}^{\mathcal{N}_Q} p_k \phi_k(x),$$

by choosing as test functions in (1.2.8) the same basis functions we obtain the following linear system

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{F} \\ \mathbf{G} \end{pmatrix}, \tag{1.2.14}$$

where $A_{ij} = \mathcal{A}(\underline{\varphi}_i, \underline{\varphi}_j)$, $B_{km} = \mathcal{B}(\underline{\varphi}_m, \phi_k)$, $(\mathbf{F})_i = \langle F, \varphi_i \rangle$, $(\mathbf{G})_k = \langle G, \phi_k \rangle$, $(\mathbf{x})_i = x_i$ and $(\mathbf{p})_k = p_k$. The $(\mathcal{N}_X + \mathcal{N}_Q) \times (\mathcal{N}_X + \mathcal{N}_Q)$ matrix in (1.2.14) is symmetric, indefinite and uniformly invertible with respect to $\mathcal{N}$ thanks to the assumptions (1.2.10) and (1.2.11). Note that, since this linear system arise from the discretization of the optimization problem (1.2.7), the block B contains the PDE operator acting as constraint, while the block A comes from the discretization of the functional $\mathcal{J}(\cdot)$.

Remark 1.3. In algebraic form the constrained optimization problem (1.2.7) can be formulated as:

$$\text{minimize } \frac{1}{2}\mathbf{x}^T A \mathbf{x} - \boldsymbol{F}^T \mathbf{x} \qquad \text{subject to } B\mathbf{x} = \boldsymbol{G}. \tag{1.2.15}$$

1.2.3 Quadratic optimization with elliptic coercive equations constraints

We now apply the results of Section 1.2 to quadratic optimization problems constrained by elliptic coercive PDEs. We identify (y, u) with the variable $\underline{x}$ of Section 1.2, where y denote the state variable and u the control variable.

Let $\Omega \subset \mathbb{R}^d$ $(d = 1, 2, 3)$ be an open bounded domain with boundary $\partial\Omega$ such that $\Gamma_D \cap \Gamma_N = \emptyset$ and $\Gamma_D \cup \Gamma_N = \partial\Omega$. We define the functional spaces Y (state space) and Q (adjoint space) such that $H_0^1(\Omega) \subset Y \subset H^1(\Omega)$ and $Q = Y$, respectively; the control space is $U = L^2(\omega)$, where ω can be the whole domain, a subdomain or a boundary. Moreover let $\mathcal{Z}$ be the observation space: typically if we observe on a portion of the domain, say $\hat{\Omega}$, then

$\mathcal{Z} = L^2(\hat{\Omega})$, while if the observation is on a portion of the boundary $\hat{\Gamma}$, then $\mathcal{Z} = L^2(\hat{\Gamma})$. We consider the following optimal control problem

$$\begin{cases} \min\limits_{u \in U} \; J(y,u) = \dfrac{1}{2}\|\mathcal{Q}y - y_d\|^2_{\mathcal{Z}} + \dfrac{\alpha}{2} n(u,u), & \text{subject to} \\ a(y,q) = c(u,q) + \langle G, q\rangle_{Q^*,Q} \qquad \forall q \in Q, & \end{cases} \tag{1.2.16}$$

where $\alpha > 0$ is a given constant, $\mathcal{Q} \in \mathcal{L}(Y, \mathcal{Z})$ and $y_d \in \mathcal{Z}$ a given function. Expressing the constraint equation in weak form permits us to consider different kinds of problems, e.g. distributed or boundary control

$$\begin{cases} \tilde{\mathcal{A}}y = f + u & \text{in } \Omega, \\ y = g_D & \text{on } \Gamma_D, \\ \partial_{\tilde{\mathcal{A}}} y = g_N & \text{on } \Gamma_N, \end{cases} \qquad \begin{cases} \tilde{\mathcal{A}}y = f & \text{in } \Omega, \\ y = g_D & \text{on } \Gamma_D, \\ \partial_{\tilde{\mathcal{A}}} y = u & \text{on } \Gamma_N, \end{cases}$$

where $\tilde{\mathcal{A}} \in \mathcal{L}(Y, Y^*)$ is the elliptic operator associated with the bilinear form $a(\cdot,\cdot)$, i.e. $\langle \tilde{\mathcal{A}}\varphi, \psi\rangle_{Y^*,Q} = a(\varphi,\psi)$, $f \in Y^*$ and the bilinear form $c(\cdot,\cdot)$ is given, respectively, by

$$c(u,q) = \int_\Omega uq \, d\Omega, \qquad c(u,q) = \int_{\Gamma_N} uq \, d\Gamma.$$

Note that the source term $G(q)$ in the weak form of the state equation takes in account also for the non-homogeneous Dirichlet boundary conditions, i.e.

$$\langle G, q\rangle_{Q^*,Q} = \int_\Omega fq \, d\Omega - a(R_g, q),$$

being $R_g \in H^1(\Omega)$ a lift function such that $R_g|_{\Gamma_D} = g_D$.

In order to formulate problem (1.2.16) in the form (1.2.7) and to prove the fulfilment of the hypotheses of Proposition 1.1, let us assume that $\|\cdot\|_U = n(u,u)$ and require the following assumptions (recalling that $Y \equiv Q$):

(i) the bilinear form $a : Y \times Q \to \mathbb{R}$ is bounded and strongly coercive, i.e. there exist two positive constants $\tilde{\gamma}_1$ and $\tilde{\alpha}$ such that

$$a(z,q) \le \tilde{\gamma}_1 \|z\|_Y \|q\|_Q \quad \forall z, \in Y, q \in Q \qquad \text{and} \qquad a(z,z) \ge \tilde{\alpha}\|z\|^2_Y \quad \forall z \in Y \equiv Q,$$

(ii) the bilinear form $c : U \times Q \to \mathbb{R}$ is symmetric and bounded, i.e. there exists a positive constant $\tilde{\gamma}_2$ such that

$$c(u,q) \le \tilde{\gamma}_2 \|u\|_U \|q\|_Y \quad \forall u \in U, \, q \in Q$$

(iii) the bilinear form $n : U \times U \to \mathbb{R}$ is symmetric, bounded and coercive, i.e. there exist two positive constants $\tilde{\gamma}_n$ and $\tilde{\alpha}_n$ such that

$$n(u,v) \le \tilde{\gamma}_n \|u\|_U \|v\|_U \quad \forall u, v \in U \qquad \text{and} \qquad n(u,u) \ge \tilde{\alpha}_n \|u\|^2_U \quad \forall u \in U.$$

Let us denote with $X = Y \times U$ the product space between the *state space* Y and the *control space* U, equipped with the scalar product $(\underline{x}, \underline{w})_X = (y,z)_Y + (u,v)_U$, being $\underline{x} = (y,u) \in X$ and $\underline{w} = (z,v) \in X$. We now define the *aggregated* bilinear forms $\mathcal{A}(\cdot,\cdot)$ and $\mathcal{B}(\cdot,\cdot)$

Definition 1.1. Let $\underline{x} = (y, u) \in X$, $\underline{w} = (z, v) \in X$, the bilinear form $\mathcal{A}(\cdot,\cdot)\colon X \times X \to \mathbb{R}$ is defined as follows

$$\mathcal{A}(\underline{x}, \underline{w}) := (\mathcal{Q}y, \mathcal{Q}z)_{\mathcal{Z}} + \alpha n(u, v). \tag{1.2.17}$$

Definition 1.2. Let $\underline{w} = (z, v) \in X$, $q \in Q$, the bilinear form $\mathcal{B}(\cdot,\cdot)\colon X \times Q \to \mathbb{R}$ is defined as follows

$$\mathcal{B}(\underline{w}, q) := a(z, q) - c(v, q). \tag{1.2.18}$$

Moreover we define the functional $\underline{F} = (\mathcal{Q}^*\Lambda_{\mathcal{Z}} y_d, 0) \in X^*$, where $\Lambda_{\mathcal{Z}}\colon \mathcal{Z} \to \mathcal{Z}^*$ is the Riesz isomorphism and $\mathcal{Q}^* \in \mathcal{L}(\mathcal{Z}^*, Y^*)$ is the adjoint of the operator $\mathcal{Q}$. We can now re-formulate the optimal control problem (1.2.16) as

$$\begin{cases} \min \mathcal{J}(\underline{x}) = \dfrac{1}{2}\mathcal{A}(\underline{x}, \underline{x}) - \langle \underline{F}, \underline{x} \rangle_{X^*,X}, & \text{subject to} \\ \mathcal{B}(\underline{x}, q) = \langle G, q \rangle_{Q^*,Q} \qquad \forall q \in Q. & \end{cases} \tag{1.2.19}$$

If we can prove that the assumptions of Proposition 1.1 hold, then we obtain the equivalence between the control problem (1.2.16) and the following saddle-point problem: find $(\underline{x}, p) \in X \times Q$ such that

$$\begin{cases} \mathcal{A}(\underline{x}, \underline{w}) + \mathcal{B}(\underline{w}, p) = \langle \underline{F}, \underline{w} \rangle & \forall \underline{w} \in X, \\ \mathcal{B}(\underline{x}, q) = \langle G, q \rangle & \forall q \in Q. \end{cases} \tag{1.2.20}$$

Moreover, holding those assumptions, the problem stated above is well posed.

Lemma 1.1. *The bilinear forms $\mathcal{A}(\cdot,\cdot)$ and $\mathcal{B}(\cdot,\cdot)$ satisfy the assumptions of Proposition 1.1.*

Proof. We denote with $\|\mathcal{Q}\| = \|\mathcal{Q}\|_{\mathcal{L}(Y,\mathcal{Z})}$, the norm of the observation operator. We will mostly make use of the hypotheses (i)-(iii) on the bilinear forms $a(\cdot,\cdot)$, $c(\cdot,\cdot)$ and $n(\cdot,\cdot)$.

1. Continuity of the bilinear form $\mathcal{A}(\cdot,\cdot)$ on $X \times X$:

$$\begin{aligned} |\mathcal{A}(\underline{x}, \underline{w})| &\le \|\mathcal{Q}y\|_{\mathcal{Z}}\|\mathcal{Q}z\|_{\mathcal{Z}} + \alpha\|u\|_U\|v\|_U \le \|\mathcal{Q}\|^2\|y\|_Y\|z\|_Y + \alpha\|u\|_U\|v\|_U \\ &\le \big(\|\mathcal{Q}\|^2 + \alpha\big)\|\underline{x}\|_X\|\underline{w}\|_X. \end{aligned}$$

2. Strong coercivity of the bilinear form $\mathcal{A}(\cdot,\cdot)$ on X_0: let $\underline{w} = (z, v) \in X_0$, i.e.

$$\mathcal{B}(\underline{w}, q) = 0 \quad \forall q \in Q \qquad \Longleftrightarrow \qquad a(z, q) = c(v, q) \quad \forall q \in Q,$$

thanks to hypothesis (i), using Lax-Milgram lemma we have $\|v\|_U \ge \frac{\tilde{\alpha}}{\tilde{\gamma}_2}\|z\|_Y$, thus

$$\begin{aligned} \mathcal{A}(\underline{w}, \underline{w}) = \|\mathcal{Q}z\|^2_{\mathcal{Z}} + \alpha\|v\|^2_U &\ge \|\mathcal{Q}z\|^2_{\mathcal{Z}} + \frac{\alpha}{2}\|v\|^2_U + \frac{\alpha}{2}\|v\|^2_U \\ &\ge \|\mathcal{Q}z\|^2_{\mathcal{Z}} + \frac{\alpha\tilde{\alpha}^2}{2\tilde{\gamma}_2^2}\|z\|^2_Y + \frac{\alpha}{2}\|v\|^2_U \\ &\ge \frac{\alpha}{2}\max\Big\{1, \frac{\tilde{\alpha}^2}{\tilde{\gamma}_2^2}\Big\}\big(\|z\|^2_Y + \|v\|^2_U\big) = \alpha_0\,\|\underline{w}\|^2_X. \end{aligned}$$

3. Continuity of the bilinear form $\mathcal{B}(\cdot,\cdot)$ on $X \times Q$:

$$|\mathcal{B}(\underline{w}, q)| \le |a(z, q)| + |c(v, q)| \le \tilde{\gamma}_1\|z\|_Y\|q\|_Q + \tilde{\gamma}_2\|v\|_U\|q\|_Q \le \big(\tilde{\gamma}_1 + \tilde{\gamma}_2\big)\|\underline{w}\|_X\|q\|_Q.$$

4. Inf-sup condition for the bilinear form $\mathcal{B}(\cdot,\cdot)$:

$$\sup_{0\neq \underline{w}\in X}\frac{\mathcal{B}(\underline{w},q)}{\|\underline{w}\|_X}=\sup_{0\neq(z,v)\in Y\times U}\frac{a(z,q)-c(v,q)}{(\|z\|_Y^2+\|v\|_U^2)^{1/2}}\underset{(z,v)=(q,0)}{\geq}\frac{a(q,q)}{\|q\|_Y}\geq\tilde{\alpha}\|q\|_Y=\tilde{\alpha}\|q\|_Q.$$

The symmetry and non-negativity of $\mathcal{A}(\cdot,\cdot)$ are trivial. □

Proposition 1.2. *Let the assumptions (i)-(iii) hold, then the optimal control problem (1.2.19) has a unique solution $(\underline{x},p)\in X\times Q$ given by the saddle-point problem (1.2.20).*

Proof. The result follows from Theorem A.2, Proposition 1.1 and Lemma 1.1. □

Remark 1.4. Using the definitions of the bilinear forms $\mathcal{A}(\cdot,\cdot)$ and $\mathcal{B}(\cdot,\cdot)$ it is easy to obtain the *state-adjoint-optimality conditions* formulation from the saddle-point formulation (1.2.20):

state equation:	$a(y,q)=c(u,q)$	$\forall q\in Q,$
adjoint equation:	$a(z,p)=(y_d-\mathcal{Q}y,\mathcal{Q}z)_{\mathcal{Z}}$	$\forall z\in Y,$
optimality condition:	$c(v,p)=\alpha n(u,v)$	$\forall v\in U.$

Let us now introduce the Galerkin approximation of the saddle-point problem (1.2.20). We consider two families of finite dimensional subspaces $X^{\mathcal{N}}$ and $Q^{\mathcal{N}}$ of the space X and Q, note that we are implicitly considering two spaces $Y^{\mathcal{N}}\subset Y$ and $U^{\mathcal{N}}\subset U$ such that $X^{\mathcal{N}}=Y^{\mathcal{N}}\times U^{\mathcal{N}}$, with $Y^{\mathcal{N}}\equiv Q^{\mathcal{N}}$. The Galerkin-FE approximation of problem (1.2.1) has the following form: find $(\underline{x}^{\mathcal{N}},p^{\mathcal{N}})\in X^{\mathcal{N}}\times Q^{\mathcal{N}}$ such that

$$\begin{cases}\mathcal{A}(\underline{x}^{\mathcal{N}},\underline{w})+\mathcal{B}(\underline{w},p^{\mathcal{N}})=\langle\underline{F},\underline{w}\rangle & \forall\underline{w}\in X^{\mathcal{N}},\\ \mathcal{B}(\underline{x}^{\mathcal{N}},q)=0 & \forall q\in Q^{\mathcal{N}}.\end{cases}\tag{1.2.22}$$

To guarantee the well-posedness of problem (1.2.22) we have to check the fulfilment of the assumptions (1.2.10) and (1.2.11), i.e. the strong coercivity of $\mathcal{A}(\cdot,\cdot)$ on $X_0^{\mathcal{N}}$ and the discrete inf-sup condition on $\mathcal{B}(\cdot,\cdot)$.

Lemma 1.2. *Assume that $Y^{\mathcal{N}}\equiv Q^{\mathcal{N}}$. Then the bilinear forms $\mathcal{A}(\cdot,\cdot)$ and $\mathcal{B}(\cdot,\cdot)$ satisfy the assumptions of Theorem A.4.*

Proof. The crucial point is to check that assumptions (i)-(iii) continue to hold in the discrete case. In particular it is sufficient to note that the continuity property of the bilinear form $a(\cdot,\cdot)$ with respect to the discrete subspaces $Y^{\mathcal{N}}\times Q^{\mathcal{N}}$ is inherited from the continuity property that holds on the parents spaces. Likewise, provided $Y^{\mathcal{N}}\equiv Q^{\mathcal{N}}$, the strong coercivity property on $Y^{\mathcal{N}}\subset Y$ automatically follows from the strong coercivity property with respect to Y. The same arguments apply to the bilinear forms $c(\cdot,\cdot)$ and $n(\cdot,\cdot)$. Now we can repeat exactly the same arguments as in the proof of Lemma 1.1. □

Proposition 1.3. *Let the assumptions (i)-(iii) hold and assume that $Y^{\mathcal{N}}\equiv Q^{\mathcal{N}}$. Then the saddle-point problem (1.2.22) has a unique solution $(\underline{x}^{\mathcal{N}},p^{\mathcal{N}})\in X^{\mathcal{N}}\times Q^{\mathcal{N}}$.*

Proof. The result follows from Theorem A.4 and Lemma 1.2. □

Example 1.4. Let us reconsider the problem proposed in Example 1.1, a distributed control problem for the Laplace equation

$$\begin{aligned} &\text{minimize} \quad J(y,u) = \frac{1}{2}\int_\Omega (y-y_d)^2\, d\Omega + \frac{\alpha}{2}\int_\Omega u^2\, d\Omega, \\ &\text{s.t.} \quad \begin{cases} -\Delta y = u + f & \text{in } \Omega, \\ y = 0 & \text{on } \partial\Omega. \end{cases} \end{aligned} \tag{1.2.23}$$

We set $Y = Q = H_0^1(\Omega)$, $U = L^2(\Omega)$ and define the bilinear forms

$$a(y,q) = (\nabla y, \nabla q)_{L^2}, \qquad n(u,v) = \alpha(u,v)_{L^2}, \qquad c(u,q) = (u,q)_{L^2},$$

which satisfies the hypotheses (i)-(iii), moreover the observation operator is given by $\mathcal{Q}y = y$, hence $\|\mathcal{Q}\| = 1$. Then we can reformulate problem (1.2.23) in the form (1.2.7), provided

$$\mathcal{A}(\underline{x}, \underline{w}) = (y,z)_{L^2} + \alpha(u,v)_{L^2}, \qquad \mathcal{B}(\underline{x}, q) = (\nabla y, \nabla q)_{L^2} - (u,q)_{L^2},$$

and $G = f \in Q^*$, $\underline{F} = (y_d, 0) \in X^*$.

1.2.4 Quadratic optimization with Stokes equations constraints

We apply the results of Section 1.2.1 to quadratic optimization problems now constrained by the Stokes system as in Example 1.3 of Section 1.1.2. Here $\boldsymbol{v}$ denotes the velocity, p the pressure (note that till now p was used to denote the adjoint variable), and $\boldsymbol{u}$ the control variable acting as a body force. We consider the following distributed optimal control problem

$$\begin{aligned} &\text{minimize } J(\boldsymbol{v}, p, \boldsymbol{u}) = \frac{1}{2}\int_\Omega |\boldsymbol{v} - \boldsymbol{v}_d|^2\, d\Omega + \frac{\alpha}{2}\int_\Omega |\boldsymbol{u}|^2\, d\Omega, \qquad \text{subject to} \\ &\begin{cases} -\nu\Delta\boldsymbol{v} + \nabla p = \boldsymbol{u} & \text{in } \Omega, \\ \operatorname{div}\boldsymbol{v} = 0 & \text{in } \Omega, \\ \boldsymbol{v} = 0 & \text{on } \partial\Omega. \end{cases} \end{aligned} \tag{1.2.24}$$

We introduce the following functional spaces: $Y = [H_0^1(\Omega)]^2 \times L^2(\Omega)$ for the state variables $(\boldsymbol{v}, p)$, $U = [L^2(\Omega)]^2$ for the control variable, $Q = [H_0^1(\Omega)]^2 \times L^2(\Omega)$ for the adjoint variables $(\boldsymbol{w}, q)$ and the product space $X = Y \times U$. Using the same notations as in Example 1.3, the weak formulation of the state equation reads: find $(\boldsymbol{v}, p) \in Y$ such that

$$\begin{cases} a(\boldsymbol{v}, \boldsymbol{\xi}) + b(\boldsymbol{\xi}, p) = c(\boldsymbol{u}, \boldsymbol{\xi}) & \forall \boldsymbol{\xi} \in [H_0^1(\Omega)]^2, \\ b(\boldsymbol{v}, \tau) = 0 & \forall \tau \in L^2(\Omega). \end{cases} \tag{1.2.25}$$

Remark 1.5. The Stokes system (1.2.25) is an example of saddle-point problem, hence its analysis could be carried out in the framework of Section A.1.3 (as already discussed in Section 1.1.2). In particular the bilinear form $b(\cdot,\cdot)$ fulfils the inf-sup condition

$$\inf_{\tau \in L^2} \sup_{\boldsymbol{\xi} \in [H_0^1(\Omega)]^2} \frac{b(\boldsymbol{\xi}, \tau)}{\|\boldsymbol{\xi}\|_{H^1} \|\tau\|_{L^2}} \geq \tilde{\beta}_b > 0.$$

Note however that, as every mixed variational problem, the system (1.2.25) provides an example of weakly coercive problem, hence it could be also analysed using Nečas-Babuška theorems, see Section A.1. It is sufficient to define the bilinear form $\mathsf{A} : Y \times Y \to \mathbb{R}$ as

$$\mathsf{A}(\{\boldsymbol{v}, p\}, \{\boldsymbol{\xi}, \tau\}) = a(\boldsymbol{v}, \boldsymbol{\xi}) + b(\boldsymbol{\xi}, p) + b(\boldsymbol{v}, \tau), \tag{1.2.26}$$

and to prove that it is continuous and weakly coercive on $Y \times Y$, i.e. there exists a constant $\tilde{\beta}_{\mathsf{A}} > 0$ such that

$$\inf_{\{\boldsymbol{v},p\}\in Y} \sup_{\{\boldsymbol{\xi},\tau\}\in Y} \frac{\mathsf{A}(\{\boldsymbol{v}, p\}, \{\boldsymbol{\xi}, \tau\})}{\|\{\boldsymbol{v}, p\}\|_Y \|\{\boldsymbol{\xi}, \tau\}\|_Y} \geq \tilde{\beta}_{\mathsf{A}} \tag{1.2.27}$$

and

$$\inf_{\{\boldsymbol{\xi},\tau\}\in Y} \sup_{\{\boldsymbol{v},p\}\in Y} \frac{\mathsf{A}(\{\boldsymbol{v}, p\}, \{\boldsymbol{\xi}, \tau\})}{\|\{\boldsymbol{v}, p\}\|_Y \|\{\boldsymbol{\xi}, \tau\}\|_Y} > 0. \tag{1.2.28}$$

Actually, since $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ satisfy the hypotheses of Theorem A.2, it can be shown (see e.g. [21, 90, 35]) that the the compound form $\mathsf{A}(\cdot, \cdot)$ is continuous and weakly coercive. We will use this fact in the proof of Lemma 1.3.

Let us define the bilinear form $\mathcal{A}(\cdot, \cdot)$ and $\mathcal{B}(\cdot, \cdot)$

Definition 1.3. Let $\underline{\boldsymbol{x}} = (\{\boldsymbol{v}, p\}, \boldsymbol{u}) \in X$, $\underline{\boldsymbol{\zeta}} = (\{\boldsymbol{\varphi}, \pi\}, \boldsymbol{\lambda}) \in X$, the bilinear form $\mathcal{A}(\cdot, \cdot) \colon X \times X \to \mathbb{R}$ is defined as follows

$$\mathcal{A}(\underline{\boldsymbol{x}}, \underline{\boldsymbol{\zeta}}) := \int_\Omega \boldsymbol{v} \cdot \boldsymbol{\varphi} \, dx + \alpha \int_\Omega \boldsymbol{u} \cdot \boldsymbol{\lambda} \, dx. \tag{1.2.29}$$

Definition 1.4. Let $\underline{\boldsymbol{x}} = (\{\boldsymbol{v}, p\}, \boldsymbol{u}) \in X$, $\{\boldsymbol{\xi}, \tau\} \in Q$, the bilinear form $\mathcal{B}(\cdot, \cdot) \colon X \times Q \to \mathbb{R}$ is defined as follows

$$\mathcal{B}(\underline{\boldsymbol{x}}, \{\boldsymbol{\xi}, \tau\}) := a(\boldsymbol{v}, \boldsymbol{\xi}) + b(\boldsymbol{\xi}, p) + b(\boldsymbol{v}, \tau) - c(\boldsymbol{u}, \boldsymbol{\xi}). \tag{1.2.30}$$

Then

$$\mathcal{B}(\underline{\boldsymbol{x}}, \{\boldsymbol{\xi}, \tau\}) = \mathsf{A}(\{\boldsymbol{v}, p\}, \{\boldsymbol{\xi}, \tau\}) - c(\boldsymbol{u}, \boldsymbol{\xi}).$$

Given $\underline{\boldsymbol{F}} = (\{\boldsymbol{v}_d, 0\}, \boldsymbol{0})$ we can reformulate the optimal control problem (1.2.24) as

$$\begin{cases} \min \mathcal{J}(\underline{\boldsymbol{x}}) = \dfrac{1}{2}\mathcal{A}(\underline{\boldsymbol{x}}, \underline{\boldsymbol{x}}) - \langle \underline{\boldsymbol{F}}, \underline{\boldsymbol{x}} \rangle, & \text{subject to} \\ \mathcal{B}(\underline{\boldsymbol{x}}, \{\boldsymbol{w}, q\}) = 0 \qquad \forall \{\boldsymbol{w}, q\} \in Q, \end{cases} \tag{1.2.31}$$

If we can prove that the assumptions of Proposition 1.1 hold, then we obtain the equivalence between the control problem (1.2.24) and the following saddle-point problem: find $(\underline{\boldsymbol{x}}, \{\boldsymbol{w}, q\}) \in X \times Q$ such that

$$\begin{cases} \mathcal{A}(\underline{\boldsymbol{x}}, \underline{\boldsymbol{\zeta}}) + \mathcal{B}(\underline{\boldsymbol{\zeta}}, \{\boldsymbol{w}, q\}) = \langle \underline{\boldsymbol{F}}, \underline{\boldsymbol{\zeta}} \rangle & \forall \underline{\boldsymbol{\zeta}} \in X, \\ \mathcal{B}(\underline{\boldsymbol{x}}, \{\boldsymbol{\xi}, \tau\}) = 0 & \forall \{\boldsymbol{\xi}, \tau\} \in Q. \end{cases} \tag{1.2.32}$$

Lemma 1.3. *The bilinear forms $\mathcal{A}(\cdot, \cdot)$ and $\mathcal{B}(\cdot, \cdot)$ satisfy the assumptions of Proposition 1.1.*

Proof. Let $\underline{\boldsymbol{x}} = (\{\boldsymbol{v}, p\}, \boldsymbol{u}) \in X$, $\underline{\boldsymbol{\zeta}} = (\{\boldsymbol{\varphi}, \pi\}, \boldsymbol{\lambda}) \in X$, we use the following scalar products:

$$(\{\boldsymbol{v}, p\}, \{\boldsymbol{\varphi}, \pi\})_Y = (\boldsymbol{v}, \boldsymbol{\varphi})_{H^1} + (p, \pi)_{L^2}, \qquad (\boldsymbol{u}, \boldsymbol{\lambda})_U = (\boldsymbol{u}, \boldsymbol{\lambda})_{L^2},$$

$$(\underline{\boldsymbol{x}}, \underline{\boldsymbol{\zeta}})_X = (\{\boldsymbol{v}, p\}, \{\boldsymbol{\varphi}, \pi\})_Y + (\boldsymbol{u}, \boldsymbol{\lambda})_U.$$

The symmetry and non-negativity of $\mathcal{A}(\cdot,\cdot)$ are trivial, while the continuity of $\mathcal{A}(\cdot,\cdot)$ and $\mathcal{B}(\cdot,\cdot)$ follows immediately from the continuity of the scalar products and the continuity of the bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$. To prove the coercivity of $\mathcal{A}(\cdot,\cdot)$ on X_0 let $\underline{\boldsymbol{x}} = (\{\boldsymbol{v}, p\}, \boldsymbol{u}) \in X_0$, note that $\underline{\boldsymbol{x}} \in X_0$ if and only if $\{\boldsymbol{v}, p\}$ solves the Stokes system (1.2.25), i.e.

$$a(\boldsymbol{v}, \boldsymbol{\xi}) + b(\boldsymbol{\xi}, p) + b(\boldsymbol{v}, \tau) = (\boldsymbol{u}, \boldsymbol{\xi}) \quad \forall \{\boldsymbol{\xi}, \tau\} \in Q,$$

let $\boldsymbol{\xi} = \boldsymbol{v}$ and $\tau = p$, by Cauchy-Schwarz and Poincaré inequalities

$$\nu \|\nabla \boldsymbol{v}\|_{L^2}^2 + 2b(\boldsymbol{v}, p) = (\boldsymbol{u}, \boldsymbol{v})_{L^2} \leq \|\boldsymbol{u}\|_{L^2} \|\boldsymbol{v}\|_{L^2} \leq C_P \|\boldsymbol{u}\|_{L^2} \|\nabla \boldsymbol{v}\|_{L^2},$$

thus

$$\|\boldsymbol{u}\|_{L^2} \geq \frac{\nu}{C_P} \|\nabla \boldsymbol{v}\|_{L^2} + \frac{2b(\boldsymbol{v}, p)}{C_P \|\nabla \boldsymbol{v}\|_{L^2}}. \tag{1.2.33}$$

We want to prove that

$$\mathcal{A}(\underline{\boldsymbol{x}}, \underline{\boldsymbol{x}}) \equiv \|\boldsymbol{v}\|_{L^2}^2 + \alpha \|\boldsymbol{u}\|_{L^2}^2 \geq \alpha_0 \left[\|\boldsymbol{v}\|_{H^1}^2 + \|p\|_{L^2}^2 + \|\boldsymbol{u}\|_{L^2}^2 \right] = \alpha_0 \|\underline{\boldsymbol{x}}\|_X^2 \quad \forall \underline{\boldsymbol{x}} \in X_0,$$

by (1.2.33)

$$\begin{aligned}
\mathcal{A}(\underline{\boldsymbol{x}}, \underline{\boldsymbol{x}}) &= \|\boldsymbol{v}\|_{L^2}^2 + \alpha \|\boldsymbol{u}\|_{L^2}^2 = \|\boldsymbol{v}\|_{L^2}^2 + \frac{\alpha}{2} \|\boldsymbol{u}\|_{L^2}^2 + \frac{\alpha}{2} \|\boldsymbol{u}\|_{L^2}^2 \\
&\geq \|\boldsymbol{v}\|_{L^2}^2 + \frac{\alpha \nu^2}{2C_P^2} \|\nabla \boldsymbol{v}\|_{L^2}^2 + \frac{4\alpha}{2C_P^2} \frac{b(\boldsymbol{v}, p)^2}{\|\nabla \boldsymbol{v}\|_{L^2}^2} + \frac{\alpha}{2} \|\boldsymbol{u}\|_{L^2}^2 \\
&\geq \underbrace{\min\left\{1, \frac{\alpha \nu^2}{2C_P^2}\right\}}_{c_0} \|\boldsymbol{v}\|_{H^1}^2 + \frac{2\alpha \tilde{\beta}_b^2}{C_P^2} \frac{\|\boldsymbol{v}\|_{H^1}^2 \|p\|_{L^2}^2}{\|\nabla \boldsymbol{v}\|_{L^2}^2} + \frac{\alpha}{2} \|\boldsymbol{u}\|_{L^2}^2 \\
&\geq c_0 \|\boldsymbol{v}\|_{H^1}^2 + \frac{2\alpha \tilde{\beta}_b^2}{C_P^2} \|p\|_{L^2}^2 + \frac{\alpha}{2} \|\boldsymbol{u}\|_{L^2}^2 \\
&\geq \min\left\{c_0, \frac{2\alpha \tilde{\beta}_b^2}{C_P^2}, \frac{\alpha}{2}\right\} \left[\|\boldsymbol{v}\|_{H^1}^2 + \|p\|_{L^2}^2 + \|\boldsymbol{u}\|_{L^2}^2 \right] \\
&= \alpha_0 \|\underline{\boldsymbol{x}}\|_X^2 \qquad \forall \underline{\boldsymbol{x}} \in X_0.
\end{aligned}$$

To prove the last assumption, i.e. the fulfilment of the inf-sup condition for the bilinear form $\mathcal{B}(\cdot,\cdot)$, we make use of the weakly coercivity of the bilinear form $\mathsf{A}(\cdot,\cdot)$:

$$\begin{aligned}
\sup_{0 \neq \underline{\boldsymbol{x}} \in X} \frac{\mathcal{B}(\underline{\boldsymbol{x}}, \{\boldsymbol{w}, q\})}{\|\underline{\boldsymbol{x}}\|_X} &= \sup_{0 \neq (\{\boldsymbol{v}, p\}, \boldsymbol{u}) \in Y \times U} \frac{\mathsf{A}(\{\boldsymbol{v}, p\}, \{\boldsymbol{w}, q\}) - c(\boldsymbol{u}, \boldsymbol{w})}{(\|\{\boldsymbol{v}, p\}\|_Y^2 + \|\boldsymbol{u}\|_U^2)^{1/2}} \\
&\underset{\boldsymbol{u}=0}{\geq} \sup_{0 \neq \{\boldsymbol{v}, p\} \in Y} \frac{\mathsf{A}(\{\boldsymbol{v}, p\}, \{\boldsymbol{w}, q\})}{\|\{\boldsymbol{v}, p\}\|_Y} \geq \tilde{\beta}_{\mathsf{A}} \|\{\boldsymbol{w}, q\}\|_Y = \tilde{\beta}_{\mathsf{A}} \|\{\boldsymbol{w}, q\}\|_Q. \qquad \square
\end{aligned}$$

Proposition 1.4. *The optimal control problem (1.2.24) has a unique solution* $(\underline{\boldsymbol{x}}, \{\boldsymbol{w}, q\}) \in X \times Q$ *given by the solution to the saddle-point problem (1.2.32).*

Proof. The result follows from Theorem A.2, Proposition 1.1 and Lemma 1.3. □

Let us now introduce the Galerkin approximation of the saddle-point problem (1.2.32). We consider two families of finite dimensional subspaces $X^{\mathcal{N}}$ and $Q^{\mathcal{N}}$ of the space X and Q, note that we are implicitly considering two spaces $Y^{\mathcal{N}} \subset Y$ and $U^{\mathcal{N}} \subset U$ such that $X^{\mathcal{N}} = Y^{\mathcal{N}} \times U^{\mathcal{N}}$, with $Y^{\mathcal{N}} \equiv Q^{\mathcal{N}}$. The Galerkin-FE approximation of problem (1.2.1) has the following form: find $(\underline{\boldsymbol{x}}^{\mathcal{N}}, \{\boldsymbol{w}^{\mathcal{N}}, q^{\mathcal{N}}\}) \in X^{\mathcal{N}} \times Q^{\mathcal{N}}$ such that

$$\begin{cases} \mathcal{A}(\underline{\boldsymbol{x}}^{\mathcal{N}}, \underline{\boldsymbol{\zeta}}) + \mathcal{B}(\underline{\boldsymbol{\zeta}}, \{\boldsymbol{w}^{\mathcal{N}}, q^{\mathcal{N}}\}) = \langle \underline{\boldsymbol{F}}, \underline{\boldsymbol{\zeta}} \rangle, & \forall \underline{\boldsymbol{\zeta}} \in X^{\mathcal{N}}, \\ \mathcal{B}(\underline{\boldsymbol{x}}^{\mathcal{N}}, \{\boldsymbol{\xi}, \tau\}) = 0 & \forall \{\boldsymbol{\xi}, \tau\} \in Q^{\mathcal{N}}. \end{cases} \tag{1.2.34}$$

To guarantee the well-posedness of problem (1.2.22) we have to check the fulfilment of the assumptions (1.2.10) and (1.2.11), i.e. the strong coercivity of $\mathcal{A}(\cdot,\cdot)$ on $X_0^{\mathcal{N}}$ and the discrete inf-sup condition on $\mathcal{B}(\cdot,\cdot)$. In particular a necessary condition is that $Y \subset Y^{\mathcal{N}}$ be an *inf-sup stable* subspace of Y for the Stokes system (1.2.25), i.e. denoting with $V^{\mathcal{N}} \subset [H_0^1(\Omega)]^2$ and $M^{\mathcal{N}} \subset L_0^2(\Omega)$ the approximation spaces for the velocity and the pressure respectively, with $Y^{\mathcal{N}} = V^{\mathcal{N}} \times M^{\mathcal{N}}$, we require that $V^{\mathcal{N}}$ and $M^{\mathcal{N}}$ satisfy the discrete inf-sup condition

$$\inf_{\tau \in M^{\mathcal{N}}} \sup_{\boldsymbol{\xi} \in V^{\mathcal{N}}} \frac{b(\boldsymbol{\xi}, \tau)}{\|\boldsymbol{\xi}\|_{H^1} \|\tau\|_{L^2}} \geq \tilde{\beta}_b^{\mathcal{N}} > 0, \tag{1.2.35}$$

for a suitable constant $\tilde{\beta}_b^{\mathcal{N}}$ (possibly dependent on $\mathcal{N}$).

Lemma 1.4. *Assume that $Y^{\mathcal{N}} \subset Y = [H_0^1(\Omega)]^2 \times L_0^2(\Omega)$ be an inf-sup stable subspace of Y for the Stokes system (1.2.25) and that $Q^{\mathcal{N}} \equiv Y^{\mathcal{N}}$. Then the bilinear forms $\mathcal{A}(\cdot,\cdot)$ and $\mathcal{B}(\cdot,\cdot)$ satisfy the assumptions of Theorem A.4.*

Proof. The continuity property of the bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$, as well as the coercivity property of $a(\cdot,\cdot)$ are inherited from the continuity and coercivity properties that holds on the parents spaces. Since $Y^{\mathcal{N}}$ is assumed to be inf-sup stable the condition (1.2.35) hold. Moreover condition (1.2.35) plus the hypothesis $Y^{\mathcal{N}} \equiv Q^{\mathcal{N}}$ implies that the bilinear form $\mathsf{A}(\cdot,\cdot)$ satisfies the discrete counterpart of (1.2.27), i.e.

$$\inf_{\{\boldsymbol{v},p\} \in Y^{\mathcal{N}}} \sup_{\{\boldsymbol{\xi},\tau\} \in Q^{\mathcal{N}}} \frac{\mathsf{A}(\{\boldsymbol{v}, p\}, \{\boldsymbol{\xi}, \tau\})}{\|\{\boldsymbol{v}, p\}\|_Y \|\{\boldsymbol{v}, p\}\|_Q} \geq \tilde{\beta}_{\mathsf{A}}^{\mathcal{N}} > 0.$$

Now we can repeat exactly the same arguments as in the proof of Lemma 1.3. □

Proposition 1.5. *Assume that the hypotheses of Lemma 1.4 hold. Then, the saddle-point problem (1.2.34) has a unique solution $(\underline{\boldsymbol{x}}^{\mathcal{N}}, \{\boldsymbol{w}^{\mathcal{N}}, q^{\mathcal{N}}\}) \in X^{\mathcal{N}} \times Q^{\mathcal{N}}$.*

Proof. The result follows from Theorem A.4 and Lemma 1.4. □

Chapter 2

Numerical methods: iterative and one-shot approach

In this Chapter we discuss some numerical methods to solve PDE-constrained optimization problems with quadratic functional and linear state equation. Before tackling this class of problems it is useful to consider optimal control problems of the abstract form

$$\min_{(y,u)} J(y,u) \quad \text{subject to} \quad \mathcal{E}(y,u) = 0, \tag{2.0.1}$$

where u is the control variable and y the state variable, related to the control through the (possibly) non linear state equation $\mathcal{E}(y,u) = 0$. The two most popular strategies to solve numerically this kind of problems are described below.

(i) Eliminate the PDE constraint by means of the solution operator $u \mapsto y(u)$ which solves $\mathcal{E}(y(u),u) = 0$; then one can replace y by $y(u)$ and keep only the control variable as optimization variable. In this way the PDE-constrained optimization problem is recast into an unconstrained optimization problem, consisting in the minimization of the reduced cost functional $\hat{J}(u) = J(y(u),u)$. The idea is to use in a suitable way the standard algorithms of non-linear optimization such as gradient, conjugate gradient, Newton and quasi-Newton methods (see e.g. [60, 27]). This approach is called *iterative* or *reduced space* method (see for instance the monographs [40, 89, 47]).

(ii) Treat both the control and the state variables $x = (y,u)$ as independent optimization variables, coupled through the PDE constraint. In this case the problem is naturally analyzed in the framework of equality constrained non-linear optimization; in fact the Lagrangian formalism leads to a (possibly) non-linear optimality system, which has to be solved through appropriate linearization procedures (like Sequential Quadratic Programming methods) or modern penalty methods (like Augmented Lagrangian methods), see e.g. [40, 47]. This is the so called *one-shot* or *all-at-once* approach.

As in Chapter 1, we limit ourselves to consider the simple (yet significant and non-trivial) case of optimal control problems with quadratic functional and linear state equation. In Section 2.1 we introduce the iterative approach and its application to this class of problems, in particular we discuss the steepest descent, conjugate gradient, Newton and quasi-Newton methods. We underline the equivalence between applying these iterative optimization methods to the reduced cost functional or using iterative linear schemes to solve the Schur complement system obtained by block elimination of the state and adjoint variables in the optimality conditions system. Hence the *iterative* approach leads to the solution of a a symmetric positive definite system for determining the control variables, the matrix of this linear system is referred to as *reduced Hessian*. We perform some numerical tests using different linear solvers like Richardson, conjugate gradient and GMRES methods, and we briefly discuss the issue of preconditioning.

In Section 2.2 we focus on the one-shot approach that in the case of linear-quadratic problem yields to solve a quadratic programming problem. In fact, as already mentioned in Section 1.2.2, at the discrete level the optimization problem can be formulated as

$$\min \frac{1}{2}x^T Ax - f^T x \quad \text{subject to } Bx = g, \tag{2.0.2}$$

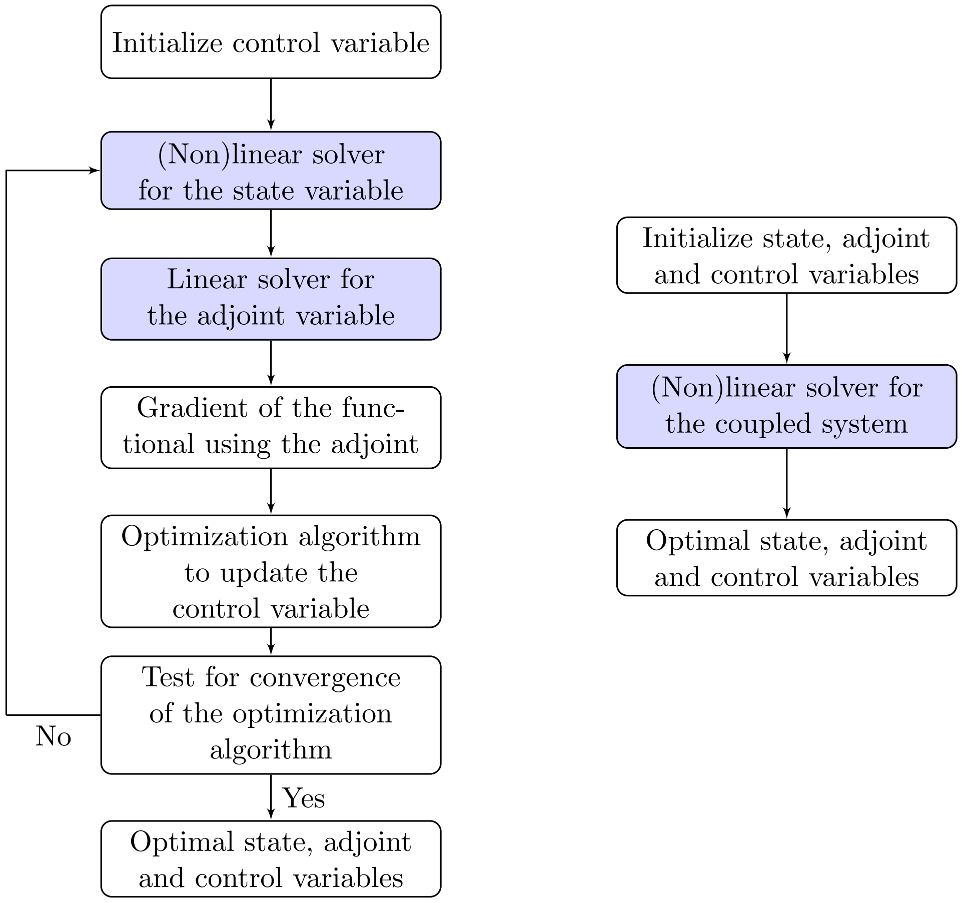
i.e. as a quadratic programming problem with equality constraint (see [60, 27]), whose optimality conditions are given by the linear system

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} x \\ p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}, \tag{2.0.3}$$

being p the adjoint variable (Lagrange multiplier). The matrix of this linear system in saddle-point form is symmetric, indefinite, usually ill-conditioned and often very large and sparse, so for the resolution of (2.0.3) we have to use some efficient preconditioned iterative linear system solvers such as Krylov subspace methods. The two main strategies for preconditioning are domain decomposition (DD, see references cited in [1]) and multigrid techniques (MG, see [9, 10]), which have been proven to be among the fastest methods for solving large-scale PDE problems. Here we will concentrate on the second one. Multigrid techniques can be applied to system (2.0.3) in two ways: either directly on the whole matrix or as approximate solvers in block preconditioning schemes. In particular while the first approach [9, 83] requires a careful design of *ad hoc* multigrid algorithms for the system (2.0.3), the second one allows to use as building blocks of the preconditioner existing and well-evaluated multigrid algorithms for the matrices A and B, in particular optimal multigrid preconditioners for the matrix B, i.e. the discretized PDE operator. In this work we will focus on the last approach referring principally to the ideas proposed in [70, 71, 72, 93]. In particular we introduce some block diagonal MG preconditioners for the optimal control of both the Laplace equation and the Stokes system.

We want to clarify that the purpose of the chapter is neither to be rigorous (we avoid to discuss the spectral properties and the theoretical convergence behaviour of the numerical schemes we are going to introduce) nor to propose original results. Rather, we aim to introduce the main ideas behind the two most popular strategies employed to tackle this kind of problems, and consequently to discuss, through some simple numerical experiments, their main differences, properties and limits.

As a final remark let us underline that even if the problems considered here are the most simple in the wider class of problems described by (2.0.1), they nonetheless represent a first step towards the resolution of more general and involved problems. In fact, as mentioned above, when tackling non-linear PDE-constrained optimization problems, even with additional inequality constraints on the control and/or state, various techniques like SQP, Penalty methods and Active Set strategies have to be adopted. These methods often require solving a sequence of simpler problems with the same structure of (2.0.3). So developing efficient methods for solving these simple problems is crucial to design efficient methods for solving even more challenging problems.


Initialize control variable
(Non)linear solver for the state variable
Linear solver for the adjoint variable
Gradient of the functional using the adjoint
Optimization algorithm to update the control variable
Test for convergence of the optimization algorithm
No
Yes
Optimal state, adjoint and control variables
Initialize state, adjoint and control variables
(Non)linear solver for the coupled system
Optimal state, adjoint and control variables

(a) *Iterative* approach.

(b) *One-shot* approach.

Figure 2.1: Schemes of the two approaches for generic optimal control problems, either linear or nonlinear.

2.1 Iterative methods for optimal control problems

As noticed above, in the *iterative* approach the starting idea is to consider the state variable as a function of u and then to recast the problem into an unconstrained optimization problem with respect only to the control variable u, i.e.

$$\min_u \hat{J}(u) = J(y(u), u)$$

where, given the control u, the state variable $y(u)$ is computable solving the state equation. In other words, an existing algorithm for the solution of the state equation is embedded into an optimization loop. Since efficient optimization requires gradients, the solver for the state equation has to be augmented with a routine which provides the gradient of the state with respect to the optimization variables. Employing a finite difference approximation of the gradient may seem attractive due to its ease of implementation, but it has limited accuracy and it is costly in the presence of many design variables; thus it is usually preferable to calculate the gradient using a sensitivity or adjoint approach (e.g [33]). As mentioned above, just limiting to consider unconstrained finite-optimization, several classes of algorithms can be adopted in PDE-constrained optimization, see for instance the monographs [40, 89, 47] and the numerous references therein.

In Section 2.1.1 we introduce the basics on nonlinear unconstrained optimization, in particular we discuss the steepest descent, conjugate gradient, Newton and quasi-Newton methods, referring principally to [27, 60]. Then in Section 2.1.2 we apply the algorithms of unconstrained optimization to PDE-constrained optimization problems. As it will be shown in Section 2.1.3, it turns out that applying these iterative optimization methods to the reduced cost functional is equivalent to use an iterative linear schemes to solve the Schur complement system, obtained by block elimination of the state and adjoint variables in the optimality conditions system. In Section 2.1.4 we perform some numerical experiments comparing different solvers.

2.1.1 Iterative methods for unconstrained optimization

Let us consider the following finite-dimensional optimization problem:

$$\min_{\mathbf{u}\in U_{\text{ad}}} J(\mathbf{u}) \tag{2.1.1}$$

where $U_{ad} \subseteq U := \mathbb{R}^n$ and $J\colon U \to \mathbb{R}$ is a smooth function (at least twice differentiable). Note that if $U_{ad} \equiv U$ the optimization problem is said to be unconstrained, while if $U_{ad} \subset U$ is said to be constrained. We will consider only the unconstrained case. Also we denote with $\mathbf{g}(\mathbf{u}) \in \mathbb{R}^n$ and $H(\mathbf{u}) \in \mathbb{R}^{n\times n}$ respectively the gradient and the Hessian of the cost functional. The basic idea for solving this minimization problem is to build an iterative algorithm that, given a starting point $\mathbf{u}_0$ supplied by the user, generates a sequence of iterates $\{\mathbf{u}_k\}$ typically such that $J(\mathbf{u}_{k+1}) < J(\mathbf{u}_k)$; the algorithm ends when a suitable stopping criterium is fulfilled. There are several possibilities to move from the current iterate $\mathbf{u}_k$ to the new iterate $\mathbf{u}_{k+1}$, each of them falling into one of these two fundamental strategy:

- the *line search* type methods, where the algorithm chooses a direction $\mathbf{d}_k$ and searches along this direction from the current iterate $\mathbf{u}_k$ for a new iterate with $J(\mathbf{u}_{k+1}) < J(\mathbf{u}_k)$;
- the *trust region* type methods, where the information gathered about J is used to construct a model function whose behaviour in a given region (the *trust region*) near the current point $\mathbf{u}_k$ is similar to that of the actual objective function J. The size of the region is modified during the iteration process.

Here we consider only line search type methods, for which the minimization problem can be stated as follows: given an initial guess $\mathbf{u}_0 \in U$, the method consists in finding iteratively a sequence $\{\mathbf{u}_k\}$ such that

$$\mathbf{u}_{k+1} = \mathbf{u}_k + \tau_k \mathbf{d}_k \qquad k = 1, 2, \ldots$$

where $\mathbf{d}_k$ represents a descent direction, that is a vector that satisfies $\mathbf{d}_k^T \cdot \nabla J(\mathbf{u}_k) < 0$, and τ_k is the *step length* that can be chosen in a variety of ways, as we discuss later. The general scheme of a line search type method can be summarized as in Algorithm 2.1. Typically the search direction $\mathbf{d}_k$ could depend on the gradient and the Hessian of the objective function $J(\cdot)$; several algorithms are available with different choices of $\mathbf{d}_k$, each of them with some advantages and disadvantages as regards computational and storage costs. In this section we briefly introduce some of these algorithms: the steepest descent method, the conjugate gradient method, and finally Newton and quasi-Newton methods (in particular the limited-Memory BFGS). Let us observe that for each of them we can choose different strategies for the computation of the step length τ_k, we only mention that one of the most popular choice

```
Input: Given a tolerance $\varepsilon > 0$ and an initial guess $\mathbf{u}_0$,
set $k = 0$
repeat
    compute the search direction $\mathbf{d}_k$
    compute the step length $\tau_k$
    set $\mathbf{u}_{k+1} = \mathbf{u}_k + \tau_k \mathbf{d}_k$
    set $k = k + 1$
until $\|\nabla J(\mathbf{u}_k)\| < \varepsilon$ or $|J(\mathbf{u}_{k+1}) - J(\mathbf{u}_k)| < \varepsilon$ or $\|\mathbf{u}_{k+1} - \mathbf{u}_k\| < \varepsilon$.
```

Algorithm 2.1: Line search methods

consist in performing an inexact line search with the Armijo rule, see [27, 60]. However, when the functional is quadratic, i.e. of the form $J(\mathbf{u}) = 1/2\mathbf{u}^T H\mathbf{u} - \mathbf{b}^T\mathbf{u}$, one can easily perform an exact line search: minimizing $J(\mathbf{u}_k + \tau_k\mathbf{d}_k)$ with respect to τ_k we obtain

$$\tau_k = -\frac{\mathbf{g}_k^T\mathbf{d}_k}{\mathbf{d}_k^T H\mathbf{d}_k}.$$

Steepest-descent method

The steepest descent method (also called gradient method) is a line search method that moves along the direction

$$\mathbf{d}_k = -\nabla J(\mathbf{u}_k)$$

at every step. Some advantages of this method are that it only requires calculation of the gradient $\nabla J(\mathbf{u}_k)$ but not of the Hessian, so it needs a low amount of memory and is globally convergent; the main disadvantage is that it converges only linearly, hence the computational cost could be relevant.

Conjugate gradient method

The conjugate gradient (CG) method improves the idea of gradient method moving along directions given by

$$\mathbf{d}_k = -\nabla J(\mathbf{u}_k) + \beta_k\mathbf{d}_{k-1},$$

where the scalar β_k is to be determined by the requirement that $\mathbf{d}_{k-1}$ and $\mathbf{d}_k$ must be conjugate with respect to $H(\mathbf{u}_k)$, i.e. $\mathbf{d}_k^T H(\mathbf{u}_k)\mathbf{d}_{k-1} = 0$. When the functional J is quadratic the CG method is said to be linear and β_k is given by

$$\beta_k = \frac{\nabla J(\mathbf{u}_k)^T H\mathbf{d}_{k-1}}{\mathbf{d}_{k-1}^T H\mathbf{d}_{k-1}}.$$

When $J(\mathbf{u})$ is a general nonlinear function, different *nonlinear* CG methods, which corresponds to different choices of β_k, are used; possible choice of β_k are given by the Fletcher-Reeves and the Polak-Ribière formulae, see [27, 60] fo further details.

Newton and quasi-Newton methods

Newton and quasi-Newton methods are probably, among line search type algorithms, the most used. The idea of Newton method is to minimize the quadratic model $m(\mathbf{d}_k)$ of $J(\mathbf{u}_k)$

$$m(\mathbf{d}_k) = J(\mathbf{u}_k) + \mathbf{d}_k^T\nabla J(\mathbf{u}_k) + \frac{1}{2}\mathbf{d}_k^T H(\mathbf{u}_k)\mathbf{d}_k$$

in a suitable neighbourhood of $\mathbf{u}_k$. If the Hessian matrix is positive definite, the vector $\mathbf{d}_k$ that minimize $m(\mathbf{d}_k)$ is given by

$$\mathbf{d}_k = -(H(\mathbf{u}_k))^{-1}\nabla J(\mathbf{u}_k),$$

i.e. in Newton methods the descent direction is the solution of the linear system

$$H(\mathbf{u}_k)\mathbf{d}_k = -\nabla J(\mathbf{u}_k). \tag{2.1.2}$$

Newton methods converge quadratically (only locally) but require at each iteration to compute the Hessian $H_k = H(\mathbf{u}_k)$ and to solve the linear system (2.1.2). Quasi-Newton methods allow us to avoid the calculation of the Hessian, replacing it with a suitable approximation, i.e. an approximated inverse of the Hessian matrix, say B_k^{-1}, is used to replace H_k^{-1} in equation (2.1.2). One of the most popular formula for updating the Hessian approximation B_k is the so called BFGS (from Broyden-Fletcher-Goldfarb-Shanno) update:

$$B_{k+1} = B_k - \frac{B_k \boldsymbol{s}_k \boldsymbol{s}_k^T B_k}{\boldsymbol{s}_k^T B_k \boldsymbol{s}_k} + \frac{\mathbf{y}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \boldsymbol{s}_k}, \tag{2.1.3}$$

where

$$\boldsymbol{s}_k = \mathbf{u}_{k+1} - \mathbf{u}_k, \qquad \mathbf{y}_k = \nabla J(\mathbf{u}_{k+1}) - \nabla J(\mathbf{u}_k).$$

In practical implementations we avoid to factorize B_k at each iteration by updating directly the inverse of B_k instead of B_k itself, using the following formula

$$C_{k+1} = (I - \rho_k \boldsymbol{s}_k \mathbf{y}_k^T) C_k (I - \rho_k \mathbf{y}_k \boldsymbol{s}_k^T) + \rho_k \boldsymbol{s}_k \boldsymbol{s}_k^T, \qquad \rho_k = \frac{1}{\mathbf{y}_k^T \boldsymbol{s}_k},$$

where $C_k := B_k^{-1}$ and usually $B_0 = I$. While the BFGS method permits us to avoid the calculation and inversion of the Hessian matrix, we still have to store at each iteration the full matrices C_k and C_{k+1}, storage that could be troublesome in case of large scale optimization problems, as optimal control of PDEs could be. The *limited-memory BFGS method* circumvent this difficulty, storing, instead of a fully dense $n \times n$ approximation matrix, only a few vectors of length n.

2.1.2 Iterative optimization methods for optimal control of PDEs

We now want to introduce an iterative algorithm well suited for solving an optimal control problem governed by a PDE. Let us consider the abstract problem

$$\min_{(y,u)} J(y,u) \quad \text{subject to } \mathcal{E}(y,u) = 0, \tag{2.1.4}$$

and assume that we have at our disposal the state operator (also called control-to-state map) $u \mapsto y(u)$ which solves $\mathcal{E}(y(u),u) = 0$, so that we can eliminate the PDE constraint $\mathcal{E}(\cdot,\cdot)$ in (2.1.4). The constrained minimization problem (2.1.4) can be recast in the following reduced problem

$$\min_{u} \hat{J}(u) = J(y(u),u). \tag{2.1.5}$$

To apply the algorithms introduced in the previous section we only need an explicit expression for the gradient of $J'(u)$, that is computable using the adjoint technique, as we have already seen in Section 1.1.1. More precisely we have

$$\hat{J}'(u) = \nabla J(u) = J_u + \mathcal{E}_u^* p,$$

where p solves the adjoint problem and $\mathcal{E}_u$ is the Fréchet derivative of the state equation with respect to u.

We now apply these ideas to the case of linear-quadratic problems, in particular, to make things more clear, we consider a simple example, the optimal control problem for the Laplace equation already discussed in Example 1.1 (Section 1.1.2). We recall here the problem at hand with the state equation in weak form

$$\begin{aligned} \text{minimize} \quad & J(y,u) = \frac{1}{2}\int_\Omega (y-y_d)^2\, d\Omega + \frac{\alpha}{2}\int_\Omega u^2\, d\Omega, \\ \text{s.t.} \quad & a(y,q) = (u,q)_{L^2} + (f,q)_{L^2} \qquad \forall q \in Y, \end{aligned} \tag{2.1.6}$$

where $Y = H_0^1(\Omega)$ and $a(y,q) = (\nabla y, \nabla q)_{L^2}$. In this example the adjoint state is defined as the solution of

$$a(z,p) = (y_d - y, z)_{L^2} \qquad \forall z \in Y,$$

while the gradient of the cost functional is given by $\nabla J(u) = -p + \alpha u$. Obviously all the optimization procedures described in the previous section can only be carried out in connection with a suitable discretization, hence let us now introduce the finite element (FE) discretization of the optimization problem. Let $\{\mathcal{T}_\mathcal{N}\}$ be a triangulation of the domain Ω, we denote $V_\mathcal{N}^r$ the $\mathbb{P}_r$-conforming finite element space associated with the triangulation $\{\mathcal{T}_\mathcal{N}\}$. Moreover we define $Y^\mathcal{N} = Y \cap V_\mathcal{N}^r$ and $U_\mathcal{N} = U \cap V_\mathcal{N}^r$ in such a way that $Y^\mathcal{N} \subset Y$, $U^\mathcal{N} \subset U$ are sequences of FE approximation spaces. The FE discretization of (2.1.6) reads:

$$\min\ J(y^\mathcal{N}, u^\mathcal{N}) \quad \text{subject to} \quad a(y^\mathcal{N}, q) = (u^\mathcal{N}, q)_{L^2} + (f,q)_{L^2} \quad \forall q \in Y^\mathcal{N}.$$

Denoting with $\mathbf{u}$ and $\mathbf{y}$ the coefficients vectors in the expansion of $u^\mathcal{N}$ and $y^\mathcal{N}$ in terms of the standard nodal bases functions for $U^\mathcal{N}$ and $Y^\mathcal{N}$, we give the algebraic formulation of the FE discretization:

$$\begin{aligned} \text{minimize} \quad & J(\mathbf{y},\mathbf{u}) = \frac{1}{2}\mathbf{y}^T M\mathbf{y} - \mathbf{y}^T M\mathbf{y}_d + \frac{\alpha}{2}\mathbf{u}^T M\mathbf{u} + \frac{1}{2}\mathbf{y}_d^T M\mathbf{y}_d, \\ \text{s.t.} \quad & K\mathbf{y} = M\mathbf{u} + \mathbf{f}, \end{aligned} \tag{2.1.7}$$

where K is the stiffness matrix arising from the discretization of the bilinear form $a(\cdot,\cdot)$ and M is the mass matrix on Ω. As well as the discretized state equation is given by

$$K\mathbf{y} = M\mathbf{u} + \mathbf{f},$$

we can easily obtain the discretized adjoint equation

$$K^T\mathbf{p} = -M\mathbf{y} + M\mathbf{y}_d,$$

and the gradient of the discrete cost functional $\nabla J(\mathbf{u}) = -\mathbf{p} + \alpha\mathbf{u}$. Now it should be more clear how to adapt the iterative algorithms of unconstrained optimization to PDE-constrained minimization: given an initial guess $\mathbf{u} = \mathbf{u}_0$ we compute $\mathbf{y}$ by solving the state equation and $\mathbf{p}$ by solving the adjoint equation; then we compute the cost functional gradient ∇J and the value of J, and apply a suitable convergence criterium. If this criterium is not fulfilled we update the control $\mathbf{u}$ according to one of the optimization methods introduced in the previuos section and repeat all the steps. The iterative process ends when the convergence criterium is fulfilled. In Algorithm 2.2 we summarized this process with more details.

```
Input: Given an initial guess $\mathbf{u}_0$, compute $\mathbf{y}_0$ by solving state equation with $\mathbf{u} = \mathbf{u}_0$, and
    $\mathbf{p}_0$ by solving adjoint equation with $\mathbf{y} = \mathbf{y}_0$, compute $J_0$ and $\nabla J_0$. Set $k = 0$.
1: while $\|\nabla J(\mathbf{u}_k)\|/\|\nabla J(\mathbf{u}_0)\| >$ tol or $|J(\mathbf{u}_k) - J(\mathbf{u}_{k-1})| >$ tol do
2:     compute search direction $\boldsymbol{d}_k$ according with one of the optimization methods (SD,
       CG, BFGS, ...)
3:     compute step length $\tau_k$ with a line search routine
4:     set $\mathbf{u}_{k+1} = \mathbf{u}_k + \tau_k \boldsymbol{d}_k$
5:     compute $\mathbf{y}_{k+1}$ by solving state equation with $\mathbf{u} = \mathbf{u}_{k+1}$
6:     compute $\mathbf{p}_{k+1}$ by solving adjoint equation with $\mathbf{y} = \mathbf{y}_{k+1}$
7:     compute $J_{k+1}$ and $\nabla J_{k+1}$
8:     $k = k + 1$
9: end while
```

Algorithm 2.2: Iterative algorithm for optimal control problem

Remark 2.1. While in classical unconstrained optimization the evaluation of the cost functional J in an arbitrary point $\bar{\mathbf{u}}$ has a low computational cost, in the context of PDE-constrained optimization, since the cost functional depends on both $\mathbf{u}$ and $\mathbf{y}$, the evaluation of $J(\bar{\mathbf{u}})$ is an expensive operation, because it requires to solve the state equation in order to compute $\bar{\mathbf{y}} = \mathbf{y}(\bar{\mathbf{u}})$ and then compute $J(\bar{\mathbf{y}}, \bar{\mathbf{u}})$.

Remark 2.2. As remarked in [89, Sec. 2.12.1], for linear-quadratic problems the computation of the *optimal step size* τ_k can be made analytically, thanks to the fact that J is quadratic. In fact (subscript 2 denote the discrete $L^2(\Omega)$ norms and scalar product)

$$\begin{aligned}\phi(\tau_k) := J(\mathbf{u}_k + \tau_k \mathbf{d}_k) &= \frac{1}{2}\|\mathbf{y}_k + \tau_k \mathbf{y}(\mathbf{d}_k) - \mathbf{y}_d\|_2^2 + \frac{\alpha}{2}\|\mathbf{u}_k + \tau_k \mathbf{d}_k\|_2^2 \\ &= \frac{\tau_k^2}{2}\Big(\|\mathbf{y}(\mathbf{d}_k)\|_2^2 + \alpha\|\mathbf{d}_k\|_2^2\Big) + \tau_k\Big((\mathbf{y}_k - \mathbf{y}_d, \mathbf{y}(\mathbf{d}_k))_2 + \alpha(\mathbf{u}_k, \mathbf{d}_k)_2\Big) + t\end{aligned}$$

where the constant t does not depend on $\mathbf{d}_k$ and τ_k, and $\mathbf{y}(\mathbf{d}_k)$ denote the solution of the state equation when $\mathbf{u} = \mathbf{d}_k$. Being a parabola, the problem of minimizing $\phi(\tau)$ can be solved by hand:

$$\tau_k = \arg\min_{\tau>0} \phi(\tau) = \frac{-\alpha(\mathbf{u}_k, \mathbf{d}_k)_2 - (\mathbf{y}_k - \mathbf{y}_d, \mathbf{y}(\mathbf{d}_k))_2}{\|\mathbf{y}(\mathbf{d}_k)\|_2^2 + \alpha\|\mathbf{d}_k\|_2^2}. \tag{2.1.8}$$

Hence, in step 3 of Algorithm 2.2 we can avoid the use of a line search routine, we just compute the optimal step size (2.1.8).

