Politecnico di Milano
IV Facoltà di Ingegneria
Corso di laurea in Ingegneria Matematica
Dipartimento di Matematica

Basi ridotte: mappe transfinite per domini parametrici e leggi di conservazione

Relatori: Prof.ssa Simona Perotto
Prof. Sandro Salsa
Prof. Stefano Berrone

Tesi di Laurea di:
Tommaso Taddei Matr. 767253

Anno Accademico 2011-2012

A Sofia

Rovistando tra i futuri più probabili,
voglio solo futuri inverosimili.

Abstract

This thesis can be inserted in the framework of *Reduced Order Modelling* (ROM) techniques; more precisely, it focuses on the Reduced Basis method for a rapid and reliable solution of parametrized partial differential equations.
Even though the pioneering works on the Reduced Basis method date back to the seventies, during the last decade the method has been deeply analysed and developed to ensure an efficient and rigorous approximation of the solution for a wide class of partial differential equations.
Moving from a sound review of the available literature, the present work is essentially centered on geometric reduction strategies to deal with differential equations defined on parametrized domains and on the extension of the Reduced Basis methodology to nonlinear scalar conservation laws.
The first issue has been widely analysed in recent years, whereas the examples of applications of the Reduced Basis method to nonlinear hyperbolic problems are very few: as we will explain in this thesis, the typical structure of the solutions to these equations introduces a number of additional criticalities that require a substantial modification to some steps of the standard methodology.
In this thesis, a new geometric reduction technique, particularly suited to the treatment of roto-translations of the domain boundaries and a reduced order strategy to deal with nonlinear conservation laws in the presence of shocks are proposed and motivated both from a theoretical and computational point of view.
Finally, some suggestions for future developments concerning a possible extension of the proposed methodology to more general problems are offered.

Keywords: Reduced Order Modelling, Reduced Basis Method, Shape Parametrization Techniques, Conservation Laws.

Sommario

Questo lavoro di tesi si inserisce nell'ambito delle tecniche per la Riduzione di Modello (*Reduced Order Modelling*) e, più in particolare, si focalizza sul metodo delle Basi Ridotte per la risoluzione di problemi differenziali dipendenti da un insieme di parametri.
Sebbene i primi lavori sul metodo delle Basi Ridotte risalgano agli anni Settanta, nell'ultimo decennio il metodo è stato oggetto di notevoli sviluppi che lo hanno portato ad essere in grado di garantire un'approssimazione efficiente e rigorosa di una vasta gamma di problemi differenziali.
Partendo da un'attenta analisi della letteratura più recente, il presente lavoro è incentrato da un lato sulle strategie di riduzione geometrica nell'ambito dell'approssimazione di equazioni differenziali definite su domini parametrici; dall'altro sull'estensione del metodo delle Basi Ridotte a leggi di conservazione scalari non lineari.
Se la prima tematica è stata largamente analizzata negli ultimi anni, poche sono le applicazioni del metodo della Basi Ridotte a problemi iperbolici non lineari: come verrà ampiamente motivato in questa tesi, la struttura intrinseca delle soluzioni di tali equazioni introduce difficoltà aggiuntive che richiedono un ripensamento sostanziale di alcuni passi del metodo standard.
In questa tesi vengono proposte una nuova tecnica di riduzione geometrica pensata specificatamente per gestire il caso di roto-traslazioni di componenti del bordo del dominio ed una strategia di riduzione di modello per leggi di conservazione in presenza di shock; queste procedure sono state analizzate sia da un punto di vista teorico che da un punto di vista computazionale.
Nella parte finale della tesi, sono discussi alcuni possibili sviluppi futuri concernenti l'estensione del metodo a problemi più generali.

Parole chiave: Riduzione di Modello, Basi Ridotte, Tecniche di Parametrizzazione di Forma, Leggi di Conservazione.

Ringraziamenti

Questo lavoro di tesi si è sviluppato fra l'Ottobre 2011 ed il 28 Novembre 2012. In questo arco di tempo numerose persone vi hanno contribuito direttamente ed indirettamente.

Vorrei anzitutto ringraziare la Professoressa Simona Perotto ed il Professor Sandro Salsa. A loro va la mia più grande stima e riconoscenza per avermi sostenuto in ogni momento e per avere interpretato e guidato le mie idee anche quando erano confuse e largamente contraddittorie.

Un grandissimo ringraziamento va al Professor Alfio Quarteroni che mi ha concesso l'enorme opportunità di trascorrere un periodo di studio sotto la sua supervisione all'EPFL e che mi saputo indirizzare verso la definizione dell'attuale algoritmo qui proposto per la risoluzione delle leggi di conservazione.

Grazie a Dott. Ing. Gianluigi Rozza che mi ha introdotto al metodo delle Basi Ridotte e mi ha supportato durante il mio periodo Losannese (e non solo)e grazie anche a Dott. Ing. Luca Dedè, Dott. Ing. Toni Lassila ed Dott. Ing. Andrea Manzoni con cui ho avuto la possibilità di discutere a lungo del mio lavoro.

Desidero infine ricordare che se sono riuscito a portare a termine questa tesi, è merito di tutti quei Professori di Analisi Matematica ed Analisi Numerica che negli anni mi hanno formato. Grazie a loro, a partire dal 20 Dicembre 2012 potrò definirmi un Ingegnere Matematico. E ne andrò orgogliosissimo.

Abbandonando per un poco i toni ufficiali, vorrei ringraziare la mia famiglia ed in particolare i miei genitori per non avermi mai fatto mancare il loro supporto in tutti questi anni.

Fra le persone che mi sono state vicine non posso non ricordare (in rigoroso ordine alfabetico) i compagni milanesi - Daniele, Silvia, Paolo, Paolo, Roberto, Marianna, Paolo, Luca, Valentina - e quelli delle fugaci apparizioni fiorentine - Leonardo, Niccolò, Leonardo, Federico, Lorenzo, Gloria, Martino, Andrea - ed i miei carissimi e dolcissimi coinquilini - Massimo e Nicola. Infine vorrei ricordare le persone con cui ho condiviso la mia esperienza a Losanna -Laura, Vasilis, Matteo, Alessandro, Cristiano, Eleonora, Paolo, Anna, Francesco.

In ultimo, vorrei ringraziare Sofia, unico motivo plausibile per cui cinque anni fa ho deciso di venire a Milano a fare l'unica facoltà che a Firenze non c'era ed, in larga parte, ragione di quello che di buono ho fatto finora (a parte le Equazioni Differenziali, ma anche su questo ci si può lavorare...).

Un giorno cominceremo a prendere i treni nella direzione corretta con regolarità, non tritureremo più automobili pensando a dimostrazioni di teoremi sbagliati e lasceremo i nostri personali sacchetti dell'immondizia nel cassonetto a loro allocato.

Quel giorno saremo, semplicemente, invecchiati.

Introduction

Since the beginning of '70s, several strategies to speed up the solution of parametrized PDEs have been proposed: starting from the full scale problem these methodologies aim at deriving a Reduced Order Model (ROM) that is potential for at least near real time analysis.

Many of the first cost reduction schemes neither were based on a rigorous mathematical framework nor were showed to be of general use.

However, in the last ten years, an intense research, focused on both a theoretical foundation and the definition of new suitable algorithms, has led to great improvements in the field of reduced order modelling.

Reduced Basis (RB) method ([80, 93, 106]) is one among the different strategies proposed; it was developed during the last decade in particular for a rapid and reliable evaluation of input-output relationships based on the solution to a parametrized partial differential equation.

In particular this thesis focuses on the development, the analysis and the implementation of two tools dealing with open issues in the Reduced Basis framework. In more details, we propose:

- *a geometrical reduction technique for parametrized domains*: in order to deal with PDEs defined on parameter dependent domains, it is necessary to describe the deformation of the domain through a small number of degrees of freedom. For this purpose, in recent years several methods based on the introduction of suitable maps between the parameter dependent domain and a parameter independent configuration have been developed. In this work a new strategy, based on the well-known Gordon Hall transfinite map ([43]) is proposed;

- *a reduced order strategy for nonlinear scalar conservation laws*: as we will explain in the third chapter, the treatment of hyperbolic problems in a RB framework is particularly involved and requires some modifications to the standard strategy. This is why in this work a new algorithm, introduced as an adaptation of the RB method to conservation laws, is presented. A great attention is paid to the mathematical foundations behind the proposed procedure and to the implementative aspects.

The new ingredients proposed in this thesis can be potentially coupled in order to solve conservation laws in higher than one space dimension. This represents a possible future research field for us. However, in this work the two topics are dealt with separately.

The thesis is organized in three distinct chapters.

- **Chapter 1** provides a presentation of the main features of the Reduced Basis method for elliptic and parabolic linear equations. The offline-online decomposition, the sampling strategy and the a posteriori error estimation are addressed in detail; then the so-called *Empirical Interpolation method* ([7]) is explained and motivated. In

view of the successive application to hyperbolic problems, some criticalities of the standard RB approach are highlighted.

- **Chapter 2** deals with the definition of suitable geometrical reduction strategies to face PDEs in parametrized domains in the context of the RB method. After an overview of the state-of-the-art, a new methodology is explained and deeply analysed. It is also shown the importance of geometrical reduction for the treatment of multi-dimensional parametrized conservation laws.

- In **Chapter 3** the algorithm for scalar parametrized conservation laws is explained and soundly analysed: great attention is paid to the mathematical motivations behind the procedure. At the end of the theoretical discussion, some numerical tests are presented to show the performances of the method.

Some concluding remarks and perspectives on future developments are offered in the final section of this thesis while some important results used throughout the work are stated in `Appendix A`. and `Appendix B`.

The numerical validation in this work has been performed via `Matlab`® software ([84]).

More precisely, in the first and in the second chapters an enhanced version of `rbMIT`© [102] (developed at CMCS-EPFL) together with the Finite Element library `MLife`, [108], has been used. In the third chapter, the case studies are performed through a `Matlab` program completely developed while carrying out this work.

Finally, I would like to remark that this thesis has greatly benefited from a two month internship at École Polytechnique Féderale de Lausanne at the Mathematics Institute of Computational Science and Engineering.

Contents

1 Reduced Basis Method for parametrized PDEs **1**
1.1 Introduction . . . 1
1.2 Reduced basis formulation for elliptic problems . . . 3
1.2.1 Formulation of the problem . . . 3
1.2.2 Reduced model . . . 4
1.3 Reduced basis formulation for parabolic problems . . . 6
1.3.1 Petrov-Galerkin formulation for parabolic linear equations . . . 6
1.3.2 Formulation of the problem . . . 7
1.3.3 Reduced model . . . 9
1.4 Sampling strategy . . . 10
1.4.1 Orthogonalization procedures . . . 11
1.4.2 Kolmogorov N-Width . . . 11
1.4.3 Proper Orthogonal Decomposition . . . 12
1.4.4 Greedy sampling . . . 13
1.4.5 Sampling methods for elliptic and parabolic equations . . . 14
1.4.6 A brief overview on convergent rates of the Greedy and POD-Greedy methods . . . 15
1.5 A posteriori error estimation . . . 17
1.5.1 Residual-based a posteriori error estimator . . . 18
1.5.2 Residual definition for elliptic and parabolic equations . . . 19
1.5.3 Inf-sup lower bound for elliptic equations . . . 20
1.5.4 Inf-sup lower bound for parabolic equations . . . 23
1.5.5 A posteriori error estimation for the output . . . 23
1.6 The whole method: offline-online decomposition . . . 24
1.7 Two numerical examples . . . 25
1.7.1 A diffusion-reaction problem . . . 25
1.7.2 Graetz problem . . . 28
1.8 The treatment of non-affine problems: the EIM . . . 30
1.8.1 Description of the algorithm . . . 31
1.8.2 Error analysis and a posteriori error estimator . . . 31
1.9 Conclusions . . . 33

2 Reduced Basis Method for PDEs in parametrized domains **35**
2.1 Introduction and motivations . . . 35
2.2 Affine geometrical parametrization . . . 36
2.2.1 Affine mappings: single subdomains . . . 37
2.2.2 Piecewise affine mappings: multiple subdomains . . . 41
2.2.3 Formulation on the reference domain . . . 41
2.2.4 Two numerical examples . . . 43

2.3 Non-parametrically affine maps: three different approaches 46
2.3.1 Free Form Deformation 46
2.3.2 Methods based on radial basis functions 49
2.3.3 Transfinite maps 50
2.3.4 Formulation on the reference domain: Offline-Online decomposition 52
2.4 Transfinite maps for small deformations 53
2.4.1 Theoretical framework and first properties 54
2.4.2 Useful properties of transfinite transformations 55
2.4.3 The approximation of the derivatives 58
2.5 Numerical simulations 59
2.5.1 Hypothesis of small deformations 59
2.5.2 NACA profile 60
2.5.3 NACA profile with support 61
2.5.4 An advection-diffusion problem around a rotating symmetric NACA profile 63
2.6 Conclusions 65

3 Reduced Basis Techniques for Conservation Laws 69
3.1 Introduction 69
3.1.1 An introductive example 69
3.1.2 Overall strategy and structure of the chapter 72
3.2 Some preliminaries 73
3.2.1 The underlined truth approximation for the problem 73
3.2.2 Smooth-Jump decomposition algorithm 75
3.2.3 Finite volume projection algorithm 80
3.2.4 Smoothing 81
3.3 Main features of the methodology 82
3.3.1 Rankine-Hugoniot condition and the "special" formulation 83
3.3.2 Shock capturing algorithm 85
3.3.3 A RB approach for the smooth problems: offline-online decomposition 87
3.3.4 Sampling strategy 90
3.3.5 Input-output relationships 90
3.3.6 The whole algorithm 91
3.4 A posteriori error estimation: some preliminary comments 92
3.4.1 Error estimation for strong solutions 93
3.5 Numerical simulations 97
3.5.1 First example: convergence study with respect to the number of basis functions 97
3.5.2 Second example: analysis of the convergence in the presence of a shock 98
3.5.3 Third example: the input-output relation 99
3.6 Conclusions 102

A Some theoretical results 107
A.1 Existence and uniqueness for linear variational problems 107
A.1.1 The Lax-Milgram and Babuska theorems 107
A.1.2 Elliptic problems with L^2 boundary data: the transposition method 108
A.1.3 A useful comparison result 108
A.2 BV and SBV spaces 109
A.3 Mathematical analysis of scalar conservation laws 111

B Preliminary results for the BRR theory **115**
B.1 Some notations and basic lemmas 115
B.1.1 Implicit and inverse function theorems 116

Chapter 1

Reduced Basis Method for parametrized PDEs

1.1 Introduction

In this chapter we introduce the Reduced Basis (RB) method for the rapid and reliable evaluation of engineering outputs associated with elliptic and parabolic equations that depend on a set of parameters, say $\boldsymbol{\mu} \in \mathcal{D}$ where $\mathcal{D} \subset \mathbb{R}^P$ is a compact set.

In formulas we consider a single output of interest:

$$s(\boldsymbol{\mu}) = l(u(\boldsymbol{\mu})) \quad \boldsymbol{\mu} \in \mathcal{D} \tag{1.1.1a}$$

where $l \in X'$ and $u(\boldsymbol{\mu})$ is the solution of the following (elliptic or parabolic) equation:

$$\mathcal{L}(u(\boldsymbol{\mu}), \boldsymbol{\mu}) = 0 \quad \text{in } X' \quad \mathcal{L} : X \times \mathcal{D} \to X'. \tag{1.1.1b}$$

In order to make a clear presentation of the topic, we first introduce some notation and hypotheses and then we try to justify the main ideas of the RB method.

The solutions to the equation (1.1.1b), for each $\boldsymbol{\mu} \in \mathcal{D}$, define the parametric manifold:

$$\mathcal{M} = \{u(\boldsymbol{\mu}) \in X : u(\boldsymbol{\mu}) \text{ is the solution to (1.1.1b), } \boldsymbol{\mu} \in \mathcal{D}\}. \tag{1.1.2}$$

Given a Finite Element[1], [18, 97], approximation space $X^{\mathcal{N}} \subset X$ of dimension $\mathcal{N}$, we define $u^{\mathcal{N}}(\boldsymbol{\mu})$ as the solution to the approximate equation:

$$\mathcal{L}^{\mathcal{N}}(u^{\mathcal{N}}(\boldsymbol{\mu}), \boldsymbol{\mu}) = 0 \text{ in } X^{\mathcal{N}\prime} \tag{1.1.3}$$

With respect to $\mathcal{M}$ defined in (1.1.2), we indicate with $\mathcal{M}^{\mathcal{N}}$ the parametric manifold induced by the FE approximation:

$$\mathcal{M}^{\mathcal{N}} = \{u^{\mathcal{N}}(\boldsymbol{\mu}) \in X^{\mathcal{N}} : u^{\mathcal{N}}(\boldsymbol{\mu}) \text{ is the solution to (1.1.3), } \boldsymbol{\mu} \in \mathcal{D}\} \tag{1.1.4}$$

If we assume that the FE grid is sufficiently fine, we can consider negligible the difference between the solution $u(\boldsymbol{\mu})$ of (1.1.1b) and the solution of (1.1.3) for each value of the parameter and thus the manifold $\mathcal{M}^{\mathcal{N}}$ can be seen as a *truth approximation* of $\mathcal{M}$.

[1] We point out that the choice of finite elements is arbitrary and does not influence the method: for instance in [49] the RB method was applied within a finite volume context and in [75] within the context of spectral methods.

We also introduce, given $N_{max} \in \mathbb{N}$, an associated sequence of N-dimensional subspaces $X_N^{\mathcal{N}}$ - $N = 1, \cdots, N_{max}$ - with the following *hierarchical property*:

$$X_1^{\mathcal{N}} \subset X_2^{\mathcal{N}} \subset \cdots \subset X_{N_{max}}^{\mathcal{N}} \subset X^{\mathcal{N}}. \tag{1.1.5}$$

If the manifold $\mathcal{M}^{\mathcal{N}}$ is low dimensional and smooth[2], we can expect to well approximate the entire manifold through a very low dimensional space $X_N^{\mathcal{N}}$. This gives reasons for the application of the Reduced Basis method that is based on the following essential components.

1. *Rapidly convergent approximations*: the key points to address in order to provide good results are the generation of suitable subspaces $\{X_N^{\mathcal{N}}\}_N$ and an efficient methodology to generate the reduced solution. The former issue regards the definition of suitable optimality criteria that lead to effective sampling strategies for the approximation of the manifold. The latter issue consists in the introduction of a suitable reduced problem obtained through a (Galerkin) projection[3] of the original equation onto the low dimensional subspace.

2. *Rigorous a posteriori error estimation procedures*: after solving the reduced problem, we need to estimate the error between the truth and the reduced solutions. This must be done in an *inexpensive* (i.e., independent of the underlined mesh), *rigorous* (i.e., the estimation must constitute an upper bound for the actual error) and possibly *effective* (i.e., the ratio of the error bound to the true error is reasonably tight) way.

3. *Offline/Online computational procedures*: the main idea of the RB approach is to decouple the work into two different stages: in the first stage, performed once, the generation of the RB approximation and the computation and storing of all the structures needed for the reduced problem are addressed; in the second stage, repeated many times, only the solution of the reduced equation and the estimation of the error are performed.

In order to deal with the components defined above in the case of elliptic and parabolic problems, we subdivide the presentation of the general methodology into three different parts:

- in sections 1.2 and 1.3 the formulation for the elliptic and parabolic cases will be introduced;
- the sampling strategy will be discussed in section 1.4;
- the a posteriori estimation will be introduced in section 1.5.

For each section, we detail how it is possible to split the work into the two different stages, offline and online. At the end of this preliminary presentation, in 1.6 we present the whole algorithm and in 1.7 we present two numerical examples.

Then, an important tool to generalize the range of applicability of the methodology, the so-called Empirical Interpolation method, is extensively explained in section 1.8.

[2] The concept of smoothness is critical. In general it is a very hard task to assess the smoothness of the parametric manifold *a priori*. See [106] (Proposition 1 section 8) for a proof-of-concept.

[3] Galerkin projection is not the only option: for instance in [94] another approach is presented. As regards this work, we apply Galerkin projection for the elliptic case in chapter 1 and another approach for conservation laws.

Before concluding this introduction, we summarize some historical aspects about the development of the Reduced Basis method[4]. In '70s and '80s, the first works on Reduced Basis method appeared in the context of many query design evaluation for linear structural applications ([39]) and in the context of nonlinear structural analysis problems ([89]). In the next decade the approach was applied to different classes of equations such as the incompressible Navier-Stokes equation ([58]).

The approach here presented[5] is rather different from these early works especially because it is focused on the a posteriori error estimation ([125]) and on effective sampling strategies for higher than one dimensional parameter samples ([88]).

Unlike the early methods, the latter approach is designed to be not only particularly effective in the *many-query* and *real time* contexts but also a reliable methodology.

1.2 Reduced basis formulation for elliptic problems

This section deals with the main features of the RB method for input-output relationships based on elliptic equations as proposed; e.g. in [93, 106, 80].

First, we introduce the formulation and we discuss the preliminary hypotheses; then the reduced model is obtained through the standard Galerkin projection onto the reduced space. In the following the truth space $X^{\mathcal{N}}$ and the truth FE solution $u^{\mathcal{N}}(\boldsymbol{\mu})$ is simply indicated as X and $u(\boldsymbol{\mu})$. As already stated above, $\mathcal{N}$ is the dimension of the FE space.

1.2.1 Formulation of the problem

Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain and $X = X(\Omega)$ be the truth approximation of a suitable Hilbert space. Let $a(\cdot,\cdot,\boldsymbol{\mu}) : X \times X \to \mathbb{R}$ and $F(\cdot,\boldsymbol{\mu}) : X \to \mathbb{R}$ be *parametric affine* bilinear and linear forms, respectively:

$$\begin{cases} a(u,v,\boldsymbol{\mu}) = \displaystyle\sum_{q=1}^{Q_a} \Theta_q^a(\boldsymbol{\mu}) a^q(u,v), \\ F(v,\boldsymbol{\mu}) = \displaystyle\sum_{q=1}^{Q_f} \Theta_q^f(\boldsymbol{\mu}) f^q(v), \end{cases} \tag{1.2.1}$$

where $\Theta_q^a : \mathcal{D} \to \mathbb{R}$ and $\Theta_q^f : \mathcal{D} \to \mathbb{R}$ are given smooth functions.

Now we have all the elements to state the general problem (1.1.1) in our case:

$$\text{Given } \boldsymbol{\mu} \in \mathcal{D}, \text{ find } s(\boldsymbol{\mu}) = l(u(\boldsymbol{\mu})), \tag{1.2.2a}$$

where $u(\boldsymbol{\mu})$ is the solution to the following elliptic equation:

$$a(u(\boldsymbol{\mu}),v,\boldsymbol{\mu}) = F(v,\boldsymbol{\mu}) \quad \forall\, v \in X. \tag{1.2.2b}$$

We introduce now a suitable inner product and we define the continuity and the coercivity constants. In the analysis below, the following inner product is selected[6]:

$$(w,v)_X = a_S(w,v;\bar{\boldsymbol{\mu}}) + \tau(w,v)_{L^2(\Omega)}, \quad \|w\|_X^2 = (w,w)_X \quad \forall\, w,v \in X, \quad \tau = \inf_{v \in X} \frac{a_S(v,v;\bar{\boldsymbol{\mu}})}{\|v\|_{L^2(\Omega)}^2} \tag{1.2.3}$$

[4] See [106] for more references.

[5] For a survey on the method we mainly refer, e.g., to [93, 106, 80].

[6] This choice is justified by the Successive Constraint Method proposed in section 1.5.

where $\bar{\boldsymbol{\mu}} \in \mathcal{D}$ and $a_S(\cdot,\cdot,\bar{\boldsymbol{\mu}})$ is the symmetric part of the bilinear form.

Given $\boldsymbol{\mu} \in \mathcal{D}$ we define the following constants:

$$\begin{cases} \alpha_a(\boldsymbol{\mu}) = \inf\limits_{w\in X} \dfrac{a(w,w,\boldsymbol{\mu})}{\|w\|_X^2}, \\ \gamma_a(\boldsymbol{\mu}) = \sup\limits_{w\in X}\sup\limits_{v\in X} \dfrac{a(w,v,\boldsymbol{\mu})}{\|w\|_X\|v\|_Y}, \\ \gamma_f(\boldsymbol{\mu}) = \sup\limits_{w\in X} \dfrac{F(w,\boldsymbol{\mu})}{\|w\|_X}, \end{cases} \tag{1.2.4}$$

i.e., the coercivity and continuity constants associated with $a(\cdot,\cdot,\boldsymbol{\mu})$ and the continuity constant associated with $F(\cdot,\boldsymbol{\mu})$.

The coercivity constant is supposed to be uniformly strictly positive (i.e., $\alpha_a(\boldsymbol{\mu}) \geq \alpha_0 > 0$, $\forall\, \boldsymbol{\mu} \in \mathcal{D}$). In addition, in order to guarantee that the RB approximation is stable for $\mathcal{N} \to \infty$, we require that all the constants do not depend on the truth approximation chosen.

Thanks to the hypotheses above, the state problem is well-posed[7] for all the values of the parameter.

1.2.2 Reduced model

Let us consider the low dimensional space $X_N = \text{span}\{\zeta_j \in X : j = 1,\cdots,N\}$ where $\{\zeta_j\}_{j=1}^N$ is an orthonormal basis with respect to the inner product $(\cdot,\cdot)_X$ defined in (1.2.3).

By projecting equation (1.2.2b) onto the reduced space X_N, the output can be approximated through:

$$s_{RB,N}((\boldsymbol{\mu}) = l(u_{RB,N}(\boldsymbol{\mu})) \tag{1.2.5a}$$

where $u_{RB,N}(\boldsymbol{\mu}) \in X_N$ is the solution to[8]

$$a(u_{RB,N}(\boldsymbol{\mu}), v, \boldsymbol{\mu}) = F(v,\boldsymbol{\mu}) \quad \forall\, v \in X_N. \tag{1.2.5b}$$

Expanding the RB solution as

$$u_{RB,N}(\boldsymbol{\mu}) = \sum_{j=1}^{N} u_{N,j}(\boldsymbol{\mu})\zeta_j;$$

we can restate (1.2.5) in the following algebraic form:

$$\begin{cases} s_{RB,N}(\boldsymbol{\mu}) = \mathbf{l}^T \mathbf{u}_N(\boldsymbol{\mu}) \\ (\mathbf{l})_m = l(\zeta_m) \end{cases}$$

$$\begin{cases} A_N(\boldsymbol{\mu})\mathbf{u}_N(\boldsymbol{\mu}) = \mathbf{F}_N(\boldsymbol{\mu}) \\ (A_N)_{m,n} = a(\zeta_n,\zeta_m;\boldsymbol{\mu}), \quad (\mathbf{F}_N)_m = F(\zeta_m;\boldsymbol{\mu}) \end{cases}$$

Thanks to the parametric affinity of the linear forms, (1.2.1), we obtain the final algebraic formulation of the reduced model:

$$\text{Given } \boldsymbol{\mu} \in \mathcal{D} \text{ find } s_{RB,N}(\boldsymbol{\mu}) = \mathbf{l}^T \mathbf{u}_N(\boldsymbol{\mu}) \tag{1.2.6a}$$

[7] The proof is a straightforward application of Lax-Milgram Lemma [109].

[8] In the present work we always refer to the solution of the reduced problem as to $u_{RB,N}(\boldsymbol{\mu})$.

where $\mathbf{u}_N(\boldsymbol{\mu}) \in \mathbb{R}^N$ is the solution to the following linear system:

$$\begin{cases} \sum_{q=1}^{Q_a} \Theta_k^a(\boldsymbol{\mu}) A^q \mathbf{u}_N(\boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \Theta_k^f(\boldsymbol{\mu}) \mathbf{F}^q(\boldsymbol{\mu}) \\ (A^q)_{m,n} = a^q(\zeta_n, \zeta_m) \quad (\mathbf{F}^q)_m = f^q(\zeta_k) \end{cases} \tag{1.2.6b}$$

Equation (1.2.6) guarantees an efficient offline-online decomposition. In fact in the offline stage, A^q, $\mathbf{F}^q$ and $\mathbf{l}$ are built. In the online stage we just assemble the reduced matrix and vector ($\mathcal{O}(Q_a N^2 + Q_f N)$), solve the system and compute the output $\mathcal{O}(N^3 + N)$. Also the amount of memory required is modest: we just need for Q_a $N \times N$ matrices and $Q_f + 1$ N-dimensional vectors.

We conclude this section with some remarks.

Remark 1.1. *Thanks to the choice of the orthonormal basis $\{\zeta_n\}_{n=1}^N$ it is easy to verify that*[9]

$$cond(A_N(\boldsymbol{\mu})) \leq \frac{\gamma_a(\boldsymbol{\mu})}{\alpha_a(\boldsymbol{\mu})}. \tag{1.2.7}$$

This guarantees that the reduced problem is well-conditioned also when N grows up.

Remark 1.2. *Thanks to the well-known Galerkin orthogonality it is possible to prove that:*

$$\|u(\boldsymbol{\mu}) - u_{RB,N}(\boldsymbol{\mu})\|_X \leq \frac{\gamma_a(\boldsymbol{\mu})}{\alpha_a(\boldsymbol{\mu})} \inf_{w_N \in X_N} \|u(\boldsymbol{\mu}) - w_N\|_X \tag{1.2.8a}$$

and so that:

$$|s(\boldsymbol{\mu}) - s_{RB,N}(\boldsymbol{\mu})| \leq \|l\|_{X'} \frac{\gamma_a(\boldsymbol{\mu})}{\alpha_a(\boldsymbol{\mu})} \inf_{w_N \in X_N} \|u(\boldsymbol{\mu}) - w_N\|_X. \tag{1.2.8b}$$

We observe that, due to the hierarchical property (1.1.5), the error bounds are monotone with respect to the dimension of the reduced space X_N. If the problem is compliant *(i.e., $a(\cdot, \cdot, \boldsymbol{\mu})$ is symmetric and $l = F$), it is possible to prove that*[10]*:*

$$\|u(\boldsymbol{\mu}) - u_{RB,N}(\boldsymbol{\mu})\|_X \leq \sqrt{\frac{\gamma_a(\boldsymbol{\mu})}{\alpha_a(\boldsymbol{\mu})}} \inf_{w_N \in X_N} \|u(\boldsymbol{\mu}) - w_N\|_X \tag{1.2.9a}$$

and so that:

$$0 \leq s(\boldsymbol{\mu}) - s_{RB,N}(\boldsymbol{\mu}) \leq \gamma_a(\boldsymbol{\mu}) \inf_{w_N \in X_N} \|u(\boldsymbol{\mu}) - w_N\|_X^2. \tag{1.2.9b}$$

Remark 1.1 gives reasons for the orthonormalization procedure in section 1.4 while both the remarks justify the application of Galerkin projection: the methodology is capable to provide a quasi-optimal approximation of the solution in a stable and automatic way; in addition, as we explain in section 1.5, it is ideally suited for the output error estimation. On the other hand, other techniques, such as the *output interpolation*, can even provide better results -especially when the dimension P of the parameter set is low- but the selection of an efficient approximate output and the a posteriori error estimation are much more involved[11].

We observe that the output error estimate can be improved through a primal-dual approach as we will discuss in section 1.5.

[9] For the proof see [93] proposition 3B.

[10] The proofs are straightforward and are contained in [93] proposition 3A.

[11] See the discussion in [106] section 8.1.5 for further details and some examples.

1.3 Reduced basis formulation for parabolic problems

In this section the Reduced Basis Method is applied to an input-output relationship based on a parabolic state equation. In [120] a new approach based on the Petrov-Galerkin formulation, [97, 21], has been proposed. As we will explain, the parabolic equation is restated as a general non-coercive problem in the Babuska framework [4]: this simplifies the a posteriori error analysis and permits to state a Galerkin optimality result in a natural way. In order to follow this approach, we first review the Petrov-Galerkin formulation for a parabolic problem and then we introduce the Reduced Basis method.

1.3.1 Petrov-Galerkin formulation for parabolic linear equations

Let (V, H, V') be a Hilbert triplet where $H = L^2(\Omega)$ and $V \hookrightarrow H \hookrightarrow V'$ and both the embeddings are dense and compact. We define $A : V \to V'$ such that $\langle Au, v\rangle_{V'\times V} = a(u, v)$ where $a : V \times V \to \mathbb{R}$ is a γ_a-continuous and (λ_a, α_a)-weakly coercive bilinear form:

$$\begin{cases} |a(\phi,\psi)| \le \gamma_a \|\psi\|_V \|\phi\|_V & \forall\, \phi, \psi \in V, \\ |a(\phi,\phi)| + \lambda_a \|\phi\|_H^2 \ge \alpha_a \|\phi\|_V^2 & \forall\, \phi \in V. \end{cases} \tag{1.3.1}$$

Finally we consider $F : [0, T] \to V'$.

Let us consider the following linear parabolic equation written in an operatorial form:

$$\begin{cases} \dot{u}(t) + Au(t) = F(t) \text{ in } V \quad \text{for a.e.} t \in I = (0, T] \\ u(0) = 0 \text{ in } H. \end{cases} \tag{1.3.2}$$

In order to introduce the Petrov Galerkin formulation, we have to define the variational equivalent to (1.3.2). First of all we define the following spaces[12]

$$\begin{cases} \mathcal{X} = \{v \in L^2(I; V) : \dot{v} \in L^2(I; V')\}, \\ \mathcal{Y} = L^2(I; V), \end{cases} \tag{1.3.3}$$

and the following bilinear and linear forms:

$$\begin{cases} b : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} & b(u, v) = \int_I \langle \dot{u}(t), v(t)\rangle_{V'\times V}\, dt + \int_I a(u(t), v(t))\, dt, \\ f : \mathcal{Y} \to \mathbb{R} & f(v) = \int_I \langle F(t), v(t)\rangle_{V'\times V}\, dt. \end{cases} . \tag{1.3.4}$$

Therefore, problem (1.3.2) can be restated in the following variational form:

$$\text{Find } u \in \mathcal{X} \text{ such that: } b(u, v) = f(v) \quad \forall\, v \in \mathcal{Y}. \tag{1.3.5}$$

We are now ready to introduce the Petrov Galerkin formulation. Following [120], we define the spatial and temporal triangulations $\mathcal{T}_h^{space}$ and $\mathcal{T}_{\Delta t}^{time} = \{k\Delta t\}_{k=1}^{\mathbb{K}}$ and, subsequently the following finite element spaces:

$$\begin{cases} S_{\Delta t} := \{\sigma \in C([0, T]) : \sigma_{|(t^k, t^{k+1})} \in P^1(t^k, t^{k+1})\ \forall\, k = 0, \cdots, \mathbb{K} - 1\}, \\ Q_{\Delta t} := \{\tau \in L^\infty(0, T) : \tau_{|(t^k, t^{k+1})} \in P^0(t^k, t^{k+1})\ \forall\, k = 0, \cdots, \mathbb{K} - 1\}, \\ V_h := \{\phi \in C(\bar{\Omega}) : \phi_{|K} \in P^1(K)\ \forall\, K \in \mathcal{T}_h^{space}\}. \end{cases} \tag{1.3.6}$$

[12] We refer to [35] for further details about time dependent FE spaces.

Finally the space-time finite element spaces are built as resultants of the tensor product between the above spaces, i.e.:

$$\begin{cases} \mathcal{X}_\delta := V_h \otimes S_{\Delta t} = \{\psi \otimes \phi : \phi \in V_h \ \psi \in S_{\Delta t}\}, \\ \mathcal{Y}_\delta := V_h \otimes Q_{\Delta t} = \{\tau \otimes \phi : \phi \in V_h \ \tau \in Q_{\Delta t}\}, \end{cases} \tag{1.3.7}$$

where $\delta = (\Delta t, h)$ and $\psi \otimes \phi(t, \mathbf{x}) = \psi(t)\phi(\mathbf{x})$. Now we have the elements to state the Petrov Galerkin approximation of (1.3.5):

$$\text{Find } u_\delta \in \mathcal{X}_\delta \text{ such that: } b(u_\delta, v_\delta) = f(v_\delta) \quad \forall\, v_\delta \in \mathcal{Y}_\delta \tag{1.3.8}$$

In order to write the algebraic formulation of (1.3.8), let $\{\phi_j\}_{j=1}^{\mathcal{N}}$ be the nodal basis for V_h with respect to $\mathcal{T}_h^{space}$, $\{\sigma^k\}_{k=1}^{\mathbb{K}}$ be the nodal basis for $S_{\Delta t}$ with respect to $\mathcal{T}_{\Delta t}^{time}$ and finally $\{\tau^k = \chi_{(t^{k-1},t^k)}\}_{k=1}^{\mathbb{K}}$ be the basis for $Q_{\Delta t}$.