2.1.3 Iterative optimization methods as linear solvers for the reduced Hessian

We want to investigate the relation between the iterative approach discussed above and the structure of the discrete optimality conditions system. To simplify the discussion we refer again to the example (2.1.6), the optimality system reads:

$$\text{adjoint:} \qquad K^T\mathbf{p} + M\mathbf{y} = M\mathbf{y}_d \tag{2.1.9a}$$

$$\text{optimality:} \qquad \alpha M\mathbf{u} - M\mathbf{p} = 0 \tag{2.1.9b}$$

$$\text{state:} \qquad K\mathbf{y} - M\mathbf{u} = \mathbf{f}, \tag{2.1.9c}$$

in matrix-vector notation,

$$\begin{pmatrix} M & 0 & K^T \\ 0 & \alpha M & -M \\ K & -M & 0 \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} M\mathbf{y}_d \\ \mathbf{0} \\ \mathbf{f} \end{pmatrix}. \tag{2.1.10}$$

We want to eliminate the state and adjoint equations and variables, and then analyze the system in the remaining control space (the so called *reduced system*), i.e. we compute the Schur complement of the linear system (2.1.10) with respect to the control variable. We substitute in the optimality equation the adjoint variable $\mathbf{p}$ derived from the adjoint equation, then we do the same for the state variable $\mathbf{y}$, rearranging we obtain the reduced system for the control variable:

$$H\mathbf{u} = \boldsymbol{b}, \tag{2.1.11}$$

where

$$H = \alpha M + MK^{-T}MK^{-1}M, \qquad \boldsymbol{b} = -MK^{-T}MK^{-1}\boldsymbol{f} + MK^{-T}M\boldsymbol{y}_d,$$

and H is referred to as the *reduced Hessian* matrix. It can be shown that applying the iterative algorithms of Section 2.1.1 to the reduced cost functional $\hat{J}(u)$ is equivalent to solve with an iterative method the linear system (2.1.11). For example the steepest descent algorithm for the minimization of the cost functional is equivalent to the following non-stationary preconditioned Richardson scheme for the solution of the reduced system: given $\mathbf{u}_0$, for $k > 0$

$$\mathbf{u}^{k+1} = \mathbf{u}^k + \tau_k M^{-1}\boldsymbol{r}_k, \tag{2.1.12}$$

where $\boldsymbol{r}_k$ is the residual defined as $\boldsymbol{r}_k = \boldsymbol{b} - H\mathbf{u}_k$. In fact, let us consider the steepest descent update

$$\mathbf{u}_{k+1} = \mathbf{u}_k + \tau_k\mathbf{d}_k,$$

where $\mathbf{d}_k = -\nabla J(\mathbf{u}_k) = \mathbf{p}_k - \alpha\mathbf{u}_k$, using the adjoint equation we get $\mathbf{p}_k = K^{-T}(-M\mathbf{y}_k + M\boldsymbol{y}_d)$, thus

$$\mathbf{u}_{k+1} = \mathbf{u}_k + \tau_k\big(K^{-T}M\boldsymbol{y}_d - \alpha\mathbf{u}_k - K^{-T}M\mathbf{y}_k\big),$$

by the state equation $\mathbf{y}_k = K^{-1}(M\mathbf{u}_k + \boldsymbol{f})$, substituting and rearranging we obtain

$$\begin{aligned} \mathbf{u}_{k+1} &= \mathbf{u}_k + \tau_k\big(K^{-T}M\boldsymbol{y}_d - K^{-T}MK^{-1}\boldsymbol{f} - \alpha\mathbf{u}_k - K^{-T}MK^{-1}M\mathbf{u}_k\big) \\ &= \mathbf{u}_k + \tau_k M^{-1}\big(-MK^{-T}MK^{-1}\boldsymbol{f} + MK^{-T}M\boldsymbol{y}_d - \alpha M\mathbf{u}_k - MK^{-T}MK^{-1}M\mathbf{u}_k\big) \\ &= \mathbf{u}_k + \tau_k M^{-1}(\boldsymbol{b} - H\mathbf{u}_k). \end{aligned}$$

Actually, this result is not surprising if we write the algebraic counterpart of the reduced cost functional $\hat{J}(u)$: substituting $\mathbf{y}(\mathbf{u}) = K^{-1}(M\mathbf{u} + \boldsymbol{f})$ in $J(\mathbf{y}, \mathbf{u})$ we obtain

$$\begin{aligned} \hat{J}(\mathbf{u}) = J(\mathbf{y}(\mathbf{u}), \mathbf{u}) &= \frac{1}{2}\mathbf{u}^T MK^{-T}MK^{-1}M\mathbf{u} - \mathbf{u}^T MK^{-T}M\mathbf{y}_d + \frac{\alpha}{2}\mathbf{u}^T M\mathbf{u} \\ &\quad + \mathbf{u}^T MK^{-T}MK^{-1}\boldsymbol{f} + \frac{1}{2}\mathbf{y}_d^T M\mathbf{y}_d + \frac{1}{2}\boldsymbol{f}K^{-T}MK^{-1}\boldsymbol{f} - \boldsymbol{f}^T K^{-T}M\mathbf{y}_d \\ &= \frac{1}{2}\mathbf{u}^T H\mathbf{u} - \mathbf{u}^T\boldsymbol{b} + t, \end{aligned}$$

where the constant term $t = \frac{1}{2}\mathbf{y}_d^T M\mathbf{y}_d + \frac{1}{2}\boldsymbol{f}K^{-T}MK^{-1}\boldsymbol{f} - \boldsymbol{f}^T K^{-T}M\mathbf{y}_d$ does not affect the minimizer of $\hat{J}(\cdot)$. Therefore, being H a symmetric positive definite matrix, the minimization of the quadratic functional $\hat{J}(\mathbf{u})$ is equivalent to the solution of the linear system $\nabla \hat{J}(\mathbf{u}) = H\mathbf{u} - \boldsymbol{b} = 0$ (e.g. [68, 81]), i.e.

$$\text{minimize } \hat{J}(\mathbf{u}) \qquad \Longleftrightarrow \qquad \text{solve } H\mathbf{u} = \boldsymbol{b}.$$

Being H the Hessian of the reduced cost functional it is evident why it is referred to as the *reduced Hessian* matrix. Moreover, it is clear that for the solution of this linear system we are no more limited to consider the optimization algorithms introduced above, but we can also use other iterative methods, for example those based on Krylov subspaces. We remark that, independently from the choice of the solver, at each iteration, to compute the matrix-vector multiplication $H\boldsymbol{r}$ for given $\boldsymbol{r}$, we have to perform the state solution

$$\mathbf{y}_r = K^{-1}M\boldsymbol{r}$$

the adjoint solution

$$\mathbf{p}_r = -K^{-T}M\mathbf{y}_r$$

and the evaluation of the expression

$$H\boldsymbol{r} = \alpha M\boldsymbol{r} - M\mathbf{p}_r.$$

Moreover, once the solver converges and we get the optimal control $\mathbf{u}$, we still have to perform a state solution and an adjoint solution to get the optimal $\mathbf{y}$ and $\mathbf{p}$.

Note also that, due to the difficulty in forming explicitly the matrix H, one can not try to solve the linear system $H\mathbf{u} = \boldsymbol{b}$ with a direct method.

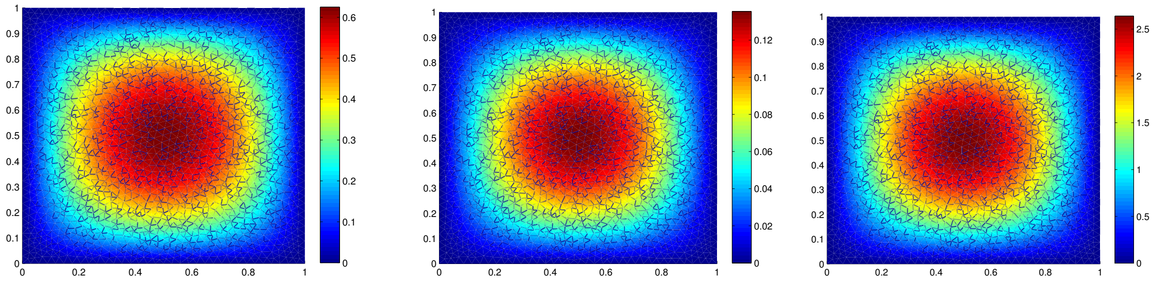
2.1.4 A numerical example: test and comparison

We test the algorithms introduced above on the simple problem (2.1.6), i.e.

$$\begin{aligned} &\text{minimize} \quad J(y,u) = \frac{1}{2}\int_\Omega (y - y_d)^2\, d\Omega + \frac{\alpha}{2}\int_\Omega u^2\, d\Omega, \\ &\text{s.t.} \quad \begin{cases} -\Delta y = u + f & \text{in } \Omega, \\ y = 0 & \text{on } \partial\Omega. \end{cases} \end{aligned} \tag{2.1.13}$$

where $\Omega = (0,1)^2$, $U = L^2(\Omega)$, $f = 0$ and the desired state is given by $y_d = 10x_1(1-x_1)x_2(1-x_2)$ (this is a test case proposed in [16]). For the discretization we use $\mathbb{P}_1$ continuous finite element for the state, adjoint and control variables. A plot of the desired state y_d is given in Figure 2.2a, as well as plots of the optimal state y and control u are given in Figure 2.2b and 2.3 for different values of the regularization parameter α.

In the following we will compare the performance of some linear solvers, in particular the preconditioned non-stationary Richardson method, the conjugate gradient (CG) method and the GMRES. We compare the performance of each method for different values of the regularization parameter α and on different unstructured meshes (the size of the reduced system related to each level of refinement is given in Table 2.1); the tolerance for the stopping criterion in the linear solvers have been set to 10^{-7}. The main optimal control solver has been implemented in MATLAB using the `Mlife` library [82], for the GMRES and CG methods we rely on the native MATLAB implementations while the Richardson scheme has been implemented. We also remark that all the state and adjoint problem solves required to compute (at each iteration) the matrix-vector multiplication $H\boldsymbol{r}$, are performed using the



(a) Desired state y_d.

(b) Optimal state (left) and control (right) solutions for $\alpha = 10^{-2}$, mesh size $h = 0.03$; the value of the cost functional is $J = 4.41 \cdot 10^{-4}$.

Figure 2.2

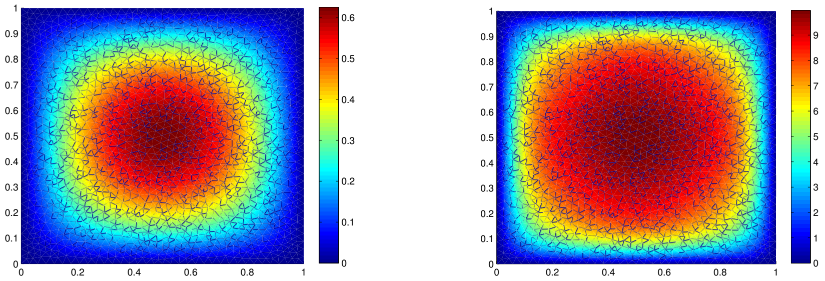


Figure 2.3: Optimal state (left) and control (right) solutions for $\alpha = 10^{-5}$, mesh size $h = 0.03$; the value of the cost functional is $J = 2.32 \cdot 10^{-4}$.

direct method embedded in MATLAB's backslash '\'. An alternative could be to use for example a Cholesky decomposition, or, in view of more challenging problems, to use some suitable inner preconditioned iterative method, for further details see [1].

As a first test we compare the number of iterations and CPU-time using the preconditioned Richardson method (i.e. the steepest descent method), the conjugate gradient and the GMRES, the last two non preconditioned. The results for $\alpha = 10^{-1}$ and $\alpha = 10^{-3}$, given in Table 2.2, suggest some comments:

- for $\alpha = 10^{-1}$ the preconditioned Richardson method outperform the CG and GMRES solver, both in terms of number of iterations and CPU-time;

- for $\alpha = 10^{-3}$ (hence smaller than before) the performance of the CG and GMRES methods remains stable, while the Richardson method shows poor convergence; in general, as α becomes smaller, the three methods requires an higher number of iterations to reach convergence (see for example Table 2.4);

- the mesh size (at least in the range tested here) does not affect considerably the performances of the solvers.

From these observations, inspired by the Richardson method, we decided to test the preconditioned CG and GMRES solver using as preconditioner the mass matrix M. The results given in Table 2.3 show the benefits given by the preconditioner, both for the number of iterations required to reach the convergence and the computational time. In Figure 2.4 we report two plots comparing the convergence behaviour of the solvers for different values of

h	0.08	0.04	0.02	0.01
$\mathcal{N}$	452	1998	8364	33922

Table 2.1: Size of the reduced system related to each level of refinement considered.

h	CG		GMRES		Richardson	
	s	iter	s	iter	s	iter
0.08	0.085	21	0.073	21	0.039	5
0.04	0.315	23	0.324	23	0.171	5
0.02	2.35	19	2.33	19	1.15	5
0.01	11.25	24	11.28	23	5.64	5

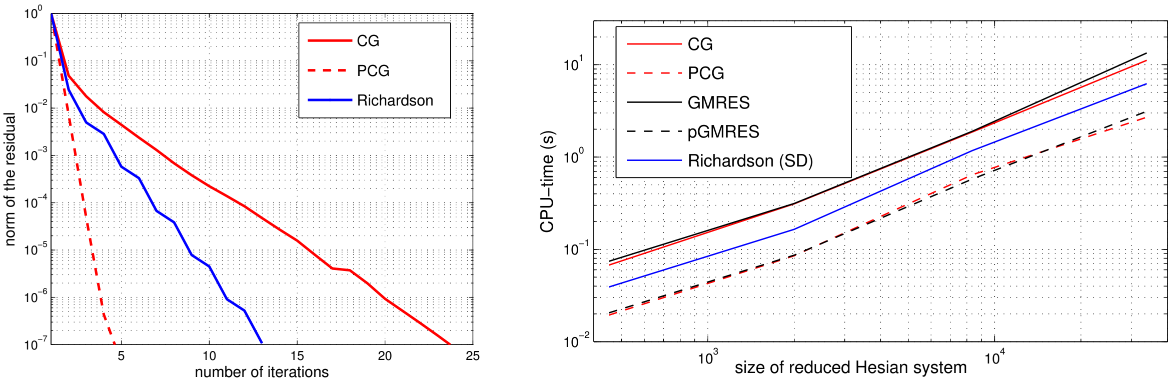
(a) $\alpha = 10^{-1}$, $J = 5.4 \cdot 10^{-2}$.

h	CG		GMRES		Richardson	
	s	iter	s	iter	s	iter
0.08	0.073	21	0.088	20	0.104	16
0.04	0.315	23	0.306	22	0.44	16
0.02	1.856	19	1.817	18	3.452	18
0.01	11.28	23	10.81	22	19.74	20

(b) $\alpha = 10^{-3}$, $J = 1.5 \cdot 10^{-2}$.

Table 2.2: Comparison of CPU time and number of iterations using the CG and GMRES method non preconditioned, and the preconditioned Richardson method.

α; we can conclude that CG and GMRES solvers have the same performances, always better than the Richardson method when opportunely preconditioned.



(a) Comparison of residual norms using CG, PCG (preconditioned with the mass matrix) and Richardson methods with $\alpha = 10^{-3}$ and mesh size $h = 0.04$.

(b) Comparison of CPU-times for $\alpha = 10^{-1}$ (P-GMRES indicates GMRES preconditioned with the mass matrix).

Figure 2.4

Then we have repeated the same test with $\alpha = 10^{-5}$ (see Table 2.4): the use of the preconditioner still helps in having a faster convergence, but we notice an increasing number of iterations. Once again, as showed in Table 2.5, as α becomes smaller the number of iterations increases. To understand the reason for which decreasing α the preconditioner looses its effectiveness, we need to analyse deeply the structure of the reduced Hessian matrix

$$H = \alpha M + MK^{-T}MK^{-1}M.$$

As noted in [30], when α is sufficiently large then H is spectrally equivalent to M and therefore M will be an effective preconditioner for H, while when α is sufficiently small then the matrix $MK^{-T}MK^{-1}M$ will be an effective preconditioner for H. For intermediate values of α, however, neither M nor $MK^{-T}MK^{-1}M$ may be effective preconditioners.

h	PCG		P-GMRES	
	s	iter	s	iter
0.08	0.025	2	0.02	2
0.04	0.084	2	0.09	2
0.02	0.76	2	0.68	2
0.01	2.79	2	2.87	2

Table 2.3: Comparison of CPU time and number of iterations with $\alpha = 10^{-1}$ using CG and GMRES solvers preconditioned with the mass matrix M.

h	CG		PCG		GMRES		P-GMRES	
	s	iter	s	iter	s	iter	s	iter
0.08	0.13	35	0.06	14	0.09	27	0.06	13
0.04	0.57	41	0.29	14	0.45	30	0.26	13
0.02	3.24	36	1.66	13	2.53	27	1.46	12
0.01	18.1	41	8.29	13	13.7	29	8.7	12

Table 2.4: Comparison of CPU-time and number of iterations with $\alpha = 10^{-5}$ using CG and GMRES solvers non-preconditioned and preconditioned with the mass matrix.

Though we are not interested here in going more deeply in the details of the design of robust preconditioner for the Hessian matrix, we want to highlight one of the most typical features that one can encounter in solving numerically this kind of problems: the different behaviour showed by the solution algorithms with respect to the value of the regularization parameter. Nevertheless this *singular* behaviour should not be surprising if we recall the role of the regularization term in the functional, i.e. it ensures the well-posedness of the problem, hence as α becomes smaller we can expect a deterioration in the convergence behaviour of the solvers. These observations will be confirmed and discussed further in the next section in the context of one-shot methods.

2.2 One-shot approach: multigrid preconditioning

As already mentioned in the introduction of the Chapter, linear-quadratic optimization problem can be generally formulated as

$$\min \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \boldsymbol{f}^T\mathbf{x} \quad \text{subject to } B\mathbf{x} = \boldsymbol{g}, \tag{2.2.1}$$

i.e. as a quadratic programming problem with equality constraint, whose optimality conditions are given by the linear system

$$\underbrace{\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}}_{\mathcal{K}} \begin{pmatrix} \mathbf{x} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{g} \end{pmatrix}. \tag{2.2.2}$$

The matrix of this linear system in saddle-point form is symmetric, indefinite, usually ill-conditioned and often very large and sparse, so for the resolution of (2.2.2) we have to use some efficient preconditioned iterative linear system solver such as Krylov subspace methods. The discussion of efficient preconditioning will be the topic of the next sections, in particular the problem will be to find a matrix $\mathcal{P}$ such that

α	10^2	10^0	10^{-2}	10^{-4}	10^{-6}	10^{-8}
s	0.55	0.56	0.64	1.1	2.75	10.01
iter	2	2	3	7	23	95

Table 2.5: Computational time and number of iterations for different values of α using PCG solver with the mass matrix as preconditioner, mesh size $h = 0.02$.

(i) $\mathcal{P}^{-1}\mathcal{K}$ has better spectral properties than $\mathcal{K}$ (i.e. lowest conditioning number and/or clustered eigenvalues),

(ii) $\mathcal{P}^{-1}\mathbf{v}$ is cheap to evaluate for any given vector $\mathbf{v}$.

The preconditioner solver will then solve the equivalent system

$$\mathcal{P}^{-1}\mathcal{K}\mathbf{v} = \mathcal{P}^{-1}\mathbf{b},$$

clearly the construction of an efficient preconditioner will be a compromise between the instances (i) and (ii). For a general introduction to numerical solution and preconditioning for saddle-point system see [8], while for an introduction to preconditioning topics and multigrid techniques see [25] and [13, 37] respectively. Various preconditioned Krylov methods have been proposed for the solution of optimality systems in PDE-constrained optimization, we only mention some recent contributions, for further details see the references cited therein. The two main strategies for the preconditioning of the KKT matrix are domain decompositions and multigrid techniques. Here we focus on the second one, some useful references for the first approach could be found in [1]. From the point of view of the multigrid preconditioning, we can roughly distinguish between two different alternatives: either the direct multigrid method, where the multigrid algorithm is implemented directly on the whole KKT system (see for instance [9, 10, 83]) or the use of multigrid schemes as inner solvers (or preconditioners) for some blocks of the KKT matrix within an outer iterative solver. We followed this strategy, making a comparison between some preconditioners recently proposed in literature. In [70, 71, 72] the authors proposed to use MINRES method with a block diagonal preconditioner with multigrid cycles (both geometrical and algebraic) applied to scalar Poisson equation and to the Stokes system, both in the stationary case; also in [70] a constraint preconditioner for PPCG (preconditioned projected conjugate gradient) method with multigrid iterations is proposed; for the time-dependent case see [86], with application to heat equation, and [87] for Stokes flows. Another different approach is to consider a non-standard inner product preconditioned conjugate gradient method, see [84] for the unconstrained case, [38] for the constrained one, also [93] for a wider perspective, and [23] where other existing approaches are interpreted in this framework. See also [39] for a space-time multigrid preconditioning method for the optimal control of the Navier-Stokes equations.

In Sections 2.2.1, 2.2.2 and 2.2.3 we introduce the one-shot approach for elliptic problems, in particular considering the example of control problem for the Laplace equation already faced using the iterative approach; we present two different block diagonal preconditioners following the work in [70, 93], then we provide some numerical experiments. Similarly, in Sections 2.2.4 and 2.2.5 we discuss the one-shot approach for the optimal control of the Stokes equations referring principally to [72].

2.2.1 Problems governed by elliptic equations

We start considering the optimal control problem (2.1.6). As showed in Example 1.4 (Section 1.2.3), setting $Y = Q = H^1_0(\Omega)$, $U = L^2(\Omega)$, $\underline{x} = (y, u)$ and defining the appropriate bilinear and linear forms, problem (2.1.6) can be formulated in the form

$$\begin{cases} \min \mathcal{J}(\underline{x}) = \dfrac{1}{2}\mathcal{A}(\underline{x}, \underline{x}) - \langle \underline{F}, \underline{x}\rangle_{X^*,X}, & \text{subject to} \\ \mathcal{B}(\underline{x}, q) = \langle G, q\rangle_{Q^*,Q} \qquad \forall q \in Q, \end{cases} \tag{2.2.3}$$

whose corresponding finite element approximation reads

$$\min \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \boldsymbol{F}^T\mathbf{x} \quad \text{subject to } B\mathbf{x} = \boldsymbol{G}. \tag{2.2.4}$$

The optimality conditions are given by

$$\underbrace{\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}}_{\mathcal{K}} \begin{pmatrix} \mathbf{x} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{F} \\ \mathbf{G} \end{pmatrix}, \tag{2.2.5}$$

where the blocks of the matrix $\mathcal{K}$ are given by

$$A = \begin{pmatrix} M & 0 \\ 0 & \alpha M \end{pmatrix}, \quad B = \begin{pmatrix} K & -M \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} \mathbf{y} & \mathbf{p} \end{pmatrix}^T, \quad \boldsymbol{G} = \boldsymbol{f}, \quad \boldsymbol{F} = \begin{pmatrix} M\mathbf{y}_d & \mathbf{0} \end{pmatrix}^T,$$

being, as usual, K the stiffness matrix resulting from the finite element discretization of the differential operator, M the mass matrix and $(\mathbf{y}, \mathbf{u}, \mathbf{p})$ the discretization of the state, control and adjoint variables. Exploiting the blocks structure in (2.2.5) we (obviously) obtain the same KKT system as in Section 2.1.3, i.e.

$$\begin{pmatrix} M & 0 & K^T \\ 0 & \alpha M & -M \\ K & -M & 0 \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} M\mathbf{y}_d \\ \mathbf{0} \\ \mathbf{f} \end{pmatrix}. \tag{2.2.6}$$

There are a number of classes of preconditioners available in the literature explicitly designed to solve different saddle-point problems arising in many areas of engineering (for an overview see [8]). However, quoting from [8, Sec. 10], "for saddle point problems, the construction of high-quality preconditioners necessitates exploiting the block structure of the problem, together with detailed knowledge about the origin and structure of the various blocks. Because the latter varies greatly from application to application, there is no such thing as the 'best' preconditioner for saddle point problems". Among the different alternatives we chose the class of block diagonal preconditioners.

2.2.2 Block diagonal preconditioners

The basic block diagonal preconditioner for system with saddle-point structure is given by (see [8])

$$\mathcal{P}_d = \begin{pmatrix} A & 0 \\ 0 & -S \end{pmatrix} \tag{2.2.7}$$

where $S = -BA^{-1}B^T$ is the Schur complement; note that assuming that both A and $\mathcal{K}$ are nonsingular implies S to be nonsingular. In our case

$$\mathcal{P}_d = \begin{pmatrix} A & 0 \\ 0 & BA^{-1}B^T \end{pmatrix} = \begin{pmatrix} M & 0 & 0 \\ 0 & \alpha M & 0 \\ 0 & 0 & \frac{1}{\alpha}M + KM^{-1}K^T \end{pmatrix}.$$

The matrix $\mathcal{T} = \mathcal{P}_d^{-1}\mathcal{K}$ is nonsingular and it is shown in [58] that it satisfies

$$(\mathcal{T} - I)\Big(\mathcal{T} - \frac{1}{2}(1+\sqrt{5})I\Big)\Big(\mathcal{T} - \frac{1}{2}(1-\sqrt{5})I\Big) = 0,$$

hence $\mathcal{T}$ is diagonalizable and has at most three distinct eigenvalues, namely $\big\{1, \frac{1}{2}(1+\sqrt{5}), \frac{1}{2}(1-\sqrt{5})\big\}$; this means that GMRES or MINRES algorithm applied to the preconditioned system with matrix $\mathcal{T}$ will terminate after at most 3 iterations. At this point it seems that we have satisfied the first request (i), i.e. the preconditioned system has better spectral properties than the original one. But what about the second request (ii)? Unfortunately forming the preconditioned system is essentially as expensive as computing the inverse of $\mathcal{K}$ using an appropriate factorization. Therefore the exact preconditioner $\mathcal{P}_d$ needs to be replaced by a suitable approximation,

$$\hat{\mathcal{P}}_d = \begin{pmatrix} \hat{A} & 0 \\ 0 & -\hat{S} \end{pmatrix} \tag{2.2.8}$$

being $\hat{A}$ and $\hat{S}$ approximations of A and S. We first note that our attention should be focused on the Schur block $S = -\frac{1}{\alpha}M - KM^{-1}K^T$; this is the only block of $\mathcal{P}_d$ that contains the PDE matrix, since the blocks (1,1) and (2,2) are simple mass matrices that we can invert without difficulties. We follow the work in [70] observing that for $10^{-1} \lesssim \alpha \lesssim 10^{-7}$ the term $\frac{1}{\alpha}M$ is smaller[1] than $KM^{-1}K^T$, hence we can consider the following approximated preconditioner

$$\hat{\mathcal{P}}_d = \begin{pmatrix} M & 0 & 0 \\ 0 & \alpha M & 0 \\ 0 & 0 & KM^{-1}K^T \end{pmatrix}.$$

Solving the preconditioned system $\hat{\mathcal{P}}_d^{-1}\mathcal{K}\mathbf{v} = \hat{\mathcal{P}}_d^{-1}\mathbf{b}$, requires, at each iteration of the Krylov subspace method, to compute the preconditioned residual $\mathbf{z}$ by solving

$$\mathcal{P}_d\mathbf{z} = \mathbf{r},$$

denoting with $\mathbf{r} = (\mathbf{r}_y, \mathbf{r}_u, \mathbf{r}_p)^T$ the residual, we obtain explicitly

$$\mathbf{z} = \begin{pmatrix} M^{-1} & 0 & 0 \\ 0 & \frac{1}{\alpha}M^{-1} & 0 \\ 0 & 0 & K^{-T}MK^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{r}_y \\ \mathbf{r}_u \\ \mathbf{r}_p \end{pmatrix},$$

still requiring to solve for K and K^T, which is equivalent to solving *exactly* the state and the adjoint problem. Being the preconditioner just an *approximate solver*, the idea is to replace the exact matrices M and K with suitable approximations. Thus we define the following preconditioner

[1] The relation could be verified comparing the order of magnitude of the entries in the two matrices in terms of the mesh size h.

$$\hat{\mathcal{P}}_d = \begin{pmatrix} \hat{M} & 0 & 0 \\ 0 & \alpha\hat{M} & 0 \\ 0 & 0 & \hat{K}M^{-1}\hat{K}^T \end{pmatrix}, \tag{2.2.9}$$

where $\hat{K}$ e $\hat{M}$ are suitable approximations of K and M, respectively. Due to its mesh independent condition number, it is not difficult to find a good approximation $\hat{M}$ for the mass matrix M, for example we could simply use the lumped mass matrix or even the diagonal of the mass matrix, or more involved approximation, like performing some symmetric Gauss-Seidel iterations or some iterations of a simple iterative method accelerated by the Chebyschev semi-iteration (see [70, 84] for further details). As regards $\hat{K}$, i.e the approximation of the PDE operator, the idea is to use some well-known and effective preconditioner for the matrix K. A fixed number of multigrid iterations is known to be an optimal preconditioner for the Laplace equation (e.g. [25, 13, 37]), hence, following [70], we can choose $\hat{K}$ and $\hat{M}$ such that:

- $\hat{K}$ denotes k Algebraic Multigrid (AMG) V-cycles (alternatively k Geometric Multigrid V-cycles, not considered here),
- $\hat{M}$ denotes m steps of the symmetric Gauss-Seidel method (in short SGS(m)).

In particular for the algebraic multigrid we use the `HSL_MI20` package [11] applied via a Matlab interface, while the symmetric Gauss-Seidel have been implemented as reported in Algorithm 2.3.

```
**Input:** $A$, $b$, $\mathbf{x}_0$, $m$
set $D = \mathtt{diag}(A)$, $E = -\mathtt{tril}(A)$, $F = -\mathtt{triu}(A)$, $k = 0$
**while** $k < m$ **do**
    $\mathbf{x}^{(k+1/2)} = (D - E)\backslash(F\mathbf{x}^{(k)}) + (D - E)\backslash\mathbf{b}$
    $\mathbf{x}^{(k+1)} \;\;= (D - F)\backslash(E\mathbf{x}^{(k+1/2)}) + (D - F)\backslash\mathbf{b}$
    $k = k + 1$
**end while**
```

Algorithm 2.3: Symmetric Gauss-Seidel (`SGS`)

Remark 2.1. As already mentioned, for the approximation of the mass matrix M there are other available alternatives, like using simply the lumped mass matrix, which is less costly and time-consuming than applying a fixed number of symmetric Gauss-Seidel iterations. However, using a poor approximation for the mass matrix often results in a higher number of iterations of the outer solver, i.e. the Krylov subspace solver. Hence using a more accurate approximation for the the mass matrix seems to be a good compromise to bring down the number of iterations of the outer solver, as well as the overall CPU-time.

Note that both $\hat{K}$ and $\hat{M}$ are not formed explicitly but are defined implicitly, i.e. are defined through the action of their inverse; in fact when we have to solve the system $\hat{K}w = z$ instead of forming the matrix $\hat{K}$ we apply a routine, say $w = \mathtt{amg\text{-}precon}(K, z)$, that implement the multigrid method. In Algorithm 2.4 we summarize the main steps required by the application of the preconditioner $\hat{\mathcal{P}}_d$ at each iteration of the Krylov subspace method.

```
Input: $\mathbf{r} = (\mathbf{r}_u\, \mathbf{r}_y\, \mathbf{r}_p)^T$
$\mathbf{z}_y = \texttt{SGS}(M, \mathbf{r}_y, m)$
$\mathbf{z}_u = \alpha^{-1}\texttt{SGS}(M, \mathbf{r}_u, m)$
$\mathbf{w}_1 = \texttt{amg-precon}(K^T, \mathbf{r}_p)$
$\mathbf{w}_2 = M\mathbf{w}_1$
$\mathbf{z}_p = \texttt{amg-precon}(K, \mathbf{w}_2)$
```

Algorithm 2.4: Application of the block preconditioner $\hat{\mathcal{P}}_d$: $\hat{\mathcal{P}}_d \mathbf{z} = \mathbf{r}$

The block diagonal preconditioner $\hat{\mathcal{P}}_d$ is proved to be an optimal preconditioner [70], in the sense that its performances are independent of the mesh size h, property that often is referred to as robustness with respect to h. However, as we are going to verify with some numerical tests in the next section, the preconditioner $\hat{\mathcal{P}}_d$ is not robust with respect to the regularization parameter α, in particular as α becomes smaller the number of iterations of the outer solver increases. Since in many applications, for instance in the context of data assimilation, the regularization parameter α could be very small, we wish to have a preconditioner robust with respect to both h and α. Among the different alternatives, we mention the preconditioners proposed in [84] and [93]. In the first work the authors, exploiting the saddle-point structure (also at the continuous level) of the problem, developed a robust preconditioner to be used within a non-standard inner product conjugate gradient method; in the successive work [93] the author, applying similar ideas (i.e. finding suitable norms that give robust stability estimates in Theorem A.4), proposed a robust block diagonal preconditioner.

Here we only introduce the last one without going in the details of its derivation, for which we refer to the original work. The main difference with the previous approach is that the preconditioner is designed for solving the reduced system

$$\underbrace{\begin{pmatrix} M & K^T \\ K & -\alpha^{-1}M \end{pmatrix}}_{\mathcal{K}_2} \begin{pmatrix} \mathbf{y} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} M\mathbf{y}_d \\ \mathbf{f} \end{pmatrix}. \tag{2.2.10}$$

obtained substituting $\mathbf{u} = \alpha^{-1}\mathbf{p}$ in the original optimality system (2.2.6). Then the block diagonal preconditioner proposed is

$$\mathcal{P}_{dr} = \begin{pmatrix} M + \alpha^{1/2}K & \\ 0 & \alpha^{-1}M + \alpha^{-1/2}K^T \end{pmatrix},$$

and it is shown to be a robust preconditioner with respect to both h and α for the matrix $\mathcal{K}_2$. Clearly the application of $\mathcal{P}_{dr}$ still requires the exact inversion of the two blocks; in order to ensure an efficient evaluation of $\mathcal{P}_{dr}^{-1}\mathbf{r}$, the two diagonal blocks are replaced by suitable multigrid iterations (the approximation of $\mathcal{P}_{dr}$ will be denoted with $\hat{\mathcal{P}}_{dr}$).

To make a correct comparison between the two preconditioner $\hat{\mathcal{P}}_d$ and $\hat{\mathcal{P}}_{dr}$ we consider the analogue of $\mathcal{P}_d$ when dealing with system (2.2.10)

$$\mathcal{P}_{d2} = \begin{pmatrix} M & 0 \\ 0 & \alpha^{-1}M + KM^{-1}K^T \end{pmatrix},$$

which can be approximated as

$$\hat{\mathcal{P}}_{d2} = \begin{pmatrix} \hat{M} & 0 \\ 0 & \hat{K}M^{-1}\hat{K}^T \end{pmatrix}.$$

In the next section we propose some numerical comparisons between the two preconditioners.

2.2.3 Comparison and numerical results

We test the performances of the preconditioners on the example (2.1.13). As in Section 2.1.4 for the discretization we use $\mathbb{P}_1$ continuous finite element both for the state, the adjoint and the control variables. Firstly we make some tests to set up the multigrid routine, then we run some tests with different values of the regularization parameter α and of the mesh size h (the size of the saddle-point system related to each level of refinement is given in Table 2.7b); the tolerance in the stopping criterion of the iterative solver is fixed to 10^{-7}. The benchmark will be the sparse direct solver provided by MATLAB, we anticipate that for 2D problems of moderate size sparse direct solvers demonstrate to be very competitive, while for 3D problems they lose their effectiveness due to the intrinsic storage and computational limitations, thus in this case it is necessary to turn to iterative methods (see e.g. [38, 70] for significant numerical examples).

h	Smooth. iter = 2		Smooth. iter = 6		Smooth. iter = 10	
	s	iter	s	iter	s	iter
0.08	0.085	22	0.054	12	0.051	12
0.04	0.398	28	0.262	14	0.31	14
0.02	2.57	48	1.15	20	1.19	18
0.01	14.1	64	4.8	22	4.3	20

(a) 1 V-cycle

h	Smooth. iter = 2		Smooth. iter = 6		Smooth. iter = 10	
	s	iter	s	iter	s	iter
0.08	0.05	12	0.046	10	0.051	10
0.04	0.27	14	0.24	10	0.28	10
0.02	1.24	22	1.04	12	0.88	10
0.01	5.91	28	3.13	14	3.34	12

(b) 2 V-cycle

Table 2.6: Comparison of CPU time and iterations with $\alpha = 10^{-2}$ using GMRES with $\hat{\mathcal{P}}_d$ as preconditioner on the full (3x3) system. In table (a) we fix the number of V-Cycles equal to 1 and change the number of smoothing iterations. In table (b) we use 2 V-cycles and change the number of smoothing iterations. In both the tables we use 8 Symmetric Gauss-Seidel iterations to approximate the mass matrix and the damped Jacobi as smoother for the AMG routine.

First of all we made some tests using the GMRES solver preconditioned with $\hat{\mathcal{P}}_d$ with different settings for the AMG (Algebraic MultiGrid) routine provided by the HSL_MI20 library, which allows to change the most significant parameters like the number of V-cycle, the number of smoothing iterations, the coarse solver and the smoother. In particular we made a comparison varying the number of V-cycles and the number of smoothing iterations, the results given in Table 2.6 show that a good compromise is to take 2 V-cycles with 6 smoothing iterations. For the approximation of the mass matrix we find that 8 iterations of Symmetric Gauss-Seidel (SGS) guarantees a sufficient accuracy. The final setup for the preconditioner $\hat{\mathcal{P}}_d$ is shown in Table 2.7a. Note that with these settings, in the preliminary test with $\alpha = 10^{-2}$ reported in Table 2.6 we obtain a number of iterations approximately independent of the mesh size.

# V-cycle iterations	2
smoother	damped Jacobi
# pre-smoothing iterations	6
# post-smoothing iterations	6
coarse solver	Gauss-Seidel
# SGS iterations	8

(a) Setup of multigrid and SGS routines for the preconditioner $\hat{\mathcal{P}}_d$.

h	0.08	0.04	0.02	0.01
$\mathcal{N}$	1356	5994	25092	101766

(b) Size of the full (3x3) saddle-point system in correspondence of different levels of refinement.

Table 2.7

Afterwards we made a test comparing the number of iterations and the computational times for different values of α using the preconditioned GMRES solver and the sparse direct solver on the full system. Some comments about the results given in Table 2.8:

- the number of iterations for both the values of α is independent of the mesh size h, confirming the robustness of the preconditioner with respect to h;
- as α becomes smaller the number of iterations increases and the overall performance of the preconditioner deteriorate, confirming that the preconditioner $\hat{\mathcal{P}}_d$ is not robust with respect to α;
- the comparison with the direct solver is quite embarrassing, fixing $h = 0.01$ for $\alpha = 10^{-3}$ the direct solver is three times faster than the preconditioned GMRES, for $\alpha = 10^{-5}$ eleven times faster. Not only, the performances of the direct solver seems not to be affected by the value of α.

h	backslash 3x3		$\hat{\mathcal{P}}_d$	
	s	iter	s	iter
0.08	0.005	-	0.06	14
0.04	0.038	-	0.32	14
0.02	0.18	-	1.22	14
0.01	1.06	-	3.74	16

(a) $\alpha = 10^{-3}$.

h	backslash 3x3		$\hat{\mathcal{P}}_d$	
	s	iter	s	iter
0.08	0.0062	-	0.16	40
0.04	0.041	-	0.98	40
0.02	0.227	-	3.56	42
0.01	0.99	-	11.5	42

(b) $\alpha = 10^{-5}$.

Table 2.8: Comparison of computational time and number of iterations using GMRES preconditioned with $\hat{\mathcal{P}}_d$ and the sparse direct solver (needless to say, there is no *number of iterations* for the last one).

Then we consider the reduced optimality system, i.e. the optimality system without the optimality equation (in some tables will be referred to as ‘2x2 system’). We compare the performances of the preconditioner $\hat{\mathcal{P}}_{d2}$ and $\hat{\mathcal{P}}_{dr}$, having as benchmark the computational time of the direct solver (now on the 2x2 system, not on the full system). Some results are given in Table 2.9 and in Figure 2.5. In particular, as expected, while the preconditioner $\hat{\mathcal{P}}_{d2}$ has the same limits of $\hat{\mathcal{P}}_d$, the preconditioner $\hat{\mathcal{P}}_{dr}$ shows to be robust with respect to both α and h. Also the computational times are more competitive and, most important, independent of the regularization parameter α.

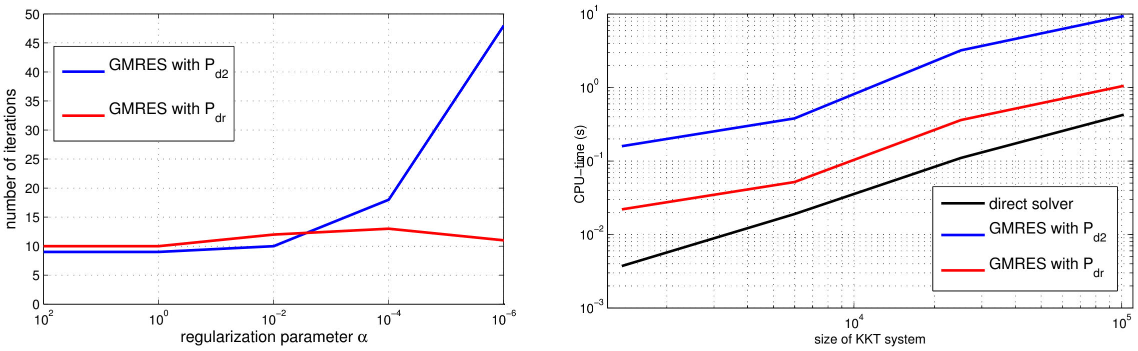
h	backslash 2x2		$\hat{\mathcal{P}}_{d2}$		$\hat{\mathcal{P}}_{dr}$	
	s	iter	s	iter	s	iter
0.08	0.0035	-	0.037	11	0.019	12
0.04	0.015	-	0.10	11	0.065	13
0.02	0.12	-	0.82	12	0.34	13
0.01	0.55	-	2.55	12	1.33	14

(a) $\alpha = 10^{-3}$.

backslash 2x2		$\hat{\mathcal{P}}_{d2}$		$\hat{\mathcal{P}}_{dr}$	
s	iter	s	iter	s	iter
0.0037	-	0.094	28	0.023	11
0.019	-	0.25	27	0.05	11
0.095	-	1.55	30	0.28	11
0.51	-	5.28	30	1.21	13

(b) $\alpha = 10^{-5}$.

Table 2.9: Comparison of CPU-time and number of iterations using GMRES preconditioned with $\hat{\mathcal{P}}_{d2}$ and $\hat{\mathcal{P}}_{dr}$ and the direct solver on the reduced (2x2) system.



(a) Comparison of number of iterations for different values of α, $\hat{\mathcal{P}}_{d2}$ vs $\hat{\mathcal{P}}_{dr}$ (mesh size h = 0.02).

(b) Comparison of CPU-times for $\alpha = 10^{-6}$ using GMRES preconditioned with $\hat{\mathcal{P}}_{d2}$, $\hat{\mathcal{P}}_{dr}$ and the direct solver on the 2x2 system.

Figure 2.5

In Figure 2.6 we compare the computational times for different solvers, using both the one-shot and the reduced Hessian approach. In particular, with a fixed $\alpha = 10^{-3}$, we have tested the following solvers: the Richardson method and the CG method for the reduced Hessian system both preconditioned with the mass matrix (see Section 2.1.4), the GMRES solver on the full KKT system ('3x3' system) preconditioned with $\hat{\mathcal{P}}_d$, the GMRES solver on the reduced '2x2' system preconditioned with $\hat{\mathcal{P}}_{dr}$, and finally the direct solver on the '2x2' and '3x3' systems.

2.2.4 Problems governed by the Stokes system: a block diagonal preconditioner

We now extend the ideas introduced in the previous section to an optimal control problem for the Stokes equations, in particular we follow the recent work [72]. Let us consider the following optimal control problem for the Stokes equations, for the sake of simplicity with control and observation on the whole domain Ω,

$$\begin{aligned} &\text{minimize} \quad J(\boldsymbol{v}, p, \boldsymbol{u}) = \frac{1}{2}\|\boldsymbol{v} - \boldsymbol{v}_d\|^2_{L^2(\Omega)} + \frac{\delta}{2}\|p - p_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|\boldsymbol{u}\|^2_{L^2(\Omega)}, \\ &\text{s.t.} \quad \begin{cases} -\nu\Delta\boldsymbol{v} + \nabla p = \boldsymbol{u} & \text{in } \Omega, \\ \operatorname{div}\boldsymbol{v} = 0 & \text{in } \Omega, \\ \boldsymbol{v} = \boldsymbol{g}_D & \text{on } \partial\Omega. \end{cases} \end{aligned} \tag{2.2.11}$$

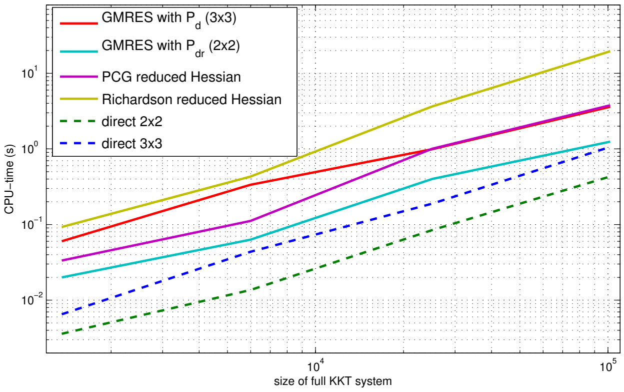


Figure 2.6: Comparison between some solvers discussed in the context of reduced Hessian methods and the one-shot approach. The regularization parameter is fixed to $\alpha = 10^{-3}$. The results confirm the initial observation that for 2D problems of moderate size direct solvers outperform the iterative ones.

Note that we are considering a three terms functional where, in addition to the usual two terms for the state velocity and the control velocity, we are considering a new term depending on the state pressure, in particular the constant $\delta > 0$ enable us to penalize the pressure. The importance of this new terms will become clear in the following.

After the discretization with a stable pair of finite element spaces for the velocity and the pressure (a detailed description can be found in [69]) the discretized state equation in algebraic form reads

$$\underbrace{\begin{pmatrix} A_s & B_s^T \\ B_s & 0 \end{pmatrix}}_{K} \begin{pmatrix} \mathbf{v} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} M_v \\ 0 \end{pmatrix} \mathbf{u} + \underbrace{\begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix}}_{\mathbf{f}_s},$$

where $(\mathbf{v}, \mathbf{p}, \mathbf{u})$ are the discrete velocity, pressure and control variables, respectively, M_v is the velocity mass matrix, A_s denotes the stiffness matrix representing the vector Laplace operator, B_s denotes the matrix representation of the divergence operator, and $\mathbf{f}$ and $\mathbf{g}$ take in account for the non-homogeneous Dirichlet boundary condition for the velocity. The discrete counterpart of (2.2.11) reads

$$\begin{aligned} \text{minimize} \quad & J(\mathbf{v}, \mathbf{p}, \mathbf{u}) = \frac{1}{2}\mathbf{v}^T M_v \mathbf{v} - \mathbf{v}^T M_v \mathbf{v}_d + \frac{\delta}{2}\mathbf{p}^T M_p \mathbf{p} + \frac{\alpha}{2}\mathbf{u}^T M_v \mathbf{u} + \frac{1}{2}\mathbf{v}_d^T M_v \mathbf{v}_d, \\ \text{s.t.} \quad & \begin{cases} A_s \mathbf{v} + B_s^T \mathbf{p} = M_v \mathbf{u} + \mathbf{f} \\ B_s \mathbf{v} = \mathbf{g}, \end{cases} \end{aligned} \tag{2.2.12}$$

being M_p the pressure mass matrix. Introducing the discrete adjoint velocity $\mathbf{w}$ and pressure $\mathbf{q}$ we obtain the optimality system

$$\begin{pmatrix} M_v & 0 & 0 & A_s & B_s^T \\ 0 & \delta M_p & 0 & B_s & 0 \\ 0 & 0 & \alpha M_v & -M_v & 0 \\ A_s & B_s^T & -M_v & 0 & 0 \\ B_s & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ \mathbf{p} \\ \mathbf{u} \\ \mathbf{w} \\ \mathbf{q} \end{pmatrix} = \begin{pmatrix} M_v \mathbf{v}_d \\ \delta M_p \mathbf{p}_d \\ \mathbf{0} \\ \mathbf{f} \\ \mathbf{g} \end{pmatrix},$$

that we can rewrite with respect to the *aggregated* variables $\mathbf{V}$, $\mathbf{U}$, $\mathbf{W}$ as

$$\begin{pmatrix} M & 0 & K \\ 0 & \alpha M_v & -E^T \\ K & -E & 0 \end{pmatrix} \begin{pmatrix} \mathbf{V} \\ \mathbf{U} \\ \mathbf{W} \end{pmatrix} = \begin{pmatrix} \mathbf{f}_a \\ \mathbf{0} \\ \mathbf{f}_s \end{pmatrix}, \tag{2.2.13}$$

where

$$M = \begin{pmatrix} M_v & 0 \\ 0 & \delta M_p \end{pmatrix}, \qquad E = \begin{pmatrix} M_v \\ 0 \end{pmatrix}, \qquad \mathbf{V} = \begin{pmatrix} \mathbf{v} \\ \mathbf{p} \end{pmatrix}, \qquad \mathbf{U} = \mathbf{u}, \qquad \mathbf{W} = \begin{pmatrix} \mathbf{w} \\ \mathbf{q} \end{pmatrix},$$

i.e. $\mathbf{V}$ is the state variable (velocity and pressure), $\mathbf{U}$ is the control variable and $\mathbf{W}$ is the adjoint variable (velocity and pressure). Obviously the optimality system could have been obtained using the saddle-point formulation (1.2.32), thus resulting in the saddle-point system

$$\underbrace{\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}}_{\mathcal{K}} \begin{pmatrix} \mathbf{X} \\ \mathbf{W} \end{pmatrix} = \begin{pmatrix} \mathbf{F} \\ \mathbf{G} \end{pmatrix},$$

where the blocks A and B are given by

$$A = \begin{pmatrix} M & 0 \\ 0 & \alpha M_v \end{pmatrix}, \qquad B = \begin{pmatrix} K & -E \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} \mathbf{V} \\ \mathbf{U} \end{pmatrix},$$

and $\mathbf{X}$ denotes the state and control variables. Noting the analogies between system (2.2.13) and the optimality system of the control problem for the Laplace equation discussed in the previous sections, we look for a block diagonal preconditioner of the form

$$\mathcal{P}_d = \begin{pmatrix} A & 0 \\ 0 & -S \end{pmatrix} \tag{2.2.14}$$

being $S = -BA^{-1}B^T$ the Schur complement.

Remark 2.2. The matrix S is nonsingular thanks to the non-singularity of the matrix $\mathcal{K}$ and A, in particular while the matrix $\mathcal{K}$ is invertible thanks to the well-posedness of the discrete problem, the matrix A has full rank thanks to the the fact that we are observing on the whole domain Ω and the presence of the pressure term in the functional. If we consider problems with observation only on a portion of the domain or without the pressure term in the functional, the diagonal preconditioner we are going to discuss can not be used as it is, since it requires the Schur complement to be nonsingular.

Exploiting the blocks structure in (2.2.14) we obtain

$$\mathcal{P}_d = \begin{pmatrix} A & 0 \\ 0 & BA^{-1}B^T \end{pmatrix} = \begin{pmatrix} M & 0 & 0 \\ 0 & \alpha M_v & 0 \\ 0 & 0 & \frac{1}{\alpha} E M_v^{-1} E^T + K M^{-1} K^T \end{pmatrix}.$$

As in the elliptic case, the exact preconditioner $\mathcal{P}_d$, in order to be efficiently evaluable, needs to be replaced by a suitable approximation

$$\hat{\mathcal{P}}_d = \begin{pmatrix} \hat{A} & 0 \\ 0 & -\hat{S} \end{pmatrix}, \tag{2.2.15}$$

being $\hat{A}$ and $\hat{S}$ approximations of A and S. Once again, we focus our attention on the Schur block S, i.e. $\alpha^{-1}EM_vE^T + KM^{-1}K^T$, since this is the only block of $\mathcal{P}_d$ that contains the PDE operator. Observing that for all but the smallest values of α the term $KM^{-1}K^T$ is dominant respect to $\alpha^{-1}EM_vE^T$ (see [72]), we can consider the following approximated preconditioner

$$\hat{\mathcal{P}}_d = \begin{pmatrix} M & 0 & 0 \\ 0 & \alpha M_v & 0 \\ 0 & 0 & KM^{-1}K^T \end{pmatrix}.$$

The last step consist in substituting the matrix M and K with suitable approximations $\hat{M}$ and $\hat{K}$, resulting in the preconditioner

$$\hat{\mathcal{P}}_d = \begin{pmatrix} \hat{M} & 0 & 0 \\ 0 & \alpha \hat{M} & 0 \\ 0 & 0 & \hat{K}M^{-1}\hat{K}^T \end{pmatrix}. \tag{2.2.16}$$

What still remains to be discussed is the choice of the approximations $\hat{M}$ and $\hat{K}$. As usual the idea is to use some effective existing preconditioners, in the case of the mass matrix M, as already discussed in Section 2.2.2, we could simply use the lumped mass matrix or even the diagonal of the mass matrix, or more involved approximation like performing some symmetric Gauss-Seidel iterations. As regards $\hat{K}$, we recall that K represents the full Stokes operator, i.e. a saddle-point matrix for which the task of preconditioning is not straightforward. Moreover, we have to guarantee that not only $\hat{K}$ be an effective preconditioner for K, but also that $\hat{K}M^{-1}\hat{K}^T$ be a good approximation of $KM^{-1}K^T$. Using some results proved in [12], the authors in [72] shows that a good choice is to take $\hat{K}$ implicitly defined by the application of a preconditioned iterative method. For example one can use a preconditioned Krylov subspace method, for which several effective preconditioners are available. However using a Krylov subspace method as inner solver for the Stokes operator leads to a nonstationary preconditioner for the whole system, that is permissible as long as we use a flexible outer method, i.e. for example flexible GMRES. A valid alternative is to use as inner solver a stationary iterative method, like the Uzawa method (see e.g. [8]). In this case to solve the system $K\mathbf{w} = \mathbf{d}$ the action of the approximation $\hat{K}$ is implicitly defined by the following iterative scheme:

$$\mathbf{w}^{(m+1)} = \mathbf{w}^{(m)} + \mathcal{M}^{-1}\mathbf{r}^{(m)},$$

where the matrix $\mathcal{M}$ is given by

$$\mathcal{M} = \begin{pmatrix} \hat{A}_s & 0 \\ B_s & -\hat{M}_p \end{pmatrix},$$

and the residual is defined as $\mathbf{r}^m = \mathbf{d} - K\mathbf{w}^m$. Once again $\hat{A}_s$ and $\hat{M}_p$ are approximation of A_s and M_p (being M_p itself an approximation of the Schur complement for the Stoke system, $S_s = -B_sA_s^{-1}B_s^T$). In particular for $\hat{M}_p$ we can use one of the several alternatives already mentioned, while for $\hat{A}_s$ an effective choice is to use a fixed number of multigrid iterations. The preconditioner $\hat{\mathcal{P}}_d$ is proved [72] to be robust with respect to the mesh size h but not to the regularization parameter α and the parameter δ. In Algorithms 2.5 and 2.6 we summarize the main steps required by the application of the preconditioner $\mathcal{P}_d$, for the sake of simplicity we do not include the approximations of the mass matrices, using instead the exact ones.

```
Input: $\mathbf{r} = (\mathbf{r}_v\, \mathbf{r}_p\, \mathbf{r}_u\, \mathbf{r}_w\, \mathbf{r}_q)^T$
$\mathbf{z}_v = M_v \backslash \mathbf{r}_v$
$\mathbf{z}_p = \delta^{-1} M_p \backslash \mathbf{r}_p$
$\mathbf{z}_u = \alpha^{-1} M_v \backslash \mathbf{r}_u$
set $\mathbf{r}_{wq} = (\mathbf{r}_w\ \mathbf{r}_q)^T$
$\mathbf{w}_{wq} = \texttt{uzawa(}K, \mathbf{r}_{wq}, m_{max}\texttt{)}$
$\mathbf{w}_{wq} = M \backslash \mathbf{w}_{wq}$
$\mathbf{z}_{wq} = \texttt{uzawa(}K, \mathbf{w}_{wq}, m_{max}\texttt{)}$
```

Algorithm 2.5: Application of the block preconditioner $\hat{\mathcal{P}}_d$: $\hat{\mathcal{P}}_d \mathbf{z} = \mathbf{r}$

```
Input: $\mathbf{b}_{wq} = (\mathbf{b}_w\ \mathbf{b}_q)^T$
set $\mathbf{x}^{(0)} = \mathbf{0}$, $\mathbf{r}^{(0)} = \mathbf{b} - K\mathbf{x}^{(0)}$, $m = 1$
while $m < m_{max}$ do
    solve the linear system $\mathcal{M}\mathbf{z}^{(m)} = \mathbf{r}^{(m)}$, i.e.
        $\mathbf{z}_w^{(m)} = \texttt{amg-precon(}A_s, \mathbf{r}_w^{(m)}\texttt{)}$
        $\mathbf{z}_q^{(m)} = M_p \backslash (B_s \mathbf{z}_w^{(m)} - \mathbf{r}_q^{(m)})$
    update the solution $\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} + \mathbf{z}^{(m)}$
    update the residual $\mathbf{r}^{(m+1)} = \mathbf{r}^{(m)} - K\mathbf{z}^{(m)}$
    set $m = m + 1$
end while
```

Algorithm 2.6: Uzawa method: $\mathbf{x}_{wq} = \texttt{uzawa(}K, \mathbf{b}_{wq}, m_{max}\texttt{)}$

Remark 2.3. As in the case of control problems governed by elliptic equations, one can also consider preconditioners for the reduced system obtained eliminating the optimality equation in the full KKT system. Note also that using this formulation it is also possible to avoid to add the pressure term in the functional and is possible to consider observation of the velocity on a portion of the domain (instead of the whole domain), see for instance [93].

2.2.5 Comparision and numerical tests

Let $\Omega = (0,1)^2$, we consider the following problem (a slightly modified numerical example proposed in [93])

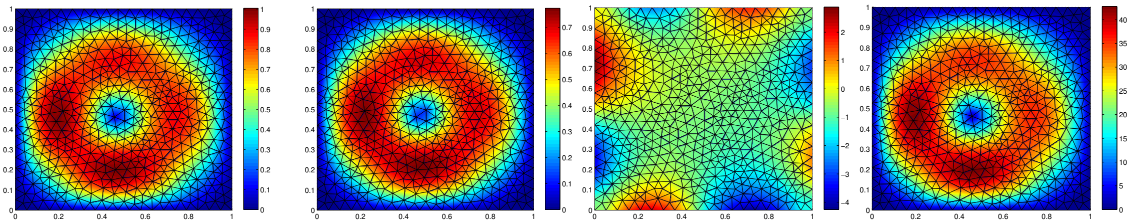
$$\begin{aligned} &\text{minimize} \quad J(\boldsymbol{v}, p, \boldsymbol{u}) = \frac{1}{2}\|\boldsymbol{v} - \boldsymbol{v}_d\|^2_{L^2(\Omega)} + \frac{\delta}{2}\|p - p_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|\boldsymbol{u}\|^2_{L^2(\Omega)}, \\ &\text{s.t.} \quad \begin{cases} -\nu\Delta\boldsymbol{v} + \nabla p = \boldsymbol{u} & \text{in } \Omega \\ \operatorname{div}\boldsymbol{v} = 0 & \text{in } \Omega \\ \boldsymbol{v} = \mathbf{0} & \text{on } \partial\Omega, \end{cases} \end{aligned} \tag{2.2.17}$$

where $p_d = 0$ and the desired velocity is given by $\boldsymbol{v}_d = (v_{d1}, v_{d2})$ with

$$v_{d1} = 10\frac{\partial}{\partial x_2}(\varphi(x_1)\varphi(x_2)), \qquad v_{d2} = -10\frac{\partial}{\partial x_1}(\varphi(x_1)\varphi(x_2)),$$

and $\varphi(z) = (1 - \cos(0.8\pi z))(1 - z^2)$. The discretized KKT system reads

$$\begin{pmatrix} M_v & 0 & 0 & A_s & B_s^T \\ 0 & \delta M_p & 0 & B_s & 0 \\ 0 & 0 & \alpha M_v & -M_v & 0 \\ A_s & B_s^T & -M_v & 0 & 0 \\ B_s & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ \mathbf{p} \\ \mathbf{u} \\ \mathbf{w} \\ \mathbf{q} \end{pmatrix} = \begin{pmatrix} M_v \mathbf{v}_d \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}.$$



(a) Desired velocity $\boldsymbol{v}_d$.

(b) Optimal velocity (left), pressure (middle) and control (right) solutions for $\alpha = 10^{-4}$ and $\delta = 0$, mesh size $h = 0.05$; the value of the cost functional is $J = 2.01 \cdot 10^{-2}$.

Figure 2.7

For the discretization we use the Taylor-Hood pair of finite element spaces consisting of continuous piecewise quadratic polynomials for the velocity and continuous piecewise linear polynomials for the pressure (see [65]). A plot of the modulus of the desired velocity $\boldsymbol{v}_d$ is given in Figure 2.7a, as well as plots of the optimal velocity $\boldsymbol{v}$, pressure p and control $\boldsymbol{u}$ are given in Figure 2.7b.

We have implemented the preconditioner $\hat{\mathcal{P}}_d$ using $m_{max} = 6$ iterations of inexact Uzawa method and $k = 3$ V-cycles in the AMG routine. The numerical experiments show (as expected) a considerable variability in the performance of the preconditioner depending on the value of the regularization parameters α and δ. We report some results in Table 2.10.

$\mathcal{N}$	backslash		$\hat{\mathcal{P}}_d$	
	s	iter	s	iter
7308	0.25	-	0.87	22
16780	0.85	-	1.71	23
37404	3.17	-	4.44	22
86040	7.30	-	10.50	26

(a) $\alpha = 10^{-2}$, $\delta = 10^{-3}$.

$\mathcal{N}$	backslash		$\hat{\mathcal{P}}_d$	
	s	iter	s	iter
7308	0.25	-	1.22	37
16780	0.56	-	2.42	37
37404	2.48	-	6.34	39
86040	6.20	-	19.56	44

(b) $\alpha = 10^{-2}$, $\delta = 10^{-1}$.

Table 2.10: Comparison of CPU-time and number of iterations using GMRES preconditioned with $\hat{\mathcal{P}}_d$ and the direct solver.

Chapter 3

Reduced Basis Method for Parametrized Elliptic PDEs

In this Chapter we introduce the reduced basis (RB) approximation and a posteriori error estimation methods for the rapid and reliable solutions of parametrized partial differential equations (PPDEs). The interest in developing efficient numerical methods for the solution of this kind of problems arise in many engineering contexts. In fact, a large part of engineering problems involve the solution of partial differential equations possibly depending on a set of *input* parameters which identify a given configuration of the system, representing physical properties or geometrical variables. Since the repeated solution of this kind problems for many different parameters values can be computationally prohibitive using classical discretization techniques (like finite element method), we have to develop suitable *reduced order methods* to reduce the computational effort. The reduced basis method is one of them, particularly well-suited in *real-time* and *many-query* contexts.

Though the RB method has been applied to several classes of equations, here we shall focus on affinely parametrized linear elliptic PDEs, either coercive or noncoercive. Denoting with $\boldsymbol{\mu}$ a p-vector of parameters belonging to the parameters space $\mathcal{D} \subset \mathbb{R}^p$, we consider an abstract parametrized variational problem of the form: given $\boldsymbol{\mu} \in \mathcal{D}$, find $u(\boldsymbol{\mu}) \in V$ such that

$$a(u(\boldsymbol{\mu}), v; \boldsymbol{\mu}) = f(v; \boldsymbol{\mu}), \qquad \forall v \in V, \tag{$P_{\boldsymbol{\mu}}$}$$

being V a suitable Hilber space, and $a(\cdot, \cdot; \boldsymbol{\mu})$ and $f(\cdot; \boldsymbol{\mu})$ the bilinear and linear forms associated to the PDE. As already mentioned, the repeated solution of problem ($P_{\boldsymbol{\mu}}$) for many different parameters values is computationally prohibitive, thus requiring a suitable model order reduction strategy. From an abstract point of view, the starting assumptions of the RB method is that the mapping $\boldsymbol{\mu} \mapsto u(\boldsymbol{\mu})$ defines a *smooth* and rather *low-dimensional* parametrically induced manifold

$$\mathcal{M} = \{u(\boldsymbol{\mu}) \in V : \boldsymbol{\mu} \in \mathcal{D}\},$$

where $u(\boldsymbol{\mu})$ is the solution of ($P_{\boldsymbol{\mu}}$). Attempting to solve numerically the problem ($P_{\boldsymbol{\mu}}$) can be seen as trials to approximate the manifold $\mathcal{M}$. In a classical discretization approach, after introducing an approximation space $X^{\mathcal{N}}$ of (typically very large) dimension $\mathcal{N}$ – e.g. a finite element (FE) space – for every value of the parameters $\boldsymbol{\mu}$ we are supposed to solve the whole problem in order to compute the solution $u^{\mathcal{N}}(\boldsymbol{\mu})$. This amounts to constructing a pointwise approximation $\mathcal{M}^{\mathcal{N}}$ of the manifold $\mathcal{M}$, thus ignoring the possibly *smooth* relation between parameters and solutions. In other words, this kind of generic approximation spaces are unnecessarily rich and hence unnecessarily expensive within the parametric framework.

A reduced (basis) approach is premised upon a classical finite element method (for example) and consists in a low-order approximation of the *truth* manifold $\mathcal{M}^{\mathcal{N}}$, based on two stages: in the former we sample some parameters values in the space $\mathcal{D}$ and compute the corresponding FE solutions, which can be seen as snapshots of the *truth* manifold $\mathcal{M}^{\mathcal{N}}$; in the latter we build a lower-dimensional approximation $\mathcal{M}^{N}$ of the *truth* manifold, by means of a suitable interpolation procedure (a Galerkin projection) of the precomputed snapshots. In particular, the main ingredients of the reduced basis (RB) methods [61, 66, 77] are the following ones:

(i) a rapidly convergent global approximation (Galerkin projection) onto a space spanned by solution of the original problem at some selected parameters value;

(ii) a rigorous a posteriori error estimation procedures which provides inexpensive yet sharp bounds for the error between the RB and the *truth* solution. The a posteriori error estimation is crucial for the certification of the method as well as for the design of the sampling (Greedy) procedure used for the construction of the reduced basis;

(iii) an Offline/Online computational procedures, i.e. a splitting between a time-consuming and parameter independent Offline stage and an inexpensive Online calculation for each new input parameter $\boldsymbol{\mu}$.

In the following we introduce the main aspects of the methodology, without going deeply in the details, but rather trying to highlight and summarize the main features of the method that we aim to extend to parametrized optimal control problems in Chapter 4. In particular, the typical problem addressed in the RB context (see [77, 66]) is the rapid and reliable evaluation of an *input-output* relationship requiring the solution of a parametrized PDE. However, in view of the application to optimal control problems, we are only interested in considering a simpler problem: the rapid and reliable solution of a parametrized PDE. Hence in our presentation of the RB method we limit to treat this aspect of the problem, avoiding to discuss all the additional issues related to the efficient and reliable outputs computation (we refer e.g. to [77, 61, 66, 76]).

In the first part of the chapter we deal with affine linear elliptic coercive PDEs. In Section 3.1 we introduce the parametrized variational formulation of the problem as well as its *truth* finite element approximation. In Section 3.2 we discuss the RB approximation and the main features of the method: the Galerkin projection, the greedy sampling procedure and the Offline-Online computational stratagem. Then in Section 3.3 we deal with the a posteriori error estimation for the RB solution. Since the abstract problem introduced in Section 3.1 is still very general, in Section 3.4 we focus on the parametrized formulation of scalar advection-diffusion-reaction equations on parametrized domains, providing a heat convection-conduction numerical example to practically illustrate the RB formulation and performances.

In the second part of the chapter (i.e. Section 3.5) we apply the RB method to the parametrized Stokes equations, a particular case of noncoercive problem. After having introduced the formulation of the problem, we discuss the RB approximation and its main features. Particular attention is devoted to the a posteriori error estimation. Finally some numerical examples will be discussed.

3.1 Problem formulation

Let $\Omega \subset \mathbb{R}^d$ $(d = 1, 2, 3)$ be a spatial domain with Lipschitz boundary $\partial\Omega$, $V = V(\Omega)$ a suitable Hilbert space. Let $a(\cdot, \cdot; \boldsymbol{\mu}) : V \times V \to \mathbb{R}$ be a bilinear form, $f(\cdot; \boldsymbol{\mu}) : V \to \mathbb{R}$ and $l(\cdot; \boldsymbol{\mu}) : V \to \mathbb{R}$ continuous functionals. The typical problem considered in the RB context is the following [77, 66]: given $\boldsymbol{\mu} \in \mathcal{D} \subset \mathbb{R}^p$, evaluate the output of interest

$$s(\boldsymbol{\mu}) = l(u(\boldsymbol{\mu}); \boldsymbol{\mu}) \tag{3.1.1}$$

where $u(\boldsymbol{\mu}) \in V(\Omega)$ satisfies

$$a(u(\boldsymbol{\mu}), v; \boldsymbol{\mu}) = f(v; \boldsymbol{\mu}), \qquad \forall v \in V.$$

However, as already mentioned, in view of the application of the RB method to optimal control problems, we shall focus only on the simpler problem: given $\boldsymbol{\mu} \in \mathcal{D}$, find $u(\boldsymbol{\mu}) \in V(\Omega)$ such that

$$a(u(\boldsymbol{\mu}), v; \boldsymbol{\mu}) = f(v; \boldsymbol{\mu}), \qquad \forall v \in V. \tag{3.1.2}$$

Let us define a convenient inner product and norm on the space V,

$$(w, v)_V = a_S(w, v; \bar{\boldsymbol{\mu}}) + \tau(w, v)_{L^2(\Omega)}, \quad \|w\|_V^2 = (w, w)_V, \qquad \forall w, v \in V$$

where $\bar{\boldsymbol{\mu}} \in \mathcal{D}$, $\tau > 0$ large enough such that the resulting norm is well-defined, and $a_S(\cdot, \cdot; \boldsymbol{\mu})$ denotes the symmetric part of the bilinear form a. We assume that the bilinear form $a(\cdot, \cdot; \boldsymbol{\mu}) : V \times V \to \mathbb{R}$ is continuous and coercive over V for all $\boldsymbol{\mu} \in \mathcal{D}$, i.e.

$$\gamma(\boldsymbol{\mu}) = \sup_{w \in V} \sup_{v \in V} \frac{a(w, v; \boldsymbol{\mu})}{\|w\|_V \|v\|_V} < +\infty, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{3.1.3}$$

and there exists a positive constant α_0 such that

$$\alpha(\boldsymbol{\mu}) = \inf_{w \in V} \frac{a(w, w; \boldsymbol{\mu})}{\|v\|_V^2} \geq \alpha_0, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}. \tag{3.1.4}$$

Moreover we assume that $f(\cdot; \boldsymbol{\mu})$ is continuous over V for all $\boldsymbol{\mu} \in \mathcal{D}$. Holding these assumptions, problem (3.1.2) admits a unique solution, thanks to Lax-Milgram lemma (see Lemma A.1). We shall make an additional assumption, crucial to Offline-Online procedures, by assuming the bilinear and linear forms to be affine in the parameter $\boldsymbol{\mu}$, i.e. for some finite Q_a and Q_f they can be expressed as

$$a(w, v; \boldsymbol{\mu}) = \sum_{q=1}^{Q_a} \Theta_a^q(\boldsymbol{\mu})\, a^q(w, v), \qquad \forall v, w \in V, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{3.1.5}$$

$$f(v; \boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \Theta_f^q(\boldsymbol{\mu})\, f^q(v), \qquad \forall v \in V, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{3.1.6}$$

for given *smooth* $\boldsymbol{\mu}$-dependent functions Θ_a^q, Θ_f^q, and continuous $\boldsymbol{\mu}$-independent bilinear and linear forms $a^q(\cdot, \cdot), f^q(\cdot)$.