With this notation, problem (1.3.8) could be rewritten as:

$$\text{Find } u_\delta \in \mathcal{X}_\delta \text{ such that: } b(u_\delta, \tau^l \otimes \phi_j) = f(\tau^l \otimes \phi_j) \quad \forall\, j = 1, \cdots, \mathcal{N} \ \forall\, l = 1, \cdots, \mathbb{K}. \tag{1.3.9}$$

If we use a trapezoidal approximation of the right-hand side temporal integration:

$$f(\tau^l \otimes \phi_j) = \int_I \langle F(t), \tau^l \otimes \phi_j \rangle_{V' \times V}\, dt = \int_{t^{l-1}}^{t^l} \langle F(t), \phi_j \rangle_{V' \times V}\, dt \simeq \frac{\Delta t}{2} \langle F(t^l) + F(t^{l-1}), \phi_j \rangle_{V' \times V},$$

and we expand the solution as $u_\delta = \sum_{j,k} u_{\delta,j}^k \sigma^k \otimes \phi_j$, it is easy to verify (see [120]) that the (1.3.8) is equivalent to:

$$\begin{cases} \frac{1}{\Delta t} M(\mathbf{u}_\delta^l - \mathbf{u}_\delta^{l-1}) + \frac{1}{2} A(\mathbf{u}_\delta^l + \mathbf{u}_\delta^{l-1}) = \frac{1}{2}(\mathbf{F}^l + \mathbf{F}^{l-1}) \quad \text{for } l = 1, \cdots, \mathbb{K} \\ \mathbf{u}_\delta^0 = \mathbf{0} \\ M_{i,j} := (\phi_j, \phi_i)_{L^2(\Omega)} \quad A_{i,j} := a(\phi_j, \phi_i) \\ (\mathbf{F}^l)_j := \langle F(t^l), \phi_j \rangle_{V' \times V}. \end{cases} \tag{1.3.10}$$

Remark 1.3. *We observe that the scheme is equivalent to the well-known Crank Nicolson method.*

1.3.2 Formulation of the problem

After presenting the Petrov-Galerkin formulation for a general parabolic equation, we consider a parametric affine parabolic equation. First of all, we define the following forms:

$$\begin{cases} m : V' \times V \times \mathcal{D} \to \mathbb{R} & m(w, v, \boldsymbol{\mu}) = \sum_{q=1}^{Q_m} \Theta_q^m(\boldsymbol{\mu}) m^q(w, v) \\ a : V \times V \times \mathcal{D} \to \mathbb{R} & a(w, v, \boldsymbol{\mu}) = \sum_{q=1}^{Q_a} \Theta_q^a(\boldsymbol{\mu}) a^q(w, v) \\ F : [0, T] \times V \times \mathcal{D} \to \mathbb{R} & F(t, w, \boldsymbol{\mu}) = g(t) \sum_{q=1}^{Q_f} \Theta_q^f(\boldsymbol{\mu}) F^q(w), \end{cases} \tag{1.3.11}$$

where the bilinear form $m(\cdot, \cdot, \boldsymbol{\mu})$ is related to the mass matrix. In the following we will assume that $m(\cdot, \cdot, \boldsymbol{\mu})$ be $\gamma_m(\boldsymbol{\mu})$-continuous and inf-sup stable, uniformly with respect to the parameter (say $\sigma(\boldsymbol{\mu}) \geq \sigma_0 > 0$). Moreover $a(\cdot, \cdot, \boldsymbol{\mu})$ is assumed to be $\gamma_a(\boldsymbol{\mu})$-continuous and $(\lambda_a(\boldsymbol{\mu}), \alpha_a(\boldsymbol{\mu}))$-weakly coercive, uniformly with respect to the parameter. As we

did in the elliptic case, we assume that all the constants do not depend on the truth approximation chosen.

With the notation of the precedent paragraph, the problem can now be stated as:

$$\text{Find } u_\delta(\boldsymbol{\mu}) \in \mathcal{X}_\delta \text{ such that: } \begin{cases} b(u_\delta(\boldsymbol{\mu}), v_\delta, \boldsymbol{\mu}) = f(v_\delta, \boldsymbol{\mu}) \quad \forall\, v_\delta \in \mathcal{Y}_\delta \\ u_\delta(\boldsymbol{\mu})(0) = 0 \\ b(w, v, \boldsymbol{\mu}) = m(\dot{w}, v, \boldsymbol{\mu}) + a(w, v, \boldsymbol{\mu}) \\ f(w, \boldsymbol{\mu}) = \int_I F(t, w, \boldsymbol{\mu}) \end{cases} . \tag{1.3.12}$$

Using the Crank Nicolson interpretation, it is easy to write the problem in an algebraic form (where $u_\delta(\boldsymbol{\mu}) = \sum_{j,k} u^k_{\delta,j}(\boldsymbol{\mu}) \sigma^k \otimes \phi_j$):

$$\begin{cases} \frac{1}{\Delta t} M(\boldsymbol{\mu})(\mathbf{u}^l_\delta(\boldsymbol{\mu}) - \mathbf{u}^{l-1}_\delta(\boldsymbol{\mu})) + \frac{1}{2} A(\boldsymbol{\mu})(\mathbf{u}^l_\delta(\boldsymbol{\mu}) + \mathbf{u}^{l-1}_\delta(\boldsymbol{\mu})) = \frac{g(t^{l-1}) + g(t^l)}{2} \mathbf{F}(\boldsymbol{\mu}) \\ \mathbf{u}^0_\delta(\boldsymbol{\mu}) = \mathbf{0} \\ M(\boldsymbol{\mu}) = \sum_{q=1}^{Q_m} \Theta^m_q(\boldsymbol{\mu}) M^q \quad M^q_{i,j} = m^q(\phi_j, \phi_i) \\ A(\boldsymbol{\mu}) = \sum_{q=1}^{Q_a} \Theta^a_q(\boldsymbol{\mu}) A^q \quad A^q_{i,j} = a^q(\phi_j, \phi_i) \\ \mathbf{F}(\boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \Theta^f_q(\boldsymbol{\mu}) \mathbf{F}^q \quad (\mathbf{F}^q)_j = F^q(\phi_j). \end{cases} \tag{1.3.13}$$

If we introduce the output

$$J : \mathcal{X}_\delta \to \mathbb{R} \qquad J(v) = \int_I l(v(t))\, dt \quad l \in V', \tag{1.3.14}$$

the input-output relationship (1.1.1) becomes:

$$\text{Given } \boldsymbol{\mu} \in \mathcal{D}, \text{ find } s_\delta(\boldsymbol{\mu}) := J(u_\delta(\boldsymbol{\mu})) \text{ where } u_\delta(\boldsymbol{\mu}) \text{ is the solution to (1.3.13)} . \tag{1.3.15}$$

As we did in the elliptic case, given $\boldsymbol{\mu} \in \mathcal{D}$ we introduce the following scalar products[13]:

$$(w, v)_{\mathcal{X},\delta} = (\dot{w}, \dot{v})_{L^2(I;V')} + (w, v)_{L^2(I;V)} + (w(T), v(T))_{L^2(\Omega)}, \quad \|w\|_{\mathcal{X},\delta} = \sqrt{(w, w)_{\mathcal{X},\delta}} \tag{1.3.16a}$$

and

$$(w, v)_{\mathcal{Y},\delta} = (w, v)_{L^2(I;V)}, \quad \|w\|_{\mathcal{Y},\delta} = \sqrt{(w, w)_{\mathcal{Y},\delta}} \tag{1.3.16b}$$

and subsequently the following constants:

$$\begin{cases} \beta_{b,\delta}(\boldsymbol{\mu}) = \inf\limits_{w \in \mathcal{X}_\delta} \sup\limits_{v \in \mathcal{Y}_\delta} \dfrac{b(w, v, \boldsymbol{\mu})}{\|w\|_{\mathcal{X},\delta} \|v\|_{\mathcal{Y},\delta}} \\ \gamma_{b,\delta}(\boldsymbol{\mu}) = \sup\limits_{w \in \mathcal{X}_\delta} \sup\limits_{v \in \mathcal{Y}_\delta} \dfrac{b(w, v, \boldsymbol{\mu})}{\|w\|_{\mathcal{X},\delta} \|v\|_{\mathcal{Y},\delta}}. \end{cases} \tag{1.3.17}$$

Like in the elliptic case, the inf-sup constant is supposed to be strictly positive uniformly with respect to the parameter (i.e., $\beta_{b,\delta}(\boldsymbol{\mu}) \geq \beta_{b,0} > 0$ for any $\boldsymbol{\mu} \in \mathcal{D}$).

Thanks to the hypotheses above, the well-known Babuska theorem (see [4]) guarantees the well-posedness of the problem.

[13] The choice, inspired by [120], simplifies the a posteriori analysis.

1.3.3 Reduced model

In this paragraph we introduce the standard RB approximation for problem (1.3.12). We use the variational formulation in order to prove the Galerkin optimality and the Crank-Nicolson interpretation (1.3.13) of our discrete problem to get an easy-to-implement formulation.

It is possible to consider a space-time RB approximation, [113], or only space RB approximations, [80, 49]. In the following the second option is chosen. Therefore, we consider

$$V_N := \text{span}(\zeta_j \in V : j = 1, \cdots, N) \quad (\zeta_i, \zeta_j)_{\mathcal{X},\delta} = \delta_{i,j},$$

and subsequently we introduce the RB sample and test spaces as:

$$\mathcal{X}_{\Delta t,N} := S_{\Delta t} \otimes V_N, \quad \mathcal{Y}_{\Delta t,N} := Q_{\Delta t} \otimes V_N. \tag{1.3.18}$$

Finally, we define $u_{RB,N}(\boldsymbol{\mu}) \in \mathcal{X}_{\Delta t,N}$ as the solution to the projected equation:

$$\text{Find } u_{RB,N}(\boldsymbol{\mu}) \in \mathcal{X}_{\Delta t,N} \text{ such that: } \begin{cases} b(u_{RB,N}(\boldsymbol{\mu}), v_\delta, \boldsymbol{\mu}) = f(v_\delta, \boldsymbol{\mu}) & \forall\, v_\delta \in \mathcal{Y}_{\Delta t,N} \\ u_\delta(\boldsymbol{\mu})(0) = 0. \end{cases} \tag{1.3.19}$$

Using the algebraic interpretation of the problem, it is easy to verify that the reduced problem can be formulated in the following way (with $u_{RB,N}(\boldsymbol{\mu}) = \sum_{j,k} u^k_{N,j}(\boldsymbol{\mu})\sigma^k \otimes \zeta_j$):

$$\begin{cases} \frac{1}{\Delta t} M_N(\boldsymbol{\mu})(\mathbf{u}^l_N(\boldsymbol{\mu}) - \mathbf{u}^{l-1}_N(\boldsymbol{\mu})) + \frac{1}{2} A_N(\boldsymbol{\mu})(\mathbf{u}^l_N(\boldsymbol{\mu}) + \mathbf{u}^{l-1}_N(\boldsymbol{\mu})) = \frac{g(t^{l-1})+g(t^l)}{2} \mathbf{F}_N(\boldsymbol{\mu}) \\ \mathbf{u}^0_N(\boldsymbol{\mu}) = \mathbf{0} \\ M_N(\boldsymbol{\mu}) = \sum_{q=1}^{Q_m} \Theta^m_q(\boldsymbol{\mu}) M^q_N \quad (M^q_N)_{i,j} = m^q(\zeta_j, \zeta_i) \\ A_N(\boldsymbol{\mu}) = \sum_{q=1}^{Q_a} \Theta^a_q(\boldsymbol{\mu}) A^q_N \quad (A^q_N)_{i,j} = a^q(\zeta_j, \zeta_i) \\ \mathbf{F}_N(\boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \Theta^f_q(\boldsymbol{\mu}) \mathbf{F}^q_N \quad (\mathbf{F}^q_N)_j = F^q_N(\zeta_j). \end{cases} \tag{1.3.20}$$

We can now define our reduced order approximation of the input-output relationship (1.3.15) in a variational formulation:

$$\text{Given } \boldsymbol{\mu} \in \mathcal{D}, \text{ find } s_{RB,N}(\boldsymbol{\mu}) := J(u_{RB,N}(\boldsymbol{\mu})) \text{ where } u_{RB,N}(\boldsymbol{\mu}) \text{ is the solution to (1.3.19)} \tag{1.3.21}$$

and in a completely algebraic one:

$$\begin{aligned} &\text{Given } \boldsymbol{\mu} \in \mathcal{D}, \text{ find } s_{RB,N}(\boldsymbol{\mu}) := \textstyle\sum_{k=1}^{\mathbb{K}} \frac{1}{2} \mathbf{l}^T (\mathbf{u}^{k-1}_N(\boldsymbol{\mu}) + \mathbf{u}^k_N(\boldsymbol{\mu})) \\ &\text{where } \{\mathbf{u}^k_N(\boldsymbol{\mu})\}_k \text{ is the solution to (1.3.20).} \end{aligned} \tag{1.3.22}$$

Equations (1.3.20)-(1.3.22) are ideally suited for the offline-online decomposition. In the offline stage M^q, A^q and $\mathbf{F}^q$ are computed. In the online stage, for each time step, we assemble the matrix and vector and then we solve the system and update the output. In conclusion the online operation count is: $\mathcal{O}\left(((Q_a + Q_m)N^2 + Q_f N + Q_f N^3 + N)\mathbb{K}\right)$.

In order to analyse the problem, we introduce the following discrete inf-sup and continuity constants associated with the involved spaces:

$$\begin{cases} \beta_{RB,N}(\boldsymbol{\mu}) = \inf\limits_{w\in\mathcal{X}_{\Delta t,N}} \sup\limits_{v\in\mathcal{Y}_{\Delta t,N}} \dfrac{b(w,v,\boldsymbol{\mu})}{\|w\|_{\mathcal{X},\delta}\|v\|_{\mathcal{Y},\delta}} \\ \gamma_{RB,N}(\boldsymbol{\mu}) = \sup\limits_{w\in\mathcal{X}_{\Delta t,N}} \sup\limits_{v\in\mathcal{Y}_{\Delta t,N}} \dfrac{b(w,v,\boldsymbol{\mu})}{\|w\|_{\mathcal{X},\delta}\|v\|_{\mathcal{Y},\delta}} \end{cases} \tag{1.3.23}$$

We observe that, thanks to the formulation chosen, the Galerkin projection is formally equivalent to the one proposed in the elliptic case. In addition it is possible to prove the following optimality estimate that is the analogous, for non-coercive problems[14], of the one presented in Remark 1.2.

Lemma 1.1. *Let $u_{RB,N}(\boldsymbol{\mu})$ be the solution to (1.3.19) and $u_\delta(\boldsymbol{\mu})$ be the solution to (1.3.12); then the following estimate holds:*

$$\|u_\delta(\boldsymbol{\mu}) - u_{RB,N}(\boldsymbol{\mu})\|_{\mathcal{X},\delta} \le \left(1 + \frac{\gamma_b(\boldsymbol{\mu})}{\beta_{RB,N}(\boldsymbol{\mu})}\right) \inf_{w\in\mathcal{X}_\delta} \|u_\delta(\boldsymbol{\mu}) - w\|_{\mathcal{X},\delta} \tag{1.3.24}$$

where $\gamma_b(\boldsymbol{\mu})$ is defined as in (1.3.17) and $\beta_{RB,N}(\boldsymbol{\mu})$ as in (1.3.23).

Consequently, we have:

$$|s_\delta(\boldsymbol{\mu}) - s_{RB,N}(\boldsymbol{\mu})| \le \sqrt{T}\|l\|_{V'} \left(1 + \frac{\gamma_b(\boldsymbol{\mu})}{\beta_{RB,N}(\boldsymbol{\mu})}\right) \inf_{w\in\mathcal{X}_\delta} \|u_\delta(\boldsymbol{\mu}) - w\|_{\mathcal{X},\delta}. \tag{1.3.25}$$

Proof. Let $w \in \mathcal{X}_{\Delta t,N}$. From the definition of $\beta_{RB,N}(\boldsymbol{\mu})$, there exists $\xi \in \mathcal{Y}_{\Delta t,N}$ such that[15]

$$\beta_{RB,N}\|u_{RB,N}(\boldsymbol{\mu}) - w\|_{\mathcal{X},\delta}\|\xi\|_{\mathcal{Y}} \le b(u_{RB,N}(\boldsymbol{\mu}) - w, \xi, \boldsymbol{\mu})$$

Using the Galerkin orthogonality[16]:

$$b(u_{RB,N}(\boldsymbol{\mu}) - w, \xi, \boldsymbol{\mu}) = b(u_\delta(\boldsymbol{\mu}) - w, \xi, \boldsymbol{\mu}) \le \gamma_b(\boldsymbol{\mu})\|u_\delta(\boldsymbol{\mu}) - w\|_{\mathcal{X},\delta}\|\xi\|_{\mathcal{Y}}$$

where in the last inequality the definition of $\gamma_b(\boldsymbol{\mu})$ in (1.3.17) has been used. Thus in conclusion we obtain:

$$\|u_{RB,N}(\boldsymbol{\mu}) - u_\delta(\boldsymbol{\mu})\|_{\mathcal{X},\delta} \le \|u_{RB,N}(\boldsymbol{\mu}) - w\|_{\mathcal{X},\delta} + \|u_\delta(\boldsymbol{\mu}) - w\|_{\mathcal{X},\delta} \le \left(1 + \frac{\gamma_b(\boldsymbol{\mu})}{\beta_{RB,N}(\boldsymbol{\mu})}\right)\|u_\delta(\boldsymbol{\mu}) - w\|_{\mathcal{X},\delta}$$

Due to the arbitrariness of w, the first estimate holds. Inequality (1.3.25) follows in a straightforward way from (1.3.24). □

1.4 Sampling strategy

It is absolutely evident that the effectiveness of the RB method strongly depends on the approximation property of the RB space. For this reason, in recent years, a great attention has been paid to the definition of effective sampling strategies able to generate adaptive and

[14] The lemma below is a simplified version of Proposition 4 in [106].

[15] Due to the fact that the spaces are finite dimensional, it is absolutely standard to verify the existence of such ξ.

[16] In formulas:

$$b(u_{RB,N}(\boldsymbol{\mu}) - u_\delta(\boldsymbol{\mu}), v, \boldsymbol{\mu}) = 0 \quad \forall\, v \in \mathcal{Y}_{\Delta t,N}$$

hierarchical spaces which are automatically tailored to the particular problem of interest. Such techniques are particularly important when the number of parameters is higher than one[17]: in fact in this case standard tensor product approaches are prohibitively expensive.

Concerning the problems considered in this chapter, the main techniques usually employed are the so-called *Greedy* algorithm, first proposed in [124], for elliptic problems and the *POD-Greedy* algorithm, [49], for parabolic equations.

An exhaustive presentation of the topic is absolutely beyond the purposes of this chapter: we just summarize the main features of the different approaches while referring to the articles in bibliography for further discussions.

We first introduce some notation. We denote by Ξ a finite sample of points in $\mathcal{D}$ (*train set*) and we define $n_{train} := |\Xi|$ the cardinality of the train set[18]. As explained above, the sample to approximate is the truth manifold $\mathcal{M}^{\mathcal{N}}$ defined in (1.1.4) and the associated spanned space: $\text{span}\left\{\mathcal{M}^{\mathcal{N}}\right\}$. For the sake of simplicity in the following the superscript $\mathcal{N}$ is omitted.

We refer to $(\cdot,\cdot)$ and consequently to $\|\cdot\|$ as to the inner product and the associated norm defined on $\text{span}\left\{\mathcal{M}^{\mathcal{N}}\right\}$.

Following [93], we focus on Lagrangian samples, thus the search of the optimal subspace $X_N = \text{span}\left\{u(\boldsymbol{\mu}^i);\, 1 \le i \le N\right\} \subset \text{span}\left\{\mathcal{M}^{\mathcal{N}}\right\}$ will be associated with the search of the optimal sample $S_N := \{\boldsymbol{\mu}^1, \cdots, \boldsymbol{\mu}^N\} \subset \Xi$. Finally, $\bar{N}_{max}$ is defined as the maximum dimension of the reduced space.

1.4.1 Orthogonalization procedures

As explained in Remark 1.1, it is extremely important to consider an orthonormal basis for the reduced space. Starting from the basis $\{u^k\}_k := \{u(\boldsymbol{\mu}^k)\}_k$, an efficient algorithm to obtain an orthonormal basis $\{\zeta_k\}_k$ is the well-known Gramm-Schmidt algorithm [42]. Algorithm 1 summarizes the procedure.

Algorithm 1 Gramm-Schmidt orthogonalization procedure: $[\{\zeta_k\}_{k=1}^K] =$ `Gramm-Schmidt`$(\{u^k\}_{k=1}^K)$

$\zeta_1 := \frac{u^1}{\|u^1\|}$;
for $k = 2 : K$ **do**
  $z^k := u^k - \sum_{m=1}^{k-1}(\zeta_m, u^k)\zeta_m$
  $\zeta_k = \frac{z^k}{\|z^k\|}$
end for

1.4.2 Kolmogorov N-Width

In order to establish a benchmark, we define the so-called Kolmorogov "N-width" [77, 59]:

$$\bar{\epsilon}^{Kol}(X) := \sup_{\boldsymbol{\mu}\in\Xi} \inf_{w\in X} \|u(\boldsymbol{\mu}) - w\| \qquad X \subset \text{span}\left\{\mathcal{M}^{\mathcal{N}}\right\}, \tag{1.4.1}$$

[17] In [93] section 3.5, a generic a priori "quasi hierarchical" sequence of spaces is provided for parametrically coercive problems that depend on a single ($P = 1$) parameter. However, in higher dimensions no such recipe is available.

[18] These samples are typically chosen by Monte Carlo methods with respect to a uniform or log-uniform density. n_{train} should be quite large (when $P > 1$, easily as large or larger than 10^6) in order to be insensitive to further refinement of the parameter sample.

where the optimal "Kolmogorov spaces" X_N^{Kol} are given by:

$$X_N^{Kol} = \arg \inf_{X_N \subset \text{span}\{\mathcal{M}^{\mathcal{N}}\}} \inf_{\dim X_N = N} \bar{\epsilon}^{Kol}(X) \tag{1.4.2}$$

We observe that the "Kolmogorov spaces" X_N^{Kol} guarantee an optimal coverage property with respect to the "$L^\infty(\Xi)$" norm in parameter of the best fit to the field variable. For this reason X_N^{Kol} can be addressed as *the best N-dimensional subspace to approximate* $u(\boldsymbol{\mu})$ *for all* $\boldsymbol{\mu} \in \mathcal{D}$.

However, such sequence is not computable in practice because it cannot be assumed to be hierarchical and the optimization procedure in (1.4.2) is combinatorially difficult with respect to n_{train} and requires the computation of n_{train} FE solutions.

1.4.3 Proper Orthogonal Decomposition

In this paragraph the Proper Orthogonal Decomposition (POD) is introduced. The method was originally introduced in probability theory as Karhunen-Loeve transformation [62, 74] or, even earlier, in statistics as Hotelling transformation [50]. Since then, it has also been denoted as principal component analysis [31] in pattern analysis. In the field of numerical analysis for partial differential equations, it has been applied -with the notion of POD- to a variety of problems such as turbulent flows [76], fluid structure interaction [28] and non linear structural mechanics [66]. The technique is typically applied within the time-domain Reduced Order Modeling (ROM) [114, 100, 101], but it can also be applied within the parametric context [114, 47] . Here, the presentation is limited to the finite dimensional case; however, it is possible to extend the methodology to the approximation of finite and infinite dimensional manifolds that belong to infinite dimensional Hilbert spaces, [51].

With respect to the Kolmogorov-N-width, we replace the $L^\infty(\Xi)$ norm of (1.4.1) with the weaker $L^2(\Xi)$ norm:

$$\bar{\epsilon}^{POD}(X) := \sqrt{\frac{1}{n_{train}} \sum_{\boldsymbol{\mu} \in \Xi} \inf_{w \in X} \|u(\boldsymbol{\mu}) - w\|^2} \quad X \subset \text{span}\left\{\mathcal{M}^{\mathcal{N}}\right\}. \tag{1.4.3}$$

The optimal POD spaces X_N^{POD} are consequently defined as:

$$X_N^{POD} := \arg \inf_{X_N \subset \text{span}\{\mathcal{M}^{\mathcal{N}}\},\, \dim X_N = N} \bar{\epsilon}^{POD}(X_N). \tag{1.4.4}$$

Thanks to the following lemma, the optimal space can be identified through the solution to a suitable symmetric positive semidefinite eigenproblem.

Lemma 1.2. *Let* $\boldsymbol{\mu}^1, \cdots, \boldsymbol{\mu}^{n_{train}}$ *be a given ordering of* Ξ *(i.e.* $\Xi := \{\boldsymbol{\mu}^1, \cdots, \boldsymbol{\mu}^{n_{train}}\}$*) and let* $C \in \mathbb{R}^{n_{train} \times n_{train}}$ *be defined as:*

$$C_{i,j} := \frac{1}{n_{train}} (u(\boldsymbol{\mu}^i), u(\boldsymbol{\mu}^j)). \tag{1.4.5}$$

Furthermore, we refer to $(\boldsymbol{\Psi}^i, \lambda^i) \in \mathbb{R}^{n_{train}} \times \mathbb{R}$ *as to the eigenpairs of the following symmetric semidefinite positive eigenproblem:*

$$C\boldsymbol{\Psi}^i = \lambda^i \boldsymbol{\Psi}^i \quad \lambda^i \geq \lambda^{i+1} \geq 0. \tag{1.4.6}$$

Then the N-dimensional POD space defined in (1.4.4) coincides with:

$$X_N^{POD} = span\left\{\psi_i^{POD} \in span\{\mathcal{M}\};\, 1 \leq i \leq N\right\} \quad \psi_i^{POD} := \sum_{n=1}^{n_{train}} \Psi_n^i u(\boldsymbol{\mu}^i) \tag{1.4.7}$$

and the approximation error (1.4.3)

$$\bar{\epsilon}^{POD}(X_N^{POD}) = \sqrt{\sum_{k=N+1}^{n_{train}} \lambda^i}. \tag{1.4.8}$$

For the proof we refer, for instance, to [30] section 2. The matrix C in (1.4.6) is usually referred to as *Gramian* or *Kernel* matrix. Lemma 1.2 shows that POD yields hierarchical spaces with an optimality covering property at non-combinatorial Offline cost. On the other hand, in order to build the Gramian matrix it is necessary to perform n_{train} FE problems to compute $u(\boldsymbol{\mu}^i)$ and n_{train}^2 inner products to form C. For this reason the method is computationally sustainable only when $\dim\mathcal{D}$ is small ($P = 1, 2$).

Estimate (1.4.8) is extremely important: it justifies the fact that the effectivity of POD depends uniquely on the eigenvalue spectrum of the Gramian matrix. We will come back to this topic later in the chapter.

The algorithm below summarizes the POD procedure.

Algorithm 2 POD sampling algorithm: $[\{\psi_i^{POD} \in \text{span}\{\mathcal{M}\}, 1 \le i \le N\}] = \texttt{POD}(\{u(\boldsymbol{\mu}^N \in \mathcal{M}, 1 \le N \le n_{train}\}, N)$

$C_{i,j} = \frac{1}{n_{train}}(u(\boldsymbol{\mu}^i), u(\boldsymbol{\mu}^j)) \quad 1 \le i, j \le n_{train}$:

Solve $C\boldsymbol{\Psi}^i = \lambda^i \boldsymbol{\Psi}^i$, ${\boldsymbol{\Psi}^i}^T C \boldsymbol{\Psi}^j = \frac{1}{n_{train}}\delta_{i,j}$ associated with the N largest eigenvalues of C.

Compute $\psi_i^{POD} = \sum_{k=1}^{n_{train}} \Psi_k^i u(\boldsymbol{\mu}^k)$ for $1 \le i \le N$

1.4.4 Greedy sampling

In information science a greedy algorithm is an algorithm that follows the problem solving heuristic of making the locally optimal choice at each stage with the hope of finding a global optimum[19].

Thus the crucial question is how, given $\boldsymbol{\mu}^1, \boldsymbol{\mu}^2, \cdots, \boldsymbol{\mu}^{N-1}$ parameters, we choose the next one, $\boldsymbol{\mu}^N$.

In our framework, it seems to be natural to consider the following criterion:

$$\boldsymbol{\mu}^N = \arg\max_{\boldsymbol{\mu}\in\Xi} \|u(\boldsymbol{\mu}) - u_{RB,N-1}(\boldsymbol{\mu})\|$$

where $u_{RB,N-1} : \mathcal{D} \to X_{N-1}^{Greedy} = \text{span}\{u(\boldsymbol{\mu}^j) : j = 1, \cdots, N-1\}$ is the solution to the reduced problem we are considering.

With respect to POD, at this level no significant computational saving is obtained: as before n_{train} FE problems have to be solved in order to define the new value of $\boldsymbol{\mu}$. However, if we recur to an efficient estimator $\Delta_{X_N^{Greedy}}(\boldsymbol{\mu})$ for the reduced basis error $\|u(\boldsymbol{\mu}) - u_{RB,N}(\boldsymbol{\mu})\|$ the offline computational effort is potentially hugely reduced: in fact the method allows to compute truth solutions/snapshots not for all points in Ξ, as in the POD context but only for the winning candidates $\boldsymbol{\mu}^N$. Since $N_{max} \ll n_{train}$ we are able to consider larger train samples Ξ and so also higher-than-one dimension parameter spaces can be taken into account.

[19] The definition is taken from `en.wikipedia.org/wiki/Greedy_algorithm`.

In section 1.5 we discuss how to build such estimator, here we write down the Pseudo-Code for the Greedy Algorithm: we assume that an initial sample $S_{N_0}^{Greedy}$ and the associated Lagrange space $X_{N_0}^{Greedy} = \text{span}\{u(\boldsymbol{\mu}^N); 1 \leq N \leq N_0\}$ are given. The termination condition of the Greedy algorithm is set through the tolerance ϵ_{tol} referred to as the $L^\infty(\Xi)$ approximation error.

Algorithm 3 Greedy sampling algorithm: $[X_{N_{max}}^{Greedy}] = \texttt{Greedy}(S_{N_0}^{Greedy}, \Xi, \epsilon_{tol}, \bar{N}_{max},)$

given $S_{N_0}^{Greedy} = \{\mu^N \;\; N = 1, \cdots, N_0\}$ and the associated $X_{N_0}^{Greedy}$:
for $N = N_0 + 1 : \bar{N}_{max}$ **do**
  $\boldsymbol{\mu}^N = \arg\max_{\mu \in \Xi} \Delta_{X_{N-1}^{Greedy}}(\boldsymbol{\mu})$
  $\epsilon_{N-1}^{Greedy} = \Delta_{X_{N-1}^{Greedy}}(\boldsymbol{\mu}^N)$
  if $\epsilon_{N-1}^{Greedy} \leq \epsilon_{tol}$ **then**
    $\bar{N}_{max} = N - 1$
    `break`
  end if
  $S_N^{Greedy} = S_{N-1}^{Greedy} \cup \{\boldsymbol{\mu}^N\}$
  $X_N^{Greedy} = X_{N-1}^{Greedy} \cup \text{span}\{u(\boldsymbol{\mu}^N)\}$
end for

We observe that the algorithm provides hierarchical spaces and it is optimized with respect to the strong $L^\infty(\Xi)$ norm. Numerical simulations show that -although it is a short horizon heuristic that is sub-optimal with respect to the $L^\infty(\Xi)$ norm- in practical the approach provides rapidly convergent approximations.

As it is shown in [93], short horizon greedy and global POD approaches perform commensurately if measured in comparable norms: as might be forecast, each is better in the native norm on Ξ which defines the respective objective functions.

1.4.5 Sampling methods for elliptic and parabolic equations

After presenting the POD and the Greedy sampling strategies from a general point of view, we apply them to our framework.

The Greedy approach is the most efficient sampling strategy for stationary equations. As explained above, it is on one hand capable to provide rapidly convergent approximations and on the other hand - unlike POD - it does not suffer too much high dimension parameter samples.

The success of the Greedy approach originates from the absence of interactions between the RB approximations for different parameter values. In the time-domain context there are of course interactions between different times and, as a result, the Greedy algorithm may not perform as well as the more global POD optimization procedure, able to capture the interactions between the solution at different times.

This gives reasons for the use, in the parabolic case, of a mixed strategy[20] (the so-called *POD-Greedy*) that combines the POD in time- to adequately capture the causality associated with the evolution equation- with the Greedy procedure to manage the high dimension of parameter variation.

[20] As we stated above in section 1.3.3, here we focus on time-independent RB approximations for evolution equations. For this reason our discussion is limited to sampling approaches that aim at producing spatial approximation subspaces.

To explain the computational procedure[21], detailed in the algorithm 4, we first define $\boldsymbol{\mu}^1$ and ϵ_{tol} the initial sample point and a given termination tolerance, respectively. Furthermore, we define two suitable integers M_1 and M_2; M_1 represents the number of basis we extract at each step from the POD procedure and it is usually defined such that (1.4.8) be under a given tolerance. On the other hand, $M_2 \leq M_1$ is chosen to avoid duplication in the RB space.

Finally, as in the Greedy strategy $\Delta_X(\boldsymbol{\mu})$ is the a posteriori error estimator between the truth solution and reduced basis solution with respect to a suitable norm.

Algorithm 4 Pod Greedy sampling algorithm:
$[X^{PG}_{N_{max}}, S^{PG}] = \texttt{Greedy}(\boldsymbol{\mu}^1, \Xi, \epsilon_{tol}, \bar{N}_{max}, M_1, M_2)$

Set $S^{PG} = \{\boldsymbol{\mu}^1\}$, $\boldsymbol{\mu}^\star = \boldsymbol{\mu}^1$
while $N \leq \bar{N}_{max}$ **do**
  $[\{\psi_i \in \text{span}\{\mathcal{M}\}, 1 \leq i \leq M_1\}] = \texttt{POD}\ (\{u(\boldsymbol{\mu}^\star, t^k), 0 \leq k \leq \mathbb{K}\}, M_1)$

  $[\{\xi_i\}, 1 \leq i \leq N + M_2] = \texttt{POD}\ (X^{PG}_N \cup \{\psi_i \in \text{span}\{\mathcal{M}\}, 1 \leq i \leq M_1\}, N + M_2)$

  $N = N + M_2$
  $X^{PG}_N = \text{span}\{\xi_i\}$

  $\boldsymbol{\mu}^\star = \arg\max_{\boldsymbol{\mu} \in \Xi} \Delta_{X^{PG}_N}(\boldsymbol{\mu})$
  $S^{PG} = S^{PG} \cup \{\boldsymbol{\mu}^\star\}$
end while

Before concluding this section, we observe that, unlike in a pure POD approach, the operation count for the POD-Greedy algorithm is additive and not multiplicative in n_{train} and $\mathcal{N}$. As a result, in the latter approach n_{train} can be taken relatively large.

The following example shows the meaning of M_1 and M_2.

Example 1.1. *Let us consider the following manifold*

$$\mathcal{M} := \{\tilde{u} + \epsilon u(\boldsymbol{\mu}) : \quad \boldsymbol{\mu} \in \mathcal{D}\}$$

where $\tilde{u} : \Omega \to \mathbb{R}$ *and* $u : \mathcal{D} \times \Omega \times (0, T) \to \mathbb{R}$ *be smooth functions and* $|\epsilon| \ll 1$.

Then it is easy to observe that for all values of the parameter POD returns $\tilde{u}$ *as first eigenvalue. Therefore, if we consider* $M_1 = M_2$*, at each step of the POD Greedy algorithm we would try to add the same eigenvector: as a result, the algorithm would determine an ill-conditioned space or would generate an error.*

1.4.6 A brief overview on convergent rates of the Greedy and POD-Greedy methods

Since its introduction in the early 2000s, several numerical simulations have shown the approximation properties of the Greedy method[22]. However, the a priori convergence results are much more recent, [12, 8, 48]. Such results permit to compare the Greedy suboptimal approach with the benchmark Kolmogorov spaces: essentially, they all state

[21] As mentioned above, the POD Greedy algorithm was first proposed in [49]; the procedure here presented, taken from [80], is a slight modification of the original approach.

[22] See [106] for some examples in the stationary case.

that, if the error sequence of the Greedy or POD-Greedy algorithm is slowly decaying, then also the Kolmogorov N-width at a certain iteration must be large.

Here we limit to state the most recent result available for the Greedy sampling strategy for stationary coercive equations, [8]. In [48], the result is extended to the POD-Greedy algorithm for time dependent problems.

Theorem 1.1. *Let us suppose that the a posteriori error estimator $\Delta_{X_N} : \mathcal{D} \to \mathbb{R}$ in the algorithm 3 satisfies the following inequality:*

$$c\Delta_{X_N}(\boldsymbol{\mu}) \leq \|u(\boldsymbol{\mu}) - P_{X_N} u(\boldsymbol{\mu})\| \leq C\Delta_{X_N}(\boldsymbol{\mu}) \tag{1.4.9}$$

for some constants $c, C > 0$ and where $P_{X_N} : \mathcal{M} \to X_N$ is the projection operator.

We define

$$d_N(\mathcal{M}) = \bar{\epsilon}^{Kol}(X_N^{Kol}), \quad \sigma_N(\mathcal{M}) = \bar{\epsilon}^{Kol}(X_N^{Greedy}),$$

where $\bar{\epsilon}^{Kol}$ is defined in (1.4.1) and X_N^{Kol} as in (1.4.2).

Then if

$$d_N(\mathcal{M}) \leq M(N)^{-\alpha} \quad \forall\, N \geq 0,\ M, \alpha > 0$$

we have

$$\sigma_N(\mathcal{M}) \leq KM(N)^{-\alpha} \tag{1.4.10}$$

where $K = \sqrt{q}(4q)^{\alpha}$ and $q = \left[2^{\alpha+1}\frac{C}{c}\right]^2$.

Otherwise if:

$$d_N(\mathcal{M}) \leq Me^{-aN^{\alpha}} \quad \forall\, N \geq 0\ a, M, \alpha > 0$$

we have:

$$\sigma_N(\mathcal{M}) \leq \inf_{\theta \in (0,1)} K(\theta)\, M(\theta) e^{-k(\theta)N^{\tilde{\beta}}} \tag{1.4.11}$$

where $\tilde{\beta} = \frac{\alpha}{\alpha+1}$, $K(\theta) = \max\{e^{cN_0^{\tilde{\beta}}}, \sqrt{q}\}$, $k(\theta) = \min\{|log(\theta)|, (4q)^{-\alpha}a\}$, $q(\theta) = \left[\frac{2c}{\gamma\theta C}\right]^2$ and $N_0(\theta) = \left\lceil (8q)^{\alpha+1} \right\rceil$.

These results represent the theoretical foundation behind the Greedy and the POD-Greedy approaches; the convergence rate is associated with the approximation properties of Kolmogorov spaces.

For several classes of problems, we expect that the Kolmogorov N-width decay rate is very fast. However, especially in the time-dependent case, there are significant examples in which the convergence is extremely slow.

Example 1.2. ***(Example 3.4 from [48])*** *Let us consider the following non parametric advection problem:*

$$\begin{cases} \dfrac{\partial \psi}{\partial t} + \dfrac{\partial \psi}{\partial x} = 0 & (t, x) \in (0, K] \times (0, K+1) = (0, T] \times \Omega \\ \psi = \psi_0 & x \in \Omega. \end{cases} \tag{1.4.12}$$

We consider the mesh $t^k = k$ and $x_k = k$ for $k = 0, \cdots, K$ and the initial data $\psi_0(x_k) = \delta_{0k}$. Then an upwind finite difference discretization yields $u = \{\mathbf{u}^k\}_{k=0}^K$ with $u_i^k = \delta_{i,k}$.

It is straightforward to verify that in this case the Gramian matrix C is in spectral form and $\lambda_0 = \cdots = \lambda_K = 1$. Hence, at each step of the POD-Greedy algorithm, one single mode is inserted and the error decays by an identical decrement. We observe that we have linear convergence of the error with respect to the number of elements in the reduced basis.

We conclude this section with some general comments. The theoretical results presented above represent an important step towards the theoretical explanation about why the Greedy and the POD-Greedy algorithm work well in practice. On the other hand, a crucial question is to provide estimates of the Kolmogorov N-width decay rate for certain classes of time dependent parametric PDEs: the above example shows that there exist relevant cases of not decaying eigenvalue spectrum (transport of discontinuities).

1.5 A posteriori error estimation

In the previous sections, the necessity for an a posteriori error estimator, say Δ_{RB}, came up. One of the main goals of Reduced Basis is to decrease significantly the online computational effort without losing the reliability of the RB approximation. Furthermore, in section 1.4 we introduced the Greedy sampling strategy that takes advantage of an efficient and inexpensive error indicator in order to consider larger training sets $\Xi \subset \mathcal{D}$ and so to provide a better parameter space exploration at greatly reduced Offline computational cost.

The reasons above justify the following requirements on the error estimator:

- *Rigour*: the a posteriori error bound must be rigorous i.e.:

$$\|u(\boldsymbol{\mu}) - u_{RB}(\boldsymbol{\mu})\| \leq \Delta_{RB}(\boldsymbol{\mu}) \quad \forall\, \boldsymbol{\mu} \in \mathcal{D}. \tag{1.5.1}$$

  Although even non-rigorous indicators may be considered during the sampling, the rigour of the error bound is fundamental in order to assess the online reliability of the approximation.

- *Sharpness*: the a posteriori error estimation must be close to the real error i.e.:

$$\Delta_{RB}(\boldsymbol{\mu}) \leq C\|u(\boldsymbol{\mu}) - u_{RB}(\boldsymbol{\mu})\| \quad \forall\, \boldsymbol{\mu} \in \mathcal{D} \quad \text{for some } C > 1. \tag{1.5.2}$$

  As theorem 1.1 shows, an overly conservative error bound deteriorates the convergence order of the Greedy sampling- with respect to the equation (1.4.9) $\gamma = C$ - and so it easily determines inefficient approximation spaces.

- *Efficiency*: the aim at reducing the Online operation count and storage and the need of taking into account large training sets during the sampling stage justify the fact that the evaluation time of the estimator must be independent of the mesh size $\mathcal{N}$ and should be commensurate with the computational time associated with the RB output prediction.

In this section the a posteriori error estimation for elliptic and parabolic equations is dealt with. As regards elliptic problems, the discussion here presented is classic, [80, 93, 106]; on the other hand, the results for parabolic problems are more recent, [120].

In order to simplify the explanation and to express the main ideas, the discussion is organized as follows:

- first, the a posteriori error estimation is introduced in an abstract setting for a general non coercive linear problem;

- then, the a posteriori estimator for parametrically affine elliptic and parabolic equations is obtained;

- at the end of the section, we briefly address the problem of the output estimation, perhaps the main goal of the RB approach.

1.5.1 Residual-based a posteriori error estimator

Let us consider the following problem:

$$\text{Find } u \in X \text{ such that } b(u,v) = f(v) \quad \forall\, v \in Y; \tag{1.5.3}$$

where X, Y are Hilbert spaces, $f \in Y'$ and $b : X \times Y \to \mathbb{R}$ is a γ-continuous and inf-sup stable, with β the corresponding constant, bilinear form[23].

Now let us consider the subspaces $X_{RB} \subset X$ and $Y_{RB} \subset Y$ and we call u_{RB} the solution to the reduced problem:

$$\text{Find } u_{RB} \in X_{RB} \text{ such that } b(u_{RB},v) = f(v) \quad \forall\, v \in Y_{RB}. \tag{1.5.4}$$

Even for the reduced problem, we suppose that the bilinear form is still inf-sup stable, with β_{RB} the corresponding constant, with respect to the reduced spaces.

Under these hypotheses, problems (1.5.3) and (1.5.4) admit a unique solution that depends continuously on data.

Let us now consider the weak residual $r : Y \to \mathbb{R}$:

$$r(v) := f(v) - b(u_{RB},v) \tag{1.5.5}$$

It is absolutely straightforward that $r \in Y'$, so, for the well-known Riesz theorem, there exists $\hat{e} \in Y$ such that:

$$(\hat{e},v)_Y = r(v), \qquad \|\hat{e}\|_Y = \|r\|_{Y'}.$$

By substracting (1.5.4) to (1.5.3), we obtain:

$$b(u - u_{RB},v) = r(v).$$

And so for the well-known Babuska theorem, [4], we get:

$$\|u - u_{RB}\|_X \le \frac{1}{\beta}\|r\|_{Y'} = \frac{1}{\beta}\|\hat{e}\|_Y. \tag{1.5.6}$$

Estimate (1.5.6) gives reasons for the introduction of our residual based estimator:

$$\Delta_{RB} := \frac{1}{\beta_{LB}}\|\hat{e}\|_Y \text{ where } \beta_{LB} \le \beta. \tag{1.5.7}$$

The rigour of this estimator is guaranteed by (1.5.6); as regards the sharpness, if we suppose Y to be reflexive[24], the residual estimator satisfies (1.5.2) too. In fact, thanks to the reflexivity[25], $\exists\, v^\star \in Y$ such that $\|v^\star\|_Y = 1$, $r(v^\star) = \|r\|_{Y'}$. Thus:

$$\Delta_{RB} = \frac{1}{\beta_{LB}}\|r\|_{Y'} = \frac{1}{\beta_{LB}}r(v^\star) = \frac{1}{\beta_{LB}}b(u - u_{RB},v^\star) \le \frac{\gamma}{\beta_{LB}}\|u - u_{RB}\|_X \tag{1.5.8}$$

that proves (1.5.2) with $C = \frac{\gamma}{\beta_{LB}}$. It is now clear that, in order to obtain a rigorous, efficient and sharp residual-based estimator it is necessary to:

1. propose a rapid and reliable methodology to estimate β_{LB}. This is the goal of section 1.5.3;

2. determine a procedure to compute efficiently the residual. The next section will deal with this topic for both elliptic and parabolic problems.

[23] We observe that both (1.2.2b) and (1.3.5) are addressed by this general setting.

[24] This assumption, that is satisfied in both our cases, can be overcome by applying a limit procedure.

[25] The fact is a notable consequence of the Banach-Alaoglu-Bourbaki theorem [10].

1.5.2 Residual definition for elliptic and parabolic equations

In this sections suitable formulas for the rapid evaluation of the residual are obtained for both the elliptic case (1.2.1)-(1.2.2b) and the parabolic one (1.3.11)-(1.3.12).

In the elliptic case, mimicking (1.2.2b) into (1.5.5) we obtain:

$$\begin{aligned} r(v;\boldsymbol{\mu}) &= \sum_{q=1}^{Q_f} \Theta_q^f(\boldsymbol{\mu}) f^q(v) - \sum_{q=1}^{Q_a} \Theta_q^a(\boldsymbol{\mu}) a^q(u_{RB,N}(\boldsymbol{\mu}), v) \\ &= \sum_{q=1}^{Q_f} \Theta_q^f(\boldsymbol{\mu}) f^q(v) - \sum_{q=1}^{Q_a} \sum_{j=1}^{N} \Theta_q^a(\boldsymbol{\mu}) u_{N,j}(\boldsymbol{\mu}) a^q(\zeta_j, v) \end{aligned}$$

So, if we define Riesz representatives $\mathcal{C}_q$ $q = 1, \cdots, Q_f$ and $\mathcal{L}_n^{q'} \in X$, $n = 1, \cdots, N$ and $q' = 1, \cdots, Q_a$ such that:

$$(\mathcal{C}_q, v)_X = f^q(v) \ \forall\, v \in X \quad (\mathcal{L}_n^{q'}, v)_X = -a^{q'}(\zeta_n, v) \ \forall\, v \in X, \tag{1.5.9}$$

thanks to the linearity, we obtain

$$\hat{e}(\boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \Theta_q^f(\boldsymbol{\mu}) \mathcal{C}_q + \sum_{q=1}^{Q_a} \sum_{n=1}^{N} \Theta_q^a(\boldsymbol{\mu}) u_{N,n}(\boldsymbol{\mu}) \mathcal{L}_n^q. \tag{1.5.10}$$

Finally, the residual is

$$\begin{aligned} \|\hat{e}(\boldsymbol{\mu})\|_X^2 = &\sum_{q,q'=1}^{Q_f} \Theta_q^f(\boldsymbol{\mu}) \Theta_{q'}^f(\boldsymbol{\mu}) (\mathcal{C}_q, \mathcal{C}_{q'})_X + \sum_{q,q'=1}^{Q_a} \sum_{n,n'=1}^{N} \Theta_q^a(\boldsymbol{\mu}) \Theta_{q'}^a(\boldsymbol{\mu}) u_{N,n}(\boldsymbol{\mu}) u_{N,n'}(\boldsymbol{\mu}) (\mathcal{L}_n^q, \mathcal{L}_{n'}^{q'})_X \\ &+ 2 \sum_{q=1}^{Q_a} \sum_{q'=1}^{Q_f} \sum_{n=1}^{N} \Theta_{q'}^f(\boldsymbol{\mu}) \Theta_q^a(\boldsymbol{\mu}) u_{N,n}(\boldsymbol{\mu}) (\mathcal{L}_n^q, \mathcal{C}_{q'})_X. \end{aligned} \tag{1.5.11}$$

The previous formula separates the parameter-dependent terms from the parameter independent ones and thus allows the computation of $\|\hat{e}(\boldsymbol{\mu})\|_X$ in an offline-online framework. In the offline stage $(\mathcal{C}_q, \mathcal{C}_{q'})_X$,$(\mathcal{L}_n^q, \mathcal{L}_{n'}^{q'})_X$ and $(\mathcal{L}_n^q, \mathcal{C}_{q'})_X$ are computed. Then in the online stage the $NQ_a^2 + Q_f^2 + Q_a Q_f + Q_f^2$ parameter dependent terms are evaluated for the new value of the parameter and the residual is computed.

Concerning the parabolic problem (1.3.12) we are considering[26], it is possible to perform the same steps as in the steady case.

First of all, the residual is:

$$\begin{aligned} r(v, \boldsymbol{\mu}) = \sum_{k=0}^{\mathbb{K}} \Delta t \Bigg\{ &\sum_{q=1}^{Q_f} \Theta_q^f(\boldsymbol{\mu}) F^q(w) g(t^k) - \sum_{q=1}^{Q_a} \sum_{j=1}^{N} \Theta_q^a(\boldsymbol{\mu}) u_{N,j}^k(\boldsymbol{\mu}) a^q(\zeta_j, v) \\ &- \sum_{q=1}^{Q_m} \sum_{j=1}^{N} \Theta_q^m(\boldsymbol{\mu}) u_{N,j}^k(\boldsymbol{\mu}) m^q(\dot{\zeta}_j, v) \Bigg\}. \end{aligned} \tag{1.5.12}$$

Thus, if we define $\mathcal{C}_q \in V_h$ for $q = 1, \cdots, Q_f$ and $\mathcal{L}_n^{q'} \in V_h$, $n = 1, \cdots, N$ and $q' = 1, \cdots, Q_a$ such that

$$\begin{cases} (\mathcal{C}_q, v)_V = F^q(v) \ q = 1, \cdots, Q_f \\ (\mathcal{L}_n^q, v)_V = -a^q(\zeta_n, v) \quad q = 1, \cdots, Q_a \\ (\mathcal{L}_n^q, v)_V = -m^{q-Q_a}(\dot{\zeta}_n, v) \quad q = Q_a + 1, \cdots, Q_a + Q_m, \end{cases} \tag{1.5.13}$$

[26] The hypotheses on the functional and on the bilinear form make the formulation much simpler; see [120] section 3.3. for further comments.

we obtain that $\hat{e}(\boldsymbol{\mu}) \in \mathcal{Y}_\delta$ is such that:

$$\hat{e}(\boldsymbol{\mu})(t^k) = g(t^k) \sum_{q=1}^{Q_f} \Theta_q^f(\boldsymbol{\mu}) \mathcal{C}_q + \sum_{q=1}^{Q_a+Q_m} \sum_{j=1}^{N} \Theta_q^a(\boldsymbol{\mu}) u_{N,n}^k(\boldsymbol{\mu}) \mathcal{L}_n^q. \tag{1.5.14}$$

In conclusion,

$$\|\hat{e}(\boldsymbol{\mu})\|_{\mathcal{Y}}^2 = \sum_{k=0}^{\mathbb{K}} \Delta_k^2 \Delta t \tag{1.5.15a}$$

where:

$$\begin{aligned}\Delta_k^2 =& g(t^k)^2 \sum_{q,q'=1}^{Q_f} \Theta_q^f(\boldsymbol{\mu}) \Theta_{q'}^f(\boldsymbol{\mu}) (\mathcal{C}_q, \mathcal{C}_{q'})_X + \sum_{q,q'=1}^{Q_a+Q_m} \sum_{n,n'=1}^{N} \Theta_q^a(\boldsymbol{\mu}) \Theta_{q'}^a(\boldsymbol{\mu}) u_{N,n}^k(\boldsymbol{\mu}) u_{N,n'}^k(\boldsymbol{\mu}) (\mathcal{L}_n^q, \mathcal{L}_{n'}^q)_X \\ &+ 2g(t^k) \sum_{q=1}^{Q_a+Q_f} \sum_{q'=1}^{Q_f} \sum_{n=1}^{N} \Theta_{q'}^f(\boldsymbol{\mu}) \Theta_q^a(\boldsymbol{\mu}) u_{N,n}^k(\boldsymbol{\mu}) (\mathcal{L}_n^q, \mathcal{C}^{q'})_X.\end{aligned} \tag{1.5.15b}$$

Like in the elliptic case, a separation between parameter dependent and independent terms is reached.

1.5.3 Inf-sup lower bound for elliptic equations

In this section we deal with an efficient procedure to build an effective lower bound for the discrete coercivity constant (1.2.4):

$$\alpha(\boldsymbol{\mu}) = \inf_{w \in X} \frac{a(w, w, \boldsymbol{\mu})}{\|w\|_X^2} \quad \forall\, \boldsymbol{\mu} \in \mathcal{D} \tag{1.5.16}$$

where X is the above mentioned FE space ($\dim(X) = \mathcal{N}$). The discrete coercivity constant is a generalized minimum eigenvalue as stated in the following lemma.

Lemma 1.3. *Let us consider $\alpha(\boldsymbol{\mu})$ defined in (1.5.16). Let us suppose that $\|\cdot\|_X$ is induced by an inner product, say $(\cdot,\cdot)_X$, and that $\{\phi_j\}_{j=1}^{\mathcal{N}}$ is a basis for X. Then the following holds:*

$$\alpha(\boldsymbol{\mu}) = \min \left\{ \lambda \in \mathbb{R} : \frac{1}{2}(A(\boldsymbol{\mu}) + A^T(\boldsymbol{\mu}))\mathbf{w} = \lambda B \mathbf{w} \text{ for some } \mathbf{w} \in \mathbb{R}^{\mathcal{N}} \right\} \tag{1.5.17}$$

where $A(\boldsymbol{\mu}), B \in \mathbb{R}^{\mathcal{N} \times \mathcal{N}}$ are such that $A_{i,j}(\boldsymbol{\mu}) = a(\phi_j, \phi_i, \boldsymbol{\mu})$ and $B_{i,j} = (\phi_j, \phi_i)_X$.

The proof, based on Lagrangian multipliers, is here omitted[27].

There are many classical techniques for the estimation of minimum eigenvalues or minimum singular values such as Gershgorin's theorem and variants [57] or methods based on eigenfunction/eigenvalue approximation and subsequent residual evaluation [92]; however, these methods do not satisfy our requirements concerning reliability and efficiency.

For this reason in the context of Reduced Basis other methods have been proposed. The most successful one is surely the so-called Successive Constraint Method [107, 106] that we are going to present.

[27] We refer to [93] for a detailed discussion on this topic.

Successive Constraint Method

Let us consider the bilinear form in (1.2.1):

$$a(u,v,\boldsymbol{\mu}) = \sum_{q=1}^{Q_a} \Theta_q^a(\boldsymbol{\mu}) a^q(u,v)$$

In order to state the algorithm, we introduce the following objective function

$$\mathcal{J}^{obj}(\boldsymbol{\mu},\mathbf{y}) = \sum_{q=1}^{Q_a} \Theta_q^a(\boldsymbol{\mu}) y_q \tag{1.5.18a}$$

and the space.

$$\mathcal{Y} = \left\{ \mathbf{y} \in \mathbb{R}^{Q_a} : \exists\, w \in X \text{ such that, for all } q = 1,\cdots,Q_a,\ y_q = a^q(w,w) \right\}. \tag{1.5.18b}$$

It is absolutely straightforward to verify that:

$$\alpha(\boldsymbol{\mu}) = \inf_{\mathbf{y}\in\mathcal{Y}} \mathcal{J}^{obj}(\boldsymbol{\mu},\mathbf{y}). \tag{1.5.19}$$

We have recast the problem of finding the coercivity constant into an optimization problem based on a linear objective function. Therefore, if we are able to define a suitable convex polyhedron, say $\mathcal{Y}_{LB}$, such that $\mathcal{Y} \subset \mathcal{Y}_{LB}$ the lower bound can be obtained through the solution of a linear program.

In order to introduce such space, we define the following quantities. First of all, we consider the *constraint sample*

$$C_J = \{\boldsymbol{\mu}^1,\cdots,\boldsymbol{\mu}^J\} \tag{1.5.20}$$

and the *continuity constraint* box

$$\mathcal{B} = \prod_{q=1}^{Q_a} \left[\inf_{w\in X} \frac{a^q(w,w,\boldsymbol{\mu})}{\|w\|_X^2}, \sup_{w\in X} \frac{a^q(w,w,\boldsymbol{\mu})}{\|w\|_X^2} \right]; \tag{1.5.21}$$

finally, given $\boldsymbol{\mu} \in \mathcal{D}$, we refer to $C_J(\boldsymbol{\mu},M) \subset C_J$ as the set of the M closest points to $\boldsymbol{\mu}$ in C_J.

We have now the elements to introduce the lower bound set

$$\mathcal{Y}_{LB}(C_J(\boldsymbol{\mu},M)) = \left\{ \mathbf{y} \in \mathbb{R}^{Q_a} : \mathbf{y} \in \mathcal{B},\quad \sum_{q=1}^{Q_a} \Theta_q^a(\boldsymbol{\mu}') y_q \geq \alpha(\boldsymbol{\mu}'),\ \forall\, \boldsymbol{\mu}' \in C_J(\boldsymbol{\mu},M) \right\}, \tag{1.5.22a}$$

as well as the upper bound set

$$\mathcal{Y}_{UB}(C_J(\boldsymbol{\mu},M)) = \left\{ \mathbf{y}^\star(\boldsymbol{\mu}') \in \mathbb{R}^{Q_a} :\ \mathbf{y}^\star(\boldsymbol{\mu}') = \arg\inf_{\mathbf{y}\in\mathcal{Y}} \mathcal{J}^{obj}(\boldsymbol{\mu},\mathbf{y}) \right\}. \tag{1.5.22b}$$

Consequently we define the following quantities

$$\alpha_{LB}(\boldsymbol{\mu},C_J,M) = \inf_{\mathbf{y}\in\mathcal{Y}_{LB}(C_J(\boldsymbol{\mu},M))} \mathcal{J}^{obj}(\boldsymbol{\mu},\mathbf{y}) \tag{1.5.23a}$$

and

$$\alpha_{UB}(\boldsymbol{\mu},C_J,M) = \inf_{\mathbf{y}\in\mathcal{Y}_{UB}(C_J(\boldsymbol{\mu},M))} \mathcal{J}^{obj}(\boldsymbol{\mu},\mathbf{y}). \tag{1.5.23b}$$

The following lemma justifies the approach[28]

[28] See [106] for the proof.

Lemma 1.4. *Given $C_J \subset \mathcal{D}$ and $M \in \mathbb{N}$ we have that:*

$$\mathcal{Y}_{UB}(C_J(\boldsymbol{\mu}, M)) \subset \mathcal{Y}(C_J(\boldsymbol{\mu}, M)) \subset \mathcal{Y}_{LB}(C_J(\boldsymbol{\mu}, M)).$$

So it readily follows that:

$$\alpha_{LB}(\boldsymbol{\mu}, C_J, M) \leq \alpha(\boldsymbol{\mu}) \leq \alpha_{UB}(\boldsymbol{\mu}, C_J, M). \tag{1.5.24}$$

The effectivity of the lower bound depends on the coercivity constraint sample C_J. In [106] a Greedy algorithm -similar to the sampling Greedy procedure 3- has been proposed. In Algorithm 5 we gather the entire procedure.

Let $\Xi \subset \mathcal{D}$ be a train sample equivalent to the one defined in section 1.4, $\epsilon_{SCM} \in (0,1)$ and $M \in \mathbb{N}$ be a given tolerance and a given positive integer, respectively; finally let $\boldsymbol{\mu}^1_{SCM} \in \Xi$ be given in order to initialize the algorithm.

Algorithm 5 Greedy selection of C_J: $[C_J] = \texttt{SCM-Greedy}(\boldsymbol{\mu}^1_{SCM}, \Xi, \epsilon_{SCM}, M)$

Set $C_1 = \{\boldsymbol{\mu}\}$ and $J = 1$.

$\eta(\boldsymbol{\mu}, C_J, M) := \dfrac{\alpha_{UB}(\boldsymbol{\mu}, C_J, M) - \alpha_{LB}(\boldsymbol{\mu}, C_J, M)}{\alpha_{UB}(\boldsymbol{\mu}, C_J, M)}$

Compute the continuity constraint box $\mathcal{B}$ in (1.5.21).

while $\max_{\boldsymbol{\mu} \in \Xi} \eta(\boldsymbol{\mu}, C_J, M) > \epsilon_{SCM}$ **do**

$\boldsymbol{\mu}^{J+1} = \arg\max_{\mu \in \Xi} \eta(\boldsymbol{\mu}, C_J, M)$

Compute $\alpha(\boldsymbol{\mu}^{J+1})$ and $\mathbf{y}^\star(\boldsymbol{\mu}^{J+1})$

$C_{J+1} = C_J \cup \{\boldsymbol{\mu}^{J+1}\}$

$J = J + 1$

end while

Essentially, at each iteration of the Greedy procedure we add to our coercivity constraint the point for which the difference between the current lower bound and upper bound estimate is largest. Due to the fact that $\alpha_{UB}(\boldsymbol{\mu}, C_J, M) = \alpha_{LB}(\boldsymbol{\mu}, C_J, M)\ \forall\, \boldsymbol{\mu} \in C_J$, it follows from the continuity of the coercivity constant, that our error tolerance will be honoured for J sufficiently large.

The choice of the stopping criterion permits us to bound the ratio:

$$\begin{aligned}
\frac{\alpha(\boldsymbol{\mu})}{\alpha_{LB}(\boldsymbol{\mu}, C_J, M)} &= \frac{\alpha(\boldsymbol{\mu})}{\alpha_{UB}(\boldsymbol{\mu}, C_J, M) - (\alpha_{UB}(\boldsymbol{\mu}, C_J, M) - \alpha_{LB}(\boldsymbol{\mu}, C_J, M))} \\
&= \frac{\alpha(\boldsymbol{\mu})}{\alpha_{UB}(\boldsymbol{\mu}, C_J, M)} \frac{1}{1 - \eta(\boldsymbol{\mu}, C_J, M)} \\
&\leq \frac{1}{1 - \epsilon_{SCM}} \quad \forall\, \boldsymbol{\mu} \in \Xi
\end{aligned}$$

Therefore, the residual estimator[29] defined in (1.5.7) satisfies the sharp condition (1.5.2) with $C = \frac{1}{1-\epsilon_{SCM}} \frac{\gamma(\boldsymbol{\mu})}{\alpha(\boldsymbol{\mu})}$.

We summarize the online and offline computational costs:

- in the offline stage we have to solve $2Q_a$ eigenproblems over X in order to build $\mathcal{B}$, J eigenproblems over X to form $\{\alpha(\boldsymbol{\mu}') : \boldsymbol{\mu}' \in C_J\}$ and finally to solve $n_{train} J$ linear programs of size at maximum $Q_a + M$ for the computation of $\eta(\boldsymbol{\mu}, C_J, M)$;

[29] $\Delta_{RB}(\boldsymbol{\mu}) = \frac{1}{\alpha_{LB}(\boldsymbol{\mu})} \|\hat{e}(\boldsymbol{\mu})\|_X$, the use of α instead of β is explained by the coercivity of the problem.

- 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 $\alpha_{LB}(\boldsymbol{\mu}, C_J, M)$.

The eigenproblems associated with the calculation of the $\alpha(\boldsymbol{\mu}')$, with $\boldsymbol{\mu}' \in C_J$, can be solved efficiently through the Lanczos method [117]. Thanks to our particular choice of the inner product- see (1.2.3)- it can be shown that the lowest eigenvalue is well separated from the second one. This ensures rapid convergence of the above mentioned procedure.

Before concluding, we give some references. The approach here presented for coercive problems can be extended to non coercive linear equations; for instance in [79] the methodology has been applied to the Stokes equation. Recently a two level affine decomposition variant has been proposed for general parametrized elliptic equations in [70].

1.5.4 Inf-sup lower bound for parabolic equations

Following the strategy proposed in [120], we deduce a lower bound for the inf-sup stability constant associated with (1.3.12).

With respect to the inner products defined in section 1.3.2, we introduce the following quantities:

$$\begin{cases} \rho_h = \sup\limits_{\phi \in V_h} \dfrac{\|\phi\|_H}{\|w\|_V} \\ \beta^{\star}_{a,h}(\boldsymbol{\mu}) = \inf\limits_{\phi \in V_h} \sup\limits_{\psi \in V_h} \dfrac{a(\psi, \phi, \boldsymbol{\mu})}{\|\psi\|_V \|\phi\|_V} \\ (\lambda_h(\boldsymbol{\mu}), \alpha_{a,h}(\boldsymbol{\mu})) \text{ such that } a(\psi, \psi, \boldsymbol{\mu}) + \lambda_h(\boldsymbol{\mu})\|\psi\|_H^2 \geq \alpha_{a,h}(\boldsymbol{\mu})\|\psi\|_V \quad \forall\, \psi \in V_h \\ \gamma_{a,h}(\boldsymbol{\mu}) = \sup\limits_{\phi \in V_h} \sup\limits_{\psi \in V_h} \dfrac{a(\psi, \phi, \boldsymbol{\mu})}{\|\psi\|_V \|\phi\|_V} \end{cases} \tag{1.5.25}$$

where we suppose that the constants are stable for $h \to 0$. We have now the elements to state the main result from [120].

Lemma 1.5. *Let us consider the equation (1.3.12) for a given $\boldsymbol{\mu} \in \mathcal{D}$. If $\alpha_{a,h}(\boldsymbol{\mu}) - \lambda_{a,h}(\boldsymbol{\mu})\rho_h^2 > 0$ the following estimate holds:*

$$\beta_{b,\delta}(\boldsymbol{\mu}) \geq \beta_{h,LB}(\boldsymbol{\mu}) := \frac{\min\{1, \alpha_{h,a}(\boldsymbol{\mu}) - \lambda_h(\boldsymbol{\mu})\rho_h^2\} \min\{1, \gamma_{a,h}(\boldsymbol{\mu})^{-2}\}}{\sqrt{2}\max\{1, (\beta_{a,h}(\boldsymbol{\mu})^{\star})^{-1}\}} \tag{1.5.26}$$

Otherwise:

$$\beta_{b,\delta}(\boldsymbol{\mu}) \geq \beta_{h,LB}(\boldsymbol{\mu}) := \frac{e^{-2\lambda T}}{\max\{\sqrt{1 + 2\lambda_h^2 \rho_h^4}, \sqrt{2}\}} \frac{\min\{1, \alpha_{a,h}(\boldsymbol{\mu}) \min\{1, \gamma_{a,h}^{-2}(\boldsymbol{\mu})\}}{\sqrt{2}\max\{1, (\beta^{\star}_{a,h})^{-1}(\boldsymbol{\mu})\}} \tag{1.5.27}$$

where $\beta_{b,\delta}$ is the inf-sup constant defined in (1.3.17).

Lemma 1.5 permits us to perform the estimation of the inf-sup constant $\beta_{b,\delta}(\boldsymbol{\mu})$ through the same algorithm used in the steady case and explained in the previous paragraph.

1.5.5 A posteriori error estimation for the output

Given problem (1.2.2):

$$s(\boldsymbol{\mu}) = l(u(\boldsymbol{\mu})) \text{ where } u \in X_N \text{ solves } a(u(\boldsymbol{\mu}), v, \boldsymbol{\mu}) = f(v, \boldsymbol{\mu})\ \forall\, v \in X$$

in section 1.2.2 we introduced the following output estimation (*Primal Approximation*):

$s_{RB,N}(\boldsymbol{\mu}) = l(u_{RB,N}(\boldsymbol{\mu}))$ where $u_{RB,N} \in X_N$ solves $a(u_{RB,N}(\boldsymbol{\mu}), v, \boldsymbol{\mu}) = f(v, \boldsymbol{\mu})\ \forall\, v \in X_N$.

Thanks to the a posteriori estimator $\Delta_{RB}(\boldsymbol{\mu})$ detailed in section 1.5.1 we have the following bound for the output error:

$$|s_{RB,N}(\boldsymbol{\mu}) - s(\boldsymbol{\mu})| \leq \|l\|_{X'} \Delta_{RB,N}(\boldsymbol{\mu}). \tag{1.5.28}$$

A significant improvement to this estimate can be obtained through a *Primal-Dual approximation*, [106]. Let us define the dual problem:

$$\text{Find } \psi(\boldsymbol{\mu}) \in X \text{ such that} \quad a(v, \psi(\boldsymbol{\mu}), \boldsymbol{\mu}) = -l(v) \quad \forall\, v \in X. \tag{1.5.29}$$

Exactly as in the previous case, we introduce the reduced spaces, X_N^{pr} and X_M^{du}. Consequently we define the projected problems:

$$\begin{cases} \text{Find } u_{RB,N}(\boldsymbol{\mu}) \in X_N^{pr} \text{ such that } a(u_{RB,N}(\boldsymbol{\mu}), v, \boldsymbol{\mu}) = f(v, \boldsymbol{\mu})\ \forall\, v \in X_N^{pr} \\ \text{Find } \psi_{RB,M}(\boldsymbol{\mu}) \in X_M^{du} \text{ such that } a(v, \psi_{RB,N}(\boldsymbol{\mu}), \boldsymbol{\mu}) = -l(v)\ \forall\, v \in X_M^{du} \end{cases} \tag{1.5.30}$$

and the following output approximation:

$$s_{RB,N,M}(\boldsymbol{\mu}) = l(u_{RB,N}(\boldsymbol{\mu})) - r^{pr}(\psi_{RB,M}; \boldsymbol{\mu}), \tag{1.5.31}$$

where $r^{pr}(\cdot; \boldsymbol{\mu})$ is the residual of the primal problem.

For the latter approximation it is possible to prove the following (see proposition 4 in [106]).

Lemma 1.6. *Given $\boldsymbol{\mu} \in \mathcal{D}$, we consider $s(\boldsymbol{\mu})$ in (1.2.2) and $s_{RB,N,M}(\boldsymbol{\mu})$ as in (1.5.31); we have*

$$|s(\boldsymbol{\mu}) - s_{RB,N,M}(\boldsymbol{\mu})| \leq \gamma_a(\boldsymbol{\mu}) \|u(\boldsymbol{\mu}) - u_{RB,N}(\boldsymbol{\mu})\|_X, \|\psi(\boldsymbol{\mu}) - \psi_{RB,M}(\boldsymbol{\mu})\|_X \tag{1.5.32}$$

with $\gamma_a(\boldsymbol{\mu})$ as in (1.2.4).

Thanks to Lemma 1.6, after introducing the estimators $\Delta_{RB,N}^{pr}(\boldsymbol{\mu})$ and $\Delta_{RB,M}^{du}(\boldsymbol{\mu})$ we obtain the following a posteriori estimate:

$$|s(\boldsymbol{\mu}) - s_{RB,N,M}(\boldsymbol{\mu})| \leq \gamma(\boldsymbol{\mu}) \Delta_{RB,N}^{pr}(\boldsymbol{\mu}) \Delta_{RB,M}^{du}(\boldsymbol{\mu}) =: \Delta_{RB,N,M}^{pr,du}(\boldsymbol{\mu}) \tag{1.5.33}$$

In order to motivate the approach, let us assume $N = M$; by (1.5.33) we have that the online computational effort is duplicated, but now the output error is proportional to the square of the field error.

For further details about the Primal-Dual formulation and for the extension of this approach to the parabolic context we refer to [106, 120]

1.6 The whole method: offline-online decomposition

At the end of the presentation of the main tools behind the Reduced Basis method for elliptic and parabolic equations, we summarize here the whole algorithm. We limit ourselves to the elliptic case, the parabolic one being very similar. Furthermore, we refer to [106] for the Primal-Dual formulation.

Algorithm 6 Offline-Online decomposition

Offline stage

1: SCM offline stage, (Algorithm 5).
2: Greedy Algorithm, (Algorithm 3); at each step we update the matrices and vectors A^q, $\mathbf{F}^q$ and $\mathbf{l}^q$ via (1.2.6) and the components of the residual via (1.5.10).

At the end of this stage the following quantities are saved: C_J and $\alpha(\boldsymbol{\mu}')$, for all $\boldsymbol{\mu}' \in C_J$, via SCM; A^q, $\mathbf{F}^q$ and $\mathbf{l}^q$ thanks to (1.2.6) and finally the components of the residual by exploiting (1.5.10).

Online stage

1: Given $\boldsymbol{\mu} \in \mathcal{D}$, we assemble the matrix $A(\boldsymbol{\mu})$ and the vector $\mathbf{F}(\boldsymbol{\mu})$ in (1.2.6b) and we solve the reduced system.
2: We compute the output through (1.2.6a).
3: Using the precomputed parameter-independent quantities we define the residual $\|\hat{e}(\boldsymbol{\mu})\|_X$ as in (1.5.10).
4: We solve the linear program to compute $\alpha_{LB}(\boldsymbol{\mu}, C_J, M)$ via (1.5.22a).
5: We compute the residual based a posteriori estimator and eventually the output estimator through (1.5.28).

1.7 Two numerical examples

In this section we apply the RB approach to two different problems. The first one is a steady diffusion-reaction problem; the second one is a steady advection-diffusion problem.

We will see how for the first one the convergence is significantly faster than for the second one even though the regularity in space is the same for both the test cases.

1.7.1 A diffusion-reaction problem

Let us consider the following input-output relation:

$$\text{Given } \boldsymbol{\mu} \in \mathcal{D} = [1,10] \times [0.1,100], \quad \text{compute } s(\boldsymbol{\mu}) = \int_{\Gamma_3} T(\boldsymbol{\mu})\, dx_2, \tag{1.7.1a}$$

where $T(\boldsymbol{\mu})$ is the solution to the following problem:

$$\begin{cases} -\dfrac{1}{\mu_1}\Delta T(\boldsymbol{\mu}) + T(\boldsymbol{\mu}) = 0 & \text{in } \Omega(\mu_1) \\ T(\boldsymbol{\mu}) = \begin{cases} 0 & \text{on } \Gamma_i(\mu_1)\ i = 1,5,6 \\ 1 & \text{on } \Gamma_i(\mu_1)\ i = 2,4 \end{cases} & \\ \dfrac{\partial T(\boldsymbol{\mu})}{\partial n} = \dfrac{\partial T(\boldsymbol{\mu})}{\partial x_1} = 0 & \text{in } \Gamma_3 \end{cases} \tag{1.7.1b}$$

In Figure 1.1, the domain is plotted:

Due to the discontinuity in the boundary data, the mathematical analysis of the well-posedness of problem (1.7.1b) is quite involved; through the *transposition method* [73] it is possible to prove the wellposedness in $L^2(\Omega(\mu_1))$. Nonetheless, from the numerical viewpoint we treat the problem as a standard second-order elliptic equation.

As we will see in the second chapter, in order to apply the methodology it is necessary to refer the problem to a parameter independent configuration, say $\Omega = \Omega(\boldsymbol{\mu}_{ref})$.