3.1.1 Truth approximation

We now introduce the *truth* approximation, on which we will construct our reduced basis approximation, and with respect to which we will measure the corresponding error. The problem (3.1.2) can be approximated by any kind of Galerkin method, here we consider the finite element (FE) method. Let $V^{\mathcal{N}} \subset V$ a sequence of FE approximation subspaces of V, i.e. such that $\dim(V^{\mathcal{N}}) = \mathcal{N} < +\infty$. The *truth* FE-Galerkin approximation of (3.1.2) reads: find $u^{\mathcal{N}}(\boldsymbol{\mu}) \in V^{\mathcal{N}}$ such that

$$a(u^{\mathcal{N}}(\boldsymbol{\mu}), v; \boldsymbol{\mu}) = f(v; \boldsymbol{\mu}), \qquad \forall v \in V^{\mathcal{N}}. \tag{3.1.7}$$

Let us define precisely the FE continuity and coercivity constants respectively as

$$\gamma^{\mathcal{N}}(\boldsymbol{\mu}) = \sup_{w \in V^{\mathcal{N}}} \sup_{v \in V^{\mathcal{N}}} \frac{a(w, v; \boldsymbol{\mu})}{\|w\|_V \|v\|_V}, \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{3.1.8}$$

and

$$\alpha^{\mathcal{N}}(\boldsymbol{\mu}) = \inf_{w \in V^{\mathcal{N}}} \frac{a(w, w; \boldsymbol{\mu})}{\|v\|_V^2}, \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{3.1.9}$$

note that, as remarked in Section A.2, we have $\alpha^{\mathcal{N}}(\boldsymbol{\mu}) \geq \alpha(\boldsymbol{\mu}) > 0$ and $\gamma^{\mathcal{N}}(\boldsymbol{\mu}) \leq \gamma(\boldsymbol{\mu})$, $\forall \boldsymbol{\mu} \in \mathcal{D}$, and the problem (3.1.7) is well posed.

3.2 Reduced basis approximation

The RB method efficiently computes an approximation of $u^{\mathcal{N}}(\boldsymbol{\mu})$ by using approximation spaces made up of well-chosen solutions of (3.1.7), i.e. corresponding to specific choices of the parameter values. As already mentioned in the introduction to the chapter, the main assumption is that the solution of (3.1.7) depends *smoothly* on the parameters, thus implying the parametric manifold $\mathcal{M}^{\mathcal{N}}$ to be smooth and approximable by selecting some *snapshot* FE solutions.

3.2.1 Formulation and main features

Assuming that we are given a FE approximation space $V^{\mathcal{N}}$ of dimension $\mathcal{N}$, we introduce, given a positive integer $N_{\max}$, an associated sequence of approximation spaces: for $N = 1, \dots, N_{\max}$, $V_N^{\mathcal{N}}$ is a N-dimensional subspace of $V^{\mathcal{N}}$. We assume that these spaces are hierarchical, i.e.

$$V_1^{\mathcal{N}} \subset V_2^{\mathcal{N}} \subset \cdots \subset V_{N_{\max}}^{\mathcal{N}} \subset V^{\mathcal{N}}, \tag{3.2.1}$$

a crucial property to ensure efficiency of the resulting RB approximation. Several alternatives have been proposed to define suitable RB approximating spaces satisfying property (3.2.1), in particular Lagrange, Taylor [64] and Hermite [48] spaces. While the Lagrange spaces are based only on *snapshot* FE solutions, Taylor and Hermite spaces take into account also partial derivatives of these basis solutions. In this work we focus on Lagrange spaces.

In order to define a sequence of Lagrange RB spaces $V_N^{\mathcal{N}}$, we first define a set of properly selected parameter points $\boldsymbol{\mu}^n \in \mathcal{D}$, $1 \leq n \leq N_{\max}$ and the corresponding Lagrange parameter samples

$$S_N = \{\boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^N\}, \tag{3.2.2}$$

for given $N \in \{1, \dots, N_{\max}\}$. Then we define the associated Lagrange RB spaces as

$$V_N^{\mathcal{N}} = \operatorname{span}\{u^{\mathcal{N}}(\boldsymbol{\mu}^n),\ 1 \le n \le N\}, \tag{3.2.3}$$

where the $u^{\mathcal{N}}(\boldsymbol{\mu})$ are the so called *snapshots* of the manifold $\mathcal{M}^{\mathcal{N}}$, i.e. FE solutions computed for selected parameters values $\boldsymbol{\mu}^n$. Note that by construction the Lagrange spaces $V_N^{\mathcal{N}}$ are hierarchical. The sampling strategy used to build the set S_N will be discussed in Section 3.2.3.

We now introduce the reduced basis approximation of problem (3.1.2): given $\boldsymbol{\mu} \in \mathcal{D}$, find $u_N^{\mathcal{N}}(\boldsymbol{\mu}) \in V_N$ such that

$$a(u_N^{\mathcal{N}}(\boldsymbol{\mu}), v; \boldsymbol{\mu}) = f(v; \boldsymbol{\mu}), \qquad \forall v \in V_N^{\mathcal{N}}. \tag{3.2.4}$$

In the following we omit the superscript $^{\mathcal{N}}$ to denote the RB space and the RB approximation of the solution, i.e. $V_N = V_N^{\mathcal{N}}$ and $u_N(\boldsymbol{\mu}) = u_N^{\mathcal{N}}(\boldsymbol{\mu})$, thus omitting explicitly the fact that the reduced space is built upon the truth approximation.

We now consider the discrete equation associated with the Galerkin approximation (3.2.4). In order to recover an orthonormal well-conditioned set of basis functions and to guarantee a good algebraic stability, we apply the Gram-Schmidt process [29] in the $(\cdot,\cdot)_V$ inner product to the snapshots $u^{\mathcal{N}}(\boldsymbol{\mu}^n)$, $1 \le n \le N_{\max}$. As a result we obtain the orthonormalized set of basis functions $\{\zeta_n^{\mathcal{N}}\}$ satisfying the orthogonality condition

$$(\zeta_i^{\mathcal{N}}, \zeta_j^{\mathcal{N}})_X = \delta_{ij}, \qquad 1 \le i, j \le N_{\max},$$

being δ_{ij} the Kronecker symbol. Then, expanding the RB solution

$$u_N(\boldsymbol{\mu}) = \sum_{j=1}^{N} u_{Nj}(\boldsymbol{\mu}) \zeta_m^{\mathcal{N}}, \tag{3.2.5}$$

inserting the expansion in the problem (3.2.4) and choosing $v = \zeta_i^{\mathcal{N}}$, $1 \le i \le N$, we obtain the set of linear algebraic equation

$$\sum_{j=1}^{N} a(\zeta_j^{\mathcal{N}}, \zeta_i^{\mathcal{N}}; \boldsymbol{\mu}) u_{Nj}(\boldsymbol{\mu}) = f(\zeta_i^{\mathcal{N}}; \boldsymbol{\mu}), \qquad 1 \le i \le N \tag{3.2.6}$$

for the reduced basis coefficients u_{Nj}, $1 \le j \le N$. The linear system (3.2.6) can be expressed in matrix form as

$$A_N(\boldsymbol{\mu}) \mathbf{u}_N(\boldsymbol{\mu}) = \mathbf{f}_N(\boldsymbol{\mu}), \tag{3.2.7}$$

where $(\mathbf{u}_N(\boldsymbol{\mu}))_j = u_{Nj}(\boldsymbol{\mu})$ and the matrix A_N and the vector $\mathbf{f}_N$ are given respectively by

$$(A_N(\boldsymbol{\mu}))_{ij} = a(\zeta_j^{\mathcal{N}}, \zeta_i^{\mathcal{N}}; \boldsymbol{\mu}), \qquad (\mathbf{f}_N(\boldsymbol{\mu}))_i = f(\zeta_i^{\mathcal{N}}; \boldsymbol{\mu}).$$

In the next section we describe the Offline-Online procedure that permits to solve efficiently, i.e. independently of $\mathcal{N}$, the linear system (3.2.7).

3.2.2 Offline-Online procedure

Although the linear system (3.2.7) is a low-dimensional system of size $N \times N$, the formation of the matrix $A_N(\boldsymbol{\mu})$ still involves the basis functions $\zeta_i^{\mathcal{N}}$ associated to the FE high-dimensional approximation. Thanks to the assumption of affine parameter dependence, we can decouple the formation of the matrix $A_N(\boldsymbol{\mu})$ in two stages, the Offline and Online stages, that enable the efficient resolution of the system (3.2.7) for each new parameter $\boldsymbol{\mu}$. In particular, thanks to (3.1.5) and (3.1.6), system (3.2.6) can be expressed as

$$\sum_{j=1}^{N}\left(\sum_{q=1}^{Q_a}\Theta_a^q(\boldsymbol{\mu})a^q(\zeta_j^{\mathcal{N}},\zeta_i^{\mathcal{N}})\right)u_{Nj}(\boldsymbol{\mu})=\sum_{q=1}^{Q_f}\Theta_f^q(\boldsymbol{\mu})f^q(\zeta_i^{\mathcal{N}}),$$

for $1\le i\le N$. The equivalent matrix form is

$$\left(\sum_{q=1}^{Q_a}\Theta_a^q(\boldsymbol{\mu})A_N^q\right)\mathbf{u}_N(\boldsymbol{\mu})=\sum_{q=1}^{Q_f}\Theta_f^q(\boldsymbol{\mu})\mathbf{f}_N^q, \tag{3.2.8}$$

where

$$(A_N^q)_{ij}=a^q(\zeta_j^{\mathcal{N}},\zeta_i^{\mathcal{N}}), \qquad (\mathbf{f}_N^q)_i=f^q(\zeta_i^{\mathcal{N}}).$$

The Offline-Online procedure is now clear:

1. in the Offline stage, performed only once, we first compute and store the basis function $\zeta_i^{\mathcal{N}}$, $1\le i\le N$, and form the matrices A_N^q, $1\le q\le Q_a$, and the vectors $\mathbf{f}_N^q$, $1\le q\le Q_f$. The operation count depends on N, Q_a, Q_f and $\mathcal{N}$;

2. in the Online stage, performed for each new value $\boldsymbol{\mu}$, we use the precomputed matrices A_N^q and vectors $\mathbf{f}_N^q$ to assemble the matrix A_N and the vector $\mathbf{f}_N$ appearing in (3.2.7), with
$$A_N=\sum_{q=1}^{Q_a}\Theta_a^q(\boldsymbol{\mu})A_N^q, \qquad \mathbf{f}_N=\sum_{q=1}^{Q_f}\Theta_f^q(\boldsymbol{\mu})\mathbf{f}^q;$$
we then solve the resulting system to obtain $\mathbf{u}_N$. The Online operation count depends on N, Q_a, Q_f but is independent of $\mathcal{N}$. In particular we need $O(Q_aN^2)$ and $O(Q_fN)$ operations to assemble matrices and vectors, and $O(N^3)$ operations to solve the RB linear system (3.2.7).

Let us now specify how the the RB matrices and vectors A_N^q and $\mathbf{f}_N^q$ are related to the corresponding FE quantities. We denote $\{\phi_r\}_{r=1}^{\mathcal{N}}$ a basis of the FE space $V^{\mathcal{N}}$; since each basis functions $\zeta_i^{\mathcal{N}}$ belongs to $V^{\mathcal{N}}$, they can be expanded in terms of the FE basis functions, i.e.

$$\zeta_i^{\mathcal{N}}=\sum_{r=1}^{\mathcal{N}}\zeta_{ir}^{\mathcal{N}}\phi_r, \qquad 1\le i\le N_{\max}.$$

Therefore we have that

$$a^q(\zeta_j^{\mathcal{N}},\zeta_i^{\mathcal{N}})=\sum_{r=1}^{\mathcal{N}}\sum_{s=1}^{\mathcal{N}}\zeta_{ir}a(\phi_s,\phi_r)\zeta_{js}, \qquad \mathbf{f}_N^q=\sum_{r=1}^{\mathcal{N}}\zeta_{ir}f^q(\phi_r),$$

in matrix form

$$A_N^q=Z^TA_{\mathcal{N}}^qZ, \qquad \mathbf{f}_N^q=Z^T\mathbf{f}_{\mathcal{N}}^q,$$

being $(A_{\mathcal{N}}^q)_{rs}=a^q(\phi_s,\phi_r)$, $(\mathbf{f}_{\mathcal{N}}^q)_r=f^q(\phi_r)$ and $Z=\begin{bmatrix}\boldsymbol{\zeta}_1 & \cdots & \boldsymbol{\zeta}_N\end{bmatrix}\in\mathbb{R}^{\mathcal{N}\times N}$, $1\le N\le N_{\max}$.

3.2.3 Sampling strategy: the greedy algorithm

In this section we discuss how to choose the sample points $\boldsymbol{\mu}^n$, $1 \leq n \leq N$ for a given N in an optimal way, i.e. such that the accuracy of the corresponding RB approximation is maximized. Let $\Xi_{\text{train}} \subset \mathcal{D}$ be a finite dimensional sample set, called the set of *train* samples. The cardinality of Ξ_{train} will be denoted with n_{train}, that we assume to be sufficiently large such that Ξ_{train} be a good approximation of the set $\mathcal{D}$ (a finite dimensional surrogate for $\mathcal{D}$). The idea of the greedy procedure is that, starting with a train sample Ξ_{train}, we *adaptively* select (in the sense of minimizing a suitable error indicator) N parameters $\boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^N$ and we form the hierarchical sequence of reduced basis space V_N as

$$V_N = \text{span}\{\zeta_n = u^{\mathcal{N}}(\boldsymbol{\mu}^n)\ 1 \leq n \leq N\}.$$

The crucial question is how, given the first $\boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^{N-1}$ parameters, we choose the next one, $\boldsymbol{\mu}^N$. The greedy algorithm, in each iteration N, appends to the previously *retained* snapshots $\{u^{\mathcal{N}}(\boldsymbol{\mu}^n)\}_{n=1}^{N-1}$ that particular candidate – over all candidate snapshots $u^{\mathcal{N}}(\boldsymbol{\mu})$, $\boldsymbol{\mu} \in \Xi_{\text{train}}$– which is least well approximated by V_{N-1}. In theory, to find the least well approximated $u^{\mathcal{N}}(\boldsymbol{\mu})$ we should compute for each parameter in Ξ_{train} the norm of the error $\|u^{\mathcal{N}}(\boldsymbol{\mu}) - u_{N-1}(\boldsymbol{\mu})\|_V$, being $u_{N-1}(\boldsymbol{\mu}) \in V_{N-1}$. Clearly this approach is computationally unaffordable, hence we need a sharp, rigorous and efficient estimator $\Delta_N(\boldsymbol{\mu})$ for the reduced basis error $\|u^{\mathcal{N}}(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_V$, where $u_N(\boldsymbol{\mu})$ is the RB approximated solution associated with the generic RB space V_N. The a posteriori error estimator Δ_N will be described in detail in Section 3.3.

Supposing now to have at our disposal such an estimator, we can state precisely the steps required by the greedy algorithm. Let us denote by ε_{tol} a chosen tolerance for the stopping criterium, the greedy sampling strategy can be implemented as reported in Algorithm 3.1.

```
$S_1 = \{\boldsymbol{\mu}^1\}$, compute $u^{\mathcal{N}}(\boldsymbol{\mu}^1)$
for $N = 2 : N_{\max}$ do
    $\boldsymbol{\mu}^N = \arg\max_{\boldsymbol{\mu} \in \Xi_{\text{train}}} \Delta_{N-1}(\boldsymbol{\mu})$
    $\varepsilon_{N-1} = \Delta_{N-1}(\boldsymbol{\mu})$
    if $\varepsilon_{N-1} \leq \varepsilon_{\text{tol}}$
        $N_{\max} = N - 1$
    end if
    compute $u^{\mathcal{N}}(\boldsymbol{\mu}^N)$
    $S_N = S_{N-1} \cup \{\boldsymbol{\mu}^N\}$
    $V_N = V_{N-1} \cup \text{span}\{u^{\mathcal{N}}(\boldsymbol{\mu}^N)\}$
end for
```

Algorithm 3.1: Greedy algorithm

We underline again that the key point in the algorithm is to exploit an a posteriori error bound $\Delta_N(\boldsymbol{\mu})$ efficiently computable, since at each iteration the algorithm requires to evaluate the error for all $\boldsymbol{\mu} \in \Xi_{\text{train}}$.

3.3 A posteriori error estimation

Effective a posteriori estimator for the error on the RB approximation of the field variable are crucial for both the *efficiency* and the *reliability* of the method. As regards efficiency, the error bound plays a central role in the sampling procedure: the application of the error bound permits an exhaustive exploration of the parameters domain in order to select properly the basis functions, i.e. in such a way to achieve the better accuracy with the smaller number of basis functions. Many recent works address improvements and new ideas for efficient greedy parametric exploration, see for instance [24]. As regards reliability, at the Online stage for each new value of parameter $\boldsymbol{\mu} \in \mathcal{D}$, the a posteriori estimator permits to bound the error of the RB approximation with respect to the underlying truth approximation.

In order to ensure efficiency and reliability we require the error bound to be: *rigorous*, i.e. valid for all $N \in \{1, \cdots N_{max}\}$ and for all $\boldsymbol{\mu} \in \mathcal{D}$; *sharp*, to avoid inefficient approximations by taking N too large; *efficient*, i.e. the Online operation count to compute the RB error bound must be independent of $\mathcal{N}$. The two main ingredients on which is based the construction of such an estimator are [61, 77]: the calculation of the dual norm of the residual and an effective calculation of a lower bound for the $\alpha^{\mathcal{N}}(\boldsymbol{\mu})$ coercivity constant.

3.3.1 Basic ingredients

The first ingredient we need is a suitable equation for the reduced basis approximation error (relative to the *truth* approximation). Let us define the error between the *truth* and the RB approximations $e(\boldsymbol{\mu}) := u^{\mathcal{N}}(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu}) \in V^{\mathcal{N}}$ and the residual

$$r(v; \boldsymbol{\mu}) = f(v; \boldsymbol{\mu}) - a(u_N, v; \boldsymbol{\mu}), \qquad \forall v \in V^{\mathcal{N}},$$

note that $r(\cdot; \boldsymbol{\mu}) \in (V^{\mathcal{N}})'$, i.e. the residual is a bounded linear functional on $V^{\mathcal{N}}$. The problem statements for $u^{\mathcal{N}}(\boldsymbol{\mu})$ (3.1.7) and $u_N(\boldsymbol{\mu})$ (3.2.4) and the bilinearity of $a(\cdot, \cdot; \boldsymbol{\mu})$ imply that the error satisfies the following equation

$$a(e(\boldsymbol{\mu}), v; \boldsymbol{\mu}) = r(v; \boldsymbol{\mu}), \qquad \forall v \in V^{\mathcal{N}}. \tag{3.3.1}$$

Let us write the error residual equation (3.3.1) as

$$a(e(\boldsymbol{\mu}), v) = (\hat{e}(\boldsymbol{\mu}), v)_V, \qquad \forall v \in V^{\mathcal{N}}, \tag{3.3.2}$$

where $\hat{e}(\boldsymbol{\mu}) \in V^{\mathcal{N}}$ is the Riesz representation of $r(\cdot; \boldsymbol{\mu})$, that is

$$(\hat{e}(\boldsymbol{\mu}), v)_V = r(v; \boldsymbol{\mu}), \qquad \forall v \in V^{\mathcal{N}}. \tag{3.3.3}$$

Note that the dual norm of the residual can be evaluated through its Riesz representation:

$$\|r(\cdot; \boldsymbol{\mu})\|_{V'} = \sup_{v \in V} \frac{r(v; \boldsymbol{\mu})}{\|v\|_V} = \|\hat{e}(\boldsymbol{\mu})\|_V, \tag{3.3.4}$$

this will be crucial for the Offline-Online stratagem, see Section 3.3.2. Then from the coercivity property of the bilinear form $a(\cdot, \cdot)$ immediately follows that

$$\|e(\boldsymbol{\mu})\|_V \leq \frac{\|r(\cdot; \boldsymbol{\mu})\|_{V'}}{\alpha^{\mathcal{N}}(\boldsymbol{\mu})} = \frac{\|\hat{e}(\boldsymbol{\mu})\|_V}{\alpha^{\mathcal{N}}(\boldsymbol{\mu})}. \tag{3.3.5}$$

Therefore a rigorous estimator for the error $e(\boldsymbol{\mu})$ is given by the ratio between the norm of the residual and the coercivity constant. However, both the norm of the residual and the coercivity constant depends on the parameters $\boldsymbol{\mu}$, hence we should provide a procedure ensuring an efficient (i.e. independent of $\mathcal{N}$) Online computation of these two quantities for each value of $\boldsymbol{\mu}$. As already mentioned, the Offline-Online stratagem for the computation of the norm of the residual will be discussed in Section 3.3.2. As regards the efficient evaluation of the stability factor $\alpha^{\mathcal{N}}(\boldsymbol{\mu})$, we have to introduce the second ingredient: a (positive) lower bound for the coercivity constant. In particular we require a lower bound $\alpha^{\mathcal{N}}_{\mathrm{LB}}(\boldsymbol{\mu}) : \mathcal{D} \to \mathbb{R}$ such that

$$0 < \alpha^{\mathcal{N}}_{\mathrm{LB}}(\boldsymbol{\mu}) \leq \alpha^{\mathcal{N}}(\boldsymbol{\mu}), \qquad \forall \boldsymbol{\mu} \in \mathcal{D},$$

and the Online computational time to evaluate $\boldsymbol{\mu} \to \alpha^{\mathcal{N}}_{\mathrm{LB}}(\boldsymbol{\mu})$ is independent of $\mathcal{N}$. In Section 3.3.3 we provide a methodology (the so-called Successive Constraint Method [46, 77]) to construct the requested lower bound. Supposing for the moment to have at our disposal the lower bound $\alpha^{\mathcal{N}}_{\mathrm{LB}}(\boldsymbol{\mu})$, we can define the error estimator

$$\Delta_N(\boldsymbol{\mu}) = \frac{\|\hat{e}(\boldsymbol{\mu})\|_V}{\alpha^{\mathcal{N}}_{\mathrm{LB}}(\boldsymbol{\mu})}, \tag{3.3.6}$$

from the inequality (3.3.5) we obtain immediately

$$\|u^{\mathcal{N}}(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_V \leq \Delta_N(\boldsymbol{\mu}). \tag{3.3.7}$$

Let us define the effectivity associated to the error estimator $\Delta_N(\boldsymbol{\mu})$,

$$\eta_N(\boldsymbol{\mu}) = \frac{\Delta_N(\boldsymbol{\mu})}{\|u^{\mathcal{N}}(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_V}, \tag{3.3.8}$$

for rigour, we shall insist upon effectivity ≥ 1, for sharpness we desire the effectivity as close to unity as possible. We can prove the following [61]

Proposition 3.1. *For any $N = 1, \dots, N_{max}$, the effectivity $\eta_N(\boldsymbol{\mu})$ satisfies*

$$1 \leq \eta_N(\boldsymbol{\mu}) \leq \frac{\gamma(\boldsymbol{\mu})}{\alpha^{\mathcal{N}}_{LB}(\boldsymbol{\mu})}. \tag{3.3.9}$$

Proof. The left inequality of (3.3.9) follows directly from (3.3.7). To prove the right inequality it suffices to choose $v = \hat{e}(\boldsymbol{\mu}) \in V^{\mathcal{N}}$ in (3.3.2), the continuity of the bilinear form $a(\cdot, \cdot; \boldsymbol{\mu})$ implies $\|\hat{e}(\boldsymbol{\mu})\|_V \leq \gamma(\boldsymbol{\mu})\|e(\boldsymbol{\mu})\|_V$, hence

$$\eta_N(\boldsymbol{\mu}) = \frac{\|\hat{e}(\boldsymbol{\mu})\|_V}{\alpha^{\mathcal{N}}_{\mathrm{LB}}(\boldsymbol{\mu})\|e(\boldsymbol{\mu})\|_V} \leq \frac{\gamma(\boldsymbol{\mu})}{\alpha^{\mathcal{N}}_{\mathrm{LB}}(\boldsymbol{\mu})}. \qquad \square$$

3.3.2 Offline-Online procedure

The main component of the error bound is the computation of the dual norm of the residual $\|\hat{e}(\boldsymbol{\mu})\|_V$. To develop the Offline-Online procedure we introduce the residual expansion

$$\begin{aligned} r(v; \boldsymbol{\mu}) &= f(v; \boldsymbol{\mu}) - a\left(\sum_{n=1}^{N} u_{Nn}(\boldsymbol{\mu})\zeta_n^{\mathcal{N}}, v; \boldsymbol{\mu}\right) \\ &= \sum_{q=1}^{Q_f} \Theta_f^q(\boldsymbol{\mu}) f^q(v) - \sum_{n=1}^{N} u_{Nn}(\boldsymbol{\mu}) \sum_{q=1}^{Q_a} \Theta_q^a(\boldsymbol{\mu}) a^q(\zeta_n^{\mathcal{N}}, v), \end{aligned} \tag{3.3.10}$$

obtained exploiting the affine assumption (3.1.5) and (3.1.6) and the RB representation (3.2.5). Moreover, we have from (3.3.10) and (3.3.3) that

$$(\hat{e}(\boldsymbol{\mu}), v)_V = \sum_{q=1}^{Q_f} \Theta_f^q(\boldsymbol{\mu}) f^q(v) - \sum_{n=1}^{N} u_{Nn}(\boldsymbol{\mu}) \sum_{q=1}^{Q_a} \Theta_q^a(\boldsymbol{\mu}) a^q(\zeta_n^{\mathcal{N}}, v), \tag{3.3.11}$$

and consequently

$$\hat{e}(\boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \Theta_f^q(\boldsymbol{\mu}) \mathcal{F}^q - \sum_{n=1}^{N} u_{Nn}(\boldsymbol{\mu}) \sum_{q=1}^{Q_a} \Theta_q^a(\boldsymbol{\mu}) \mathcal{L}_n^q,$$

where $\forall v \in V^{\mathcal{N}}$

$$(\mathcal{F}^q, v)_V = f^q(v), \qquad 1 \le q \le Q_f, \tag{3.3.12}$$

$$(\mathcal{L}_n^q, v)_V = -a^q(\zeta_n^{\mathcal{N}}, v), \qquad 1 \le q \le Q_a, \quad 1 \le n \le N. \tag{3.3.13}$$

Note that $\mathcal{F}^q$ is the Riesz representation of $f^q(\cdot)$ and $\mathcal{L}_n^q$ is the Riesz representation of $A_n^q(\cdot) = a^q(\zeta_n^{\mathcal{N}}, \cdot)$, computable solving parameter-independent Poisson-like problems. Finally we obtain

$$\begin{aligned} \|\hat{e}(\boldsymbol{\mu})\|_V^2 = \sum_{q=1}^{Q_f} \sum_{q'=1}^{Q_f} \Theta_f^q(\boldsymbol{\mu}) \Theta_f^{q'}(\boldsymbol{\mu}) (\mathcal{F}^q, \mathcal{F}^{q'})_V + \sum_{q=1}^{Q_a} \sum_{n=1}^{N} \Theta_a^q(\boldsymbol{\mu}) u_{Nn}(\boldsymbol{\mu}) \Bigg\{ \\ 2 \sum_{q'=1}^{Q_f} \Theta_f^{q'}(\boldsymbol{\mu}) (\mathcal{F}^{q'}, \mathcal{L}_n^q)_V + \sum_{q'=1}^{Q_a} \sum_{n'=1}^{N} \Theta_a^{q'}(\boldsymbol{\mu}) u_{Nn'}(\boldsymbol{\mu}) (\mathcal{L}_n^q, \mathcal{L}_n^{q'})_V \Bigg\} \end{aligned} \tag{3.3.14}$$

from which we can calculate the dual norm of the residual through (3.3.4). Let us summarize the Offline-Online decomposition:

1. in the Offline stage we first compute the Q_f terms $\mathcal{F}^q$ and the NQ_a terms $\mathcal{L}_n^q$ solving problems (3.3.12) and (3.3.13) respectively; then we store the scalar products $(\mathcal{F}^q, \mathcal{F}^{q'})_V$, $(\mathcal{F}^{q'}, \mathcal{L}_n^q)_V$ and $(\mathcal{L}_n^q, \mathcal{L}_n^q)_V$. The Offline operation count depends then on N, Q_a, Q_f and $\mathcal{N}$.

2. in the Online stage, for each new value of $\boldsymbol{\mu}$, we simply evaluate the sum (3.3.14) in terms of the $\Theta^q(\boldsymbol{\mu})$, u_{Nn} and the precomputed scalar products. The Online operation count is $O(N^2 Q_a^2 + 2NQ_aQ_f + NQ_f^2)$, independent of $\mathcal{N}$.

3.3.3 SCM coercivity constant lower bounds

As introduced in Section 3.3.1, the evaluation the a posteriori error estimator $\Delta_N(\boldsymbol{\mu})$ involves the computation of lower bounds for the coercivity constant $\alpha^{\mathcal{N}}(\boldsymbol{\mu})$ whose discrete version is a generalized eigenvalue problem (see Section A.3). We introduce here the Successive Constraint Method (SCM), an approach to the construction of such lower bounds that provide an efficient Offline-Online strategy which makes the Online complexity independent of $\mathcal{N}$ [46, 61, 77].

As anticipated, we want to compute $\alpha_{\mathrm{LB}}^{\mathcal{N}}(\boldsymbol{\mu})$ such that $0 < \alpha_{\mathrm{LB}}^{\mathcal{N}}(\boldsymbol{\mu}) < \alpha^{\mathcal{N}}(\boldsymbol{\mu})$, $\forall \boldsymbol{\mu} \in \mathcal{D}$ and the evaluation $\boldsymbol{\mu} \rightarrow \alpha_{\mathrm{LB}}^{\mathcal{N}}(\boldsymbol{\mu})$ be $\mathcal{N}$-independent. Let us first introduce an objective function $\mathcal{J}^{obj} : \mathcal{D} \times \mathbb{R}^{Q_a} \rightarrow \mathbb{R}$ given by

$$\mathcal{J}^{obj}(\boldsymbol{\mu}; y) = \sum_{q=1}^{Q_a} \Theta_a^q(\boldsymbol{\mu}) y_q, \tag{3.3.15}$$

where $y = (y_1, \dots, y_{Q_a})$. Thanks to the affine dependence assumption, the coercivity constant may be expressed as

$$\alpha^{\mathcal{N}}(\boldsymbol{\mu}) = \inf_{y \in \mathcal{Y}} \mathcal{J}^{obj}(\boldsymbol{\mu}; y), \tag{3.3.16}$$

where the set $\mathcal{Y} \subset \mathbb{R}^{Q_a}$ is defined by

$$\mathcal{Y} = \left\{ y \in \mathbb{R}^{Q_a} \middle| \exists w_y \in V^{\mathcal{N}} \text{ s.t. } y_q = \frac{a^q(w_y, w_y)}{\|w_y\|_V^2}, \, 1 \leq q \leq Q_a \right\}.$$

We next introduce the *continuity constraint* box

$$\mathcal{B} = \prod_{q=1}^{Q_a} \left[\inf_{w \in V^{\mathcal{N}}} \frac{a^q(w,w)}{\|w\|_V^2}, \, \sup_{w \in V^{\mathcal{N}}} \frac{a^q(w,w)}{\|w\|_V^2} \right],$$

that is bounded thanks to the continuity hypothesis. Finally, let us define the *coercivity constraint* sample as

$$C_J = \{\boldsymbol{\mu}_{\mathrm{SCM}}^1, \dots, \boldsymbol{\mu}_{\mathrm{SCM}}^J\} \subset \mathcal{D}.$$

We denote by $C_J^{M,\boldsymbol{\mu}}$ the set of M (≥ 1) points in $\mathcal{C}_J$ closest to a given $\boldsymbol{\mu}$ (if $M > J$, then $\mathcal{C}_J^{M,\boldsymbol{\mu}} = \mathcal{C}_J$).

We can now construct the lower bound. For given $\mathcal{C}_J$, $M \in \mathbb{N}$, and any $\boldsymbol{\mu} \in \mathcal{D}$, we define the *lower bound set* $\mathcal{Y}_{\mathrm{LB}}(\boldsymbol{\mu}; \mathcal{C}_J, M)$ as

$$\mathcal{Y}_{\mathrm{LB}}(\boldsymbol{\mu}; \mathcal{C}_J, M) = \left\{ y \in \mathbb{R}^{Q_a} \middle| y \in \mathcal{B}, \, \sum_{q=1}^{Q_a} \Theta_a^q(\boldsymbol{\mu}') y_q \geq \alpha^{\mathcal{N}}(\boldsymbol{\mu}'), \; \forall \boldsymbol{\mu}' \in \mathcal{C}_J^{M,\boldsymbol{\mu}} \right\}. \tag{3.3.17}$$

Since it can be proved [77] that $\mathcal{Y} \subset \mathcal{Y}_{\mathrm{LB}}(\boldsymbol{\mu}; \mathcal{C}_J, M)$, we can define our lower bound as

$$\alpha_{\mathrm{LB}}^{\mathcal{N}}(\boldsymbol{\mu}; \mathcal{C}_J, M) = \inf_{y \in \mathcal{Y}_{\mathrm{LB}}(\boldsymbol{\mu}; \mathcal{C}_J, M)} \mathcal{J}^{obj}(\boldsymbol{\mu}; y). \tag{3.3.18}$$

Hence, for given $C_J \subset \mathcal{D}$, $M \in \mathbb{N}$, it readily follows that

$$\alpha_{\mathrm{LB}}^{\mathcal{N}}(\boldsymbol{\mu}) \leq \alpha^{\mathcal{N}}(\boldsymbol{\mu}), \quad \forall \boldsymbol{\mu} \in \mathcal{D}.$$

It is important to note that the lower bound (3.3.18) is in fact a Linear-Program (LP) with Q_a design variables and $2Q_a + M$ inequality constraints. Moreover, given $\mathcal{B}$ and the set $\{\alpha^{\mathcal{N}}(\boldsymbol{\mu}') | \boldsymbol{\mu}' \in \mathcal{C}_j\}$, the operation count to evaluate $\boldsymbol{\mu} \rightarrow \alpha_{\mathrm{LB}}^{\mathcal{N}}(\boldsymbol{\mu})$ is $\mathcal{N}$-independent.

In order to construct a suitable *coercivity constraint* sample C_J, we also require an upper bound for the coercivity constant. For given C_J, $M \in \mathbb{N}$, and any $\boldsymbol{\mu} \in \mathcal{D}$, we introduce the *upper bound set* $\mathcal{Y}_{\mathrm{UB}}(\boldsymbol{\mu}; \mathcal{C}_J, M) \subset \mathbb{R}^{Q_a}$ as

$$\mathcal{Y}_{\mathrm{UB}}(\boldsymbol{\mu}; \mathcal{C}_J, M) = \{y^*(\boldsymbol{\mu}') \,|\, \boldsymbol{\mu}' \in \mathcal{C}_J^{M,\boldsymbol{\mu}}\}, \tag{3.3.19}$$

where

$$y^*(\boldsymbol{\mu}) = \arg \inf_{y \in \mathcal{Y}} \mathcal{J}^{obj}(\boldsymbol{\mu}; y).$$

We define the upper bound as

$$\alpha^{\mathcal{N}}_{\mathrm{UB}}(\boldsymbol{\mu}; \mathcal{C}_J, M) = \min_{y \in \mathcal{Y}_{\mathrm{UB}}(\boldsymbol{\mu}; \mathcal{C}_J, M)} \mathcal{J}^{obj}(\boldsymbol{\mu}; y). \tag{3.3.20}$$

Since $\mathcal{Y}_{\mathrm{UB}}(\boldsymbol{\mu}; \mathcal{C}_J, M) \subset \mathcal{Y}$, for given $C_J \subset \mathcal{D}$, $M \in \mathbb{N}$, we have that $\alpha^{\mathcal{N}}_{\mathrm{UB}}(\boldsymbol{\mu}) \geq \alpha^{\mathcal{N}}(\boldsymbol{\mu})$, $\forall \boldsymbol{\mu} \in \mathcal{D}$. Note that, given the set $\{y^*(\boldsymbol{\mu}') \,|\, \boldsymbol{\mu}' \in \mathcal{C}_J\}$, the operation count for the Online stage to evaluate $\boldsymbol{\mu} \to \alpha^{\mathcal{N}}_{\mathrm{UB}}(\boldsymbol{\mu})$ is independent of $\mathcal{N}$.

We now present a greedy algorithm (to be performed Offline) for the construction of the set $\mathcal{C}_J$. We require a *train* sample

$$\Xi_{\mathrm{train,SCM}} = \left\{\boldsymbol{\mu}^1_{\mathrm{train,SCM}}, \ldots, \boldsymbol{\mu}^{n_{\mathrm{train,SCM}}}_{\mathrm{train,SCM}}\right\} \subset \mathcal{D},$$

of $n_{\mathrm{train,SCM}}$ parameters point, and a tolerance $\varepsilon_{\mathrm{SCM}}$ which shall control the error in the lower bound prediction. The greedy procedure is given in Algorithm 3.2.

```
**Input:** $\Xi_{\mathrm{train,SCM}} \subset \mathcal{D}$, tolerance $\varepsilon_{\mathrm{SCM}} \in (0,1)$
  set $J = 1$ and choose $\mathcal{C}_1 = \{\boldsymbol{\mu}^1_{\mathrm{SCM}}\}$ arbitrarily

  compute $\eta_{M,J}(\boldsymbol{\mu}) = \dfrac{\alpha^{\mathcal{N}}_{\mathrm{UB}}(\boldsymbol{\mu}; \mathcal{C}_J, M) - \alpha^{\mathcal{N}}_{\mathrm{LB}}(\boldsymbol{\mu}; \mathcal{C}_J, M)}{\alpha^{\mathcal{N}}_{\mathrm{UB}}(\boldsymbol{\mu}; \mathcal{C}_J, M)}$

  **while** $\max_{\boldsymbol{\mu} \in \Xi_{\mathrm{train,SCM}}} \eta_{M,J}(\boldsymbol{\mu}) > \varepsilon_{\mathrm{SCM}}$ **do**
      $\boldsymbol{\mu}^{J+1}_{\mathrm{SCM}} = \arg \max_{\boldsymbol{\mu} \in \Xi_{\mathrm{train,SCM}}} \eta_{M,J}(\boldsymbol{\mu})$
      $\mathcal{C}_{J+1} = \mathcal{C} \cup \boldsymbol{\mu}^{J+1}_{\mathrm{SCM}}$
      $J \leftarrow J + 1$
      $\eta_{M,J}(\boldsymbol{\mu}) = \dfrac{\alpha^{\mathcal{N}}_{\mathrm{UB}}(\boldsymbol{\mu}; \mathcal{C}_J, M) - \alpha^{\mathcal{N}}_{\mathrm{LB}}(\boldsymbol{\mu}; \mathcal{C}_J, M)}{\alpha^{\mathcal{N}}_{\mathrm{UB}}(\boldsymbol{\mu}; \mathcal{C}_J, M)}$
  **end while**
  set $J_{max} = J$.
```

Algorithm 3.2: SCM algorithm.

As already mentioned, the choice of the stopping criterion $\varepsilon_{\mathrm{SCM}}$ permits to bound the ratio between the coercivity constant and corresponding lower bound in the following way [77]:

$$\frac{\alpha^{\mathcal{N}}(\boldsymbol{\mu})}{\alpha^{\mathcal{N}}_{\mathrm{LB}}(\boldsymbol{\mu}; \mathcal{C}_J, M)} \leq \frac{1}{1 - \varepsilon_{\mathrm{SCM}}}, \qquad \forall \boldsymbol{\mu} \in \Xi_{\mathrm{train,SCM}}.$$

We briefly summarize the Offline and Online computational costs:

1. in the Offline stage we have to solve $2Q_a$ eigenproblems over $V^{\mathcal{N}}$ to form $\mathcal{B}$ and J_{max} eigenproblems over $V^{\mathcal{N}}$ to form $\{\alpha^{\mathcal{N}}(\boldsymbol{\mu}') \,|\, \boldsymbol{\mu}' \in \mathcal{C}_{J_{max}}\}$, to compute $J_{max} Q_a$ inner products over $V^{\mathcal{N}}$ to form $\{y^*(\boldsymbol{\mu}') \,|\, \boldsymbol{\mu}' \in \mathcal{C}_{J_{max}}\}$ and finally to solve $n_{\mathrm{train,SCM}} J_{max}$ linear programs of *size* $Q_a + M$;

2. in the Online stage, given a new value of $\boldsymbol{\mu}$, we have to solve a linear program of *size* $Q_a + M$ to evaluate $\alpha_{\mathrm{LB}}(\boldsymbol{\mu})$. The Online operation count is thus $\mathcal{N}$-independent.

3.4 Diffusion-advection-reaction equations in parametrized domains

The class of problems described by the variational problem (3.1.2) is clearly very general, therefore in the following we restrict ourself to consider the relatively simple (yet relevant to many applications) class of second order scalar PDEs in two spatial dimensions. In particular we consider diffusion-convection-reaction type equations. We want to show how, for this class of problems, starting from a parametrized PDE defined on a parametrized geometry, one can obtain a problem in the form (3.1.2) satisfying the affinity assumption.

We first define an *original* problem (subscript o), posed over a parameter-dependent domain; let $\mathcal{D} \subset \mathbb{R}^p$ be the parameters space, $\Omega_o = \Omega_o(\boldsymbol{\mu})$ be, for each $\boldsymbol{\mu} \in \mathcal{D}$, a regular bounded spatial domain in $\mathbb{R}^2$, $\Gamma_D^o = \partial\Omega_o$ its boundary. We introduce a domain decomposition (called RB *triangulation*) of the original domain

$$\Omega_o(\boldsymbol{\mu}) = \bigcup_{r=1}^{R} \Omega_o^r(\boldsymbol{\mu}),$$

where the $\Omega_o^r(\boldsymbol{\mu})$, $1 \le r \le R$, are mutually nonoverlapping subdomains, i.e. for all $\boldsymbol{\mu} \in \mathcal{D}$

$$\Omega_o^r(\boldsymbol{\mu}) \cap \Omega_o^{r'} = \emptyset, \quad 1 \le r, r' \le R,\ r \ne r'.$$

Then we consider the following equation, for the sake of simplicity with non-homogeneous Dirichlet condition on the whole boundary:

$$\begin{cases} -\dfrac{\partial}{\partial x_{oi}}\left(k_{o,r}^{ij}(\boldsymbol{\mu})\dfrac{\partial u_o(\boldsymbol{\mu})}{\partial x_{oj}}\right) + b_{o,r}^{j}(\boldsymbol{\mu})\dfrac{\partial u_o(\boldsymbol{\mu})}{\partial x_{oj}} & \\ \qquad\qquad + \dfrac{\partial}{\partial x_{oi}}\left(a_{o,r}^{i}(\boldsymbol{\mu})u_o(\boldsymbol{\mu})\right) + c_{o,r}(\boldsymbol{\mu})u_o(\boldsymbol{\mu}) = f_{o,r}, & \text{in } \Omega_o(\boldsymbol{\mu}), \\ u_o(\boldsymbol{\mu}) = g_D, & \text{on } \Gamma_D^o(\boldsymbol{\mu}), \end{cases}$$

where $\boldsymbol{x}_o = (x_{o1}, x_{o2})$ denotes a point in $\Omega_o(\boldsymbol{\mu})$ and summation over indices $1 \le i, j \le 2$ is understood. We now consider the weak formulation of the PDE. Let us introduce a lift function $R_D \in H^1(\Omega_o)$ for the non-homogeneous Dirichlet boundary condition and denote $\hat{u}_o = u_o - R_D$, so that $\hat{u}_o|_{\Gamma_D^o} = 0$; for the sake of simplicity we still denote $\hat{u}_o$ with u_o. Let $V_o = V(\Omega_o) = H_0^1(\Omega_o)$, $a_o(\cdot,\cdot;\boldsymbol{\mu}) : V \times V \to \mathbb{R}$ a bilinear form and $f_o(\cdot;\boldsymbol{\mu}) \in V'$ a linear functional. The weak formulation reads: given $\boldsymbol{\mu} \in \mathcal{D} \subset \mathbb{R}^p$, find $u_o(\boldsymbol{\mu}) \in V_o$ such that

$$a_o(u_o, v; \boldsymbol{\mu}) = f_o(v; \boldsymbol{\mu}), \qquad \forall v \in V_o, \tag{3.4.1}$$

where the bilinear form can be expressed as

$$a_o(w, v; \boldsymbol{\mu}) = \sum_{r=1}^{R} \int_{\Omega_0^r(\boldsymbol{\mu})} \begin{bmatrix} \frac{\partial w}{\partial x_{o1}} & \frac{\partial w}{\partial x_{o2}} & w \end{bmatrix} \boldsymbol{K}_{o,r}(\boldsymbol{\mu}) \begin{bmatrix} \frac{\partial v}{\partial x_{o1}} \\ \frac{\partial v}{\partial x_{o2}} \\ v \end{bmatrix} d\Omega_o, \tag{3.4.2}$$

where $\boldsymbol{K}_{o,r} : \mathcal{D} \to \mathbb{R}^{3x3}$ is, for each subdomain r, the positive definite matrix given by

$$\boldsymbol{K}_{o,r} = \begin{pmatrix} k_{o,r}^{11} & k_{o,r}^{12} & b_{o,r}^{1} \\ k_{o,r}^{21} & k_{o,r}^{22} & b_{o,r}^{2} \\ a_{o,r}^{1} & a_{o,r}^{2} & c_{o,r} \end{pmatrix}, \qquad 1 \leq r \leq R.$$

The upper 2×2 principal submatrix is the usual diffusivity/conductivity tensor, the $(3,3)$ element represents the identity operator (a reaction term resulting in a mass matrix), and the $(3,1),(3,2),(1,3),(2,3)$ elements represent first derivatives operators (convection terms). Hence, we have simply expressed in a compact and subdomain-dependent way a standard diffusion-advection-reaction equation. Similarly, the functional $f_o(\cdot;\boldsymbol{\mu})$ can be expressed as

$$f_o(v) = \sum_{r=1}^{R} \int_{\Omega_o^r(\boldsymbol{\mu})} f_{o,r}(\boldsymbol{\mu}) v \, d\Omega_o - a_o(R_D, v; \boldsymbol{\mu}). \tag{3.4.3}$$

It should be clear that the RB method requires a parameter independent domain Ω in order to compute and combine FE solutions that will be used as bases of the RB approximation space. For this reason, we need to map the original domain $\Omega_o(\boldsymbol{\mu})$ to a reference domain $\Omega = \Omega(\boldsymbol{\mu}_{\text{ref}})$ in order to recast the problem (3.4.1) in the form (3.1.2).

3.4.1 Affine geometrical parametrization

We denote with $\Omega = \Omega_o(\boldsymbol{\mu}_{\text{ref}})$, $\boldsymbol{\mu}_{\text{ref}} \in \mathcal{D}$, the reference domain and we identify $\Omega^r = \Omega_o^r(\boldsymbol{\mu}_{\text{ref}})$. We want to build a mapping $T(\cdot,\boldsymbol{\mu})\colon \Omega^r \to \Omega_o^r(\boldsymbol{\mu})$, $1 \leq r \leq R$, such that

$$\Omega_o^r(\boldsymbol{\mu}) = T^r(\Omega^r; \boldsymbol{\mu});$$

these maps must be individually bijective and collectively continuous, i.e. they have to fulfill the following interface condition:

$$T^r(\boldsymbol{x};\boldsymbol{\mu}) = T^{r'}(\boldsymbol{x};\boldsymbol{\mu}), \qquad \forall \boldsymbol{x} \in \Omega^r \cap \Omega^{r'}, \; 1 \leq r < r' \leq R.$$

In the affine case considered here these mappings have the general form

$$T_i^r(\boldsymbol{x},\boldsymbol{\mu}) = C_i^r(\boldsymbol{\mu}) + \sum_{j=1}^{2} G_{ij}^r(\boldsymbol{\mu}) x_j, \qquad 1 \leq j \leq 2, \tag{3.4.4}$$

for given translation vectors $\boldsymbol{C}^r\colon \mathcal{D} \to \mathbb{R}^2$ and linear transformations matrices $\boldsymbol{G}^r\colon \mathcal{D} \to \mathbb{R}^{2\times 2}$. The linear transformation matrices allow rotations, scaling and shear and have to be invertible, that means that the associated Jacobians $J^r(\boldsymbol{\mu}) = |\det(\boldsymbol{G}^r(\boldsymbol{\mu}))|$ should be strictly positive. For the details about the construction of such mappings and the implementation in the `rbMIT` software package we use in this work see [77, 44, 61].

In the (more realistic) case of non-affinely parametrized transformations of the domain, the mappings need to be approximated by affinely parametrized tensors through the empirical interpolation method [4, 31], in order to ensure the feasibility of the Offline/Online computational strategy. Since shape representation is highly problem-dependent, various methods have been proposed; common strategies for shape deformation involve the use of *(i)* the coordinates of the boundary points as design variables (*local boundary variation*) or *(ii)* some families of basis shapes combined by means of a set of control point (*polynomial* *boundary parametrizations*). These techniques are not well suited within the RB framework, since a global mapping $T(\cdot;\boldsymbol{\mu})$ is needed, rather than a boundary representation. A more versatile parametrization can be introduced by exploiting the *free-form deformation* (FFD) techniques, in which the deformations of an initial design, rather than the geometry itself, are parametrized [51]. Other different techniques based on interpolation properties may be introduced, in particular we mention the *radial basis functions* (RBF) techniques [55].

3.4.2 Parametrized formulation on a reference domain

By identifying $u(\boldsymbol{\mu}) = u_o(\boldsymbol{\mu}) \circ T(\cdot;\boldsymbol{\mu})$ and tracing (3.4.1) back on the reference domain Ω, the problem can be written as: find $u(\boldsymbol{\mu}) \in V$ such that

$$a(u(\boldsymbol{\mu}), v; \boldsymbol{\mu}) = f(v;\boldsymbol{\mu}) \qquad \forall v \in V,$$

where

$$a(w,v;\boldsymbol{\mu}) = \sum_{r=1}^{R} \int_{\Omega^r(\boldsymbol{\mu})} \begin{bmatrix} \frac{\partial w}{\partial x_1} & \frac{\partial w}{\partial x_2} & w \end{bmatrix} \boldsymbol{K}_r(\boldsymbol{\mu}) \begin{bmatrix} \frac{\partial v}{\partial x_1} \\ \frac{\partial v}{\partial x_2} \\ v \end{bmatrix} d\Omega, \tag{3.4.5}$$

and

$$f(v) = \sum_{r=1}^{R} \int_{\Omega^r(\boldsymbol{\mu})} F_r(\boldsymbol{\mu}) v \, d\Omega - a(R_D, w; \boldsymbol{\mu}). \tag{3.4.6}$$

The transformation coefficients for the linear and bilinear forms $\boldsymbol{K}_r(\boldsymbol{\mu}) : \mathcal{D} \to \mathbb{R}^{3x3}$ and $F_r(\boldsymbol{\mu}) : \mathcal{D} \to \mathbb{R}$ are defined as follows:

$$\boldsymbol{K}_r(\boldsymbol{\mu}) = J^r(\boldsymbol{\mu}) \tilde{\boldsymbol{G}}^r(\boldsymbol{\mu}) \boldsymbol{K}_{o,l}(\boldsymbol{\mu}) (\tilde{\boldsymbol{G}}^r(\boldsymbol{\mu}))^T, \qquad F_r(\boldsymbol{\mu}) = J^r(\boldsymbol{\mu}) \boldsymbol{F}_{o,r}(\boldsymbol{\mu}), \tag{3.4.7}$$

with

$$\tilde{\boldsymbol{G}}^r(\boldsymbol{\mu}) = \begin{pmatrix} (\boldsymbol{G}^r(\boldsymbol{\mu}))^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & 1 \end{pmatrix}, \quad 1 \le r \le R.$$

The affine decomposition (3.1.5) for the bilinear form $a(\cdot,\cdot;\boldsymbol{\mu})$ can be derived by expanding the expression (3.4.5) in terms of the subdomains Ω^r and the entries K_r^{ij}, i.e.

$$a(w,v,;\boldsymbol{\mu}) = K_1^{11}(\boldsymbol{\mu}) \int_{\Omega^1} \frac{\partial w}{\partial x_1} \frac{\partial v}{\partial x_1} + K_1^{12}(\boldsymbol{\mu}) \int_{\Omega^1} \frac{\partial w}{\partial x_1} \frac{\partial v}{\partial x_2} + K_1^{13}(\boldsymbol{\mu}) \int_{\Omega^1} \frac{\partial w}{\partial x_1} v + \dots$$

Then for each term in the expression above, the pre-factor K_r^{ij} represents $\Theta^q(\boldsymbol{\mu})$, while the $\boldsymbol{\mu}$-independent integral represents the bilinear forms $a^q(w,v)$. Needless to say, $f(\cdot;\boldsymbol{\mu})$ admits a similar treatment.

Note that the procedure described can be easily extended in order to consider different boundary conditions, for instance Robin conditions or Neumann conditions. Not only, we can also permit affine polynomial dependence on $\boldsymbol{x}_o$ in the coefficients $k_{o,r}^{ij}(\boldsymbol{x}_o;\boldsymbol{\mu})$, $b_{o,r}^{j}(\boldsymbol{x}_o;\boldsymbol{\mu})$, $a_{o,r}^{i}(\boldsymbol{x}_o;\boldsymbol{\mu})$, $c_{o,r}(\boldsymbol{x}_o;\boldsymbol{\mu})$ and $f_{o,r}(\boldsymbol{x}_o;\boldsymbol{\mu})$ and still ensure the affine development. In the following numerical example we show some of these extensions.

3.4.3 A Graetz conduction-convection problem

This example deals with steady forced heat convection combined with heat conduction in a straight duct, whose walls can be kept at fixed temperature or insulated or characterized by

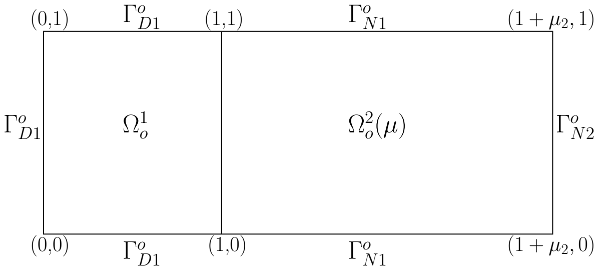


Figure 3.1: Original domain $\Omega_o(\boldsymbol{\mu})$.

heat exchange [2]. The flow has an imposed temperature at the inlet and a known convection field (a Poiseuille flow, i.e. a given parabolic velocity profile). Variants of the problem have been already treated with the RB method, see e.g. [66]. Here, we consider the original domain shown in Figure 3.1, where μ_2 is a geometrical parameter, i.e. the length of the second portion of the channel; moreover let us denote $\tilde{D}$ the thermal diffusion coefficient of the air flowing in the duct, $\tilde{h}$ the channel width and $\tilde{U}$ the reference velocity for the convection field. The Péclet number is given by the ratio $\mathrm{Pe} = \tilde{U}\tilde{h}/\tilde{D}$. The temperature field $u_o(\boldsymbol{\mu})$ satisfies the following steady convection-diffusion equation:

$$\begin{cases} -\dfrac{1}{\mu_1}\Delta u_o(\boldsymbol{\mu}) + x_{o2}(1-x_{o2})\dfrac{\partial u_o(\boldsymbol{\mu})}{\partial x_{o1}} = 0 & \text{in } \Omega_o(\boldsymbol{\mu}) \\ u_o(\mu) = 0 & \text{on } \Gamma^o_{D1} \\ \dfrac{1}{\mu_1}\nabla u_o(\boldsymbol{\mu})\cdot\mathbf{n} = \mu_3 & \text{on } \Gamma^o_{N1}(\boldsymbol{\mu}) \\ \dfrac{1}{\mu_1}\nabla u_o(\boldsymbol{\mu})\cdot\mathbf{n} = 0 & \text{on } \Gamma^o_{N2}(\boldsymbol{\mu}), \end{cases} \tag{3.4.8}$$

where $\mu_1 = \mathrm{Pe}$. Hence we are imposing the temperature on Γ^o_{D1}, while we consider an insulated wall at the outflow (zero heat flux on Γ^o_{N2}) and heat exchange at the rate μ_3 on the boundaries Γ^o_{N1}. The parameter domain is given by $\mathcal{D} = [0.1, 50] \times [1, 5] \times [0.2, 1]$. The weak formulation of (3.4.8) can be easily stated as in (3.4.2) by defining

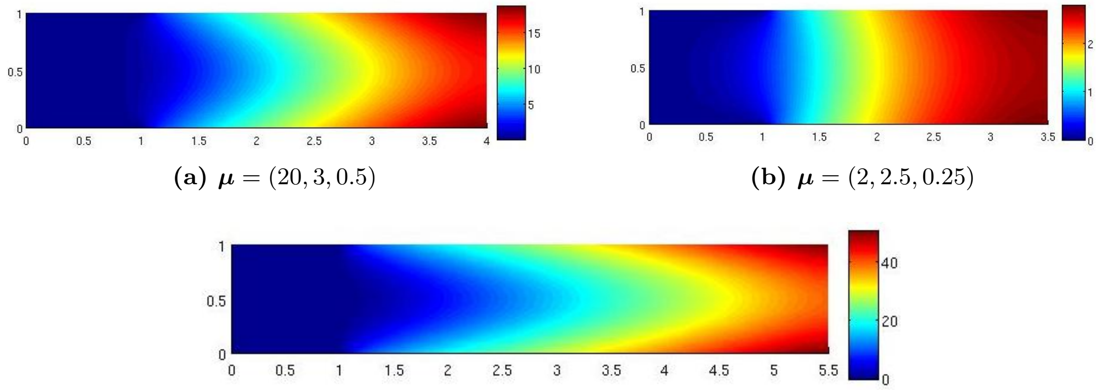
$$\boldsymbol{K}_{o,r} = \begin{pmatrix} -\frac{1}{\mu_1} & 0 & x_{o2}(1-x_{o2}) \\ 0 & -\frac{1}{\mu_1} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad 1 \le r \le 2.$$

The problem is then mapped to a fixed reference domain $\Omega = \Omega(\boldsymbol{\mu}_{\mathrm{ref}})$ with $\boldsymbol{\mu}_{\mathrm{ref}} = [1, 1, 0.5]$ and discretized by piecewise linear finite elements. Being the first subdomain parameter independent the corresponding affine geometrical mapping is trivial

$$\boldsymbol{C}^1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \boldsymbol{G}^1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad J^r = 1,$$

while for the second subdomain we obtain

$$\boldsymbol{C}^2 = \begin{bmatrix} 1-\mu_2 \\ 0 \end{bmatrix}, \qquad \boldsymbol{G}^2 = \begin{bmatrix} \mu_2 & 0 \\ 0 & 1 \end{bmatrix}, \qquad J^r = \mu_2.$$



(a) $\boldsymbol{\mu} = (20, 3, 0.5)$

(b) $\boldsymbol{\mu} = (2, 2.5, 0.25)$

(c) $\boldsymbol{\mu} = (50, 4.5, 0.8)$

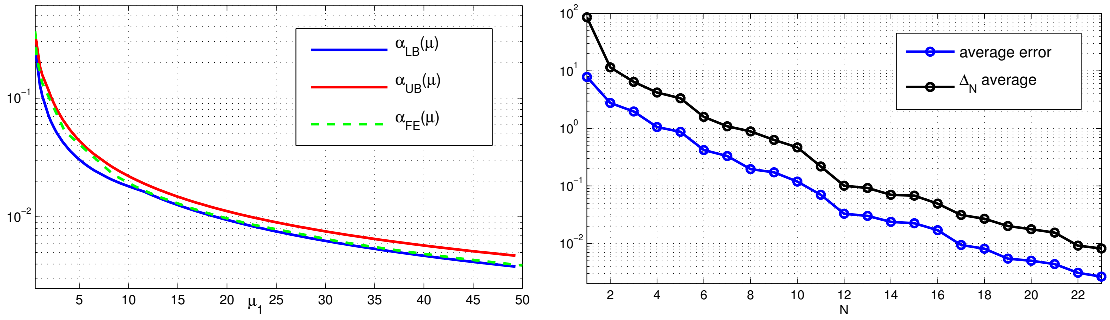
Figure 3.2: Temperature field $u(\boldsymbol{\mu})$ for different values of the parameters.

As regards the affine decomposition we have $Q_a = 4$, $Q_f = 5$,

$$\begin{aligned}
\Theta_a^1(\boldsymbol{\mu}) &= \frac{1}{\mu_1}, & a^1(w,v) &= \int_{\Omega_1} \frac{\partial w}{\partial x_1}\frac{\partial v}{\partial x_1}\, d\Omega + \int_{\Omega_1} \frac{\partial w}{\partial x_2}\frac{\partial v}{\partial x_2}\, d\Omega \\
\Theta_a^2(\boldsymbol{\mu}) &= \frac{1}{\mu_2\mu_1}, & a^2(w,v) &= \int_{\Omega_2} \frac{\partial w}{\partial x_1}\frac{\partial v}{\partial x_1}\, d\Omega \\
\Theta_a^3(\boldsymbol{\mu}) &= \frac{\mu_2}{\mu_1}, & a^3(w,v) &= \int_{\Omega_2} \frac{\partial w}{\partial x_2}\frac{\partial v}{\partial x_2}\, d\Omega \\
\Theta_a^4(\boldsymbol{\mu}) &= 1, & a^4(w,v) &= \int_{\Omega_1} x_2(1-x_2)\frac{\partial w}{\partial x_1} v\, d\Omega + \int_{\Omega_2} x_2(1-x_2)\frac{\partial w}{\partial x_1} v\, d\Omega \\
\Theta_1^f(\mu) &= \mu_2\mu_3, & f^5(v) &= \int_{\Gamma_{N1}} v\, d\Gamma.
\end{aligned}$$

The dimension of the FE space used is $\mathcal{N} = 4153$. In Figure 3.2 we report some plots of the solution for different values of the parameters. Depending on the value of the Péclet number the solution shows thermal boundary layers near the walls with heat exchange (at the rate μ_3). With a fixed tolerance $\varepsilon_{\text{tol}} = 10^{-3}$ $N = 23$ basis have been selected by the Greedy algorithm. In Figure 3.3a we show the upper and lower bound for the coercivity constant $\alpha^{\mathcal{N}}(\boldsymbol{\mu})$ obtained using the SCM algorithm; in Figure 3.3b we compare the a posteriori error bound $\Delta_N(\boldsymbol{\mu})$ with the true error $\|u^{\mathcal{N}}(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_V$, in particular we show the average of these two quantities over a sample of 250 parameters.

The Offline computational time is equal to $t_{RB}^{offline} = 510s$, the (average) Online solution time is $t_{RB}^{online} = 0.035s$ comprehensive of the evaluation of the a posteriori error estimation; the computational time needed to build and solve the RB linear system alone is very small, less than $10^{-3}s$ (in fact we are solving a system of size 23×23). The evaluation time for the FE approximation is equal to $t_{FE} = 6.3s$ taking into account the time needed for assembling the FE matrices and vectors; exploiting the affine decompositions also for the FE approximation, the computational time decrease to about $t_{FE} = 1.1$s, i.e. the time needed to sum the matrices and vectors in the affine decomposition and solve the linear system of size $\mathcal{N} \times \mathcal{N}$. The computational speedup defined as $\mathcal{S} = t_{FE}/t_{RB}^{online}$ is about 182 while the break-even point defined as $\mathcal{Q}_{BE} = t_{RB}^{offline}/t_{FE}$ is about 75 (in a many-query context, it



(a) Upper (red) and lower (blue) bound for the coercivity constant $\alpha^{\mathcal{N}}(\boldsymbol{\mu})$ (green) as a function of μ_1 with μ_2 and μ_3 fixed. Computed by the SCM algorithm using a tolerance $\varepsilon_{SCM} = 0.75$.

(b) A posteriori error bound. Comparison of the average computed error (between the *truth* FE solution and the RB approximation $\|u^{\mathcal{N}}(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_V$) and the estimator $\Delta(\boldsymbol{\mu})$, for $N = 1, \cdots, N_{max} = 23$.

Figure 3.3

represents the number of online queries to justify the use of reduced basis instead of FE approximation for the online calculations).

3.5 Non-coercive problems: the Stokes equations

The reduced basis framework can be effectively applied also to non-coercive problems (also called weakly coercive problems, see A.1), i.e. problems which do not satisfy the coercivity assumption. In these cases the stability is in fact fulfilled in the more general sense of the *inf-sup* constant. An example of such problems of particular interest is provided by the Stokes equations [69]. In this section we introduce the reduced basis method for Stokes equations in domains with affine parametric dependence, referring principally to [73, 79, 76, 28]. For extensions to non-affine problems see [75, 56].

3.5.1 Problem formulation

We consider the following steady Stokes problem [69, 65] in the *original* domain $\Omega_o = \Omega_o(\boldsymbol{\mu}) \subset \mathbb{R}^2$

$$\begin{cases} -\nu\Delta \boldsymbol{v}_o + \nabla p_o = \boldsymbol{f}_o & \text{in } \Omega_o, \\ \operatorname{div} \boldsymbol{v}_o = 0 & \text{in } \Omega_o, \\ -p_o\boldsymbol{n} + \nu\dfrac{\partial \boldsymbol{v}_o}{\partial \boldsymbol{n}} = \boldsymbol{g}_N & \text{on } \Gamma_N^o, \\ \boldsymbol{v}_o = \boldsymbol{0} & \text{on } \Gamma_{D_0}^o, \\ \boldsymbol{v}_o = \boldsymbol{g}_D & \text{on } \Gamma_{D_g}^o, \end{cases} \tag{3.5.1}$$

where $\boldsymbol{v}$ is the velocity, p the pressure, $\boldsymbol{f}$ a force field, $\boldsymbol{n}$ the normal unit vector, and the three components of the boundary $\Gamma_N^o, \Gamma_{D_0}^o, \Gamma_{D_g}^o$ are such that $\Gamma_N^o \cap \Gamma_{D_0}^o \cap \Gamma_{D_g}^o = \emptyset$ and $\Gamma_N^o \cup \Gamma_{D_0}^o \cup \Gamma_{D_g}^o = \partial\Omega_o$. We introduce a lift function $R_{\boldsymbol{g}} \in [H^1(\Omega_o)]^2$ such that $R_{\boldsymbol{g}}|_{\Gamma_D^o} = \boldsymbol{g}_D$ and $\boldsymbol{v}_o = \tilde{\boldsymbol{v}}_o + R_{\boldsymbol{g}}$; we still denote $\tilde{\boldsymbol{v}}_o$ with $\boldsymbol{v}_o$ in the sequel. We define the appropriate functional spaces for the velocity and pressure, respectively $V_o = [H^1_{\Gamma_D^o}(\Omega_o)]^2$ and $M_o = L^2(\Omega_o)$. The weak formulation of the state equation reads [69]: find $(\boldsymbol{v}, p) \in V_o \times M_o$ such that

$$\begin{cases} \nu \displaystyle\int_{\Omega_o} \nabla \boldsymbol{v}_o \cdot \nabla \boldsymbol{\xi} \, d\Omega - \int_{\Omega_o} p_o \nabla \cdot \boldsymbol{\xi} \, d\Omega = & \\ \qquad\qquad = \displaystyle\int_{\Omega_o} \boldsymbol{f}_o \cdot \boldsymbol{\xi} \, d\Omega + \int_{\Gamma_N^o} \boldsymbol{g}_N \cdot \boldsymbol{\xi} \, d\Gamma + \langle F_0^o, \boldsymbol{\xi} \rangle, & \forall \boldsymbol{\xi} \in V_o, \\ \displaystyle\int_{\Omega_o} \tau \nabla \cdot \boldsymbol{v}_o \, d\Omega = \langle G_0^o, \tau \rangle, & \forall \tau \in M_o, \end{cases} \tag{3.5.2}$$

where F_0^o, G_0^o take in account for the non homogeneous Dirichlet boundary condition. Similarly to what we have already done in Section 3.4, we assume that the original domain is made up of R mutually nonoverlapping subdomains

$$\Omega_o(\boldsymbol{\mu}) = \bigcup_{r=1}^{R} \Omega_o^r(\boldsymbol{\mu}),$$

so that the bilinear and linear forms in (3.5.2) can be expressed as

$$a_o(\boldsymbol{v}, \boldsymbol{\xi}) = \sum_{r=1}^{R} \int_{\Omega_o^r} \nu_{ij}^o \frac{\partial \boldsymbol{v}}{\partial x_{oi}} \cdot \frac{\partial \boldsymbol{\xi}}{\partial x_{oj}} \, d\Omega, \qquad b_o(\boldsymbol{\xi}, \tau) = -\sum_{r=1}^{R} \int_{\Omega_o^r} \tau \nabla \cdot \boldsymbol{\xi} \, d\Omega_o,$$

where $1 \le i, j \le 2$, $\nu_{ij}^o = \nu \delta_{ij}$ and summation over i and j is understood, and $\langle F^o, \boldsymbol{\xi} \rangle = \langle F_s^o, \boldsymbol{\xi} \rangle + \langle F_0^o, \boldsymbol{\xi} \rangle$, with

$$\langle F_s^o, \boldsymbol{\xi} \rangle = \sum_{r=1}^{R} \int_{\Omega_o^r} \boldsymbol{f}_o \cdot \boldsymbol{\xi} \, d\Omega_o + \sum_{r=1}^{R} \int_{\Gamma_N^{o,r}} \boldsymbol{g}_N \cdot \boldsymbol{\xi} \, d\Gamma_o, \qquad \langle F_0^o, \boldsymbol{\xi} \rangle = -a_o(R_{\boldsymbol{g}}, \boldsymbol{\xi}),$$

and $\langle G^o, \tau \rangle = -b_o(R_{\boldsymbol{g}}, \tau)$. We can write equivalently the weak formulation (3.5.2) as: find $(\boldsymbol{v}, p) \in V_o \times M_o$ such that

$$\begin{cases} a_o(\boldsymbol{v}_o, \boldsymbol{\xi}) + b_o(\boldsymbol{\xi}, p_o) = \langle F^o, \boldsymbol{\xi} \rangle & \forall \boldsymbol{\xi} \in V_o, \\ b_o(\boldsymbol{v}_o, \tau) = \langle G^o, \tau \rangle & \forall \tau \in M_o. \end{cases} \tag{3.5.3}$$

Parametrized formulation

Denoting with $\Omega = \Omega_o(\boldsymbol{\mu}_{\text{ref}})$ the reference domain, we can trace (3.5.3) back to this reference domain by the affine mapping $T(\cdot; \boldsymbol{\mu})$ already described in Section 3.4. Denoting $V = [H^1_{\Gamma_D^o}(\Omega)]^2$ and $M = L^2(\Omega)$, we have the following parametrized formulation: given $\boldsymbol{\mu} \in \mathcal{D}$, find $(\boldsymbol{v}(\boldsymbol{\mu}), p(\boldsymbol{\mu})) \in V \times M$ such that

$$\begin{cases} a(\boldsymbol{v}(\boldsymbol{\mu}), \boldsymbol{\xi}; \boldsymbol{\mu}) + b(\boldsymbol{\xi}, p(\boldsymbol{\mu}); \boldsymbol{\mu}) = \langle F(\boldsymbol{\mu}), \boldsymbol{\xi} \rangle & \forall \boldsymbol{\xi} \in V, \\ b(\boldsymbol{v}(\boldsymbol{\mu}), \tau; \boldsymbol{\mu}) = \langle G(\boldsymbol{\mu}), \tau \rangle & \forall \tau \in M, \end{cases} \tag{3.5.4}$$

where the following affine decomposition for the linear and bilinear forms hold (for the details we refer to [79, 76]):

$$a(\boldsymbol{v}, \boldsymbol{\xi}; \boldsymbol{\mu}) = \sum_{q=1}^{Q_a} \Theta_a^q(\boldsymbol{\mu}) \, a^q(\boldsymbol{v}, \boldsymbol{\xi}), \qquad b(\boldsymbol{\xi}, \tau; \boldsymbol{\mu}) = \sum_{q=1}^{Q_b} \Theta_b^q(\boldsymbol{\mu}) \, b^q(\boldsymbol{\xi}, \tau) \tag{3.5.5}$$

$$\langle G(\boldsymbol{\mu}), \tau \rangle = -b(R_{\boldsymbol{g}}, \tau; \boldsymbol{\mu}), \qquad \langle F(\boldsymbol{\mu}), \boldsymbol{\xi} \rangle = \sum_{q=1}^{Q_f} \Theta_f^q(\boldsymbol{\mu}) \, \langle F^q, \boldsymbol{\xi} \rangle, \tag{3.5.6}$$

where, for instance,

$$a^{q(i,j,r)}(\boldsymbol{v},\boldsymbol{\xi}) = \int_{\Omega^r} \frac{\partial \boldsymbol{v}}{\partial x_i} \cdot \frac{\partial \boldsymbol{\xi}}{\partial x_j}\, d\Omega, \qquad b^{q(i,j,r)}(\boldsymbol{\xi},\tau) = -\int_{\Omega^r} \tau \frac{\partial w_i}{\partial x_j}\, d\Omega,$$

and

$$\Theta_a^{q(i,j,r)}(\boldsymbol{\mu}) = \nu_{ij}^r = (G^r(\boldsymbol{\mu}))^{-1}_{ii'} \nu^o_{i'j'} (G^r(\boldsymbol{\mu}))^{-1}_{jj'} J^r(\boldsymbol{\mu}), \quad 1 \le i,i',j,j' \le 2,$$

$$\Theta_b^{q(i,j,r)}(\boldsymbol{\mu}) = \xi_{ij}^r = (G^r(\boldsymbol{\mu}))^{-1}_{ij} J^r(\boldsymbol{\mu}), \quad 1 \le i,j \le 2.$$

The well-posedness of (3.5.4) is ensured by (Brezzi) Theorem A.2 (see, e.g., [14, 69]), we recall here the definition of the required continuity, coercivity and inf-sup constants in the parametrized context. The bilinear form $a(\cdot,\cdot;\boldsymbol{\mu})$ is continuous over $V \times V$ and coercive over V, i.e.

$$\gamma_a(\boldsymbol{\mu}) = \sup_{\boldsymbol{\xi}\in V} \sup_{\boldsymbol{v}\in V} \frac{a(\boldsymbol{v},\boldsymbol{\xi};\boldsymbol{\mu})}{\|\boldsymbol{v}\|_V \|\boldsymbol{\xi}\|_V} < +\infty, \qquad \forall \boldsymbol{\mu} \in \mathcal{D},$$

and there exists $\alpha_0 > 0$ such that

$$\alpha(\boldsymbol{\mu}) = \inf_{\boldsymbol{\xi}\in V} \sup_{\boldsymbol{v}\in V} \frac{a(\boldsymbol{v},\boldsymbol{\xi};\boldsymbol{\mu})}{\|\boldsymbol{v}\|_V \|\boldsymbol{\xi}\|_V} > \alpha_0, \qquad \forall \boldsymbol{\mu} \in \mathcal{D};$$

the bilinear form $b(\cdot,\cdot;\boldsymbol{\mu})$ is continuous over $V \times M$

$$\gamma_b(\boldsymbol{\mu}) = \sup_{\boldsymbol{\xi}\in V} \sup_{\tau\in M} \frac{b(\boldsymbol{\xi},\tau;\boldsymbol{\mu})}{\|\tau\|_M \|\boldsymbol{\xi}\|_V} < +\infty, \qquad \forall \boldsymbol{\mu} \in \mathcal{D},$$

and satisfies the following inf-sup condition

$$\beta(\boldsymbol{\mu}) = \inf_{\tau\in M} \sup_{\boldsymbol{\xi}\in V} \frac{b(\boldsymbol{\xi},\tau;\boldsymbol{\mu})}{\|\tau\|_M \|\boldsymbol{\xi}\|_V} \ge \beta_0, \qquad \forall \boldsymbol{\mu} \in \mathcal{D},$$

where $\beta_0 > 0$.

***Truth* approximation**

We consider the Galerkin-FE approximation of (3.5.4). We denote with $V^{\mathcal{N}}$-$M^{\mathcal{N}}$ a stable pair of finite element spaces, i.e. ensuring the fulfilment of the discrete inf-sup condition (see Section A.1 and [69] for examples of such spaces). In particular $V^{\mathcal{N}} \subset V$ and $M^{\mathcal{N}} \subset M$ are two sequences of FE approximating spaces of global dimension $\mathcal{N} = \mathcal{N}_V + \mathcal{N}_M$. The truth FE approximation reads: find $(\boldsymbol{v}^{\mathcal{N}}(\boldsymbol{\mu}), p^{\mathcal{N}}(\boldsymbol{\mu})) \in V^{\mathcal{N}} \times M^{\mathcal{N}}$ such that

$$\begin{cases} a(\boldsymbol{v}^{\mathcal{N}}(\boldsymbol{\mu}),\boldsymbol{\xi};\boldsymbol{\mu}) + b(\boldsymbol{\xi},p^{\mathcal{N}}(\boldsymbol{\mu});\boldsymbol{\mu}) = \langle F(\boldsymbol{\mu}),\boldsymbol{\xi}\rangle & \forall \boldsymbol{\xi} \in V^{\mathcal{N}}, \\ b(\boldsymbol{v}^{\mathcal{N}}(\boldsymbol{\mu}),\tau;\boldsymbol{\mu}) = \langle G(\boldsymbol{\mu}),\tau\rangle, & \forall \tau \in M^{\mathcal{N}}, \end{cases} \tag{3.5.7}$$

The bilinear form $a(\cdot,\cdot;\boldsymbol{\mu})$ remains continuous over $V^{\mathcal{N}} \times V^{\mathcal{N}}$

$$\gamma_a^{\mathcal{N}}(\boldsymbol{\mu}) = \sup_{\boldsymbol{\xi}\in V^{\mathcal{N}}} \sup_{\boldsymbol{v}\in V^{\mathcal{N}}} \frac{a(\boldsymbol{v},\boldsymbol{\xi};\boldsymbol{\mu})}{\|\boldsymbol{v}\|_V \|\boldsymbol{\xi}\|_V} \le \gamma_a(\boldsymbol{\mu}), \qquad \forall \boldsymbol{\mu} \in \mathcal{D},$$

and coercive over $V^{\mathcal{N}}$

$$\alpha^{\mathcal{N}}(\boldsymbol{\mu}) = \inf_{\boldsymbol{\xi}\in V^{\mathcal{N}}} \sup_{\boldsymbol{v}\in V^{\mathcal{N}}} \frac{a(\boldsymbol{v},\boldsymbol{\xi};\boldsymbol{\mu})}{\|\boldsymbol{v}\|_V \|\boldsymbol{\xi}\|_V} \ge \alpha(\boldsymbol{\mu}) > \alpha_0, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}.$$

The bilinear form $b(\cdot,\cdot;\boldsymbol{\mu})$ remains continuous over $V^{\mathcal{N}} \times M^{\mathcal{N}}$

$$\gamma_b^{\mathcal{N}}(\boldsymbol{\mu}) = \sup_{\boldsymbol{\xi}\in V^{\mathcal{N}}} \sup_{\tau\in M^{\mathcal{N}}} \frac{b(\boldsymbol{\xi},\tau;\boldsymbol{\mu})}{\|\tau\|_M\|\boldsymbol{\xi}\|_V} \le \gamma_b(\boldsymbol{\mu}), \qquad \forall \boldsymbol{\mu} \in \mathcal{D},$$

moreover, thanks to the choice of the approximation spaces, there exists a constant $\beta_0^{\mathcal{N}} > 0$ [69] such that

$$\beta^{\mathcal{N}}(\boldsymbol{\mu}) = \inf_{\tau\in M^{\mathcal{N}}} \sup_{\boldsymbol{\xi}\in V^{\mathcal{N}}} \frac{b(\boldsymbol{\xi},\tau;\boldsymbol{\mu})}{\|\tau\|_M\|\boldsymbol{\xi}\|_V} \ge \beta_0 > 0, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}.$$

3.5.2 Reduced basis approximation

Let us take, for given $N \in \{1,\dots,N_{\max}\}$, a set of parameter values $S_N = \{\boldsymbol{\mu}^1,\dots,\boldsymbol{\mu}^N\}$ and consider the corresponding FE solutions $\{(\boldsymbol{v}^{\mathcal{N}}(\boldsymbol{\mu}^n), p^{\mathcal{N}}(\boldsymbol{\mu}^n)),\ n = 1,\dots,N\}$. We define the reduced basis pressure spaces as

$$M_N = \text{span}\{\varphi_n := p^{\mathcal{N}}(\boldsymbol{\mu}^n), \quad n = 1,\dots,N\}. \tag{3.5.8}$$

As regards the reduced basis velocity space, in order to verify an equivalent Brezzi reduced basis inf-sup condition, it is necessary to introduce a particular recipe [79, 76]. Let us firstly introduce the following (pressure) supremizer operator $T_p^{\boldsymbol{\mu}} : M^{\mathcal{N}} \to V^{\mathcal{N}}$ defined as follows:

$$(T_p^{\boldsymbol{\mu}}\tau, \boldsymbol{\xi})_V = b(\boldsymbol{\xi},\tau;\boldsymbol{\mu}), \qquad \forall \boldsymbol{\xi} \in V^{\mathcal{N}}. \tag{3.5.9}$$

Then we build the reduced basis velocity space enriching the velocity space with the supremizer solutions, i.e. we define

$$V_N^{\boldsymbol{\mu}} = \text{span}\{\boldsymbol{\zeta}_n := \boldsymbol{v}^{\mathcal{N}}(\boldsymbol{\mu}^n),\ T_p^{\boldsymbol{\mu}}\varphi_n, \quad n = 1,\dots,N\}. \tag{3.5.10}$$

By using Galerkin projection onto $V_N^{\boldsymbol{\mu}} \times M_N$, we obtain the following reduced basis approximation: find $(\boldsymbol{v}_N(\mu), p_N(\mu)) \in V_N^{\boldsymbol{\mu}} \times M_N$ such that

$$\begin{cases} a(\boldsymbol{v}_N(\boldsymbol{\mu}),\boldsymbol{\xi};\boldsymbol{\mu}) + b(\boldsymbol{\xi}, p_N(\boldsymbol{\mu});\boldsymbol{\mu}) = \langle F(\boldsymbol{\mu}),\boldsymbol{\xi}\rangle & \forall \boldsymbol{\xi} \in V_N^{\boldsymbol{\mu}}, \\ b(\boldsymbol{v}_N(\boldsymbol{\mu}),\tau;\boldsymbol{\mu}) = \langle G(\boldsymbol{\mu}),\tau\rangle, & \forall \tau \in M_N, \end{cases} \tag{3.5.11}$$

Since $V_N^{\boldsymbol{\mu}} \subset V^{\mathcal{N}}$ and $M_N \subset M^{\mathcal{N}}$, the bilinear form $a(\cdot,\cdot;\boldsymbol{\mu})$ remains continuous over $V_N^{\boldsymbol{\mu}} \times V_N^{\boldsymbol{\mu}}$ and coercive over $V_N^{\boldsymbol{\mu}}$, and the bilinear form $b(\cdot,\cdot;\boldsymbol{\mu})$ remains continuous over $V_N^{\boldsymbol{\mu}} \times M_N$. Moreover, thanks to the definition (3.5.10), the bilinear form $b(\cdot,\cdot;\boldsymbol{\mu})$ fulfils an equivalent RB inf-sup condition; in fact, by defining

$$\beta_N(\boldsymbol{\mu}) = \inf_{\tau\in M_N} \sup_{\boldsymbol{\xi}\in V_N^{\boldsymbol{\mu}}} \frac{b(\boldsymbol{\xi},\tau;\boldsymbol{\mu})}{\|\tau\|_M\|\boldsymbol{\xi}\|_V}, \qquad \forall \boldsymbol{\mu} \in \mathcal{D},$$

it can be proved [79] that the following inequality holds:

$$\beta_N(\boldsymbol{\mu}) \ge \beta^{\mathcal{N}}(\boldsymbol{\mu}) \ge \beta_0 > 0, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}. \tag{3.5.12}$$

Hence the RB approximation (3.5.11) is well-posed thanks to (Brezzi) Theorem A.4.

It is important to note that, because of the definition of the supremizer operator $T_p^{\boldsymbol{\mu}}$, the RB velocity space $V_N^{\boldsymbol{\mu}}$ still depends on the parameters $\boldsymbol{\mu}$, thus affecting a bit the efficient decoupling of the Offline and Online stages. In order to express in a $\boldsymbol{\mu}$-independent way the RB velocity space, several alternative constructions have been proposed [76], related also to different orthonormalization procedures. In the next section we limit ourselves to introduce one of the possible alternatives, in particular the one that seems to represent the better compromise between rigour and computational efficiency. For a detailed discussion we refer to [76].

3.5.3 Algebraic formulation

Let us firstly note that from the affine assumption (3.5.5) it follows that the suprimizer operator can be expressed as

$$T_p^{\boldsymbol{\mu}}\tau = \sum_{q=1}^{Q_b} \Theta_b^q(\boldsymbol{\mu}) T_p^q \tau, \qquad \forall \tau \in M^{\mathcal{N}}, \tag{3.5.13}$$

where $(T_p^q\tau, \boldsymbol{\xi})_V = b^q(\boldsymbol{\xi}, \tau)$, $\forall \boldsymbol{\xi} \in V^{\mathcal{N}}$, $1 \leq q \leq Q_b$. In order to define a $\boldsymbol{\mu}$-independent RB velocity space, the crucial idea is to enrich the velocity space with supremizers built upon summation using the same $\boldsymbol{\mu}^n$ values used to store velocity $\boldsymbol{\zeta}_n(\boldsymbol{\mu}^n)$ and pressure solutions $\varphi_n(\boldsymbol{\mu}^n)$. This leads to define the reduced basis space for the velocity as [76]

$$V_N = \text{span}\left\{ \boldsymbol{\sigma}_n = \sum_{k=1}^{\bar{Q}_b} \Theta_b^k(\boldsymbol{\mu}^j)\boldsymbol{\sigma}_{kn}, \ n = 1, \dots, 2N \right\}, \tag{3.5.14}$$

where $\bar{Q}_b = Q_b + 1$, $\Theta_b^{\bar{Q}_b} = 1$ and, for $n = 1, \dots, N$

$$\boldsymbol{\sigma}_{kn} = \begin{cases} 0, & \text{for } k = 1, \dots, Q_b \\ \boldsymbol{\zeta}_n, & \text{for } k = \bar{Q}_b \end{cases}$$

while for $n = N + 1, \dots, 2N$

$$\boldsymbol{\sigma}_{kn} = \begin{cases} T_p^k \varphi_{n-N}, & \text{for } k = 1, \dots, Q_b \\ 0, & \text{for } k = \bar{Q}_b. \end{cases}$$

We can thus express the RB velocity and pressure solutions as

$$\mathbf{v}_N(\boldsymbol{\mu}) = \sum_{j=1}^{2N} v_{Nj}(\boldsymbol{\mu}) \left(\sum_{k=1}^{\bar{Q}_b} \Theta_b^k(\boldsymbol{\mu}^j)\boldsymbol{\sigma}_{kj} \right), \qquad p_N(\boldsymbol{\mu}) = \sum_{l=1}^{N} p_{Nl}(\boldsymbol{\mu})\varphi_l.$$

Hence, for a new parameter $\boldsymbol{\mu}$, the RB solution of the problem (3.5.11) can be written as a combination of basis functions with weights given by the following reduced basis linear system

$$\begin{cases} \displaystyle\sum_{j=1}^{2N}\sum_{q=1}^{Q_a} \Theta_a^q(\boldsymbol{\mu}) A_{ij}^q\, v_{Nj}(\boldsymbol{\mu}) + \sum_{l=1}^{N}\sum_{q=1}^{Q_b} \Theta_b^q(\boldsymbol{\mu}) B_{li}^q\, p_{Nl}(\boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \Theta_f^q(\boldsymbol{\mu}) F_i^q, & 1 \leq i \leq 2N, \\ \displaystyle\sum_{j=1}^{2N}\sum_{q=1}^{Q_b} \Theta_b^q(\boldsymbol{\mu}) B_{lj}^q\, v_{Nj}(\boldsymbol{\mu}) = \sum_{q=1}^{Q_g} \Theta_g^q(\boldsymbol{\mu}) G_l^q, & 1 \leq l \leq N, \end{cases} \tag{3.5.15}$$

where the submatrices A^q and B^q are given by

$$A^q_{ij} = a^q(\boldsymbol{\sigma}_j, \boldsymbol{\sigma}_i), \qquad B^q_{li} = b^q(\boldsymbol{\sigma}_i, \varphi_l), \qquad 1 \le i, j \le 2N, \quad 1 \le l \le N,$$

and the vectors F^q, G^q by

$$F^q_i = \langle F^q, \boldsymbol{\sigma}_i \rangle, \qquad G^q_l = \langle G^q, \varphi_l \rangle, \qquad 1 \le i \le 2N, \quad 1 \le l \le N.$$

Finally, denoting with $A_N(\boldsymbol{\mu}) = \sum \Theta^q_a A^q$, $B_N(\boldsymbol{\mu}) = \sum \Theta^q_b B^q$, we can rewrite problem (3.5.15) as

$$\underbrace{\begin{pmatrix} A_N(\boldsymbol{\mu}) & B_N^T(\boldsymbol{\mu}) \\ B_N(\boldsymbol{\mu}) & 0 \end{pmatrix}}_{K_N(\boldsymbol{\mu})} \begin{pmatrix} \mathbf{v}_N(\boldsymbol{\mu}) \\ \mathbf{p}_N(\boldsymbol{\mu}) \end{pmatrix} = \begin{pmatrix} \mathbf{F}_N(\boldsymbol{\mu}) \\ \mathbf{G}_N(\boldsymbol{\mu}) \end{pmatrix}. \tag{3.5.16}$$

where $\mathbf{v}_N$ and $\mathbf{p}_N$ are the column vectors of the linear combination coefficient $\{v_{Nj}\}_{j=1}^{2N}$ and $\{p_{Nl}\}_{l=1}^{N}$ respectively. In order to analyze more deeply the structure of the linear system with saddle-point structure (3.5.16) let us define the *basis matrices*

$$Z_v = \begin{pmatrix} \boldsymbol{\sigma}_1 & \cdots & \boldsymbol{\sigma}_{2N} \end{pmatrix} \in \mathbb{R}^{\mathcal{N}_V \times 2N}, \qquad Z_p = \begin{pmatrix} \boldsymbol{\varphi}_1 & \cdots & \boldsymbol{\varphi}_N \end{pmatrix} \in \mathbb{R}^{\mathcal{N}_M \times N},$$

$$Z = \begin{pmatrix} Z_v & 0 \\ 0 & Z_p \end{pmatrix} \in \mathbb{R}^{(2\mathcal{N}_V + \mathcal{N}_M) \times 3N},$$

where $\boldsymbol{\sigma}_n$ and $\boldsymbol{\varphi}_n$ now denote the FE expansions of the RB basis functions. Then $K_N = Z^T K Z$ is given by

$$K_N = \begin{pmatrix} A_N & B_N^T \\ B_N & 0 \end{pmatrix} = \begin{pmatrix} Z_v^T A Z_v & Z_v^T B^T Z_p \\ Z_p^T B Z_v & 0 \end{pmatrix}. \tag{3.5.17}$$

3.5.4 Offline-Online computational procedure and sampling strategies

Thanks to the assumption of affine parameter dependence, we can decouple the formation of the matrix $K_N(\boldsymbol{\mu})$ in two stages, the Offline and Online stages, that enable the efficient resolution of the system (3.5.16) for each new parameter $\boldsymbol{\mu}$. In particular:

1. in the Offline stage, performed only once, we first compute and store the basis function $\{\boldsymbol{\sigma}_i\}_{i=1}^{2N}$ and $\{\varphi_j\}_{j=1}^{N}$, and form the $\boldsymbol{\mu}$-independent matrices A^q_N, $1 \le q \le Q_a$, B^q_N, $1 \le q \le Q_b$ and the vectors F^q_N, $1 \le q \le Q_f$, G^q_N, $1 \le q \le Q_g$. The operation count depends on N, Q_a, Q_b, Q_f, Q_g and $\mathcal{N}$;

2. in the Online stage, performed for each new value $\boldsymbol{\mu}$, we use the precomputed matrices A^q_N, B^q_N and vectors F^q_N, G^q_N to assemble the (full) matrix K_N and the vectors $\mathbf{F}_N$, $\mathbf{G}_N$ appearing in (3.5.5) (3.5.6), with

$$A_N(\boldsymbol{\mu}) = \sum_{q=1}^{Q_a} \Theta^q_a(\boldsymbol{\mu}) A^q_N, \qquad B_N(\boldsymbol{\mu}) = \sum_{q=1}^{Q_b} \Theta^q_b(\boldsymbol{\mu}) B^q_N,$$

$$\mathbf{F}_N(\boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \Theta^q_f(\boldsymbol{\mu}) F^q_N, \qquad \mathbf{G}_N(\boldsymbol{\mu}) = \sum_{q=1}^{Q_g} \Theta^q_f(\boldsymbol{\mu}) G^q_N;$$

we then solve the resulting system to obtain $(\mathbf{v}_N, \mathbf{p}_N)$. The Online operation count depends on N, Q_a, Q_b, Q_f, Q_g but is independent of $\mathcal{N}$. In particular we need $O((Q_a+Q_b)N^2)$ and $O((Q_f+Q_g)N)$ operations to assemble matrices and vectors, and $O((3N)^3)$ operations to solve the RB linear system (3.5.16).

For the construction of the hierarchical Lagrange RB approximation spaces we rely again on the sampling strategy based on the greedy algorithm described in Section 3.2.3. In particular, in each iteration, given the parameter samples $S_N = \{\boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^N\}$, the new sample point $\boldsymbol{\mu}^{N+1}$ to be added is such that

$$\boldsymbol{\mu}^{N+1} = \arg \max_{\boldsymbol{\mu} \in \Xi_{\text{train}}} \Delta_N(\boldsymbol{\mu}),$$

where $\Delta_N(\boldsymbol{\mu})$ is a rigorous, sharp and inexpensive a posteriori error bound for the error on the velocity and pressure variables, i.e.

$$\left(\|\mathbf{v}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathbf{v}_N(\boldsymbol{\mu})\|^2_{H^1} + \|p^{\mathcal{N}}(\boldsymbol{\mu}) - p_N(\boldsymbol{\mu})\|^2_{L^2}\right)^{1/2} \leq \Delta_N(\boldsymbol{\mu}). \tag{3.5.18}$$

The next section is devoted to the construction of such an error estimator.

3.5.5 A posteriori error estimation

The construction of the estimator $\Delta_N(\boldsymbol{\mu})$ will be carried out in the Babuška framework, as proposed in [76]. In fact, as observed in Section A.1, saddle-points problems are a particular case of weakly coercive problem, for which the stability analysis can be carried out by using the Nečas-Babuška theorem. Therefore, to construct the error estimator $\Delta_N(\boldsymbol{\mu})$ it is sufficient to exploit this alternative point of view and rewrite the problem as a weakly coercive problem, for which RB a posteriori error estimates techniques are already available.

In order to formulate the problem (3.5.4) in the standard form of weakly coercive problems (see Section A.1.2), it suffices to denote $Y = V \times M$ and define the bilinear form $\mathsf{A}(\cdot,\cdot;\boldsymbol{\mu})\colon Y \times Y \to \mathbb{R}$ given by

$$\mathsf{A}(\mathsf{v},\mathsf{w};\boldsymbol{\mu}) := a(\boldsymbol{v},\boldsymbol{\xi};\boldsymbol{\mu}) + b(\boldsymbol{\xi},p;\boldsymbol{\mu}) + b(\boldsymbol{v},\tau;\boldsymbol{\mu}), \tag{3.5.19}$$

and the linear continuous functional $\mathsf{F}\colon Y \to \mathbb{R}$

$$\mathsf{F}(\mathsf{w};\boldsymbol{\mu}) = \langle \underline{F}(\boldsymbol{\mu}), \boldsymbol{\xi}\rangle + \langle G(\boldsymbol{\mu}), \tau\rangle. \tag{3.5.20}$$

where $\mathsf{v} = (\boldsymbol{v}, p) \in Y$ and $\mathsf{w} = (\boldsymbol{\xi}, \tau) \in Y$. Then, we can formulate equivalently the problem (3.5.4) as:

$$\text{find } \mathsf{v} \in Y \text{ s.t:} \qquad \mathsf{A}(\mathsf{v},\mathsf{w};\boldsymbol{\mu}) = \mathsf{F}(\mathsf{w};\boldsymbol{\mu}) \qquad \forall \mathsf{w} \in Y. \tag{3.5.21}$$

The problem (3.5.21) is well posed if and only if the following conditions hold (see Section A.1):

1. the bilinear form $\mathsf{A}(\cdot,\cdot;\boldsymbol{\mu})$ is continous, i.e. there exists a constant $\gamma(\boldsymbol{\mu}) > 0$ such that

$$\mathsf{A}(\mathsf{v},\mathsf{w};\boldsymbol{\mu}) \leq \gamma(\boldsymbol{\mu})\|\mathsf{v}\|_Y\|\mathsf{w}\|_Y, \qquad \forall \boldsymbol{\mu} \in \mathcal{D};$$

2. $\mathsf{A}(\cdot,\cdot;\boldsymbol{\mu})$ satisfies the inf-sup condition, i.e. there exists a constant $\tilde{\beta}_0 > 0$ such that

$$\tilde{\beta}(\boldsymbol{\mu}) := \inf_{\mathsf{w}\in Y} \sup_{\mathsf{v}\in Y} \frac{\mathsf{A}(\mathsf{v},\mathsf{w};\boldsymbol{\mu})}{\|\mathsf{v}\|_Y\|\mathsf{w}\|_Y} \geq \tilde{\beta}_0, \qquad \forall \boldsymbol{\mu} \in \mathcal{D};$$

moreover, for each $\boldsymbol{\mu} \in \mathcal{D}$, the unique solution satisfies

$$\|\mathsf{v}(\boldsymbol{\mu})\|_Y \leq \frac{1}{\tilde{\beta}(\boldsymbol{\mu})}\|\mathsf{F}(\cdot;\boldsymbol{\mu})\|_{Y'}, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}. \tag{3.5.22}$$

Actually, since the bilinear forms $a(\cdot,\cdot;\boldsymbol{\mu})$ and $b(\cdot,\cdot;\boldsymbol{\mu})$ satisfy the hypotheses of (Brezzi) Theorem A.2, it can be shown (see e.g. [21, 90, 35]) that the the compound form $\mathsf{A}(\cdot,\cdot;\boldsymbol{\mu})$ is continuous and weakly coercive. Similarly, the FE and RB approximations satisfy the same inf-sup condition,

$$\tilde{\beta}^{\mathcal{N}}(\boldsymbol{\mu}) := \inf_{\mathsf{w}\in Y^{\mathcal{N}}} \sup_{\mathsf{v}\in Y^{\mathcal{N}}} \frac{\mathsf{A}(\mathsf{v},\mathsf{w};\boldsymbol{\mu})}{\|\mathsf{v}\|_Y\|\mathsf{w}\|_Y} > 0, \qquad \forall \boldsymbol{\mu}\in\mathcal{D},$$

$$\tilde{\beta}_N(\boldsymbol{\mu}) := \inf_{\mathsf{w}\in Y_N} \sup_{\mathsf{v}\in Y_N} \frac{\mathsf{A}(\mathsf{v},\mathsf{w};\boldsymbol{\mu})}{\|\mathsf{v}\|_Y\|\mathsf{w}\|_Y} > 0, \qquad \forall \boldsymbol{\mu}\in\mathcal{D},$$

where $Y^{\mathcal{N}} = V^{\mathcal{N}}\times M^{\mathcal{N}}$ and $Y_N = V_N^{\boldsymbol{\mu}}\times M_N$. Moreover the stability estimate (3.5.22) holds also for the FE and RB approximations, in particular

$$\|\mathsf{v}^{\mathcal{N}}(\boldsymbol{\mu})\|_Y \le \frac{1}{\tilde{\beta}^{\mathcal{N}}(\boldsymbol{\mu})}\|\mathsf{F}(\cdot;\boldsymbol{\mu})\|_{Y'}, \qquad \forall \boldsymbol{\mu}\in\mathcal{D}. \tag{3.5.23}$$

In the following we will refer to the inf-sup constant $\tilde{\beta}(\boldsymbol{\mu})$ as the Babuška inf-sup constant (in contrast to the Brezzi inf-sup constant $\beta(\boldsymbol{\mu})$).

Once we have guaranteed the well-posedness of the problem (3.5.21) and its FE and RB approximations, for the construction of the a posteriori error estimator we need the usual two main ingredients: an effective calculation of a lower bound for the Babuška inf-sup constant $\tilde{\beta}^{\mathcal{N}}(\boldsymbol{\mu})$ and the (standard) calculation of the dual norm of the residual. For the first one we assume that we can calculate a $\boldsymbol{\mu}$-dependent lower bound $\tilde{\beta}_{\mathrm{LB}}(\boldsymbol{\mu})$ for the inf-sup constant $\tilde{\beta}^{\mathcal{N}}(\boldsymbol{\mu})$, i.e. $\tilde{\beta}^{\mathcal{N}}(\boldsymbol{\mu}) \ge \tilde{\beta}_{\mathrm{LB}}(\boldsymbol{\mu}) \ge \tilde{\beta}_0 > 0$, $\forall \boldsymbol{\mu}\in\mathcal{D}$. The calculation of $\tilde{\beta}_{\mathrm{LB}}(\boldsymbol{\mu})$ will be carried out using the *Natural Norm Successive Constraint Method*, an improvement of the already discussed SCM algorithm specifically tailored for noncoercive problems, see [43, 76].

Let us firstly define the errors between the FE and the RB approximations:

$$\boldsymbol{e_v}(\boldsymbol{\mu}) = \boldsymbol{v}^{\mathcal{N}}(\boldsymbol{\mu}) - \boldsymbol{v}_N(\boldsymbol{\mu}), \qquad e_p(\boldsymbol{\mu}) = p^{\mathcal{N}}(\boldsymbol{\mu}) - p_N(\boldsymbol{\mu}),$$

and

$$\mathsf{e}(\boldsymbol{\mu}) = (\boldsymbol{e_v}(\boldsymbol{\mu}), e_p(\boldsymbol{\mu})) = \mathsf{v}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathsf{v}_N(\boldsymbol{\mu}). \tag{3.5.24}$$

Then we define the residuals

$$r_{\boldsymbol{v}}(\boldsymbol{\xi};\boldsymbol{\mu}) = \langle \underline{F}(\boldsymbol{\mu}),\boldsymbol{\xi}\rangle - a(\boldsymbol{v}_N,\boldsymbol{\xi};\boldsymbol{\mu}) - b(\boldsymbol{\xi},p_N;\boldsymbol{\mu}) \qquad \forall \boldsymbol{\xi}\in V^{\mathcal{N}},$$

$$r_p(\tau;\boldsymbol{\mu}) = \langle G(\boldsymbol{\mu}),\tau\rangle - b(\boldsymbol{v}_N,\tau;\boldsymbol{\mu}) \qquad \forall \tau\in M^{\mathcal{N}},$$

and the *global* residual

$$\mathsf{r}(\mathsf{w};\boldsymbol{\mu}) = \mathsf{F}(\mathsf{w};\boldsymbol{\mu}) - \mathsf{A}(\mathsf{v}_N,\mathsf{w};\boldsymbol{\mu}) \equiv r_{\boldsymbol{v}}(\boldsymbol{\xi};\boldsymbol{\mu}) + r_p(\tau;\boldsymbol{\mu}) \qquad \forall \mathsf{w}\in Y^{\mathcal{N}};$$

note that $\mathsf{r}(\cdot;\boldsymbol{\mu})\in(Y^{\mathcal{N}})'$. The problem statement for $(\boldsymbol{v}^{\mathcal{N}},p^{\mathcal{N}})$ (3.5.7) and $(\boldsymbol{v}_N,p_N)$ (3.5.11) and the bilinearity of $a(\cdot,\cdot;\boldsymbol{\mu})$ and $b(\cdot,\cdot;\boldsymbol{\mu})$ imply that the errors satisfy the following equations

$$\begin{cases} a(\boldsymbol{e}_v(\boldsymbol{\mu}),\boldsymbol{\xi};\boldsymbol{\mu}) + b(\boldsymbol{\xi},e_p(\boldsymbol{\mu});\boldsymbol{\mu}) = r_{\boldsymbol{v}}(\boldsymbol{\xi},\boldsymbol{\mu}) & \forall \boldsymbol{\xi}\in V^{\mathcal{N}},\\ b(\boldsymbol{e}_v(\boldsymbol{\mu}),\tau;\boldsymbol{\mu}) = r_p(\tau;\boldsymbol{\mu}) & \forall \tau\in Q^{\mathcal{N}}, \end{cases} \tag{3.5.25}$$

or equivalently

$$\mathsf{A}(\mathsf{e},\mathsf{w};\boldsymbol{\mu}) = \mathsf{r}(\mathsf{w};\boldsymbol{\mu}) \qquad \forall \mathsf{w}\in Y^{\mathcal{N}}.$$

By using the stability estimate (3.5.23) we obtain the following residual-based estimation

$$\|\mathsf{e}(\boldsymbol{\mu})\|_Y \leq \frac{1}{\tilde{\beta}^{\mathcal{N}}(\boldsymbol{\mu})}\|\mathsf{r}(\cdot;\boldsymbol{\mu})\|_{Y'}, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{3.5.26}$$

exploiting the lower bound for the inf-sup constant

$$\|\mathsf{e}(\boldsymbol{\mu})\|_Y \leq \frac{1}{\tilde{\beta}_{\mathrm{LB}}(\boldsymbol{\mu})}\|\mathsf{r}(\cdot;\boldsymbol{\mu})\|_{Y'} := \Delta_N(\boldsymbol{\mu}). \tag{3.5.27}$$

We can rewrite (3.5.27) equivalently as

$$\|\boldsymbol{v}^{\mathcal{N}}(\boldsymbol{\mu}) - \boldsymbol{v}_N^{\mathcal{N}}(\boldsymbol{\mu})\|_V^2 + \|p^{\mathcal{N}}(\boldsymbol{\mu}) - p_N^{\mathcal{N}}(\boldsymbol{\mu})\|_M^2 \\ \leq \frac{1}{\tilde{\beta}_{\mathrm{LB}}^2(\boldsymbol{\mu})}\Big(\|r_{\boldsymbol{v}}(\cdot;\boldsymbol{\mu})\|_{V'}^2 + \|r_p(\cdot;\boldsymbol{\mu})\|_{M'}^2\Big) \equiv \big(\Delta_N(\boldsymbol{\mu})\big)^2.$$

Offline-Online procedure

We introduce the Riesz representation of $\mathsf{r}(\cdot;\boldsymbol{\mu})$: $\hat{\mathsf{e}}(\boldsymbol{\mu}) \in Y^{\mathcal{N}}$ satisfies

$$(\hat{\mathsf{e}}(\boldsymbol{\mu}), \mathsf{w})_Y = \mathsf{r}(\mathsf{w};\boldsymbol{\mu}), \qquad \forall \mathsf{w} \in Y^{\mathcal{N}}. \tag{3.5.28}$$

The dual norm of the residuals can be evaluated through its Riesz representation:

$$\|\mathsf{r}(\cdot;\boldsymbol{\mu})\|_{Y'} = \sup_{\mathsf{w}\in Y^{\mathcal{N}}} \frac{\mathsf{r}(\mathsf{w};\boldsymbol{\mu})}{\|\mathsf{w}\|_Y} = \|\hat{\mathsf{e}}(\boldsymbol{\mu})\|_Y,$$

From the affine decompositions of the bilinear forms (3.5.5) we can write equivalently

$$\mathsf{A}(\mathsf{v},\mathsf{w};\boldsymbol{\mu}) = \sum_{q=1}^{Q_a+2Q_b} \Theta_A^q(\boldsymbol{\mu})\mathsf{A}^q(\mathsf{v},\mathsf{w}) \tag{3.5.29}$$

where

$$\begin{aligned} &\Theta_A^q(\boldsymbol{\mu}) = \Theta_a^q(\boldsymbol{\mu}), && \mathsf{A}^q(\mathsf{v},\mathsf{w}) = a^q(\boldsymbol{v},\boldsymbol{\xi}), && 1 \leq q \leq Q_a, \\ &\Theta_A^q(\boldsymbol{\mu}) = \Theta_b^q(\boldsymbol{\mu}), && \mathsf{A}^q(\mathsf{v},\mathsf{w}) = b^q(\boldsymbol{\xi},p), && Q_a+1 \leq q \leq Q_a+Q_b, \\ &\Theta_A^q(\boldsymbol{\mu}) = \Theta_b^q(\boldsymbol{\mu}), && \mathsf{A}^q(\mathsf{v},\mathsf{w}) = b^q(\boldsymbol{v},\tau), && Q_a+Q_b+1 \leq q \leq Q_a+2Q_b. \end{aligned}$$

Similarly, using (3.5.6), the linear functional $\mathsf{F}(\cdot;\boldsymbol{\mu})$ can be expressed as

$$\mathsf{F}(\mathsf{w};\boldsymbol{\mu}) = \sum_{q=1}^{Q_f+Q_g} \Theta_F^q(\boldsymbol{\mu})\mathsf{F}^q(\mathsf{w}), \tag{3.5.30}$$

where

$$\begin{aligned} &\Theta_F^q(\boldsymbol{\mu}) = \Theta_f^q(\boldsymbol{\mu}), && \mathsf{F}^q(\mathsf{w}) = \langle F^q, \boldsymbol{\xi}\rangle, && 1 \leq q \leq Q_f, \\ &\Theta_F^q(\boldsymbol{\mu}) = \Theta_g^q(\boldsymbol{\mu}), && \mathsf{F}^q(\mathsf{w}) = \langle G^q, \tau\rangle, && Q_f+1 \leq q \leq Q_f+Q_g. \end{aligned}$$

In this way, recalling that $\mathsf{v}_N(\boldsymbol{\mu}) = (\boldsymbol{v}_N(\boldsymbol{\mu}), p_N(\boldsymbol{\mu})) \in \mathbb{R}^{3N}$ denotes the global vector of the RB components, the residual can be expressed as

$$\begin{aligned} \mathsf{r}(\mathsf{w};\boldsymbol{\mu}) &= \mathsf{F}(\mathsf{w};\boldsymbol{\mu}) - \mathsf{A}\left(\sum_{n=1}^{3N} \mathsf{v}_{Nn}(\boldsymbol{\mu})\Phi_n, \mathsf{w}; \boldsymbol{\mu}\right) \\ &= \sum_{q=1}^{Q_F} \Theta_F^q(\boldsymbol{\mu})\mathsf{F}^q(\mathsf{w}) - \sum_{n=1}^{3N} \mathsf{v}_{Nn}(\boldsymbol{\mu}) \sum_{q=1}^{Q_A} \Theta_A^q(\boldsymbol{\mu})\mathsf{A}^q(\Phi_n, \mathsf{w}), \end{aligned} \tag{3.5.31}$$

where $Q_A = Q_a + 2Q_b$, $Q_F = Q_f + Q_g$ and

$$\Phi_n = (\boldsymbol{\sigma}_n, 0), \quad 1 \le n \le 2N, \qquad \Phi_n = (0, \varphi_n), \quad 2N+1 \le n \le 3N.$$

Recalling the relation (3.5.28), we thus may write $\hat{\mathsf{e}}(\boldsymbol{\mu}) \in Y^{\mathcal{N}}$ as

$$\hat{\mathsf{e}}(\boldsymbol{\mu}) = \sum_{q=1}^{Q_F} \Theta_F^q(\boldsymbol{\mu}) \mathcal{F}^q - \sum_{n=1}^{3N} \mathsf{v}_{Nn}(\boldsymbol{\mu}) \sum_{q=1}^{Q_A} \Theta_A^q(\boldsymbol{\mu}) \mathcal{L}_n^q,$$

where $\forall \mathsf{w} \in Y^{\mathcal{N}}$

$$(\mathcal{F}^q, \mathsf{w})_Y = \mathsf{F}^q(\mathsf{w}), \qquad 1 \le q \le Q_F, \tag{3.5.32}$$

$$(\mathcal{L}_n^q, v)_Y = -\mathsf{A}^q(\Phi_n, \mathsf{w}), \qquad 1 \le q \le Q_A, \quad 1 \le n \le 3N. \tag{3.5.33}$$

Note that $\mathcal{F}^q$ is the Riesz representation of $\mathsf{F}^q(\cdot)$ and $\mathcal{L}_n^q$ is the Riesz representation of $\mathsf{A}^q(\Phi_n, \cdot)$, computable solving parameter-independent Poisson-like problems. Finally we obtain

$$\begin{aligned} \|\hat{\mathsf{e}}(\boldsymbol{\mu})\|_Y^2 = \sum_{q=1}^{Q_F} \sum_{q'=1}^{Q_F} \Theta_F^q(\boldsymbol{\mu}) \Theta_F^{q'}(\boldsymbol{\mu}) (\mathcal{F}^q, \mathcal{F}^{q'})_Y + \sum_{q=1}^{Q_A} \sum_{n=1}^{3N} \Theta_A^q(\boldsymbol{\mu}) \mathsf{v}_{Nn}(\boldsymbol{\mu}) \Bigg\{ \\ 2 \sum_{q'=1}^{Q_F} \Theta_F^{q'}(\boldsymbol{\mu}) (\mathcal{F}^{q'}, \mathcal{L}_n^q)_Y + \sum_{q'=1}^{Q_A} \sum_{n'=1}^{3N} \Theta_A^{q'}(\boldsymbol{\mu}) \mathsf{v}_{Nn'}(\boldsymbol{\mu}) (\mathcal{L}_n^q, \mathcal{L}_n^{q'})_Y \Bigg\} \end{aligned} \tag{3.5.34}$$

from which we can calculate the dual norm of the residual. Let us summarize the Offline-Online decomposition:

1. in the Offline stage we first compute the Q_F terms $\mathcal{F}^q$ and the $3NQ_A$ terms $\mathcal{L}_n^q$ solving problems (3.5.32) and (3.5.33) respectively; then we store the scalar products $(\mathcal{F}^q, \mathcal{F}^{q'})_Y$, $(\mathcal{F}^{q'}, \mathcal{L}_n^q)_Y$ and $(\mathcal{L}_n^q, \mathcal{L}_n^q)_Y$. The Offline operation count depends then on N, Q_A, Q_F and $\mathcal{N}$.

2. in the Online stage, for each new value of $\boldsymbol{\mu}$, we simply evaluate the sum (3.5.34) in terms of the $\Theta^q(\boldsymbol{\mu})$, v_{Nn} and the precomputed scalar products. The Online operation count is $O(9N^2Q_A^2 + 6NQ_AQ_f + 3NQ_F^2)$, independent of $\mathcal{N}$.

3.5.6 Numerical example: Couette flow

This example deals with a Couette flow in a pipe of uniform cross-section. Namely, we refer to a slightly different version of one of the test cases proposed in [76]. We consider the physical domain shown in Figure 3.4 and the following problem:

$$\begin{cases} -\mu_3 \Delta \boldsymbol{v}_o + \nabla p_o = \boldsymbol{f}_o(\boldsymbol{\mu}) & \text{in } \Omega_o(\boldsymbol{\mu}) \\ \operatorname{div} \boldsymbol{v}_o = 0 & \text{in } \Omega_o(\boldsymbol{\mu}) \\ v_{o1} = x_{o2}, \quad v_{o2} = 0 & \text{on } \Gamma_D^o(\boldsymbol{\mu}) \\ v_{o2} = 0, \quad -p_o \boldsymbol{n}_1 + \mu_3 \dfrac{\partial v_{o1}}{\partial x_{o1}} \boldsymbol{n}_1 = 0 & \text{on } \Gamma_N^o(\boldsymbol{\mu}), \end{cases} \tag{3.5.35}$$

where the forcing term is $\boldsymbol{f}_o(\boldsymbol{\mu}) = (0, -\mu_2)$, $\mu_3 = \nu$ is the kinematic viscosity and μ_1 is the height of the channel (and thus also the value of the horizontal velocity v_{o1} imposed on the upper boundary). The parameter domain is given by $\mathcal{D} = [0.5, 2] \times [0.5, 1.5] \times [0.1, 1]$.

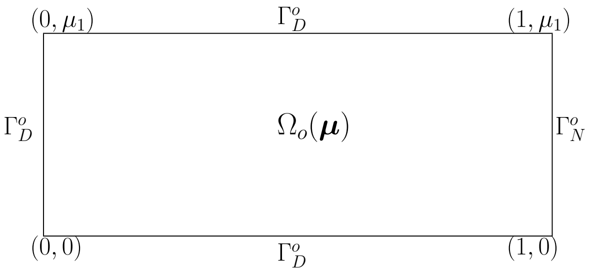


Figure 3.4: Original domain $\Omega_o(\boldsymbol{\mu})$ for the Couette flow.

The problem is then mapped to a fixed reference domain $\Omega = \Omega(\boldsymbol{\mu}_{\text{ref}})$ with $\boldsymbol{\mu}_{\text{ref}} = [1, 1, 0.5]$ and discretized by $\mathbb{P}^2 - \mathbb{P}^1$ Taylor-Hood finite elements [69]. We have only one subdomain, the corresponding affine geometrical mapping is simply given by

$$\boldsymbol{C}^1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \boldsymbol{G}^1 = \begin{bmatrix} 1 & 0 \\ 0 & \mu_1 \end{bmatrix}, \qquad J^r = \mu_1.$$

As regards the affine decomposition we have $Q_a = 2$, $Q_b = 2$, $Q_f = 1$, $Q_g = 4$ and, for instance,

$$\begin{aligned}
\Theta_a^1(\boldsymbol{\mu}) &= \mu_1\mu_3, & a^1(\boldsymbol{v},\boldsymbol{\xi}) &= \int_\Omega \frac{\partial \boldsymbol{v}}{\partial x_1} \cdot \frac{\partial \boldsymbol{\xi}}{\partial x_1}\, d\Omega, \\
\Theta_a^2(\boldsymbol{\mu}) &= \frac{\mu_3}{\mu_1}, & a^2(\boldsymbol{v},\boldsymbol{\xi}) &= \int_\Omega \frac{\partial \boldsymbol{v}}{\partial x_2} \cdot \frac{\partial \boldsymbol{\xi}}{\partial x_2}\, d\Omega, \\
\Theta_b^1(\boldsymbol{\mu}) &= \mu_1, & b^1(\boldsymbol{\xi},\tau) &= -\int_\Omega \tau \frac{\partial \xi_1}{\partial x_1}\, d\Omega, \\
\Theta_b^2(\boldsymbol{\mu}) &= 1, & b^2(\boldsymbol{\xi},\tau) &= -\int_\Omega \tau \frac{\partial \xi_2}{\partial x_2}\, d\Omega.
\end{aligned}$$

The dimension of the FE space $Y^{\mathcal{N}}$ is $\mathcal{N} = 7416$. In Figure 3.5 and 3.6 we report some plots of the solution for different values of the parameters. With a fixed tolerance $\varepsilon_{\text{tol}} = 5 \cdot 10^{-4}$, $N_{max} = 9$ basis functions have been selected by the Greedy algorithm. In Figure 3.7 we show the lower bound for the Babuška inf-sup constant $\tilde{\beta}^{\mathcal{N}}(\boldsymbol{\mu})$ obtained using the SCM algorithm; in Figure 3.8 we compare the a posteriori error bound $\Delta_N(\boldsymbol{\mu})$ with the true error $\|\mathrm{v}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathrm{v}_N(\boldsymbol{\mu})\|_Y$, in particular we show the average of these two quantities over a sample of 250 parameters.

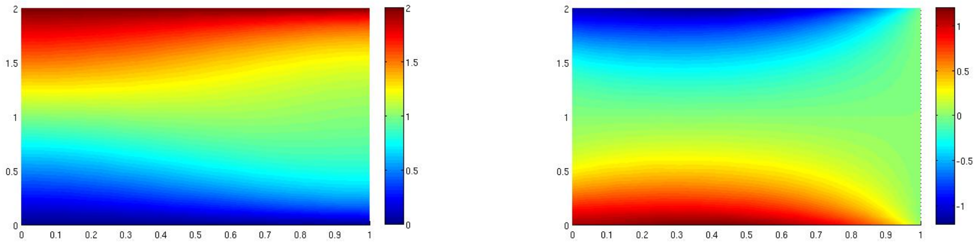


Figure 3.5: Couette flow: representative solution for $\boldsymbol{\mu} = (2, 1.5, 0.3)$. Velocity on the left, pressure on the right.

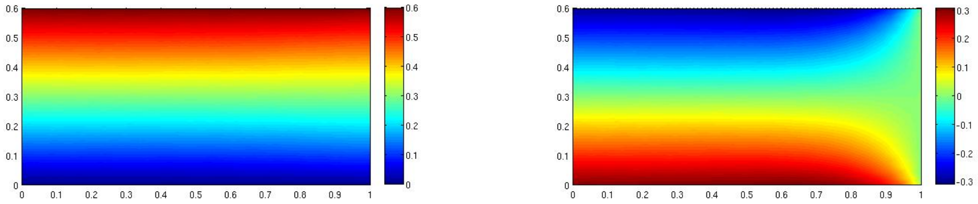
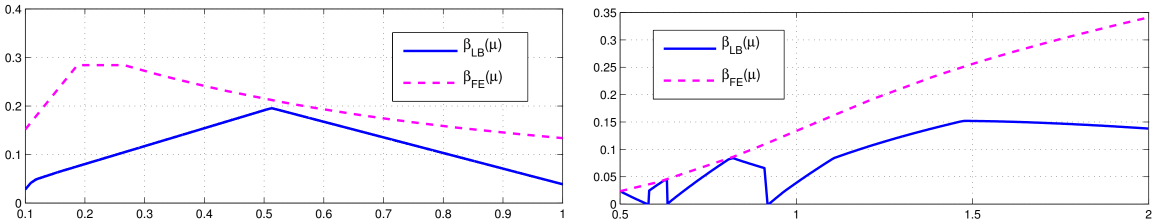


Figure 3.6: Couette flow: representative solution for $\boldsymbol{\mu} = (0.6, 1, 0.7)$. Velocity on the left, pressure on the right.



(a) $\tilde{\beta}(\boldsymbol{\mu})$ as a function of the viscosity μ_3, $(\mu_1, \mu_2) = (1, 1.4)$ fixed.

(b) $\tilde{\beta}(\boldsymbol{\mu})$ as a function of channel height μ_1, $(\mu_2, \mu_3) = (1, 1)$ fixed.

Figure 3.7: Couette flow: lower bound for the Babuška inf-sup constant $\tilde{\beta}(\boldsymbol{\mu})$.

The Offline computational time is equal to $t_{RB}^{offline} = 1900s$, in particular we underline that the SCM algorithm takes about the 70% of the overall Offline computational time (requiring to solve in this case 43 eigenproblems). The Online solution time is of order 10^{-2} seconds, comprehensive of the time needed to evaluate the a posteriori error estimator; the computational time needed to build and solve the RB linear system alone is very small, less than $10^{-3}s$ (in fact we are solving a system of size 27×27).

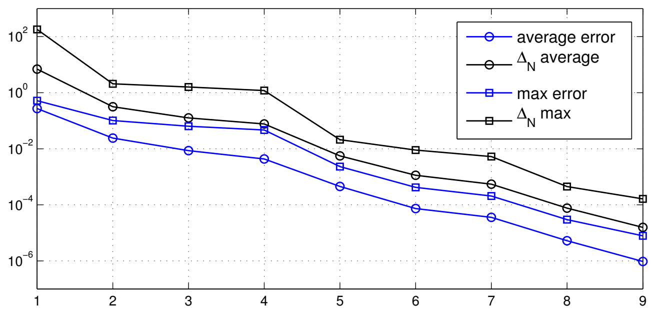


Figure 3.8: Couette flow: a posteriori error bound. Comparison of the average computed error (between the *truth* FE solution and the RB approximation $\|\mathsf{v}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathsf{v}_N(\boldsymbol{\mu})\|_Y$) and the estimator $\Delta(\boldsymbol{\mu})$, for $N = 1, \cdots, N_{max} = 9$.

Chapter 4

Reduced Basis Method for Parametrized Elliptic Optimal Control Problems

We provide here a reduced basis framework for the efficient solution of parametrized linear/quadratic optimal control problems governed by coercive second-order elliptic PDEs. We thus consider the abstract problem ($OCP_{\boldsymbol{\mu}}$) stated in the Introduction under the form discussed in Chapter 1:

$$\begin{cases} \min \ \mathcal{J}(\underline{x};\boldsymbol{\mu}) = \dfrac{1}{2}\mathcal{A}(\underline{x},\underline{x};\boldsymbol{\mu}) - \langle \underline{F}(\boldsymbol{\mu}), \underline{x}\rangle, & \text{subject to} \\ \mathcal{B}(\underline{x},q;\boldsymbol{\mu}) = 0 \qquad \forall q \in Q. \end{cases} \tag{4.0.1}$$

being $\mathcal{B}(\cdot,\cdot;\boldsymbol{\mu})$ the bilinear form associated to the PDE (i.e. the state equation). In analogy to the results given in Chapter 1, the parametrized optimality conditions system in saddle-point form reads: find $(\underline{x}(\boldsymbol{\mu}), p(\boldsymbol{\mu})) \in X \times Q$ such that

$$\begin{cases} \mathcal{A}(\underline{x},\underline{w};\boldsymbol{\mu}) + \mathcal{B}(\underline{w},p;\boldsymbol{\mu}) = \langle \underline{F}(\boldsymbol{\mu}),\underline{w}\rangle & \forall \underline{w} \in X, \\ \mathcal{B}(\underline{x},q;\boldsymbol{\mu}) = 0 & \forall q \in Q. \end{cases} \tag{4.0.2}$$

In this case the goal of a reduced basis (RB) method is the computation of parameter-dependent solutions of the optimal control problem by solving a saddle-point problem of low dimension. The main features of our RB approach will be the following.

1. A reduced basis made by the FE solutions $(\underline{x}^{\mathcal{N}}(\boldsymbol{\mu}^i), p^{\mathcal{N}}(\boldsymbol{\mu}^i))$ of (4.0.2) for some selected parameter values $S = \{\boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^N\}$ (by a proper greedy procedure), giving the spaces $X_N = \text{span}\{\underline{x}^{\mathcal{N}}(\boldsymbol{\mu}^i)\}_i$ for the state and control variables and $Q_N = \text{span}\{p^{\mathcal{N}}(\boldsymbol{\mu}^i)\}_i$ for the adjoint.

2. A reduced approximation of the saddle-point problem obtained as a Galerkin projection onto the reduced spaces: find $(\underline{x}_N(\boldsymbol{\mu}), p_N(\boldsymbol{\mu})) \in X_N \times Q_N$ such that

$$\begin{cases} \mathcal{A}(\underline{x}_N(\boldsymbol{\mu}),\underline{w};\boldsymbol{\mu}) + \mathcal{B}(\underline{w},p_N(\boldsymbol{\mu});\boldsymbol{\mu}) = \langle \underline{F}(\boldsymbol{\mu}),\underline{w}\rangle, & \forall \underline{w} \in X_N \\ \mathcal{B}(\underline{x}_N(\boldsymbol{\mu}),q;\boldsymbol{\mu}) = 0, & \forall q \in Q_N. \end{cases}$$

   In particular we prove existence, uniqueness and stability of the RB approximation by means of Brezzi theorem, in particular in order to fulfill an appropriate inf-sup condition we investigate various enrichment strategy for the RB spaces.

3. A *rigorous*, sharp and inexpensive a posteriori error estimator $\Delta_N(\boldsymbol{\mu})$ for both the control, state and adjoint variables, i.e. such that

$$\left(\|\underline{x}^{\mathcal{N}}(\boldsymbol{\mu}) - \underline{x}_N(\boldsymbol{\mu})\|_X^2 + \|p^{\mathcal{N}}(\boldsymbol{\mu}) - p_N(\boldsymbol{\mu})\|_Q^2\right)^{1/2} \leq \Delta_N(\boldsymbol{\mu}).$$

   In particular for the a posteriori error analysis we follow the guidelines of Stokes problems (see Section 3.5), i.e. we develop the analysis in the framework of noncoercive problems. Moreover, following the approach proposed in [17], we obtain a rigorous estimator for the cost functional

$$|\mathcal{J}(\underline{x}^{\mathcal{N}}(\boldsymbol{\mu}); \boldsymbol{\mu}) - \mathcal{J}(\underline{x}_N(\boldsymbol{\mu}); \mu)| \leq \Delta_N^J(\boldsymbol{\mu}).$$

4. The standard Offline/Online decomposition stratagem [77], that enables to decouple the generation and projection stages of the RB approximation. In particular we rely on the affine parameter dependence assumption on the operator defining the cost functional and the state equation.

The chapter is structured as follows. In Section 4.1 we introduce the formulation of parametrized linear/quadratic optimal control problems (governed by elliptic coercive PDEs) with affine parameter dependence; we briefly discuss also the FE approximation, recalling the necessary assumptions to ensure the well-posedness. In Section 4.2 we discuss the RB approximation and the main features of the method, focusing on the corresponding stability condition for the RB approximation. Then in Section 4.3 we deal with the a posteriori error estimation for the RB solution and functional based on the Babuška stability theory. Finally, in Section 4.4 some numerical examples are presented.

4.1 Problem formulation

We consider the parametrized version of the optimal control problems introduced in Section 1.2.3. Let $\Omega \subset \mathbb{R}^d$ $(d = 1, 2, 3)$ be a spatial domain with Lipschitz boundary $\partial\Omega$, we define the functional spaces Y (state space) and Q (adjoint space) such that $H_0^1(\Omega) \subset Y \subset H^1(\Omega)$ and $Q \equiv Y$, respectively. The control space is $U = L^2(\omega)$, where ω can be the whole domain Ω, a subdomain or a boundary. Moreover $\mathcal{Z}$ shall denote the observation space. We consider the following parametrized optimal control problem:

$$\begin{aligned} \text{minimize} \quad & J(y, u) = \frac{1}{2} m(y - y_d(\boldsymbol{\mu}), y - y_d(\boldsymbol{\mu}); \boldsymbol{\mu}) + \frac{\alpha}{2} n(u, u; \boldsymbol{\mu}), \\ \text{s.t.} \quad & a(y, q; \boldsymbol{\mu}) = c(u, q; \boldsymbol{\mu}) + \langle G(\boldsymbol{\mu}), q \rangle \qquad \forall q \in Q. \end{aligned} \tag{4.1.1}$$

Let us precise the hypotheses on the linear and bilinear forms (see also Section 1.2.3). We assume that the bilinear form $a(\cdot, \cdot; \boldsymbol{\mu}) : Y \times Q \to \mathbb{R}$ is bounded over $Y \times Q$. i.e.

$$\tilde{\gamma}_1(\boldsymbol{\mu}) = \sup_{z \in Y} \sup_{q \in Q} \frac{a(z, q; \boldsymbol{\mu})}{\|z\|_Y \|q\|_Q} < +\infty, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{4.1.2}$$

and coercive over $Y \equiv Q$, i.e. there exists $\tilde{\alpha}_0 > 0$ such that

$$\tilde{\alpha}(\boldsymbol{\mu}) = \inf_{z \in Y} \frac{a(z, z; \boldsymbol{\mu})}{\|z\|_Y^2} = \inf_{q \in Q} \frac{a(q, q; \boldsymbol{\mu})}{\|q\|_Q^2} \geq \tilde{\alpha}_0, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}. \tag{4.1.3}$$

We assume that the bilinear form $c(\cdot,\cdot;\boldsymbol{\mu}) : U \times Q \to \mathbb{R}$ is symmetric and bounded over $U \times Q$, i.e.

$$\tilde{\gamma}_2(\boldsymbol{\mu}) = \sup_{u\in U}\sup_{q\in Q} \frac{c(u,q;\boldsymbol{\mu})}{\|u\|_U\|q\|_Q} < +\infty, \qquad \forall\boldsymbol{\mu}\in\mathcal{D}$$

and the bilinear form $n(\cdot,\cdot;\boldsymbol{\mu}) : U \times U \to \mathbb{R}$ is symmetric, bounded over $U \times U$, i.e.

$$\tilde{\gamma}_n(\boldsymbol{\mu}) = \sup_{u\in U}\sup_{v\in U} \frac{n(u,v;\boldsymbol{\mu})}{\|u\|_U\|v\|_U} < +\infty, \qquad \forall\boldsymbol{\mu}\in\mathcal{D}$$

and coercive over U

$$\tilde{\alpha}_n(\boldsymbol{\mu}) = \inf_{v\in U} \frac{n(v,v;\boldsymbol{\mu})}{\|v\|_U^2} \geq \tilde{\alpha}_{n0}, \qquad \forall\boldsymbol{\mu}\in\mathcal{D}.$$

Moreover we assume the bilinear form $m(\cdot,\cdot;\boldsymbol{\mu})$ to be symmetric, continuous and positive in the norm induced on the space $\mathcal{Z}$, $y_d(\boldsymbol{\mu}) \in \mathcal{Z}$ and $G(\boldsymbol{\mu}) \in Q', \forall\boldsymbol{\mu}\in\mathcal{D}$. We shall make an additional assumptions, crucial to Offline-Online procedures, by assuming the bilinear and linear forms to be affine in the parameter $\boldsymbol{\mu}$, i.e. for some finite $\tilde{Q}_*$, $* \in \{a,c,n,m,g\}$, they can be expressed as

$$\begin{aligned}
a(z,q;\boldsymbol{\mu}) &= \sum_{q=1}^{\tilde{Q}_a} \tilde{\Theta}_a^q(\boldsymbol{\mu})\, a^q(z,q), & c(v,q;\boldsymbol{\mu}) &= \sum_{q=1}^{\tilde{Q}_c} \tilde{\Theta}_c^q(\boldsymbol{\mu})\, c^q(v,q),\\
m(y,z;\boldsymbol{\mu}) &= \sum_{q=1}^{\tilde{Q}_m} \tilde{\Theta}_m^q(\boldsymbol{\mu})\, m^q(y,z), & n(u,v;\boldsymbol{\mu}) &= \sum_{q=1}^{\tilde{Q}_n} \tilde{\Theta}_n^q(\boldsymbol{\mu})\, n^q(u,v),\\
\langle G(\boldsymbol{\mu}),q\rangle &= \sum_{q=1}^{\tilde{Q}_g} \tilde{\Theta}_g^q(\boldsymbol{\mu})\, \langle G^q,q\rangle,
\end{aligned} \tag{4.1.4}$$

for given *smooth* $\boldsymbol{\mu}$-dependent function $\tilde{\Theta}_*^q(\boldsymbol{\mu})$ and continuous $\boldsymbol{\mu}$-independent bilinear and linear forms $a^q(\cdot,\cdot)$, $c^q(\cdot,\cdot)$, etc. In order to formulate the optimal control problem (4.1.1) as a saddle-point problem, we denote $X = Y \times U$, $\underline{x} = (y,u) \in X$, $\underline{w} = (z,v) \in X$, $p,q \in Q$ and define the bilinear form $\mathcal{A}(\cdot,\cdot;\boldsymbol{\mu}) : X \times X \to \mathbb{R}$ as

$$\mathcal{A}(\underline{x},\underline{w};\boldsymbol{\mu}) = m(y,z;\boldsymbol{\mu}) + \alpha n(u,v;\boldsymbol{\mu}), \qquad \forall \underline{x},\underline{w}\in X,$$

and the bilinear form $\mathcal{B}(\cdot,\cdot;\boldsymbol{\mu}) : X \times Q \to \mathbb{R}$ as

$$\mathcal{B}(\underline{w},q;\boldsymbol{\mu}) = a(z,q;\boldsymbol{\mu}) - c(v,q;\boldsymbol{\mu}), \qquad \forall \underline{w}\in X, q\in Q.$$

Defining $\underline{F}(\boldsymbol{\mu}) = m(y_d(\boldsymbol{\mu}),\cdot) \in X'$, we can reformulate the problem (4.1.1) as (for the details see Section 1.2.3): given $\boldsymbol{\mu}\in\mathcal{D}$,

$$\begin{cases} \min\ \mathcal{J}(\underline{x};\boldsymbol{\mu}) = \dfrac{1}{2}\mathcal{A}(\underline{x},\underline{x};\boldsymbol{\mu}) - \langle \underline{F}(\boldsymbol{\mu}),\underline{x}\rangle, & \text{subject to}\\ \mathcal{B}(\underline{x},q;\boldsymbol{\mu}) = \langle G(\boldsymbol{\mu}),q\rangle \qquad \forall q\in Q. \end{cases} \tag{4.1.5}$$

Recalling the results proved in Section 1.2.3, we know that the assumptions made on the linear and bilinear forms in (4.1.1) guarantees the fulfilment of the hypotheses of Brezzi theorem (see Theorem A.2 and Proposition 1.1), which implies the equivalence between

(4.1.5) and the following saddle-point problem: given $\boldsymbol{\mu} \in \mathcal{D}$, find $(\underline{x}(\boldsymbol{\mu}), p(\boldsymbol{\mu})) \in X \times Q$ such that

$$\begin{cases} \mathcal{A}(\underline{x}(\boldsymbol{\mu}), \underline{w}; \boldsymbol{\mu}) + \mathcal{B}(\underline{w}, p(\boldsymbol{\mu}); \boldsymbol{\mu}) = \langle \underline{F}(\boldsymbol{\mu}), \underline{w} \rangle & \forall \underline{w} \in X, \\ \mathcal{B}(\underline{x}(\boldsymbol{\mu}), q; \boldsymbol{\mu}) = \langle G(\boldsymbol{\mu}), q \rangle & \forall q \in Q. \end{cases} \tag{4.1.6}$$

In particular the bilinear forms $\mathcal{A}(\cdot, \cdot; \boldsymbol{\mu})$ and $\mathcal{B}(\cdot, \cdot; \boldsymbol{\mu})$ satisfy the following assumptions:

1. the bilinear form $\mathcal{A}(\cdot, \cdot; \boldsymbol{\mu})$ is continuous over $X \times X$:

$$\gamma_a(\boldsymbol{\mu}) = \sup_{\underline{x} \in X} \sup_{\underline{w} \in X} \frac{\mathcal{A}(\underline{x}, \underline{w}; \boldsymbol{\mu})}{\|\underline{w}\|_X \|\underline{x}\|_X} < +\infty, \qquad \forall \boldsymbol{\mu} \in \mathcal{D};$$

2. the bilinear form $\mathcal{A}(\cdot, \cdot; \boldsymbol{\mu})$ is coercive over $X_0 = \{\underline{w} \in X : \mathcal{B}(\underline{w}, q; \boldsymbol{\mu}) = 0 \quad \forall q \in Q\} \subset X$, i.e. there exists a constant $\alpha_0 > 0$ such that

$$\alpha(\boldsymbol{\mu}) = \inf_{\underline{x} \in X_0} \frac{\mathcal{A}(\underline{x}, \underline{x}; \boldsymbol{\mu})}{\|\underline{x}\|_X^2} \geq \alpha_0, \qquad \forall \boldsymbol{\mu} \in \mathcal{D};$$

3. the bilinear form $\mathcal{B}(\cdot, \cdot; \boldsymbol{\mu})$ is continuous over $X \times Q$

$$\gamma_b(\boldsymbol{\mu}) = \sup_{\underline{w} \in X} \sup_{q \in Q} \frac{\mathcal{B}(\underline{w}, q; \boldsymbol{\mu})}{\|\underline{w}\|_X \|q\|_Q} < +\infty, \qquad \forall \boldsymbol{\mu} \in \mathcal{D};$$

4. the bilinear form $\mathcal{B}(\cdot, \cdot)$ satisfies the inf-sup condition over $X \times Q$, i.e. there exists a constant $\beta_0 > 0$ such that

$$\beta(\boldsymbol{\mu}) = \inf_{q \in Q} \sup_{\underline{w} \in X} \frac{\mathcal{B}(\underline{w}, q; \boldsymbol{\mu})}{\|\underline{w}\|_X \|q\|_Q} \geq \beta_0, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{4.1.7}$$

5. the bilinear form $\mathcal{A}(\cdot, \cdot; \boldsymbol{\mu})$ is symmetric and non-negative over X.