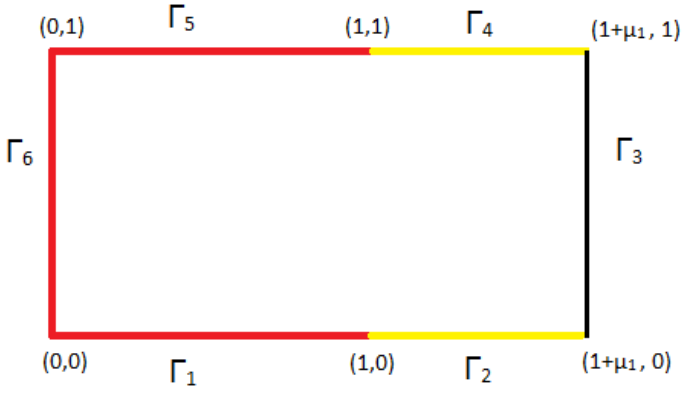


Figure 1.1: problem domain: red boundaries are associated to the homogeneous Dirichlet condition whereas yellow boundaries are associated to non-homogeneous Dirichlet condition.

For this reason, we define $\boldsymbol{\mu}_{ref} = [1,\ 1]$ and consequently the map $\mathcal{T} : \mathcal{D} \times \mathbb{R}^2 \to \mathbb{R}^2$ such that:

$$\mathcal{T}(\boldsymbol{\mu}, \mathbf{x}) = \begin{bmatrix} \dfrac{1+\mu_1}{2} x_1 \\ x_2 \end{bmatrix} \tag{1.7.2}$$

Then we can write the problem in the reference configuration in such a way:

$$\text{Given } \boldsymbol{\mu} \in \mathcal{D} = [1, 10] \times [0.1, 100], \quad \text{compute } s(\boldsymbol{\mu}) = \int_{\Gamma_3} \tilde{T}(\boldsymbol{\mu})\, dx_2 \tag{1.7.3a}$$

where $\tilde{T}(\boldsymbol{\mu}) : \Omega \to \Omega(\boldsymbol{\mu})$, $\tilde{T}(\boldsymbol{\mu}) = T(\boldsymbol{\mu}) \circ \mathcal{T}(\boldsymbol{\mu})$ is the solution to the following problem:

$$\begin{cases} -\text{div}\left(\begin{bmatrix} \dfrac{2}{\mu_2(1+\mu_1)} & 0 \\ 0 & \dfrac{1+\mu_1}{2\mu_2} \end{bmatrix} \nabla \tilde{T}(\boldsymbol{\mu}) \right) + \frac{1+\mu_1}{2} \tilde{T}(\boldsymbol{\mu}) = 0 \quad \text{in } \Omega \\ \tilde{T}(\boldsymbol{\mu}) = \begin{cases} 0 & \text{on } \Gamma_i \ i = 1, 5, 6 \\ 1 & \text{on } \Gamma_i \ i = 2, 4 \end{cases} \\ \dfrac{\partial \tilde{T}(\boldsymbol{\mu})}{\partial n} = \dfrac{\partial u}{\partial x_1} = 0 \qquad \text{in } \Gamma_3 \end{cases} \tag{1.7.3b}$$

Here we gather some plots and data. First we consider the convergence of the Greedy algorithm in Figure 1.2.

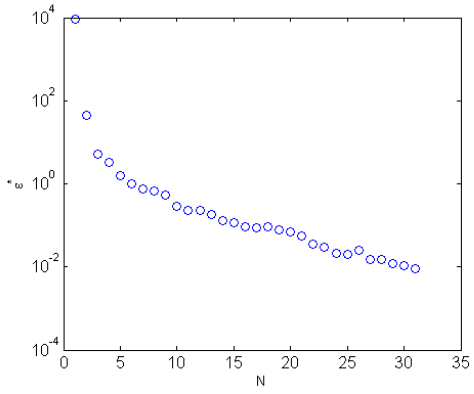
By applying the exponential fitting to the data[30] we obtain:

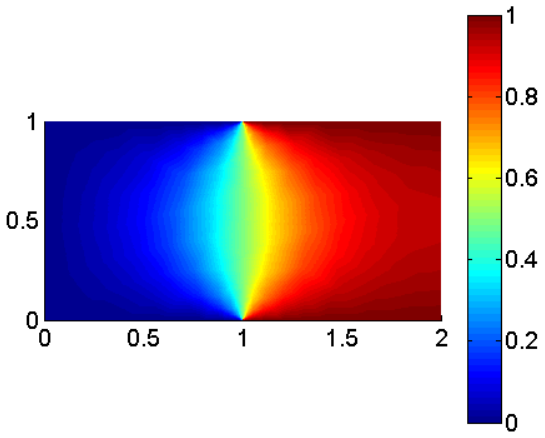
$$\|\Delta_N(\cdot)\|_{L^\infty(\Xi)} \simeq 33.56 e^{-0.9937 N}$$

We observe that, despite the lack of regularity in space, the convergence is extremely fast. In Figure 1.3 the solution for two extreme values of the parameter sample are plotted. We observe that the methodology is able to deal with large variations in the shape of the solutions associated with different values of the parameter through a very small number of bases.

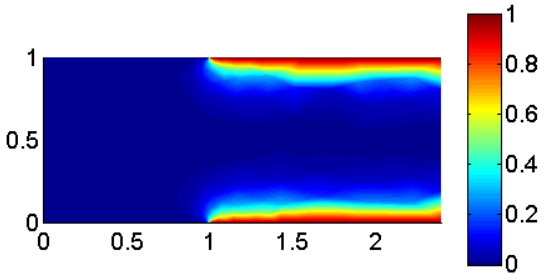
[30] The fitting is made by the `Matlab` tool `fit` [84]. See the documentation for further details about the algorithm used.

Figure 1.2: convergence of the Greedy algorithm for the primal problem





(a) $\mu_1 = 1.4$ $\mu_2 = 0.1$



(b) $\mu_1 = 1.4$ $\mu_2 = 100$

Figure 1.3: reduced solutions for two values of the parameter

In Table 1.1 we gather the a posteriori error estimation and the real error in the output estimation for a given parameter[31]. We observe that the effectivity of the a posteriori error

[31] The value obtained for $X^{pr}_{N^{pr}}$, $X^{du}_{N^{du}}$ $N_{pr} = 15$ and $N^{du} = 10$ is assumed to be exact. For these spaces the a posteriori estimator is lower than 10^{-7}.

$N \setminus M$	1	5	10
1	0.3892 (0.0044)	0.0142 ($6.3 \cdot 10^{-4}$)	$4.43 \cdot 10^{-4}$ ($1.8 \cdot 10^{-5}$)
5	0.0134 ($4.3 \cdot 10^{-5}$)	$4 \cdot 10^{-4}$ ($9.6 \cdot 10^{-5}$)	$1.53 \cdot 10^{-5}$ ($9.2 \cdot 10^{-7}$)
10	0.0021 ($3.4 \cdot 10^{-5}$)	$7.67 \cdot 10^{-5}$ ($7.4 \cdot 10^{-7}$)	$2.4 \cdot 10^{-6}$ ($1.87 \cdot 10^{-8}$)

Table 1.1: estimated and real output error for different primal and dual reduced spaces. $\boldsymbol{\mu} = [2.1,\ 1]$

bound defined as:

$$\eta(\boldsymbol{\mu}) = \frac{\Delta^{pr,du}_{RB,N,M}(\boldsymbol{\mu})}{|s(\boldsymbol{\mu}) - s_{RB,N,M}(\boldsymbol{\mu})|} \tag{1.7.4}$$

is $\mathcal{O}$ (100).

1.7.2 Graetz problem

Graetz problem[32] is a classical problem in literature concerning forced heat convection combined with heat conduction in a duct with walls at different temperature. The flow enters a given tube at a temperature T_0 and encounters a wall temperature T_1 which can be larger or smaller than T_0. A simple version of the problem was first analysed by Graetz [45] in 1883.

The domain is still the one plotted in Figure 1.1; on the other hand, the problem we are going to solve is now:

$$\text{Given } \boldsymbol{\mu} \in \mathcal{D} = [1,10] \times [0.1,100], \quad \text{compute } s(\boldsymbol{\mu}) = \int_{\Gamma_3} T(\boldsymbol{\mu})\, dx_2, \tag{1.7.5a}$$

where $T(\boldsymbol{\mu})$ is the solution of the following problem:

$$\begin{cases} -\dfrac{1}{\mu_1}\Delta T + 10x_2(1-x_2)\dfrac{\partial}{\partial x_1}T = 0 \quad \text{in } \Omega(\mu_1) \\ T = \begin{cases} 0 \quad \text{on } \Gamma_i(\mu_1)\ i = 1,5,6 \\ 1 \quad \text{on } \Gamma_i(\mu_1)\ i = 2,4 \end{cases} \\ \dfrac{\partial T}{\partial n} = \dfrac{\partial T}{\partial x_1} = 0 \qquad \text{in } \Gamma_3. \end{cases} \tag{1.7.5b}$$

If we refer the problem to a parameter independent configuration through the map $\mathcal{T}$ defined in (1.7.2) we obtain:

$$\text{Given } \boldsymbol{\mu} \in \mathcal{D} = [1,10] \times [0.1,100], \quad \text{compute } s(\boldsymbol{\mu}) = \int_{\Gamma_3} \tilde{T}(\boldsymbol{\mu})\, dx_2, \tag{1.7.6a}$$

where $\tilde{T}(\boldsymbol{\mu}) : \Omega \to \mathbb{R}$ is the solution to the following problem:

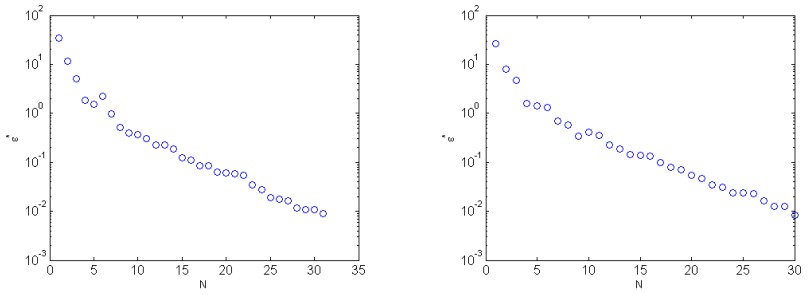
$$\begin{cases} -\text{div}\left(\begin{bmatrix} \frac{2}{\mu_2(1+\mu_1)} & 0 \\ 0 & \frac{1+\mu_1}{2\mu_2} \end{bmatrix} \nabla \tilde{T}(\boldsymbol{\mu})\right) + 10x_2(1-x_2)\frac{\partial}{\partial x_1}\tilde{T}(\boldsymbol{\mu}) = 0 \qquad \text{in } \Omega \\ T = \begin{cases} 0 \quad \text{on } \Gamma_i\ i = 1,5,6 \\ 1 \quad \text{on } \Gamma_i\ i = 2,4 \end{cases} \\ \dfrac{\partial T}{\partial n} = \dfrac{\partial T}{\partial x_1} = 0 \qquad \text{in } \Gamma_3. \end{cases} \tag{1.7.6b}$$

[32] The example is taken from the Worked problems at `augustine.mit.edu`. The simulations have been performed through the software `rbMIT` [102].

As for the previous problem, it is possible to prove the well-posedness in $L^2(\Omega)$.

We have all the elements to apply the Reduced basis method. As before we first plot the convergence of the Greedy algorithm. We note that convergence rate does not depend on the underlined mesh, but it is significantly lower than in the previous case. By applying the exponential fitting to the data as before, we obtain:

$$\|\Delta_N(\cdot)\|_{L^\infty(\Xi)} \simeq 21.82e^{-0.4891N}.$$



(a) $\mathcal{N} = 480$ (b) $\mathcal{N} = 891$

Figure 1.4: convergence rates for the Greedy algorithm for two different underlined meshes

As regards the adaptive choice of the parameters, we observe that the sample set is far from a tensor-product form: as we may expect, the highest clustering is near to $\mu_2 = 0.1$ corresponding to regions in $\mathcal{D}$ where the parametric sensitivity is largest.

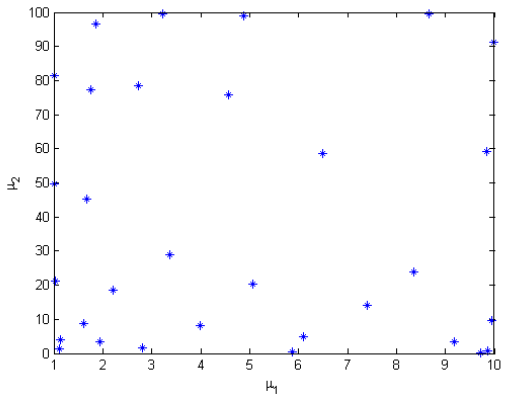


Figure 1.5: parameters chosen by the Greedy Algorithm

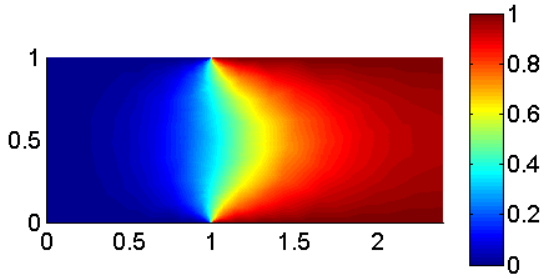
As before, we also gather the results for a given value of the parameters for different reduced spaces, see Table 1.2.

We observe that the effectivity defined in (1.7.4) is in this case around $\mathcal{O}$ (200).

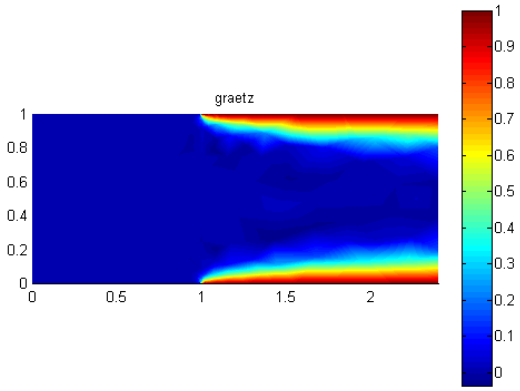
Finally, we plot in Figure 1.6 the solution for two different values of the parameter. We observe that the boundary layer is well detected through a low number of bases: this is possible because of the fact that the position of the layer does not depend on the parameters.

$N \setminus M$	1	10	20
1	5.42 (0.067)	0.216 (0.0012)	0.011 $(1.1 \cdot 10^{-4})$
5	0.33 (0.004)	0.0133 $(2.4 \cdot 10^{-4})$	0.0018 $(7.9 \cdot 10^{-6})$
10	0.046 $(2.4 \cdot 10^{-4})$	0.0018 $(3.69 \cdot 10^{-5})$	$2.51 \cdot 10^{-4}$ $(1.02 \cdot 10^{-5})$
20	0.0011 $(7.93 \cdot 10^{-5})$	$2.23 \cdot 10^{-4}$ $(9.0 \cdot 10^{-7})$	$3.03 \cdot 10^{-5}$ $(1.76 \cdot 10^{-7})$

Table 1.2: estimated and real output error for different primal and dual reduced spaces. $\boldsymbol{\mu} = [2.1,\ 50]$



(a) $\mu_1 = 1.4\ \mu_2 = 1$



(b) $\mu_1 = 1.4\ \mu_2 = 100$

Figure 1.6: reduced solutions for two values of the parameter

1.8 The treatment of non-affine problems: the EIM

In order to manage an efficient offline-online computational decomposition as the one described in the previous sections, it is necessary to deal with partial differential equations with affine parameter dependence. Therefore, in order to face problems with nonaffine parameter dependence, we have to resort to a method that replaces non-affine coefficient functions with a collateral reduced basis expansion which allows us to perform an effectively offline-online computational decomposition.

Example 1.3. *Let us consider the following non affine bilinear form and the variational problem associated with it:*

$$\textit{Find} u \in X:\ a(u,v;\boldsymbol{\mu}) = f(v)\ \forall\, v \in X, \quad \textit{with}\ a(u,v;\boldsymbol{\mu}) = \int_\Omega g(\boldsymbol{\mu},\mathbf{x}) \nabla u(\mathbf{x}) \cdot \nabla v(x)\, d\mathbf{x} \tag{1.8.1}$$

where for the sake of simplicity g is assumed to be arbitrary regular, say $g \in \mathcal{C}^\infty(\mathcal{D}, L^\infty(\Omega))$. In order to deal with problem (1.8.1) in the RB offline-online framework proposed above, we need to replace $g(\boldsymbol{\mu}, \mathbf{x})$ with a parametrically affine surrogate $g_M(\boldsymbol{\mu}, \mathbf{x}) = \sum_k \Theta_k^g(\boldsymbol{\mu}) q(\mathbf{x})$
.

$$Find\, u \in X:\ a_M(u, v; \boldsymbol{\mu}) = \sum_{k=1}^{M} \Theta_k^g(\boldsymbol{\mu}) \int_\Omega q(\mathbf{x}) \nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x})\, d\mathbf{x} = f(v) \quad \forall v \in X \qquad (1.8.2)$$

The Empirical Interpolation Method (EIM)[7, 32] provides a methodology to efficiently compute the parametrically affine surrogate g_M.

1.8.1 Description of the algorithm

As the RB framework requires, the EIM consists in an offline and in an online stage. Like in the sampling strategy, we consider an approximation space of the form:

$$W_M^g = \text{span}\{g(\boldsymbol{\mu}_M^g, \cdot):\ \boldsymbol{\mu}_M^g \in \mathcal{D}\}$$

Despite of the sampling strategy, in this case the approximate solution is obtained through an interpolation strategy instead of a Galerkin projection.

The offline stage consists of two steps:

- in the first we build an approximation space, say W_M^g, by a greedy strategy: at each step $\boldsymbol{\mu}_M^g$ is chosen in order to maximize the error in the L^∞-norm;
- in the second we find an interpolatory basis for the space W_M^g. Even in this case, a Greedy approach is used in order to find the set of suitably chosen interpolation points.

Therefore, the offline algorithm takes a function $g : \mathcal{D} \times \Omega \to \mathbb{R}$ and returns a set of functions $\{q_j\}$ that represents a basis for the approximation space W_M^g, an interpolation matrix $B \in \mathbb{R}^{M \times M}$ and the so called *magic points* $\mathbf{t}_M \in \Omega$ such that $B_{i,j} = q_j(\mathbf{t}_i)$ and $q_j(\mathbf{t}_i) = 0$ if $j > i$ and $B_{i,j} = q_j(\mathbf{t}_i) = 1$ if $j = i$. On the other hand, in the online stage we evaluate g in the magic points and we compute the coefficients of the basis by solving a linear system.

In order to state the entire algorithm, we introduce a discretization of the parameter domain, say $\Xi^g \subset \mathcal{D}$.

Algorithm 7 gathers the Pseudo Code of the procedure as presented in [7].

1.8.2 Error analysis and a posteriori error estimator

After the presentation of the algorithm, we briefly outline some theoretical aspects about the convergence properties and the a posteriori error analysis associated with EIM.

Let us suppose that $\dim(\text{span}\{g(\boldsymbol{\mu}, \cdot):\ \boldsymbol{\mu} \in \mathcal{D}\}) \geq M_{max}$. Thanks to the previous hypothesis, it is possible to prove - see [7] - that for all $M \leq M_{max}$ $\dim W_M^g = M$. Then the construction of the interpolation points is well defined and the functions $\{q_1, \cdots, q_M\}$ form a basis for W_M^g.

So the algorithm presented is well defined.

In order to deal with the error analysis we introduce some notation. First, we indicate with d the dimension of the space: $\Omega \subset \mathbb{R}^d$. Then we define the Lebesgue costant, [98], $\Lambda_M = \sup_{\mathbf{x} \in \Omega} \sum_{m=1}^{M} \left| V_m^M(\mathbf{x}) \right|$ where $\{V_m^M\}$ are the Lagrange basis polynomials associated with $\{\mathbf{t}_1, \cdots, t_M\}$.

We have now the elements to state the following result from [33].

Algorithm 7 Empirical Interpolation Method:

Offline stage:
$[\{q_k(\cdot) : 1 \le k \le M\}, \{\mathbf{t}_j : 1 \le j \le M_{Max}\}, \{B_{i,j} : 1 \le i, j \le M\}]$ = EIM $\{g(\boldsymbol{\mu}, \cdot), M_{Max}\}$

Given $S_1^g = \{\boldsymbol{\mu}_1^g\}$, $\xi_1(\cdot) \equiv g(\boldsymbol{\mu}_1^g, \cdot)$, $W_1^g = \text{span}\{\xi_1\}$:
Part I: Construction of the Approximation Space
for $M = 2 : M_{max}$ **do**
  $\boldsymbol{\mu}_M^g = \arg\max_{\mu \in \Xi^g} \inf_{z \in W_{M-1}^g} \|g(\cdot, \mu) - z\|_{L^\infty(\Omega)}$
  $S_M^g = S_{M-1}^g \cup \{\boldsymbol{\mu}_M^g\}$, $\xi_M(\cdot) \equiv g(\boldsymbol{\mu}_M^g, \cdot)$, $W_M^g = W_{M-1}^g \cup \text{span}\{\xi_M\}$
end for
Part II: Construction of the basis
$\mathbf{t}_1 = \arg \operatorname{ess\,sup}_{\mathbf{x} \in \Omega} \|\xi_1(\mathbf{x})\|$, $q_1 = \frac{\xi_1}{\xi_1(\mathbf{t}_1)}$, $B_{11} = 1$
for $M = 2 : M_{max}$ **do**
  Find $\boldsymbol{\sigma} \in \mathbb{R}^{M-1} : \sum_{j=1}^{M-1} \sigma_j q_j(\mathbf{t}_i) = \xi_M(\mathbf{t}_i)$ for $1 \le i \le M-1$
  $r_M(\mathbf{x}) = \xi_M(\mathbf{x}) - \sum_{j=1}^{M-1} \sigma_j q_j(\mathbf{x})$
  $\mathbf{t}_M = \arg \operatorname{ess\,sup}_{\mathbf{x} \in \Omega} \|r_M(\mathbf{x})\|$, $q_M(\mathbf{x}) = \frac{r_M(\mathbf{x})}{r_M(\mathbf{t}_M)}$, $B_{i,j}^M = q_j(\mathbf{t}_i)$
end for

Online stage: $[g_M(\boldsymbol{\mu})]$ = EIM $\{\boldsymbol{\mu}, M\}$

Compute $\boldsymbol{\beta}(\boldsymbol{\mu}) \in \mathbb{R}^M$ such that: $B_M \boldsymbol{\beta}(\boldsymbol{\mu}) = \mathbf{g}(\boldsymbol{\mu})$ where $g_i(\boldsymbol{\mu}) = g(\boldsymbol{\mu}, \mathbf{t}_i)$ $i = 1, \cdot, M$.
Define

$$g_M(\boldsymbol{\mu}, \cdot) = \sum_{m=1}^{M} \beta_m(\boldsymbol{\mu}) q_m(\cdot)$$

Lemma 1.7. *Given a multi-index $\beta \in \mathbb{N}^d$, let us suppose that $g(\boldsymbol{\mu}, \cdot) \in C^{|\beta|}$. Then the following estimate holds:*

$$\|D^\beta g(\boldsymbol{\mu}, \cdot) - D^\beta g_M(\boldsymbol{\mu}, \cdot)\|_{L^\infty(\Omega)} \le (1 + \Lambda_M) \min_{z \in W_M^g} \|D^\beta g(\boldsymbol{\mu}, \cdot) - D^\beta z\|_{L^\infty(\Omega)} \quad \forall\, \boldsymbol{\mu} \in \mathcal{D} \tag{1.8.3}$$

Moreover, $\Lambda_M \le 2^M - 1$.

Even if this lemma provides some notion of stability, it is quite pessimistic and not relevant from an applicative point of view. In [33] several numerical simulations show that, in practice, the Lebesgue constant does not grow so fast for $M \to \infty$.

In [7] Maday et al. provide a very inexpensive but not rigorous a posteriori estimator that has been shown to be quite reliable numerically.

Given an approximation $g_M(\boldsymbol{\mu}, \cdot)$ for $M \le M_{max} - 1$ let $\mathcal{E}_M(\boldsymbol{\mu}, \cdot) \equiv \hat{\epsilon}_M(\mu) q_{M+1}(\cdot)$ where $\hat{\epsilon}_M(\mu) = |g(\boldsymbol{\mu}, \mathbf{t}_{M+1}) - g_M(\boldsymbol{\mu}, \mathbf{t}_{M+1})|$. It is possible to show that:

Lemma 1.8. *If $g(\boldsymbol{\mu}, \cdot) \in W_{M+1}^g$, then*

- $g(\boldsymbol{\mu}, \cdot) - g_M(\boldsymbol{\mu}, \cdot) = \pm \mathcal{E}_M(\boldsymbol{\mu}, \cdot)$
- $\|g(\boldsymbol{\mu}, \cdot) - g_M(\boldsymbol{\mu}, \cdot)\|_{L^\infty(\Omega)} \le \hat{\epsilon}_M(\mu)$

Obviously it is not so frequent that $g(\boldsymbol{\mu}, \cdot) \in W_{M+1}^g$; then the estimator is not rigorous, but, if $\|g(\boldsymbol{\mu}, \cdot) - g_M(\boldsymbol{\mu}, \cdot)\|_{L^\infty(\Omega)} \to 0$ for $M \to \infty$ very fast, we can expect that $\|g(\boldsymbol{\mu}, \cdot) - g_M(\boldsymbol{\mu}, \cdot)\|_{L^\infty(\Omega)} \approx \hat{\epsilon}_M(\boldsymbol{\mu})$.

Some extensions and comments

In this section we have introduced the Empirical Interpolation Method for the treatment of non-affine parametrized PDEs. The approximation of the variational form introduces some additional complications in the a posteriori estimation. We refer to [115] (proposition 4.5.1) for the definition of a rigorous estimator for parametrically non-affine linear elliptic equations[33]. In view of the application to hyperbolic problem, we state an example.

Example 1.4. *Let us consider the following two problems ($X = H_0^1(\Omega)$):*

$$\text{Find } u \in X \quad a(u,v) = \int_\Omega \nu \nabla u \nabla v \, dx + \int_\Omega (\mathbf{b}\cdot\nabla u) v \, dx = f(v) = \int_\Omega f v \, dx \quad \forall\, v \in X \tag{1.8.4a}$$

$$\text{Find } u^\epsilon \in X \quad a(u,v) = \int_\Omega \nu^\epsilon \nabla u \nabla v \, dx + \int_\Omega (\mathbf{b}^\epsilon\cdot\nabla u) v \, dx = f(v) = \int_\Omega f v \, dx \quad \forall\, v \in X \tag{1.8.4b}$$

Let α and α^ϵ be the coercivity constants associated with the two problems. With standard calculations, it is possible to prove the following estimate valid for $\Omega \subset \mathbb{R}^d$ $d = 2, 3$:

$$\|u - u^\epsilon\|_X \leq \frac{C_\Omega}{\alpha\alpha^\epsilon} \left(\|\nu - \nu^\epsilon\|_{L^\infty} + C'_\Omega \|\mathbf{b} - \mathbf{b}^\epsilon\|_{L^6} \right) \|f\|_{L^2} \tag{1.8.5}$$

where C_Ω is the Sobolev constant associated with the embedding $H_0^1 \subset L^2$ and C'_Ω with the embedding $H_0^1 \subset L^6$.

Inequality (1.8.5) shows how the required accuracy for the interpolation process is highly problem dependent and more precisely it is related to the coercivity constant: this means that for advection-dominated problems the required number of terms rapidly becomes unaffordable.

Before concluding, we give some references. In [17] the Empirical Interpolation method has been adapted to treat PDEs with highly non-linear terms (*discrete empirical interpolation method* DEIM). More recently in [34] a multi-domain EIM (*hp*-EIM) has been tested: first a partition of the original domain is constructed (*h-refinement*); then the standard EIM is applied to each subdomain independently. Finally in [115] the EIM has been applied to non-affine tensorial functions (*Multi-Component Empirical Interpolation method* MCEIM).

1.9 Conclusions

In this chapter the main features of the Reduced Basis method for elliptic and parabolic equations have been introduced. Before concluding, we make some observations related to advection-diffusion problems at the hyperbolic limit.

- Example 1.2 leads us to expect that the Kolmogorov N-width does not converge as fast as we need when we have discontinuities. For this reason we expect that the Greedy strategy will be weakened by the presence of discontinuities that depends on the parameters.
- Example 1.4 represents another important point to be addressed: how many terms in the affine expansion do we need in order to guarantee satisfactory approximation properties to the surrogate solution? In addition we expect that in the non-linear case the number of terms required in the approximation would be even larger because we have to take into account the application of the DEIM to treat the non linear terms.

[33] In [15] the estimator is extended to the non linear case.

These two observations justify the approach we will propose for hyperbolic conservation laws in chapter 3.

Chapter 2

Reduced Basis Method for PDEs in parametrized domains

2.1 Introduction and motivations

In this chapter the Reduced Basis (RB) method is applied to partial differential equations in parametrized domains[1], say $\Omega(\boldsymbol{\mu})$ where $\boldsymbol{\mu} \in \mathcal{D}$ with $\mathcal{D}$ a set of (not necessary) geometrical parameters. The RB recipe requires that Ω is a parameter independent domain: indeed, if we wish to consider linear combinations of snapshots, these snapshots must be defined on a common spatial configuration. Therefore, in order to transform the problem statement over the original domain into an equivalent problem statement over the reference domain, it is necessary to define a transformation between the pre-image parameter independent Ω and the parameter dependent actual domain $\Omega(\boldsymbol{\mu})$:

$$\boldsymbol{T} : \mathcal{D} \times \mathbb{R}^d \to \mathbb{R}^d \quad \boldsymbol{T}(\boldsymbol{\mu}, \cdot) : \Omega \to \Omega(\boldsymbol{\mu}) \tag{2.1.1}$$

Furthermore, as we have already explained in the first chapter, in order to perform a computational offline-online decomposition, it is crucial that the bilinear form and the load functional associated with the original problem are in the following *parametrically affine form*:

$$a(u, v, \boldsymbol{\mu}) = \sum_{q=1}^{Q_a} \Theta_q^a(\boldsymbol{\mu}) a^q(u, v) \quad F(v, \boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \Theta_q^f(\boldsymbol{\mu}) F^q(v) \tag{2.1.2}$$

A formulation of this type in the reference domain requires that the map $\boldsymbol{T}$ is a global bijective piecewise-affine transformation [106]. For many applications it is possible to consider directly these maps but for more complex geometries this hypothesis must be relaxed.

Common strategies for shape deformation involve the use of (i) the coordinates of the boundary points as design variables (*local boundary variation* [85]) or (ii) some families of basis shapes combined by means of a set of control points (*polynomial boundary parametrizations* [23]). These techniques focus on the representation of the boundary and do not build up a global transformation from the reference configuration to the deformed

[1] In shape optimization or fluid-structure interaction problems the introduction of a reference configuration is crucial in order to avoid remeshing operations. This is why in recent years a great effort has been made in order to develop efficient and reliable techniques for this kind of problems. The first works on the topic were [103, 105] and the Ph. D. theses [104, 121]. Several further references are given in the following.

one. This is why in the RB context- in which we aim at representing smooth global deformations of the reference domains with a relatively small number of parameters- other strategies are preferred.

In recent years, much research has been focused on defining efficient strategies to deal with parametrized domains in the context of Reduced Basis method[2]. For this reason, the aim of the present chapter is first to present the state of the art and then to introduce a new approach based on a slight modification of one of the standard techniques.

The structure of the chapter is the following one:

- in section 2.2, we review the classic approach based on piecewise affine maps, [106]. In the presentation we also derive the formulation on the reference domain for a simple second order-scalar problem;
- in section 2.3, we analyse three shape deformation techniques that are proved to be well suited within a RB framework: Free Form Deformation (FFD), parametrizations based on Radial Basis Functions (RBF) and transfinite maps. The first two maps are well known in Computer Graphics [112, 119] and only recently applied to the shape optimization context [111, 81]. On the other hand, transfinite maps constitute a generalization of the Gordon Hall formula [43];
- in section 2.4, we introduce the technique we propose based on the above mentioned transfinite maps. Some theoretical evidence will be stated in order to motivate this approach.
- at the end of this theoretical presentation, some numerical examples are provided in section 2.5 to compare the different approaches;
- in section 2.6, some conclusions and a general overview of the possible applications to hyperbolic problems are given.

Before starting the presentation, we introduce the notation.

The actual (or physical) domain is referred to as $\Omega(\boldsymbol{\mu})$, whereas the reference domain is referred to as Ω; we refer to the generic point of Ω as $\boldsymbol{x}$ and to the generic point of $\Omega(\boldsymbol{\mu})$ as $\boldsymbol{y}$. In addition, $X(\boldsymbol{\mu})$ and X indicate a suitable Hilbert space defined onto $\Omega(\boldsymbol{\mu})$ and Ω, respectively.

The map between the reference and the actual configuration is referred to as $\boldsymbol{T}$ and its Jacobian as $\boldsymbol{J_T}$.

2.2 Affine geometrical parametrization

Let $\Omega(\boldsymbol{\mu}) \subset \mathbb{R}^d$ be a connected set expressed as:

$$\bar{\Omega}(\boldsymbol{\mu}) = \bigcup_{l=1}^{K_{dom}} \bar{\Omega}_l(\boldsymbol{\mu}) \tag{2.2.1}$$

where the $\Omega_k(\boldsymbol{\mu})$ are mutually non-overlapping open regions[3] ($\Omega_k(\boldsymbol{\mu}) \cap \Omega_{k'}(\boldsymbol{\mu}) = \varnothing$ if $k \neq k'$).

We now choose a value $\boldsymbol{\mu}_{ref} \in \mathcal{D}$ and define our reference domain as $\Omega = \Omega(\boldsymbol{\mu}_{ref})$. Clearly:

$$\bar{\Omega} = \bigcup_{k=1}^{K_{dom}} \bar{\Omega}_k \tag{2.2.2}$$

[2] We refer to [78, 55] for a comprehensive presentation of the main techniques proposed.

[3] Tipically the different regions correspond to different PDE coefficients; however, as we will see in the examples, they can also be introduced for algorithmic purposes.

where $\Omega_k = \Omega_k(\boldsymbol{\mu}_{ref})$.

The partition $\{\Omega_k\}_k$ represents a coarse domain decomposition of the global domain. Thus we build our FE approximation on a fine $\mathcal{N}$-subtriangulation of $\{\Omega_k\}$. This FE triangulation ensures that the FE approximation treats the discontinuities in PDE parameters associated with different regions and, as we can show in section 2.2.3, it plays an important role in the generation of the affine representation (2.1.2).

Now we have the ingredients to state the so-called *Affine Geometry Precondition*. We are able to deal with any original domain $\Omega(\boldsymbol{\mu})$ and associated regions as in (2.2.1) that admits a domain decomposition such that:

1. $\bar{\Omega}_k(\boldsymbol{\mu}) = \boldsymbol{T}^{\mathrm{aff},k}(\boldsymbol{\mu}, \bar{\Omega}_k)$, for $1 \le k \le K_{dom}$ and for affine maps $\boldsymbol{T}^{\mathrm{aff},k}(\boldsymbol{\mu}) : \Omega_k \to \Omega_k(\boldsymbol{\mu})$.

2. each $\boldsymbol{T}^{\mathrm{aff},k}$ is *individually bijective* and *collectively continuous* (i.e., $\boldsymbol{T}^{\mathrm{aff},k}(\boldsymbol{\mu}, \boldsymbol{x}) = \boldsymbol{T}^{\mathrm{aff},k'}(\boldsymbol{\mu}, \boldsymbol{x})$ for $\boldsymbol{x} \in \bar{\Omega}_k \cap \bar{\Omega}_{k'}$, for each $k \neq k'$).

The Affine Geometric Precondition is necessary for an affine parameter dependence as defined in (2.1.2). We observe that K_{dom} does not depend on the FE subtriangulation. Therefore, it could be also defined with respect to the exact problem.

In order to make some calculations, we introduce an explicit notation for the affine maps:

$$\boldsymbol{T}_i^{\mathrm{aff},k}(\boldsymbol{\mu}, \mathbf{x}) = C_i^{\mathrm{aff},k}(\boldsymbol{\mu}) + \sum_{j=1}^{d} G_{i,j}^{\mathrm{aff},k}(\boldsymbol{\mu}) x_j \quad 1 \le i \le N \quad C^{\mathrm{aff},k} : \mathcal{D} \to \boldsymbol{R}^d,\ G^{\mathrm{aff},k} : \mathcal{D} \to \boldsymbol{R}^{d,d} \tag{2.2.3}$$

We point out that we can interpret the local maps in terms of a global transformation $\boldsymbol{T}^{\mathrm{aff}}(\boldsymbol{\mu}) : \Omega \to \Omega(\boldsymbol{\mu})$:

$$\boldsymbol{T}^{\mathrm{aff}}(\boldsymbol{\mu}, \boldsymbol{x}) = \boldsymbol{T}^{\mathrm{aff},k}(\boldsymbol{\mu}, \boldsymbol{x}) \qquad k = \min_{k' : \boldsymbol{x} \in \bar{\Omega}_{k'}} k' \tag{2.2.4}$$

We highlight the one to one property- the arbitrariness of the min-choice in (2.2.4) is a consequence of the collectively continuity- and that this global continuous mapping is compatible with the second order PDE variational formulation (indeed, as a consequence of the Fubini-Tonelli theorem, $w \in H^1(\Omega_o(\boldsymbol{\mu})) \Leftrightarrow w \circ \boldsymbol{T}^{\mathrm{aff}}(\mu) \in H^1(\Omega)$). This proves that the mapped problem belongs to the classical "conforming" variety.