Moreover, thanks to the affine parameter dependence assumption (4.1.4), an affine decomposition holds also for the bilinear and linear forms in (4.1.6), i.e. for some finite Q_a, Q_b, Q_f, Q_g, they can be expressed as

$$\mathcal{A}(\underline{x}, \underline{w}; \boldsymbol{\mu}) = \sum_{q=1}^{Q_a} \Theta_a^q(\boldsymbol{\mu})\, \mathcal{A}^q(\underline{x}, \underline{w}), \qquad \mathcal{B}(\underline{w}, p; \boldsymbol{\mu}) = \sum_{q=1}^{Q_b} \Theta_b^q(\boldsymbol{\mu})\, \mathcal{B}^q(\underline{w}, p) \tag{4.1.8}$$

$$\langle G(\boldsymbol{\mu}), q \rangle = \sum_{q=1}^{Q_g} \Theta_g^q(\boldsymbol{\mu})\, \langle G^q, q \rangle, \qquad \langle \underline{F}(\boldsymbol{\mu}), \underline{w} \rangle = \sum_{q=1}^{Q_f} \Theta_f^q(\boldsymbol{\mu})\, \langle \underline{F}^q, \underline{w} \rangle, \tag{4.1.9}$$

where the coefficients $\Theta^q(\boldsymbol{\mu})$ and the $\boldsymbol{\mu}$-independent linear and bilinear forms are related to those appearing in (4.1.4). For example, $Q_a = \tilde{Q}_m + \tilde{Q}_n$, $\Theta_a^q(\boldsymbol{\mu}) = \tilde{\Theta}_m^q(\boldsymbol{\mu})$ and $\mathcal{A}^q(\underline{x}, \underline{w}) = m^q(y, z)$ for $1 \leq q \leq \tilde{Q}_m$, while $\Theta_a^{q+\tilde{Q}_m}(\boldsymbol{\mu}) = \tilde{\Theta}_n^q(\boldsymbol{\mu})$ and $\mathcal{A}^{q+\tilde{Q}_m}(\underline{x}, \underline{w}) = n^q(u, v)$ for $1 \leq q \leq \tilde{Q}_n$.

4.1.1 Truth approximation

Let $\{\mathcal{T}_\mathcal{N}\}$ be a triangulation of the domain Ω, we denote

$$V^r_\mathcal{N} = \left\{\psi_\mathcal{N} \in \mathrm{C}^0(\overline{\Omega}) : \psi_\mathcal{N}|_K \in \mathbb{P}_r,\ \forall K \in \mathcal{T}_\mathcal{N}\right\}$$

the space of globally continuous function that are polynomials of degree r on the single elements of the triangulation. Moreover we define $Y^\mathcal{N} = Y \cap V^r_\mathcal{N}$, $Q^\mathcal{N} = Y^\mathcal{N}$ and $U_\mathcal{N} = U \cap V^r_\mathcal{N}$ in such a way that $Y^\mathcal{N} \subset Y$, $U^\mathcal{N} \subset U$, $X^\mathcal{N} = Y^\mathcal{N} \times U^\mathcal{N} \subset X$, $Q^\mathcal{N} \subset Q$ are sequences of FE approximation spaces; we indicate with $\mathcal{N}$ the global dimension of the product space $X^\mathcal{N} \times Q^\mathcal{N}$, i.e. $\mathcal{N} = \mathcal{N}_X + \mathcal{N}_Q$ where $\mathcal{N}_X = \mathcal{N}_Y + \mathcal{N}_U$ and $\mathcal{N}_Y = \mathcal{N}_Q$. The *truth* Galerkin-FE approximation reads: given $\boldsymbol{\mu} \in \mathcal{D}$, find $(\underline{x}^\mathcal{N}(\boldsymbol{\mu}), p^\mathcal{N}(\boldsymbol{\mu})) \in X^\mathcal{N} \times Q^\mathcal{N}$ such that

$$\begin{cases} \mathcal{A}(\underline{x}^\mathcal{N}(\boldsymbol{\mu}), \underline{w}; \boldsymbol{\mu}) + \mathcal{B}(\underline{w}, p^\mathcal{N}(\boldsymbol{\mu}); \boldsymbol{\mu}) = \langle \underline{F}(\boldsymbol{\mu}), \underline{w} \rangle & \forall \underline{w} \in X^\mathcal{N}, \\ \mathcal{B}(\underline{x}^\mathcal{N}(\boldsymbol{\mu}), q; \boldsymbol{\mu}) = \langle G(\boldsymbol{\mu}), q \rangle & \forall q \in Q^\mathcal{N}. \end{cases} \tag{4.1.10}$$

As already proved in Lemma 1.2, provided $Y^\mathcal{N} \equiv Q^\mathcal{N}$, the bilinear form $\mathcal{A}(\cdot,\cdot;\boldsymbol{\mu})$ remains continuous over $X^\mathcal{N} \times X^\mathcal{N}$ and coercive over $X^\mathcal{N}_0 = \{\underline{w} \in X^\mathcal{N} : \mathcal{B}(\underline{w}, q; \boldsymbol{\mu}) = 0 \quad \forall q \in Q^\mathcal{N}\}$, i.e.

$$\gamma^\mathcal{N}_a(\boldsymbol{\mu}) = \sup_{\underline{x} \in X^\mathcal{N}} \sup_{\underline{w} \in X^\mathcal{N}} \frac{\mathcal{A}(\underline{x}, \underline{w}; \boldsymbol{\mu})}{\|\underline{x}\|_X \|\underline{w}\|_X} \leq \gamma_a(\boldsymbol{\mu}) < +\infty, \qquad \forall \boldsymbol{\mu} \in \mathcal{D},$$

$$\alpha^\mathcal{N}(\boldsymbol{\mu}) = \inf_{\underline{x} \in X^\mathcal{N}_0} \frac{\mathcal{A}(\underline{x}, \underline{x}; \boldsymbol{\mu})}{\|\underline{x}\|^2_X} \geq \alpha(\boldsymbol{\mu}) \geq \alpha_0, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}.$$

Similarly, the bilinear form $\mathcal{B}(\cdot,\cdot;\boldsymbol{\mu})$ remains continuous

$$\gamma^\mathcal{N}_b(\boldsymbol{\mu}) = \sup_{q \in Q^\mathcal{N}} \sup_{\underline{w} \in X^\mathcal{N}} \frac{\mathcal{B}(\underline{w}, q; \boldsymbol{\mu})}{\|\underline{w}\|_X \|q\|_Q} \leq \gamma_b(\boldsymbol{\mu}) < +\infty, \qquad \forall \boldsymbol{\mu} \in \mathcal{D},$$

and inf-sup stable over $X^\mathcal{N} \times Q^\mathcal{N}$, i.e. there exists a constant $\beta_0 > 0$ such that

$$\beta^\mathcal{N}(\boldsymbol{\mu}) = \inf_{q \in Q^\mathcal{N}} \sup_{\underline{w} \in X^\mathcal{N}} \frac{\mathcal{B}(\underline{w}, q; \boldsymbol{\mu})}{\|\underline{w}\|_X \|q\|_Q} \geq \beta_0, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}. \tag{4.1.11}$$

In particular we recall that we proved in Lemma 1.2 the estimate $\beta^\mathcal{N}(\boldsymbol{\mu}) \geq \tilde{\alpha}^\mathcal{N}(\boldsymbol{\mu})$, being $\tilde{\alpha}^\mathcal{N}(\boldsymbol{\mu})$ the coercivity constant of the bilinear form $a(\cdot,\cdot;\boldsymbol{\mu})$. Therefore the FE approximation (4.1.10) is well-posed, see Proposition 1.3.

At the algebraic level we obtain the linear system already introduced in Chapter 1 and widely discussed in Chapter 2, i.e.

$$\underbrace{\begin{pmatrix} A(\boldsymbol{\mu}) & B^T(\boldsymbol{\mu}) \\ B(\boldsymbol{\mu}) & 0 \end{pmatrix}}_{\mathcal{K}(\boldsymbol{\mu})} \begin{pmatrix} \mathbf{x}^\mathcal{N}(\boldsymbol{\mu}) \\ \mathbf{p}^\mathcal{N}(\boldsymbol{\mu}) \end{pmatrix} = \begin{pmatrix} \mathbf{F}(\boldsymbol{\mu}) \\ \mathbf{G}(\boldsymbol{\mu}) \end{pmatrix}.$$

Note that we have the same affine decompositions (4.1.8) and (4.1.9) for the matrices A, B,

$$A(\boldsymbol{\mu}) = \sum_{q=1}^{Q_a} \Theta^q_a(\boldsymbol{\mu})\, A^q, \qquad B(\boldsymbol{\mu}) = \sum_{q=1}^{Q_b} \Theta^q_b(\boldsymbol{\mu})\, B^q,$$

and for the right-hand sides

$$\mathbf{F}(\boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \Theta_f^q(\boldsymbol{\mu})\,\mathbf{F}^q, \qquad \mathbf{G}(\boldsymbol{\mu}) = \sum_{q=1}^{Q_g} \Theta_g^q(\boldsymbol{\mu})\,\mathbf{G}^q,$$

where the matrices and vectors A^q, B^q, $\mathbf{F}^q$, $\mathbf{G}^q$ represent the discrete counterparts of the corresponding bilinear and linear forms.

4.2 The reduced basis approximation

As in the case of parametrized elliptic equations discussed in Chapter 3, the idea of the RB method is to efficiently compute an approximation of $(\underline{x}^{\mathcal{N}}(\boldsymbol{\mu}), p^{\mathcal{N}}(\boldsymbol{\mu})$ by using approximation spaces made up of well-chosen solutions of (4.1.10), i.e. corresponding to specific choices of the parameter values. As already mentioned in the introduction to the chapter, the main assumption is that the solution of (4.1.10) depends *smoothly* on the parameters, thus implying the parametric manifold $\mathcal{M}^{\mathcal{N}}$ to be smooth and approximable by selecting some *snapshot* FE solutions.

4.2.1 Formulation

Let us take, for given $N \in \{1, \dots, N_{\max}\}$, a set of parameter values $S_N = \{\boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^N\}$ and consider the corresponding FE solutions $\{(\underline{x}^{\mathcal{N}}(\boldsymbol{\mu}^n), p^{\mathcal{N}}(\boldsymbol{\mu}^n)),\ n = 1, \dots, N\}$. We (naively) define the reduced basis spaces for the state, control and adjoint variables respectively as

$$\begin{aligned} Y_N &= \mathrm{span}\{\zeta_n := y^{\mathcal{N}}(\boldsymbol{\mu}^n), \quad n = 1, \dots, N\}, \\ U_N &= \mathrm{span}\{\lambda_n := u^{\mathcal{N}}(\boldsymbol{\mu}^n), \quad n = 1, \dots, N\}, \\ Q_N &= \mathrm{span}\{\xi_n := p^{\mathcal{N}}(\boldsymbol{\mu}^n), \quad n = 1, \dots, N\}, \end{aligned}$$

and denote $X_N = Y_N \times U_N$. By using Galerkin projection onto $X_N \times Q_N$, we obtain the following reduced basis approximation: find $(\underline{x}_N(\boldsymbol{\mu}), p_N(\boldsymbol{\mu})) \in X_N \times Q_N$ such that

$$\begin{cases} \mathcal{A}(\underline{x}_N(\boldsymbol{\mu}), \underline{w}; \boldsymbol{\mu}) + \mathcal{B}(\underline{w}, p_N(\boldsymbol{\mu}); \boldsymbol{\mu}) = \langle \underline{F}(\boldsymbol{\mu}), \underline{w} \rangle & \forall \underline{w} \in X_N, \\ \mathcal{B}(\underline{x}_N(\boldsymbol{\mu}), q; \boldsymbol{\mu}) = \langle G(\boldsymbol{\mu}), q \rangle & \forall q \in Q_N. \end{cases} \tag{4.2.1}$$

Let us discuss the well-posedness of the RB approximation. While the continuity property of the bilinear forms over the RB spaces are automatically inherited from the parents spaces (i.e. the FE spaces), the coercivity property of the bilinear form $\mathcal{A}(\cdot, \cdot; \boldsymbol{\mu})$ over

$$X_0^N = \{\underline{w} \in X_N : \mathcal{B}(\underline{w}, q; \boldsymbol{\mu}) = 0 \quad \forall q \in Q_N\}$$

and the fulfillment of the inf-sup condition of $\mathcal{B}(\cdot, \cdot; \boldsymbol{\mu})$ should be proved. In particular, the problem (4.2.1) should satisfy the following reduced basis inf-sup condition: there exists $\beta_0 > 0$ such that

$$\beta_N(\boldsymbol{\mu}) = \inf_{q \in Q_N} \sup_{\underline{w} \in X_N} \frac{\mathcal{B}(\underline{w}, q; \boldsymbol{\mu})}{\|\underline{w}\|_X \|q\|_Q} \geq \beta_0, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}. \tag{4.2.2}$$

The first idea in order to prove the fulfillment of (4.2.2) is to mimic the proof already used for the continuous problem and its FE approximation, see Lemma 1.1 and 1.2. We try to repeat the same (simple) arguments: exploiting the definition of $\mathcal{B}(\cdot,\cdot;\boldsymbol{\mu})$ we can write

$$\sup_{0\neq \underline{w}\in X_N} \frac{\mathcal{B}(\underline{w},q;\boldsymbol{\mu})}{\|\underline{w}\|_X} = \sup_{0\neq (z,v)\in Y_N\times U_N} \frac{a(z,q;\boldsymbol{\mu}) - c(v,q;\boldsymbol{\mu})}{(\|z\|_Y^2 + \|v\|_U^2)^{1/2}}$$

now we would like to choose $(z,v) = (q,0)$ to obtain

$$\sup_{0\neq (z,v)\in Y_N\times U_N} \frac{a(z,q;\boldsymbol{\mu}) - c(v,q;\boldsymbol{\mu})}{(\|z\|_Y^2 + \|v\|_U^2)^{1/2}} \underset{(z,v)=(q,0)}{\geq} \frac{a(q,q;\boldsymbol{\mu})}{\|q\|_Y},$$

and exploit the coercivity of the bilinear form $a(\cdot,\cdot;\boldsymbol{\mu})$. However, since $z \in Y_N$, $q \in Q_N$ and $Y_N \neq Q_N$, we cannot use this trick. Hence, while in the continuous case (respectively for the FE approximation) the state and adjoint spaces Y and Q (respectively $Y^{\mathcal{N}}$ and $Q^{\mathcal{N}}$) are equivalent, with the choice made above we lose this property on the corresponding RB spaces. Note that this is a particular feature of RB the approximation, since it uses basis functions specific for the problem instead of generic basis functions like the FE approximation.

It is clear that to recover the stability of the RB approximation we need to enrich in some way at least one of the RB spaces involved. This is not surprising when dealing with the RB approximation of a saddle-point problem, in fact there are at least other two examples where a similar treatment demonstrates to be mandatory: the application of the RB method to the Stokes problem (as already discussed in Section 3.5, see also [73, 79, 76, 75]) and to parametrized variational inequalities [36]. The two main strategies that we have analysed to reach the stability of the approximation are either the use of a suitable supremizer operator or the use of the same (properly defined) space for the state and adjoint variables. While the first option can be seen as a trial to mimic what has been done in the case of the Stokes problem, the second option follows naturally from the discussion above. We have made some preliminary investigations (both from the theoretical and the numerical point of view) comparing the two strategies, and we choose to pursue the second one, in this way we can avoid to introduce (less standard) supremizer operators. Moreover this choice has some advantages in the context of a posteriori error estimation for the cost functional, as already pointed out in [17]. In any case we are aware that these issues deserve further investigations in order to explore also other strategies, maybe even more convenient on the computational point of view.

However, before discussing the enrichment strategy, it is useful to analyze the algebraic structure of the RB approximation.

4.2.2 Algebraic formulation

Let us investigate the structure of the algebraic system associated with the RB approximation (4.2.1). We can express the RB state, adjoint and control solutions as

$$y_N(\boldsymbol{\mu}) = \sum_{j=1}^{N} y_{Nj}(\boldsymbol{\mu})\zeta_j, \qquad u_N(\boldsymbol{\mu}) = \sum_{j=1}^{N} u_{Nj}(\boldsymbol{\mu})\lambda_j, \qquad p_N(\boldsymbol{\mu}) = \sum_{j=1}^{N} p_{Nj}(\boldsymbol{\mu})\xi_j,$$

we also define a reduced basis for the product space $X_N = \text{span}\{\underline{\sigma}_j,\ j = 1,\dots,2N\}$ where

$$\underline{\sigma}_j = \begin{cases} (\zeta_j, 0), & j = 1,\dots,N \\ (0,\lambda_j), & j = N+1,\dots,2N, \end{cases}$$

note that

$$\underline{x}_N(\boldsymbol{\mu}) = (y_N(\boldsymbol{\mu}), u_N(\boldsymbol{\mu})) = \left(\sum_{j=1}^{N} y_{Nj}(\boldsymbol{\mu})\zeta_j, \sum_{j=1}^{N} u_{Nj}(\boldsymbol{\mu})\lambda_j \right) = \sum_{j=1}^{2N} x_{Nj}(\boldsymbol{\mu})\underline{\sigma}_j.$$

Hence, for a new parameter $\boldsymbol{\mu}$, the RB solution of the problem (4.2.1) can be written as a combination of basis functions with weights given by the following reduced basis linear system

$$\left\{ \begin{aligned} &\sum_{j=1}^{2N}\sum_{q=1}^{Q_a} \Theta_a^q(\boldsymbol{\mu}) A_{ij}^q \, x_{Nj}(\boldsymbol{\mu}) + \sum_{l=1}^{N}\sum_{q=1}^{Q_b} \Theta_b^q(\boldsymbol{\mu}) B_{li}^q \, p_{Nl}(\boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \Theta_f^q(\boldsymbol{\mu}) F_i^q, && 1 \le i \le 2N, \\ &\sum_{j=1}^{2N}\sum_{q=1}^{Q_b} \Theta_b^q(\boldsymbol{\mu}) B_{lj}^q \, x_{Nj}(\boldsymbol{\mu}) = \sum_{q=1}^{Q_g} \Theta_g^q(\boldsymbol{\mu}) G_l^q, && 1 \le l \le N, \end{aligned} \right. \tag{4.2.3}$$

where the submatrices A^q and B^q are given by

$$A_{ij}^q = \mathcal{A}^q(\underline{\sigma}_j, \underline{\sigma}_i), \qquad B_{li}^q = \mathcal{B}^q(\underline{\sigma}_i, \xi_l), \qquad 1 \le i, j \le 2N, \quad 1 \le l \le N,$$

and the vectors F^q, G^q by

$$F_i^q = \langle F^q, \underline{\sigma}_i \rangle, \qquad G_l^q = \langle G^q, \xi_l \rangle \qquad 1 \le i \le 2N, \quad 1 \le l \le N.$$

Finally, denoting with $A_N(\boldsymbol{\mu}) = \sum \Theta_a^q A^q$, $B_N(\boldsymbol{\mu}) = \sum \Theta_b^q B^q$, we can rewrite problem (4.2.3) as

$$\underbrace{\begin{pmatrix} A_N(\boldsymbol{\mu}) & B_N^T(\boldsymbol{\mu}) \\ B_N(\boldsymbol{\mu}) & 0 \end{pmatrix}}_{\mathcal{K}_N(\boldsymbol{\mu})} \begin{pmatrix} \mathbf{x}_N(\boldsymbol{\mu}) \\ \mathbf{p}_N(\boldsymbol{\mu}) \end{pmatrix} = \begin{pmatrix} \mathbf{F}_N(\boldsymbol{\mu}) \\ \mathbf{G}_N(\boldsymbol{\mu}) \end{pmatrix}, \tag{4.2.4}$$

where $\mathbf{x}_N$ and $\mathbf{p}_N$ are the column vectors of the linear combination coefficient $\{x_{Nj}\}_{j=1}^{2N}$ and $\{p_{Nl}\}_{l=1}^{N}$ respectively. In order to analyze more in-depth the structure of the linear system with saddle-point structure (4.2.4) let us define the *basis matrices*

$$Z_y = \begin{pmatrix} \boldsymbol{\zeta}_1 & \cdots & \boldsymbol{\zeta}_N \end{pmatrix}, \qquad Z_p = \begin{pmatrix} \boldsymbol{\xi}_1 & \cdots & \boldsymbol{\xi}_N \end{pmatrix}, \qquad Z_u = \begin{pmatrix} \boldsymbol{\lambda}_1 & \cdots & \boldsymbol{\lambda}_N \end{pmatrix},$$

$$Z_x = \begin{pmatrix} Z_y & 0 \\ 0 & Z_u \end{pmatrix}, \qquad Z = \begin{pmatrix} Z_y & 0 & 0 \\ 0 & Z_u & 0 \\ 0 & 0 & Z_p \end{pmatrix},$$

where $Z_y, Z_u, Z_p \in \mathbb{R}^{\mathcal{N} \times N}$, $Z_x \in \mathbb{R}^{2\mathcal{N} \times 2N}$ and $Z \in \mathbb{R}^{3\mathcal{N} \times 3N}$. Then $\mathcal{K}_N = Z^T \mathcal{K} Z$ is given by

$$\mathcal{K}_N = \begin{pmatrix} A_N & B_N^T \\ B_N & 0 \end{pmatrix} = \begin{pmatrix} Z_x^T A Z_x & Z_x^T B^T Z_p \\ Z_p^T B Z_x & 0 \end{pmatrix}, \tag{4.2.5}$$

highlighting the structure of each block

$$\mathcal{K}_N = \begin{pmatrix} M_N & 0 & K_N^T \\ 0 & \alpha N_N & -C_N \\ K_N & -C_N & 0 \end{pmatrix} = \begin{pmatrix} Z_y^T M Z_y & 0 & Z_y^T K^T Z_p \\ 0 & \alpha Z_u^T N Z_u & -Z_u^T C Z_p \\ Z_p^T K Z_y & -Z_p^T C Z_u & 0 \end{pmatrix}, \tag{4.2.6}$$

where the matrices K, C, M, N are those induced by the corresponding bilinear forms $a(\cdot, \cdot)$, $c(\cdot, \cdot)$, $m(\cdot, \cdot)$ and $n(\cdot, \cdot)$ respectively.

Remark 4.1. The RB matrix $\mathcal{K}_N$ features the same symmetry and saddle-point structure of the matrix arising from the finite element approximation.

Remark 4.2. Let us write separately the state, adjoint and optimality equations

$$\begin{aligned} \text{state:} \quad & (Z_p^T K Z_y)\mathbf{y}_N && - (Z_p^T C Z_u)\mathbf{u}_N && = Z_p^T \mathbf{F}_s, \\ \text{adjoint:} \quad & (Z_y^T K^T Z_y)\mathbf{p}_N && + (Z_y^T M Z_y)\mathbf{y}_N && = Z_y^T \mathbf{F}_a, \\ \text{optimality:} \quad & \alpha(Z_u^T N Z_u)\mathbf{u}_N && - (Z_u^T C Z_p)\mathbf{p}_N && = 0. \end{aligned} \tag{4.2.7}$$

Looking at the state equation it is clear that with this choice of the RB spaces we are using different trial and test spaces: in particular we are testing the state equation using the adjoint basis. The same arguments hold for the adjoint equation, where the test functions belong to the state space. Different spaces for trial and test functions means that we are approximating the state and adjoint equations with a Petrov-Galerkin scheme, obviously this property can be seen directly on the formulation (4.2.1) exploiting the definition of the bilinear forms.

4.2.3 Approximation stability: enriched spaces

To ensure the stability of the RB approximation we define the following *aggregated* space for the state and adjoint variables

$$Z_N = \text{span}\{\zeta_n := y^{\mathcal{N}}(\boldsymbol{\mu}^n),\ \xi_n := p^{\mathcal{N}}(\boldsymbol{\mu}), \quad n = 1, \dots, N\}.$$

Now let $Y_N = Z_N$, $X_N = Z_N \times U_N$ and $Q_N = Z_N$, the RB approximation reads: find $(\underline{x}_N(\boldsymbol{\mu}), p_N(\boldsymbol{\mu})) \in X_N \times Q_N$ such that

$$\begin{cases} \mathcal{A}(\underline{x}_N(\boldsymbol{\mu}), \underline{w}; \boldsymbol{\mu}) + \mathcal{B}(\underline{w}, p_N(\boldsymbol{\mu}); \boldsymbol{\mu}) = \langle \underline{F}(\boldsymbol{\mu}), \underline{w} \rangle & \forall \underline{w} \in X_N, \\ \mathcal{B}(\underline{x}_N(\boldsymbol{\mu}), q; \boldsymbol{\mu}) = \langle G(\boldsymbol{\mu}), q \rangle & \forall q \in Q_N. \end{cases} \tag{4.2.8}$$

We can now easily prove the fulfillment of the required inf-sup condition.

Lemma 4.1. *The bilinear form $\mathcal{B}(\cdot, \cdot; \boldsymbol{\mu})$ satisfies the inf-sup condition (4.2.2). Moreover we have the estimate*

$$\beta_N(\boldsymbol{\mu}) \geq \tilde{\alpha}^{\mathcal{N}}(\boldsymbol{\mu}), \qquad \forall \boldsymbol{\mu} \in \mathcal{D},$$

where $\tilde{\alpha}^{\mathcal{N}}(\boldsymbol{\mu})$ is the coercivity constant associated to the FE approximation of the bilinear form $a(\cdot, \cdot; \boldsymbol{\mu})$.

Proof. It suffices to mimic the proof of Lemma 1.2. In fact,

$$\begin{aligned} \beta_N(\boldsymbol{\mu}) = \inf_{q \in Q_N} \sup_{\underline{w} \in X_N} \frac{\mathcal{B}(\underline{w}, q; \boldsymbol{\mu})}{\|\underline{w}\|_X \|q\|_Q} &= \inf_{q \in Z_N} \sup_{(z,v) \in Z_N \times U_N} \frac{a(z, q; \boldsymbol{\mu}) - c(v, q; \boldsymbol{\mu})}{\|(z, v)\|_X \|q\|_Q} \\ \underset{(z,v)=(q,0)}{\geq} \inf_{q \in Z_N} \frac{a(q, q; \boldsymbol{\mu})}{\|q\|_Q} &= \tilde{\alpha}_N(\boldsymbol{\mu}) \geq \tilde{\alpha}^{\mathcal{N}}(\boldsymbol{\mu}) > 0. \end{aligned}$$

Note that now the choice $z = q$ is allowed because both z and q belong to the space Z_N. □

Proposition 4.1. *The RB saddle-point problem (4.2.8) has a unique solution $(\underline{x}_N(\boldsymbol{\mu}), p_N(\boldsymbol{\mu})) \in X_N \times Q_N$ for all $\boldsymbol{\mu} \in \mathcal{D}$.*

Proof. It suffices to check that the assumptions of Theorem A.4 hold. As already mentioned, the continuity properties of the bilinear and linear forms over the RB space are automatically inherited from the parents spaces (i.e. the FE spaces). The fulfillment of the inf-sup condition of the bilinear form $\mathcal{B}(\cdot, \cdot; \boldsymbol{\mu})$ has been proved in Lemma 4.1, while the fulfillment of the coercivity condition of the bilinear form $\mathcal{A}(\cdot, \cdot; \boldsymbol{\mu})$ can be proved using the same arguments as in Lemma 1.1 and Lemma 1.2. □

4.2.4 Algebraic formulation with enriched spaces

Let us now investigate the algebraic formulation associated to the enriched spaces introduced in the previous section. Let $\{\tau_j\}_{j=1}^{2N} = \{\zeta_j\}_{j=1}^{N} \cup \{\xi_j\}_{j=1}^{N}$ such that $Z_N = \text{span}\{\tau_j,\, j = 1,\dots,2N\}$, we can express the RB state, adjoint and control solutions as

$$y_N(\boldsymbol{\mu}) = \sum_{j=1}^{2N} y_{Nj}(\boldsymbol{\mu})\tau_j, \qquad u_N(\boldsymbol{\mu}) = \sum_{j=1}^{N} u_{Nj}(\boldsymbol{\mu})\lambda_j, \qquad p_N(\boldsymbol{\mu}) = \sum_{j=1}^{2N} p_{Nj}(\boldsymbol{\mu})\tau_j,$$

we define in the usual way a reduced basis for the product space $X_N = \text{span}\{\underline{\sigma}_j,\, j = 1,\dots,3N\}$ where

$$\underline{\sigma}_j = \begin{cases} (\tau_j, 0), & j = 1,\dots,2N \\ (0, \lambda_j), & j = 2N+1,\dots,3N, \end{cases}$$

note that

$$\underline{x}_N(\boldsymbol{\mu}) = (y_N(\boldsymbol{\mu}), u_N(\boldsymbol{\mu})) = \left(\sum_{j=1}^{2N} y_{Nj}(\boldsymbol{\mu})\zeta_j, \sum_{j=1}^{N} u_{Nj}(\boldsymbol{\mu})\lambda_j\right) = \sum_{j=1}^{3N} x_{Nj}(\boldsymbol{\mu})\underline{\sigma}_j.$$

Hence, for a new parameter $\boldsymbol{\mu}$, the RB solution of the problem (4.2.8) can be written as a combination of basis functions with weights given by the following reduced basis linear system:

$$\begin{cases} \displaystyle\sum_{j=1}^{3N}\sum_{q=1}^{Q_a} \Theta_a^q(\boldsymbol{\mu}) A_{ij}^q\, x_{Nj}(\boldsymbol{\mu}) + \sum_{l=1}^{2N}\sum_{q=1}^{Q_b} \Theta_b^q(\boldsymbol{\mu}) B_{li}^q\, p_{Nl}(\boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \Theta_f^q(\boldsymbol{\mu}) F_i^q, & 1 \le i \le 3N, \\ \displaystyle\sum_{j=1}^{3N}\sum_{q=1}^{Q_b} \Theta_b^q(\boldsymbol{\mu}) B_{lj}^q\, x_{Nj}(\boldsymbol{\mu}) = \sum_{q=1}^{Q_g} \Theta_g^q(\boldsymbol{\mu}) G_l^q, & 1 \le l \le 2N, \end{cases} \tag{4.2.9}$$

where the submatrices A_N^q and B_N^q (we have omitted the subscript $_N$ in (4.2.9)) are given by

$$(A_N)_{ij}^q = \mathcal{A}^q(\underline{\sigma}_j, \underline{\sigma}_i), \qquad (B_N)_{li}^q = \mathcal{B}^q(\underline{\sigma}_i, \tau_l), \qquad 1 \le i,j \le 3N, \quad 1 \le l \le 2N,$$

and

$$(F_N)_i^q = \langle F^q, \underline{\sigma}_i \rangle, \qquad (G_N)_l^q = \langle G^q, \tau_l \rangle \qquad 1 \le i \le 3N, \quad 1 \le l \le 2N.$$

Finally, denoting with $A_N(\boldsymbol{\mu}) = \sum \Theta_a^q(\boldsymbol{\mu}) A_N^q$, $B_N(\boldsymbol{\mu}) = \sum \Theta_b^q(\boldsymbol{\mu}) B_N^q$, we can rewrite problem (4.2.9) as

$$\underbrace{\begin{pmatrix} A_N(\boldsymbol{\mu}) & B_N^T(\boldsymbol{\mu}) \\ B_N(\boldsymbol{\mu}) & 0 \end{pmatrix}}_{\mathcal{K}_N(\boldsymbol{\mu})} \begin{pmatrix} \mathbf{x}_N(\boldsymbol{\mu}) \\ \mathbf{p}_N(\boldsymbol{\mu}) \end{pmatrix} = \begin{pmatrix} \mathbf{F}_N(\boldsymbol{\mu}) \\ \mathbf{G}_N(\boldsymbol{\mu}) \end{pmatrix}. \tag{4.2.10}$$

Let us define the *basis matrices* $Z_z = \begin{pmatrix} \boldsymbol{\tau}_1 & \cdots & \boldsymbol{\tau}_N \end{pmatrix} \in \mathbb{R}^{\mathcal{N}\times 2N}$, $Z_u = \begin{pmatrix} \boldsymbol{\lambda}_1 & \cdots & \boldsymbol{\lambda}_N \end{pmatrix} \in \mathbb{R}^{\mathcal{N}\times N}$ and

$$Z_x = \begin{pmatrix} Z_z & 0 \\ 0 & Z_u \end{pmatrix} \in \mathbb{R}^{2\mathcal{N}\times 3N}, \qquad Z = \begin{pmatrix} Z_z & 0 & 0 \\ 0 & Z_u & 0 \\ 0 & 0 & Z_z \end{pmatrix} \in \mathbb{R}^{3\mathcal{N}\times 5N},$$

then $\mathcal{K}_N = Z^T \mathcal{K} Z$ is given by

$$\mathcal{K}_N = \begin{pmatrix} A_N & B_N^T \\ B_N & 0 \end{pmatrix} = \begin{pmatrix} Z_x^T A Z_x & Z_x^T B^T Z_z \\ Z_z^T B Z_x & 0 \end{pmatrix}, \tag{4.2.11}$$

highlighting the structure of each block

$$\mathcal{K}_N = \begin{pmatrix} M_N & 0 & K_N^T \\ 0 & \alpha N_N & -C_N \\ K_N & -C_N & 0 \end{pmatrix} = \begin{pmatrix} Z_z^T M Z_z & 0 & Z_z^T K^T Z_z \\ 0 & \alpha Z_u^T N Z_u & -Z_u^T C Z_z \\ Z_z^T K Z_z & -Z_z^T C Z_u & 0 \end{pmatrix}. \tag{4.2.12}$$

Thus the matrix $\mathcal{K}_N$ is still symmetric, with saddle-point structure and has dimension $5N \times 5N$. To keep under control the condition number of the matrix $\mathcal{K}_N$ we have adopted the Gram-Schmidt (GS) orthonormalization procedure already introduced in Chapter 3. In particular we apply the GS procedure separately on the basis functions of the space Z_N and on the basis functions of the space U_N.

4.2.5 Offline-Online procedure and sampling strategy

Thanks to the assumption of affine parameter dependence, we can decouple the formation of the matrix $\mathcal{K}_N(\boldsymbol{\mu})$ in two stages, the Offline and Online stages, that enable the efficient resolution of the system (4.2.10) for each new parameter $\boldsymbol{\mu}$. In particular:

1. in the Offline stage, performed only once, we first compute and store the basis function $\{\tau_i\}_{i=1}^{2N}$ and $\{\lambda_j\}_{j=1}^{N}$, and form the $\boldsymbol{\mu}$-independent matrices A_N^q, $1 \le q \le Q_a$, B_N^q, $1 \le q \le Q_b$ and the vectors F_N^q, $1 \le q \le Q_f$, G_N^q, $1 \le q \le Q_g$. The operation count depends on N, Q_a, Q_b, Q_f, Q_g and $\mathcal{N}$;

2. in the Online stage, performed for each new value $\boldsymbol{\mu}$, we use the precomputed matrices A_N^q, B_N^q and vectors F_N^q, G_N^q to assemble the (full) matrix $\mathcal{K}_N$ and the vectors $\mathbf{F}_N$, $\mathbf{G}_N$ appearing in (4.2.10), with

$$A_N(\boldsymbol{\mu}) = \sum_{q=1}^{Q_a} \Theta_a^q(\boldsymbol{\mu}) A_N^q, \qquad B_N(\boldsymbol{\mu}) = \sum_{q=1}^{Q_b} \Theta_b^q(\boldsymbol{\mu}) B_N^q,$$

$$\mathbf{F}_N(\boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \Theta_f^q(\boldsymbol{\mu}) F_N^q, \qquad \mathbf{G}_N(\boldsymbol{\mu}) = \sum_{q=1}^{Q_g} \Theta_f^q(\boldsymbol{\mu}) G_N^q;$$

we then solve the resulting system to obtain $(\mathbf{x}_N, \mathbf{p}_N)$. The Online operation count depends on N, Q_a, Q_b, Q_f, Q_g but is independent of $\mathcal{N}$. In particular we need $O((Q_a+Q_b)N^2)$ and $O((Q_f+Q_g)N)$ operations to assemble matrices and vectors, and $O((5N)^3)$ operations to solve the RB linear system (4.2.10).

For the construction of the hierarchical Lagrange RB approximation spaces we rely again on the sampling strategy based on the greedy algorithm described in Chapter 3. In particular, in each iteration, given the parameter samples $S_N = \{\boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^N\}$, the new sample point $\boldsymbol{\mu}^{N+1}$ to be added is such that

$$\boldsymbol{\mu}^{N+1} = \arg \max_{\boldsymbol{\mu} \in \Xi_{\text{train}}} \Delta_N(\boldsymbol{\mu}),$$

where $\Delta_N(\boldsymbol{\mu})$ is a rigorous, sharp and inexpensive a posteriori error bound for the error on the state, control and adjoint variables, i.e.

$$\left(\|\underline{x}^{\mathcal{N}}(\boldsymbol{\mu}) - \underline{x}_N(\boldsymbol{\mu})\|_X^2 + \|p^{\mathcal{N}}(\boldsymbol{\mu}) - p_N(\boldsymbol{\mu})\|_Q^2\right)^{1/2} \leq \Delta_N(\boldsymbol{\mu}), \tag{4.2.13}$$

where $\|\underline{x}^{\mathcal{N}}(\boldsymbol{\mu}) - \underline{x}_N(\boldsymbol{\mu})\|_X^2 = \|y^{\mathcal{N}}(\boldsymbol{\mu}) - y_N(\boldsymbol{\mu})\|_Y^2 + \|u^{\mathcal{N}}(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_U^2$. The next section is devoted to the construction of such an error estimator.

4.3 A posteriori error estimation

In Section 4.3.1 we construct a rigorous, sharp and inexpensive (i.e. $\mathcal{N}$-independent) a posteriori error bound $\Delta_N(\boldsymbol{\mu})$ such that

$$\left(\|\underline{x}^{\mathcal{N}}(\boldsymbol{\mu}) - \underline{x}_N(\boldsymbol{\mu})\|_X^2 + \|p^{\mathcal{N}}(\boldsymbol{\mu}) - p_N(\boldsymbol{\mu})\|_Q^2\right)^{1/2} \leq \Delta_N(\boldsymbol{\mu}). \tag{4.3.1}$$

In Section 4.3.2 we describe the Offline-Online strategy that permits the efficient evaluation of the proposed estimator. Then in Section 4.3.3, using the same ingredients, we construct a rigorous, sharp and inexpensive a posteriori error bound $\Delta_N^J(\boldsymbol{\mu})$ for the error on the cost functional, i.e.

$$|J(y^{\mathcal{N}}(\boldsymbol{\mu}), u^{\mathcal{N}}(\boldsymbol{\mu}); \boldsymbol{\mu}) - J(y_N(\boldsymbol{\mu}), u_N(\boldsymbol{\mu}); \boldsymbol{\mu})| \leq \Delta_N^J(\boldsymbol{\mu}). \tag{4.3.2}$$

4.3.1 Bound for the solution: Babuška framework

The construction of the estimator $\Delta_N(\boldsymbol{\mu})$ will be carried out in the Babuška framework, as already done for the Stokes equations in Section 3.5.5. In fact, as observed in Section A.1, saddle-points problems are a particular case of weakly coercive problem, for which the stability analysis can be carried out by using the Nečas-Babuška theorem. Therefore, to construct the error estimator $\Delta_N(\boldsymbol{\mu})$ it is sufficient to exploit this alternative point of view and rewrite the problem as a weakly coercive problem, for which RB a posteriori error estimates techniques are already available.

In order to formulate the problem (4.1.6) in the standard form (P_2) on page 134, it suffices to denote $\mathcal{X} = X \times Q$ and define the bilinear form $\mathsf{B}(\cdot, \cdot; \boldsymbol{\mu}) \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ given by

$$\mathsf{B}(\mathsf{x}, \mathsf{w}; \boldsymbol{\mu}) := \mathcal{A}(\underline{x}, \underline{w}; \boldsymbol{\mu}) + \mathcal{B}(\underline{w}, p; \boldsymbol{\mu}) + \mathcal{B}(\underline{x}, q; \boldsymbol{\mu}), \tag{4.3.3}$$

and the linear continuous functional $\mathsf{F} \colon \mathcal{X} \to \mathbb{R}$

$$\mathsf{F}(\mathsf{w}; \boldsymbol{\mu}) = \langle \underline{F}(\boldsymbol{\mu}), \underline{w} \rangle + \langle G(\boldsymbol{\mu}), q \rangle, \tag{4.3.4}$$

where $\mathsf{x} = (\underline{x}, p) \in \mathcal{X}$ and $\mathsf{w} = (\underline{w}, q) \in \mathcal{X}$. Then, we can formulate equivalently the problem (4.1.6) as: given $\boldsymbol{\mu} \in \mathcal{D}$,

$$\text{find } \mathsf{x} \in \mathcal{X} \text{ s.t:} \qquad \mathsf{B}(\mathsf{x}, \mathsf{w}; \boldsymbol{\mu}) = \mathsf{F}(\mathsf{w}; \boldsymbol{\mu}) \qquad \forall \mathsf{w} \in \mathcal{X}. \tag{4.3.5}$$

The problem (4.3.5) is well posed if and only if the following conditions hold (see Section A.1):

1. the bilinear form $\mathsf{B}(\cdot, \cdot; \boldsymbol{\mu})$ is continous, i.e. there exists a constant $\gamma(\boldsymbol{\mu}) > 0$ such that

$$\mathsf{B}(\mathsf{x}, \mathsf{w}; \boldsymbol{\mu}) \leq \gamma(\boldsymbol{\mu}) \|\mathsf{x}\|_{\mathcal{X}} \|\mathsf{w}\|_{\mathcal{X}}, \qquad \forall \boldsymbol{\mu} \in \mathcal{D};$$

2. the bilinear form $\mathsf{B}(\cdot,\cdot;\boldsymbol{\mu})$ is weakly coercive, i.e. there exists a constant $\hat{\beta}_0 > 0$ such that
$$\hat{\beta}(\boldsymbol{\mu}) = \inf_{\mathsf{w}\in\mathcal{X}} \sup_{\mathsf{x}\in\mathcal{X}} \frac{\mathsf{B}(\mathsf{x},\mathsf{w};\boldsymbol{\mu})}{\|\mathsf{x}\|_{\mathcal{X}}\|\mathsf{w}\|_{\mathcal{X}}} \geq \hat{\beta}_0, \qquad \forall\boldsymbol{\mu}\in\mathcal{D}.$$

Moreover, for each $\boldsymbol{\mu}\in\mathcal{D}$, the unique solution satisfies
$$\|\mathsf{x}(\boldsymbol{\mu})\|_{\mathcal{X}} \leq \frac{1}{\hat{\beta}(\boldsymbol{\mu})}\|\mathsf{F}(\cdot;\boldsymbol{\mu})\|_{\mathcal{X}'}. \tag{4.3.6}$$

Actually, since the bilinear forms $\mathcal{A}(\cdot,\cdot;\boldsymbol{\mu})$ and $\mathcal{B}(\cdot,\cdot;\boldsymbol{\mu})$ satisfy the hypotheses of (Brezzi) Theorem A.2, it can be shown (see e.g. [21, 90, 35]) that the the compound form $\mathsf{B}(\cdot,\cdot;\boldsymbol{\mu})$ is continuous and weakly coercive. Similarly, the FE and RB approximations satisfy the same inf-sup condition,
$$\hat{\beta}^{\mathcal{N}}(\boldsymbol{\mu}) := \inf_{\mathsf{w}\in\mathcal{X}^{\mathcal{N}}} \sup_{\mathsf{x}\in\mathcal{X}^{\mathcal{N}}} \frac{\mathsf{B}(\mathsf{x},\mathsf{w};\boldsymbol{\mu})}{\|\mathsf{x}\|_{\mathcal{X}}\|\mathsf{w}\|_{\mathcal{X}}} > 0, \qquad \forall\boldsymbol{\mu}\in\mathcal{D},$$
$$\hat{\beta}_N(\boldsymbol{\mu}) := \inf_{\mathsf{w}\in\mathcal{X}_N} \sup_{\mathsf{x}\in\mathcal{X}_N} \frac{\mathsf{B}(\mathsf{x},\mathsf{w};\boldsymbol{\mu})}{\|\mathsf{x}\|_{\mathcal{X}}\|\mathsf{w}\|_{\mathcal{X}}} > 0, \qquad \forall\boldsymbol{\mu}\in\mathcal{D},$$

where $\mathcal{X}^{\mathcal{N}} = X^{\mathcal{N}} \times Q^{\mathcal{N}}$ and $\mathcal{X}_N = X_N \times Q_N$. Moreover the stability estimate (4.3.6) holds also for the FE and RB approximations, in particular
$$\|\mathsf{x}^{\mathcal{N}}(\boldsymbol{\mu})\|_{\mathcal{X}} \leq \frac{1}{\hat{\beta}^{\mathcal{N}}(\boldsymbol{\mu})}\|\mathsf{F}(\cdot;\boldsymbol{\mu})\|_{\mathcal{X}'}, \qquad \forall\boldsymbol{\mu}\in\mathcal{D}. \tag{4.3.7}$$

In the following we will refer to the inf-sup constant $\hat{\beta}(\boldsymbol{\mu})$ as the Babuška inf-sup constant (in contrast to the Brezzi inf-sup constant $\beta(\boldsymbol{\mu})$).

Once we have guaranteed the well-posedness of the optimal control problem in the form (4.3.5), for the construction of the a posteriori error estimator we need the usual two main ingredients: an effective calculation of a lower bound for the Babuška inf-sup constant $\hat{\beta}^{\mathcal{N}}(\boldsymbol{\mu})$ and the (standard) calculation of the dual norm of the residual. For the first one we assume that we can calculate a $\boldsymbol{\mu}$-dependent lower bound $\hat{\beta}_{\text{LB}}(\boldsymbol{\mu})$ for the inf-sup constant $\hat{\beta}^{\mathcal{N}}(\boldsymbol{\mu})$, i.e. $\hat{\beta}^{\mathcal{N}}(\boldsymbol{\mu}) \geq \hat{\beta}_{\text{LB}}(\boldsymbol{\mu}) \geq \hat{\beta}_0 > 0, \forall\boldsymbol{\mu}\in\mathcal{D}$. The calculation of $\hat{\beta}_{\text{LB}}(\boldsymbol{\mu})$ will be carried out using the *Natural Norm Successive Constraint Method*, as in the case of Stokes problem (see Section 3.5.5).

Let us firstly define the errors between the FE and the RB approximations:
$$e_y(\boldsymbol{\mu}) = y^{\mathcal{N}}(\boldsymbol{\mu}) - y_N(\boldsymbol{\mu}), \qquad e_u(\boldsymbol{\mu}) = u^{\mathcal{N}}(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu}),$$
$$e_p(\boldsymbol{\mu}) = p^{\mathcal{N}}(\boldsymbol{\mu}) - p_N(\boldsymbol{\mu}), \qquad \underline{e}_x(\boldsymbol{\mu}) = (e_y(\boldsymbol{\mu}), e_u(\boldsymbol{\mu})) = \underline{x}^{\mathcal{N}}(\boldsymbol{\mu}) - \underline{x}_N(\boldsymbol{\mu})$$

and the *global* error
$$\mathsf{e}(\boldsymbol{\mu}) = (\underline{e}_x(\boldsymbol{\mu}), e_p(\boldsymbol{\mu})) = \mathsf{x}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathsf{x}_N(\boldsymbol{\mu}). \tag{4.3.8}$$

Then we define the residuals
$$r_{\underline{x}}(\underline{w};\boldsymbol{\mu}) = \langle \underline{F}(\boldsymbol{\mu}), \underline{w}\rangle - \mathcal{A}(\underline{x}_N, \underline{w};\boldsymbol{\mu}) - \mathcal{B}(\underline{w}, p_N;\boldsymbol{\mu}) \qquad \forall\underline{w}\in X^{\mathcal{N}}$$
$$r_p(q;\boldsymbol{\mu}) = \langle G(\boldsymbol{\mu}), q\rangle - \mathcal{B}(\underline{x}_N, q;\boldsymbol{\mu}) \qquad \forall q\in Q^{\mathcal{N}},$$

and the *global* residual

$$\mathsf{r}(\mathsf{w};\boldsymbol{\mu}) = \mathsf{F}(\mathsf{w};\boldsymbol{\mu}) - \mathsf{B}(\mathsf{x}_N,\mathsf{w};\boldsymbol{\mu}) \equiv r_{\underline{x}}(\underline{w};\boldsymbol{\mu}) + r_p(q;\boldsymbol{\mu}) \qquad \forall \mathsf{w} \in \mathcal{X}^{\mathcal{N}};$$

note that $\mathsf{r}(\cdot;\boldsymbol{\mu}) \in (\mathcal{X}^{\mathcal{N}})'$. The problem statements for $(\underline{x}^{\mathcal{N}}, p^{\mathcal{N}})$ (4.1.10) and $(\underline{x}_N, p_N)$ (4.2.8) and the bilinearity of $\mathcal{A}(\cdot,\cdot;\boldsymbol{\mu})$ and $\mathcal{B}(\cdot,\cdot;\boldsymbol{\mu})$ imply that the errors satisfy the following equations

$$\begin{cases} \mathcal{A}(\underline{e}_x(\boldsymbol{\mu}),\underline{w};\boldsymbol{\mu}) + \mathcal{B}(\underline{w},e_p(\boldsymbol{\mu});\boldsymbol{\mu}) = r_{\underline{x}}(\underline{w},\boldsymbol{\mu}), & \forall \underline{w} \in X^{\mathcal{N}}, \\ \mathcal{B}(\underline{e}_x(\boldsymbol{\mu}),q;\boldsymbol{\mu}) = r_p(q;\boldsymbol{\mu}), & \forall q \in Q^{\mathcal{N}}. \end{cases} \tag{4.3.9}$$

Equivalently the global error $\mathsf{e}(\boldsymbol{\mu})$ satisfies

$$\mathsf{B}(\mathsf{e},\mathsf{w};\boldsymbol{\mu}) = \mathsf{r}(\mathsf{w};\boldsymbol{\mu}) \qquad \forall \mathsf{w} \in \mathcal{X}^{\mathcal{N}}.$$

By using the stability estimate (4.3.6) we obtain the following residual-based estimation

$$\|\mathsf{e}(\boldsymbol{\mu})\|_{\mathcal{X}} \le \frac{1}{\hat{\beta}^{\mathcal{N}}(\boldsymbol{\mu})} \|\mathsf{r}(\cdot;\boldsymbol{\mu})\|_{\mathcal{X}'}, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{4.3.10}$$

exploiting the lower bound for the inf-sup constant

$$\|\mathsf{e}(\boldsymbol{\mu})\|_{\mathcal{X}} \le \frac{1}{\hat{\beta}_{\mathrm{LB}}(\boldsymbol{\mu})} \|\mathsf{r}(\cdot;\boldsymbol{\mu})\|_{\mathcal{X}'} := \Delta_N(\boldsymbol{\mu}), \qquad \forall \boldsymbol{\mu} \in \mathcal{D}. \tag{4.3.11}$$

Note that we can rewrite (4.3.11) equivalently as

$$\begin{aligned} \|y^{\mathcal{N}}(\boldsymbol{\mu}) - y_N^{\mathcal{N}}(\boldsymbol{\mu})\|_Y^2 + \|u^{\mathcal{N}}(\boldsymbol{\mu}) - u_N^{\mathcal{N}}(\boldsymbol{\mu})\|_U^2 + \|p^{\mathcal{N}}(\boldsymbol{\mu}) - p_N^{\mathcal{N}}(\boldsymbol{\mu})\|_Q^2 & \\ \le \frac{1}{\hat{\beta}_{\mathrm{LB}}^2(\boldsymbol{\mu})} \Big(\|r_{\underline{x}}(\cdot;\boldsymbol{\mu})\|_{X'}^2 + \|r_p(\cdot;\boldsymbol{\mu})\|_{Q'}^2 \Big) = \big(\Delta_N(\boldsymbol{\mu})\big)^2. & \end{aligned}$$

4.3.2 Offline-Online procedure

We introduce the Riesz representation of $\mathsf{r}(\cdot;\boldsymbol{\mu})$: $\hat{\mathsf{e}}(\boldsymbol{\mu}) \in \mathcal{X}^{\mathcal{N}}$ satisfies

$$(\hat{\mathsf{e}}(\boldsymbol{\mu}),\mathsf{w})_{\mathcal{X}} = \mathsf{r}(\mathsf{w};\boldsymbol{\mu}), \qquad \forall \mathsf{w} \in \mathcal{X}^{\mathcal{N}}. \tag{4.3.12}$$

The dual norm of the residuals can be evaluated through its Riesz representation:

$$\|\mathsf{r}(\cdot;\boldsymbol{\mu})\|_{\mathcal{X}'} = \sup_{\mathsf{w}\in\mathcal{X}^{\mathcal{N}}} \frac{\mathsf{r}(\mathsf{w};\boldsymbol{\mu})}{\|\mathsf{w}\|_{\mathcal{X}}} = \|\hat{\mathsf{e}}(\boldsymbol{\mu})\|_{\mathcal{X}}.$$

From the affine decompositions of the bilinear forms (4.1.8) we can write equivalently

$$\mathsf{B}(\mathsf{x},\mathsf{w};\boldsymbol{\mu}) = \sum_{q=1}^{Q_a+2Q_b} \Theta_B^q(\boldsymbol{\mu}) \mathsf{B}^q(\mathsf{x},\mathsf{w}), \tag{4.3.13}$$

where

$$\begin{aligned} &\Theta_B^q(\boldsymbol{\mu}) = \Theta_a^q(\boldsymbol{\mu}), && \mathsf{B}^q(\mathsf{x},\mathsf{w}) = \mathcal{A}^q(\underline{x},\underline{w}), && 1 \le q \le Q_a, \\ &\Theta_B^q(\boldsymbol{\mu}) = \Theta_b^q(\boldsymbol{\mu}), && \mathsf{B}^q(\mathsf{x},\mathsf{w}) = \mathcal{B}^q(\underline{w},p), && Q_a+1 \le q \le Q_a+Q_b, \\ &\Theta_B^q(\boldsymbol{\mu}) = \Theta_b^q(\boldsymbol{\mu}), && \mathsf{B}^q(\mathsf{x},\mathsf{w}) = \mathcal{B}^q(\underline{x},q), && Q_a+Q_b+1 \le q \le Q_a+2Q_b. \end{aligned}$$

Similarly, using (4.1.9), the linear functional $\mathsf{F}(\cdot;\boldsymbol{\mu})$ can be expressed as

$$\mathsf{F}(\mathsf{w};\boldsymbol{\mu}) = \sum_{q=1}^{Q_f+Q_g} \Theta_F^q(\boldsymbol{\mu})\mathsf{F}^q(\mathsf{w}), \tag{4.3.14}$$

where

$$\begin{aligned} &\Theta_F^q(\boldsymbol{\mu}) = \Theta_f^q(\boldsymbol{\mu}), && \mathsf{F}^q(\mathsf{w}) = \langle F^q, \underline{w}\rangle, && 1 \le q \le Q_f, \\ &\Theta_F^q(\boldsymbol{\mu}) = \Theta_g^q(\boldsymbol{\mu}), && \mathsf{F}^q(\mathsf{w}) = \langle G^q, q\rangle, && Q_f + 1 \le q \le Q_f + Q_g. \end{aligned}$$

In this way, recalling that $\mathsf{x}_N(\boldsymbol{\mu}) = (\underline{x}_N(\boldsymbol{\mu}), p_N(\boldsymbol{\mu})) \in \mathbb{R}^{5N}$ denotes the global vector of the RB components, the residual can be expressed as

$$\begin{aligned} \mathsf{r}(\mathsf{w};\boldsymbol{\mu}) &= \mathsf{F}(\mathsf{w};\boldsymbol{\mu}) - \mathsf{B}\Big(\sum_{n=1}^{5N} \mathsf{x}_{Nn}(\boldsymbol{\mu})\Phi_n, \mathsf{w}; \boldsymbol{\mu}\Big) \\ &= \sum_{q=1}^{Q_F} \Theta_F^q(\boldsymbol{\mu})\mathsf{F}^q(\mathsf{w}) - \sum_{n=1}^{5N} \mathsf{x}_{Nn}(\boldsymbol{\mu}) \sum_{q=1}^{Q_B} \Theta_B^q(\boldsymbol{\mu})\mathsf{B}^q(\Phi_n, \mathsf{w}), \end{aligned} \tag{4.3.15}$$

where $Q_B = Q_a + 2Q_b$, $Q_F = Q_f + Q_g$ and

$$\Phi_n = (\underline{\sigma}_n, 0), \quad 1 \le n \le 3N, \qquad \Phi_n = (0, \tau_n), \quad 3N + 1 \le n \le 5N.$$

Recalling the relation (4.3.12), we thus may write $\hat{\mathsf{e}}(\boldsymbol{\mu}) \in \mathcal{X}^{\mathcal{N}}$ as

$$\hat{\mathsf{e}}(\boldsymbol{\mu}) = \sum_{q=1}^{Q_F} \Theta_F^q(\boldsymbol{\mu})\mathcal{F}^q - \sum_{n=1}^{5N} \mathsf{x}_{Nn}(\boldsymbol{\mu}) \sum_{q=1}^{Q_B} \Theta_B^q(\boldsymbol{\mu})\mathcal{L}_n^q,$$

where $\forall \mathsf{w} \in \mathcal{X}^{\mathcal{N}}$

$$(\mathcal{F}^q, \mathsf{w})_{\mathcal{X}} = \mathsf{F}^q(\mathsf{w}), \qquad 1 \le q \le Q_F, \tag{4.3.16}$$

$$(\mathcal{L}_n^q, v)_{\mathcal{X}} = -\mathsf{B}^q(\Phi_n, \mathsf{w}), \qquad 1 \le q \le Q_B, \quad 1 \le n \le 5N. \tag{4.3.17}$$

Note that $\mathcal{F}^q$ is the Riesz representation of $\mathsf{F}^q(\cdot)$ and $\mathcal{L}_n^q$ is the Riesz representation of $\mathsf{B}^q(\Phi_n, \cdot)$, computable solving parameter-independent Poisson-like problems. Finally we obtain

$$\begin{aligned} \|\hat{\mathsf{e}}(\boldsymbol{\mu})\|_{\mathcal{X}}^2 = \sum_{q=1}^{Q_F}\sum_{q'=1}^{Q_F} \Theta_F^q(\boldsymbol{\mu})\Theta_F^{q'}(\boldsymbol{\mu})(\mathcal{F}^q, \mathcal{F}^{q'})_{\mathcal{X}} + \sum_{q=1}^{Q_B}\sum_{n=1}^{5N} \Theta_B^q(\boldsymbol{\mu})\mathsf{x}_{Nn}(\boldsymbol{\mu})\Bigg\{ \\ 2\sum_{q'=1}^{Q_F} \Theta_F^{q'}(\boldsymbol{\mu})(\mathcal{F}^{q'}, \mathcal{L}_n^q)_{\mathcal{X}} + \sum_{q'=1}^{Q_B}\sum_{n'=1}^{5N} \Theta_B^{q'}(\boldsymbol{\mu})\mathsf{x}_{Nn'}(\boldsymbol{\mu})(\mathcal{L}_n^q, \mathcal{L}_n^{q'})_{\mathcal{X}} \Bigg\} \end{aligned} \tag{4.3.18}$$

from which we can calculate the dual norm of the residual. Let us summarize the Offline-Online decomposition:

1. in the Offline stage we first compute the Q_F terms $\mathcal{F}^q$ and the $5NQ_B$ terms $\mathcal{L}_n^q$ solving problems (4.3.16) and (4.3.17) respectively; then we store the scalar products $(\mathcal{F}^q, \mathcal{F}^{q'})_{\mathcal{X}}$, $(\mathcal{F}^{q'}, \mathcal{L}_n^q)_{\mathcal{X}}$ and $(\mathcal{L}_n^q, \mathcal{L}_n^q)_{\mathcal{X}}$. The Offline operation count depends then on N, Q_B, Q_F and $\mathcal{N}$.

2. in the Online stage, for each new value of $\boldsymbol{\mu}$, we simply evaluate the sum (4.3.18) in terms of the $\Theta^q(\boldsymbol{\mu})$, x_{Nn} and the precomputed scalar products. The Online operation count is $O(25N^2Q_B^2 + 10NQ_BQ_f + 5NQ_F^2)$, independent of $\mathcal{N}$.

4.3.3 A posteriori error bound for the cost functional

We aim to develop an a posteriori error bound on the cost functional $J(y,u;\boldsymbol{\mu})$. We firstly observe that this is equivalent to provide an estimator for the error on $\mathcal{J}(\underline{x};\boldsymbol{\mu})$, since $\mathcal{J}(\cdot;\boldsymbol{\mu})$ and $J(\cdot,\cdot;\boldsymbol{\mu})$ differ only in a constant term. Although the cost functional $\mathcal{J}(\cdot;\boldsymbol{\mu})$ is a *quadratic* functional, thanks to the structure of the optimal control problem we can avoid to use the techniques of error estimation for quadratic outputs already proposed in the RB context, see for instance [45, 85, 42]. Rather, following the work in [17] we may use the recipes of goal-oriented analysis, a standard tool for the development of a posteriori error estimates for optimal control problems (see for example [5] in the context of mesh adaptivity).

Let us recall the notation introduced above: we denote with $\mathsf{x} = (\underline{x}, p) = (y,u,p) \in \mathcal{X}$ the solution of the optimal control problem (i.e. state, adjoint and control variables), with $\mathsf{w} = (\underline{w}, q) = (z,v,q)$ the generic test function in the space $\mathcal{X}$. The error on the cost functional evaluated with respect to the FE and RB approximations will be denoted with

$$\mathcal{J}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathcal{J}_N(\boldsymbol{\mu}) = J(y^{\mathcal{N}}(\boldsymbol{\mu}), u^{\mathcal{N}}(\boldsymbol{\mu});\boldsymbol{\mu}) - J(y_N(\boldsymbol{\mu}), u_N(\boldsymbol{\mu});\boldsymbol{\mu}),$$

while the Lagrangian functional associated with the optimal control problem reads

$$\mathcal{L}(\mathsf{x};\boldsymbol{\mu}) = \mathcal{J}(\underline{x};\boldsymbol{\mu}) + \mathcal{B}(\underline{x},p;\boldsymbol{\mu}) - \langle G(\boldsymbol{\mu}), p\rangle. \tag{4.3.19}$$

Note that the first order optimality conditions system given by $\nabla\mathcal{L}(\mathsf{x};\boldsymbol{\mu})[\mathsf{w}] = 0$ coincides with the saddle-point problem (just different formalism), i.e.

$$\nabla\mathcal{L}(\mathsf{x};\boldsymbol{\mu})[\mathsf{w}] = \mathsf{B}(\mathsf{x},\mathsf{w};\boldsymbol{\mu}) - \mathsf{F}(\mathsf{w};\boldsymbol{\mu}), \qquad \forall \mathsf{w} \in \mathcal{X}. \tag{4.3.20}$$

The crucial result is given by the following Proposition, the proof follows standard arguments, see e.g. [6, 5, 19, 17].

Proposition 4.2. *The RB error on the cost functional is equal to*

$$\mathcal{J}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathcal{J}_N(\boldsymbol{\mu}) = \frac{1}{2}\nabla\mathcal{L}(\mathsf{x}_N(\boldsymbol{\mu});\boldsymbol{\mu})[\mathsf{x}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathsf{x}_N(\boldsymbol{\mu})]. \tag{4.3.21}$$

Proof. The problem statements for $(\underline{x}^{\mathcal{N}}, p^{\mathcal{N}})$ (4.1.10) and $(\underline{x}_N, p_N)$ (4.2.8) imply that $\mathcal{J}^{\mathcal{N}}(\boldsymbol{\mu}) = \mathcal{L}(\mathsf{x}^{\mathcal{N}};\boldsymbol{\mu})$ and $\mathcal{J}_N(\boldsymbol{\mu}) = \mathcal{L}(\mathsf{x}_N;\boldsymbol{\mu})$. By applying the fundamental theorem of calculus and thanks to the linearity of the Lagrangian functional, we have that

$$\begin{aligned}\mathcal{J}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathcal{J}_N(\boldsymbol{\mu}) &= \mathcal{L}(\mathsf{x}^{\mathcal{N}};\boldsymbol{\mu}) - \mathcal{L}(\mathsf{x}_N;\boldsymbol{\mu}) \\ &= \int_{\mathsf{x}_N}^{\mathsf{x}^{\mathcal{N}}} \nabla\mathcal{L}(\mathsf{x};\boldsymbol{\mu})\cdot d\mathbf{x} = \int_0^1 \nabla\mathcal{L}(s\mathsf{x}^{\mathcal{N}} + (1-s)\mathsf{x}_N);\boldsymbol{\mu})[\mathsf{x}^{\mathcal{N}} - \mathsf{x}_N]\,ds \\ &= \frac{1}{2}\nabla\mathcal{L}(\mathsf{x}^{\mathcal{N}};\boldsymbol{\mu})[\mathsf{x}^{\mathcal{N}} - \mathsf{x}_N] + \frac{1}{2}\nabla\mathcal{L}(\mathsf{x}_N;\boldsymbol{\mu})[\mathsf{x}^{\mathcal{N}} - \mathsf{x}_N].\end{aligned}$$

Since $\mathcal{X}_N \subset \mathcal{X}^{\mathcal{N}}$ we obtain that $\nabla\mathcal{L}(\mathsf{x}^{\mathcal{N}};\boldsymbol{\mu})[\mathsf{x}^{\mathcal{N}} - \mathsf{x}_N] = 0$, the result (4.3.21) follows. □

Thanks to (4.3.20) we have that

$$\nabla\mathcal{L}(\mathsf{x}_N;\boldsymbol{\mu})[\mathsf{x}^{\mathcal{N}} - \mathsf{x}_N] = \mathsf{B}(\mathsf{x}_N, \mathsf{x}^{\mathcal{N}} - \mathsf{x}_N;\boldsymbol{\mu}) - \mathsf{F}(\mathsf{x}^{\mathcal{N}} - \mathsf{x}_N;\boldsymbol{\mu}) = \mathsf{r}(\mathsf{x}^{\mathcal{N}} - \mathsf{x}_N;\boldsymbol{\mu}),$$

exploiting the estimate (4.3.11) we finally obtain the following bound for the error on the cost functional:

$$
\begin{aligned}
|\mathcal{J}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathcal{J}_N(\boldsymbol{\mu})| &= \frac{1}{2}\nabla\mathcal{L}(\mathsf{x}_N(\boldsymbol{\mu});\boldsymbol{\mu})[\mathsf{x}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathsf{x}_N(\boldsymbol{\mu})] = \frac{1}{2}\mathsf{r}(\mathsf{x}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathsf{x}_N(\boldsymbol{\mu});\boldsymbol{\mu}) \\
&\leq \frac{1}{2}\|\mathsf{r}(\cdot;\boldsymbol{\mu})\|_{\mathcal{X}'}\|\mathsf{e}(\boldsymbol{\mu})\|_{\mathcal{X}} \leq \frac{1}{2}\frac{\|\hat{\mathsf{e}}(\boldsymbol{\mu})\|^2_{\mathcal{X}}}{\hat{\beta}_{\mathrm{LB}}(\boldsymbol{\mu})} := \Delta_N^J(\boldsymbol{\mu}).
\end{aligned}
\tag{4.3.22}
$$

Note that the error estimator $\Delta_N^J(\boldsymbol{\mu})$ does not need any additional ingredients than those already discussed: the efficient computation of the dual norm of the residual and the calculation of a lower bound for the Babuška inf-sup constant.

4.4 Numerical examples

In this section we present three numerical examples to test the performances of the RB method for optimal control problems. In the cases in which we consider a parametrized geometry we proceed as in Chapter 3: we firstly define an original problem posed over a parameters dependent domain, then we trace back the problem to a reference domain through the affine mappings (see Section 3.4.1) in order to recover the formulation (4.1.6).

The implementation of the method has been carried out in the MATLAB environment using an enhanced version[1] of the `rbMIT` library [44, 61], for the FE assembling stage we have exploited the `MLife` library [82]. Since the problems we deal with are of moderate size, all the required linear system solves will be done using the sparse direct solver provided by MATLAB. In particular, in the Offline stage, the direct solver will be used to solve in *one-shot* the required N_{max} finite element saddle-point problems. Note that, as the dimensions of the optimal control problems increase, one should rely on suitably preconditioned iterative solvers (like the preconditioned Krylov subspace methods discussed in Section 2.2).

4.4.1 Test 1: distributed optimal control for the Laplace equation with geometrical parametrization

The original domain $\Omega_o(\boldsymbol{\mu}) = \Omega_o^1 \cup \Omega_o^2(\boldsymbol{\mu})$ is a rectangle separated in two subdomains ($R = 2$), with the first one parameter independent while the second parameter dependent, as shown in Figure 4.1. We consider two parameters $\boldsymbol{\mu} = (\mu_1, \mu_2)$, being μ_1 the geometrical parameter and μ_2 such that

$$
y_d(\boldsymbol{\mu}) = \begin{cases} 1, & x \in \Omega_o^1, \\ \mu_2, & x \in \Omega_o^2(\boldsymbol{\mu}), \end{cases}
$$

i.e. the desired function is parameter dependent (constant on each subdomain). The parameter domain is given by $\mathcal{D} = [1, 3.5] \times [0.5, 2.5]$. We consider the following optimal control problem

$$
\begin{aligned}
&\min_{u_o \in U_o} J(y_o(\boldsymbol{\mu}), u_o(\boldsymbol{\mu}); \boldsymbol{\mu}) = \frac{1}{2}\|y_o(\boldsymbol{\mu}) - y_d(\boldsymbol{\mu})\|^2_{L^2(\Omega_o)} + \frac{\alpha}{2}\|u_o(\boldsymbol{\mu})\|^2_{U_o}, \\
&\text{s.t.} \begin{cases} -\Delta y_o(\boldsymbol{\mu}) = u_o(\boldsymbol{\mu}) & \text{in } \Omega_o(\boldsymbol{\mu}), \\ y_o(\boldsymbol{\mu}) = g_D & \text{on } \Gamma_D^o(\boldsymbol{\mu}) = \partial\Omega_o(\boldsymbol{\mu}), \end{cases}
\end{aligned}
\tag{4.4.1}
$$

[1] Co-developed at CMCS (Chair of Modelling and Scientific Computing), EPFL, based on the official released version of `rbMIT`.

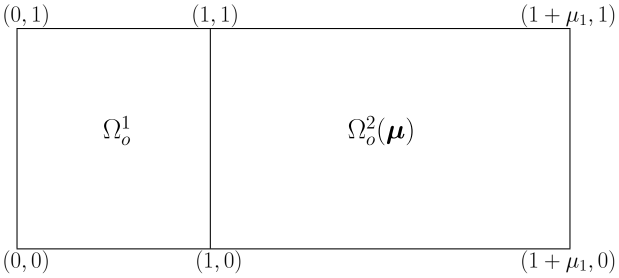


Figure 4.1: Test 1: original domain $\Omega_o(\boldsymbol{\mu})$.

where y_o and u_o are the state and control functions defined on the original domain, the Dirichlet boundary condition is given by $g_D = 1$. We denote with Y_o and U_o the spaces $H_0^1(\Omega_o)$ and $L^2(\Omega_o)$ respectively, moreover $Q_o \equiv Y_o$. We also introduce a lift function $R_g \in H^1(\Omega_o)$ such that $R_g|_{\Gamma_D^o} = g_D$ and $y_o = \tilde{y}_o + R_g$; for the sake of simplicity, we still denote $\tilde{y}_o$ with y_o in the sequel. Hence, the weak formulation of the state equation reads: find $y_o \in Y_o$ such that

$$a_o(y_o, q; \boldsymbol{\mu}) = c_o(u_o, q; \boldsymbol{\mu}) + \langle G_o, q\rangle \qquad \forall q \in Q_o,$$

where the bilinear form $a_o\colon Y_o \times Q_o \to \mathbb{R}$ and $c_o\colon U_o \times Q_o \to \mathbb{R}$ are defined as follows

$$a_o(z, q; \boldsymbol{\mu}) = \sum_{r=1}^{R} \int_{\Omega_o^r} \nabla z \cdot \nabla q \, d\Omega_o, \qquad c_o(v, q; \boldsymbol{\mu}) = \sum_{r=1}^{R} \int_{\Omega_o^r} vq \, d\Omega_o,$$

and the term $G_o(\boldsymbol{\mu})$ is due to non-homogeneous Dirichlet boundary condition on Γ_D^o, i.e.

$$\langle G_o(\boldsymbol{\mu}), q\rangle = -a(R_g, q; \boldsymbol{\mu}).$$

Moreover the bilinear forms $m_o(\cdot,\cdot;\boldsymbol{\mu}) : Y_o \times Y_o \to \mathbb{R}$ and $n_o(\cdot,\cdot;\boldsymbol{\mu}) : U_o \times U_o \to \mathbb{R}$ are given by

$$m_o(y, z; \boldsymbol{\mu}) = \sum_{r=1}^{R} \int_{\Omega_o^r} y_o z \, d\Omega_o, \qquad n_o(u, v; \boldsymbol{\mu}) = \sum_{r=1}^{R} \int_{\Omega_o^r} \alpha u_o v \, d\Omega_o.$$

In order to formulate the optimal control problem as a saddle-point problem, let $X_o = Y_o \times U_o$, $\underline{x}_o = (y_o, u_o) \in X_o$, $\underline{w} = (z, v) \in X_o$, $p_o, q \in Q_o$ and define the bilinear forms

$$\begin{aligned} \mathcal{A}_o(\underline{x}_o, \underline{w}; \boldsymbol{\mu}) &= m_o(y_o, z; \boldsymbol{\mu}) + n_o(u_o, v; \boldsymbol{\mu}), \\ \mathcal{B}_o(\underline{w}, q; \mu) &= a_o(z, q; \boldsymbol{\mu}) - c_o(v, q; \boldsymbol{\mu}), \end{aligned}$$

and the linear functional

$$\langle \underline{F}_o(\boldsymbol{\mu}), \underline{w}\rangle = \sum_{r=1}^{R} \int_{\Omega_o^r} y_d(\boldsymbol{\mu}) z \, d\Omega_o.$$

We denote with $\Omega = \Omega_o(\boldsymbol{\mu}_{\text{ref}})$ the reference domain, with the choice $\boldsymbol{\mu}_{\text{ref}} = (1,1)$. Being the first subdomain parameter independent the corresponding affine geometrical mapping is trivial

$$\boldsymbol{C}^1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \boldsymbol{G}^1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad J^r = 1,$$

while for the second subdomain we obtain

$$\boldsymbol{C}^2 = \begin{bmatrix} 1-\mu_1 \\ 0 \end{bmatrix}, \qquad \boldsymbol{G}^2 = \begin{bmatrix} \mu_1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad J^r = \mu_1.$$

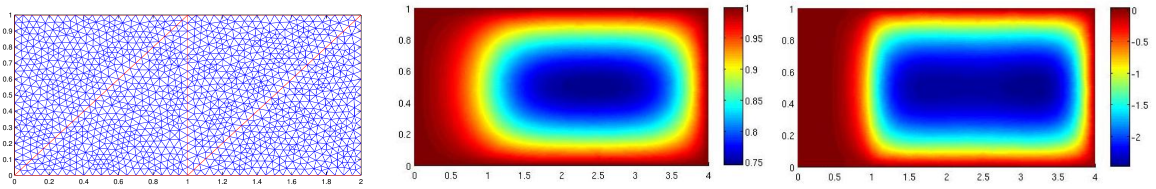
By tracing the problem back to the reference domain we obtain the parametrized formulation (4.1.6) where the affine decompositions (4.1.8) (4.1.9) of the bilinear forms are given by: $Q_a = 2$, $Q_b = 3$, $Q_f = 2$, $Q_g = 3$ and

$$\begin{aligned}
&\Theta_a^1(\boldsymbol{\mu}) = 1, && \mathcal{A}^1(\underline{x}, \underline{w}) = \int_{\Omega_1} yz\, d\Omega + \int_{\Omega_1} \alpha uv\, d\Omega, \\
&\Theta_a^2(\boldsymbol{\mu}) = \mu_1, && \mathcal{A}^2(\underline{x}, \underline{w}) = \int_{\Omega_2} yz\, d\Omega + \int_{\Omega_2} \alpha uv\, d\Omega, \\
&\Theta_b^1(\boldsymbol{\mu}) = 1, && \mathcal{B}^1(\underline{w}, p) = \int_{\Omega_1} \frac{\partial z}{\partial x_1}\frac{\partial p}{\partial x_1}\, d\Omega + \int_{\Omega_1} \frac{\partial z}{\partial x_2}\frac{\partial p}{\partial x_2}\, d\Omega - \int_{\Omega_1} vp\, d\Omega, \\
&\Theta_b^2(\boldsymbol{\mu}) = \frac{1}{\mu_1}, && \mathcal{B}^2(\underline{w}, p) = \int_{\Omega_2} \frac{\partial z}{\partial x_1}\frac{\partial p}{\partial x_1}\, d\Omega, \\
&\Theta_b^3(\boldsymbol{\mu}) = \mu_1, && \mathcal{B}^3(\underline{w}, p) = \int_{\Omega_2} \frac{\partial z}{\partial x_2}\frac{\partial p}{\partial x_2}\, d\Omega - \int_{\Omega_2} vp\, d\Omega,
\end{aligned}$$

$$\Theta_f^1(\boldsymbol{\mu}) = 1, \qquad \langle \underline{F}^1, \underline{w} \rangle = \int_{\Omega_1} z\, d\Omega, \qquad \Theta_2^f(\boldsymbol{\mu}) = \mu_1\mu_2, \qquad \langle \underline{F}^2, \underline{w} \rangle = \int_{\Omega_2} z\, d\Omega,$$

and $\langle G(\boldsymbol{\mu}), q \rangle = -\mathcal{B}(\{R_g, 0\}, q; \boldsymbol{\mu})$.

We fixed $\alpha = 0.01$ and we used piecewise linear finite elements for the FE approximation, the dimension of the global FE space $\mathcal{X}^{\mathcal{N}}$ used is $\mathcal{N} = 4857$. The computational (reference) domain as well as plots of the solution for different values of the parameters are given in Figure 4.2 and 4.3.



(a) Computational reference domain.

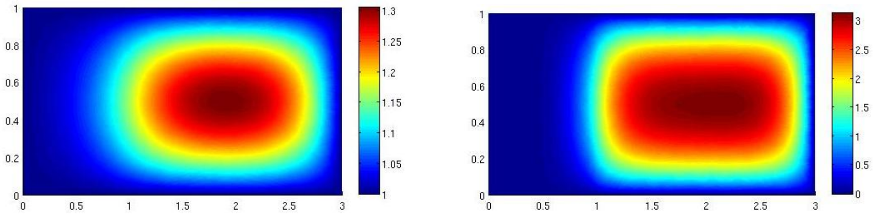
(b) state $y(\boldsymbol{\mu})$

(c) control $u(\boldsymbol{\mu})$

Figure 4.2: Test 1: on the left the computational domain, on the right representative solution for $\boldsymbol{\mu} = (0.6, 3)$ (the value of the cost functional is $J = 0.054$).

With a fixed tolerance $\varepsilon_{tol} = 5 \cdot 10^{-4}$, $N_{max} = 9$ basis functions have been selected, thus resulting in a RB linear system of dimension 45×45. In Figure 4.4 we show the lower bound for the Babuška inf-sup constant $\hat{\beta}^{\mathcal{N}}(\boldsymbol{\mu})$ obtained using the natural norm SCM algorithm with a tolerance $\varepsilon_{\mathrm{SCM}} = 0.85$ and a uniform train sample of size $n_{\mathrm{train,SCM}} = 1000$; SCM requires in this case the solution of $10 + 2(Q_a + 2Q_b)$ eigenproblems. In Figure 4.4 is reported also the RB Babuška inf-sup constant $\hat{\beta}_N(\boldsymbol{\mu})$, in particular we can observe that $\hat{\beta}_N(\boldsymbol{\mu}) \geq \hat{\beta}^{\mathcal{N}}(\boldsymbol{\mu})$, thus indicating the good stability property of the RB approximation.

Furthermore, as regards the stability properties, in Figure 4.5 we give some numerical results on the Brezzi inf-sup constants $\beta^{\mathcal{N}}(\boldsymbol{\mu})$ and $\beta_N(\boldsymbol{\mu})$, also compared with the coercivity constant



(a) state $y(\boldsymbol{\mu})$ **(b)** control $u(\boldsymbol{\mu})$

Figure 4.3: Test 1: representative solutions for $\boldsymbol{\mu} = (1.5, 2)$. The value of the cost functional is $J = 0.086$.

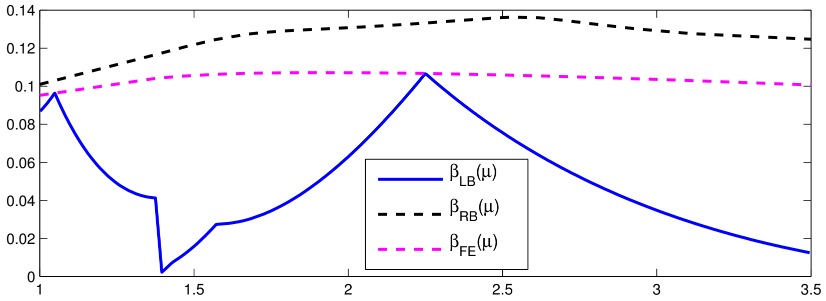


Figure 4.4: Test1: lower bound for the Babuška inf-sup constant $\hat{\beta}(\boldsymbol{\mu})$ as a function of the geometrical parameter μ_1 ($\hat{\beta}_{\text{LB}}(\boldsymbol{\mu})$ in blue, $\hat{\beta}^{\mathcal{N}}(\boldsymbol{\mu})$ in magenta, $\hat{\beta}_N(\boldsymbol{\mu})$ in black).

$\tilde{\alpha}(\boldsymbol{\mu})$ of the bilinear form $a(\cdot, \cdot; \boldsymbol{\mu})$ in the state equation. Let us explain and comment the two figures:

- in Figure 4.5a we report some results obtained in a preliminary numerical investigation without any enrichment option, i.e. using different RB spaces Y_N and Q_N (see Section 4.2.1). We compare the Brezzi inf-sup condition $\beta(\boldsymbol{\mu})$ and the coercivity constant $\tilde{\alpha}(\boldsymbol{\mu})$ for the FE and RB approximation. We can confirm that, as proved in Lemma 1.2, $\beta^{\mathcal{N}}(\boldsymbol{\mu}) \geq \tilde{\alpha}^{\mathcal{N}}(\boldsymbol{\mu})$. Moreover we observe that

$$\tilde{\alpha}_N(\boldsymbol{\mu}) \leq \beta_N(\boldsymbol{\mu}) \leq \tilde{\alpha}^{\mathcal{N}}(\boldsymbol{\mu}) \leq \beta^{\mathcal{N}}(\boldsymbol{\mu}),$$

  hence (as expected) we cannot bound from below the RB inf-sup constant $\beta_N(\boldsymbol{\mu})$ with similar quantities related to the FE approximations. We note also that in this case, as already mentioned in Remark 4.2, the coercivity constant $\tilde{\alpha}_N(\boldsymbol{\mu})$ is in fact an inf-sup constant since we are approximating the state equation with a Petrov-Galerkin scheme, i.e.

$$\tilde{\alpha}_N(\boldsymbol{\mu}) = \inf_{q \in Q_N} \sup_{y \in Y_N} \frac{a(y, q; \boldsymbol{\mu})}{\|q\|_Q \|y\|_Y}, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}.$$

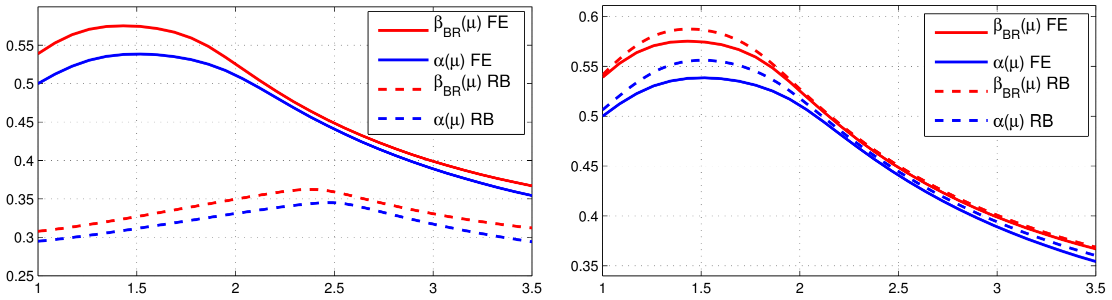
- in Figure 4.5b we compare the RB stability factors obtained using the aggregated space Z_N for the state and adjoint variables. In this case we have confirmed numerically that

$$\beta_N(\boldsymbol{\mu}) \geq \tilde{\alpha}_N(\boldsymbol{\mu}) \geq \tilde{\alpha}^{\mathcal{N}}(\boldsymbol{\mu}), \qquad \forall \boldsymbol{\mu} \in \mathcal{D},$$

  as proved in Lemma 4.1.

Finally in Figure 4.6 we compare the a posteriori error bound $\Delta_N(\boldsymbol{\mu})$ with the true error $\|\mathsf{x}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathsf{x}_N(\boldsymbol{\mu})\|_{\mathcal{X}}$ and the a posteriori error bound $\Delta_N^J(\boldsymbol{\mu})$ with the true error on the cost functional $|\mathcal{J}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathcal{J}_N(\boldsymbol{\mu})|$.

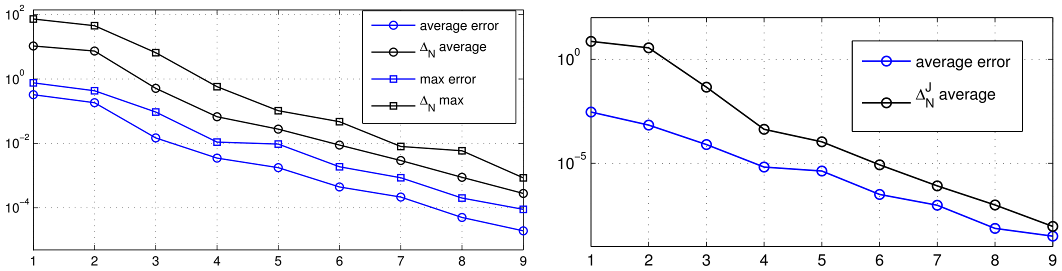
As regards the computational performances, the Offline computational time is equal to $t_{RB}^{offline} = 1620s$, the (average) Online solution time is $t_{RB}^{online} = 0.08s$ comprehensive of the evaluation of the a posteriori error estimation. As already noted in Section 3.5.6 for the Stokes problem, most of the Offline time is spent performing the SCM algorithm, since it requires about the 70% of the overall Offline computational time. The evaluation time for the FE approximation is equal to about $t_{FE} = 13s$ taking into account the time needed for assembling the FE matrices and vectors. The computational speedup defined as $\mathcal{S} = t_{FE}/t_{RB}^{online}$ is about 160 while the break-even point defined as $\mathcal{Q}_{BE} = t_{RB}^{offline}/t_{FE}$ is about 120.



(a) No enrichment: $Y_N \neq Q_N$

(b) Aggregated space: $Z_N = Y_N \cup Q_N$

Figure 4.5: Test 1: comparison of Brezzi inf-sup constant $\beta(\boldsymbol{\mu})$ and coercivity constant of the state equation $\tilde{\alpha}(\boldsymbol{\mu})$ for the FE and RB approximations. The two quantities are given as function only of μ_1, since μ_2 does not appear in the affine expansion of $\mathcal{B}(\cdot,\cdot;\boldsymbol{\mu})$.



(a) Average and max computed errors and bounds between the *truth* FE solution and the RB approximation, for $N = 1, \cdots, N_{max}$.

(b) Average computed errors and bounds $\Delta_N^J(\mu)$ between $J^{\mathcal{N}}(\mu)$ and $J_N(\mu)$, for $N = 1, \cdots, N_{max}$.

Figure 4.6: Test 1: a posteriori error bounds. Here Ξ_{train} is a sample of size $n_{train} = 1000$ and $N_{max} = 9$.

4.4.2 Test 2: distributed optimal control for a Graetz convection-diffusion problem with physical parametrization

As a second example we consider a distributed optimal control problem for the Graetz conduction-convection equation introduced in Section 3.4.3. With respect to the previous test we consider a simple physical parametrization instead of a geometrical one, in particular μ_1 will be the Péclet number, while μ_2, μ_3 similarly to the previous example are such that

$$\mu_1 = \mathrm{Pe}, \qquad y_d(\boldsymbol{\mu}) = \begin{cases} \mu_2, & x \in \hat{\Omega}_1 \\ \mu_3, & x \in \hat{\Omega}_2, \end{cases}$$

where the domain (shown in Figure 4.7) is the rectangle $\Omega = [0, 2.5] \times [0, 1]$ and the observation subdomains are given by $\hat{\Omega}_1 = [0.2, 0.8] \times [0.3, 0.7]$, $\hat{\Omega}_2 = [1.2, 2.5] \times [0.3, 0.7]$. The parameter domain is $\mathcal{D} = [3, 20] \times [0.5, 1.5] \times [1.5, 2.5]$.

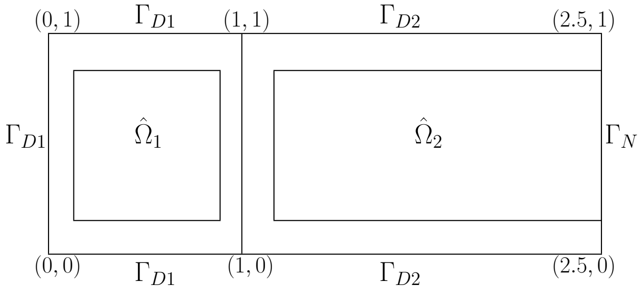


Figure 4.7: Test 2: domain Ω. $\hat{\Omega}_1$ and $\hat{\Omega}_2$ are the observation domains.

We consider the following distributed optimal control problem:

$$\min_{u \in U} J(y, u; \boldsymbol{\mu}) = \frac{1}{2} \|y(\boldsymbol{\mu}) - y_d(\boldsymbol{\mu})\|^2_{L^2(\hat{\Omega})} + \frac{\alpha}{2} \|u(\boldsymbol{\mu})\|^2_{L^2(\Omega)}, \qquad \text{subject to}$$

$$\begin{cases} -\dfrac{1}{\mu_1} \Delta y(\boldsymbol{\mu}) + x_2(1 - x_2) \dfrac{\partial y(\boldsymbol{\mu})}{\partial x_1} = u(\boldsymbol{\mu}) & \text{in } \Omega \\ y(\boldsymbol{\mu}) = 1 & \text{on } \Gamma_{D1} \\ y(\boldsymbol{\mu}) = 2 & \text{on } \Gamma_{D2} \\ \dfrac{1}{\mu_1} \nabla y(\mu) \cdot \boldsymbol{n} = 0 & \text{on } \Gamma_N. \end{cases} \tag{4.4.2}$$

where $y(\boldsymbol{\mu})$ is the temperature field, the control $u(\boldsymbol{\mu})$ acts as a heat source and $\hat{\Omega} = \hat{\Omega}_1 \cup \hat{\Omega}_2$ is the observation domain. We have the usual affine decompositions with $Q_a = 1$, $Q_b = 2$, $Q_f = 2$, $Q_g = 2$ and

$$\Theta^1_a(\boldsymbol{\mu}) = 1, \qquad \mathcal{A}^1(\underline{x}, \underline{w}) = \int_{\hat{\Omega}} yz \, d\Omega + \int_{\Omega} \alpha uv \, d\Omega,$$

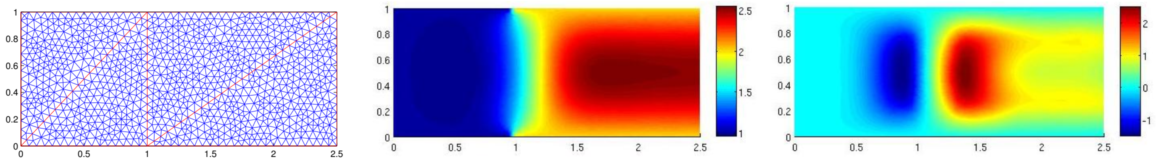
$$\Theta^1_b(\boldsymbol{\mu}) = \frac{1}{\mu_1}, \qquad \mathcal{B}^1(\underline{w}, p) = \int_{\Omega} \frac{\partial z}{\partial x_1} \frac{\partial p}{\partial x_1} \, d\Omega + \int_{\Omega} \frac{\partial z}{\partial x_2} \frac{\partial p}{\partial x_2} \, d\Omega,$$

$$\Theta^2_b(\boldsymbol{\mu}) = 1, \qquad \mathcal{B}^2(\underline{w}, p) = \int_{\Omega} x_2(1 - x_2) \frac{\partial z}{\partial x_1} p \, d\Omega - \int_{\Omega} vp \, d\Omega,$$

$$\Theta^1_f(\boldsymbol{\mu}) = \mu_2, \qquad \langle \underline{F}^1, \underline{w} \rangle = \int_{\hat{\Omega}_1} z \, d\Omega, \qquad \Theta^2_f(\boldsymbol{\mu}) = \mu_3, \qquad \langle \underline{F}^2, \underline{w} \rangle = \int_{\hat{\Omega}_2} z \, d\Omega,$$

and $\langle G(\boldsymbol{\mu}), q \rangle = -\mathcal{B}(\{R_g, 0\}, q; \boldsymbol{\mu})$.

For the computation we fixed $\alpha = 0.01$ and used piecewise linear finite elements for the FE approximation The computational domain as well as plots of the solution for different values of the parameters are given in Figure 4.8 and 4.9.

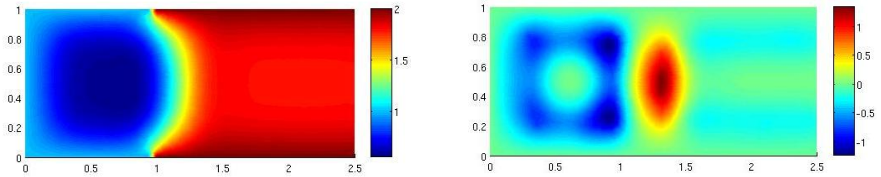


(a) Computational reference domain.

(b) state $y(\boldsymbol{\mu})$

(c) control $u(\boldsymbol{\mu})$

Figure 4.8: Test 2: on the left the computational domain, on the right representative solution for $\boldsymbol{\mu} = (5, 1, 2.5)$ (the value of the cost functional is $J = 1.57 \cdot 10^{-2}$).

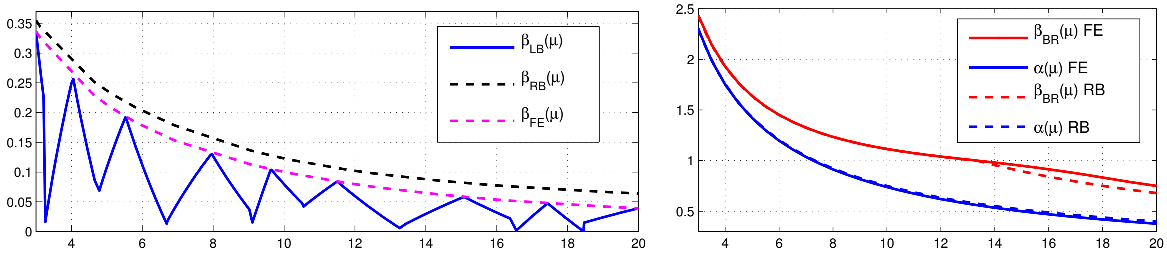


(a) state $y(\boldsymbol{\mu})$

(b) control $u(\boldsymbol{\mu})$

Figure 4.9: Test 2: representative solutions for $\boldsymbol{\mu} = (15, 0.6, 18)$. The value of the cost functional is $J = 2.7 \cdot 10^{-3}$.

With a fixed tolerance $\varepsilon_{tol} = 5 \cdot 10^{-4}$, $N_{max} = 17$ basis functions have been selected, thus resulting in a RB linear system of dimension 85×85. In Figure 4.10a we show the lower bound for the Babuška inf-sup constant $\hat{\beta}^{\mathcal{N}}(\boldsymbol{\mu})$ obtained using the natural norm SCM algorithm with a tolerance $\varepsilon_{\text{SCM}} = 0.85$ and a uniform train sample of size $n_{\text{train,SCM}} = 1000$; SCM requires in this case the solution of $28 + 2(Q_a + 2Q_b)$ eigenproblems. Once again we can observe that $\hat{\beta}_N(\boldsymbol{\mu}) \geq \hat{\beta}^{\mathcal{N}}(\boldsymbol{\mu})$, thus indicating the good stability property of the RB approximation.



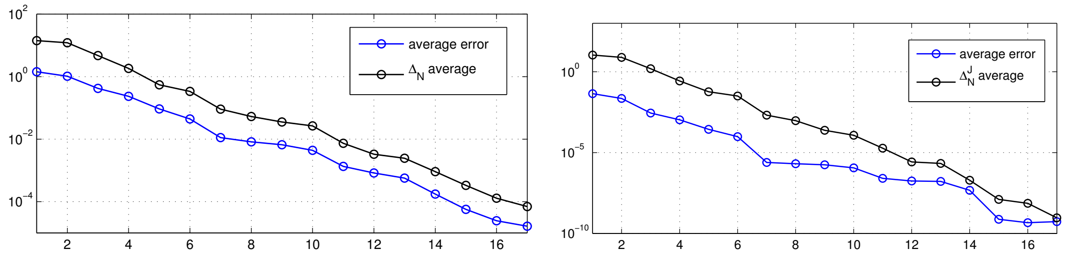
(a) Lower bound for the Babuška inf-sup constant $\hat{\beta}(\boldsymbol{\mu})$ as a function of the physical parameter μ_1 ($\hat{\beta}_{\text{LB}}(\boldsymbol{\mu})$ in blue, $\hat{\beta}^{\mathcal{N}}(\boldsymbol{\mu})$ in magenta, $\beta_N(\boldsymbol{\mu})$ in black).

(b) Comparison of Brezzi inf-sup constant $\beta(\boldsymbol{\mu})$ and coercivity constant of the state equation $\tilde{\alpha}(\boldsymbol{\mu})$ for the FE and RB approximations as functions of μ_1.

Figure 4.10: Test 2: stability properties.

In Figure 4.10b we compare the Brezzi inf-sup constants $\beta^{\mathcal{N}}(\boldsymbol{\mu})$ and $\beta_N(\boldsymbol{\mu})$ and the coercivity constants $\tilde{\alpha}^{\mathcal{N}}(\boldsymbol{\mu})$ and $\tilde{\alpha}_N(\boldsymbol{\mu})$ of the bilinear form $a(\cdot, \cdot; \boldsymbol{\mu})$. As in the previous example we have confirmed numerically that $\beta_N(\boldsymbol{\mu}) \geq \tilde{\alpha}_N(\boldsymbol{\mu}) \geq \tilde{\alpha}^{\mathcal{N}}(\boldsymbol{\mu})$. Finally in Figure 4.11 we compare the a posteriori error bound $\Delta_N(\boldsymbol{\mu})$ with the true error $\|\mathsf{x}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathsf{x}_N(\boldsymbol{\mu})\|_{\mathcal{X}}$ and the a posteriori error bound $\Delta_N^J(\boldsymbol{\mu})$ with the true error on the cost functional $|\mathcal{J}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathcal{J}_N(\boldsymbol{\mu})|$.

With regard to the computational performances, the Offline computational time is equal

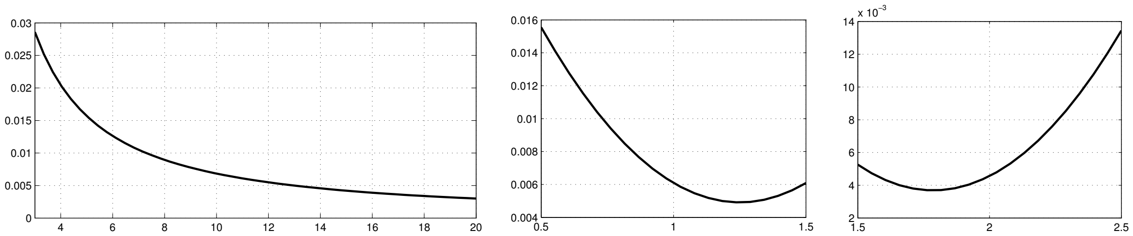


(a) Average computed errors and bounds between the *truth* FE solution and the RB approximation, for $N = 1, \cdots, N_{max}$.

(b) Average computed errors and bounds $\Delta^J_N(\mu)$ between $J^{\mathcal{N}}(\mu)$ and $J_N(\mu)$, for $N = 1, \cdots, N_{max}$.

Figure 4.11: Test 2: a posteriori error bounds. Here Ξ_{train} is a sample of size $n_{train} = 1000$ and $N_{max} = 17$.

to $t^{offline}_{RB} = 2420s$, the (average) Online solution time is about $t^{online}_{RB} = 0.085s$ comprehensive of the evaluation of the a posteriori error estimation while the evaluation time for the FE approximation is equal to about $t_{FE} = 12.3s$, thus resulting again in a speedup of two orders of magnitude. In Figure 4.12 we show the graphs of the cost functional $J(\boldsymbol{\mu})$ as a function of one parameter at a time, keeping the others fixed.



(a) $J(\boldsymbol{\mu})$ function of μ_1; $\mu_2 = 1$, $\mu_3 = 2.5$.

(b) $J(\boldsymbol{\mu})$ function of μ_2; $\mu_1 = 7$, $\mu_3 = 2.3$.

(c) $J(\boldsymbol{\mu})$ function of μ_3; $\mu_1 = 7$, $\mu_2 = 0.8$.

Figure 4.12: Test 2: value of the cost functional $J(\boldsymbol{\mu})$ as a function of the parameters.

4.4.3 Test 3: boundary optimal control for a Graetz flow with physical and geometrical parametrization

This third example deals again with a control problem for the Graetz equation, however this time we consider a boundary control instead of a distributed one and we consider both a geometrical and physical parametrization. The original domain is shown in Figure 4.13, we consider 3 parameters: μ_1 is the Péclet number, μ_2 is the geometrical parameter (the length of second portion of the channel) and μ_3 is such that $y_d(\boldsymbol{\mu}) = \mu_3 \chi_{\hat{\Omega}_o}$, being $\hat{\Omega}_o(\boldsymbol{\mu})$ the observation domain $\hat{\Omega}_o(\boldsymbol{\mu}) \subset \Omega^2_o(\boldsymbol{\mu})$. The parameter domain is $\mathcal{D} = [6, 20] \times [1, 3] \times [0.5, 3]$.

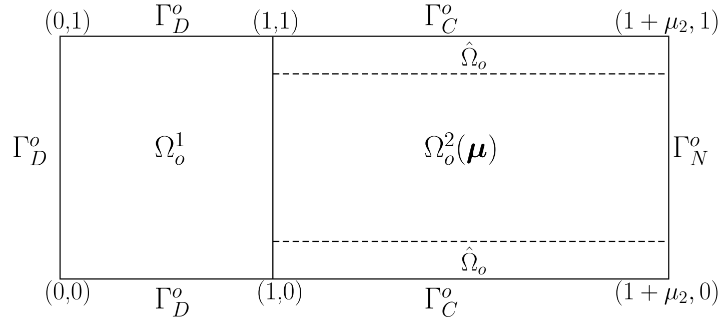


Figure 4.13: Original domain for Test 3.

We consider the following optimal control problem

$$\min_{u_o \in U_o} J(y_o(\boldsymbol{\mu}), u_o(\boldsymbol{\mu}); \boldsymbol{\mu}) = \frac{1}{2}\|y_o(\boldsymbol{\mu}) - y_d(\boldsymbol{\mu})\|^2_{L^2(\hat{\Omega}_o)} + \frac{\alpha}{2}\|u_o(\boldsymbol{\mu})\|^2_{U_o}, \qquad \text{subject to}$$

$$\begin{cases} -\dfrac{1}{\mu_1}\Delta y_o(\boldsymbol{\mu}) + x_{o2}(1-x_{o2})\dfrac{\partial y_o(\boldsymbol{\mu})}{\partial x_{o1}} = 0 & \text{in } \Omega_o(\boldsymbol{\mu}) \\ y_o(\boldsymbol{\mu}) = 1 & \text{on } \Gamma_D^o \\ \dfrac{1}{\mu_1}\nabla y_o(\boldsymbol{\mu}) \cdot \mathbf{n} = u_o(\boldsymbol{\mu}) & \text{on } \Gamma_C^o(\boldsymbol{\mu}) \\ \dfrac{1}{\mu_1}\nabla y_o(\boldsymbol{\mu}) \cdot \mathbf{n} = 0 & \text{on } \Gamma_N^o(\boldsymbol{\mu}), \end{cases} \tag{4.4.3}$$

where y_o and u_o are state and control functions defined on the original domain; we impose constant Dirichlet conditions on the inlet boundary of the channel, homogeneous Neumann condition on the outlet boundary and finally Neumann condition equal to the control function u_o on Γ_C^o. We set the regularization parameter $\alpha = 0.07$. We denote with Y_o and U_o the spaces $H_0^1(\Omega_o)$ and $L^2(\Gamma_C^o)$ respectively, moreover $Q_o = Y_o$. We also introduce a lift function $R_g \in H^1(\Omega_o)$ such that $R_g|_{\Gamma_D^o} = g_D$ and $y_o = \tilde{y}_o + R_g$; for the sake of simplicity, we still denote $\tilde{y}_o$ with y_o in the sequel. Hence, the weak formulation of the state equation reads: find $y_o \in Y_o$ such that

$$a_o(y_o, q; \boldsymbol{\mu}) = c_o(u_o, q; \boldsymbol{\mu}) + \langle G_o(\boldsymbol{\mu}), q\rangle \qquad \forall q \in Q_o$$

where the bilinear form $a_o \colon Y_o \times Q_o \to \mathbb{R}$ and $c_o \colon U_o \times Q_o \to \mathbb{R}$ are defined as follows

$$a_o(z, q; \boldsymbol{\mu}) = \sum_{r=1}^{2} \int_{\Omega_o^r} \left\{\frac{1}{\mu_1}\nabla z \cdot \nabla q + x_{o2}(1-x_{o2})\frac{\partial z}{\partial x_{o1}} q\right\} d\Omega_o, \qquad c_o(v, q; \boldsymbol{\mu}) = \int_{\Gamma_C^o} vq \, d\Gamma_o,$$

and the term G_o is due to non-homogeneous Dirichlet boundary condition on Γ_D^o, i.e.

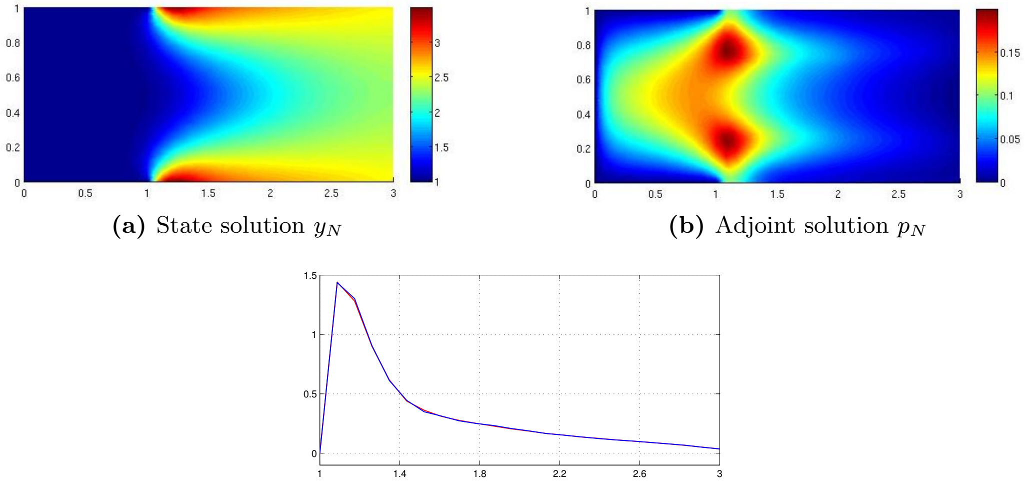
$$\langle G_o(\boldsymbol{\mu}), q\rangle = -a_o(R_g, q; \boldsymbol{\mu}).$$

We denote with $\Omega = \Omega_o(\boldsymbol{\mu}_{\text{ref}})$ the reference domain, with the choice $\boldsymbol{\mu}_{\text{ref}} = (\cdot, 1, \cdot)$. The affine geometrical mappings are the same as in Test 1, in particular for the second subdomain we have

$$\boldsymbol{C}^2 = \begin{bmatrix} 1-\mu_2 \\ 0 \end{bmatrix}, \qquad \boldsymbol{G}^2 = \begin{bmatrix} \mu_2 & 0 \\ 0 & 1 \end{bmatrix}, \qquad J^r = \mu_2.$$

By tracing the problem back to the reference domain we obtain the parametrized formulation (4.1.6) where the affine decompositions (4.1.8) (4.1.9) of the linear and bilinear forms are given by: $Q_a = 1$, $Q_b = 5$, $Q_f = 1$, $Q_g = 4$ and

$$\begin{aligned}
&\Theta_a^1(\boldsymbol{\mu}) = \mu_2, && \mathcal{A}^1(\underline{x},\underline{w}) = \int_{\Omega_2} \chi_{\hat{\Omega}} yz\, d\Omega + \int_{\Gamma_C} \alpha uv\, d\Gamma,\\
&\Theta_b^1(\boldsymbol{\mu}) = \frac{1}{\mu_1}, && \mathcal{B}^1(\underline{w},p) = \int_{\Omega_1} \frac{\partial z}{\partial x_1}\frac{\partial p}{\partial x_1}\, d\Omega + \int_{\Omega_1} \frac{\partial z}{\partial x_2}\frac{\partial p}{\partial x_2}\, d\Omega,\\
&\Theta_b^2(\boldsymbol{\mu}) = \frac{1}{\mu_2\mu_1}, && \mathcal{B}^2(\underline{w},p) = \int_{\Omega_2} \frac{\partial z}{\partial x_1}\frac{\partial p}{\partial x_1}\, d\Omega,\\
&\Theta_b^3(\boldsymbol{\mu}) = \frac{\mu_2}{\mu_1}, && \mathcal{B}^3(\underline{w},p) = \int_{\Omega_2} \frac{\partial z}{\partial x_2}\frac{\partial p}{\partial x_2}\, d\Omega,\\
&\Theta_b^4(\boldsymbol{\mu}) = 1, && \mathcal{B}^4(\underline{w},p) = \int_{\Omega_1} x_2(1-x_2)\frac{\partial z}{\partial x_1} p\, d\Omega + \int_{\Omega_2} x_2(1-x_2)\frac{\partial z}{\partial x_1} p\, d\Omega,\\
&\Theta_b^5(\boldsymbol{\mu}) = \mu_2, && \mathcal{B}^5(\underline{w},p) = \int_{\Gamma_C} vp\, d\Gamma,\\
&\Theta_f^1(\boldsymbol{\mu}) = \mu_2\mu_3, && \langle \underline{F}^1, \underline{w}\rangle = \int_{\Omega_2} \chi_{\hat{\Omega}} z\, d\Omega.
\end{aligned}$$



(a) State solution y_N

(b) Adjoint solution p_N

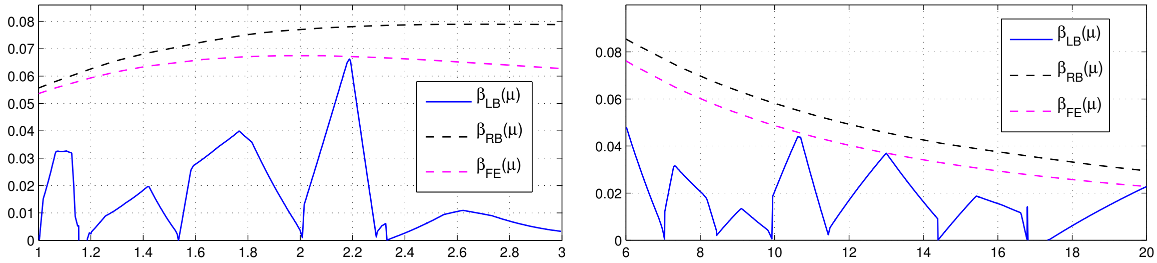
(c) Optimal controls u_N on Γ_C^o

Figure 4.14: Test 3: representative solution for $\boldsymbol{\mu} = (12, 2, 2.5)$.

For the computation we fixed $\alpha = 0.07$ and used piecewise linear finite elements for the FE approximation The computational domain as well as plots of the solution for different values of the parameters are given in Figure 4.14.

With a fixed tolerance $\varepsilon_{tol} = 5 \cdot 10^{-4}$, $N_{max} = 39$ basis functions have been selected. In Figure 4.15 we show the lower bound for the Babuška inf-sup constant $\hat{\beta}^{\mathcal{N}}(\boldsymbol{\mu})$ obtained using the natural norm SCM algorithm with a tolerance $\varepsilon_{\text{SCM}} = 0.85$ and a uniform train sample of size $n_{\text{train,SCM}} = 2000$; SCM requires in this case the solution of $143 + 2(Q_a + 2Q_b)$ eigenproblems.

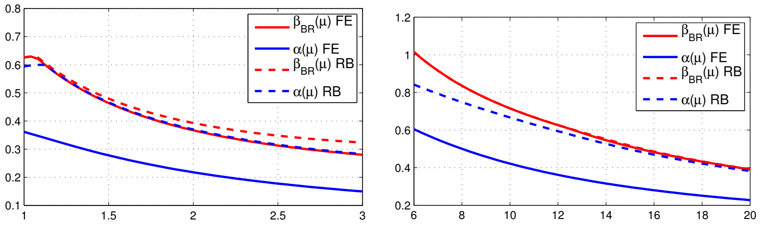
In Figure 4.16 we compare the Brezzi inf-sup constants $\beta^{\mathcal{N}}(\boldsymbol{\mu})$ and $\beta_N(\boldsymbol{\mu})$ and the coercivity constants $\tilde{\alpha}^{\mathcal{N}}(\boldsymbol{\mu})$ and $\tilde{\alpha}_N(\boldsymbol{\mu})$of the bilinear form $a(\cdot,\cdot;\boldsymbol{\mu})$. As in the previous example we



(a) $\hat{\beta}(\boldsymbol{\mu})$ as a function of μ_2, $(\mu_1, \mu_3) = (7, 1.8)$ fixed. **(b)** $\hat{\beta}(\boldsymbol{\mu})$ as a function of μ_1, $(\mu_2, \mu_3) = (2, 1.8)$ fixed.

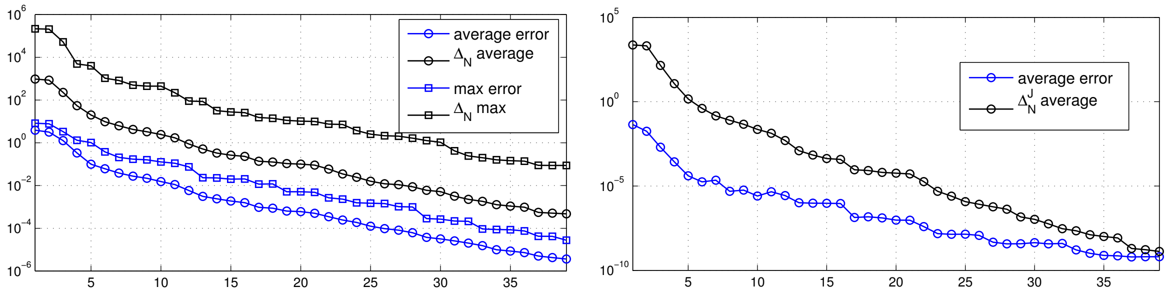
Figure 4.15: Test 3: lower bound for Babuška inf-sup constant $\hat{\beta}(\boldsymbol{\mu})$ ($\hat{\beta}_{\text{LB}}(\boldsymbol{\mu})$ in blue, $\hat{\beta}^{\mathcal{N}}(\boldsymbol{\mu})$ in magenta, $\beta_N(\boldsymbol{\mu})$ in black).

have confirmed numerically that $\beta_N(\boldsymbol{\mu}) \geq \tilde{\alpha}_N(\boldsymbol{\mu}) \geq \tilde{\alpha}^{\mathcal{N}}(\boldsymbol{\mu})$. Finally in Figure 4.11 we compare the a posteriori error bound $\Delta_N(\boldsymbol{\mu})$ with the true error $\|\mathsf{x}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathsf{x}_N(\boldsymbol{\mu})\|_{\mathcal{X}}$ and the a posteriori error bound $\Delta_N^J(\boldsymbol{\mu})$ with the true error on the cost functional $|\mathcal{J}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathcal{J}_N(\boldsymbol{\mu})|$.



(a) $\beta(\boldsymbol{\mu})$ as a function of μ_2, $(\mu_1, \mu_3) = (12, 1.8)$ fixed. **(b)** $\beta(\boldsymbol{\mu})$ as a function of μ_1, $(\mu_2, \mu_3) = (1, 2.2)$ fixed. **(c)** $\beta(\boldsymbol{\mu})$ as a function of μ_1, $(\mu_2, \mu_3) = (2.5, 2.2)$ fixed.

Figure 4.16: Test 3: comparison of Brezzi inf-sup constant $\beta(\boldsymbol{\mu})$ and coercivity constant of the state equation $\tilde{\alpha}(\boldsymbol{\mu})$ for the FE and RB approximations. In **(b)** the $\beta(\boldsymbol{\mu})$ inf-sup constant of the FE and RB approximations coincide.



(a) Average and max computed error and bound between the *truth* FE solution and the RB approximation, for $N = 1, \cdots, N_{max}$. **(b)** Average computed error and bound $\Delta_N^J(\boldsymbol{\mu})$ between $J^{\mathcal{N}}(\boldsymbol{\mu})$ and $J_N(\boldsymbol{\mu})$, for $N = 1, \cdots, N_{max}$.

Figure 4.17: Test 3: a posteriori error bounds. Here Ξ_{train} is a sample of size $n_{train} = 2500$ and $N_{max} = 39$.

Chapter 5

Reduced Basis Method for Parametrized Stokes Optimal Control problems

We provide here a reduced basis framework for the efficient solution of parametrized linear/quadratic optimal control problems governed by Stokes equations. We shall adapt the methodology developed in Chapter 5 to the case with Stokes system as constraint. The main features of our RB approach will be the same: (i) a reduced basis made by the FE solutions for some selected (by a proper greedy procedure) parameter values $S = \{\boldsymbol{\mu}^1, \ldots, \boldsymbol{\mu}^N\}$; (ii) a reduced approximation of the saddle-point problem obtained as a Galerkin projection onto the reduced spaces (iii) a *rigorous*, sharp and inexpensive a posteriori error estimator $\Delta_N(\boldsymbol{\mu})$ for the control, state and adjoint variables; (iv) the standard Offline/Online decomposition stratagem [77], that enables to decouple the generation and projection stages of the RB approximation.
After having tested the properties of the method on two simple numerical examples, we apply it to an inverse problem in haemodynamics: a data assimilation problem for an arterial bifurcation.

The chapter is structured as follows. In Section 5.1 we introduce the formulation of parametrized linear/quadratic optimal control problems governed by Stokes system with affine parameter dependence; we briefly discuss also the FE approximation, recalling the necessary assumptions to ensure the well-posedness. In Section 5.2 we discuss the RB approximation and the main features of the method, focusing on the corresponding stability condition for the RB approximation. Then in Section 5.3 we deal with the a posteriori error estimation for the RB solution and functional based on the Babuška stability theory. In Section 5.4 two numerical examples are presented, and finally in Section 5.5 we show the potentialities of the method with the application to a data assimilation problem.

5.1 Problem formulation

We consider the parametrized version of the optimal control problems for the Stokes equations introduced in Section 1.2.4. Let $\Omega \subset \mathbb{R}^2$ be a spatial domain with Lipschitz boundary $\partial\Omega$, we introduce the functional spaces for the velocity and pressure variables, $[H_0^1(\Omega)]^2 \subset V \subset [H^1(\Omega)]^2$ and $M = L^2(\Omega)$ respectively. Then we define the spaces $Y = V \times M$ for the state variables $\{\boldsymbol{v}, p\}$, $Q \equiv Y$ for the adjoint variables $\{\boldsymbol{w}, q\}$, $U = [L^2(\omega)]^2$ for the control variable $\boldsymbol{u}$, where ω can be the whole domain Ω, a subdomain or a boundary. Moreover $\mathcal{Z}$ shall denote the observation space, typically $\mathcal{Z} = [L^2(\hat{\omega})]^2$, being $\hat{\omega}$ either the whole domain Ω, a subdomain or a boundary. We consider the following parametrized optimal control problem:

$$\text{minimize } \; J(\boldsymbol{v}, p, \boldsymbol{u}; \boldsymbol{\mu}) = \frac{1}{2} m(\boldsymbol{v} - \boldsymbol{v}_d(\boldsymbol{\mu}), \boldsymbol{v} - \boldsymbol{v}_d(\boldsymbol{\mu}); \boldsymbol{\mu}) + \frac{\alpha}{2} n(\boldsymbol{u}, \boldsymbol{u}; \boldsymbol{\mu}) \qquad \text{subject to}$$
$$\begin{cases} a(\boldsymbol{v}, \boldsymbol{\xi}; \boldsymbol{\mu}) + b(\boldsymbol{\xi}, p; \boldsymbol{\mu}) = \langle F(\boldsymbol{\mu}), \boldsymbol{\xi} \rangle + c(\boldsymbol{u}, \boldsymbol{\xi}; \boldsymbol{\mu}) & \forall \boldsymbol{\xi} \in V, \\ b(\boldsymbol{v}, \tau; \boldsymbol{\mu}) = \langle G(\boldsymbol{\mu}), \tau \rangle & \forall \tau \in M, \end{cases} \tag{5.1.1}$$

where $\boldsymbol{v}_d \in \mathcal{Z}$ and the linear and bilinear forms $F(\boldsymbol{\mu}) \in V'$, $G(\boldsymbol{\mu}) \in M'$, $a(\cdot, \cdot; \boldsymbol{\mu})$ and $b(\cdot, \cdot; \boldsymbol{\mu})$ are those defined in Section 3.5.1. We assume the bilinear form $c(\cdot, \cdot; \boldsymbol{\mu}) : U \times Q \to \mathbb{R}$ to be symmetric and bounded over $U \times V$, and the bilinear form $n(\cdot, \cdot; \boldsymbol{\mu}) : U \times U \to \mathbb{R}$ to be symmetric, bounded over $U \times U$ and coercive over U. Moreover we assume the bilinear form $m(\cdot, \cdot; \boldsymbol{\mu})$ to be symmetric, continuous and positive in the norm induced on the space $\mathcal{Z}$. Let us rewrite problem (5.1.1) in the form

$$\begin{aligned} \text{minimize } \; & J(\boldsymbol{v}, p, \boldsymbol{u}; \boldsymbol{\mu}) = \frac{1}{2} m(\boldsymbol{v} - \boldsymbol{v}_d(\boldsymbol{\mu}), \boldsymbol{v} - \boldsymbol{v}_d(\boldsymbol{\mu}); \boldsymbol{\mu}) + \frac{\alpha}{2} n(\boldsymbol{u}, \boldsymbol{u}; \boldsymbol{\mu}) \\ \text{s.t. } \; & \mathsf{A}(\{\boldsymbol{v}, p\}, \{\boldsymbol{\xi}, \tau\}; \boldsymbol{\mu}) = \langle \boldsymbol{G}(\boldsymbol{\mu}), \{\boldsymbol{\xi}, \tau\} \rangle + \mathsf{C}(\boldsymbol{u}, \boldsymbol{\xi}; \boldsymbol{\mu}) \qquad \forall \{\boldsymbol{\xi}, \tau\} \in Q, . \end{aligned} \tag{5.1.2}$$

where the bilinear form $\mathsf{A}(\cdot, \cdot; \boldsymbol{\mu}) : Y \times Q \to \mathbb{R}$ is given by

$$\mathsf{A}(\{\boldsymbol{v}, p\}, \{\boldsymbol{\xi}, \tau\}; \boldsymbol{\mu}) = a(\boldsymbol{v}, \boldsymbol{\xi}; \boldsymbol{\mu}) + b(\boldsymbol{\xi}, p; \boldsymbol{\mu}) + b(\boldsymbol{v}, \tau; \boldsymbol{\mu}), \tag{5.1.3}$$

the linear continuous functional $\boldsymbol{G}(\boldsymbol{\mu}) \in Q'$ is given by

$$\langle \boldsymbol{G}(\boldsymbol{\mu}), \{\boldsymbol{\xi}, \tau\} \rangle = \langle F(\boldsymbol{\mu}), \boldsymbol{\xi} \rangle + \langle G(\boldsymbol{\mu}), \tau \rangle, \tag{5.1.4}$$

while the bilinear form $\mathsf{C}(\cdot, \cdot; \boldsymbol{\mu}) : U \times Q \to \mathbb{R}$ is simply defined as

$$\mathsf{C}(\boldsymbol{u}, \{\boldsymbol{\xi}, \tau\}; \boldsymbol{\mu}) = c(\boldsymbol{u}, \boldsymbol{\xi}; \boldsymbol{\mu}). \tag{5.1.5}$$

Note that the bilinear form $\mathsf{C}(\cdot, \cdot; \boldsymbol{\mu})$ is bounded over $U \times Q$. We shall make an additional assumption, crucial to Offline-Online procedures, by assuming the bilinear forms $m(\cdot, \cdot; \boldsymbol{\mu})$, $n(\cdot, \cdot, \boldsymbol{\mu})$ and $\mathsf{C}(\cdot, \cdot, \boldsymbol{\mu})$ to be affine in the parameter $\boldsymbol{\mu}$, i.e. for some finite $\tilde{Q}_*$, $* \in \{c, m, n\}$, they can be expressed as

$$\begin{gathered} m(\boldsymbol{v}, \boldsymbol{\varphi}; \boldsymbol{\mu}) = \sum_{q=1}^{\tilde{Q}_m} \tilde{\Theta}_m^q(\boldsymbol{\mu})\, m^q(\boldsymbol{v}, \boldsymbol{\varphi}), \qquad n(\boldsymbol{u}, \boldsymbol{\lambda}; \boldsymbol{\mu}) = \sum_{q=1}^{\tilde{Q}_n} \tilde{\Theta}_n^q(\boldsymbol{\mu})\, n^q(\boldsymbol{u}, \boldsymbol{\lambda}), \\ \mathsf{C}(\boldsymbol{u}, \{\boldsymbol{\xi}, \tau\}; \boldsymbol{\mu}) = \sum_{q=1}^{\tilde{Q}_c} \tilde{\Theta}_c^q(\boldsymbol{\mu})\, \mathsf{C}^q(\boldsymbol{u}, \{\boldsymbol{\xi}, \tau\}), \end{gathered} \tag{5.1.6}$$

for given *smooth* $\boldsymbol{\mu}$-dependent function $\tilde{\Theta}_*^q(\boldsymbol{\mu})$ and continuous $\boldsymbol{\mu}$-independent bilinear and linear forms $m^q(\cdot, \cdot)$, $\mathsf{C}^q(\cdot, \cdot)$, $n^q(\cdot, \cdot)$. As regards the bilinear and linear forms $\mathsf{A}(\cdot, \cdot; \boldsymbol{\mu})$ and

$\boldsymbol{G}(\cdot;\boldsymbol{\mu})$ we rely on the affine decompositions given in Section 3.5.1 and Section 3.5.5, i.e. we assume that

$$\begin{aligned} \mathsf{A}(\{\boldsymbol{v},p\},\{\boldsymbol{\xi},\tau\};\boldsymbol{\mu}) &= \sum_{q=1}^{\tilde{Q}_A} \tilde{\Theta}_A^q(\boldsymbol{\mu})\,\mathsf{A}^q(\{\boldsymbol{v},p\},\{\boldsymbol{\xi},\tau\}), \\ \langle \boldsymbol{G}(\boldsymbol{\mu}),\{\boldsymbol{\xi},\tau\}\rangle &= \sum_{q=1}^{\tilde{Q}_g} \tilde{\Theta}_g^q(\boldsymbol{\mu})\langle \boldsymbol{G}^q,\{\boldsymbol{\xi},\tau\}\rangle. \end{aligned} \tag{5.1.7}$$

In order to formulate the optimal control problem (5.1.2) as a saddle-point problem, we denote $X = Y \times U$, $\underline{\boldsymbol{x}} = (\{\boldsymbol{v},p\},\boldsymbol{u}) \in X$, $\underline{\boldsymbol{\zeta}} = (\{\boldsymbol{\varphi},\pi\},\boldsymbol{\lambda}) \in X$, $\{\boldsymbol{\xi},\tau\} \in Q$ and define the bilinear form $\mathcal{A}(\cdot,\cdot;\boldsymbol{\mu}) : X \times X \to \mathbb{R}$ as

$$\mathcal{A}(\underline{\boldsymbol{x}},\underline{\boldsymbol{\zeta}};\boldsymbol{\mu}) = m(\boldsymbol{v},\boldsymbol{\varphi};\boldsymbol{\mu}) + \alpha n(\boldsymbol{u},\boldsymbol{\lambda};\boldsymbol{\mu}), \qquad \forall \underline{\boldsymbol{x}},\underline{\boldsymbol{\zeta}} \in X,$$

and the bilinear form $\mathcal{B}(\cdot,\cdot;\boldsymbol{\mu}) : X \times Q \to \mathbb{R}$ as

$$\mathcal{B}(\underline{\boldsymbol{x}},\{\boldsymbol{\xi},\tau\};\boldsymbol{\mu}) = \mathsf{A}(\{\boldsymbol{v},p\},\{\boldsymbol{\xi},\tau\};\boldsymbol{\mu}) - \mathsf{C}(\boldsymbol{u},\{\boldsymbol{\xi},\tau\};\boldsymbol{\mu}), \qquad \forall \underline{\boldsymbol{x}} \in X, \{\boldsymbol{\xi},\tau\} \in Q.$$

Defining $\underline{\boldsymbol{F}}(\boldsymbol{\mu}) = m(\boldsymbol{v}_d(\boldsymbol{\mu}),\cdot) \in X'$, we can reformulate the problem (5.1.2) as (for the details see Section 1.2.4): given $\boldsymbol{\mu} \in \mathcal{D}$,

$$\begin{cases} \min \mathcal{J}(\underline{\boldsymbol{x}};\boldsymbol{\mu}) = \dfrac{1}{2}\mathcal{A}(\underline{\boldsymbol{x}},\underline{\boldsymbol{x}};\boldsymbol{\mu}) - \langle \underline{\boldsymbol{F}}(\boldsymbol{\mu}),\underline{\boldsymbol{x}}\rangle, & \text{subject to} \\ \mathcal{B}(\underline{\boldsymbol{x}},\{\boldsymbol{\xi},\tau\};\boldsymbol{\mu}) = \langle \boldsymbol{G}(\boldsymbol{\mu}),\{\boldsymbol{\xi},\tau\}\rangle \qquad \forall\{\boldsymbol{\xi},\tau\} \in Q. \end{cases} \tag{5.1.8}$$

Recalling the results proved in Section 1.2.4, we know that the assumptions made on the linear and bilinear forms in (5.1.2) guarantees the fulfillment of the hypotheses of Brezzi theorem (see Theorem A.2 and Proposition 1.1), which implies the equivalence between (5.1.8) and the following saddle-point problem: given $\boldsymbol{\mu} \in \mathcal{D}$, find $(\underline{\boldsymbol{x}}(\boldsymbol{\mu}),\{\boldsymbol{w}(\boldsymbol{\mu}),q(\boldsymbol{\mu})\}) \in X \times Q$ such that

$$\begin{cases} \mathcal{A}(\underline{\boldsymbol{x}},\underline{\boldsymbol{\zeta}};\boldsymbol{\mu}) + \mathcal{B}(\underline{\boldsymbol{\zeta}},\{\boldsymbol{w},q\};\boldsymbol{\mu}) = \langle \underline{\boldsymbol{F}}(\boldsymbol{\mu}),\underline{\boldsymbol{\zeta}}\rangle & \forall \underline{\boldsymbol{\zeta}} \in X, \\ \mathcal{B}(\underline{\boldsymbol{x}},\{\boldsymbol{\xi},\tau\};\boldsymbol{\mu}) = \langle \boldsymbol{G}(\boldsymbol{\mu}),\{\boldsymbol{\xi},\tau\}\rangle & \forall\{\boldsymbol{\xi},\tau\} \in Q. \end{cases} \tag{5.1.9}$$

In particular the bilinear forms $\mathcal{A}(\cdot,\cdot;\boldsymbol{\mu})$ and $\mathcal{B}(\cdot,\cdot;\boldsymbol{\mu})$ satisfy the following assumptions:

1. the bilinear form $\mathcal{A}(\cdot,\cdot;\boldsymbol{\mu})$ is continuous over $X \times X$:

$$\gamma_a(\boldsymbol{\mu}) = \sup_{\underline{\boldsymbol{x}}\in X}\sup_{\underline{\boldsymbol{\zeta}}\in X} \frac{\mathcal{A}(\underline{\boldsymbol{x}},\underline{\boldsymbol{\zeta}};\boldsymbol{\mu})}{\|\underline{\boldsymbol{\zeta}}\|_X\|\underline{\boldsymbol{x}}\|_X} < +\infty, \qquad \forall \boldsymbol{\mu} \in \mathcal{D};$$

2. the bilinear form $\mathcal{A}(\cdot,\cdot;\boldsymbol{\mu})$ is coercive over $X_0 = \{\underline{\boldsymbol{\zeta}} \in X : \mathcal{B}(\underline{\boldsymbol{\zeta}},\{\boldsymbol{\xi},\tau\};\boldsymbol{\mu}) = 0 \quad \forall\{\boldsymbol{\xi},\tau\} \in Q\} \subset X$, i.e. there exists a constant $\alpha_0 > 0$ such that

$$\alpha(\boldsymbol{\mu}) = \inf_{\underline{\boldsymbol{x}}\in X_0} \frac{\mathcal{A}(\underline{\boldsymbol{x}},\underline{\boldsymbol{x}};\boldsymbol{\mu})}{\|\underline{\boldsymbol{x}}\|_X^2} \geq \alpha_0, \qquad \forall \boldsymbol{\mu} \in \mathcal{D};$$

3. the bilinear form $\mathcal{B}(\cdot,\cdot;\boldsymbol{\mu})$ is continuous over $X \times Q$

$$\gamma_b(\boldsymbol{\mu}) = \sup_{\underline{\boldsymbol{\zeta}}\in X}\sup_{\{\boldsymbol{\xi},\tau\}\in Q} \frac{\mathcal{B}(\underline{\boldsymbol{\zeta}},\{\boldsymbol{\xi},\tau\};\boldsymbol{\mu})}{\|\underline{\boldsymbol{\zeta}}\|_X\|\{\boldsymbol{\xi},\tau\}\|_Q} < +\infty, \qquad \forall \boldsymbol{\mu} \in \mathcal{D};$$

4. the bilinear form $\mathcal{B}(\cdot,\cdot)$ satisfies the inf-sup condition over $X \times Q$, i.e. there exists a constant $\beta_0 > 0$ such that

$$\beta(\boldsymbol{\mu}) = \inf_{\{\boldsymbol{\xi},\tau\}\in Q} \sup_{\underline{\boldsymbol{\zeta}}\in X} \frac{\mathcal{B}(\underline{\boldsymbol{\zeta}}, \{\boldsymbol{\xi},\tau\};\boldsymbol{\mu})}{\|\underline{\boldsymbol{\zeta}}\|_X \|\{\boldsymbol{\xi},\tau\}\|_Q} \geq \beta_0, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{5.1.10}$$

5. the bilinear form $\mathcal{A}(\cdot,\cdot;\boldsymbol{\mu})$ is symmetric and non-negative over X.

Moreover, thanks to the affine parameter dependence assumption (5.1.6) and (5.1.7), an affine decomposition holds also for the bilinear and linear forms in (5.1.9), i.e. for some finite Q_a, Q_b, Q_f, Q_g, they can be expressed as

$$\mathcal{A}(\underline{\boldsymbol{x}},\underline{\boldsymbol{\zeta}};\boldsymbol{\mu}) = \sum_{q=1}^{Q_a} \Theta_a^q(\boldsymbol{\mu})\, \mathcal{A}^q(\underline{\boldsymbol{x}},\underline{\boldsymbol{\zeta}}), \qquad \mathcal{B}(\underline{\boldsymbol{x}},\{\boldsymbol{\xi},\tau\};\boldsymbol{\mu}) = \sum_{q=1}^{Q_b} \Theta_b^q(\boldsymbol{\mu})\, \mathcal{B}^q(\underline{\boldsymbol{x}},\{\boldsymbol{\xi},\tau\}), \tag{5.1.11}$$

$$\langle \boldsymbol{G}(\boldsymbol{\mu}),\{\boldsymbol{\xi},\tau\}\rangle = \sum_{q=1}^{Q_g} \Theta_g^q(\boldsymbol{\mu})\, \langle \boldsymbol{G}^q,\{\boldsymbol{\xi},\tau\}\rangle, \qquad \langle \underline{\boldsymbol{F}}(\boldsymbol{\mu}),\underline{\boldsymbol{\zeta}}\rangle = \sum_{q=1}^{Q_f} \Theta_f^q(\boldsymbol{\mu})\, \langle \underline{\boldsymbol{F}}^q,\underline{\boldsymbol{\zeta}}\rangle, \tag{5.1.12}$$

where the coefficients $\Theta^q(\boldsymbol{\mu})$ and the $\boldsymbol{\mu}$-independent linear and bilinear forms are related to those appearing in (5.1.6) and (5.1.7).