Now, in order to get familiar with the scope of the affine mappings, we consider in the next subsection the case of a single subdomain; then we extend the method to the case of several subdomains.

2.2.1 Affine mappings: single subdomains

In this section we deal with parametrized domains that could be described by an affine transformation map of the type (2.2.3):

We point out that an affine transformation maps straight lines into straight lines, thus a n-hedron is mapped into a n-hedron. In addition also ellipsoids are mapped into ellipsoids. These properties are crucial for the description of domains relevant in the applications.

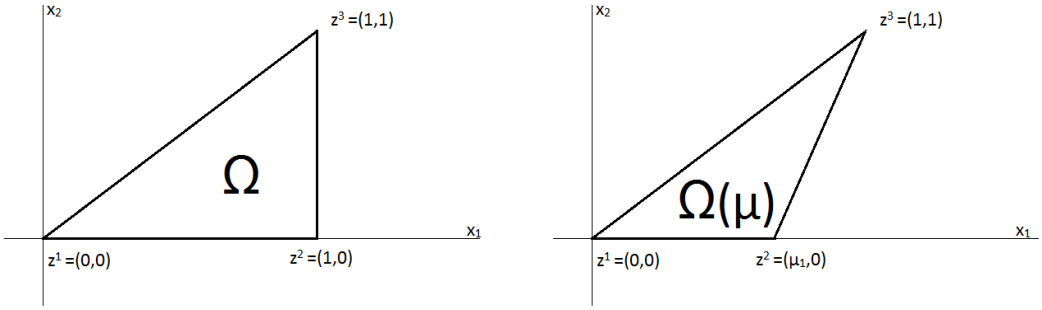
We now restrict ourselves to the bidimensional case ($d = 2$). In order to completely define an affine transformation we have to prescribe 6 coefficients; we can univocally identify C^{aff} and G^{aff} from the relationship between 3 non-collinear points in the reference domain

Ω - say $\bar{\boldsymbol{z}}^j$, $1 \le j \le 3$ - and the respective 3 parametrized image points in $\Omega(\boldsymbol{\mu})$ -say $\boldsymbol{z}^j(\boldsymbol{\mu})$, $1 \le j \le 3$. By simple calculations we obtain that:

$$\begin{bmatrix} C_1^{\text{aff}}(\boldsymbol{\mu}) \\ C_2^{\text{aff}}(\boldsymbol{\mu}) \\ G_{1,1}^{\text{aff}}(\boldsymbol{\mu}) \\ G_{1,2}^{\text{aff}}(\boldsymbol{\mu}) \\ G_{2,1}^{\text{aff}}(\boldsymbol{\mu}) \\ G_{2,2}^{\text{aff}}(\boldsymbol{\mu}) \end{bmatrix} = \begin{bmatrix} 1 & 0 & \bar{z}_1^1 & \bar{z}_2^1 & 0 & 0 \\ 0 & 1 & 0 & 0 & \bar{z}_1^1 & \bar{z}_2^1 \\ 1 & 0 & \bar{z}_1^2 & \bar{z}_2^2 & 0 & 0 \\ 0 & 1 & 0 & 0 & \bar{z}_1^2 & \bar{z}_2^2 \\ 1 & 0 & \bar{z}_1^3 & \bar{z}_2^3 & 0 & 0 \\ 0 & 1 & 0 & 0 & \bar{z}_1^3 & \bar{z}_2^3 \end{bmatrix}^{-1} \begin{bmatrix} z_1^1(\boldsymbol{\mu}) \\ z_2^1(\boldsymbol{\mu}) \\ z_1^2(\boldsymbol{\mu}) \\ z_2^2(\boldsymbol{\mu}) \\ z_1^3(\boldsymbol{\mu}) \\ z_2^3(\boldsymbol{\mu}) \end{bmatrix}, \tag{2.2.5}$$

where $z_i^j(\boldsymbol{\mu})$ and $\bar{z}_i^j$ are the components of the vectors $\boldsymbol{z}^j(\boldsymbol{\mu})$ and $\bar{\boldsymbol{z}}^j$, respectively. We observe that the matrix is $\boldsymbol{\mu}$-independent: the parametric dependence is limited to the image points.

In order to clarify the technique presented, we provide an illustrative example from [106].



(a) (b)

Figure 2.1: (a) reference configuration Ω, (b) actual configuration $\Omega(\boldsymbol{\mu})$.

Example 2.1. *Let us consider the configurations in Figure 2.1, relation (2.2.5) consequently becomes:*

$$\begin{bmatrix} C_1^{\text{aff}}(\mu_1) \\ C_2^{\text{aff}}(\mu_1) \\ G_{1,1}^{\text{aff}}(\mu_1) \\ G_{1,2}^{\text{aff}}(\mu_1) \\ G_{2,1}^{\text{aff}}(\mu_1) \\ G_{2,2}^{\text{aff}}(\mu_1) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ \mu_1 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

and hence:

$$C^{\text{aff}}(\mu_1) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad G^{\text{aff}}(\mu_1) = \begin{bmatrix} \mu_1 & 1-\mu_1 \\ 0 & 1 \end{bmatrix}$$

When dealing with standard triangles, we did not need any hypothesis on the shape of the reference domain Ω. More in general, we can consider (sub)domains of arbitrary shape. The only requirement we need is that the transformation is of the form (2.2.3).

In the following we focus on the so-called "elliptical triangles". At the end of the section we briefly consider the more general "curvy triangles".

An "elliptical triangle" is defined by three nodes $\bar{\boldsymbol{z}}^1(\boldsymbol{\mu})$, $\bar{\boldsymbol{z}}^2(\boldsymbol{\mu})$ and $\bar{\boldsymbol{z}}^3(\boldsymbol{\mu})$, by the two straight edges $\overline{\bar{\boldsymbol{z}}^1(\boldsymbol{\mu})\bar{\boldsymbol{z}}^2(\boldsymbol{\mu})}$ and $\overline{\bar{\boldsymbol{z}}^1(\boldsymbol{\mu})\bar{\boldsymbol{z}}^3(\boldsymbol{\mu})}$ and by the elliptical arc:

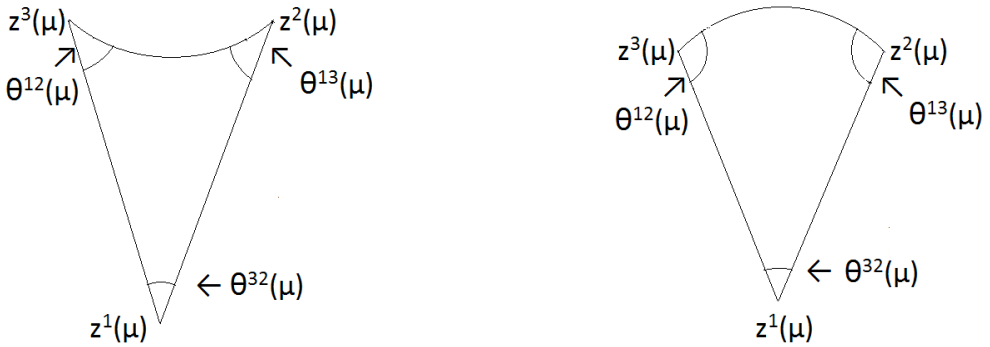
$$\overline{\bar{\boldsymbol{z}}^2(\boldsymbol{\mu})\bar{\boldsymbol{z}}^3(\boldsymbol{\mu})}^{arc} = \left\{ O(\boldsymbol{\mu}) + Q_{rot}(\boldsymbol{\mu})S(\boldsymbol{\mu}) \begin{bmatrix} cos(t) \\ sin(t) \end{bmatrix} \middle| \, t \in [t_2, t_3] \right\},$$

where

$$Q_{rot}(\boldsymbol{\mu}) = \begin{bmatrix} cos(\phi(\boldsymbol{\mu})) & -sin(\phi(\boldsymbol{\mu})) \\ sin(\phi(\boldsymbol{\mu})) & cos(\phi(\boldsymbol{\mu})) \end{bmatrix} \qquad S(\boldsymbol{\mu}) = \begin{bmatrix} \rho_1((\boldsymbol{\mu})) & 0 \\ 0 & \rho_2((\boldsymbol{\mu})) \end{bmatrix}$$

and $O(\boldsymbol{\mu})$ is the origin of the reference coordinate system.

There are two types of elliptical triangles: inwards (convex) and outwards(concave). Figure 2.2 shows the differences.



(a) (b)

Figure 2.2: (a) inwards elliptical triangle, (b) outwards elliptical triangle

Now we write the first point as:

$$\boldsymbol{z}^1(\boldsymbol{\mu}) = O(\boldsymbol{\mu}) + \omega Q_{rot}(\boldsymbol{\mu})S(\boldsymbol{\mu}) \begin{bmatrix} cos(t^1) \\ sin(t^1) \end{bmatrix}$$

thus we can express the three image points $\boldsymbol{z}_o^m(\boldsymbol{\mu})$ in the compact form ($\omega^1 = \omega, \omega^2 = \omega^3 = 1$):

$$\boldsymbol{z}^m(\boldsymbol{\mu}) = O(\boldsymbol{\mu}) + \omega^m Q_{rot}(\boldsymbol{\mu})S(\boldsymbol{\mu}) \begin{bmatrix} cos(t^m) \\ sin(t^m) \end{bmatrix}, \text{ for } m = 1, 2, 3. \tag{2.2.6}$$

We restrict our attention to *proper elliptical triangles*[4] i.e. triangles such that $0 < \theta^{12}, \theta^{13}, \theta^{32} < \pi$.

In [106] it is proven that:

- the transformation associated with the deformation (2.2.6) can be described through a parametric affine map of the form (2.2.3) with the following mapping coefficients:

$$C^{\text{aff}}(\boldsymbol{\mu}) = O(\mu) - Q_{rot}(\boldsymbol{\mu})S(\boldsymbol{\mu})S(\boldsymbol{\mu}_{ref})^{-1}Q_{rot}(\boldsymbol{\mu}_{ref})^T O(\mu_{ref})$$
$$G^{\text{aff}}(\boldsymbol{\mu}) = Q_{rot}(\boldsymbol{\mu})S(\boldsymbol{\mu})S(\boldsymbol{\mu}_{ref})^{-1}Q_{rot}(\boldsymbol{\mu}_{ref})^T;$$

[4] If we considered non-proper triangles in the multidomain context, we would have to introduce an additional control parameter to avoid the risk of discontinuous global mapping. This is not absolutely convenient; on one hand, the implementation would become more involved and on the other hand - with respect to our knowledge - there are not examples of parametrized domains that requires the use on non-proper triangles.

- the elliptical triangle is *proper* if and only if

$$\boldsymbol{z}^1 \in \begin{cases} \mathcal{R}_{In}(\boldsymbol{\mu}) & \text{inwards case} \\ \mathcal{R}_{Out}(\boldsymbol{\mu}) & \text{outwards case,} \end{cases} \tag{2.2.7}$$

where denoted by $\boldsymbol{n}^2(\mu)$ and $\boldsymbol{n}^3(\mu)$ the outward-oriented normals to the ellipse at $\bar{\boldsymbol{z}}^2(\mu)$ and $\bar{\boldsymbol{z}}^3(\mu)$, respectively and by $\boldsymbol{n}^{2,3}(\mu)$ the outward normal to the line segment $\overline{\bar{\boldsymbol{z}}^2(\boldsymbol{\mu})\bar{\boldsymbol{z}}^3(\boldsymbol{\mu})}$ in the middle point;

$$\mathcal{R}_{In}(\boldsymbol{\mu}) = \left\{ \boldsymbol{z}^1(\boldsymbol{\mu}) \in \mathbb{R}^2 \text{ such that } \begin{array}{l} (\boldsymbol{z}^1(\boldsymbol{\mu}) - \boldsymbol{z}^2(\boldsymbol{\mu}))^T n^2(\boldsymbol{\mu}) < 0, \\ (\boldsymbol{z}^1(\boldsymbol{\mu}) - \boldsymbol{z}^3(\boldsymbol{\mu}))^T n^3(\boldsymbol{\mu}) < 0 \\ (\boldsymbol{z}^1(\boldsymbol{\mu}) - \boldsymbol{z}^{2,3}(\boldsymbol{\mu}))^T n^{2,3}(\boldsymbol{\mu}) < 0 \end{array} \right\}$$

while

$$\mathcal{R}_{Out}(\boldsymbol{\mu}) = \left\{ \boldsymbol{z}^1(\boldsymbol{\mu}) \in \mathbb{R}^2 \text{ such that } \begin{array}{l} (\boldsymbol{z}^1(\boldsymbol{\mu}) - \boldsymbol{z}^2(\boldsymbol{\mu}))^T n^2(\boldsymbol{\mu}) > 0, \\ (\boldsymbol{z}^1(\boldsymbol{\mu}) - \boldsymbol{z}^3(\boldsymbol{\mu}))^T n^3(\boldsymbol{\mu}) > 0 \end{array} \right\}$$

Figure 2.3 illustrates condition (2.2.7).

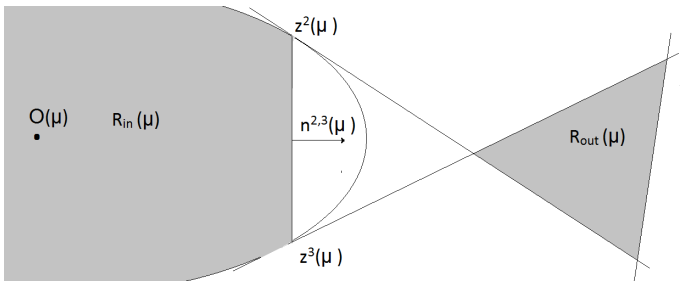


Figure 2.3: regions where $\boldsymbol{z}^1(\boldsymbol{\mu})$ has to be located in order to satisfy the proper condition.

We conclude this section with some remarks.

1. It is evident that the construction of proper elliptical triangles requires some care to ensure controlled elliptical arcs, continuous mappings and well defined internal angles. On the other hand, it is quite simple to derive explicit conditions on ω such that the angle condition is satisfied in (2.2.6)and, consequently, such that the condition (2.2.7) is *parameter independent* (that means that could be expressed in terms of the reference configuration). In addition elliptical triangles are consistent under refinement: if we split either a straight edge or the elliptical edge of a proper elliptical triangle, we obtain two daughter elliptical triangles that are also proper[5].

2. The extension from elliptical triangles to general curvy triangles is formally straightforward: we simply substitute $cos(t)$ and $sin(t)$ with $g_1(t)$ and $g_2(t)$ where, with respect to strictly convex (inwards) or concave (outwards) curvy triangles, we can demonstrate that the internal angle condition (2.2.7) is applicable, parameter independent and reducible to a small set of algebraic conditions for a proper choice of the center[6]. However, it is not in general possible to find a simple closed form for the internal angle condition.

[5] This edge split consistency plays an important role in the algorithm proposed for the multidomain case.

[6] In the numerical examples in section 2.2.4 we will use curvy triangles.

3. All the discussion has been set in a bidimensional environment: the extension to the tridimensional case is possible but not trivial [41, 118]. In [106] there is an example for a linear elasticity problem.

2.2.2 Piecewise affine mappings: multiple subdomains

In order to deal with more complex geometries, it is necessary to consider piecewise affine mappings. In this section we consider "Elliptical-edge" domains, i.e., domains and associated regions whose boundary and internal interfaces can be represented by either straight edges or the elliptical arcs described previously.

In [106] the following mapping process has been proposed:

- we define the RB triangulation (2.2.1). The triangulation must be compatible with the existence of a piecewise globally continuous affine map;
- we build up the affine maps for each subdomain: by taking into account the condition (2.2.7).

We observe that, since the point selection in elliptical triangles is not arbitrary, the first two steps are coupled.

Here we present the algorithm implemented in `rbMIT` for the construction of the RB triangulation of the whole domain[7]. We point out that the following algorithm does not ensure non-singular and efficient transformations. However, several numerical tests demonstrate that a proper initialization of the algorithm determines a well-behaved triangulation. In section 2.2.4 we will come back to this topic with two numerical examples.

- **Stage 1:** starting from the control points given by the user, the software focuses on all elliptical edges making part of the domain; for each of them an elliptical triangle is introduced. In order to preserve (2.2.7), splittings of elliptical triangles are performed. For each splitting, the new point created by the introduction of the new elliptical triangle is denoted by *interior control point*[8].
- **Stage 2:** the remain part of the domain is meshed via standard triangle:
  1. we introduce a Delaunay triangulation [22] moving from all the control points (the ones initially provided by the user and the interior ones);
  2. we search for the edges that belong to the domain boundary but do not belong to the Delaunay triangulation;
  3. we split all these edges by creating boundary/internal edges or interior control points;
  4. we iterate points 1-3 procedure until no selected edges remain.

2.2.3 Formulation on the reference domain

Here we formally derive the problem formulation with respect to the reference domain. In section 2.3.4 we generalize this discussion to more general non-parametrically affine maps. Let $X(\boldsymbol{\mu}) \subset H^1(\Omega(\boldsymbol{\mu}))$ be a given Hilbert space. We consider the following problem[9]

$$\text{Given } \boldsymbol{\mu} \in \mathcal{D} \quad \text{compute } s(\boldsymbol{\mu}) = f(u(\boldsymbol{\mu}), \boldsymbol{\mu}) \tag{2.2.8}$$

[7] The other two steps are local. Therefore the procedure is the same described for the single subdomain.

[8] The interior control points constitute the subset of the control point that are linked to the elliptical triangles.

[9] By writing the problem directly in the weak form, we do not have to deal explicitly with boundary conditions. This critical aspect is briefly addressed at the end of the section.

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

$$a(u(\boldsymbol{\mu}), v, \boldsymbol{\mu}) = f(v, \boldsymbol{\mu}) \quad \forall\, v \in X(\boldsymbol{\mu}).$$

We consider $a : X(\boldsymbol{\mu}) \times X(\boldsymbol{\mu}) \times \mathcal{D} \to \mathbb{R}$ of the form

$$a(u, v, \boldsymbol{\mu}) = \sum_{k=1}^{K_{dom}} \int_{\Omega_k(\boldsymbol{\mu})} \begin{bmatrix} \frac{\partial u}{\partial y_1} & \frac{\partial u}{\partial y_2} & u \end{bmatrix} \mathcal{K}_{o,k}^{ij}(\boldsymbol{\mu}) \begin{bmatrix} \frac{\partial v}{\partial y_1} \\ \frac{\partial v}{\partial y_2} \\ v \end{bmatrix} d\boldsymbol{y}. \tag{2.2.9}$$

Here $\mathcal{K}_{o,k} : \mathcal{D} \to \mathbb{R}^{d \times d}$ is a symmetric semi-positive definite matrix such that the bilinear form $a(\cdot, \cdot, \boldsymbol{\mu})$ is coercive. Similarly, $f : X(\boldsymbol{\mu}) \times \mathcal{D} \to \mathbb{R}$ is of the form:

$$f(v, \boldsymbol{\mu}) = \sum_{k=1}^{K_{dom}} \int_{\Omega_k(\boldsymbol{\mu})} \mathcal{F}_{o,l}(\boldsymbol{\mu}) v \, d\boldsymbol{y} \tag{2.2.10}$$

At the end of the section we discuss about restrictions (2.2.9) and (2.2.10).

Thanks to the chain rule, we obtain that

$$\frac{\partial}{\partial x_i} = \left((G^{\text{aff},k})^{-1} \right)_{i,j} \frac{\partial}{\partial y_j};$$

therefore the transformed bilinear form a can be expressed as:

$$a(u, v, \boldsymbol{\mu}) = \sum_{k=1}^{K_{dom}} \int_{\Omega_k} \begin{bmatrix} \frac{\partial u}{\partial x_1} & \frac{\partial u}{\partial x_2} & u \end{bmatrix} \mathcal{K}_k^{ij}(\boldsymbol{\mu}) \begin{bmatrix} \frac{\partial v}{\partial x_1} \\ \frac{\partial v}{\partial x_2} \\ v \end{bmatrix} d\boldsymbol{x}; \tag{2.2.11}$$

where $\mathcal{K}_k : \mathcal{D} \to \mathbb{R}^{d,d}$ is given by:

$$\mathcal{K}_k(\boldsymbol{\mu}) = \det \mathcal{G}^k(\boldsymbol{\mu}) \, \mathcal{G}^k(\boldsymbol{\mu}) \mathcal{K}_{o,k}(\boldsymbol{\mu}) (\mathcal{G}^k(\boldsymbol{\mu}))^T$$

with:

$$\mathcal{G}^k(\boldsymbol{\mu}) = \begin{pmatrix} (G^{\text{aff},k})^{-1} & 0 \\ 0 & 1 \end{pmatrix}.$$

In the same way, the transformed linear form in the reference configuration is

$$f(v, \boldsymbol{\mu}) = \sum_{k=1}^{K_{dom}} \int_{\Omega_k} \mathcal{F}^k(\boldsymbol{\mu}) v, \tag{2.2.12}$$

with $\mathcal{F}^k(\boldsymbol{\mu}) = \det \mathcal{G}^k(\boldsymbol{\mu}) \mathcal{F}_o^k(\boldsymbol{\mu})$.

At this point it is straightforward to write the bilinear form and the load in the parametrically affine form (2.1.2):

$$a(v, w, \boldsymbol{\mu}) = \mathcal{K}_{1,1}^1(\mu) \int_{\Omega_1} \frac{\partial v}{\partial x_1} \frac{\partial w}{\partial x_1} \, d\boldsymbol{x} + \mathcal{K}_{1,2}^1(\mu) \int_{\Omega_1} \frac{\partial v}{\partial x_1} \frac{\partial w}{\partial x_2} \, d\boldsymbol{x} + \cdots + \mathcal{K}_{d,d}^{K_{dom}}(\mu) \int_{\Omega_{K_{dom}}} vw \, d\boldsymbol{x}$$

We conclude this section with some comments.

- This discussion provides a constructive proof that the Affine Geometry Precondition is sufficient to obtain a parametrically affine bilinear form;
- in practical situations many entries in $\mathcal{K}_{o,k}$ are zero or may be redundant: this is why a good choice of the user-provided initial control points for the RB triangulation are fundamental in the preprocessing to simplify the tensor; this explains why symbolic manipulation techniques are particularly useful in the simplification of the bilinear form found after applying the domain decomposition stage[10].
- The hypotheses on the structure of the bilinear form can be relaxed: for instance, we can admit affine polynomial dependence on $\boldsymbol{x}$ in both $\mathcal{K}_{o,k}$ and $\mathcal{F}_{o,k}$ because a composition between a polynomial and a parametrically affine function is parametrically affine[11]. Furthermore, in the absence of geometric variation in a particular region $\Omega_l(\boldsymbol{\mu})$ every separable form in $\boldsymbol{x}$ and $\boldsymbol{\mu}$ is allowed. As regards Neumann and Robin boundary conditions, it is trivial that homogeneous Neumann conditions create no problem; for other choices we get equations in the affine form (2.1.2) only in the case of straight or circular edges[12].

2.2.4 Two numerical examples

In this paragraph we apply the methodology presented above to two different problems[13]: in the first one, we just verify that, in the case of elliptical arcs, the method is extremely efficient. In the second test case, we highlight the fact that, for more complex shapes, the procedure could produce a bad triangulation or fail.

In the following we do not consider a particular differential equation but we just focus on the parametrization procedure.

A half-elliptic obstacle

We consider the following domain where $\mathcal{E} = \mathcal{E}(\mu_1, \mu_2)$ is the half elliptic obstacle centered in $(0,0)$ with semi-principal axes of length μ_1, μ_2.

$$\Omega(\boldsymbol{\mu}) = \{\boldsymbol{x} \in (-2,2)^2 : \boldsymbol{x} \notin \mathcal{E}(\mu_1, \mu_2)\} \tag{2.2.13}$$

The deformation to the obstacle is described by the following linear tensor.

$$\boldsymbol{T}(\boldsymbol{\mu}, \boldsymbol{x}) = \begin{bmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{2.2.14}$$

[10] For instance the software `rbMit`[102] uses extensively this kind of techniques.

[11] On the other hand a composition between a parametrically affine not polynomial function and a parametrically affine function is in general not affine.

[12] Suppose that the map $\boldsymbol{\gamma}^k : \mathcal{D} \times [0,1]^{N-1} \to \mathbb{R}^N$ describes the boundary of $\partial\Omega_k(\boldsymbol{\mu})$ and that $\bar{\boldsymbol{\gamma}}^k(\boldsymbol{t}) := \boldsymbol{\gamma}^k(\boldsymbol{\mu}_{ref}, \boldsymbol{t})$. Then, under the hypothesis of affine mappings we have that:

$$\boldsymbol{\gamma}^k(\boldsymbol{\mu}, \boldsymbol{t}) = \boldsymbol{C}^{\mathrm{aff},k}(\boldsymbol{\mu}) + \boldsymbol{G}^{\mathrm{aff},k}(\boldsymbol{\mu})\bar{\boldsymbol{\gamma}}^k(\boldsymbol{t})$$

Integrating by part we obtain ($\boldsymbol{J}_{\bar{\boldsymbol{\gamma}}^k}$ is the Jacobian of $\bar{\boldsymbol{\gamma}}^k$)

$$\int_{\partial\Omega(\boldsymbol{\mu})} f v \, d\sigma = \sum_k \int_{\partial\Omega(\boldsymbol{\mu}) \cap \bar{\Omega}_{o,k}(\boldsymbol{\mu})} f v \, d\sigma = \sum_k \int_{[0,1]^{N-1}} f(\boldsymbol{\gamma}(\boldsymbol{t})) v(\boldsymbol{\gamma}(\boldsymbol{t}) \sqrt{1 + |\boldsymbol{G}^{\mathrm{aff},k}(\boldsymbol{\mu}) \boldsymbol{J}_{\bar{\boldsymbol{\gamma}}^k}(\boldsymbol{t})|^2} \det \boldsymbol{G}^{\mathrm{aff},k}(\boldsymbol{\mu}) \, d\boldsymbol{t}$$

Clearly the last summation could be written in a parametrically affine way if and only if for each boundary edge of $\bar{\Omega}_k(\boldsymbol{\mu})$ we have that the matrix $\boldsymbol{G}^{\mathrm{aff},k}(\boldsymbol{\mu})$ is orthogonal (that means that the edge is spherical) or $\boldsymbol{J}_{\bar{\boldsymbol{\gamma}}^k}(\cdot) = cost$ (that means that the edge is straight).

[13] The simulations have been performed using `rbMIT` [102].

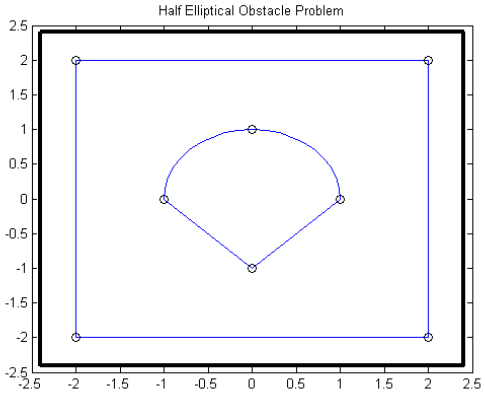
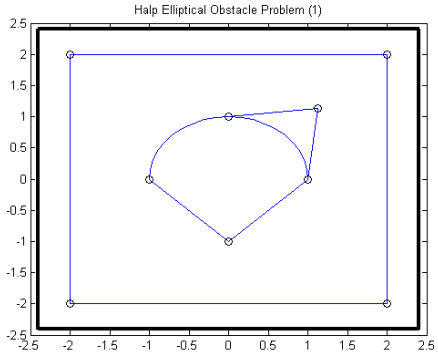
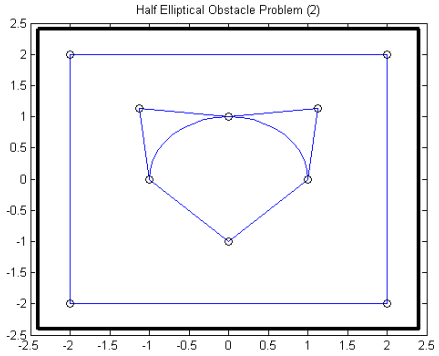


Figure 2.4: half ellipse obstacle problem

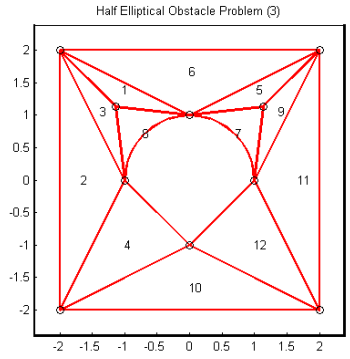
Figure 2.4 shows the reference configuration. Figure 2.5 explains the different steps of the procedure. The algorithm starts by considering the elliptical arcs: it constructs an elliptical triangle for each one and then completes the triangulation. With respect to Stage 2, in this case we just verify that the triangulation coincides with the Delaunay grid. Moreover, we observe that in this case the underlined FE mesh does not suffer from the constraints imposed by the RB triangulation.



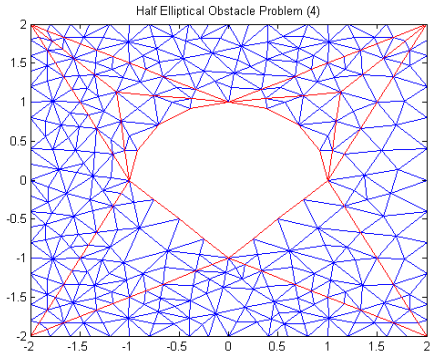
(a) First elliptical triangle



(b) Conclusion of the first stage



(c) Conclusion of the second stage



(d) Final Mesh

Figure 2.5: elliptical obstacle problem: definition of the RB triangulation (in red) and of the underlined FE triangulation (in blue).

NACA symmetric profile

The second problem we consider consists in the rotation of a NACA symmetric profile. The chord length is fixed to 1, the thickness is μ_1. The centre of rotation is $(0,0)$ and corresponds to the barycentre of the wing.

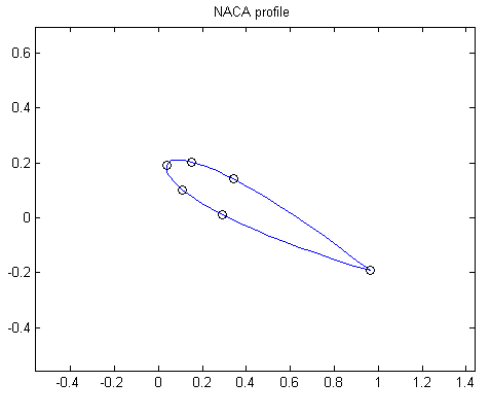


Figure 2.6: NACA airfoil.

The airfoil could be parametrized in the following way[14]

$$\boldsymbol{\gamma} = \begin{bmatrix} cos(\mu_1) & -sin(\mu_1) \\ sin(\mu_1) & cos(\mu_1) \end{bmatrix} \boldsymbol{p}(t) \tag{2.2.15}$$

where for $t \in [0, \sqrt{0.3}]$:

$$\boldsymbol{p}(t) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} -1 & 0 \\ 0 & \pm\mu_1 \end{bmatrix} \begin{bmatrix} 1 - t^2 \\ 0.2969t - 0.1260t^2 - 0.3520t^4 + 0.2832t^6 - 0.1021t^8 \end{bmatrix} \tag{2.2.16}$$

while for $t \in [\sqrt{0.3}, 1]$

$$\boldsymbol{p}(t) = \begin{bmatrix} 1 & 0 \\ 0 & \pm\mu_1 \end{bmatrix} \begin{bmatrix} t^2 \\ 0.2969t - 0.1260t^2 - 0.3520t^4 + 0.2832t^6 - 0.1021t^8 \end{bmatrix} \tag{2.2.17}$$

Figure 2.6 shows the reference configuration. Figure 2.7 describes the different steps of the procedure. In this case the domain decomposition $\{\Omega_k\}_k$ - see (2.2.2)- we get is locally extremely stretched; as a consequence, the FE mesh generator exhibits some difficulties to build a conformal FE triangulation with respect to the elements Ω_k induced by the RB procedure[15].

Before concluding, there are two important points to discuss.

- With a "blind" parametrization the algorithm would not be able to provide the RB triangulation; this is why, to obtain the previous results, two different tricks have been used: first the parametrization for the airfoil is chosen in order to eliminate the singularity of the derivative in $(0,0)$; then we split the boundary in different segments and we define $(0,0)$ with respect to the origin $(1,0)$. These two tricks are not evident and are strictly linked to the parametrized domain under consideration. Therefore, particular care is required to deal with complex shapes.

[14] The numerical example is taken from [46], (see also [83]).

[15] We cannot exclude that the RB triangulation could be improved through a different choice of the initial points. However, we think that the problems associated with the FE mesh generator are extremely difficult to be solved.

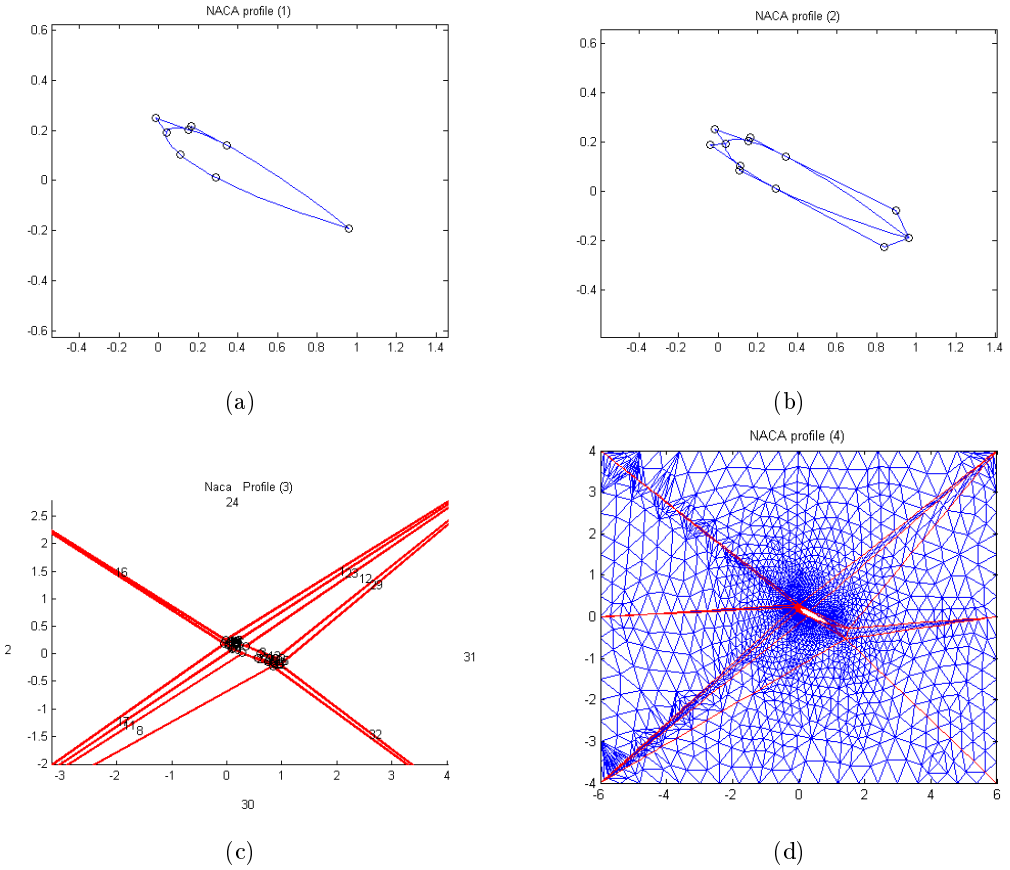


Figure 2.7: flux around a NACA airfoil: definition of the RB and FE triangulation. (a)first curvy triangle, (b)conclusion of the first stage, (c)conclusion of the second stage, (d)final mesh

- Let us assume that we want to solve a differential problem that exhibits a boundary layer. In this case we could be interested in using grid adaptive procedures. As Figure 2.7 shows, the RB triangulation limits our possibilities to adapt the grid: thus, due to the fact that the FE grid has to be conformal with the RB triangulation, our possibilities to adapt the grid are limited[16]. This is why in certain cases a "free-grid" transformation (i.e., a transformation that does not impose constraints to the grid) could be preferable.

2.3 Non-parametrically affine maps: three different approaches

As the previous examples show, the affine parametrizations work well only for problems with simple domains and pure sizing deformations. For more complex problems (industrial applications, non-Cartesian geometries, etc.) we need to resort to non-affine maps.

2.3.1 Free Form Deformation

The Free Form Deformation (FFD) [112] is a powerful tool for representing smooth global deformations of the reference domain and leading to the reduction of a great number of shape parameters. It was developed for free boundary problems, computer graphics and

[16] On the other hand some regularization techniques could be applied.

for the parametrization and optimal design of aerodynamics surfaces such as wings; more recently, it has been applied to the reduced basis framework.

Here we only define and motivate the map while referring to [69, 5, 63] for numerical simulations and further details.

Due to the complexity of the procedure, we first summarize all the steps.

- We first define a control volume D such that $\Omega \subset D$.
- We introduce the map $\boldsymbol{\psi} : D \to \hat{D}$ between the control volume and the reference configuration $\hat{D} = [0,1]^d$.
- We define the map $\hat{T} : \mathcal{D} \times \hat{D} \to \hat{D}(\boldsymbol{\mu})$. The parameters $\boldsymbol{\mu}$ represent the perturbation of a given set of control points.
- By referring the map to the control volume D, we obtain $\tilde{\boldsymbol{T}} : \mathcal{D} \times D \to D(\boldsymbol{\mu})$, $\tilde{\boldsymbol{T}} = \boldsymbol{\psi}^{-1} \circ \hat{\boldsymbol{T}}$.
- Finally, we define $\boldsymbol{T}$ as the restriction of $\tilde{\boldsymbol{T}}$ to Ω.