5.1.1 Truth approximation

Let us now introduce the FE-Galerkin approximation of the saddle-point problem (5.1.9). We assume that $Y^{\mathcal{N}} \subset Y$ be an inf-sup stable family of FE subspaces of Y for the Stokes system; moreover we assume that $U^{\mathcal{N}}$ be a family of FE subspaces of U and that $Q^{\mathcal{N}} \equiv Y^{\mathcal{N}}$. Note that these assumptions imply that $X^{\mathcal{N}} \equiv Y^{\mathcal{N}} \times U^{\mathcal{N}} \subset Y \times U \equiv X$. We indicate with $\mathcal{N}$ the global dimension of the product space $X^{\mathcal{N}} \times Q^{\mathcal{N}}$, i.e. $\mathcal{N} = \mathcal{N}_X + \mathcal{N}_Q$ where $\mathcal{N}_X = \mathcal{N}_Y + \mathcal{N}_U$ and $\mathcal{N}_Y = \mathcal{N}_Q$.
The *truth* Galerkin-FE approximation reads: given $\boldsymbol{\mu} \in \mathcal{D}$, find $(\underline{\boldsymbol{x}}^{\mathcal{N}}(\boldsymbol{\mu}), \{\boldsymbol{w}^{\mathcal{N}}(\boldsymbol{\mu}), q^{\mathcal{N}}(\boldsymbol{\mu})\}) \in X^{\mathcal{N}} \times Q^{\mathcal{N}}$ such that

$$\begin{cases} \mathcal{A}(\underline{\boldsymbol{x}}^{\mathcal{N}},\underline{\boldsymbol{\zeta}};\boldsymbol{\mu}) + \mathcal{B}(\underline{\boldsymbol{\zeta}},\{\boldsymbol{w}^{\mathcal{N}},q^{\mathcal{N}}\};\boldsymbol{\mu}) = \langle \underline{\boldsymbol{F}}(\boldsymbol{\mu}),\underline{\boldsymbol{\zeta}}\rangle & \forall \underline{\boldsymbol{\zeta}} \in X^{\mathcal{N}}, \\ \mathcal{B}(\underline{\boldsymbol{x}}^{\mathcal{N}},\{\boldsymbol{\xi},\tau\};\boldsymbol{\mu}) = \langle \boldsymbol{G}(\boldsymbol{\mu}),\{\boldsymbol{\xi},\tau\}\rangle & \forall \{\boldsymbol{\xi},\tau\} \in Q^{\mathcal{N}}. \end{cases} \tag{5.1.13}$$

As already proved in Lemma 1.4, provided $Y^{\mathcal{N}} \equiv Q^{\mathcal{N}}$, the bilinear form $\mathcal{A}(\cdot,\cdot;\boldsymbol{\mu})$ remains continuous over $X^{\mathcal{N}} \times X^{\mathcal{N}}$ and coercive over $X_0^{\mathcal{N}} = \{\underline{\boldsymbol{\zeta}} \in X^{\mathcal{N}} : \mathcal{B}(\underline{\boldsymbol{\zeta}},\{\boldsymbol{\xi},\tau\};\boldsymbol{\mu}) = 0 \quad \forall \{\boldsymbol{\xi},\tau\} \in Q^{\mathcal{N}}\}$, i.e.

$$\alpha^{\mathcal{N}}(\boldsymbol{\mu}) = \inf_{\underline{\boldsymbol{x}}\in X_0^{\mathcal{N}}} \frac{\mathcal{A}(\underline{\boldsymbol{x}},\underline{\boldsymbol{x}};\boldsymbol{\mu})}{\|\underline{\boldsymbol{x}}\|_X^2} \geq \alpha(\boldsymbol{\mu}) \geq \alpha_0, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}.$$

Similarly, the bilinear form $\mathcal{B}(\cdot,\cdot;\boldsymbol{\mu})$ remains continuous and inf-sup stable over $X^{\mathcal{N}} \times Q^{\mathcal{N}}$, i.e. there exists a constant $\beta_0 > 0$ such that

$$\beta^{\mathcal{N}}(\boldsymbol{\mu}) = \inf_{\{\boldsymbol{\xi},\tau\}\in Q^{\mathcal{N}}} \sup_{\underline{\boldsymbol{\zeta}}\in X^{\mathcal{N}}} \frac{\mathcal{B}(\underline{\boldsymbol{\zeta}}, \{\boldsymbol{\xi},\tau\};\boldsymbol{\mu})}{\|\underline{\boldsymbol{\zeta}}\|_X \|\{\boldsymbol{\xi},\tau\}\|_Q} \geq \beta_0, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}. \tag{5.1.14}$$

In particular we recall that we proved in Lemma 1.4 the estimate $\beta^{\mathcal{N}}(\boldsymbol{\mu}) \geq \tilde{\beta}^{\mathcal{N}}(\boldsymbol{\mu})$, being $\tilde{\beta}^{\mathcal{N}}(\boldsymbol{\mu})$ the Babuška inf-sup constant of the bilinear form $\mathsf{A}(\cdot,\cdot;\boldsymbol{\mu})$. Therefore the FE approximation (5.1.13) is well-posed, see Proposition 1.5.

At the algebraic level we obtain the linear system already introduced in Chapter 1 and widely discussed in Chapter 2, i.e.

$$\underbrace{\begin{pmatrix} A(\boldsymbol{\mu}) & B^T(\boldsymbol{\mu}) \\ B(\boldsymbol{\mu}) & 0 \end{pmatrix}}_{\mathcal{K}(\boldsymbol{\mu})} \begin{pmatrix} \mathbf{X}^{\mathcal{N}}(\boldsymbol{\mu}) \\ \mathbf{W}^{\mathcal{N}}(\boldsymbol{\mu}) \end{pmatrix} = \begin{pmatrix} \mathbf{F}(\boldsymbol{\mu}) \\ \mathbf{G}(\boldsymbol{\mu}) \end{pmatrix}.$$

where $\mathbf{V}$ is the state variable (velocity and pressure), $\mathbf{U}$ is the control variable, $\mathbf{W}$ is the adjoint variable (velocity and pressure) and $\mathbf{X} = \begin{pmatrix} \mathbf{V} & \mathbf{U} \end{pmatrix}^T$ denotes the state and control variables. Note that we have the same affine decompositions (5.1.11) and (5.1.12) for the matrices A, B,

$$A(\boldsymbol{\mu}) = \sum_{q=1}^{Q_a} \Theta_a^q(\boldsymbol{\mu})\, A^q, \qquad B(\boldsymbol{\mu}) = \sum_{q=1}^{Q_b} \Theta_b^q(\boldsymbol{\mu})\, B^q,$$

and for the right-hand sides

$$\mathbf{F}(\boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \Theta_f^q(\boldsymbol{\mu})\, \mathbf{F}^q, \qquad \mathbf{G}(\boldsymbol{\mu}) = \sum_{q=1}^{Q_g} \Theta_g^q(\boldsymbol{\mu})\, \mathbf{G}^q,$$

where the matrices and vectors A^q, B^q, $\mathbf{F}^q$, $\mathbf{G}^q$ represent the discrete counterparts of the corresponding bilinear and linear forms.

5.2 The reduced basis approximation

Once again, the aim of the RB method is to efficiently compute an approximation of $(\underline{\boldsymbol{x}}(\boldsymbol{\mu}), \{\boldsymbol{w}(\boldsymbol{\mu}), q(\boldsymbol{\mu})\}) \in X \times Q$ by using approximation spaces made up of well-chosen solutions of (5.1.13), i.e. corresponding to specific choices of the parameter values.

5.2.1 Formulation

Let us take, for given $N \in \{1, \dots, N_{\max}\}$, a set of parameter values $S_N = \{\boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^N\}$ and consider the corresponding FE solutions $\{(\underline{\boldsymbol{x}}^{\mathcal{N}}(\boldsymbol{\mu}^n), \{\boldsymbol{w}(\boldsymbol{\mu}^n), p^{\mathcal{N}}(\boldsymbol{\mu}^n)\}\}$, $n = 1, \dots, N\}$. In order to define suitable (i.e. stable) RB spaces Y_N, U_N and Q_N for the state, control and adjoint variables, respectively, we recall that:

1. as discussed in Section 3.5.2, the well-posedness of the RB approximation of the Stokes system is ensured by enrichnment of the velocity space with suitably defined supremizer solutions (3.5.10);

2. as discussed in Section 4.2 in the context of parametrized optimal control for coercive elliptic problems, in order to fulfill an equivalent RB inf-sup condition on the bilinear form $\mathcal{B}(\cdot, \cdot; \boldsymbol{\mu})$ an effective recipe is the use of the same RB spaces for the state and adjoint variables, i.e. $Y_N \equiv Q_N$.

Therefore to ensure the stability of the RB approximation we define the following *aggregated* spaces for the pressure variables

$$M_N = \operatorname{span}\{p^{\mathcal{N}}(\boldsymbol{\mu}^n),\ q^{\mathcal{N}}(\boldsymbol{\mu}), \quad n = 1, \dots, N\}, \tag{5.2.1}$$

and for the velocity variables

$$V_N^{\boldsymbol{\mu}} = V_N^{\boldsymbol{\mu},state} \cup V_N^{\boldsymbol{\mu},adj}, \tag{5.2.2}$$

where

$$V_N^{\boldsymbol{\mu},state} = \mathrm{span}\{\boldsymbol{v}^{\mathcal{N}}(\boldsymbol{\mu}^n), T^{\boldsymbol{\mu}} p^{\mathcal{N}}(\boldsymbol{\mu}^n)\}_{n=1}^N, \qquad V_N^{\boldsymbol{\mu},adj} = \mathrm{span}\{\boldsymbol{w}^{\mathcal{N}}(\boldsymbol{\mu}^n), T^{\boldsymbol{\mu}} q^{\mathcal{N}}(\boldsymbol{\mu}^n)\}_{n=1}^N.$$

Then we define the space for the control variable

$$U_N = \mathrm{span}\{\boldsymbol{u}^{\mathcal{N}}(\boldsymbol{\mu}^n), \quad n = 1, \dots, N\}, \tag{5.2.3}$$

and the following *aggregated* space for the state and adjoint variables

$$Z_N = V_N^{\boldsymbol{\mu}} \times M_N.$$

Now let $Y_N = Z_N$, $X_N = Z_N \times U_N$ and $Q_N = Z_N$, the RB approximation reads: given $\boldsymbol{\mu} \in \mathcal{D}$, find $(\underline{\boldsymbol{x}}_N(\boldsymbol{\mu}), \{\boldsymbol{w}_N(\boldsymbol{\mu}), q_N(\boldsymbol{\mu})\}) \in X_N \times Q_N$ such that

$$\begin{cases} \mathcal{A}(\underline{\boldsymbol{x}}_N, \underline{\boldsymbol{\zeta}}; \boldsymbol{\mu}) + \mathcal{B}(\underline{\boldsymbol{\zeta}}, \{\boldsymbol{w}_N, q_N\}; \boldsymbol{\mu}) = \langle \underline{\boldsymbol{F}}(\boldsymbol{\mu}), \underline{\boldsymbol{\zeta}} \rangle & \forall \underline{\boldsymbol{\zeta}} \in X_N, \\ \mathcal{B}(\underline{\boldsymbol{x}}_N, \{\boldsymbol{\xi}, \tau\}; \boldsymbol{\mu}) = \langle \boldsymbol{G}(\boldsymbol{\mu}), \{\boldsymbol{\xi}, \tau\} \rangle & \forall \{\boldsymbol{\xi}, \tau\} \in Q_N. \end{cases} \tag{5.2.4}$$

Let us discuss the well-posedness of the RB approximation. While the continuity property of the bilinear forms over the RB spaces are automatically inherited from the parents spaces (i.e. the FE spaces), the coercivity property of the bilinear form $\mathcal{A}(\cdot, \cdot; \boldsymbol{\mu})$ over

$$X_0^N = \{\underline{\boldsymbol{\zeta}} \in X_N \colon \mathcal{B}(\underline{\boldsymbol{\zeta}}, \{\boldsymbol{\xi}, \tau\}; \boldsymbol{\mu}) = 0 \quad \forall \{\boldsymbol{\xi}, \tau\} \in Q_N\},$$

and the fulfillment of the inf-sup condition of $\mathcal{B}(\cdot, \cdot; \boldsymbol{\mu})$ should be proved. In particular, the problem (5.2.4) should satisfy the following reduced basis inf-sup condition: there exists $\beta_0 > 0$ such that

$$\beta_N(\boldsymbol{\mu}) = \inf_{\{\boldsymbol{\xi},\tau\} \in Q_N} \sup_{\underline{\boldsymbol{x}} \in X_N} \frac{\mathcal{B}(\underline{\boldsymbol{x}}, \{\boldsymbol{\xi}, \tau\}; \boldsymbol{\mu})}{\|\underline{\boldsymbol{x}}\|_X \|\{\boldsymbol{\xi}, \tau\}\|_Q} \geq \beta_0, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}. \tag{5.2.5}$$

Lemma 5.1. *The bilinear form $\mathcal{B}(\cdot, \cdot; \boldsymbol{\mu})$ satisfies the inf-sup condition (5.2.5).*

Proof. Let us firstly note that, thanks to the enrichment by supremizer solutions in the RB velocity space V_N and to the fact that $Y_N \equiv Q_N$, there exists a positive constant $\tilde{\beta}_N^0$ such that

$$\tilde{\beta}_N(\boldsymbol{\mu}) = \inf_{\{\boldsymbol{v},p\} \in Y_N} \sup_{\{\boldsymbol{\xi},\tau\} \in Q_N} \frac{\mathsf{A}(\{\boldsymbol{v}, p\}, \{\boldsymbol{\xi}, \tau\}; \boldsymbol{\mu})}{\|\{\boldsymbol{v}, p\}\|_Y \|\{\boldsymbol{\xi}, \tau\}\|_Q} = \inf_{\{\boldsymbol{\xi},\tau\} \in Q} \sup_{\{\boldsymbol{v},p\} \in Y_N} \frac{\mathsf{A}(\{\boldsymbol{v}, p\}, \{\boldsymbol{\xi}, \tau\}; \boldsymbol{\mu})}{\|\{\boldsymbol{v}, p\}\|_Y \|\{\boldsymbol{\xi}, \tau\}\|_Y} \geq \tilde{\beta}_N^0.$$

Now, as in Lemma 1.4, we make use of the weakly coercivity of the bilinear form $\mathsf{A}(\cdot, \cdot; \boldsymbol{\mu})$. In fact, given $\{\boldsymbol{\xi}, \tau\} \in Q_N$,

$$\begin{aligned} \sup_{0 \neq \underline{\boldsymbol{x}} \in X_N} \frac{\mathcal{B}(\underline{\boldsymbol{x}}, \{\boldsymbol{\xi}, \tau\}; \boldsymbol{\mu})}{\|\underline{\boldsymbol{x}}\|_X} &= \sup_{0 \neq (\{\boldsymbol{v},p\},\boldsymbol{u}) \in Y_N \times U_N} \frac{\mathsf{A}(\{\boldsymbol{v}, p\}, \{\boldsymbol{\xi}, \tau\}; \boldsymbol{\mu}) - c(\boldsymbol{u}, \boldsymbol{\xi}; \boldsymbol{\mu})}{(\|\{\boldsymbol{v}, p\}\|_Y^2 + \|\boldsymbol{u}\|_U^2)^{1/2}} \\ &\underset{\boldsymbol{u}=0}{\geq} \sup_{0 \neq \{\boldsymbol{v},p\} \in Y_N} \frac{\mathsf{A}(\{\boldsymbol{v}, p\}, \{\boldsymbol{\xi}, \tau\}; \boldsymbol{\mu})}{\|\{\boldsymbol{v}, p\}\|_Y} \geq \tilde{\beta}_N^0 \|\{\boldsymbol{\xi}, \tau\}\|_Y = \tilde{\beta}_N^0 \|\{\boldsymbol{\xi}, \tau\}\|_Q. \end{aligned}$$ □

Proposition 5.1. *The RB saddle-point problem (5.2.4) has a unique solution for all $\boldsymbol{\mu} \in \mathcal{D}$.*

Proof. It suffices to check that the assumptions of Theorem A.4 hold. As already mentioned, the continuity properties of the bilinear and linear forms over the RB space are automatically inherited from the parents spaces (i.e. the FE spaces). The fulfillment of the inf-sup condition of the bilinear form $\mathcal{B}(\cdot,\cdot;\boldsymbol{\mu})$ has been proved in Lemma 5.1, while the fulfillment of the coercivity condition of the bilinear $\mathcal{A}(\cdot,\cdot;\boldsymbol{\mu})$ can be proved using the same arguments as in Lemma 1.3 and Lemma 1.4 (note however that those arguments apply correctly thanks to the supremizer enrichment and the choice $Y_N \equiv Q_N$). □

5.2.2 Algebraic formulation

Let us now briefly discuss the algebraic formulation associated to the enriched spaces introduced above; we simply merge the constructions made in Section 3.5.3 and Section 4.2.3. We build the aggregated pressure space M_N as given by (5.2.1), while for the velocity space $V_N^{\boldsymbol{\mu}}$, in order to build a $\boldsymbol{\mu}$-independent space, we rely on the construction given in Section 3.5.3. In particular we first construct a $\boldsymbol{\mu}$-independent state velocity space V_N^{state} using the recipe given in (3.5.10) and similarly a $\boldsymbol{\mu}$-independent adjoint velocity space V_N^{adj}, then we define the aggregated RB velocity space as

$$V_N = V_N^{state} \cup V_N^{adj}.$$

Note that V_N is made of $4N$ basis functions. Now we define the aggregated spaces for the state and adjoint variables

$$Z_N = V_N \times M_N, \qquad Y_N \equiv Q_N \equiv Z_N,$$

and we denote with $\{\boldsymbol{z}_j\}_{j=1}^{6N}$ the corresponding basis functions. Finally we construct the control space U_N as in (5.2.3) and the product space $X_N = Y_N \times U_N = \text{span}\{\underline{\boldsymbol{\sigma}}_j,\ j = 1,\dots,7N\}$ with suitably defined basis functions $\underline{\boldsymbol{\sigma}}_n$ (see Section 4.2.3).

Now we can express the RB state, adjoint and control solutions as

$$\underline{\boldsymbol{x}}_N(\boldsymbol{\mu}) = \sum_{j=1}^{7N} X_{Nj}(\boldsymbol{\mu})\underline{\boldsymbol{\sigma}}_j, \qquad \{\boldsymbol{w}_N(\boldsymbol{\mu}), q_N(\boldsymbol{\mu})\} = \sum_{j=1}^{6N} W_{Nj}(\boldsymbol{\mu})\boldsymbol{z}_j.$$

Hence, for a new parameter $\boldsymbol{\mu}$, the RB solution of the problem (5.2.4) can be written as a combination of basis functions with weights given by the following reduced basis linear system:

$$\begin{cases} \displaystyle\sum_{j=1}^{7N}\sum_{q=1}^{Q_a} \Theta_a^q(\boldsymbol{\mu}) A_{ij}^q\, X_{Nj}(\boldsymbol{\mu}) + \sum_{l=1}^{6N}\sum_{q=1}^{Q_b} \Theta_b^q(\boldsymbol{\mu}) B_{li}^q\, W_{Nl}(\boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \Theta_f^q(\boldsymbol{\mu}) F_i^q, & 1 \le i \le 7N, \\ \displaystyle\sum_{j=1}^{7N}\sum_{q=1}^{Q_b} \Theta_b^q(\boldsymbol{\mu}) B_{lj}^q\, X_{Nj}(\boldsymbol{\mu}) = \sum_{q=1}^{Q_g} \Theta_g^q(\boldsymbol{\mu}) G_l^q, & 1 \le l \le 6N, \end{cases} \tag{5.2.6}$$

where the submatrices A_N^q and B_N^q (we have omitted the subscript $_N$ in (4.2.9)) are given by

$$(A_N)_{ij}^q = \mathcal{A}^q(\underline{\boldsymbol{\sigma}}_j, \underline{\boldsymbol{\sigma}}_i), \qquad (B_N)_{li}^q = \mathcal{B}^q(\underline{\boldsymbol{\sigma}}_i, \boldsymbol{z}_l), \qquad 1 \le i,j \le 7N, \quad 1 \le l \le 6N,$$

and

$$(\boldsymbol{F}_N)_i^q = \langle F^q, \underline{\boldsymbol{\sigma}}_i \rangle, \qquad (\boldsymbol{G}_N)_l^q = \langle G^q, \boldsymbol{z}_l \rangle \qquad 1 \le i \le 7N, \quad 1 \le l \le 6N.$$

Finally, denoting with $A_N(\boldsymbol{\mu}) = \sum \Theta_a^q(\boldsymbol{\mu}) A_N^q$, $B_N(\boldsymbol{\mu}) = \sum \Theta_b^q(\boldsymbol{\mu}) B_N^q$, we can rewrite problem (5.2.6) as

$$\underbrace{\begin{pmatrix} A_N(\boldsymbol{\mu}) & B_N^T(\boldsymbol{\mu}) \\ B_N(\boldsymbol{\mu}) & 0 \end{pmatrix}}_{\mathcal{K}_N(\boldsymbol{\mu})} \begin{pmatrix} \mathbf{X}_N(\boldsymbol{\mu}) \\ \mathbf{W}_N(\boldsymbol{\mu}) \end{pmatrix} = \begin{pmatrix} \mathbf{F}_N(\boldsymbol{\mu}) \\ \mathbf{G}_N(\boldsymbol{\mu}) \end{pmatrix}. \tag{5.2.7}$$

The matrix $\mathcal{K}_N$ is still symmetric, with saddle-point structure and has dimension $13N \times 13N$. To keep under control the condition number of the matrix $\mathcal{K}_N$ we have adopted a suitable Gram-Schmidt (GS) orthonormalization procedure already introduced in Chapter 3.

5.2.3 Offline-Online procedure and sampling strategy

Thanks to the assumption of affine parameter dependence, we can decouple the formation of the matrix $\mathcal{K}_N(\boldsymbol{\mu})$ in two stages, the Offline and Online stages, that enable the efficient resolution of the system (5.2.7) for each new parameter $\boldsymbol{\mu}$. In particular:

1. in the Offline stage, performed only once, we first compute and store the basis function, and form the $\boldsymbol{\mu}$-independent matrices A_N^q, $1 \leq q \leq Q_a$, B_N^q, $1 \leq q \leq Q_b$ and the vectors F_N^q, $1 \leq q \leq Q_f$, G_N^q, $1 \leq q \leq Q_g$. The operation count depends on N, Q_a, Q_b, Q_f, Q_g and $\mathcal{N}$;

2. in the Online stage, performed for each new value $\boldsymbol{\mu}$, we use the precomputed matrices A_N^q, B_N^q and vectors F_N^q, G_N^q to assemble the (full) matrix $\mathcal{K}_N$ and the vectors $\mathbf{F}_N$, $\mathbf{G}_N$ appearing in (5.2.7), with

$$A_N(\boldsymbol{\mu}) = \sum_{q=1}^{Q_a} \Theta_a^q(\boldsymbol{\mu}) A_N^q, \qquad B_N(\boldsymbol{\mu}) = \sum_{q=1}^{Q_b} \Theta_b^q(\boldsymbol{\mu}) B_N^q,$$

$$\mathbf{F}_N(\boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \Theta_f^q(\boldsymbol{\mu}) F_N^q, \qquad \mathbf{G}_N(\boldsymbol{\mu}) = \sum_{q=1}^{Q_g} \Theta_f^q(\boldsymbol{\mu}) G_N^q;$$

   we then solve the resulting system to obtain $(\mathbf{X}_N, \mathbf{W}_N)$. The Online operation count depends on N, Q_a, Q_b, Q_f, Q_g but is independent of $\mathcal{N}$. In particular we need $O((Q_a+Q_b)(7N)^2)$ and $O((Q_f+Q_g)7N)$ operations to assemble matrices and vectors, and $O((13N)^3)$ operations to solve the RB linear system (5.2.7).

For the construction of the hierarchical Lagrange RB approximation spaces we rely again on the sampling strategy based on the greedy algorithm described in Chapter 3. In particular, in each iteration, given the parameter samples $S_N = \{\boldsymbol{\mu}^1, \ldots, \boldsymbol{\mu}^N\}$, the new sample point $\boldsymbol{\mu}^{N+1}$ to be added is such that

$$\boldsymbol{\mu}^{N+1} = \arg \max_{\boldsymbol{\mu} \in \Xi_{\text{train}}} \Delta_N(\boldsymbol{\mu}),$$

where $\Delta_N(\boldsymbol{\mu})$ is a rigorous, sharp and inexpensive a posteriori error bound for the error on the state, control and adjoint variables. The next section is devoted to the construction of such an error estimator.

5.3 A posteriori error estimation

We can straightforwardly generalize the a posteriori error analysis of Section 4.3.1 to the case considered here. In particular we can easily construct a rigorous, sharp and inexpensive (i.e. $\mathcal{N}$-independent) a posteriori error bound $\Delta_N(\boldsymbol{\mu})$ such that

$$\begin{aligned}\big(\|\boldsymbol{v}^{\mathcal{N}}(\boldsymbol{\mu}) - \boldsymbol{v}_N(\boldsymbol{\mu})\|_V^2 + \|p^{\mathcal{N}}(\boldsymbol{\mu}) - p_N(\boldsymbol{\mu})\|_M^2\big) + \|\boldsymbol{u}^{\mathcal{N}}(\boldsymbol{\mu}) - \boldsymbol{u}_N(\boldsymbol{\mu})\|_U^2 \\ + \big(\|\boldsymbol{w}^{\mathcal{N}}(\boldsymbol{\mu}) - \boldsymbol{w}_N(\boldsymbol{\mu})\|_V^2 + \|q^{\mathcal{N}}(\boldsymbol{\mu}) - q_N(\boldsymbol{\mu})\|_M^2\big) \leq \Delta_N^2(\boldsymbol{\mu}).\end{aligned}$$

Using the same ingredients, we can also construct a rigorous, sharp and inexpensive a posteriori error bound $\Delta_N^J(\boldsymbol{\mu})$ for the error on the cost functional, such that

$$|J(\boldsymbol{v}^{\mathcal{N}}(\boldsymbol{\mu}), p^{\mathcal{N}}(\boldsymbol{\mu}), \boldsymbol{u}^{\mathcal{N}}(\boldsymbol{\mu}); \boldsymbol{\mu}) - J(\boldsymbol{v}_N(\boldsymbol{\mu}), p_N(\boldsymbol{\mu}), \boldsymbol{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})| \leq \Delta_N^J(\boldsymbol{\mu}). \tag{5.3.1}$$

Needless to say, we can also provide the usual Offline-Online strategy that permits the efficient evaluation of the proposed estimators.

Bound for the solution: Babuška framework

The construction of the estimator $\Delta_N(\boldsymbol{\mu})$ will be carried out in the Babuška framework, as already done in Section 3.5.5 and in Section 4.3.1. In order to formulate the problem (5.1.9) in the standard form of weakly coercive problems (see Section A.1.2), it suffices to denote $\mathcal{X} = X \times Q$ and define the bilinear form $\mathsf{B}(\cdot,\cdot;\boldsymbol{\mu})\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ given by

$$\mathsf{B}(\mathsf{x},\mathsf{w};\boldsymbol{\mu}) := \mathcal{A}(\underline{\boldsymbol{x}}, \underline{\boldsymbol{\zeta}}; \boldsymbol{\mu}) + \mathcal{B}(\underline{\boldsymbol{\zeta}}, \{\boldsymbol{w}, q\}; \boldsymbol{\mu}) + \mathcal{B}(\underline{\boldsymbol{x}}, \{\boldsymbol{\xi}, \tau\}; \boldsymbol{\mu}), \tag{5.3.2}$$

and the linear continuous functional $\mathsf{F}\colon \mathcal{X} \to \mathbb{R}$

$$\mathsf{F}(\mathsf{w};\boldsymbol{\mu}) = \langle \underline{\boldsymbol{F}}(\boldsymbol{\mu}), \underline{\boldsymbol{\zeta}} \rangle + \langle \boldsymbol{G}(\boldsymbol{\mu}), \{\boldsymbol{\xi}, \tau\} \rangle, \tag{5.3.3}$$

where

$$\mathsf{x} = (\underline{\boldsymbol{x}}, \{\boldsymbol{w}, p\}) = (\{\boldsymbol{v}, p\}, \boldsymbol{u}, \{\boldsymbol{w}, q\}) \in \mathcal{X}, \qquad \mathsf{w} = (\underline{\boldsymbol{\zeta}}, \{\boldsymbol{\xi}, \tau\}) = (\{\boldsymbol{\varphi}, \pi\}, \boldsymbol{\lambda}, \{\boldsymbol{\xi}, \tau\}) \in \mathcal{X}.$$

Then we can formulate equivalently the problem (5.1.9) as: given $\boldsymbol{\mu} \in \mathcal{D}$,

$$\text{find } \mathsf{x} \in \mathcal{X} \text{ s.t:} \qquad \mathsf{B}(\mathsf{x},\mathsf{w};\boldsymbol{\mu}) = \mathsf{F}(\mathsf{w};\boldsymbol{\mu}) \qquad \forall \mathsf{w} \in \mathcal{X}. \tag{5.3.4}$$

Since the bilinear forms $\mathcal{A}(\cdot,\cdot;\boldsymbol{\mu})$ and $\mathcal{B}(\cdot,\cdot;\boldsymbol{\mu})$ satisfy the hypotheses of (Brezzi) Theorem A.2, it can be shown (see e.g. [21, 90, 35]) that the the compound form $\mathsf{B}(\cdot,\cdot;\boldsymbol{\mu})$ is continuous and weakly coercive. Similarly, the FE and RB approximations are well-posed, in particular we have the following estimate for the FE approximation $\mathsf{x}^{\mathcal{N}}(\boldsymbol{\mu}) \in \mathcal{X}^{\mathcal{N}}$:

$$\|\mathsf{x}^{\mathcal{N}}(\boldsymbol{\mu})\|_{\mathcal{X}} \leq \frac{1}{\hat{\beta}^{\mathcal{N}}(\boldsymbol{\mu})} \|\mathsf{F}(\cdot;\boldsymbol{\mu})\|_{\mathcal{X}'}, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{5.3.5}$$

where $\mathcal{X}^{\mathcal{N}} = X^{\mathcal{N}} \times Q^{\mathcal{N}}$ and

$$\hat{\beta}^{\mathcal{N}}(\boldsymbol{\mu}) := \inf_{\mathsf{w} \in \mathcal{X}^{\mathcal{N}}} \sup_{\mathsf{x} \in \mathcal{X}^{\mathcal{N}}} \frac{\mathsf{B}(\mathsf{x},\mathsf{w};\boldsymbol{\mu})}{\|\mathsf{x}\|_{\mathcal{X}} \|\mathsf{w}\|_{\mathcal{X}}} > 0, \qquad \forall \boldsymbol{\mu} \in \mathcal{D}.$$

For the construction of the a posteriori error estimator we exploit the stability estimate (5.3.5) using the usual two ingredients: an effective calculation of a lower bound for the

Babuška inf-sup constant $\hat{\beta}^{\mathcal{N}}(\boldsymbol{\mu})$ and the (standard) calculation of the dual norm of the residual. Once again the calculation of $\hat{\beta}_{\mathrm{LB}}(\boldsymbol{\mu})$ will be carried out using the *Natural Norm Successive Constraint Method* (see Section 3.5.5).
It suffices to define the *global* error between the FE and the RB approximations

$$\mathsf{e}(\boldsymbol{\mu}) = \mathsf{x}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathsf{x}_N(\boldsymbol{\mu}),$$

and the *global* residual

$$\mathsf{r}(\mathsf{w};\boldsymbol{\mu}) = \mathsf{F}(\mathsf{w};\boldsymbol{\mu}) - \mathsf{B}(\mathsf{x}_N, \mathsf{w};\boldsymbol{\mu}) \qquad \forall \mathsf{w} \in \mathcal{X}^{\mathcal{N}}.$$

Then, as in Section 4.3.1, by using the stability estimate (5.3.5) and exploiting the lower bound for the inf-sup constant we obtain the following residual-based estimation

$$\|\mathsf{e}(\boldsymbol{\mu})\|_{\mathcal{X}} \leq \frac{1}{\hat{\beta}_{\mathrm{LB}}(\boldsymbol{\mu})} \|\mathsf{r}(\cdot;\boldsymbol{\mu})\|_{\mathcal{X}'} := \Delta_N(\boldsymbol{\mu}), \qquad \forall \boldsymbol{\mu} \in \mathcal{D}. \tag{5.3.6}$$

Offline-Online procedure

The dual norm of the residuals can be evaluated through its Riesz representation $\hat{\mathsf{e}}(\boldsymbol{\mu}) \in X^{\mathcal{N}}$:

$$\|\mathsf{r}(\cdot;\boldsymbol{\mu})\|_{\mathcal{X}'} = \sup_{\mathsf{w} \in \mathcal{X}^{\mathcal{N}}} \frac{\mathsf{r}(\mathsf{w};\boldsymbol{\mu})}{\|\mathsf{w}\|_{\mathcal{X}}} = \|\hat{\mathsf{e}}(\boldsymbol{\mu})\|_{\mathcal{X}},$$

where $\hat{\mathsf{e}}(\boldsymbol{\mu})$ satisfies $(\hat{\mathsf{e}}(\boldsymbol{\mu}), \mathsf{w})_{\mathcal{X}} = \mathsf{r}(\mathsf{w};\boldsymbol{\mu})$, $\forall \mathsf{w} \in \mathcal{X}^{\mathcal{N}}$. From the affine decompositions of the linear and bilinear forms (5.1.11) (5.1.12) we can write equivalently

$$\mathsf{B}(\mathsf{x},\mathsf{w};\boldsymbol{\mu}) = \sum_{q=1}^{Q_a+2Q_b} \Theta_B^q(\boldsymbol{\mu}) \mathsf{B}^q(\mathsf{x},\mathsf{w}), \qquad \mathsf{F}(\mathsf{w};\boldsymbol{\mu}) = \sum_{q=1}^{Q_f+Q_g} \Theta_F^q(\boldsymbol{\mu}) \mathsf{F}^q(\mathsf{w};\boldsymbol{\mu}), \tag{5.3.7}$$

where $Q_B = Q_a + 2Q_b$ and $Q_F = Q_f + Q_g$. In this way, recalling that

$$\mathsf{x}_N(\boldsymbol{\mu}) = \left(\underline{\boldsymbol{x}}_N(\boldsymbol{\mu}), \{\boldsymbol{w}_N(\boldsymbol{\mu}), q_N(\boldsymbol{\mu})\}\right) \in \mathbb{R}^{13N}$$

denotes the global vector of the RB components, the residual can be expressed as

$$\begin{aligned} \mathsf{r}(\mathsf{w};\boldsymbol{\mu}) &= \mathsf{F}(\mathsf{w};\boldsymbol{\mu}) - \mathsf{B}\left(\sum_{n=1}^{13N} \mathsf{x}_{Nn}(\boldsymbol{\mu}) \Phi_n, \mathsf{w};\boldsymbol{\mu}\right) \\ &= \sum_{q=1}^{Q_F} \Theta_F^q(\boldsymbol{\mu}) \mathsf{F}^q(\mathsf{w}) - \sum_{n=1}^{13N} \mathsf{x}_{Nn}(\boldsymbol{\mu}) \sum_{q=1}^{Q_B} \Theta_B^q(\boldsymbol{\mu}) \mathsf{B}^q(\Phi_n, \mathsf{w}), \end{aligned} \tag{5.3.8}$$

where

$$\Phi_n = (\underline{\boldsymbol{\sigma}}_n, 0), \quad 1 \leq n \leq 7N, \qquad \Phi_n = (0, \boldsymbol{z}_n), \quad 7N+1 \leq n \leq 13N.$$

Proceeding as in Section 4.3.2 we finally obtain

$$\begin{aligned} \|\hat{\mathsf{e}}(\boldsymbol{\mu})\|_{\mathcal{X}}^2 = \sum_{q=1}^{Q_F} \sum_{q'=1}^{Q_F} \Theta_F^q(\boldsymbol{\mu}) \Theta_F^{q'}(\boldsymbol{\mu}) (\mathcal{F}^q, \mathcal{F}^{q'})_{\mathcal{X}} + \sum_{q=1}^{Q_B} \sum_{n=1}^{13N} \Theta_B^q(\boldsymbol{\mu}) \mathsf{x}_{Nn}(\boldsymbol{\mu}) \Bigg\{ \\ 2 \sum_{q'=1}^{Q_F} \Theta_F^{q'}(\boldsymbol{\mu}) (\mathcal{F}^{q'}, \mathcal{L}_n^q)_{\mathcal{X}} + \sum_{q'=1}^{Q_B} \sum_{n'=1}^{13N} \Theta_B^{q'}(\boldsymbol{\mu}) \mathsf{x}_{Nn'}(\boldsymbol{\mu}) (\mathcal{L}_n^q, \mathcal{L}_n^{q'})_{\mathcal{X}} \Bigg\} \end{aligned} \tag{5.3.9}$$

from which we can calculate the dual norm of the residual. We can thus exploit the usual Offline-Online decomposition, we limit to note that the Online operation count is $O((13N)^2 Q_B^2 + 26N Q_B Q_f + 13N Q_F^2)$, independent of $\mathcal{N}$.

A posteriori error bound for the cost functional

The error on cost functional evaluated with respect to the FE and RB approximations will be denoted with

$$\mathcal{J}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathcal{J}_N(\boldsymbol{\mu}) = J(\boldsymbol{v}^{\mathcal{N}}(\boldsymbol{\mu}), p^{\mathcal{N}}(\boldsymbol{\mu}), \boldsymbol{u}^{\mathcal{N}}(\boldsymbol{\mu}); \boldsymbol{\mu}) - J(\boldsymbol{v}_N(\boldsymbol{\mu}), p_N(\boldsymbol{\mu}), \boldsymbol{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu}).$$

Proceeding exactly as in Section 4.3.3 we obtain the following bound:

$$|\mathcal{J}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathcal{J}_N(\boldsymbol{\mu})| \leq \frac{1}{2}\|\mathsf{r}(\cdot;\boldsymbol{\mu})\|_{\mathcal{X}'}\|\mathsf{e}(\boldsymbol{\mu})\|_{\mathcal{X}} \leq \frac{1}{2}\frac{\|\hat{\mathsf{e}}(\boldsymbol{\mu})\|^2_{\mathcal{X}}}{\hat{\beta}_{\mathrm{LB}}(\boldsymbol{\mu})} := \Delta_N^J(\boldsymbol{\mu}). \tag{5.3.10}$$

Note that the error estimator $\Delta_N^J(\boldsymbol{\mu})$ does not need any additional ingredients than those already discussed: the efficient computation of the dual norm of the residual and the calculation of a lower bound for the Babuška inf-sup constant.

5.4 Preliminary numerical tests

In this section we present two (very simple) numerical examples to test the stability and convergence properties, as well as the computational load of the proposed methodology. In particular we consider as state problem the Couette flow introduced in Section 3.5.6. Since the geometry is parametrized we proceed as in Chapter 3: we firstly define an original problem posed over a parameters dependent domain, then we trace back the problem to a reference domain through the affine mappings (see Section 3.4.1) in order to recover the formulation (5.1.9).

As in Section 4.4, the implementation of the method has been carried out in the MATLAB environment using an enhanced version[1] of the `rbMIT` library [44, 61], for the FE assembling stage we have exploited the `MLife` library [82].

5.4.1 Test 1: distributed optimal control for a Couette flow

The original domain is the pipe $\Omega_o(\boldsymbol{\mu}) = [0,1] \times [0,\mu_1]$ shown in Figure 5.1. We consider two parameters $\boldsymbol{\mu} = (\mu_1, \mu_2)$, being μ_1 a geometrical parameter (the channel length) and μ_2 a physical parameter in the forcing term of the state equation. The parameter domain is given by $\mathcal{D} = [0.5, 2] \times [0.5, 1.5]$.

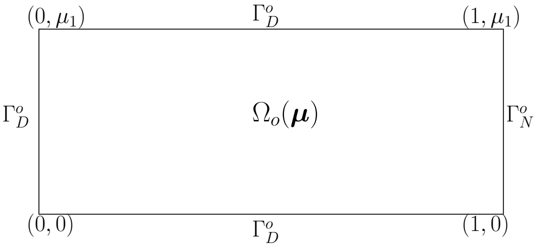


Figure 5.1: Original domain for Test 1.

[1] Co-developed at CMCS (Chair of Modelling and Scientific Computing), EPFL, based on the official released version of `rbMIT`.

We consider the following optimal control problem

$$\min_{\mathbf{u}_o\in U_o} J(\boldsymbol{v}_o(\boldsymbol{\mu}),p_o(\boldsymbol{\mu}),\boldsymbol{u}_o(\boldsymbol{\mu})) = \frac{1}{2}\|v_{o1}(\boldsymbol{\mu})-x_{o2}\|^2_{L^2(\Omega_o)} + \frac{\alpha}{2}\|\boldsymbol{u}_o(\boldsymbol{\mu})\|^2_{L^2(\Omega_o)},$$
$$\text{s.t.}\ \begin{cases} -\nu\Delta\boldsymbol{v}_o + \nabla p_o = \boldsymbol{f}_o(\boldsymbol{\mu}) + \boldsymbol{u}_o & \text{in } \Omega_o(\boldsymbol{\mu}) \\ \operatorname{div}\boldsymbol{v}_o = 0 & \text{in } \Omega_o(\boldsymbol{\mu}) \\ v_{o1} = x_{o2},\ v_{o2}=0 & \text{on } \Gamma_D^o(\boldsymbol{\mu}) \\ -p_o\boldsymbol{n}_{o1} + \nu\dfrac{\partial v_{o1}}{\partial \boldsymbol{n}_{o1}} = 0,\ v_{o2}=0 & \text{on } \Gamma_N^o(\boldsymbol{\mu}), \end{cases} \tag{5.4.1}$$

where the forcing term is given by $\boldsymbol{f}_o(\boldsymbol{\mu}) = (0,-\mu_2)$ and we observe only the first component of the velocity with observation function (desired state) equal to x_{o2}. We define the velocity and pressure spaces respectively as $V_o = [H^1_{\Gamma_D}(\Omega_o)]^2$ and $M_o = L^2(\Omega_o)$; moreover we define the state space $Y_o = V_o \times M_o$, the adjoint space $Q_o \equiv Y_o$ and the control space $U_o = [L^2(\Omega_o)]^2$.

We also introduce a lift function $R_{\boldsymbol{g}} \in [H^1(\Omega_o)]^2$ such that $\tilde{\boldsymbol{v}}_o = \boldsymbol{v}_o - R_{\boldsymbol{g}} \in V_o$; for the sake of simplicity, we still denote $\tilde{\boldsymbol{v}}_o$ with $\boldsymbol{v}_o$ in the sequel. The weak formulation of the state equation is given by

$$\begin{cases} a_o(\boldsymbol{v}_o,\boldsymbol{\xi}) + b_o(\boldsymbol{\xi},p_o) = \langle F_o(\boldsymbol{\mu}),\boldsymbol{\xi}\rangle + c_o(\boldsymbol{u}_o,\boldsymbol{\xi}), & \forall\boldsymbol{\xi}\in V_o, \\ b_o(\boldsymbol{v}_o,\tau) = \langle G_o,\tau\rangle, & \forall\tau\in M_o, \end{cases} \tag{5.4.2}$$

where the bilinear form $a_o\colon V_o\times V_o\to\mathbb{R}$ and $b_o\colon V_o\times M_o\to\mathbb{R}$ are defined as in Section 3.5.6, while the bilinear form $c_o\colon U_o\times V_o\to\mathbb{R}$ is given by

$$c_o(\boldsymbol{u}_o,\boldsymbol{\xi}) = \int_{\Omega_o} \boldsymbol{u}_o\cdot\boldsymbol{\xi}\,d\Omega_o.$$

Now we define the compound bilinear forms $\mathsf{A}_o(\cdot,\cdot;\boldsymbol{\mu}) : Y_o\times Y_o\to\mathbb{R}$ and $\mathsf{C}_o(\cdot,\cdot;\boldsymbol{\mu}) : U_o\times Q_o\to\mathbb{R}$ respectively as

$$\mathsf{A}_o(\{\boldsymbol{v}_o,p_o\},\{\boldsymbol{\xi},\tau\}) = a_o(\boldsymbol{v}_o,\boldsymbol{\xi}) + b_o(\boldsymbol{\xi},p_o) + b_o(\boldsymbol{v}_o,\tau), \qquad \mathsf{C}_o(\boldsymbol{u}_o,\{\boldsymbol{\xi},\tau\}) = c_o(\boldsymbol{u}_o,\boldsymbol{\xi}),$$

while the linear functional $\boldsymbol{G}_o(\boldsymbol{\mu})\in Q_o'$ is given by $\langle\boldsymbol{G}_o(\boldsymbol{\mu}),\{\boldsymbol{\xi},\tau\}\rangle = \langle F_o(\boldsymbol{\mu}),\boldsymbol{\xi}\rangle + \langle G_o,\tau\rangle$. Finally we define the bilinear forms $m_o(\cdot,\cdot) : V_o\times V_o\to\mathbb{R}$ and $n_o(\cdot,\cdot) : U_o\times U_o\to\mathbb{R}$ as

$$m_o(\boldsymbol{v}_o,\boldsymbol{\varphi}) = \int_{\Omega_o} v_{o1}\varphi_1\,d\Omega_o, \qquad n_o(\boldsymbol{u}_o,\boldsymbol{\lambda}) = \int_{\Omega_o} \alpha\boldsymbol{u}_o\cdot\boldsymbol{\lambda}\,d\Omega_o.$$

In order to formulate the optimal control problem in the form (5.1.8), let $X_o = Y_o\times U_o$, $\underline{\boldsymbol{x}}_o = (\{\boldsymbol{v}_o,p_o\},\boldsymbol{u}_o)\in X_o$, $\underline{\boldsymbol{\zeta}} = (\{\boldsymbol{\varphi},\pi\},\boldsymbol{\lambda})\in X_o$, $\{\boldsymbol{\xi},\tau\}\in Q_o$ and define the bilinear forms

$$\begin{aligned} \mathcal{A}_o(\underline{\boldsymbol{x}}_o,\underline{\boldsymbol{\zeta}}) &= m_o(\boldsymbol{v}_o,\boldsymbol{\varphi}) + n_o(\boldsymbol{u}_o,\boldsymbol{\lambda}), \\ \mathcal{B}_o(\underline{\boldsymbol{x}}_o,\{\boldsymbol{\xi},\tau\}) &= \mathsf{A}_o(\{\boldsymbol{v}_o,p_o\},\{\boldsymbol{\xi},\tau\}) - \mathsf{C}_o(\boldsymbol{u}_o,\{\boldsymbol{\xi},\tau\}), \end{aligned}$$

and the linear functional

$$\langle\underline{\boldsymbol{F}}_o,\underline{\boldsymbol{\zeta}}\rangle = \int_{\Omega_o} x_{o2}\varphi_2\,d\Omega_o.$$

We denote with $\Omega = \Omega_o(\boldsymbol{\mu}_{\text{ref}})$ the reference domain, with the choice $\boldsymbol{\mu}_{\text{ref}} = (1,1)$; the affine geometrical mapping is trivial (see Section 3.5.6). By tracing the problem back to the reference domain we obtain the parametrized formulation (5.1.8) and the equivalent saddle-point problem (5.1.9), where the affine decompositions (5.1.11) (5.1.12) holds with $Q_a = 1$, $Q_b = 4$, $Q_f = 1$, $Q_g = 5$.

We fixed $\alpha = 0.008$, $\nu = 0.1$ and we used $\mathbb{P}^2 - \mathbb{P}^1$ Taylor-Hood finite elements for the FE approximation of the velocity and pressure variables, the dimension of the global FE space $\mathcal{X}^{\mathcal{N}}$ used is $\mathcal{N} = 17439$. In Figure 5.2 we report a representative solution, the action of the control becomes quite evident comparing the velocity field to the uncontrolled one given in Figure 3.5.

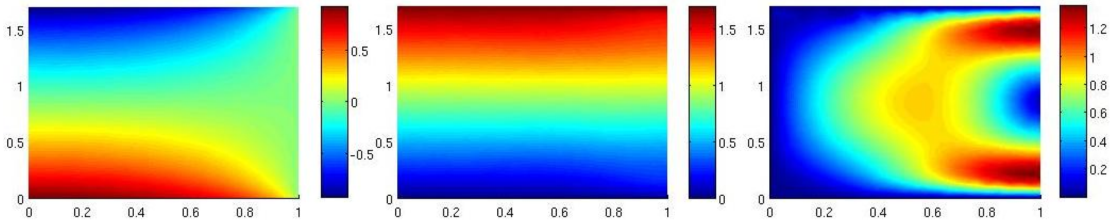
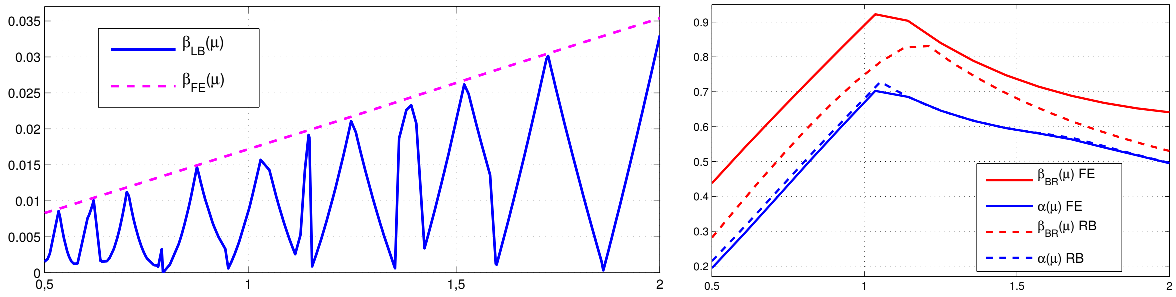


Figure 5.2: Test 1: representative solutions for $\boldsymbol{\mu} = (1.7, 1.5)$; pressure on the left, velocity in the middle, control on the right.

With a fixed tolerance $\varepsilon_{tol} = 10^{-3}$, $N_{max} = 15$ basis functions have been selected, thus resulting in a RB linear system of dimension 225×225. In Figure 5.3a we show the lower bound for the Babuška inf-sup constant $\hat{\beta}^{\mathcal{N}}(\boldsymbol{\mu})$ obtained using the natural norm SCM algorithm with a tolerance $\varepsilon_{\mathrm{SCM}} = 0.85$ and a uniform train sample of size $n_{\mathrm{train,SCM}} = 1000$; in the Offline stage SCM requires in this case the solution of $68 + 2(Q_a + 2Q_b)$ eigenproblems of size $\mathcal{N}$.



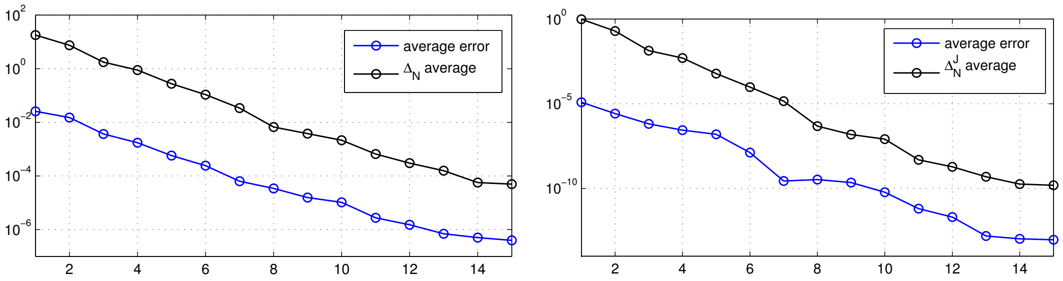
(a) Lower bound for the Babuška inf-sup constant $\hat{\beta}(\boldsymbol{\mu})$ as a function of the geometrical parameter μ_1 ($\hat{\beta}_{\mathrm{LB}}(\boldsymbol{\mu})$ in blue, $\hat{\beta}^{\mathcal{N}}(\boldsymbol{\mu})$ in magenta).

(b) Comparison of Brezzi inf-sup constant $\beta(\boldsymbol{\mu})$ and Babuška in-sup constant of the state (Stokes) equation $\tilde{\beta}(\boldsymbol{\mu})$ for the FE and RB approximations as functions of μ_1 (in the legend $\tilde{\beta}$ is denoted with α).

Figure 5.3: Test 1: stability properties (note that μ_2 does not affect the value of the stability factors since it appears only in the right hand side of the saddle-point problem).

In Figure 5.3b we compare the Brezzi inf-sup constants $\beta^{\mathcal{N}}(\boldsymbol{\mu})$ and $\beta_N(\boldsymbol{\mu})$ and the Babuška inf-sup constants $\tilde{\beta}^{\mathcal{N}}(\boldsymbol{\mu})$ and $\tilde{\beta}_N(\boldsymbol{\mu})$ of the bilinear form $\mathsf{A}(\cdot,\cdot;\boldsymbol{\mu})$. Finally in Figure 5.4 we compare the a posteriori error bound $\Delta_N(\boldsymbol{\mu})$ with the true error $\|\mathsf{x}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathsf{x}_N(\boldsymbol{\mu})\|_{\mathcal{X}}$ and the a posteriori error bound $\Delta_N^J(\boldsymbol{\mu})$ with the true error on the cost functional $|\mathcal{J}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathcal{J}_N(\boldsymbol{\mu})|$.

As regards the computational performances, the Offline computational time is equal to $t_{RB}^{offline} = 7820$s (mostly spent performing the SCM algorithm for the computation of the



(a) Average computed error and bound between the *truth* FE solution and the RB approximation, for $N = 1, \cdots, N_{max}$.

(b) Average computed errors and bounds $\Delta_N^J(\boldsymbol{\mu})$ between $J^{\mathcal{N}}(\boldsymbol{\mu})$ and $J_N(\boldsymbol{\mu})$, for $N = 1, \cdots, N_{max}$.

Figure 5.4: Test 1: a posteriori error bounds. Here Ξ_{train} is a sample of size $n_{train} = 1000$ and $N_{max} = 15$.

lower bound of the inf-sup constant $\hat{\beta}^{\mathcal{N}}(\boldsymbol{\mu})$), the (average) Online solution time for the control problem is about $t_{RB}^{online} = 0.1$s comprehensive of the evaluation of the a posteriori error estimation while the evaluation time for the FE solution of the control problem is equal to about $t_{FE} = 16.1$s, thus resulting in a speedup of two orders of magnitude.

5.4.2 Test 2: boundary optimal control for a Couette flow

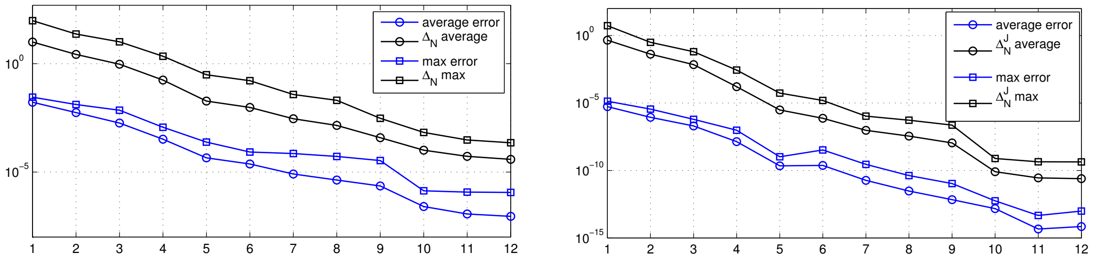
We deal again with the same problem as before, but instead of a distributed control we now consider a boundary control acting on the Neumann (outflow) boundary. The spatial domain as well as the parameter domain are the same as in the previous example. We thus consider the following boundary optimal control problem

$$\min_{\mathbf{u}_o \in U_o} J(\boldsymbol{v}_o(\boldsymbol{\mu}), p_o(\boldsymbol{\mu}), u_o(\boldsymbol{\mu})) = \frac{1}{2}\|v_{o1}(\boldsymbol{\mu}) - x_{o2}\|^2_{L^2(\Omega_o)} + \frac{\alpha}{2}\|u_o(\boldsymbol{\mu})\|^2_{L^2(\Gamma^o_N)},$$
$$\text{s.t.} \quad \begin{cases} -\nu\Delta\boldsymbol{v}_o + \nabla p_o = \boldsymbol{f}_o(\boldsymbol{\mu}) & \text{in } \Omega_o(\boldsymbol{\mu}) \\ \operatorname{div} \boldsymbol{v}_o = 0 & \text{in } \Omega_o(\boldsymbol{\mu}) \\ v_{o1} = x_{o2}, \;\; v_{o2} = 0 & \text{on } \Gamma^o_D(\boldsymbol{\mu}) \\ -p_o\boldsymbol{n}_{o1} + \nu\dfrac{\partial v_{o1}}{\partial \boldsymbol{n}_{o1}} = u_o, \;\; v_{o2} = 0 & \text{on } \Gamma^o_N(\boldsymbol{\mu}), \end{cases} \tag{5.4.3}$$

where the forcing term is given by $\boldsymbol{f}_o(\boldsymbol{\mu}) = (0, -\mu_2)$ and we observe only the first component of the velocity with observation function (desired state) equal to x_{o2}.
We can now proceed exactly as in the previous example, the only differences are the definition of the control space $U_o = L^2(\Gamma^o_N)$ and the expressions of the bilinear forms $c_o(\cdot,\cdot)\colon U_o \times V_o \to \mathbb{R}$ and $n_o(\cdot,\cdot) : U_o \times U_o \to \mathbb{R}$, now given by

$$c_o(u_o, \boldsymbol{\xi}) = \int_{\Gamma^o_N} u_o\xi_1 \, d\Gamma_o, \qquad n_o(u_o, \boldsymbol{\lambda}) = \int_{\Gamma^o_N} \alpha u_o \lambda_1 \, d\Gamma_o.$$

The performances and properties of the RB procedure are very similar to those of the previous example; with a fixed tolerance $\varepsilon_{tol} = 10^{-3}$, $N_{max} = 12$ basis functions have been selected, thus resulting in a RB linear system of dimension 156 × 156. In Figure 5.5 we compare the a posteriori error bound $\Delta_N(\boldsymbol{\mu})$ with the true error $\|\mathrm{x}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathrm{x}_N(\boldsymbol{\mu})\|_{\mathcal{X}}$ and the a posteriori error bound $\Delta_N^J(\boldsymbol{\mu})$ with the true error on the cost functional $|\mathcal{J}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathcal{J}_N(\boldsymbol{\mu})|$.



(a) Average and max computed errors and bounds between the *truth* FE solution and the RB approximation, for $N = 1, \cdots, N_{max}$.

(b) Average computed errors and bounds $\Delta_N^J(\boldsymbol{\mu})$ between $J^{\mathcal{N}}(\boldsymbol{\mu})$ and $J_N(\boldsymbol{\mu})$, for $N = 1, \cdots, N_{max}$.

Figure 5.5: Test 2: a posteriori error bounds. Here Ξ_{train} is a sample of size $n_{train} = 1000$ and $N_{max} = 12$.

5.5 Application to an inverse problem in haemodynamics: data assimilation for an arterial bifurcation

As a conclusive application we consider an inverse boundary problem in hemodynamics, inspired by the recent works [20, 62]. We consider a simplified model of an arterial bifurcation, the parametrized computational domain (see Figure 5.6 on the left) features an inflow boundary Γ_{in}, two outflow boundaries Γ_C and the physical wall of the vessel Γ_D. The variables of interest are the velocity $\boldsymbol{v}$ and the pressure p, supposed to obey the steady Stokes equations (as an approximation of the incompressible Navier-Stokes equations). We suppose to have a measured velocity profile on the red section in Figure 5.6, but not the Neumann flux on Γ_C that will be our control variable. Starting from the velocity measures we want to find the control variable in order to retrieve the velocity and pressure fields in the whole domain. We consider several possible parameters: geometrical parameters $\boldsymbol{\mu}_{geom}$ (e.g. the length of each of the two branches, the angle of the bifurcation etc.), a parametrized ($\boldsymbol{\mu}_{meas}$) measured velocity profile and a parametrized inflow velocity profile $\boldsymbol{g}(\boldsymbol{\mu}_{in})$. Figure 5.6 shows the (idealized) *real-time* data assimilation procedure obtainable via the RB method.

The problem we are going to describe represents an extremely simplification of a realistic problem: we consider a 2D trivial geometry, we deal with a steady problem also introducing several simplifications on the constitutive laws of the modelled system; moreover we suppose to have at our disposal a measured velocity profile approximable with a simple parametrized analytical profile. Note however that some of the simplifications introduced can be relaxed using more advanced existing techniques, for example we could consider more realistic geometry relying on non-affine geometrical mappings and empirical interpolation method [55, 56].

5.5.1 Mathematical modeling

The parametrized original domain $\Omega(\boldsymbol{\mu})$ is shown in Figure 5.7, in particular after having fixed the length (4 cm) and the diameter (2 cm) of the large vessel (quite unrealistic but it allows to simplify the parametrization), we have considered six geometrical parameters $\boldsymbol{\mu}_{geom}$: μ_1 (resp. $2 - \mu_1$) represents the height of the upper (resp. lower) branch, μ_2 and μ_3 are the angles of the branches with respect to the horizontal line, μ_5 and μ_6 are the length of the two branches while μ_4 is the distance of the observation line Γ_{obs}^o from the bifurcation.

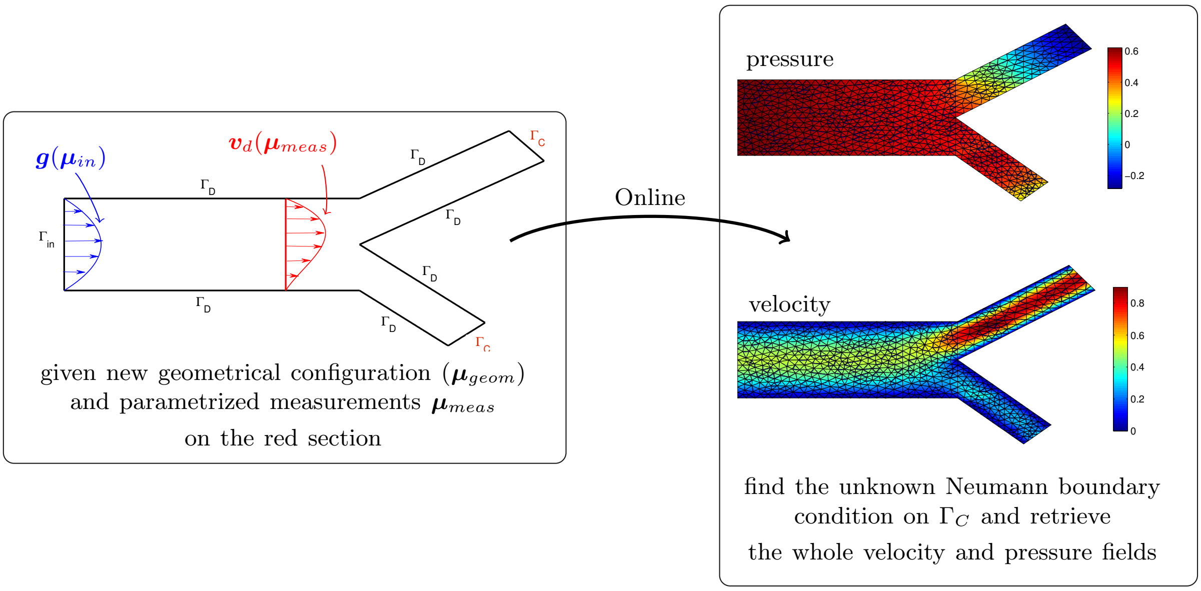
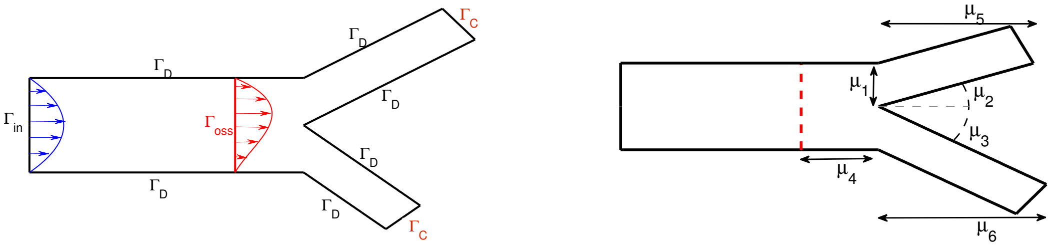

$g(\mu_{in})$
$v_d(\mu_{meas})$
Γ_D
Γ_{in}
Γ_C
given new geometrical configuration (μ_{geom})
and parametrized measurements μ_{meas}
on the red section
Online
pressure
0.6
0.4
0.2
0
−0.2
velocity
0.8
0.6
0.4
0.2
0
find the unknown Neumann boundary
condition on Γ_C and retrieve
the whole velocity and pressure fields

Figure 5.6: An (idealized) example of inverse boundary problem in haemodynamics. Given a geometrical configuration and some velocity measurements on some sections of the domain (both obtainable via medical image and data assimilation devices, e.g. MRI), we want to retrieve the whole pressure and velocity fields in order to detect possible pathologies, e.g. occlusions or flow disturbance in arterial bifurcations.

The state velocity and pressure variables $\{\boldsymbol{v}_o, p_o\}$ satisfy the following Stokes problem in the original domain $\Omega_o(\boldsymbol{\mu})$:

$$\begin{cases} -\nu\Delta\boldsymbol{v}_o + \nabla p_o = 0 & \text{in } \Omega_o(\boldsymbol{\mu}), \\ \operatorname{div}\boldsymbol{v}_o = 0 & \text{in } \Omega_o(\boldsymbol{\mu}), \\ \boldsymbol{v} = 0 & \text{on } \Gamma_D^o(\boldsymbol{\mu}), \\ \boldsymbol{v}_o = \boldsymbol{g}(\boldsymbol{\mu}_{in}) & \text{on } \Gamma_{in}^o(\boldsymbol{\mu}), \\ -p_o\boldsymbol{n}_o + \nu\dfrac{\partial\boldsymbol{v}_o}{\partial\boldsymbol{n}_o} = \boldsymbol{u}_o & \text{on } \Gamma_C^o(\boldsymbol{\mu}), \end{cases} \tag{5.5.1}$$

where $\boldsymbol{u}_o$ is the control variable; the inlet velocity profile $\boldsymbol{g}(\boldsymbol{\mu}_{in})$ is given by the following



(a) Boundary conditions: no-slip conditions on $\Gamma_D(\boldsymbol{\mu})$, Poiseuille velocity profile $\boldsymbol{g}(\boldsymbol{\mu}_{in})$ on Γ_{in}, unknown Neumann flux on the outflow sections.

(b) Parametrization of the original domain.

Figure 5.7: Original domain $\Omega_o(\boldsymbol{\mu})$ of the arterial bifurcation.

Poiseuille parabolic profile

$$\boldsymbol{g}(\boldsymbol{\mu}_{in}) = \begin{pmatrix} 10\mu_8(x_{o2}+1)(1-x_{o2}) \\ 0 \end{pmatrix},$$

with a (parametrized) peak velocity equal to $\tilde{v} = 10\mu_8$ cm s^{-1}. The kinematic viscosity is $\nu = 0.04$ cm^{2}s^{-1} [55], thus resulting in a Reynolds number Re$= \tilde{v}l/\nu$ of about 500 (taking l as the diameter of the large vessel and fixing $\mu_8 = 1$).
Then we consider the following (parametrized) cost functional to be minimized

$$J(\boldsymbol{v}_o(\boldsymbol{\mu}), p_o(\boldsymbol{\mu}), \boldsymbol{u}_o(\boldsymbol{\mu}); \boldsymbol{\mu}) = \frac{1}{2}\int_{\Gamma^o_{obs}} |\boldsymbol{v}_o(\boldsymbol{\mu}) - \boldsymbol{v}_d(\boldsymbol{\mu})|^2 \, d\Gamma_o \\ + \frac{\alpha_1}{2}\int_{\Gamma^o_C} |\nabla \boldsymbol{u}_o(\boldsymbol{\mu})\mathbf{t}_o|^2 d\Gamma_o + \frac{\alpha_2}{2}\int_{\Gamma^o_C} |\boldsymbol{u}_o(\boldsymbol{\mu})|^2 d\Gamma_o, \quad (5.5.2)$$

where

$$\int_{\Gamma^o_C} |\nabla \boldsymbol{u}_o(\boldsymbol{\mu})\mathbf{t}_o|^2 d\Gamma_o = \int_{\Gamma^o_C} \left(\frac{\partial u_{o1}(\boldsymbol{\mu})}{\partial \gamma_o}\right)^2 d\Gamma_o + \int_{\Gamma^o_C} \left(\frac{\partial u_{o2}(\boldsymbol{\mu})}{\partial \gamma_o}\right)^2 d\Gamma_o, \quad (5.5.3)$$

being $\mathbf{t}_o$ the tangential unit vector to the boundary Γ^o_C and γ_o the curvilinear abscissa of the boundary Γ^o_C. Note that the functional (5.5.2) contains two penalization terms, the fist one penalizing rapid variations of the control variable along the control boundary, the second one penalizing high values of the control variable; in the following we shall take $\alpha_2 = 0.1\alpha_1$, (i.e. we prefer to penalize rapid variations rather than high values, see also [20]) with $\alpha_1 = 10^{-3}$.
The parametrized observation velocity profile on Γ^o_{obs} is given by

$$\boldsymbol{v}_d(\boldsymbol{\mu}) = \begin{pmatrix} \mu_8\big(\mu_7\eta_1(x_{o2}) + (1-\mu_7)\eta_2(x_{o2})\big) \\ 0 \end{pmatrix},$$

where the functions $\eta_1(z) = 10(z^3 - z^2 - z + 1)$ and $\eta_2(z) = 10(-z^3 - z^2 + z + 1)$ are shown in Figure 5.8. We are thus considering a parametrized positive horizontal velocity profile (which guarantees mass conservation for all values of μ_7 and μ_8) and a null vertical velocity profile. The latter condition is quite unphysical, since when $\mu_7 \neq 0.5$ we reasonably expect a non-null vertical velocity profile on Γ_{obs}, however this condition permits us to avoid flow inversions that could arise in one of the two branches if we do not observe at all the vertical component of the velocity.
Finally, the parameter domain is given by

$$\mathcal{D} = \{\boldsymbol{\mu} = (\mu_1, \dots, \mu_8) \in \mathbb{R}^8 \, : \, \mu_i \in [\mu_{m,i}, \mu_{M,i}] \quad \forall i = 1, \dots, 8, \}$$

where

$$\boldsymbol{\mu}_m = \begin{pmatrix} 0.7 & \pi/7 & \pi/7 & 0.7 & 1.5 & 1.5 & 0.0 & 0.5 \end{pmatrix}, \\ \boldsymbol{\mu}_M = \begin{pmatrix} 1.3 & \pi/3 & \pi/3 & 1.2 & 2.5 & 2.5 & 1 & 1.5 \end{pmatrix}.$$

The optimal control problem can thus be stated as

$$\text{given } \boldsymbol{\mu} \in \mathcal{D}, \quad \text{minimize } J(\cdot; \boldsymbol{\mu}) \quad \text{subject to } (5.5.1).$$

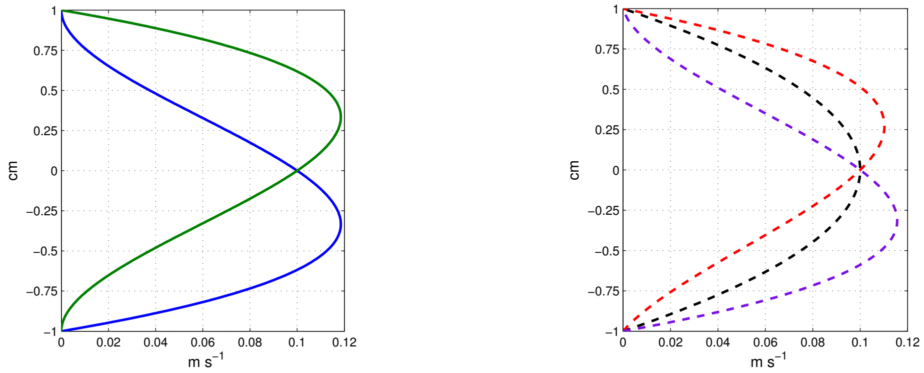


Figure 5.8: Parametrized observation $\boldsymbol{v}_d(\boldsymbol{\mu}_{meas})$. On the left the functions $\eta_1(x_{o2})$ (in blue) and $\eta_2(x_{o2})$ (in green). On the right: horizontal component of the observation function $\boldsymbol{v}_d(\boldsymbol{\mu})$ for different values of the parameter μ_7 (with μ_8=1 fixed): $\mu_7 = 0.5$ in black, $\mu_7 = 0.1$ in red, $\mu_7 = 0.95$ in violet.