Figure 2.8 provides a graphical sketch of the construction of the map[17].

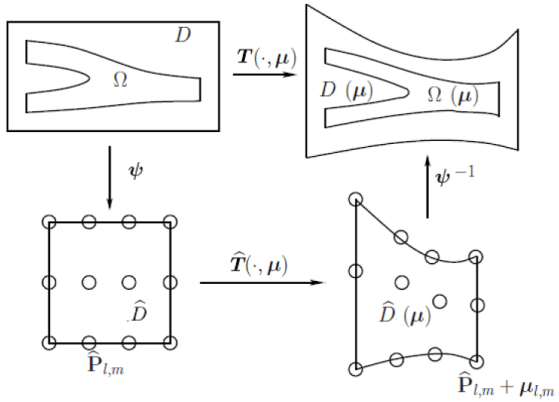


Figure 2.8: construction of the Free Form Deformation map

Let us now explain all the details behind the procedure.

Let $\Omega \subset \mathbb{R}^2$ be the reference domain and let $D = [x_1^{min}, x_1^{max}] \times [x_2^{min}, x_2^{max}]$ be the control volume such that $\Omega \subset D$. Then, the map $\boldsymbol{\psi} : D \to \hat{D} = [0,1]^2$, $\boldsymbol{x} \mapsto \hat{\boldsymbol{x}}$ can be defined as follows:

$$\psi(x_1, x_2) = \begin{bmatrix} \dfrac{x_1 - x_1^{min}}{x_1^{max} - x_1^{min}} \\ \dfrac{x_2 - x_2^{min}}{x_2^{max} - x_2^{min}} \end{bmatrix}$$

Let $K, L \in \mathbb{N}$ be two positive integers, we now select the regular grid of control points in $\hat{D}$:

$$\hat{\boldsymbol{P}}_{k,l} = \begin{bmatrix} \dfrac{k}{K} \\ \dfrac{l}{L} \end{bmatrix} \quad k = 0, \cdots, K \;\; l = 0, \cdots, L$$

[17] Figure 2.8 is taken from [82].

and we define the deformation in $\hat{D}$ through the perturbation of the control points via a set of $(L+1)(K+1)$ parameter vectors $\boldsymbol{\mu}_{l,k}$ ($\hat{\boldsymbol{P}}_{k,l}(\boldsymbol{\mu}_{l,k}) = \hat{\boldsymbol{P}}_{k,l} + \boldsymbol{\mu}_{l,k}$):

$$\hat{\boldsymbol{T}}(\boldsymbol{\mu}, \hat{\boldsymbol{x}}) = \sum_{k=0}^{K} \sum_{l=0}^{L} b_{k,l}^{K,L}(\hat{\boldsymbol{x}}) \left[\hat{\boldsymbol{P}}_{k,l}(\boldsymbol{\mu}_{l,k})\right] \tag{2.3.1}$$

where $b_{k,l}^{K,L}(\hat{\boldsymbol{x}}) = b_k^K(\hat{x}_1) b_l^L(\hat{x}_2)$ are tensor products of the one dimensional Bernstein basis polynomials[18].

Then the map $\tilde{\boldsymbol{T}}$ is defined as follows:

$$\tilde{\boldsymbol{T}}(\boldsymbol{\mu}, \boldsymbol{x}) = \boldsymbol{\psi}^{-1}\left(\sum_{k=0}^{K} \sum_{l=0}^{L} b_{k,l}^{K,L}(\boldsymbol{\psi}(\boldsymbol{x})) \left[\boldsymbol{P}_{k,l}(\boldsymbol{\mu}_{l,k})\right]\right), \tag{2.3.2}$$

and finally,

$$\boldsymbol{T} : \mathcal{D} \times \Omega \to \Omega(\boldsymbol{\mu}) \quad \boldsymbol{T}(\boldsymbol{\mu}) := \tilde{\boldsymbol{T}}(\boldsymbol{\mu})|_{\Omega} \tag{2.3.3}$$

The following Remark explains how to compute the derivatives of the map.

Remark 2.1. *It is possible to prove that the following formula for the derivatives of the polynomial basis functions holds:*

$$\nabla b_{l,k}^{L,K}(\hat{x}_1, \hat{x}_2) = \begin{bmatrix} L[b_{l-1}^{L-1}(\hat{x}_1) - b_l^{L-1}(\hat{x}_1)] b_k^K(\hat{x}_2) \\ L[b_{k-1}^{K-1}(\hat{x}_2) - b_k^{K-1}(\hat{x}_2)] b_l^L(\hat{x}_1) \end{bmatrix}. \tag{2.3.4}$$

As a consequence of (2.3.4) and (2.3.2) we can write the Jacobian of the global map in the following form:

$$\boldsymbol{J}_{\boldsymbol{T}}(\boldsymbol{\mu}, \boldsymbol{x}) = \boldsymbol{J}_{\boldsymbol{\psi}}^{-1}\left[\boldsymbol{I} + \sum_{k=0}^{K} \sum_{l=0}^{L} \nabla b_{k,l}^{K,L}(\boldsymbol{\psi}(\boldsymbol{x})) \boldsymbol{\mu}_{l,k}\right] \boldsymbol{J}_{\boldsymbol{\psi}}. \tag{2.3.5}$$

This explicit expression for the Jacobian of the transformation is extremely important: in this way the map can be easily applied to the RB framework, [78].

We conclude with some observations.

The greatest advantage of using the FFD map is that we have the possibility to fix a certain number of control points or allow some control points to move only in a predetermined direction. As a result, the number of degrees of freedom required to correctly represent the deformation(i.e., the number of parameters) is in practice reasonably low. Furthermore, as Remark 2.1 shows, the Jacobian is computable in an efficient and stable way.

However, despite its great flexibility and ease to use, FFD suffers from some limitations. Due to the fact that the deformations are applied in the square reference domain, they do not have a precise physical meaning; in addition FFD map is not interpolatory and so we do not have a direct control on the boundary of the domain. At last, in order to preserve the partition of unity property of the Bernstein polynomial, the control point grid must be uniformly distributed: due to the fact that the degree of Bernstein polynomials depends on the global lattice, the overall complexity of the map would suffer.

[18] $b_k^K(t) = \binom{K}{k}(1-t)^{K-k}t^k$. The use of Bernstein polynomials is motivated by the partition of unity property

$$\sum_{k=0}^{K} b_k^K(t) \equiv 1$$

and by positivity. Moreover they can be evaluated in a numerically stable way thanks to the de Casteljau algorithm [38]. We point out that the partition of unity property depends on the fact that the control grid is regular. Otherwise it is not in general true.

2.3.2 Methods based on radial basis functions

In several applications[19] the limitations of FFD described above, especially the one related to the description of the boundary, could be not acceptable. This justifies why in [81] a different strategy, based on Radial Basis functions[20], is proposed[21]. Let us briefly introduce the radial basis setting.

In the two dimensional case, let us consider the map $\boldsymbol{\tau} : \mathbb{R}^2 \to \mathbb{R}^2$ defined as:

$$\boldsymbol{\tau}(\boldsymbol{x}) = P(\boldsymbol{x}) + \sum_{i=1}^{k} \sigma(\|\boldsymbol{x} - \boldsymbol{X}_i\|)\boldsymbol{w}_i \tag{2.3.6}$$

where P is a low order polynomial function, $\{\boldsymbol{w}_i\}$ is a set of weights corresponding to the k control points, whose reference positions are $[\boldsymbol{X}_1, \cdots, \boldsymbol{X}_k]$ and $\sigma(\cdot)$ a radially symmetric function (RBF). Standard choices for σ for modelling bidimensional or tridimensional shapes are:

$$\sigma(h) = \begin{cases} exp(\frac{h^2}{\sigma^2}) & \text{Gaussian RBF} \\ (h^2 + \gamma^2)^{\frac{1}{2}} & \text{Multiquadratic RBF} \\ h^\gamma & (\gamma = 1, 3) - \text{power RBF} \\ h^2 \log(h) & \text{thin-plate splines.} \end{cases} \tag{2.3.7}$$

The choice for σ is usually made according to shape regularity and to the properties of the numerical method used to compute the coefficients in (2.3.6). For instance in [81] for the parametric description of carotidal bifurcations, a cubic function $\sigma(h) = h^3$ is chosen.

Concerning the choice of P in (2.3.6), usually a polynomial function of degree 1 is chosen, so that the map $\boldsymbol{\tau}$ can be written as:

$$\boldsymbol{\tau}(\boldsymbol{x}) = \boldsymbol{c} + \boldsymbol{A}\boldsymbol{x} + \mathbb{W}^T s(\boldsymbol{x}), \tag{2.3.8}$$

being $s(\boldsymbol{x}) = (\sigma(\|\boldsymbol{x} - \boldsymbol{X}_1\|), \cdots, \sigma(\|\boldsymbol{x} - \boldsymbol{X}_k\|))^T$ with $\{\boldsymbol{X}_j\}_j$ a given set of control points and $\mathbb{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_k]^T$. The map (2.3.8) has $2k+6$ unknown coefficients in the two dimensional case that are determined by imposing the interpolation constraints:

$$\boldsymbol{\tau}(\boldsymbol{X}_i) = \boldsymbol{Y}_i \tag{2.3.9}$$

and the following additional constraints:

$$\sum_{i=1}^{k} \boldsymbol{w}_i = 0 \quad \sum_{i=1}^{k} \boldsymbol{w}_i X_{i1} = \sum_{i=1}^{k} \boldsymbol{w}_i X_{i2} = 0, \tag{2.3.10}$$

where $\boldsymbol{Y}_i$ is the point in the actual configuration associated with $\boldsymbol{X}_i$ and X_{i1}, X_{i2} are the components of the $\boldsymbol{X}_i$.

---

[19] For instance in imaging problems FFD is completely inadequate because it does not guarantee a sharp representation of the boundaries.

[20] Radial basis functions were originally used in neural networks and successively applied to the context of PDEs. For a general introduction on radial basis functions from a mathematical point of view see [13], for the applications to shape optimization and mesh deformation see [86]. In [20], multivariate RB functions are applied to the reconstruction of implicit surfaces from 3D scattered data.

[21] So far, there are no exhaustive comparative studies between Free Form and RBF-based maps. In [68] the authors apply both FFD and RBF to a shape optimization problem regarding different carotid configurations. The results show that, for that kind of applications the RBF technique, is more versatile and accurate. However, the construction and the computation of the parametrized tensors is much more difficult.

We observe that (2.3.10) can represent the conservation of total force and momentum[22]. Thanks to this condition the polynomial is the affine part (rotation) of the transformation and the term depending on the control points adds the non-affine contribution. In order to fit the RBF technique in our parametrized framework we consider the deformed position of the control points as:

$$\boldsymbol{Y}_i(\boldsymbol{\mu}_i) = \boldsymbol{X}_i + \boldsymbol{\mu}_i \quad i = 1, \cdots k$$

Clearly $\boldsymbol{\mu}_k$ is the displacement of the k-th control point. If we define $\mathbb{S}_{i,j} = s_i(\boldsymbol{X}_j)$, $\mathbb{X}_{i,j} = X_{ij}$ we can rewrite the constraints in a compact form in which the parameter dependent coefficients are only at the right hand side:

$$\begin{bmatrix} \mathbb{S} & \mathbb{I}_k & \mathbb{X} \\ \mathbb{I}_k & 0 & 0 \\ \mathbb{X}^T & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbb{W} \\ \boldsymbol{c}^T \\ \mathbb{A}^T \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}(\boldsymbol{\mu}) \\ 0 \\ 0 \end{bmatrix} \tag{2.3.11}$$

By imposing some not obvious conditions on the radial functions[23], we have that $\mathbb{S}$ is symmetric definite positive so that the matrix in (2.3.11) is invertible. Due to the fact that the matrix is parameter independent the online resolution of the system is extremely efficient.

2.3.3 Transfinite maps

Transfinite maps have been used in [75] in the reduced basis element method framework and recently in [56] for the reduced basis hybrid method. A transfinite map is a not trivial generalization of the Gordon Hall formula, [43]. The main idea behind the approach is to build the image of the interior points of the actual domain as a suitable linear combination of images at the boundary points.

Let Ω be the reference domain and $\Omega(\boldsymbol{\mu})$ be the actual domain. We suppose that they have the same number of curved edges, say n. Furthermore, called Γ_i and $\Gamma_i(\boldsymbol{\mu})$ the i-th edge in the reference and actual configurations, we assume that they are numbered clockwise.

For each Γ_i we define the *weight function* ϕ_i as solution of the following problem ($j \mathrm{mod} n = j - \left[\frac{j}{n}\right] n$):

$$\begin{cases} \Delta\phi_i = 0 & \text{in } \Omega \\ \phi_i = 1 & \text{on } \Gamma_i \\ \frac{\partial\phi_i}{\partial n} = 0 & \text{on } \Gamma_{(i-1)\mathrm{mod}n} \cup \Gamma_{(i+1)\mathrm{mod}n} \\ \phi_i = 0 & \text{on } \Omega \setminus \left(\Gamma_i \cup \Gamma_{(i-1)\mathrm{mod}n} \cup \Gamma_{(i+1)\mathrm{mod}n}\right), \end{cases} \tag{2.3.12}$$

[22] We try to motivate this interpretation. Suppose that $\boldsymbol{w}_i$ is the force applied to the material point $\boldsymbol{X}_i$. Then, the resultant force applied to the system of material points is

$$\boldsymbol{R} = \sum_{i=1}^{k} \boldsymbol{w}_i \overset{(2.3.10)}{\Rightarrow} \boldsymbol{R} = 0$$

Consider now the resultant momentum with respect to the pole $\mathbf{O} = (0,0)$; the resultant force is null then the pole is arbitrary. We have that ($\boldsymbol{r}_i = \boldsymbol{X}_i - \mathbf{O}$, $\mathbf{k}$ is the versor perpendicular to the plane that contains Ω):

$$\boldsymbol{M} = \sum_{i=1}^{k} \boldsymbol{r}_i \times \boldsymbol{w}_i = \sum_{i=1}^{k} (w_{i1} X_{i2} - w_{i2} X_{i1})\, \boldsymbol{k} \overset{(2.3.10)}{\Rightarrow} \boldsymbol{M} = 0$$

[23] See [13] for the details.

and the *projection function* π_i as solution of:

$$\begin{cases} \Delta\pi_i = 0 & \text{in } \Omega \\ \pi_i = t & \text{on } \Gamma_i \\ \pi_i = 0 & \text{on } \Gamma_{(i-1)\bmod n} \\ \pi_i = 1 & \text{on } \Gamma_{(i+1)\bmod n} \\ \frac{\partial \pi_i}{\partial n} = 0 & \text{on } \Omega \setminus \left(\Gamma_i \cup \Gamma_{(i-1)\bmod n} \cup \Gamma_{(i+1)\bmod n}\right) \end{cases} \tag{2.3.13}$$

where t denotes the normalized arc-length. Figure (2.9) - taken from [56] - shows an example of these functions. Now we introduce the edge functions $\boldsymbol{\psi}_i : [0,1] \times \mathcal{D} \to \Gamma_{o,i}$ such

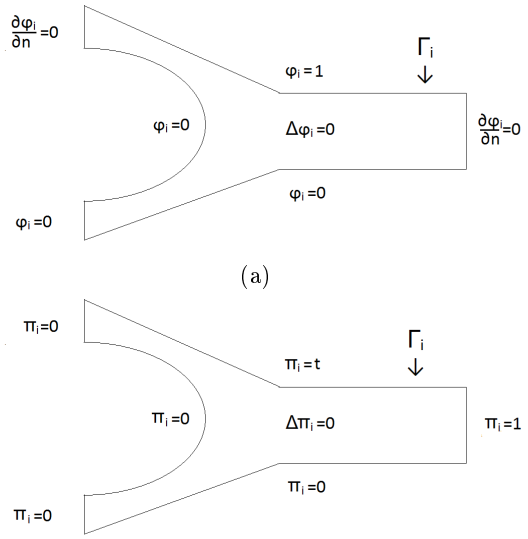


Figure 2.9: an example of (a) weight function ϕ_i (b) projection function π_i.

that $\boldsymbol{\psi}_i(1, \boldsymbol{\mu}) = \boldsymbol{x}_i$, $\boldsymbol{\psi}_i(0, \boldsymbol{\mu}) = \boldsymbol{x}_{(i-1)\bmod n}$, where $\boldsymbol{x}_i$ is the actual vertex shared by Γ_i and $\Gamma_{(i+1)\bmod n}$.

We can define the *transfinite map*[24]

$$\boldsymbol{T}(\boldsymbol{\mu}, \boldsymbol{x}) = \sum_{i=1}^{n} \{\phi_i(\boldsymbol{x})\boldsymbol{\psi}_i(\pi_i(\boldsymbol{x}), \boldsymbol{\mu}) - \phi_i(\boldsymbol{x})\phi_{i+1}(\boldsymbol{x})\boldsymbol{x}_i\} \tag{2.3.14}$$

We observe that the weight functions ϕ_i and the projection functions π_i are parameter independent thus they are computable offline. As a consequence, they are suited to the reduced basis context.

We point out that, unlike FFD and RBF, transfinite maps are not specialized in a particular area of interest but are conceived to be rather general. In addition, they are based on a precise description of the boundaries. On the other hand, the map depends on the solution of a finite element problem so we have to reconstruct the derivatives. This increases the approximation error of the truth approximation with respect to the real problem and in practice a very fine grid is necessary in order to guarantee a reasonable

[24] As stated before the image of the inner point $\hat{\boldsymbol{x}}$ is a linear combination of the boundary points $\pi_i(\hat{\boldsymbol{x}})$ and $\hat{\boldsymbol{x}}_i$.

tolerance[25]. In addition, in order to consider non-simply connected domains, it is necessary to perform a suitable domain decomposition and then to apply the technique to each subdomain. Clearly this increases the computational costs.

2.3.4 Formulation on the reference domain: Offline-Online decomposition

In this subsection we explain how parametrically non-affine maps can be introduced in the RB framework. For simplicity we consider a slightly different problem with respect to the one of section 2.2.3 and we restrict ourselves to a single subdomain and homogeneous Dirichlet boundary conditions.

We set:

$$s(\boldsymbol{\mu}) = f(u(\boldsymbol{\mu}), \boldsymbol{\mu}) \tag{2.3.15}$$

where $u(\boldsymbol{\mu}) \in X(\boldsymbol{\mu}) = H^1(\Omega(\boldsymbol{\mu}))$ satisfies:

$$a(u(\boldsymbol{\mu}), v, \boldsymbol{\mu}) = f(v, \boldsymbol{\mu}) \quad \forall\, v \in X(\boldsymbol{\mu})$$

$a : X(\boldsymbol{\mu}) \times X(\boldsymbol{\mu}) \times \mathcal{D} \to \mathbb{R}$ is of the form:

$$a(u, v, \boldsymbol{\mu}) = \int_{\Omega(\boldsymbol{\mu})} \begin{bmatrix} \dfrac{\partial u}{\partial y_1} & \dfrac{\partial u}{\partial y_2} & u \end{bmatrix} \mathcal{K}_o^{ij}(\boldsymbol{\mu}) \begin{bmatrix} \frac{\partial v}{\partial y_1} \\ \frac{\partial v}{\partial y_2} \\ v \end{bmatrix} d\mathbf{y}. \tag{2.3.16}$$

Here $\mathcal{K}_o : \mathcal{D} \to \mathbb{R}^{2\times 2}$ is:

$$\mathcal{K}_o(\boldsymbol{\mu}) = \begin{bmatrix} \boldsymbol{\nu}_o(\boldsymbol{\mu}) & 0 \\ 0 & a(\boldsymbol{\mu}) \end{bmatrix}.$$

Similarly $f : X(\boldsymbol{\mu}) \times \mathcal{D} \to \mathbb{R}$ is of the form:

$$f(v, \boldsymbol{\mu}) = \int_{\Omega(\boldsymbol{\mu})} f v \, d\mathbf{y}. \tag{2.3.17}$$

With straightforward calculations, we have that:

$$a(u, v, \boldsymbol{\mu}) = \int_\Omega \boldsymbol{\nu}_T(\boldsymbol{\mu}) \nabla u \cdot \nabla v \, d\mathbf{x} + \int_\Omega r_T(\boldsymbol{\mu}) uv \, d\mathbf{x} \quad f(v, \boldsymbol{\mu}) = \int_\Omega f \circ \boldsymbol{T}(\boldsymbol{\mu}) |\mathrm{det}\mathbf{J}_\mathbf{T}(\boldsymbol{\mu})| v \, d\mathbf{x},$$

where $\boldsymbol{\nu}_T(\boldsymbol{x}, \boldsymbol{\mu}) = \boldsymbol{J}_T^{-T} \boldsymbol{\nu}_o(\boldsymbol{\mu}) \boldsymbol{J}_T^{-1} |\mathrm{det}\mathbf{J}_\mathbf{T}(\boldsymbol{\mu})|$ and $r_T = |\mathrm{det}\mathbf{J}_\mathbf{T}(\boldsymbol{\mu})| a(\boldsymbol{\mu})$.

In order to have an efficient offline online computational decomposition, we consider an expansion constructed by the Empirical Interpolation Method[26].

$$\begin{cases} [\boldsymbol{\nu}_T]_{i,j}(\boldsymbol{x}, \boldsymbol{\mu}) = \displaystyle\sum_{m=1}^{\tilde{M}_{i,j}} \Theta_{i,j}^m(\boldsymbol{\mu}) \xi_{i,j}^m(\boldsymbol{x}) + \epsilon_{i,j}(\boldsymbol{x}, \boldsymbol{\mu}) \\ r_T(\boldsymbol{x}, \boldsymbol{\mu}) = \displaystyle\sum_{m=1}^{\tilde{M}_r} \Theta_r^m(\boldsymbol{\mu}) \xi_r^m(\boldsymbol{x}) + \epsilon_r(\boldsymbol{x}, \boldsymbol{\mu}) \\ f \circ \boldsymbol{T}(\boldsymbol{\mu}) |\mathrm{det}\mathbf{J}_\mathbf{T}(\boldsymbol{\mu})| = \displaystyle\sum_{m=1}^{\tilde{M}_f} \Theta_f^m(\boldsymbol{\mu}) {\xi_f}^m(\boldsymbol{x}) + \epsilon_f(\boldsymbol{x}, \boldsymbol{\mu}), \end{cases} \tag{2.3.18}$$

[25] This is especially true if we use low order finite elements.

[26] See section 1.8 for the details concerning this interpolation procedure.

where all the $\boldsymbol{\mu}$-dependent terms are efficiently computable. In conclusion, substituting (2.3.18) into (2.3.16) and (2.3.17) and dropping the error terms, we obtain:

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{m=1}^{\tilde{M}_{i,j}}\Theta_{i,j}^{m}(\boldsymbol{\mu})a_{i,j}^{m}(u(\boldsymbol{\mu}),v)+\sum_{m=1}^{\tilde{M}_{r}}\Theta_{r}^{m}(\boldsymbol{\mu})a_{r}^{m}(u(\boldsymbol{\mu}),v)=\sum_{m=1}^{\tilde{M}_{f}}\Theta_{f}^{m}(\boldsymbol{\mu})f^{m}(v), \qquad (2.3.19)$$

where:

$$\begin{cases} a_{i,j}^{m}(w,v) &= \displaystyle\int_{\Omega}\xi_{i,j}^{m}(\boldsymbol{x})\frac{\partial w}{\partial x_i}\frac{\partial v}{\partial x_j}\,d\mathbf{x} \\ a_{r}^{m}(w,v) &= \displaystyle\int_{\Omega}\xi_{r}^{m}(\boldsymbol{x})wv\,d\mathbf{x} \\ f^{m}(v) &= \displaystyle\int_{\Omega}\xi_{f}^{m}(\boldsymbol{x})v\,d\mathbf{x}. \end{cases}$$

Now we have an efficient offline-online procedure for the matrix assembly[27]: if $\epsilon_{i,j}$, ϵ_r and ϵ_f are below a given tolerance

$$\|\epsilon_{i,j}(\cdot,\boldsymbol{\mu})\|_{L^\infty(\Omega)},\|\epsilon_r(\cdot,\boldsymbol{\mu})\|_{L^\infty(\Omega)}\|\epsilon_f(\cdot,\boldsymbol{\mu})\|_{L^\infty(\Omega)}\leq\epsilon_{tol}^{EIM},$$

with ϵ_{tol}^{EIM} is a given tolerance, we can apply the RB methodology discussed in the first chapter directly to the (2.3.19) without significantly affecting the overall error[28].

2.4 Transfinite maps for small deformations

In this section we propose a new approach obtained by simplifying the above mentioned transfinite map under the hypothesis of small deformations[29]. First we introduce some definitions, then we explain the method and finally we provide some numerical results. As in the previous sections, we restrict ourselves to the bidimensional case[30]. Before starting, Example 2.2 motivates the approach.

Example 2.2. *Let us consider an advection-diffusion problem around an airfoil profile for different angles of incidence of the wing, for instance we consider the symmetric configuration presented in section 2.2. Let the problem be advection-dominated.*

In this case it is clear that:

- *we require a precise description of the wing profile;*
- *in order to accurately capture the boundary layer around the airfoil, we need to properly adapt the grid.*

[27] In practice $\tilde{M}_{i,j}$, $\tilde{M}_r$ and $\tilde{M}_f$ are quite modest if the number of parameters P is small.

[28] It is well-known that, named $u^{EIM}(\boldsymbol{\mu})$ the solution to a parametrically affine problem and $u(\mu)$ the solution to the original problem in the reference configuration, we have (following the same strategy as in Example 1.4):

$$\|u(\boldsymbol{\mu})-u^{EIM}(\boldsymbol{\mu})\|_{H^1}\leq C(\|\epsilon_{i,j}\|_{L^\infty(\Omega)}+\|\epsilon_r\|_{L^\infty(\Omega)}+\|\epsilon_f\|_{L^\infty(\Omega)}).$$

As explained in the first chapter, the feasibility of this requirement is strictly connected to the coercivity (inf-sup) constant: for advection dominated problems this requirement is extremely hard to be assessed.

[29] The approach is somehow similar to the one proposed in [15, 116] and could be seen as a generalization of it. We will quantify theoretically in Lemma 2.1 and then numerically in the first example the entity of the deformation.

[30] However the extension to the tridimensional case is, at least from the theoretical point of view, not particularly involved.

The method based on the curvy triangles domain decomposition imposes several restrictions on the underlined Finite Element mesh. Thus performing a suitable adaptive procedure could easily become an unaffordable task.

On the other hand, if we require an absolutely precise description of the profile, FFD and also RBF-based methods[31] *seem to be less attractive for this kind of application than transfinite mapping. However, for such an easy deformation we aim at simplifying the latter method to mitigate the drawbacks cited above.*

2.4.1 Theoretical framework and first properties

Let $\boldsymbol{g} : \mathcal{D} \times \partial\Omega \to \partial\Omega(\boldsymbol{\mu})$ be a given map that describes the deformation of the boundary of the parametrized domain such that $\partial\Omega = \partial\Omega(\boldsymbol{\mu}_{ref})$ for some $\boldsymbol{\mu}_{ref} \in \mathcal{D}$.

Then, we define $\boldsymbol{T} : \mathcal{D} \times \Omega \to \Omega(\boldsymbol{\mu})$ as follows:

$$\begin{cases} -\Delta \boldsymbol{T}(\boldsymbol{\mu}) = 0 & \text{in } \Omega \\ \boldsymbol{T}(\boldsymbol{\mu}) = \boldsymbol{g}(\boldsymbol{\mu}) & \text{on } \partial\Omega \end{cases} \tag{2.4.1}$$

In the next subsection we discuss some hypotheses to guarantee that the vector-valued function $\boldsymbol{T}$ is a change of coordinates for each value of the parameter.

It is easy to observe that due to the linearity of the equation, if $\boldsymbol{g}(\boldsymbol{\mu}, \boldsymbol{x}) = \sum_{k=1}^{Q} \Theta_k^g(\boldsymbol{\mu}) \boldsymbol{h}_k(\boldsymbol{x})$, then $\boldsymbol{T}(\boldsymbol{\mu}, \boldsymbol{x}) = \sum_{k=1}^{Q} \Theta_k^g(\boldsymbol{\mu}) \boldsymbol{T}_k(\boldsymbol{x})$ where $\boldsymbol{T}_k$ is the solution to (2.4.1) corresponding to the Dirichlet data $\boldsymbol{g}_k$.

In addition, thanks to the hypothesis on $\boldsymbol{g}$ we have that $\mathbf{T}(\boldsymbol{\mu}_{ref}, \mathbf{x}) = \mathbf{x}$.

In order to highlight the possibility to perform an offline-online decomposition in the case of parametrically affine change of coordinates, we recall a classical theorem and we make some considerations.

Theorem 2.1. ***(Cayley-Hamilton)*** *Let $A \in \mathbb{R}^{2\times 2}$. The following equation holds:*

$$A^2 - tr(A)A + detA\mathbb{I} = 0. \tag{2.4.2}$$

From this result we can deduce that:

- If $A(\boldsymbol{\mu}, \boldsymbol{x}) = \sum_{k=1}^{Q} \Theta_k(\boldsymbol{\mu}) A_k(\boldsymbol{x})$, then:

$$\det A(\boldsymbol{\mu}, \boldsymbol{x}) = \sum_{\substack{1\le i,j\le Q,\\ 0\le k,l\le 2}} \left((\Theta_i(\boldsymbol{\mu}))^k \, (\Theta_j(\boldsymbol{\mu}))^l \right) a_{i,j,k,l}(\boldsymbol{x})$$

  for some parameter independent functions $\{a_{i,j,k,l}(\cdot)\}$.

- $A^{-1} = -\frac{1}{\det A}(A - tr(A)\mathbb{I})$.

Remark 2.2. *For the offline-online decomposition presented in section 2.3.4, it is important to find an affine approximation of $\boldsymbol{\nu}_T(\boldsymbol{x}, \boldsymbol{\mu}) = \boldsymbol{J}_T^{-T}(\boldsymbol{x}, \boldsymbol{\mu}) \boldsymbol{\nu}_o(\boldsymbol{\mu}) \boldsymbol{J}_T^{-1}(\boldsymbol{x}, \boldsymbol{\mu}) |\det \boldsymbol{J}_{\boldsymbol{T}}(\boldsymbol{\mu}, \boldsymbol{x})|$.*

Thanks to the Cayley-Hamilton theorem, if $\boldsymbol{\nu}_o = \nu\mathbb{I}$ the following formula holds:

$$\boldsymbol{\nu}_T = \nu \boldsymbol{J_T}^{-T} \boldsymbol{J_T}^{-1} \det \boldsymbol{J_T} = \frac{\nu}{det \boldsymbol{J_T}} \left(tr \boldsymbol{J_T} \mathbb{I} - \boldsymbol{J}_T^T \right) \left(tr \boldsymbol{J}_T \mathbb{I} - \boldsymbol{J}_T \right) \tag{2.4.3}$$

For this reason, after computing $\boldsymbol{\nu}_T$ we can proceed by applying the EIM to the non-affine term $\frac{\nu}{det \boldsymbol{J}_T}$ or separately to each term of the matrix[32]*.*

[31] Even if the second method is interpolatory, a high number of control points on the wing causes an inefficient RB formulation and also in this way it is not possible to guarantee the perfect description of the deformation of the boundary. In aeronautics FFD has been used to describe airfoil wings in situations in which there are no strict requirements on the representation of the airfoil.

[32] When expansion is long, the application of the Empirical Interpolation Method to each term is surely

2.4.2 Useful properties of transfinite transformations

In order to motivate the proposed approach, we have to deal with some mathematical properties of the transfinite map[33]. There are several counterexamples that show that in general:

- it is not true that $\boldsymbol{T}(\boldsymbol{\mu}, \Omega) = \Omega(\boldsymbol{\mu})$;
- the map does not have the one-to-one property.

However, for small variations of the parameters, Laplace based transformations can be used and it is possible to prove the following quantitative result.

Lemma 2.1. *Suppose that Ω is a domain of class C^h and $\boldsymbol{g} \in H^{h-\frac{1}{2}}(\partial\Omega)$ with h large enough to have $\boldsymbol{T} \in C^1(\bar{\Omega})$[34]. Furthermore let us assume that $\boldsymbol{g}(\boldsymbol{\mu}, \boldsymbol{x}) = \sum_{k=1}^{Q} \Theta_k^g(\boldsymbol{\mu}) \boldsymbol{h}_k(\boldsymbol{x})$ with $\Theta_k^g(\boldsymbol{\mu}_{ref}) \neq 0$. Then:*

- $$\begin{aligned} \left|\det \boldsymbol{J}_{\boldsymbol{T}}(\boldsymbol{\mu}) - 1\right| \leq & \sum_{\substack{1 \leq i,j \leq Q, \\ 0 \leq k,l \leq 2}} \left| \left(\Theta_i(\boldsymbol{\mu})\right)^k \left(\Theta_j(\boldsymbol{\mu})\right)^l - \left(\Theta_i(\boldsymbol{\mu}_{ref})\right)^k \left(\Theta_j(\boldsymbol{\mu}_{ref})\right)^l \right| \max_{\mathbf{x} \in \Omega} |a_{i,j,k,l}(\mathbf{x})| \\ = & C(\boldsymbol{\mu}). \end{aligned} \tag{2.4.4}$$
- *Moreover, if $C(\boldsymbol{\mu}) < 1$ then $\boldsymbol{T}(\boldsymbol{\mu}) : \Omega \to \Omega(\boldsymbol{\mu})$ is bijective.*

Proof. Due to the fact that $\boldsymbol{g}(\boldsymbol{\mu}, \boldsymbol{x}) = \sum_{k=1}^{Q} \Theta_k^g(\boldsymbol{\mu}) \boldsymbol{h}_k(\boldsymbol{x})$, we have that:

$$det \boldsymbol{J}_{\boldsymbol{T}}(\boldsymbol{\mu}, \boldsymbol{x}) = \sum_{\substack{1 \leq i,j \leq Q, \\ 0 \leq k,l \leq 2}} \left(\left(\Theta_i(\boldsymbol{\mu})\right)^k \left(\Theta_j(\boldsymbol{\mu})\right)^l \right) a_{i,j,k,l}(\boldsymbol{x})$$

As already stated above, $\boldsymbol{T}(\boldsymbol{\mu}_{ref}, \boldsymbol{x}) = \boldsymbol{x}$ and so $\det \mathbf{J}_{\mathbf{T}}(\boldsymbol{\mu}_{ref}, \boldsymbol{x}) \equiv 1$. Thus we can easily conclude that:

$$\begin{aligned} & \left|\det \mathbf{J}_{\mathbf{T}}(\boldsymbol{\mu}, \boldsymbol{x}) - \det \mathbf{J}_{\mathbf{T}}(\boldsymbol{\mu}_{ref}, \boldsymbol{x})\right| = \left|\det \mathbf{J}_{\mathbf{T}}(\boldsymbol{\mu}, \boldsymbol{x}) - 1\right| \\ & = \sum_{\substack{1 \leq i,j \leq Q, \\ 0 \leq k,l \leq 2}} \left| \left(\left(\Theta_i(\boldsymbol{\mu})\right)^k \left(\Theta_j(\boldsymbol{\mu})\right)^l - \left(\Theta_i(\boldsymbol{\mu}_{ref})\right)^k \left(\Theta_j(\boldsymbol{\mu}_{ref})\right)^l \right) a_{i,j,k,l}(\mathbf{x}) \right| \\ & \leq \sum_{\substack{1 \leq i,j \leq Q, \\ 0 \leq k,l \leq 2}} \left| \left(\Theta_i(\boldsymbol{\mu})\right)^k \left(\Theta_j(\boldsymbol{\mu})\right)^l - \left(\Theta_i(\boldsymbol{\mu}_{ref})\right)^k \left(\Theta_j(\boldsymbol{\mu}_{ref})\right)^l \right| \max_{\mathbf{x} \in \Omega} |a_{i,j,k,l}(\mathbf{x})| \end{aligned}$$

This concludes the first part of the proof.

---

more advantageous; but if the number of terms is short , say 1 or 2, the application of the EIM to the inverse of the determinant $\frac{\nu}{\det \boldsymbol{J}_T}$ can be preferable. We point out that if we apply the EIM only to the inverse of the determinant, the approximation maintains the same eigenvectors of the original deformation. However if we use Q_{eim} terms for each empirical interpolation, we have an expansion of $\frac{Q(Q+1)}{2} Q_{eim}$ instead of $3Q_{eim}$. In our implementation the EIM will be applied to each term. In [115] the so-called MCEIM has been proposed to apply the empirical interpolation directly to tensors.

[33] Historically the transformations based on the Laplace equation have certainly been the most studied and discussed in the hambit of numerical grid generation. Several generalizations of the simple idea we deal with in this section have been proposed and transformation (2.3.12)-(2.3.13) represents one of these generalizations. For a comprehensive presentation of the theoretical aspects we refer to [16].

[34] In two and three dimensions this means $h \geq 3$.

We are going to prove that $\boldsymbol{T}(\boldsymbol{\mu}) : \Omega \to \Omega(\boldsymbol{\mu})$ is a change of coordinates if $C(\boldsymbol{\mu}) < 1$.

Let $p \in \partial\Omega$, then $p(\boldsymbol{\mu}) := \boldsymbol{T}(p) \in \partial\Omega(\mu)$ by construction.

By contradiction, $p \in \mathring{\Omega}$, $p(\boldsymbol{\mu}) \notin \mathring{\Omega}(\boldsymbol{\mu})$; therefore due to the continuity[35] of $\boldsymbol{T}$ in $\boldsymbol{\mu}$, there exists $\bar{\boldsymbol{\mu}} \in \mathcal{D}$ such that $p(\bar{\boldsymbol{\mu}}) \in \partial\Omega(\bar{\boldsymbol{\mu}})$. Thus $p(\bar{\boldsymbol{\mu}}) = \tilde{p}(\bar{\boldsymbol{\mu}})$ where $\tilde{p} \in \partial\Omega$.

Consider the scalar function $g(t) = (\tilde{p}-p)^T \boldsymbol{T}(\bar{\boldsymbol{\mu}}, p + t(\tilde{p}-p))$. Clearly $g(0) = g(1)$, thus thanks to Rolle theorem $g'(\bar{t}) = 0$ for some $\bar{t} \in (0,1)$.

However, $g'(t) = (\tilde{p}-p)^T \boldsymbol{J}_{\boldsymbol{T}}(\bar{\boldsymbol{\mu}}, p + t(\tilde{p}-p))(\tilde{p}-p)$ that implies that[36] $\det \boldsymbol{J}_{\boldsymbol{T}}(\bar{\boldsymbol{\mu}}, \boldsymbol{x}) = 0$ for some $\boldsymbol{x} \in \Omega$. This is a contradiction.

Until now we proved that, for all $\boldsymbol{\mu} \in \mathcal{D}$ such that $C(\boldsymbol{\mu}) < 1$, $\boldsymbol{T}(\boldsymbol{\mu}, \Omega) \subset \Omega(\boldsymbol{\mu})$ and more specifically that $\boldsymbol{T}(\boldsymbol{\mu}, \partial\Omega) = \partial\boldsymbol{T}(\boldsymbol{\mu}, \Omega) = \partial\Omega(\boldsymbol{\mu})$. But for a generalization of the *Intermediate Value Theorem*[37] we can conclude that $\boldsymbol{T}(\boldsymbol{\mu}, \cdot)$ is surjective in $\Omega(\boldsymbol{\mu})$.

For the injectivity we can proceed in the same way. □

We highlight that the estimate presented above is extremely pessimistic; however, in an offline-online context, we can check *a posteriori* whether the map is a change of coordinate and we can quantify the entity of the admissible perturbation offline in a pre-process stage.

Lemma 2.1 can be assumed just as a proof-of-concept[38]. However, through a simple modification it is possible to extend the method to a wider family of domains.

Let us start from an example: the rotation of a NACA symmetric profile.

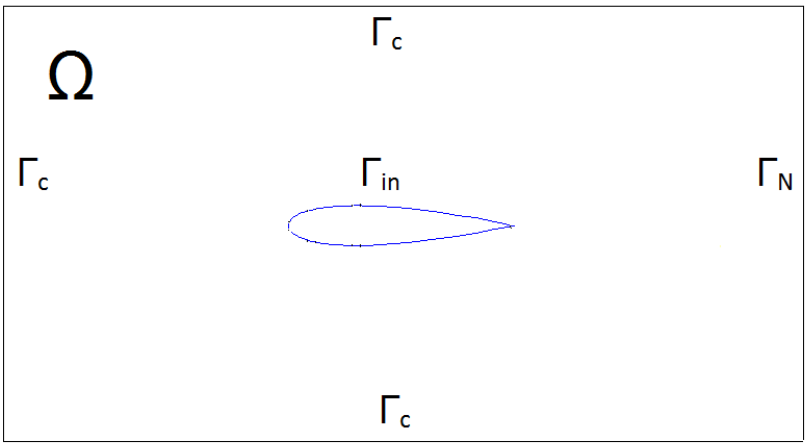


Figure 2.10: NACA airfoil. Γ_{in} is the wing boundary, whereas Γ_{out} is the square boundary

In this case the deformation is represented by:

$$\boldsymbol{T}(\mu, \boldsymbol{x}) = \boldsymbol{T}_1(\boldsymbol{x}) + cos(\mu)\boldsymbol{T}_2(\boldsymbol{x}) + sin(\mu)\boldsymbol{T}_3(\boldsymbol{x}), \tag{2.4.5a}$$

[35] In this work we do not prove this statement, we just observe that if $\Omega \in C^m$, $\boldsymbol{g}_i \in H^{m-\frac{1}{2}}$ then $\boldsymbol{T} : \mathcal{D} \to H^m(\Omega)$ is continuous (see [109], chapter 8).

[36] This statement is straightforward because for $\boldsymbol{\mu} = \boldsymbol{\mu}_{ref}$ all the eigenvalues are positive, so they remain positive for small perturbation; thus the Jacobian is definite positive.

[37] Let $f : X \to Y$, with X Y topological spaces and f a continuous function, then, if X is connect, also Y is connect.

[38] We would argue that even if from a mathematical point of view the approach is not satisfactory it could work from a numerical viewpoint. Some numerical simulations in the following section show that also in practice, when we try to refine the mesh near the corners in order to increase the accuracy, we have problems.

where $\boldsymbol{T}_1$, $\boldsymbol{T}_2$ and $\boldsymbol{T}_3$ are the solutions to:

$$\begin{cases} \Delta \boldsymbol{T}_1 = 0 & \text{in } \Omega \\ \boldsymbol{T}_1 = \boldsymbol{x} & \text{on } \Gamma_{out} \\ \boldsymbol{T}_1 = 0 & \text{on } \Gamma_{in}, \end{cases} \tag{2.4.5b}$$

$$\begin{cases} \Delta \boldsymbol{T}_2 = 0 & \text{in } \Omega \\ \boldsymbol{T}_2 = 0 & \text{on } \Gamma_{out} \\ \boldsymbol{T}_2 = \boldsymbol{x} & \text{on } \Gamma_{in}, \end{cases} \tag{2.4.5c}$$

$$\begin{cases} \Delta \boldsymbol{T}_3 = 0 & \text{in } \Omega \\ \boldsymbol{T}_3 = 0 & \text{on } \Gamma_{out} \\ \boldsymbol{T}_3 = -y\boldsymbol{i} + x\boldsymbol{j} & \text{on } \Gamma_{in}, \end{cases} \tag{2.4.5d}$$

respectively.

In this case the square vertices do not represent a problem because we are interested in studying the equation around the profile. However, the discontinuity of the derivative at the tail of the edge affects the accuracy of the solution: there the derivative of the map is not bounded, thus we cannot refine adequately the grid close to the point.

The idea we propose is the following[39]: we consider two concentric ellipses, say E_{inner} and E_{outer}, around the airfoil; then we extend the inner boundary condition of each problem to $\Omega \cap E_{inner}$ and the outer boundary condition to $\Omega \setminus E_{outer}$. Figure 2.11 shows the decomposition of the domain:

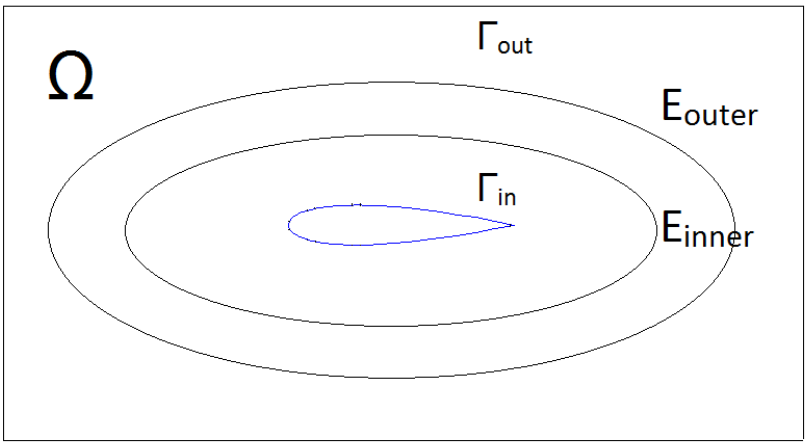


Figure 2.11: decomposition of the domain

Now let us define:

$$\boldsymbol{T}_i^\star(\boldsymbol{x}) = \begin{cases} \boldsymbol{u}_i^{inner} & \text{in } \Omega \cap E_{inner} \\ \tilde{\boldsymbol{T}}_i & \text{in } E_{outer} \setminus E_{inner} \\ \boldsymbol{u}_i^{outer} & \text{in } \Omega \setminus E_{outer} \end{cases}, \tag{2.4.6}$$

where $\tilde{\boldsymbol{T}}_i$ is the harmonic function such that $\boldsymbol{T}_i^\star$ is globally continuous[40] and $\boldsymbol{u}_i^{inner}$, $\boldsymbol{u}_i^{outer}$ are suitable extensions of the boundary data. It is absolutely trivial to prove the following.

[39] With respect to the approach proposed in [116], the main difference is that in our case we do not provide the map at hand but we use the transfinite map. This guarantees the possibility to consider more complex structures: for instance with this approach we can consider a fixed structure with a rotating part.

[40] In the case of elliptical disc we can find an analytical solution. We do not take advantage of it because for real applications we do not have the possibility to use elliptical discs.

Lemma 2.2. *The transformation* $\boldsymbol{T}^\star(\mu, \boldsymbol{x}) = \boldsymbol{T}_1^\star(\boldsymbol{x}) + cos(\mu)\boldsymbol{T}_2^\star(\boldsymbol{x}) + sin(\mu)\boldsymbol{T}_3^\star(\boldsymbol{x})$ *is globally* $K(\mu)$*-Lipschitz. Moreover the* $K(\mu)$ *constant is continuous with respect to the parameter* μ*.*

Remark 2.3. *We observe that the piecewise approach presented above is extremely indicated when the problem at hand presents a boundary layer for two reasons:*

- *with respect to the affine parametrizations described in section 2.2, it is possible to perform some grid adaptivity in order to tailor the mesh to our problem.*
- *with respect to the non-affine parametrizations described in section 2.3, we observe that the boundary is now represented without any approximation: this is extremely important because the reconstruction of the derivative and the application of the EIM -needed in order to make the problem affine with respect to the parameter- is now limited to an area where the solution is smooth. For this reason we can reasonably expect that the distance* $\|u(\boldsymbol{\mu}) - u_{EIM}(\boldsymbol{\mu})\|_{H^1(\Omega)}$ *- where* $u(\boldsymbol{\mu})$ *and* $u_{EIM}(\boldsymbol{\mu})$ *are the truth solutions referred to the real and to the approximate domains, respectively - can be under our desired tolerance through a reasonably small parametrically affine expansion.*

2.4.3 The approximation of the derivatives

Before concluding this chapter with the presentation of some numerical simulations, we briefly address the following issue: due to the fact that the transformation is obtained through the solution of a differential problem, is it possible to rely on its derivative?

This aspect is crucial in order to guarantee that our truth approximation be close to the real one.

It is well known that Finite Elements guarantee good approximation properties in $H^1(\Omega)$, however it is much more difficult to guarantee a good approximation in terms of $W_1^\infty(\Omega)$.

In [26] under some hypotheses on the mesh it is proved that the finite element solution to a second order elliptic boundary value problem satisfies the following best approximation property:

$$\|\nabla(u - u_h)\|_{L^\infty(\Omega)} \leq C \min_{\chi \in X_h} \|\nabla(u - \chi)\|_{L^\infty(\Omega)} \tag{2.4.7}$$

where X_h is the FE space. Another important topic is how to compute the finite element derivative involved in (2.4.3). For each element it is possible to obtain the correct values through the following procedure here presented for P_1-elements[41]

Let $\hat{K}$ be the reference triangle, identified by the vertices $(0,1)$, $(0,0)$, $(1,0)$. Given the element $K \in \mathcal{T}_h$, let (B_K, c_K) be the unique matrix and vector such that $K = B_K\hat{K} + c_K$ and such that the node $\boldsymbol{x}_{i_1}$ is mapped into $(0,0)$, $\boldsymbol{x}_{i_2}$ into $(1,0)$ and $\boldsymbol{x}_{i_3}$ into $(0,1)$. Thus the derivative of u_h in the element K can be computed as:

$$\nabla u_h\big|_K = B_K^{-T}\left(u_{i_1}\nabla\phi_{(0,0)} + u_{i_2}\nabla\phi_{(1,0)} + u_{i_3}\nabla\phi_{(0,1)}\right).$$

[41] This is algorithm implemented in the Matlab tool `pdegrad`. Clearly more general recovery methods are available, see for instance [127, 126].

2.5 Numerical simulations

After introducing the theory, we are going to provide some examples.

We first address the problem of quantifying the hypothesis of small deformations. For two different geometries, we show that the transfinite maps are able to represent the deformation of the domain. We also discuss the convergence of the EIM algorithm in presence of small deformations of the domain.

Then, we consider the application of the piecewise transfinite map to an advection-diffusion problem.

2.5.1 Hypothesis of small deformations

Let us consider the following problem:

$$\begin{cases} -\Delta u = f & \text{in } \Omega(\boldsymbol{\mu}) \\ u = g & \text{on } \partial\Omega_D(\boldsymbol{\mu}). \end{cases} \tag{2.5.1}$$

In order to evaluate the entity of the small deformation hypothesis[42], we introduce the following indicator

$$\xi(\boldsymbol{\mu}) = \frac{\lambda_{max}}{\lambda_{min}} \text{ where } \lambda_{max}, \lambda_{min} \text{ are the eigenvalues of } \boldsymbol{\nu}_T(\boldsymbol{\mu}) = \boldsymbol{J_T}^{-T}\boldsymbol{J_T}^{-1}. \tag{2.5.2}$$

Let us briefly motivate our choice: if we refer the problem to the reference configuration using the map $\mathbf{T}$, it is straightforward to prove that λ_{max} is equal to the continuity constant associated with the bilinear form of the problem while λ_{min} coincides with the coercivity constant. Thus, if we suppose to solve problem (2.5.1) thanks to the Céa Lemma (see [97]), we have that

$$\|u(\boldsymbol{\mu}) - u_h(\boldsymbol{\mu})\|_X \leq \sqrt{\xi(\boldsymbol{\mu})} \inf_{w \in X_h} \|u(\boldsymbol{\mu}) - w\|_X. \tag{2.5.3}$$

Therefore, by solving the problem in the reference configuration, our FE approximation is deteriorated by a factor $\sqrt{\xi(\boldsymbol{\mu})}$.

The indicator $\xi(\boldsymbol{\mu})$ depends on the differential problem and can be hard to compute in more complex situations. For this reason, another indicator, directly linked to the approximation properties of the FE space, can be proposed:

$$\tilde{\xi}(\boldsymbol{\mu}) = \left\| \frac{|\lambda_{max,T}(\boldsymbol{\mu}, \cdot)|}{|\lambda_{min,T}(\boldsymbol{\mu}, \cdot)|} \right\|_{L^\infty(\Omega)} \tag{2.5.4}$$

where $\lambda_{max,T}$ and $\lambda_{min,T}$ are the maximum and minimum (in modulus) eigenvalue associated with $\mathbf{J_T}$, respectively.

The interpretation of this second indicator is the following one: let us consider a *regular* triangulation $\mathcal{T}_h$ in the reference domain Ω such that:

$$\max_{K \in \mathcal{T}_h} \frac{h_K}{\rho_K} \leq \sigma, \quad \forall\, h > 0$$

where ρ_K and h_K are the radius of the circle inscribed and circumscribed to the element K, respectively.

[42] With respect to our knowledge, this kind of analysis is new; also the two indicators are original.

Then the mapped problem[43] is equivalent to the FE solution of the problem in the actual configuration with respect to a still regular triangulation $\mathcal{T}_h(\mu)$ such that

$$\max_{K(\mu)\in\mathcal{T}_h(\mu)} \frac{h_{K(\mu)}}{\rho_{K(\mu)}} \leq \sigma\tilde{\xi}(\mu) \quad \forall\, h > 0.$$

Therefore, as anticipated, this second indicator is strictly related to the approximation property of the FE space (see Theorem 3.4.2 in [99]).

2.5.2 NACA profile

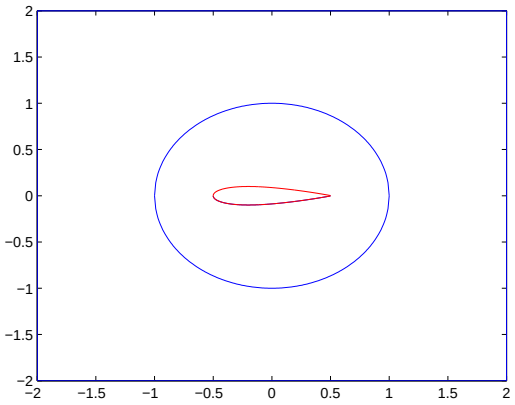


Figure 2.12: symmetric NACA profile

As first example, we consider the rotation of a NACA symmetric profile with respect to $(0,0)$. The parameter μ represents the incidence angle with respect to the horizontal axis. The length of the profile is set to one and the domain $\Omega = (-2,2)^2$. We consider the piecewise approach in which the profile is enclosed by a circle centered in $(0,0)$ with unitary radius. Figure 2.12 shows the geometry[44] for $\mu = 0$.

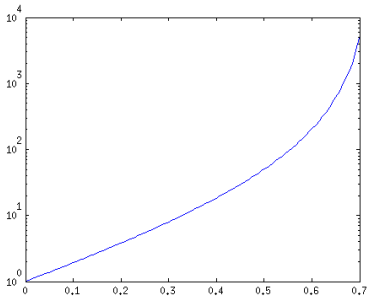


Figure 2.13: ratio $\xi(\mu)$.

[43] We assume that the approximation associated with the application of the EIM is negligible.
[44] The number of points in the outer region is 161.

it.	μ	Max Err
1	0	39.84
2	0.7	3.21
3	0.658	1.44
4	0.467	0.391
5	0.686	0.109

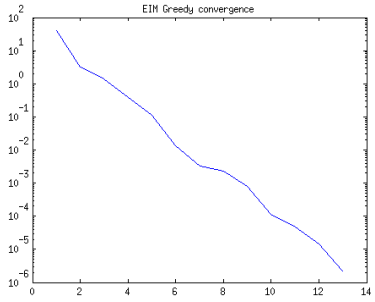


Figure 2.14: in the table we gather the μ selected at each iteration and the correspondent maximum error on $\Xi \subset [0, 0.7]$ where $|\Xi| = 100$. The angles are in radians, $0.7rad = 40.11°$. On the right we show the convergence of the EIM Greedy algorithm.

$\mathcal{D}$	EIM expansion
$\mu \in [0, 0.3]$	7
$\mu \in [0, 0.4]$	8
$\mu \in [0, 0.5]$	10
$\mu \in [0, 0.6]$	11
$\mu \in [0, 0.7]$	14
$\mu \in [0, 0.8]$	22

Table 2.1: number of terms in the EIM expansion vs size of the domain. $|\Xi| = 1000$, the tolerance is set to 10^{-5}. In the last case $\|\xi\|_{L^\infty(\Xi)} > 10^4$.

Figure 2.13 shows that for $\mu < 0.4$ (equal to 22.93°) $\xi(\mu) < 10$. Otherwise, for $\mu > 0.4$ the ratio diverges rapidly. Starting from this observation, we may state that the map can be successfully used in this framework for $\mathcal{D} \subset [0, 0.4]$.

Figure 2.14, right, analyses the performances of the EIM for the first component of the viscosity matrix $\nu_{1,1}(\mu)$, for $\mathcal{D} = [0, 0.7]$. Even if the non-affine function is extremely badly scaled, the Greedy algorithm converges very fast; as it may be expected, the first parameters chosen by the Greedy strategy are closed to $\mu = 0.7$ (see Figure 2.14, left).

Finally Table 2.1 analyses the growth of the number of terms in the EIM expansion necessary to satisfy a given tolerance. For $\mu_{max} < 0.8$ the number grows linearly: this provides further evidence about the robustness of the EIM.

2.5.3 NACA profile with support

We turn to a more involved example. The NACA profile is now fastened to a triangular fixed support. As before the length of the NACA profile is set to one and the parameter is the rotation angle μ (the rotation center is still $(0, 0)$). Figure 2.15 shows the geometry[45] for $\mu = 0$.

As we may expect, the increasing complexity in the geometry deteriorates the performances of the method. As plotted in Figure 2.16 the indicator $\xi(\mu)$ is less than 10 for $\mu < 0.15$ (=8.59°). Nevertheless, the results remain extremely positive: the maximum reachable angle is close to the critical angle of attack[46]. Furthermore, we reasonably expect that by changing the artificial inner domain that contains the profile, the performance of the map can be improved.

[45] The size of the mesh associated with the outer region is 3331.

[46] The critical angle of attack depends on the airfoil, it is in general close to $10 - 15°$ degrees [3] .

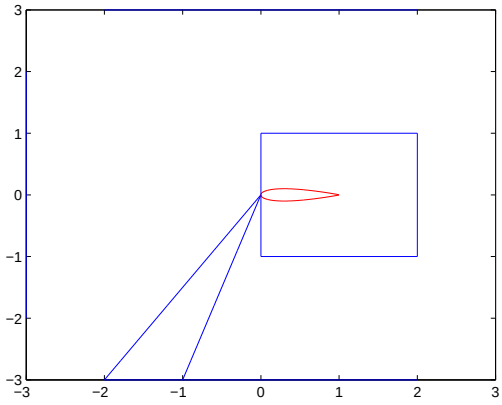


Figure 2.15: symmetric NACA profile with support

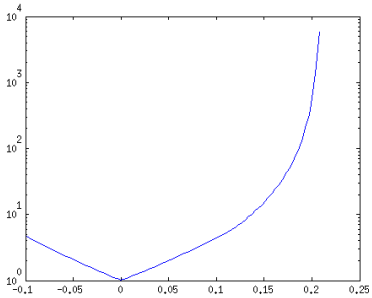


Figure 2.16: ratio $\xi(\mu)$.

Figure 2.17 shows the performances of the Greedy algorithm for the first component of the viscosity matrix $\nu_{1,1}(\mu)$, for $\mathcal{D} = [-0.1, 0.25]$. As in the previous case, the convergence is exponentially fast. On the other hand, we do not observe a polarization of the sampled values at one of the two extreme points.

μ	Max Err
-0.1	64.68
0.21	6.79
0.25	1.025
0.15	0.2423
-0.008	0.048

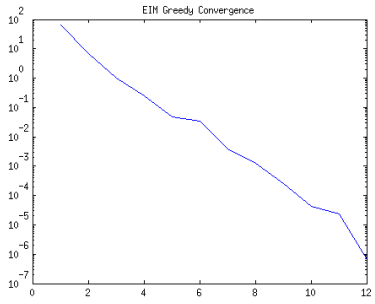


Figure 2.17: in the table we gather the μ selected at each iteration and the correspondent maximum error on $\Xi \subset [-0.1, 0.25]$ where $|\Xi| = 100$. The angles are in radians, $0.25rad = 14.32°$. On the right we show the convergence of the EIM Greedy algorithm.

Finally, as we made in Table 2.1, in Table 2.2 we analyse the growth of the number of terms in the EIM expansion with respect to the size of the domain. For $\mu_{max} \leq 0.3$ and $\mu_{min} \geq -0.1$ we observe a substantially linear dependence.

$\mathcal{D}$	EIM expansion
$\mu \in [-0.1, 0.1]$	8
$\mu \in [-0.1, 0.2]$	10
$\mu \in [-0.1, 0.3]$	13
$\mu \in [-0.1, 0.4]$	20

Table 2.2: number of terms in the EIM expansion vs size of the Domain. $|\Xi| = 1000$,the tolerance is set to 10^{-5}. In the last case $\|\xi\|_{L^\infty(\Xi)} > 10^4$.

2.5.4 An advection-diffusion problem around a rotating symmetric NACA profile

As final example we test our piecewise map on an advection-diffusion problem.

The goal of the test is to compare the FE numerical solution computed in the reference configuration with the FE numerical solution directly computed in the actual configuration.

Figure 2.18 shows the reference domain.

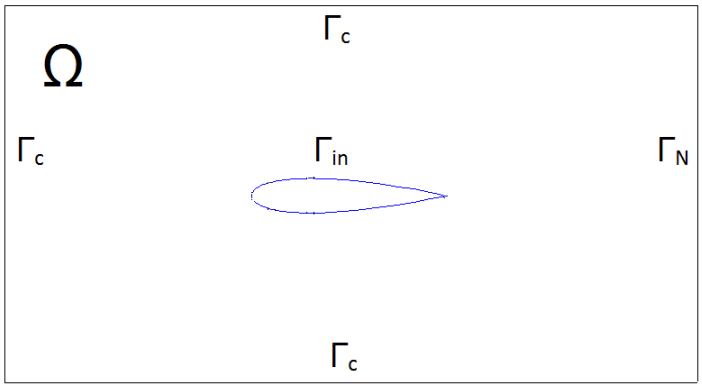


Figure 2.18: computational domain: reference configuration

The differential problem is the following one:

$$\begin{cases} -\Delta u + \boldsymbol{b} \cdot \nabla u = 0 & \text{in } \Omega = (-4,4)^2 \setminus N(\mu) \\ u = 0 & \text{on } \Gamma_C \\ \dfrac{\partial u}{\partial n} = 0 & \text{on } \Gamma_N \\ u = 1 & \text{on } \partial\Gamma_{in} \end{cases} \tag{2.5.5}$$

where $N(\mu)$ is the rotating airfoil (μ is the rotating angle). In order to deal with high Péclet numbers, we consider a strongly consistent stabilized formulation[47]. We refer to [25] for the theoretical aspects; here we just recall some details.

The formulation on the actual configuration is the following one:

$$a_h(u,v,\mu) = \int_{\Omega(\mu)} \nu \nabla u \cdot \nabla v \, d\boldsymbol{y} + \int_{\Omega(\mu)} \boldsymbol{b} \cdot \nabla u v \, d\boldsymbol{y} + \sum_{K(\mu) \in \mathcal{T}_h(\mu)} \frac{h_K(\mu)\delta}{|\boldsymbol{b}|} \int_{K(\mu)} (\boldsymbol{b} \cdot \nabla u)(\boldsymbol{b} \cdot \nabla v) \, d\boldsymbol{y} = 0. \tag{2.5.6}$$

[47] In all these simulations we have used piecewise linear finite elements so SUPG, GLS and DW are equivalent.

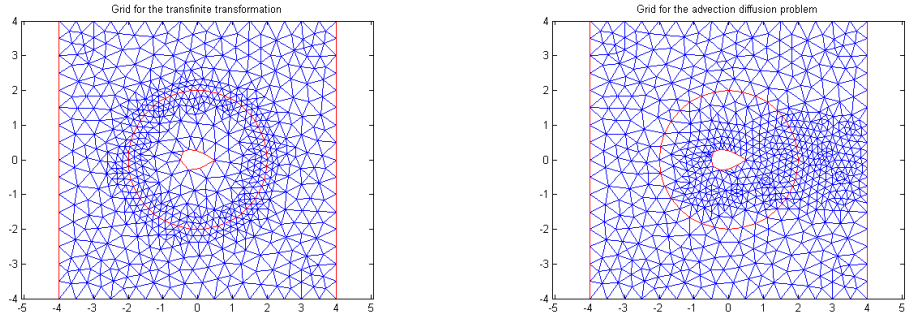
It is easy to prove that the formulation in the reference configuration ($\Omega = \Omega(0)$) is:

$$a_h(u,v,\mu) = \int_\Omega \boldsymbol{\nu}_T(\mu)\nabla u\cdot\nabla v\, d\boldsymbol{x} + \int_\Omega (\boldsymbol{J}_T^{-1}(\mu)\boldsymbol{b})\cdot\nabla u v\, d\boldsymbol{x} + \sum_{K\epsilon\mathcal{T}_h} \frac{h_K\delta}{|\boldsymbol{b}|}\int_K \boldsymbol{\Psi}_T(\mu)\nabla u\cdot\nabla v\, d\boldsymbol{x} = 0 \tag{2.5.7}$$

where:

$$\boldsymbol{\nu}_T(\mu) = \nu \boldsymbol{J}_T^{-T}(\mu)\boldsymbol{J}_T^{-1}(\mu), \quad \boldsymbol{\Psi}_T(\mu) = \begin{bmatrix} \left(\boldsymbol{J}_T^{-1}(\mu)\boldsymbol{b}\right)_1^2 & \frac{1}{2}\left(\boldsymbol{J}_T^{-1}(\mu)\boldsymbol{b}\right)_1\left(\boldsymbol{J}_T^{-1}(\mu)\boldsymbol{b}\right)_2 \\ \frac{1}{2}\left(\boldsymbol{J}_T^{-1}(\mu)\boldsymbol{b}\right)_1\left(\boldsymbol{J}_T^{-1}(\mu)\boldsymbol{b}\right)_2 & \left(\boldsymbol{J}_T^{-1}(\mu)\boldsymbol{b}\right)_2^2 \end{bmatrix}.$$

We consider two different grids (see Figure 2.19). The first one is suited to the components of the transformation and is refined on the circle that separates the two subdomains (see Figure 2.19(a)); the second one is suited to the problem and is refined near the wing and along the wake[48], (see Figure 2.19(b)).



(a) (b)

Figure 2.19: numerical grids for (a)the transfinite transformation; (b) the problem.

Here we present a comparison between the finite element solution of the problem (2.5.6) in the actual domain and the finite element solution of the problem (2.5.7) in the reference domain after the application of the EIM[49].

In Table 2.3 we report the difference between the FE solutions for two different choice of the field $\mathbf{b}$.

As we may expect, the difference depends on the Péclet number; on the other hand the incidence angle does not influence the difference. This shows that the EIM provides a good approximation for all the values of the parameter sample. On the other hand, we observe that the Peclét number deteriorates the approximation, see the estimate (1.8.5) for a mathematical explanation of the phenomenon.

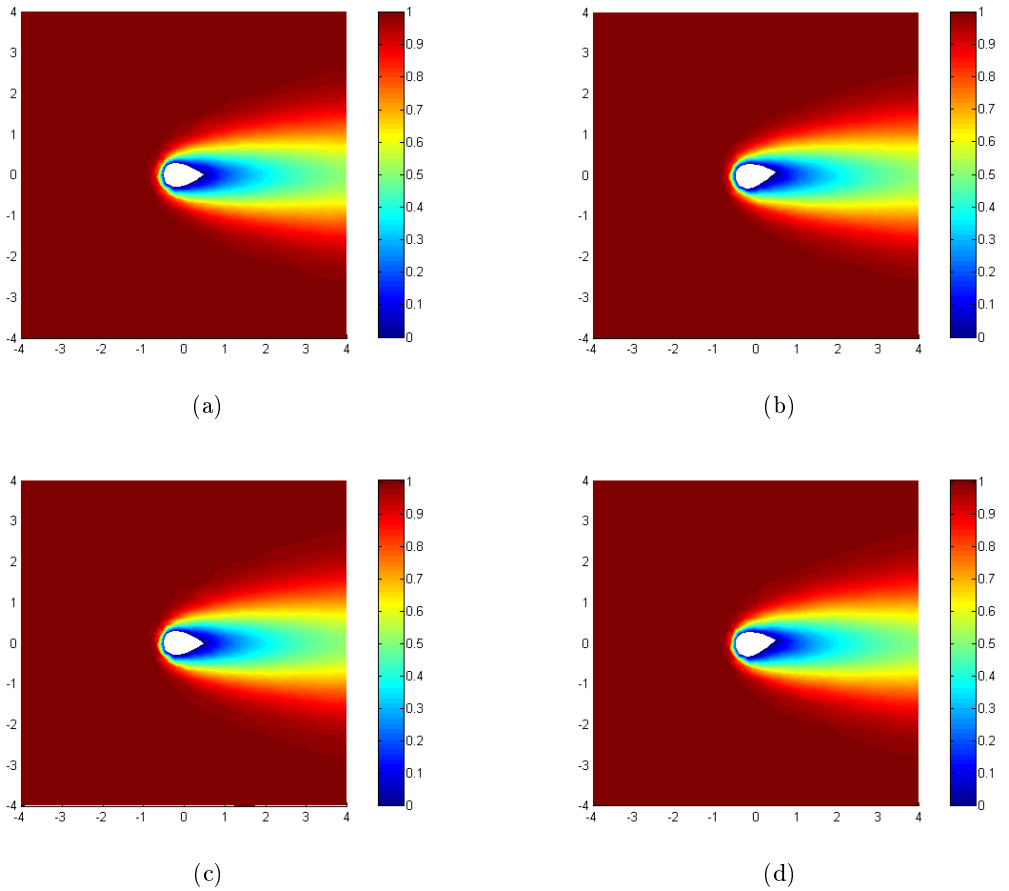
Figure 2.20 shows the solutions: we observe that the only significant differences between the two solutions are in the strong gradient region perpendicular to the airfoil. If the difference were not acceptable for our purposes, we would have the possibility to consider another inner artificial domain -instead of the unitary circle we consider and in this example - and refine the grid as much as we need. This shows the flexibility of the methodology, that is in our opinion its great advantage with respect to the multi-domain approach tested in section 2.2.4.

[48] In the differential problem, the number of grid nodes is 734.

[49] In order to evaluate the parametrization in our opinion it is better to separate the error linked to the map and the error linked to the reduced basis approximation.

	Relative (Absolute) L^2-error ($\mu = 0$)	Relative (Absolute) L^2-error ($\mu = 0.15$)
$\boldsymbol{b} = [1\ 0]$	0.0019 (0.0128)	0.0019 (0.0128)
$\boldsymbol{b} = [10\ 0]$	0.0033 (0.0285)	0.0033 (0.0246)

Table 2.3: relative (with respect to the norm of the finite element solution computed on the actual domain) and absolute difference between the finite element computed on the actual domain and the finite element solution computed on the reference domain for two different angles of incidence and two different Péclet number.



(a) (b)

(c) (d)

Figure 2.20: the advection-diffusion problem with $\boldsymbol{b} = [10\ 0]'$: comparison between the finite elements solutions. (a)-(b) FE solution in the actual domain (no mapping), (c)-(d) FE solution in the reference domain (piecewise affine mapping and EIM).

2.6 Conclusions

In this chapter we dealt with suitable parametrizations for parameter dependent domains in the context of the Reduced Basis method. After a survey on the state of the art, we focused on a new methodology: the piecewise transfinite approach for small deformations.

The mapping here proposed is a simplification of the well-known transfinite approach particularly suitable for rigid transformations of complex structures[50]. In this chapter we

[50]Theoretically the approach could be also applied to more complex geometries and transformations. When it works, i.e. the deformation is small enough, it is surely much more reliable and fast than the complete transfinite approach.

showed that:

- the hypothesis of small deformations is not truly restrictive for some applications;
- for rigid deformations of complex shapes the method is surely more indicated than the multi-domain affine approach especially because it does not impose constraints to the FE grid and reduces significantly the offline computational costs;
- as a next step, it is necessary to test the method on more involved equations and even more complex structures. The testing should be articulated into two different steps: the first one should be focused on the assessment of the small deformation hypothesis - for this tasks the two indicators proposed in (2.5.2) and (2.5.4) can be used - and on the convergence of the EIM, the second one should be directly related to the RB approximation of the parametrically affine problem.

Before concluding, let us briefly address the reasons why the study of parametrization techniques is weighty to deal with conservation laws.

Let $\Omega = (0,1)^2 \subset \mathbb{R}^2$ and let $\boldsymbol{\gamma}(\boldsymbol{\mu}) \subset \Omega$ be a curve that splits the domain into two regular subdomains, say $\Omega_1(\boldsymbol{\mu})$ and $\Omega_2(\boldsymbol{\mu})$:

$$\begin{aligned} \Omega_1(\boldsymbol{\mu}) &:= \{\boldsymbol{x} \in \Omega : \ x_1 < \gamma(\boldsymbol{\mu}, x_2)\} \\ \Omega_2(\boldsymbol{\mu}) &:= \{\boldsymbol{x} \in \Omega : \ x_1 > \gamma(\boldsymbol{\mu}, x_2)\} \end{aligned} \qquad \boldsymbol{\gamma}(\boldsymbol{\mu}, s) = \begin{bmatrix} \gamma(\boldsymbol{\mu}, s) \\ s \end{bmatrix} \tag{2.6.1}$$

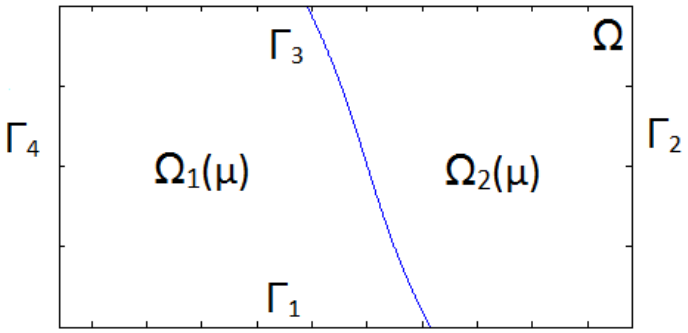


Figure 2.21: problem domain

Let us suppose that the curve γ represents a discontinuity; then, following the approach that we is discussed in the next chapter, it is necessary to solve an equation in $\Omega_1(\boldsymbol{\mu})$ and one in $\Omega_2(\boldsymbol{\mu})$. For this reason, we need two suitable applications, $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$, that map the problems into a given reference configuration $\hat{\Omega}$:

$$\boldsymbol{T}_1 : \hat{\Omega} \times \mathcal{D} \to \Omega_1(\boldsymbol{\mu}) \quad \boldsymbol{T}_2 : \hat{\Omega} \times \mathcal{D} \to \Omega_2(\boldsymbol{\mu})$$

It is quite evident that the classic multi-domain approach is inadequate to deal with such situation due to the complexity of the deformation.