5.5.2 Original formulation of the optimal control problem

Let us now define the appropriate functional spaces for the pressure and velocity variables: $M_o = L^2(\Omega_o)$ and $V_o = [H^1_{\Gamma_D}(\Omega_o)]^2$ where

$$H^1_{\Gamma_D}(\Omega_o) = \{v \in H^1(\Omega_o) \,:\, v|_{\Gamma^o_D} = 0,\, v|_{\Gamma^o_{in}} = 0\}.$$

Then we define the state space $Y_o = V_o \times M_o$, the adjoint space $Q_o \equiv Y_o$ and the control space $U_o = [H^1(\Gamma^o_C)]^2$. The weak formulation of the state equation is given by

$$\begin{cases} a_o(\boldsymbol{v}_o, \boldsymbol{\xi}) + b_o(\boldsymbol{\xi}, p_o) = \langle F_o(\boldsymbol{\mu}), \boldsymbol{\xi}\rangle + c_o(\boldsymbol{u}_o, \boldsymbol{\xi}), & \forall \boldsymbol{\xi} \in V_o, \\ b_o(\boldsymbol{v}_o, \tau) = \langle G_o(\boldsymbol{\mu}), \tau\rangle, & \forall \tau \in M_o, \end{cases} \tag{5.5.4}$$

where the bilinear form $a_o\colon V_o \times V_o \to \mathbb{R}$ and $b_o\colon V_o \times M_o \to \mathbb{R}$ are defined as in Section 3.5.1, while the bilinear form $c_o\colon U_o \times V_o \to \mathbb{R}$ is given by

$$c_o(\boldsymbol{u}_o, \boldsymbol{\xi}) = \int_{\Gamma^o_C} \boldsymbol{u}_o \cdot \boldsymbol{\xi}\, d\Gamma_o.$$

Moreover the terms $G_o(\boldsymbol{\mu})$ and $F_o(\boldsymbol{\mu})$ are due to the non-homogeneous Dirichlet boundary condition on Γ^o_{in}. Now we define the compound bilinear forms $\mathsf{A}_o(\cdot,\cdot) : Y_o \times Y_o \to \mathbb{R}$ and $\mathsf{C}_o(\cdot,\cdot) : U_o \times Q_o \to \mathbb{R}$ respectively as

$$\mathsf{A}_o(\{\boldsymbol{v}_o, p_o\}, \{\boldsymbol{\xi}, \tau\}) = a_o(\boldsymbol{v}_o, \boldsymbol{\xi}) + b_o(\boldsymbol{\xi}, p_o) + b_o(\boldsymbol{v}_o, \tau),$$

$$\mathsf{C}_o(\boldsymbol{u}_o, \{\boldsymbol{\xi}, \tau\}) = c_o(\boldsymbol{u}_o, \boldsymbol{\xi}),$$

while the linear functional $\boldsymbol{G}_o(\boldsymbol{\mu}) \in Q'_o$ is given by $\langle \boldsymbol{G}_o(\boldsymbol{\mu}), \{\boldsymbol{\xi}, \tau\}\rangle = \langle F_o(\boldsymbol{\mu}), \boldsymbol{\xi}\rangle + \langle G_o(\boldsymbol{\mu}), \tau\rangle$. Finally we define the bilinear form $m_o(\cdot,\cdot) : V_o \times V_o \to \mathbb{R}$ as

$$m_o(\boldsymbol{v}_o, \boldsymbol{\varphi}) = \int_{\Gamma^o_{obs}} \boldsymbol{v}_o \cdot \boldsymbol{\varphi}\, d\Gamma_o,$$

and the bilinear form $n_o(\cdot,\cdot) : U_o \times U_o \to \mathbb{R}$ as

$$n_o(\boldsymbol{u}_o, \boldsymbol{\lambda}) = \alpha_2 \int_{\Gamma^o_C} \boldsymbol{u}_o \cdot \boldsymbol{\lambda}\, d\Gamma_o + \alpha_1 \int_{\Gamma^o_C} \frac{\partial \boldsymbol{u}_o}{\partial \gamma_o} \cdot \frac{\partial \boldsymbol{\lambda}}{\partial \gamma_o}\, d\Gamma_o.$$

In order to formulate the optimal control problem as a saddle-point problem, let $X_o = Y_o \times U_o$, $\underline{\boldsymbol{x}}_o = (\{\boldsymbol{v}_o, p_o\}, \boldsymbol{u}_o) \in X_o$, $\underline{\boldsymbol{\zeta}} = (\{\boldsymbol{\varphi}, \pi\}, \boldsymbol{\lambda}) \in X_o$, $\{\boldsymbol{\xi}, \tau\} \in Q_o$ and define the bilinear forms

$$\begin{aligned} \mathcal{A}_o(\underline{\boldsymbol{x}}_o, \underline{\boldsymbol{\zeta}}) \quad &= m_o(\boldsymbol{v}_o, \boldsymbol{\varphi}) + n_o(\boldsymbol{u}_o, \boldsymbol{\lambda}), \\ \mathcal{B}_o(\underline{\boldsymbol{x}}_o, \{\boldsymbol{\xi}, \tau\}) \quad &= \mathsf{A}_o(\{\boldsymbol{v}_o, p_o\}, \{\boldsymbol{\xi}, \tau\}) - \mathsf{C}_o(\boldsymbol{u}_o, \{\boldsymbol{\xi}, \tau\}), \end{aligned}$$

and the linear functional

$$\langle \underline{\boldsymbol{F}}_o(\boldsymbol{\mu}), \underline{\boldsymbol{\zeta}} \rangle = \int_{\Gamma^o_{obs}} \boldsymbol{v}_d(\boldsymbol{\mu}) \cdot \boldsymbol{\varphi} \, d\Gamma_o.$$

The original optimal control problem reads: given $\boldsymbol{\mu} \in \mathcal{D}$,

$$\begin{cases} \min \, \mathcal{J}(\underline{\boldsymbol{x}}_o; \boldsymbol{\mu}) = \dfrac{1}{2}\mathcal{A}_o(\underline{\boldsymbol{x}}_o, \underline{\boldsymbol{x}}_o) - \langle \underline{\boldsymbol{F}}_o(\boldsymbol{\mu}), \underline{\boldsymbol{x}}_o \rangle, & \text{subject to} \\ \mathcal{B}_o(\underline{\boldsymbol{x}}_o, \{\boldsymbol{\xi}, \tau\}) = \langle \boldsymbol{G}_o(\boldsymbol{\mu}), \{\boldsymbol{\xi}, \tau\} \rangle \qquad \forall \{\boldsymbol{\xi}, \tau\} \in Q_o. \end{cases} \tag{5.5.5}$$

5.5.3 Parametrized formulation in the reference domain

We denote with $\Omega = \Omega_o(\boldsymbol{\mu}_{\text{ref}})$ the reference domain, with the choice (recall that μ_7 and μ_8 are not geometrical parameters)

$$\boldsymbol{\mu}_{\text{ref}} = (1, \pi/5, \pi/5, 1, 2, 2, \cdot, \cdot).$$

The reference domain with the subdomains numbering is shown in Figure 5.9. The linear transformations matrices of the affine geometrical mappings are computed automatically by the `rbMIT` software, we report here only some of them:

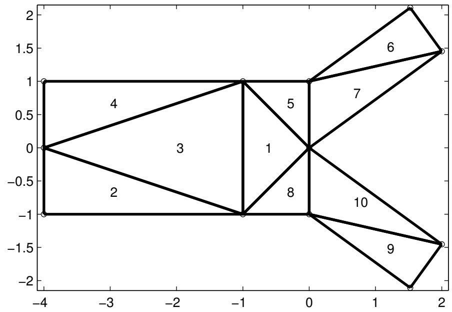


Figure 5.9: Reference domain with `rbMIT` triangulation.

$$\boldsymbol{G}^1 = \begin{pmatrix} \mu_4 & 0 \\ 1 - \mu_1 & 1 \end{pmatrix}, \qquad \boldsymbol{G}^2 = \boldsymbol{G}^3 = \boldsymbol{G}^4 = \begin{pmatrix} \frac{4-\mu_4}{3} & 0 \\ 0 & 1 \end{pmatrix},$$

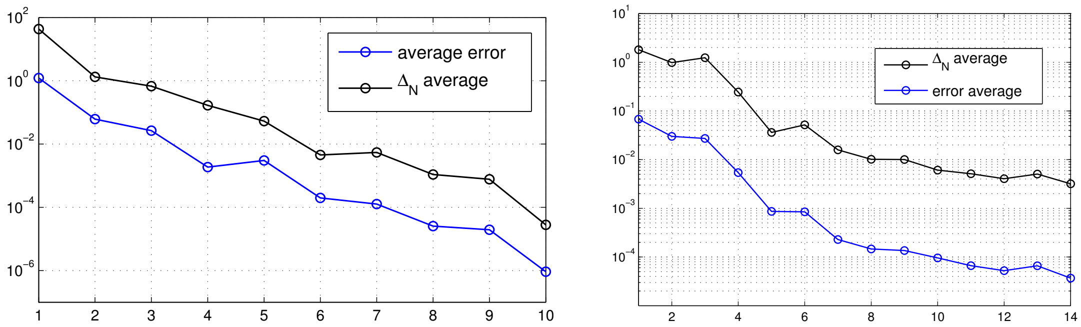
$$\boldsymbol{G}^5 = \begin{pmatrix} \mu_4 & 0 \\ 0 & \mu_1 \end{pmatrix}, \qquad \boldsymbol{G}^7 = \begin{pmatrix} \frac{\mu_5}{2} & 0 \\ \frac{\mu_5 \, \tan(\mu_2)}{2} - \mu_1 \tan(\pi/5) & \mu_1 \end{pmatrix},$$

$$\boldsymbol{G}^8 = \begin{pmatrix} \mu_4 & 0 \\ 0 & 2 - \mu_1 \end{pmatrix}, \qquad \boldsymbol{G}^{10} = \begin{pmatrix} \frac{\mu_6}{2} & 0 \\ 2\tan(\pi/5) - \frac{\mu_6 \, \tan(\mu_3)}{2} - \mu_1 \tan(\pi/5) & 2 - \mu_1 \end{pmatrix}.$$

By tracing problem (5.5.5) back to the reference domain we obtain the parametrized formulation (5.1.8) and the equivalent saddle-point problem (5.1.9), where the affine decompositions (5.1.11) (5.1.12) holds with $Q_a = 1 + 4$, $Q_b = 8 \cdot 8 + 2$, $Q_f = 4$, $Q_g = 8 \cdot 2$.

5.5.4 Numerical results

Since using a moderately refined mesh the dimension of the global FE space $\mathcal{X}^{\mathcal{N}} = X^{\mathcal{N}} \times Q^{\mathcal{N}}$ is equal to $\mathcal{N} = 39656$, quite larger than in all the previously discussed examples, we encountered some software and hardware limitations. In particular, the current implementation of the SCM algorithm when dealing with such a number of parameters requires the solution of an incredible number of (numerically challenging) eigenproblems, an unaffordable task for a desktop computer. Furthermore, performing the greedy algorithm for the bases selection guided by the a posteriori error estimator also require a huge amount of memory, resulting in a limitation of the maximum number of basis functions computable[2]. To overcome these difficulties we have firstly verified the reliability of the methodology performing some tests on a coarsest mesh varying only some parameters at a time and keeping the others fixed; in these cases we have also the advantage given by the lower number of terms in the affine decomposition (i.e. Q_a and Q_b are lower than in the complete problem), thus resulting in a lower computational effort (both Offline and Online). Then the full problem has been solved without providing the a posteriori error estimate (hence the online evaluation is not certified).



(a) First test: varying μ_2, μ_7, μ_8; $N_{max} = 10$.

(b) Second test: varying $\mu_1, \mu_4, \mu_7, \mu_8$; $N_{max} = 14$.

Figure 5.10: Average computed error and bound between the *truth* FE solution and the RB approximation, for $N = 1, \cdots, N_{max}$.

As a first test we allow the parameters μ_2, μ_7, μ_8 to vary while the others are kept fixed equal to their reference values. With a fixed tolerance $\varepsilon_{tol} = 5 \cdot 10^{-3}$, $N_{max} = 10$ basis functions have been selected; in Figure 5.10a we compare the average a posteriori error bound $\Delta_N(\boldsymbol{\mu})$ with the average true error $\|\mathrm{x}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathrm{x}_N(\boldsymbol{\mu})\|_{\mathcal{X}}$.

Similarly, as a second test we allow the parameters $\mu_1, \mu_4, \mu_7, \mu_8$ to vary while the others are kept fixed equal to their reference values. With a fixed tolerance $\varepsilon_{tol} = 5 \cdot 10^{-3}$, $N_{max} = 14$ basis functions have been selected; in Figure 5.10b we compare the average a posteriori error bound $\Delta_N(\boldsymbol{\mu})$ with the average true error $\|\mathrm{x}^{\mathcal{N}}(\boldsymbol{\mu}) - \mathrm{x}_N(\boldsymbol{\mu})\|_{\mathcal{X}}$. Similar results can be obtained with other combinations of the parameters.

As regards the computational costs, we underline that, as already noted in the numerical tests discussed in Chapter 4, most of the Offline time is spent providing the ingredients for the a posteriori error estimation. In fact, as the complexity of the problem increases (both in terms of size $\mathcal{N}$, number of parameters and number of terms in the affine decompositions),

[2] An implementation of this problem on a cluster made up of several computing processors was beyond the purposes of this work.

the time spent computing the basis functions becomes more and more marginal with respect to the time spent performing the SCM algorithm and computing the scalar products needed for the calculation of the dual norm of the residual (5.3.9). Therefore, providing the a posteriori error estimation can also require about the 90% of the overall Offline time.

Finally, once we have (partially) tested the correctness of the model and the good approximation properties of the RB method, we have considered the complete problem, allowing variations in all the eight parameters. As already mentioned, we avoided to perform the a posteriori error estimation: the basis functions have been computed in correspondence of a random set of 43 parameter samples. To check the convergence of the RB approximation we have computed the average error between the *truth* FE solution and the RB approximation, as shown in Figure 5.11.

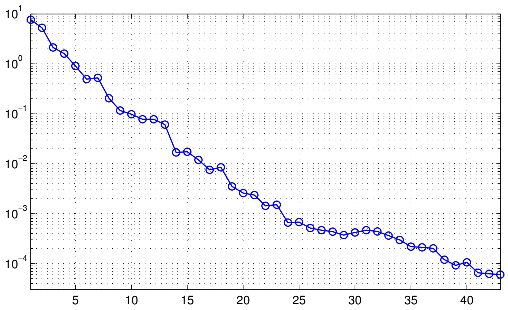


Figure 5.11: Average computed error between the *truth* FE solution and the RB approximation.

In Figure 5.12, 5.13, 5.14 we report some representative solutions. For each case we show the geometrical configuration identified by the chosen geometrical parameters as well as the inflow velocity profile and the desired velocity profile on Γ_{obs} depending on the values of μ_7 and μ_8; then we show the retrieved velocity and pressure fields. The Online RB evaluation requires less than 0.1s, since we have only to solve the low-dimensional RB linear system (however providing also the a posteriori bound should not require more than a couple of seconds).

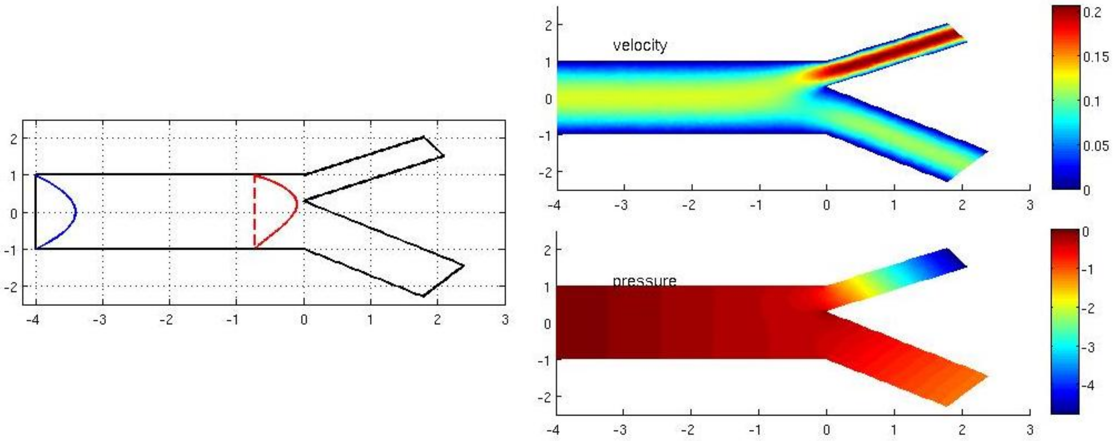


Figure 5.12: Representative solution for $\boldsymbol{\mu} = (0.7, \pi/6, \pi/5, 0.73, 2.1, 2.4, 0.25, 1.2)$. On the left the input geometrical configuration with plots of the inflow velocity profile and desired velocity profile on Γ_{obs}; on the right velocity (up, [ms^{-1}]) and pressure (bottom, [Pa]) fields.

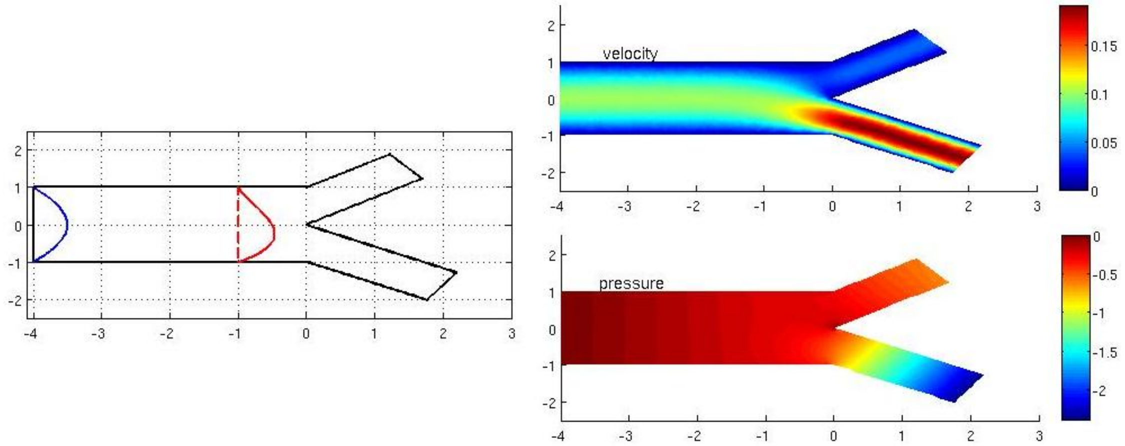


Figure 5.13: Representative solution for $\boldsymbol{\mu} = (1, \pi/5, \pi/6, 1, 1.7, 2.2, 0.8, 1)$. On the left the input geometrical configuration with plots of the inflow velocity profile and desired velocity profile on Γ_{obs}; on the right velocity (up, $[\text{ms}^{-1}]$) and pressure (bottom, [Pa]) fields.

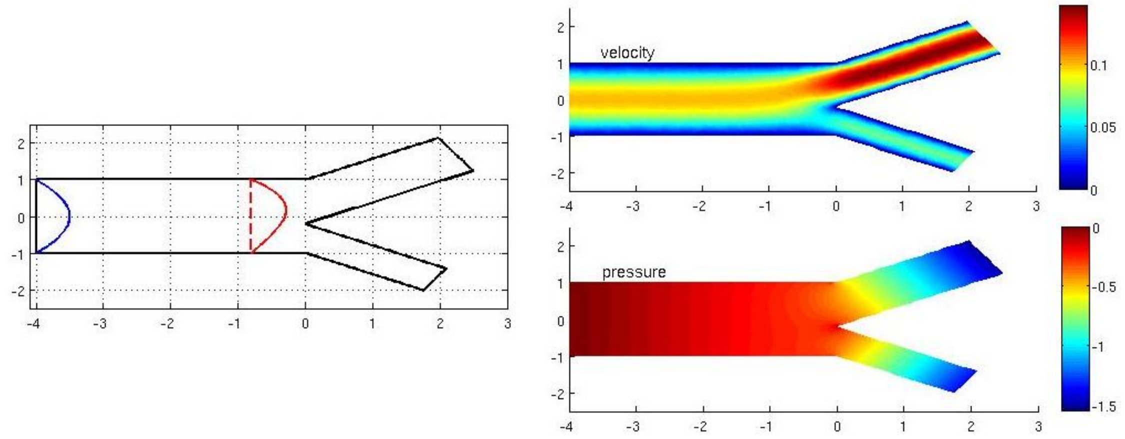


Figure 5.14: Representative solution for $\boldsymbol{\mu} = (1.2, \pi/6, \pi/6, 0.8, 2.5, 2.1, 0.3, 1)$. On the left the input geometrical configuration with plots of the inflow velocity profile and desired velocity profile on Γ_{obs}; on the right velocity (up, $[\text{ms}^{-1}]$) and pressure (bottom, [Pa]) fields.

Conclusions

In this thesis we have developed a reduced basis framework for the efficient solution of parametrized linear-quadratic optimal control problems. As already pointed out in the Introduction, some reduced order strategies have already been used for the efficient solution of parametrized optimal control problems; however most of the previous works deal with low-dimensional control variables, e.g. a set of scalars, that could be treated themselves as parameters, thus not requiring a *reduction to low dimensionality* of the whole optimal control problem, but only of the state equation. In this work we considered high-dimensional control variables, in particular we dealt with both distributed and boundary control problems, the control variable belonging to an infinite-dimensional functional space. Moreover, an efficient and rigorous a posteriori error estimation, necessary both for constructing the reduced order model and measuring its accuracy, was still missing for the general class of optimal control problems here considered. For example, the a posteriori estimators for the error in the cost functional and in the control variable proposed in [17, 18] show to be efficient in practice but unfortunately lack of rigorousness, whereas the estimator proposed in [88] is proved to be rigorous but not efficient. Only recently in [32] an efficient and rigorous estimator has been proposed, yet again in the simpler case of scalar constant control function.
In particular, we have proposed a RB method for reducing the complexity of the whole optimal control problem and not just for reducing the complexity of the state equation in order to speed up an optimization process in the Online stage. In fact, while the latter approach is undoubtedly convenient when the control variable is low-dimensional and coincide with a set of parameters (as in the case of shape optimization problems [56, 51, 78]), in the case here considered of high-dimensional control variables, it is no more viable.

The first part of the thesis, comprehensive of the first three chapters, have been in some way preparatory for the development of the reduce basis scheme. We briefly summarize here the main concepts introduced, highlighting their role in the construction of the proposed RB methodology.
The first chapter have been devoted to the introduction of the theory of optimal control problems. We have seen that using the Lagrangian formalism, the solution of the optimal control problem can be characterized by the optimality conditions system, which in our case yield a system of PDEs – state equation, adjoint equation and optimality condition. In the case of optimal linear-quadratic control problems here considered, this coupled system features a saddle-point structure. Exploiting this structure has been the starting point in order to develop our reduction scheme, since it has permitted us to suitably adapt the already well-developed theory of RB method for Stoke-type problems. To highlight this structure, we have thus recast the problem in the framework of mixed variational problems, discussing its well-posedness as well as the stability of its Galerkin approximation. This well-posedness analysis has been useful also in Chapters 4 and 5 to guide the construction of stable RB

spaces.
In the second chapter we have presented the most popular numerical methods employed to solve the Galerkin-finite element approximation of the control problem. Once again, we were interested in discussing these issues in view of the RB paradigm: firstly because in our approach the RB spaces are made of FE solutions of the whole control problem computed in correspondence of several (fixed) values of the parameters; hence in order to provide an efficient Offline stage, we have analysed the more convenient strategies to solve the FE approximation of the problem. Secondly, because at the Online stage our reduced scheme requires the solution of a RB optimality conditions system featuring the same structure of the FE one. In both the cases we have employed a one-shot approach.
In the third chapter we have briefly introduced the main ingredients of the RB method for parametrized PDEs. The main role of this chapter has been to highlight the basic features that an efficient and reliable RB scheme should provide and to introduce the computational strategies (such as the greedy algorithm for the bases selection and the SCM algorithm for the computation of a lower bound for the stability factor) that we have successively adopted also in our RB scheme.

After these preliminary chapters we had at our disposal all the basic ingredients needed to construct an efficient and reliable RB scheme for the solution of parametrized optimal control problems. In particular we developed the method for problems governed by scalar elliptic coercive equations and Stokes system. From the theoretical point of view, the well-posedness analysis has been carried out exploiting the saddle-point formulation of the problem. In fact, we have ensured the stability of the RB approximation through the definition of suitable RB spaces fulfilling an equivalent Brezzi inf-sup condition. On the other hand, the certified error bounds on the state, adjoint and control variables as well as on the cost functional have been obtained by recasting the problem in the form of weakly coercive problems and then applying standard arguments based on Nečas-Babuška stability theory. We also provided a full Offline-Online decomposition strategy ensuring the Online efficiency of the method. We have then performed several numerical tests confirming the theoretical results and showing the good reliability and efficiency properties of the proposed RB paradigm. The performances, both in terms of computational costs and accuracy, largely justify the adoption of the RB method for the solution of parametrized optimal control problems in the many-query and real-time contexts. As an example of application of the developed technique we have considered an inverse problem in haemodynamics. Beyond some implementation limitations, the method have shown its great versatility and potentiality in tackling parametrized optimal control problems that could arise in a a broad variety of application contexts, like environmental and bio-engineering/life sciences, to provide just a few examples of possible applications of recent growing interest.

We now mention some possible further developments that could be included in the proposed framework. In the context of linear-quadratic problems, possible developments guidelines are related to:

- the study of the time-dependent case, with a suitable POD approach, extending the work in [17, 18]. While from the theoretical point of view the non-stationary case does not represent a challenging task, from the computational point of view it requires very efficient numerical methods;
- to add control and/or state constrains, the first case could be treated straightforwardly

while the second would require a more involved analysis both from the theoretical and algorithmic point of view;

- to consider non-affinely parametrized transformations of the domain, that need to be approximated by affinely parametrized tensors through the empirical interpolation method [4, 31], in order to ensure the feasibility of the Offline/Online computational strategy. With this extension it would be possible to consider more realistic geometries.

Another challenging task could be to consider non-linear state equation, for which developing rigorous error estimation is not straightforward. Of particular interest the development of reduced order strategies for the optimal control of Navier-Stokes equations: we only mention the first works by Ito and Ravindran [48, 49, 50] and the recent works [22, 55] on the application of the RB method to Navier-Stokes equations (not in optimization context).

Appendices

Appendix A

Classical Variational Methods

In this appendix we introduce the most common abstract variational problems and analyse their well-posedness. We briefly review strongly coercive problems, weakly coercive problems (also called non-coercive problem) and saddle-point problems (also called mixed variational problems). Then we consider their Galerkin approximation, i.e. their restriction to finite-dimensional subspaces, focusing on the discrete counterparts of the conditions ensuring the well-posedness of each problem. Finally, in the third Section, we discuss how to compute numerically the stability factors in the discretized case, i.e. discrete coercivity and inf-sup constants.

A.1 Abstract variational problems

We briefly review strongly coercive problems, weakly coercive problems and saddle-point problems. For each of them, after having defined the functional settings and the precise statement, we present the main results ensuring the well-posedness: Lax-Milgram lemma, Nečas theorem and Brezzi theorem, respectively. We refer principally to [65, 15, 35].

A.1.1 Strongly coercive problems

Given an Hilbert space V along with its dual V^*, the bilinear form $a\colon V \times V \to \mathbb{R}$ and the linear functional $F \in V^*$, we consider the following abstract variational problem: find $u \in V$ such that

$$a(u,v) = F(v), \qquad \forall v \in V. \tag{P_1}$$

The bilinear form $a(\cdot,\cdot)$ is called continuous on $V \times V$ if

$$|a(u,v)| \leq M\|u\|_V\|v\|_V \qquad \forall u,v \in V,$$

where $M < \infty$, and is called strongly coercive (or simply coercive) if there exists a constant $\alpha > 0$ such that

$$a(v,v) \geq \alpha\|v\|_V^2 \qquad \forall v \in V.$$

The Lax-Milgram lemma shows that strongly coercive problems are well posed (see for instance [26, 69]).

Lemma A.1 (Lax-Milgram)**.** *Let V be an Hilbert space, $a(\cdot,\cdot)$ a continuous, strongly coercive bilinear form on $V \times V$, and $F(\cdot)$ a bounded linear functional on V. Then, the abstract variational problem (P_1) has a unique solution and we have the estimate:*

$$\|u\|_V \leq \frac{1}{\alpha}\|F\|_{V^*}. \tag{A.1.1}$$

A.1.2 Weakly coercive problems

Given two Hilbert space V and W along with their dual V^* and W^*, respectively, the bilinear form $\mathsf{A}\colon V \times W \to \mathbb{R}$ and the linear functional $\mathsf{F} \in W^*$, we consider the following abstract variational problem: find $u \in V$ such that

$$\mathsf{A}(u,w) = \mathsf{F}(w), \qquad \forall w \in W. \tag{P_2}$$

The bilinear form $\mathsf{A}(\cdot,\cdot)$ is called continuous on $V \times W$ if

$$|\mathsf{A}(u,v)| \leq M\|u\|_V\|v\|_W \qquad \forall u \in V, w \in W,$$

where $M < \infty$, and is called weakly coercive if there exists a constant $\beta > 0$ such that

$$\inf_{v\in V}\sup_{w\in W}\frac{\mathsf{A}(v,w)}{\|v\|_V\|w\|_W} \geq \beta, \tag{A.1.2}$$

and

$$\inf_{w\in W}\sup_{v\in V}\frac{\mathsf{A}(v,w)}{\|v\|_V\|w\|_W} > 0.$$

The Nečas theorem shows that weakly coercive problems are well posed [59].

Theorem A.1 (Nečas)**.** *Let V and W be two Hilbert space, $\mathsf{A}(\cdot,\cdot)$ a continuous, weakly coercive bilinear form on $V \times W$, and $F(\cdot)$ a bounded linear functional on W. Then, the abstract variational problem (P_2) has a unique solution and we have the estimate:*

$$\|u\|_V \leq \frac{1}{\beta}\|F\|_{W^*}. \tag{A.1.3}$$

Remark A.1. Since strong coercivity implies weakly coercivity, Lax-Milgram lemma is a special case of Nečas theorem.

A.1.3 Saddle-point problems

Given two Hilbert space X and Q along with their dual X^* and Q^*, respectively, the bilinear forms $a\colon X \times X \to \mathbb{R}$, $b\colon X \times Q \to \mathbb{R}$ and the linear functionals $f \in X^*$ and $g \in Q^*$, we consider the following abstract saddle-point problem (also called mixed variational problem): find $(x,p) \in X \times Q$ such that

$$\begin{cases} a(x,w) + b(w,p) = f(w) & \forall w \in X, \\ b(x,q) = g(q) & \forall q \in Q. \end{cases} \tag{P_3}$$

Let us define the subspace of X

$$X_0 = \{w \in X \colon b(w,q) = 0 \quad \forall q \in Q\} \subset X.$$

The Brezzi theorem [14] establishes conditions for which the saddle-point problem (P_3) is well posed.

Theorem A.2 (Brezzi)**.** *Given the Hilbert spaces X and Q, the functionals $f \in X^*$ and $g \in Q^*$, and the bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ on $X \times X$ and $X \times Q$, respectively. Assume that the bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ satisfy the following assumptions:*

1. *continuity of the bilinear form $a(\cdot,\cdot)$, i.e. there exists a constant $\gamma_a > 0$ such that*

$$|a(x,w)| \leq \gamma_a \|w\|_X \|x\|_X \qquad \forall x, w \in X;$$

2. *weakly coercivity of the bilinear form $a(\cdot,\cdot)$ on X_0, i.e. there exists a constant $\alpha_0 > 0$ such that*

$$\inf_{x\in X_0} \sup_{w \in X_0} \frac{a(x,w)}{\|x\|_X \|w\|_X} \geq \alpha_0 \qquad \textit{and} \qquad \inf_{w\in X_0} \sup_{x \in X_0} \frac{a(x,w)}{\|x\|_X \|w\|_X} > 0;$$

3. *continuity of the bilinear form $b(\cdot,\cdot)$, i.e. there exists a constant $\gamma_b > 0$ such that*

$$|b(w,q)| \leq \gamma_b \|w\|_X \|q\|_Q, \qquad \forall w \in X,\, q \in Q;$$

4. *the bilinear form $b(\cdot,\cdot)$ satisfies the inf-sup condition*

$$\beta = \inf_{q \in Q} \sup_{w \in X} \frac{b(w,q)}{\|w\|_X \|q\|_Q} \geq \beta_0, \tag{A.1.4}$$

where $\beta_0 > 0$.

Then there exists a unique solution $(x,p) \in X \times Q$ to problem (P_3) for any $f \in X^*$ and for any $g \in Q^*$. Moreover the following a priori estimates hold:

$$\|x\|_X \leq \frac{1}{\alpha}\Big[\|f\|_{X^*} + \frac{\alpha_0 + \gamma_a}{\beta_0}\|g\|_{Q^*}\Big], \tag{A.1.5a}$$

$$\|p\|_Q \leq \frac{1}{\beta}\Big[\Big(1 + \frac{\gamma_a}{\alpha_0}\Big)\|f\|_{X^*} + \frac{\gamma_a(\alpha_0 + \gamma_a)}{\alpha_0 + \beta_0}\|g\|_{Q^*}\Big]. \tag{A.1.5b}$$

Remark A.2. Note that mixed variational problems are a special case of weakly coercive problem. In fact by setting $V = W = X \times Q$, defining the bilinear form $\mathsf{A} : V \times V \to \mathbb{R}$

$$\mathsf{A}(\{x,p\},\{w,q\}) = a(x,w) + b(w,p) + b(x,q), \tag{A.1.6}$$

and the functional $\mathsf{F}(\{w,q\}) = f(w) + g(q)$, we can rewrite ($P_3$) in the form of ($P_2$): find $\{x,p\} \in X$ such that

$$\mathsf{A}(\{x,p\},\{w,q\}) = \mathsf{F}(\{w,q\}), \qquad \forall \{w,q\} \in X,$$

which could be analysed using Nečas theorem. For further details on the relations between the two theories see [21, 90].

A.2 Approximation of solutions of variational problems

A.2.1 Strongly coercive problems

Let $V_h \subset V$ be a nontrivial subspace of V, usually in applications V_h is finite dimensional and the subscript h is related to certain discretizations parameters. We consider the following variational problem: find $u_h \in V_h$ such that

$$a(u_h, v_h) = F(v_h), \qquad \forall v_h \in V_h. \tag{P_1^h}$$

The solution u_h of this problem is often known as the Galerkin approximation of u. The well-posedness of the discretized variational problem is given by the Céa lemma [15, 69, 35].

Lemma A.2 (Céa)**.** *Let the space V, the bilinear form $a(\cdot,\cdot)$ and the linear functional $f(\cdot)$ satisfy the hypotheses of Lemma A.1. Let $V_h \subset V$ be a closed subspace. Then $a(\cdot,\cdot)$ is continuous on $V_h \times V_h$ and satisfies*

$$a(v_h, v_h) \geq \alpha \|v_h\|_V^2, \qquad \forall v_h \in V_h,$$

and, for every $h > 0$, the discretized problem (P_1^h) has a unique solution $u_h \in V_h$. Moreover, that solution satisfies the stability estimate

$$\|u_h\|_V \leq \frac{1}{\alpha}\|f\|_{V^*},$$

and, if $u \in V$ denotes the unique solution of (P_1), the optimal error estimate

$$\|u - u_h\|_V \leq \frac{M}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_V.$$

A.2.2 Weakly coercive problems

Let $V_h \subset V$ and $W_h \subset W$ be two nontrivial subspaces of V and W, respectively. We consider the following variational problem: find $u_h \in V_h$ such that

$$\mathsf{A}(u_h, w_h) = \mathsf{F}(w_h), \qquad \forall w_h \in W_h. \tag{P_2^h}$$

The solution u_h of this problem is often known as the Petrov-Galerkin (often simply Galerkin) approximation of u. The well-posedness of the discretized variational problem is given by the Babuška theorem [3].

Theorem A.3 (Babuška)**.** *Let the space V and W, the bilinear form $\mathsf{A}(\cdot,\cdot)$ and the functional $\mathsf{F}(\cdot)$ satisfy the hypotheses of Theorem A.1. Let $V_h \subset V$ and $W_h \subset W$ be two closed subspaces. Then $A(\cdot,\cdot)$ is continuous on $V_h \times W_h$. Assume also that the bilinear form $\mathsf{A}(\cdot,\cdot)$ satisfies the discrete inf-sup conditions*

$$\inf_{v_h \in V_h} \sup_{w_h \in W_h} \frac{\mathsf{A}(v_h, w_h)}{\|v_h\|_V \|w_h\|_W} \geq \beta_h, \tag{A.2.1}$$

and

$$\inf_{w_h \in W_h} \sup_{v_h \in V_h} \frac{\mathsf{A}(v_h, w_h)}{\|v_h\|_V \|w_h\|_W} > 0.$$

where $\beta_h \geq \hat{\beta} > 0$. Then, for every $h > 0$, the discretized problem (P_2^h) has a unique solution $u_h \in V_h$. Moreover, that solution satisfies the stability estimate

$$\|u_h\|_V \leq \frac{1}{\hat{\beta}}\|f\|_{W^*},$$

and, if $u \in V$ denotes the unique solution of (P_2), the optimal error estimate

$$\|u - u_h\|_V \leq \left(1 + \frac{M}{\hat{\beta}}\right) \inf_{v_h \in V_h} \|u - v_h\|_V.$$

A.2.3 Saddle-point problems

Let $X_h \subset X$ and $Q_h \subset Q$ be two subspaces of X and Q, respectively. We consider the following variational problem: find $(x_h, p_h) \in X_h \times Q_h$ such that

$$\begin{cases} a(x_h, w_h) + b(w_h, p_h) = f(w_h) & \forall w_h \in X_h, \\ b(x_h, q_h) = g(q_h) & \forall q_h \in Q_h. \end{cases} \tag{P_3}$$

The solution (x_h, p_h) of this problem is the Galerkin approximation of (x, p). Let

$$X_0^h = \{w_h \in X_h \colon b(w_h, q_h) = 0 \quad \forall q_h \in Q_h\} \subset X_h.$$

The well-posedness of the discretized variational problem (P_3) is given by the discrete counterpart of Brezzi theorem [14, 69].

Theorem A.4 (Brezzi)**.** *Let the space X and Q, the bilinear forms $a(\cdot,\cdot)$, $b(\cdot,\cdot)$ and the functionals $f(\cdot)$ and $g(\cdot)$ satisfy the hypotheses of Theorem A.2. Let $X_h \subset X$ and $Q_h \subset Q$ be two finite dimensional subspaces. Then $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are continuous on $X_h \times X_h$ and $X_h \times Q_h$, respectively. Assume that the bilinear form $a(\cdot,\cdot)$ is weakly coercive on X_0^h, i.e.*

$$\inf_{x_h \in X_0^h} \sup_{w_h \in X_0^h} \frac{a(x_h, w_h)}{\|x_h\|_X \|w_h\|_X} \geq \alpha_h,$$

$$\inf_{w_h \in X_0^h} \sup_{x_h \in X_0^h} \frac{a(x_h, w_h)}{\|x_h\|_X \|w_h\|_X} > 0,$$

where $\alpha_h \geq \hat{\alpha} > 0$. Moreover suppose that the bilinear form $b(\cdot,\cdot)$ verifies the discrete inf-sup condition

$$\inf_{q_h \in Q_h} \sup_{w_h \in X_h} \frac{b(w_h, q_h)}{\|w_h\|_X \|q_h\|_Q} \geq \beta_h,$$

where $\beta_h \geq \hat{\beta} > 0$. Then, for every $h > 0$, the discretized problem (P_3) has a unique solution $(x_h, p_h) \in X_h \times Q_h$. Moreover, that solution satisfies the stability estimate

$$\|x_h\|_X \leq \frac{1}{\hat{\alpha}}\left[\|f\|_{X^*} + \frac{\hat{\alpha} + \gamma_a}{\hat{\beta}}\|g\|_{Q^*}\right], \tag{A.2.2a}$$

$$\|p_h\|_Q \leq \frac{1}{\hat{\beta}}\left[\left(1 + \frac{\gamma_a}{\hat{\alpha}}\right)\|f\|_{X^*} + \frac{\gamma_a(\hat{\alpha} + \gamma_a)}{\hat{\alpha} + \hat{\beta}}\|g\|_{Q^*}\right]. \tag{A.2.2b}$$

and, if $(x, p) \in X \times Q$ denotes the unique solution of (P_3), the optimal error estimate

$$\|x - x_h\|_X + \|p - p_h\|_Q \leq C\left(\inf_{w_h \in X_h} \|x - w_h\|_X + \inf_{q_h \in Q_h} \|p - q_h\|_Q\right),$$

where $C = C(\hat{\alpha}, \hat{\beta}, \gamma_a, \gamma_b)$ is independent of h.

A.3 On the computation of discrete stability factors

We are interested in deriving algebraic formulas to compute the discrete stability factors introduced in the Section above, i.e. discrete coercivity and inf-sup constants appearing in Céa lemma, Babuška theorem and Brezzi theorem. We refer principally to [54, 69, 92, 90].

A.3.1 Computation of discrete coercivity constant

Let V be an Hilbert spaces, $V_h \subset V$ a finite dimensional subspaces of V and consider the continuous bilinear form $a(\cdot,\cdot) : V \times V \to \mathbb{R}$. The goal is to find an algebraic formula to compute the coercivity constant

$$\alpha_h = \inf_{v_h \in V_h} \frac{a(v_h, v_h)}{\|v_h\|_V^2} > 0. \tag{A.3.1}$$

Let $\{\varphi_k\}_{k=1}^{N_v}$, where $N_v = \dim V_h$, be a basis for V_h and denote with $\mathbb{V}$ the norm matrix associated with the scalar product in V, i.e.

$$\mathbb{V}_{ij} = (\varphi_i, \varphi_j)_V,$$

and with $A \in \mathbb{R}^{N_v \times N_v}$ the matrix given by

$$A_{ij} = a(\varphi_j, \varphi_i), \qquad i, j = 1, \dots, N_v.$$

A given function $v_h \in V_h$ can be expressed through a linear combination of the basis functions of V_h in the following way

$$v_h = \sum_{i=1}^{N_v} v_i \varphi_i;$$

let $\mathbf{v} = (v_1, \dots, v_{N_v})^T$ denote the coefficients in the expansion of v_h in terms of the basis, note that $\|v_h\|_V^2 = \mathbf{v}^T \mathbb{V} \mathbf{v}$, $\forall v_h \in V_h$. We can now rewrite (A.3.1) as

$$\alpha_h = \inf_{\mathbf{v} \in \mathbb{R}^{N_v}} \frac{\mathbf{v}^T A \mathbf{v}}{\mathbf{v}^T \mathbb{V} \mathbf{v}},$$

we have thus expressed the coercivity constant α_h as a Rayleigh quotient and we can now compute it solving the following generalized eigenvalue problem: find $(\lambda, \mathbf{v})$ such that

$$A\mathbf{v} = \lambda \mathbb{V} \mathbf{v},$$

then the coercivity constant is given by $\alpha_h = \lambda_{\min}$.

A.3.2 Computation of discrete inf-sup constant

Let V and W be two Hilbert spaces, $V_h \subset V$ and $W_h \subset W$ two finite dimensional subspaces of V and W respectively, and consider the continuous bilinear form $\mathsf{A}(\cdot,\cdot) : V \times W \to \mathbb{R}$. The goal is to find an algebraic formula to compute the inf-sup constant

$$\beta_h = \inf_{v_h \in V_h} \sup_{w_h \in W_h} \frac{\mathsf{A}(v_h, w_h)}{\|v_h\|_V \|w_h\|_W} > 0. \tag{A.3.2}$$

Let us introduce the supremizer operator $T : V \to W$ defined as follows

$$(Tv, w)_W = \mathsf{A}(v, w) \qquad \forall w \in W,$$

it is easy to prove that

$$Tv = \arg \sup_{w \in W} \frac{\mathsf{A}(v, w)}{\|w\|_W}, \tag{A.3.3}$$

in fact, for all $w \in W$,

$$\frac{\mathsf{A}(v, w)}{\|w\|_W} = \frac{(Tv, w)_W}{\|w\|_W} \leq \frac{\|Tv\|_W \|w\|_W}{\|w\|_W} = \|Tv\|_W$$

and

$$\frac{\mathsf{A}(v, Tv)}{\|Tv\|_W} = \frac{(Tv, Tv)_W}{\|Tv\|_W} = \|Tv\|_W.$$

Furthermore, by definition of supremizer and using (A.3.3),

$$\beta = \inf_{v \in V} \sup_{w \in W} \frac{\mathsf{A}(v, w)}{\|v\|_V \|w\|_W} = \inf_{v \in V} \frac{\mathsf{A}(v, Tv)}{\|Tv\|_W \|v\|_V} = \inf_{v \in V} \frac{(Tv, Tv)_W}{\|Tv\|_W \|v\|_V} = \inf_{v \in V} \frac{\|Tv\|_W}{\|v\|_V},$$

or equivalently $\beta^2 = \inf_{v \in V} \frac{\|Tv\|_W^2}{\|v\|_V^2}$. Since the same arguments apply exactly also in the discrete case, we obtain

$$\beta_h^2 = \inf_{v_h \in V_h} \frac{\|Tv_h\|_W^2}{\|v_h\|_V^2}. \tag{A.3.4}$$

Now let $\{\varphi_k\}_{k=1}^{N_v}$, $\{\psi_k\}_{k=1}^{N_w}$, where $N_v = \dim\, V_h$, $N_w = \dim\, W_h$, be bases for V_h and W_h, respectively, and denote with $\mathbb{V}$ and $\mathbb{W}$ the norm matrices associated with the scalar products in V and W, i.e.

$$\mathbb{V}_{rs} = (\varphi_r, \varphi_s)_V, \qquad \mathbb{W}_{mn} = (\psi_m, \psi_n)_W,$$

and with $A \in \mathbb{R}^{N_w \times N_v}$ the matrix given by

$$A_{ij} = \mathsf{A}(\varphi_j, \psi_i), \qquad i = 1, \ldots, N_w,\ j = 1, \ldots, N_v.$$

Every functions $v_h \in V_h$ and $w_h \in W_h$ can be expressed through a linear combination of the basis functions of V_h and W_h, respectively, in the following way

$$v_h = \sum_{j=1}^{N_v} v_j \varphi_j, \qquad w_h = \sum_{i=1}^{N_w} w_i \psi_i;$$

let $\mathbf{v} = (v_1, \ldots, v_{N_v})^T$ and $\mathbf{w} = (w_1, \ldots, w_{N_w})^T$ denote the coefficients in the expansion of v_h and w_h in terms of the bases. Note that

$$\|v_h\|_V^2 = \mathbf{v}^T \mathbb{V} \mathbf{v}, \qquad \|w_h\|_W = \mathbf{w}^T \mathbb{W} \mathbf{w}, \qquad \forall v_h \in V_h, w_h \in W_h.$$

Moreover, let $\mathbf{t}$ denote the coefficients in the expansion of $Tv_h \in W_h$ with respect to W_h, by the definition of supremizer we obtain

$$\mathbf{t}^T \mathbb{W} \mathbf{w} = \mathbf{v}^T A^T \mathbf{w},$$

i.e. $\mathbb{W}\mathbf{t} = A\mathbf{v}$. Inserting $\mathbf{t} = \mathbb{W}^{-1}A\mathbf{v}$ in (A.3.4) gives

$$\beta_h^2 = \inf_{\mathbf{v}\in\mathbb{R}^{N_v}} \frac{\mathbf{t}^T\mathbb{W}\mathbf{t}}{\mathbf{v}^T\mathbb{V}\mathbf{v}} = \inf_{\mathbf{v}\in\mathbb{R}^{N_v}} \frac{\mathbf{v}^T A^T\mathbb{W}^{-T}\mathbb{W}\mathbb{W}^{-1}A\mathbf{v}}{\mathbf{v}^T\mathbb{V}\mathbf{v}} = \inf_{\mathbf{v}\in\mathbb{R}^{N_v}} \frac{\mathbf{v}^T A^T\mathbb{W}^{-1}A\mathbf{v}}{\mathbf{v}^T\mathbb{V}\mathbf{v}}, \tag{A.3.5}$$

we have thus expressed the inf-sup constant β_h^2 as a Rayleigh quotient and we can now compute it solving the following generalized eigenvalue problem: find $(\lambda, \mathbf{v})$ such that

$$A^T\mathbb{W}^{-1}A\mathbf{v} = \lambda\mathbb{V}\mathbf{v}, \tag{A.3.6}$$

then the inf-sup constant is given by $\beta_h = \sqrt{\lambda_{\min}}$.

Remark A.1. If $V = W$ and the bilinear form $\mathsf{A}(\cdot,\cdot)$ is symmetric it can be shown (e.g. [54]) that β_h is the minimum eigenvalue of the following eigenvalue problem

$$A\mathbf{v} = \lambda\mathbb{V}\mathbf{v}.$$

Remark A.2. In the case of mixed variational problems we can specialize (A.3.6) to the bilinear form $b(\cdot,\cdot) : X \times Q \to \mathbb{R}$; in this case we want to compute the slightly different inf-sup constant (known as Brezzi inf-sup constant)

$$\beta_h = \inf_{q_h\in Q_h} \sup_{w_h\in X_h} \frac{b(w_h, q_h)}{\|w_h\|_X\|q_h\|_Q}.$$

It is sufficient to define the bilinear form $\mathsf{A} : Q \times X \to \mathbb{R}$ as

$$\mathsf{A}(q, w) = b(w, q), \qquad \forall w \in X, q \in Q.$$

Now we identify V with Q and W with X, and denote with $\mathbb{X}$, $\mathbb{Q}$ the norm matrices associated with the scalar products in X and Q, respectively. Then β_h is given by the square root of the minimum eigenvalue of the following generalized eigenvalue problem: find $(\lambda, \mathbf{q})$ such that

$$B\mathbb{X}^{-1}B^T\mathbf{q} = \lambda\mathbb{Q}\mathbf{q}, \tag{A.3.7}$$

where the matrix $B = A^T$ is the matrix induced by the bilinear form $b(\cdot,\cdot)$.

Bibliography

[1] V. Akcelik, G. Biros, O. Ghattas, J. Hill, D. Keyes, and B. Waanders. Parallel Algorithms for PDE-Constrained Optimization. In M.A. Heroux, P. Raghavan, and H.D. Simon, editors, *Parallel Processing for Scientific Computing*, Philadephia, PA, 2006. SIAM.

[2] V.S. Arpaci. *Conduction Heat Transfer.* Addison-Wesley, reading ,UK, 1966.

[3] I. Babuška. Error-bounds for finite element method. *Numer. Math.*, 16:322–333, 1971.

[4] M. Barrault, Y. Maday, N. Nguyen Cuong, and A.T. Patera. An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations. *C. R. Acad. Sci. Paris. Sér. I Math.*, 339(9):667 – 672, 2004.

[5] R. Becker, H. Kapp, and R. Rannacher. Adaptive Finite Element Methods for Optimal Control of Partial Differential Equations: Basic Concept. *SIAM J. Control Optim.*, 39:113–132, 2000.

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

[7] M. Benzi, L. Ferragut, M. Pennacchio, and V. Simoncini. Solution of linear systems from an optimal control problem arising in wind simulation. *Numer. Linear Algebra Appl.*, 17(6):895–915, 2010.

[8] M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numer.*, 14(1):1–137, 2005.

[9] A. Borzì and K. Kunisch. A multigrid scheme for elliptic constrained optimal control problems. *Comput. Optim. Appl.*, 31:309–333, 2005.

[10] A. Borzì and V. Schulz. Multigrid methods for PDE optimization. *SIAM Review*, 51(2):361–395, 2009.

[11] J. Boyle, M. D. Mihajlović, and J. A. Scott. HSL MI20: an efficient AMG preconditioner, 2007.

[12] D. Braess and P. Peisker. On the numerical solution of the Biharmonic equation and the role of squaring matrices for preconditioning. *IMA Journal of Numerical Analysis*, 6(4):393–404, 1986.

[13] J.H. Bramble. *Multigrid Methods.* Pitman Research Notes in Mathematics Series, Essex, 1993.

[14] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Elements Methods.* Springer-Verlag, New York, 1991.

[15] P. Ciarlet. *The Finite Element Method for Elliptic Problems.* SIAM Classics in Applied Mathematics, Philadelphia, 2002.

[16] L. Dedé. *Adaptive and reduced basis method for optimal control problems in environmental applications.* PhD thesis, Politecnico di Milano, 2008. Available at `http://mox.polimi.it`.

[17] L. Dedé. Reduced basis method and a posteriori error estimation for parametrized linear-quadratic optimal control problems. *SIAM J. Sci. Comput.*, 32:997–1019, 2010.

[18] L. Dedé. Reduced basis method and error estimation for parametrized optimal control problems with control constraints. *Journal of Scientific Computing*, pages 1–19, 2011.

[19] L. Dedé and A. Quarteroni. Optimal control and numerical adaptivity for advection-diffusion equations. *M2AN Math. Model. Numer. Anal.*, 35:1019–1040, 2005.

[20] M. D'Elia, L. Mirabella, T. Passerini, M. Perego, M. Piccinelli, C. Vergara, and A. Veneziani. Applications of Variational Data Assimilation in Computational Hemodynamics. In D. Ambrosi, A. Quarteroni, and G. Rozza, editors, *Modelling of Physiological Flows*, volume 5 of *MS&A Series.* Springer, 2011.

[21] L. Demkowicz. Babuška = Brezzi? Technical Report 08-06, ICE, Univ. of Texas, 2006.

[22] S. Deparis and G. Rozza. Reduced basis method for multi-parameter-dependent steady Navier-Stokes equations: Applications to natural convection in a cavity. *J. Comput. Phys.*, 228:4359–4378, 2009.

[23] H.S. Dollar, N.I.M. Gould, M. Stoll, and A.J. Wathen. Preconditioning Saddle-Point Systems with Applications in Optimization. *SIAM J. Sci. Comput.*, 32:249–270, February 2010.

[24] J.L. Eftang, D.J. Knezevic, and A.T. Patera. An "hp" certified reduced basis method for parametrized parabolic partial differential equations. *Math. Comp. Model. Dyn.*, 17(4):395–422, 2011.

[25] H.C. Elman, D.J. Silvester, and A.J. Wathen. *Finite Elements and Fast Iterative Solvers with Applications in Incompressible Fluid Dynamics.* Oxford University Press, 2004.

[26] L. C. Evans. *Partial Differential Equations.* American Math Society, 1998.

[27] R. Fletcher. *Practical methods of optimization.* Wiley-Interscience (John Wiley & Sons), New York, 2nd edition, 2001.

[28] A.-L. Gerner and K. Veroy. Reduced basis a posteriori error bounds for the Stokes equations in parameterized domains: A penalty approach. *Mathematical Models and Methods in Applied Sciences (M3AS)*, 2011.

[29] G.H. Golub and C.F. Van Loan. *Matrix computations (3rd ed.).* Johns Hopkins University Press, 1996.

[30] E. Gonçalves, T. P. Mathew, M. Sarkis, and C. E. Schaerer. A Robust Preconditioner for the Hessian System in Elliptic Optimal Control Problems. In U. Langer, M. Discacciati, D. E. Keyes, O. B. Widlund, and W. Zulehner, editors, *Domain Decomposition Methods in Science and Engineering XVII*, volume 60 of *Lecture Notes in Computational Science and Engineering*, pages 527–534. Springer Berlin Heidelberg, 2008.

[31] M. Grepl, Y. Maday, N. Nguyen, and A.T. Patera. Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations. *Esaim Math. Model. Numer. Anal.*, 41(3):575–605, 2007.

[32] M.A. Grepl and M. Karcher. Reduced basis a posteriori error bounds for parametrized linear-quadratic elliptic optimal control problems. *C. R. Math. Acad. Sci. Paris*, 349(15-16):873 – 877, 2011.

[33] M.D. Gunzburger. *Perspectives in flow control and optimization.* SIAM, Philadelphia, 2003.

[34] M.D. Gunzburger and P.B. Bochev. Finite element methods for optimization and control problems for the Stokes equations. *Comp. Math. Appl.*, 48:1035–1057, 2004.

[35] M.D. Gunzburger and P.B. Bochev. *Least-Squares Finite Element Methods.* Springer, 2009.

[36] B. Haasdonk, J. Salomon, and B. Wohlmuth. A reduced basis method for parametrized variational inequalities. Technical Report 08-06, SimTech, Univ. of Stuttgart, May 2011.

[37] W. Hackbusch. *Multi-grid Methods and Applications.* Springer-Verlag, New York, 1985.

[38] R. Herzog and E. Sachs. Preconditioned conjugate gradient method for optimal control problems with control and state constraints. *SIAM. J. Matrix Anal. Appl.*, 31:2291–2317, 2010.

[39] M. Hinze, M. Koster, and S. Turek. A space-time multigrid solver for distributed control of the time-dependent Navier-Stokes system. Technical report, Priority Programme 1253, Preprint-Nr.: SPP1253-16-02, 2008.

[40] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints.* Springer, 2009.

[41] P. Holmes, J. Lumley, and G. Berkooz. *Turbulence, coherent structures, dynamical systems and symmetry.* Cambridge University Press, UK, 1996.

[42] D.B.P. Huynh. *Reduced Basis Approximation and Application to Fracture Problems.* PhD thesis, Singapore-MIT Alliance, National University of Singapore, 2007. Available at `http://augustine.mit.edu`.

[43] D.B.P. Huynh, D.J. Knezevic, Y. Chen, J.S. Hesthaven, and A.T. Patera. A natural-norm successive constraint method for inf-sup lower bounds. *Comput. Methods Appl. Mech. Engrg.*, 199(29-32):1963 – 1975, 2010.

[44] D.B.P. Huynh, N.C. Nguyen, A.T. Patera, and G. Rozza. Rapid reliable solution of the parametrized partial differential equations of continuum mechanics and transport. Available at `http://augustine.mit.edu`, ©MIT 2008-2011.

[45] D.B.P. Huynh and A.T. Patera. Reduced basis approximation and a posteriori error estimation for stress intensity factors. *Internat. J. Numer. Methods Engrg.*, 72(10):1219–1259, 2007.

[46] D.B.P Huynh, G. Rozza, S. Sen, and A.T. Patera. A successive constraint linear optimization method for lower bounds of parametric coercivity and inf-sup stability costants. *C. R. Acad. Sci. Paris. Sér. I Math.*, 345(8):473–478, 2007.

[47] K. Ito and K. Kunisch. *Lagrange Multiplier Approach to Variational Problems and Applications.* Adv. Des. Control. SIAM, 2008.

[48] K. Ito and S. S. Ravindran. A reduced-order method for simulation and control of fluid flows. *J. Comput. Phys.*, 143:403–425, July 1998.

[49] K. Ito and S. S. Ravindran. Reduced basis method for optimal control of unsteady viscous flows. *Int. J. Comput. Fluid Dyn.*, 15:97–113, 2001.

[50] K. Ito and S.S. Ravindran. A reduced basis method for control problems governed by PDEs. *Internat. Ser. Numer. Math.*, 126:153–168, 1998.

[51] T. Lassila and G. Rozza. Parametric free-form shape design with PDE models and reduced basis method. *Comput. Methods Appl. Mech. Engrg.*, 199(23–24):1583–1592, 2010.

[52] J.L. Lions. *Optimal Control of Systems governed by Partial Differential Equations.* Springer-Verlag, Berlin Heidelberg, 1971.

[53] J.L. Lions. *Some aspects of the optimal control od ditributed parameter systems.* SIAM, Philadelphia, 1972.

[54] D. S. Malkus. Eigenproblems associated with the discrete LBB-condition for incompressible finite elements. *Int. J. Engrg. Sci.*, 19:1299–1310, 1981.

[55] A. Manzoni, A. Quarteroni, and G. Rozza. Model reduction techniques for fast blood flow simulation in parametrized geometries. *Int. J. Numer. Meth. Biomed. Engng.*, 2011. Available online.

[56] A. Manzoni, A. Quarteroni, and G. Rozza. Shape optimization for viscous flows by reduced basis method and free form deformation. *Int. J. Numer. Meth. Fluids*, 2011. Available Online.

[57] Matlab®. The MathWorks. `http://www.mathworks.com`.

[58] F.M. Murphy, G.H. Golub, and A.J. Wathen. A note on preconditioning for indefinite linear systems. *SIAM J. Sci. Comput.*, 21:1969–1972, December 1999.

[59] J. Nečas. *Les Methodes Directes en Theorie des Equations Elliptiques.* Masson, Paris, 1967.

[60] J. Nocedal and S.J. Wright. *Numerical Optimization.* Springer, New York, 2006.

[61] A. Patera and G. Rozza. *Reduced Basis Approximation and A Posteriori Error Estimation for Parametrized Partial Differential Equations.* Copyright MIT, to appear in (tentative rubric) MIT Pappalardo Graduate Monographs in Mechanical Engeneering. Available at `http://augustine.mit.edu`, Version 1.0, 2006.

[62] M. Perego, A. Veneziani, and C. Vergara. A variational approach for estimating the compliance of the cardiovascular tissue: an inverse fluid-structure interaction problem. *SIAM J. Sci. Comput.*, 33(3):1181–1211, 2011.

[63] M. Picasso. Anisotropic a posteriori error estimate for an optimal control problem governed by the heat equation. *Numer. Methods Partial Differential Equations*, 22(6):1314–1336, 2006.

[64] T.A. Porsching. Estimation of the error in the reduced basis method solution of nonlinear equations. *Math. Comput.*, 45(172):487–496, 1985.

[65] A. Quarteroni. *Numerical models for differential problems*, volume 2 of *MS&A Series.* Springer, 2009.

[66] A. Quarteroni, G. Rozza, and A. Manzoni. Certified reduced basis approximation for parametrized PDE and applications. *J. Math in Industry*, 3(1), 2011.

[67] A. Quarteroni, G. Rozza, and A. Quaini. Reduced basis methods for optimal control of advection-diffusion problems. In *Advances in Numerical Mathematics*, pages 193–216, Moscow, Russia and Houston, USA, 2007.

[68] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical mathematics.* Springer, 2000.

[69] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations (1st Ed.).* Springer-Verlag, Berlin-Heidelberg, 1994.

[70] T. Rees, H. S. Dollar, and A. J. Wathen. Optimal Solvers for PDE-Constrained Optimization. *SIAM J. Sci. Comput.*, 32:271–298, 2010.

[71] T. Rees, M. Stoll, and A.J. Wathen. All at once preconditioning in pde-constrained optimization. Technical report, Oxford eprints archive, 2009.

[72] T. Rees and A. J. Wathen. Preconditioning iterative methods for the optimal control of the Stokes equation. Technical report, Oxford eprints archive, 2010.

[73] D.V. Rovas. *Reduced-basis output bound methods for parametrized partial differential equations.* PhD thesis, Massachusetts Institute of Technology, 2003.

[74] G. Rozza. On optimization, control and shape design of an arterial bypass. *Internat. J. Numer. Methods Fluids*, 47(10-11):1411–1419, 2005.

[75] G. Rozza. Reduced basis methods for Stokes equations in domains with non-affine parameter dependence. *Comput. Vis. Sci.*, 12(1):23–35, 2009.

[76] G. Rozza, D.B.P. Huynh, and A. Manzoni. Reduced basis approximation and a posteriori error estimation for Stokes flows in parametrized geometries: roles of the inf-sup stability constants. Technical Report 22.2010, MATHICSE. Submitted.

[77] G. Rozza, D.B.P. Huynh, and A.T. Patera. Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations. *Arch. Comput. Methods Eng.*, 15:229–275, 2007.

[78] G. Rozza, T. Lassila, and A. Manzoni. Reduced basis approximation for shape optimization in thermal flows with a parametrized polynomial geometric map. In *Spectral and High Order Methods for Partial Differential Equations. Selected papers from the ICOSAHOM '09 conference, June 22-26, Trondheim, Norway*, pages 307–315. Springer, Series: Lecture Notes in Computational Science and Engineering, vol. 76, J.S. Hesthaven, E.M. Rønquist (Eds.), 2011.

[79] G. Rozza and K. Veroy. On the stability of the reduced basis method for Stokes equations in parametrized domains. *Comput. Methods Appl. Mech. Engrg.*, 196(7):1244 – 1260, 2007.

[80] W. Rudin. *Real and Complex Analysis.* McGraw-Hill, 1986.

[81] Y. Saad. *Iterative Methods for Sparse Linear Systems.* SIAM, Philadelphia, 2003.

[82] F. Saleri, P. Gervasio, G. Rozza, and A. Manzoni. `MLife`, A Matlab Library for Finite Elements, tutorial (in progress). *MOX, Politecnico di Milano and CMCS, École Polytechnique Fédérale de Lausanne, 2000-2011.* ©Politecnico di Milano.

[83] J. Schöberl, R. Simon, and W. Zulehner. A robust multigrid method for elliptic optimal control problems. *SIAM J. Numer. Anal.*, 49(4):1482–1503, 2011.

[84] J. Schöberl and W. Zulehner. Symmetric Indefinite Preconditioners for Saddle Point Problems with Applications to PDE-Constrained Optimization Problems. *SIAM J. Matrix Anal. Appl.*, 29:752–773, October 2007.

[85] S. Sen. *Reduced Basis Approximation and A Posteriori Error Estimation for Non-Coercive Elliptic Problems: Application to Acoustics.* PhD thesis, Massachusetts Institute of Technology, 2007. Available at `http://augustine.mit.edu`.

[86] M. Stoll and A. J. Wathen. All-at-once solution of time-dependent PDE-constrained optimization problems. Technical report, Oxford eprints archive, 2010.

[87] M. Stoll and A. J. Wathen. All-at-once solution of time-dependent Stokes control. Technical Report MPIMD/11-03, Max Planck Institute Magdeburg Preprints, June 2011.

[88] T. Tonn, K. Urban, and S. Volkwein. Comparison of the reduced-basis and POD a-posteriori estimators for an elliptic linear-quadratic optimal control problem. *Math. Comput. Model. Dyn. Syst.*, 17(1):355–369, 2011.

[89] F. Tröltzsch. *Optimal Control of Partial Differential Equations.* Graduate Studies in Mathematics. AMS, Providence, Rhode Island, 2010.

[90] J. Xu and L. Zikatanov. Some observations on Babuška and Brezzi theories. *Numerische Mathematik*, 94:195–202, 2003.

[91] K. Yosida. *Functional Analysis.* Springer-Verlag, Berlin Heidelberg, 1974.

[92] L. Zanon. Reduced-basis approximation and a posteriori error estimation for saddle-point problems. Master's thesis, Politecnico di Torino, 2010.

[93] W. Zulehner. Nonstandard norms and robust estimates for saddle point problems. *SIAM. J. Matrix Anal. & Appl.*, 32(2):536–560, 2011.

Acknowledgements

Al termine di questo lavoro desidero ringraziare il Prof. Alfio Quarteroni, il Dr. Ing. Gianluigi Rozza e il Dott. Ing. Andrea Manzoni. La loro esemplare professionalità, unita ad una profonda umanità e benevolenza, sono state per me di continuo supporto e incoraggiamento.

Ringrazio anzitutto il Professor Alfio Quarteroni per avermi dato l'opportunità di svolgere questa tesi presso l'EPFL sotto la sua supervisione. Nonostante i numerosi impegni ha sempre trovato il tempo per seguire e indirizzare il mio lavoro.

Un ringraziamento speciale va al mio correlatore, il Dr. Ing Gianluigi Rozza, per la sua costante disponibilità, competenza e gentilezza. Grazie per aver sempre risposto *real-time* alle mie *many queries*.

Vorrei inoltre ringraziare il Dott. Ing. Andrea Manzoni: il suo nome non compare nel frontespizio, ma il suo contributo è stato davvero determinante; grazie per l'aiuto con Mlife e rbMIT, per le risposte ai piccoli e grandi dubbi, e per avermi sempre accolto in ufficio con il sorriso.

Ringrazio anche la Prof.ssa Paola Gervasio per l'utile discussione sui temi del secondo capitolo.