Let us derive the equations for the transfinite map proposed in this chapter. First, we consider $\hat{\Omega} = \Omega$ and we define

$$\boldsymbol{g}_1(\boldsymbol{x}, \boldsymbol{\mu}) = \begin{cases} \gamma(\boldsymbol{\mu}, 0) x_1 & \text{if } \boldsymbol{x} \in \Gamma_1 \\ \gamma(\boldsymbol{\mu}, x_2) & \text{if } \boldsymbol{x} \in \Gamma_2 \\ \gamma(\boldsymbol{\mu}, 1) x_1 & \text{if } \boldsymbol{x} \in \Gamma_3 \\ x_2 & \text{if } \boldsymbol{x} \in \Gamma_4 \end{cases} \qquad \boldsymbol{g}_2(\boldsymbol{x}, \boldsymbol{\mu}) = \begin{cases} \gamma(\boldsymbol{\mu}, 0) + (1 - \gamma(\boldsymbol{\mu}, 0)) x_1 & \text{if } \boldsymbol{x} \in \Gamma_1 \\ x_2 & \text{if } \boldsymbol{x} \in \Gamma_2 \\ \gamma(\boldsymbol{\mu}, 0) + (1 - \gamma(\boldsymbol{\mu}, 0)) x_1 & \text{if } \boldsymbol{x} \in \Gamma_3 \\ \gamma(\boldsymbol{\mu}, x_2) & \text{if } \boldsymbol{x} \in \Gamma_4 \end{cases}$$

where Γ_i, with $i = 1, 2, 3, 4$ are the domain boundaries (see Figure 2.21). It is easy to observe that $\boldsymbol{g}_l(\partial\Omega, \boldsymbol{\mu}) = \partial\Omega_l(\boldsymbol{\mu})$, $l = 1, 2$. Then we can define $\boldsymbol{T}_l(\cdot, \boldsymbol{\mu})$ as the solution of the following problems:

$$\begin{cases} \Delta \boldsymbol{T}_1(\boldsymbol{\mu}) = \boldsymbol{0} & \text{in } \Omega \\ \boldsymbol{T}_1(\boldsymbol{\mu}) = \boldsymbol{g}_1(\boldsymbol{\mu}) & \text{on } \partial\Omega \end{cases} \quad \begin{cases} \Delta \boldsymbol{T}_2(\boldsymbol{\mu}) = \boldsymbol{0} & \text{in } \Omega \\ \boldsymbol{T}_2(\boldsymbol{\mu}) = \boldsymbol{g}_2(\boldsymbol{\mu}) & \text{on } \partial\Omega. \end{cases} \tag{2.6.2}$$

If the curve γ is reasonably smooth, we expect that the hypothesis of small deformations holds.

In our opinion transfinite approach seems to be the most convincing geometric reduction strategy among the four here presented to deal with hyperbolic problems. In fact the boundary control - i.e., the precise description of the shock curve - is absolutely crucial in this context. Moreover, when it works, the simplified approach permits to hugely reduce the computational effort both in the offline and in the online stage.

Chapter 3

Reduced Basis Techniques for Conservation Laws

3.1 Introduction

This chapter is related to the application of the Reduced Basis (RB) method to one dimensional scalar conservation laws that depends on a set of parameters, say $\boldsymbol{\mu} \in \mathcal{D}$:

$$\begin{cases} \dfrac{\partial}{\partial t} u(\boldsymbol{\mu}) + \dfrac{\partial}{\partial x} f(u(\boldsymbol{\mu}), \boldsymbol{\mu}) = 0 \quad (t,x) \in (0, T_{max}) \times (a,b) \\ u(\boldsymbol{\mu}, 0, x) = u_0(\boldsymbol{\mu}, x) \qquad\qquad\qquad\qquad\qquad\quad x \in (a,b) \end{cases} \tag{3.1.1}$$

completed with periodic or inflow boundary conditions. In the literature the main references concerning the RB application to hyperbolic problems are [49, 29] and some works related to viscous Burgers equation[1] [123, 87, 94]. Even if our examples mainly deal with quadratic fluxes, we aim at developing an approach suited to deal with general non-linear fluxes.

In the introduction of this thesis we discussed about the RB method for both steady and time-dependent problems. The example 1.2 highlights some criticalities in the application of the RB methodology to hyperbolic problems with discontinuous initial data. The following example faces the same problem from a different point of view and provides some solutions.

3.1.1 An introductive example

Let us first analyse a simple linear problem.

$$\begin{cases} \text{Given } \mu \in \mathcal{D} = [\mu_{min}, \mu_{max}], \\ \\ \text{find } u(\mu) \in C(0,T; L^2(\mathbb{R})) \text{ such that } \begin{cases} \dfrac{\partial}{\partial t} u(\mu) + \mu \dfrac{\partial}{\partial x} u(\mu) = 0 \quad (t,x) \in (0,\infty) \times \mathbb{R} \\ u(\mu, 0, x) = \chi_{x>0}(x) \quad x \in \mathbb{R}, \end{cases} \end{cases} \tag{3.1.2}$$

whose (elementary) solution is:

$$u(\mu, t, x) = \begin{cases} 0 & x \le \mu t \\ 1 & x > \mu t. \end{cases} \tag{3.1.3}$$

[1] We also refer to [96] for the study of the Burgers equation in the presence of uncertainty.

In order to apply the Reduced Basis method, we consider a truth approximation of the solution, say $u_\delta(\mu)$, with $\delta = (\Delta t, h)$. For the sake of simplicity, we consider a piecewise constant in time approximation i.e.

$$u_\delta(\mu, t, x) = \sum_k u_h^k(\mu, x)\chi_{\{t^k, t^{k+1}\}}(t)$$

where $\{t^k\}_{k=0}^{\mathbb{K}}$ represents the temporal grid . In addition we suppose that the truth solver is fine enough to catch correctly the shock position for each time step. In conclusion the truth manifold is[2]:

$$\mathcal{M}^\delta = \left\{ u_\delta(\mu) = \sum_k H_{\{\mu t^k\}}\chi_{[t^k, t^{k+1})} : \mu \in \mathcal{D} \right\} \tag{3.1.4}$$

To apply a reduced basis strategy to problem (3.1.2), we have to choose the correct RB space. The three main options- see for instance [58]- are:

- *the Lagrange subspace*: for time-dependent problems in the RB context, the usual Lagrangian space considered is the following[3]:

$$W_N := S_{\Delta t} \otimes V_N \tag{3.1.5a}$$

where:

$$\begin{cases} S_{\Delta t} := \text{span } \left\{ \chi_{\{[t^k, t^{k+1})\}} : k = 0, \cdots, \mathbb{K} - 1 \right\} \\ V_N := \text{span } \left\{ u_h^{k_j}(\mu_j) : \text{ for some couples } (\mu_j, k_j) \in \mathcal{D} \times \{0, \cdots, \mathbb{K}\},\ j = 1, \cdots, N \right\}; \end{cases} \tag{3.1.5b}$$

- *the Taylor subspace*: let us suppose that the solution u_δ is N-time derivable in $\mu = \mu_{ref} \in \mathcal{D}$ with respect to the parameter. Thus the reduced basis subspace is defined as:

$$X_N = \text{span} \left\{ y_j : y_j = \frac{\partial^j u_\delta}{\partial \mu^j}\Big|_{\mu=\mu_{ref}} \quad j = 0, \cdots, N \right\}; \tag{3.1.6}$$

- *the Hermite subspace*: the idea is to combine the Lagrange and Taylor approaches: the reduced basis subspace is spanned by the solutions and their first order partial derivatives at various parameter values μ_j:

$$H_N = \text{span} \left\{ y_j = u_\delta(\mu_j) \text{ and } \frac{\partial u_\delta}{\partial \mu}\Big|_{\mu=\mu_j} \quad \mu_j \in \mathcal{D} \quad j = 1, \cdots, N \right\}. \tag{3.1.7}$$

It should be noticed that, even if it is not a priori impossible, the computation of derivatives could be rather involved. In addition the derivative could be irregular: for our problem it

[2] $H_{x^\star} : \mathbb{R} \to \mathbb{R}$ is the Heaviside function centered at $x^\star$.

[3] By applying the POD Greedy technique we would obtain a slight different approximation space; however, the conclusions would be substantially unchanged. As already cited in the introduction, in [113] Urban et al. have considered another Lagrangian space, based on the entire solutions to problem (3.1.2) corresponding to various parameter values μ_j i.e.:

$$W_N = \text{span} \{ y_j = u_\delta(\mu_j) : \mu_j \in \mathcal{D}\ j = 1, \cdots, N \}.$$

is easy to observe that[4]:

$$\frac{\partial u_\delta}{\partial \mu}(\mu,t,x) = -\sum_k \delta_{\mu=\frac{x}{t^k}} \chi_{\{t^k,t^{k+1}\}}.$$

Also standard Lagrange subspaces are not suitable for the approximation of the solution manifold. To show it, we consider the Lagrangian space (3.1.5)

Given a new value of the parameter, as a pure theoretical exercise we try to compute the W_N-minimizer of the L^p-distance[5] from the solution $u_\delta(\mu)$ at a fixed time t. We point out that we do not introduce any practical approach to implement the minimization: having the solution, it is possible to explicitly compute the exact minimizer. The result is:

$$u(\mu_{j^\star}, t^{j^\star}, \cdot) = \arg\inf_{w\in V_N} \|u_\delta(\mu,t,\cdot) - w(\cdot)\|_{L^p(\mathbb{R})} \qquad j^\star = \min_{j=1,\cdots,N} |\mu t - \mu_j t^j|.$$

This formula shows that the convergence is linear with respect to the density of the parameter sample. In practice the approximation error related to the Lagrangian space W_N decreases too slowly to motivate the application of the RB method.

In conclusion no approach seems to be suitable for the problem; thus some corrections are probably required.

Let us analyse, more in detail, why the Lagrangian approach does not work. On one hand the knowledge of the solution for different values of the parameter seems to provide us much information about the entire manifold; on the other hand a method based on linear combinations of snapshots is surely inadequate to take advantage of this information. By combining linearly a set of functions, the resulting singularity set[6] coincides with the union of the singularity sets of the different addends; this is why in order to reconstruct a function with a singularity in x_0 through a linear combination of pre-computed snapshots it is necessary *to have a snapshot with a singularity close to* x_0. It is obvious that it is not possible to guarantee this condition with a reasonable number of snapshots.

However, we observe that the real solution u is smooth in each subdomain of the domain decomposition induced by the jump $[0,T]\times\mathbb{R} = \bar{\Omega}_1 \cup \bar{\Omega}_2$:

$$u(\mu,t,x) = u_{smooth,1}(\mu,t,x) \equiv 0 \quad \forall\,(t,x) \in \Omega_1 = \{(t,x) \in [0,T]\times\mathbb{R} :\, x < \mu t\},$$
$$u(\mu,t,x) = u_{smooth,2}(\mu,t,x) \equiv 1 \quad \forall\,(t,x) \in \Omega_2 = \{(t,x) \in [0,T]\times\mathbb{R} :\, x > \mu t\}.$$

Furthermore, it is easy to notice that $u_{smooth,1}(\mu)$ and $u_{smooth,2}(\mu)$ depend continuously on the parameter-in this case they are even constants; for this reason we expect that they can be approximated by the corresponding smooth parts of the pre-computed solutions. The difficulty is that a priori also the domain decomposition depends on the parameter through

[4] The equation should be intended in $\mathcal{D}'((\mu_{min},\mu_{max}))$. The calculus of the derivative is correct if we consider the solution as $u_\delta : [0,T]\times[\mu_{min},\mu_{max}] \to L^\infty(\mathbb{R})$. It could be argued that -in bounded domains- other L^p norms can be considered: however the ratio:

$$\left\| \frac{u_\delta(\mu_1,t) - u_\delta(\mu_2,t)}{\mu_1-\mu_2} \right\|_{L^p} = (|\mu_1-\mu_2|t)^{\frac{1}{p}-1}$$

is unbounded for $|\mu_1-\mu_2| \to 0$ if $p > 1$ and for $p = 1$ does not converge in the space norm. Thus it is not possible to define Fréchet or Gâteaux derivatives for $u_\delta : [0,T]\times[\mu_{min},\mu_{max}] \to L^p(\mathbb{R})$.

[5] Here we consider $p < \infty$. If $p = \infty$ then the minimizer associated with $u_\delta(\mu,t,\cdot) \notin V_N$ is not unique in V_N. More precisely, if $u_\delta(\mu,t,\cdot) \notin V_N$, every $w \in V_N$, such that $0 \le w \le 1$ is a minimizer.

[6] In this work given an almost everywhere continuous function $f : \mathbb{R}^m \to \mathbb{R}^n$, we refer to the singularity set of f as $\{\bar{x} \in \mathbb{R}^m : \nexists \lim_{x\to\bar{x}} f(x)\}$.

the well-known Rankine-Hugoniot condition, [71, 109], that is -called $x(t)$ the position of the shock at the time t:

$$\dot{x}(t) = \frac{q(u(\mu,t,x^+(t)) - q(u(\mu,t,x^-(t))}{u(\mu,t,x^+(t)) - u(\mu,t,x^-(t))}, \quad \text{with } q(x) = \mu x. \tag{3.1.8}$$

However, thanks to the linearity, the equation (3.1.8) does not depend directly on the solution; thus it is straightforward to solve the ODE and find that $x(t) = \mu t$. Moreover, due to the fact that the initial condition is piecewise constant, it is possible to write the general solution as a composition between a solution related to a specific value μ_{ref} of the parameter and a suitable parameter-dependent map induced by the above mentioned Rankine-Hugoniot condition (3.1.8):

$$u(\mu,t,x) = u(\mu_{ref}, \boldsymbol{\tau}(\mu,t,x)) \qquad \text{where } \boldsymbol{\tau}(\mu,t,x) = \begin{bmatrix} x \\ \frac{\mu}{\mu_{ref}} t \end{bmatrix}.$$

This slight modification of the general approach permits to solve our problem in an efficient way. However, at this stage it is not clear how to extend the method to nonlinear fluxes and to more general initial data.

3.1.2 Overall strategy and structure of the chapter

In order to generalize the example proposed above, we introduce a new functional space[7].

Definition 3.1. *Let $I = (a,b) \subset \mathbb{R}$ be an interval and $w : I \to \mathbb{R}$ be a measurable[8] function. The total variation of w is defined by:*

$$TV(w) := \sup_{\{x_j\}_j : a<x_j<x_{j+1}<b} \left\{ \sum_j |w(x_j) - w(x_{j-1})| \right\} \tag{3.1.9}$$

Then we denote with $BV(I)$ the set of all the real valued measurable functions $w : I \to \mathbb{R}$ with bounded total variation.

If the initial condition and the boundary condition are sufficiently regular, the solution to the problem (3.1.1) could be searched[9] in the following subspace of $BV(I)$.

Definition 3.2. *We say that $w \in BV(I)$ is a special function with bounded variation, and we write $w \in SBV(I)$, if $w = w_s + w_j$ where $w_s \in W^{1,1}(I)$ is the smooth part of w while w_j is piecewise constant, i.e., $Dw_j = \sum_{x_i \in A} \gamma_i \delta_{x_i}$ where A is a countable set, δ_{x_i} are the Dirac measures at x_i and $\gamma_i \in \mathbb{R}$.*

In the sequel, we will refer to w_s and w_j as to the smooth and jump component, respectively. The example stated in the previous paragraph heuristically showed how the RB approach - that is based on linear combinations of pre-computed solutions- is not adequate to deal with the jump part of the solution. However, it is reasonable to expect that smooth components are well approximated through linear combinations of the corresponding smooth components of the pre-computed solutions. For this reason the strategy we propose here consists in:

1. building a surrogate problem for the smooth component of the solution and take advantage of the RB approach to reduce the computational effort;

[7] In Appendix A we introduce the space BV in a more general and abstract setting.
[8] In this work the measurability is always intended with respect to the Lebesgue σ-algebra.
[9] See [1, 24] for further details.

2. computing the jump component of the solution, independently;

3. adding the two components in order to find the solution.

Apart from these distinguished features, several further auxiliary ingredients will characterize our approach. For this reason in order to clearly explain the methodology, the chapter is structured in four distinct sections.

- Section 3.2 deals with some preliminary topics: the underlined truth approximation chosen and the consequent treatment of boundary conditions; an algorithm to split the smooth and the jump parts of the solutions; a procedure to change spatial variables in the Finite Volume framework.

- In section 3.3 the entire methodology is presented: first, a new formulation is derived and then the algorithms to efficiently treat the shock and the smooth parts of the solutions are presented.

- In section 3.4 the a posteriori error estimation is briefly addressed.

- In section 3.5 some numerical examples motivate the strategy.

At the end some conclusions are drawn.

3.2 Some preliminaries

The Reduced Basis method is built upon an underlined numerical scheme used to provide the so-called *truth approximation* of the solution. As motivated in section 3.3, our method is completely independent of the special scheme chosen to solve the equations[10].

In this section, we briefly introduce the well-known Godunov and Lax Friedrichs schemes in the framework of the so-called conservative methods. For a complete coverage of the topic we refer to [71, 72]. Then we discuss a general strategy used to detect the shock equation.

3.2.1 The underlined truth approximation for the problem

Given the spatial-temporal domain $Q_T = (0,T) \times (a,b)$, we consider the equispaced mesh (t^k, x_j) $k = 0, \cdots, \mathbb{K}$, $j = 0, \cdots, \mathcal{N}$ such that $x_j = a + jh$, $t^k = k\Delta t$, with $\delta = (\Delta t, h)$. A conservative method for the equation (3.1.1) is a method that can be written in the following form:

$$u_{\delta,j}^{n+1} = u_{\delta,j}^n - \frac{\Delta t}{h}\left(F(u_{\delta,j-p}^n, \cdots, u_{\delta,j+q}^n) - F(u_{\delta,j-p-1}^n, \cdots, u_{\delta,j+q-1}^n)\right) \tag{3.2.1}$$

where the solution $\{u_{\delta,j}^n\}_{n,j}$ is an approximation to the cell average of the exact solution at time t^n, i.e., $u_{\delta,j}^n \simeq \frac{1}{h}\int_{x_j-\frac{h}{2}}^{x_j+\frac{h}{2}} u(t^n, x)\, dx$.

In this work we are going to consider the Lax-Friedrichs and the Godunov fluxes:

$$\text{Lax Friedrichs Flux: } F(u,w) = \frac{h}{2\Delta t}(u-w) + \frac{1}{2}(f(u)+f(w)); \tag{3.2.2a}$$

[10] This is not true for the RB method based on the Galerkin projection: two different bilinear forms generate two different reduced models.

$$\text{Godunov Flux: } F(u,w) = \begin{cases} \min_{z\in[u,w]} f(z) & \text{if } u \le w \\ \max_{z\in[w,u]} f(z) & \text{if } w \le u. \end{cases} \tag{3.2.2b}$$

These fluxes satisfy the following important properties:

- the fluxes are consistent, i.e., $F(u,u) = f(u)$;
- they lead to monotone schemes, i.e., $v_j^n \ge u_j^n$ for any j implies that $v_j^{n+1} \ge u_j^{n+1}\ \forall\, j$. As a consequence the schemes are L^1-contractive and TVD (Total Variation Diminishing). The latter property guarantees that no spurious oscillation arises close to the shock;
- they satisfy the following *discrete entropic condition*: given a C^1-convex function η and a C^1-function ψ such that $\psi' = \eta' f'$, it is possible to define a consistent entropic flux Ψ[11] such that the numerical solution satisfies

$$\eta(u_{\delta,j}^{k+1}) \le \eta(u_{\delta,j}^k) - \frac{\Delta t}{h}(\Psi_{j+\frac{1}{2}} - \Psi_{j-\frac{1}{2}}). \tag{3.2.3}$$

In order to be stable these methods require that the mesh satisfies the following CFL (Courant, Friedrichs, Lewy) condition ([71]):

$$\frac{\Delta t}{h}\|f'(u)\|_\infty \le 1. \tag{3.2.4}$$

Under the previous hypothesis it is possible to prove that the methods are convergent to the entropic solution.

The treatment of boundary conditions

Due to the fact that in this work we focus on conservation laws in bounded domains, we briefly address here the treatment of boundary conditions that we use in our simulations.

Two different types of boundary conditions can be imposed to problem (3.1.1): a suitable condition at each inflow boundary or periodic conditions (i.e., $u(t,a) = u(t,b)$).

We remember that $x = x^\star$ is said to be an inflow boundary for problem (3.1.1) if

$$f(u(t,x^\star))n(x^\star) < 0 \quad \text{where} \quad n = \begin{cases} -1 & \text{if } x^\star = a \\ 1 & \text{if } x^\star = b. \end{cases} \tag{3.2.5}$$

As usual in the context of Finite Volume schemes[12], in this work we impose the boundary conditions through the introduction of some auxiliary points, the so-called *ghost points*[13]. Due to the fact that we use three point stencils, we just need to add a single ghost point for each boundary, i.e., $u_{\delta,-1}^k$ and $u_{\delta,\mathcal{N}+1}^k$.

[11] For the Godunov scheme the entropic flux is:

$$\Psi(u,w) = \begin{cases} \psi(u) & f'(u) > 0, \quad s = \frac{f(w)-f(u)}{w-u} > 0 \\ \psi(w) & f'(u) < 0, \quad s < 0 \\ \psi(\xi) & f'(u) < 0 < f'(w) \end{cases}$$

where ξ is the zero of the flux derivative (i.e., $f'(\xi) = 0$).

[12] The technique here presented constitutes the basis for the implementation adopted in the hyperbolic solver `Clawpack`, [72].

[13] Ghost point technique is not the only option to impose boundary conditions in Hyperbolic problems- for instance in [72] another technique based on the definition of a suitable flux at the boundary is presented- , however the technique here adopted is extremely easy to implement and constitutes a powerful tool also in more involved problems and for more accurate schemes.

- *Periodic Boundary conditions:* in this case we simply impose $u^k_{\delta,-1} = u^k_{\delta,\mathcal{N}}$ and $u^k_{\delta,\mathcal{N}+1} = u^k_{\delta,0}$.
- *Outflow boundaries*: if the condition (3.2.5) is not satisfied, the differential problem does not need any boundary condition. However, in order to compute the new solution at the outflow boundary, say $x = b$, we need to set the ghost point value. This can be done by zero-order[14] extrapolation, i.e., $u^k_{\delta,\mathcal{N}+1} = u^k_{\delta,\mathcal{N}}$.
- *Inflow boundaries:* From the mathematical point of view, the correct way to impose inflow condition is[15]:

$$u(t,a) \text{ is such that } \max_{k\in(u(t,a),\xi_0(t))} \{\text{sign}\,(u(t,a)-\xi_0(t))\,[f(u(t,a))-k]\} = 0 \quad (3.2.6)$$

  where a is inflow boundary and ξ_0 is the condition at $x = a$. In FV schemes, the ghost point can be set to[16]: $u^k_{\delta,-1} = \frac{1}{\Delta t}\int_{t^k}^{t^{k+1}} \xi_0(\tau)\,d\tau$.

3.2.2 Smooth-Jump decomposition algorithm

Given the problem (3.1.1), we suppose that a conservative scheme with a monotone flux (either Godunov or Lax-Friedrichs in the following) is used to obtain a numerical solution $\{u^k_j\}_{k,j}$ of the problem. In this section we introduce an efficient algorithm to decompose the solution in a regular part and in a piecewise constant part. For our purposes, the main goal of the method is to detect the initial condition, i.e., $(x_{shock}(t^\star),\ u^-(t^\star, x_{shock}(t^\star))$, $u^+(t^\star, x_{shock}(t^\star))$, to initialize the shock capturing scheme; however we present the algorithm from a more general viewpoint.

In order to deal with the problem, we first make few observations.

- Each numerical scheme introduces an artificial diffusion: hence a shock is in practice a region (hopefully small) in which the solution exhibits a strong gradient. For this reason, it is in practice impossible to find the right shock point, but we can assume that it is inside the above-mentioned high-derivative region.
- Some numerical schemes (such as Lax-Wendroff) can exhibit spurious oscillations close to the shock. This limits our possibilities to recognize in an automatic way the two shocks because the high derivative region becomes often larger and unstable with respect to the sign of the derivative. This is the reason why we decided to first test our method on discontinuous solutions obtained with monotone schemes. However we point out that the algorithm here presented is absolutely independent of the method used to build the solution.
- It is possible to forecast a priori whether the jump sign is positive or negative. In particular if the flux is concave, the entropy solution has only positive jumps. This is fundamental because also rarefaction waves have unbounded gradients and so, without an a priori sign condition, it would be much more difficult to distinguish numerically the two phenomena.

The main idea of the algorithm is to identify the high derivative (with respect to the right sign) regions and then to quantify the entity of the jump. In order to do this we have

[14] First order extrapolation represents a further alternative; however, it can lead to stability problems and it is in general not recommended ([72]).

[15] See [6].

[16] The motivation is based on the characteristic equation that is introduced in section 3.3.2

to set two constants: the constant K and the extension δ_{visc} of the artificial boundary layer. In the following we will consider[17]:

$$K = 30\% \frac{\Delta u}{h} \quad \delta_{visc} = \frac{h^2}{\Delta t} \quad \text{where } \Delta u := \max_{\mathbb{R}} u_0 - \min_{\mathbb{R}} u_0 \tag{3.2.7}$$

Despite all these assumptions, the problem is still quite involved. Basically, there are two difficult situations to deal with:

- the case in which the shock starting point is strictly positive;
- the case in which we have an intersection between shocks.

In the practical implementation we can deal with these two cases together at the same time. However, for the sake of simplicity, we treat the two cases separately.

Let us consider the following example[18]

$$\begin{cases} \dfrac{\partial}{\partial t} u + v \dfrac{\partial}{\partial x} (u(1-u)) = 0 & (t,x) \in [0,T) \times \mathbb{R} \\ u(0,x) = g(x) = \begin{cases} \frac{1}{3} & x \le 0 \\ \frac{1}{3} + \frac{5}{12} x & 0 \le x \le 1 \\ \frac{3}{4} & x \ge 1 \end{cases} \end{cases} \tag{3.2.8}$$

It is possible to prove that the solution is

$$u(t,x) = \begin{cases} \frac{1}{3} & x < \min\{\frac{1}{3}vt, \frac{1}{2} - \frac{1}{12}vt\} \\ \dfrac{4 + 5x - 5vt}{2(6 - 5vt)} & \frac{1}{3}vt \le x \le 1 - \frac{1}{2}vt, \quad t < \frac{6}{5v} \\ \frac{3}{4} & x > \max\{1 - \frac{1}{2}vt, \frac{1}{2} - \frac{1}{12}vt\}. \end{cases} \tag{3.2.9}$$

When the shock starts at $t = t^\star = \frac{6}{5v} > 0$ we have that the solution exhibits a derivative that becomes unbounded for $t \to t^\star$. In this case the only reasonable strategy is to check whether the maximum of the derivative is monotonically increasing in time: we expect indeed that the numerical solution has an increasing derivative before stabilizing at a certain level. In algorithm 8 this procedure is fully described.

As regards the decomposition step, some obvious modifications can be made in order to fit the method with our purposes. For instance in our case we split the two smooth components and we refer them to the global mesh. In order to test the algorithm we consider problem (3.2.8) for some specific values of the parameter. The Godunov scheme is used with $h = 0.01$ and $\Delta t = 0.002$, $T_{fin} = 5$, $\Omega = (-3, 3)$. The parameters are:

$$K = 7 \quad \delta_{visc} = \frac{h^2}{\Delta t} = 5h$$

The results are extremely good and do not depend on the choice of the parameters K and δ_{visc} that in practice are only needed to properly start the detecting procedure based on the time progression of the derivative peak. Table 3.1 summarizes them for three different values of the parameter v. For all the values of v, the numerical estimation turns out to be satisfying.

[17] For the boundary layer we use a formula somehow close to $\frac{1}{Pe}$ (see [97] chapter 5). If a finer analysis is necessary, we think that the notable Von Neumann analysis could guarantee better results. However in this work we do not study this topic.

[18] This example is taken from [110] chapter 3 exercise 2.5.

Algorithm 8 Smooth-Jump decomposition procedure (single shock case)

Given $\{u_j^k\}_{j,k}$, compute $D_j u^k = \frac{u_j^k - u_{j-1}^k}{h}$;
Compute $[M_\nabla^k, j_\nabla^k] = \max_{j=1,\cdots,\mathcal{N}} D_j u^k$, for each $k = 1, \cdots, \mathbb{K}$.

Shock starting point
if $M_\nabla^1 > K$ **then**
  $t_{shock} = 1$
else
  for $k = 1, \cdots, \mathbb{K}$ **do**
    if $M_\nabla^k > K$ AND $M_\nabla^k > \frac{c_{test}}{N_{test}} \sum_{l=1}^{N_{test}} M_\nabla^{l+k}$ **then**
      $t_{shock} = K$
      `Break`
    end if
  end for
end if

Decomposition
for $k = t_{shock}, \cdots, \mathbb{K}$ **do**
  $\Delta^k = u^k_{j_\nabla^k + \frac{\delta_{visc}}{h}} - u^k_{j_\nabla^k - \frac{\delta_{visc}}{h}}$ (**shock magnitude**)

$$u^k_{smooth,j} = \begin{cases} u_j^k & j \le j_\nabla^k - \frac{\delta_{visc}}{h} \\ u^k_{j_\nabla^k - \frac{\delta_{visc}}{h}} & j_\nabla^k - \frac{\delta_{visc}}{h} \le j \le j_\nabla^k + \frac{\delta_{visc}}{h} \\ u_j^k - \Delta^k & j \ge j_\nabla^k + \frac{\delta_{visc}}{h} \end{cases} \qquad \textbf{smooth component}$$

end for

v	CFL$= \frac{2v\Delta t}{h}$	$t(v)$ (th.)	$t(v)$ (num.)
0.6	0.12	2	2.002
1.0	0.2	1.2	1.2
1.6	0.32	0.75	0.748

Table 3.1: comparison between the exact shock starting point and the numerical estimation

Now we turn to the second problem. As working example we consider the following:

$$\begin{cases} \dfrac{\partial}{\partial t} u + v \dfrac{\partial}{\partial x} (u(1-u)) = 0 & (t,x) \in [0,T) \times \mathbb{R} \\ u(0,x) = g(x) = \begin{cases} 0 & x \le 0 \\ \frac{1}{2} & 0 \le x \le 1 \\ 1 & x \ge 1. \end{cases} \end{cases} \tag{3.2.10}$$

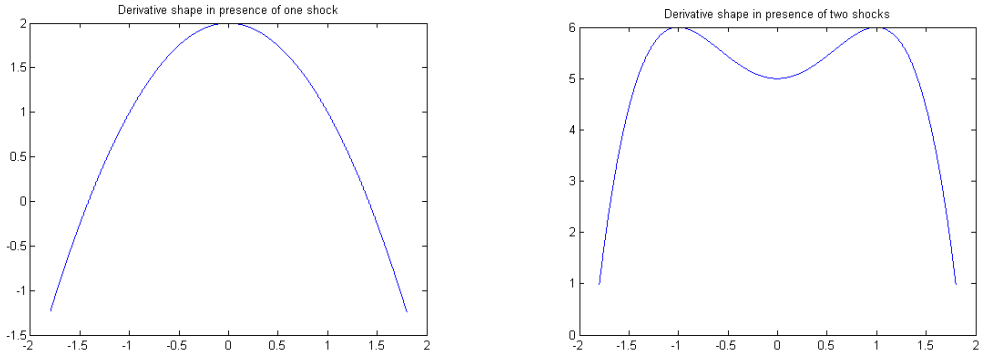
It is possible to prove that the solution is

$$u(t,x) = \begin{cases} 0 & x < \min\{\frac{1}{2}vt, \frac{1}{2}\} \\ \frac{1}{2} & \frac{1}{2}vt < x < 1 - \frac{1}{2}vt \quad t < \frac{1}{v} \\ 1 & x > \max\{1 - \frac{1}{2}vt, \frac{1}{2}\} \end{cases} \tag{3.2.11}$$

The goal of our test is to detect the exact point $(x_{shock}, t^\star(v)) = (\frac{1}{2}, \frac{1}{v})$ in which the two shocks intersect each other.

In this situation we expect that for $t < t^\star(\mu)$ we have two distinct maxima in the derivative. For $t \to t^\star(v)$ the two maxima get closer and closer. Then for $t > t^\star(v)$ only one maximum is observed.

Due to the artificial viscosity we expect also that for $t \to t^\star(\mu)$ we observe an unique boundary layer with a decreasing in time amplitude. A first strategy consists in analyzing the solution when the two high derivative regions have a not null intersection- that means (with the notation of the previous algorithm) $j^k_{\nabla,1} + \frac{\delta_{visc}}{h} \leq j^k_{\nabla,2} - \frac{\delta_{visc}}{h}$. In principle, we expect to observe a concave parabola or a quartic function with two maxima (see Figure 3.1).



(a) One shock shape (b) Two shock shape

Figure 3.1: theoretical shapes for one (a) and two (b) shocks.

However, the values of the derivatives close to the shocks are not enough reliable to distinguish the two shapes through a polynomial fitting. In our numerical tests we observed that, in practice, when the two shocks are sufficiently close, we have a unique distributed high derivative region. We think that it is not so easy to define a rigorous a priori criterium so that we focus on some data-driven strategies. Our idea is to identify a shock indicator based on the problem: first, we calibrate the indicator using the two shock branches separately and then we use such indicator to detect the time in which the two shocks intersect each other.

The indicator we chose is the ratio between the sum of the first two peaks of the derivative and the sum of the third and the fourth peak.

$$\frac{(D_j u^k)_1 + (D_j u^k)_2}{(D_j u^k)_3 + (D_j u^k)_4}. \tag{3.2.12}$$

This choice is motivated by the fact that the behaviour of the numerical solution close to the shock is linked to the numerical method, to the jump magnitude and, more in general, to the problem solution. Due to the fact that we can forecast a priori if the jump discontinuity is increasing or decreasing, we do not need to consider the absolute value of the gradient; this prevents from taking into account possible rarefaction waves that start at the time t^k. This ratio does not depend on the magnitude of the shock[19] (at least this dependence should be reasonable weak); therefore it seems to be well-suited to our problem.

[19] If for each shock branch $(D_j u^k)_l \simeq C\zeta_l \Delta u^k_{jump}$ where ζ_l depends on $l = 1, 2, 3, 4$ and $\Delta u^k_{jump} = u^k_{right} - u^k_{left}$ and $C > 0$ is a given constant, then (3.2.12) does not depend on $\Delta u^k_{jump} = u^k_{right} - u^k_{left}$. This shows that (3.2.12) does not depend (or depends weakly) on the Δu^k_{jump}.

In conclusion we define our approximation of the interaction between shocks as to:

$$t^k := \min\left\{ t^k = k\Delta t : \ \frac{(D_j u^k)_1 + (D_j u^k)_2}{(D_j u^k)_3 + (D_j u^k)_4} \geq \frac{1}{c^\star} \right\} \tag{3.2.13}$$

where $c^\star < 1$ is a given constant.

In algorithm 9 the condition is formalized while in Table 3.2 we provide some results associated with problem (3.2.10) with $h = 0.01$ and $\Delta t = 0.002$, $T_{fin} = 5$, $\Omega = (-3, 3)$.

Algorithm 9 Condition for the Smooth-Jump decomposition procedure

if $j^k_{\nabla,1} + \frac{\delta_{visc}}{h} \leq j^k_{\nabla,2} - \frac{\delta_{visc}}{h}$ **then**
  Let $j^k_{1,\star} < j^k_{2,\star}$ the indices associated with the two largest values of the vector $D_j u^k$.
  $\Delta_1 = j^k_{2,\star} - j^k_{1,\star}$
  Let $\left(\{D_j u^k\}\right)_i \geq \left(\{D_j u^k\}\right)_{i+1}$ be the derivative vector in decreasing order.

  if $\left(\{D_j u^k\}\right)_1 + \left(\{D_j u^k\}\right)_2 > \frac{1}{c^\star}\left(\left(\{D_j u^k\}\right)_3 + \left(\{D_j u^k\}\right)_4\right)$ AND $\Delta_1 \leq 1$ **then**
    $t_{shock} = k$
  end if
end if

v	$t(v)$ (th.)	$t(v)$ ($c^\star = 40\%$)	$t(v)$ ($c^\star = 50\%$)	$t(v)$ ($c^\star = 60\%$)
0.6	1.6667	n.c.	1.71	1.674
1.0	1	n.c.	1.026	1.004
1.6	0.625	n.c.	0.64	0.628

Table 3.2: estimation of the shock interaction with different values of the parameter $c^\star$ and for different values of v. n.c. indicates that the algorithm does not converge.

As we may expect, the strategy seems to be extremely precise, but it depends strongly on the choice of $c^\star$. This quantity is difficult to be chosen a priori, but it can be estimated empirically on the previous shock branches. The approach we here propose can be summarized through the following three steps:

1. we define $k^\star := \inf\{k;\ j^k_{\nabla,1} + \frac{\delta_{visc}}{h} \geq j^k_{\nabla,2} - \frac{\delta_{visc}}{h}\}$;
2. for all $k \leq k^\star$ we define:

$$c_1^{\star,k} = \left(\frac{(D_j u^k)_1 + (D_j u^k)_2}{(D_j u^k)_3 + (D_j u^k)_4} \right)^{-1}$$

where we have considered only the indices "close"[20] to $j^k_{\nabla,1}$ and:

$$c_2^{\star,k} = \left(\frac{(D_j u^k)_1 + (D_j u^k)_2}{(D_j u^k)_3 + (D_j u^k)_4} \right)^{-1}$$

where we have considered only the indices "close" to $j^k_{\nabla,2}$;

[20] Instead of considering the entire vector $\{D_j u^k\}_{j=0}^{\mathcal{N}}$ as in (3.2.12), we have considered the vector $\{D_j u^k\}_{j=j^k_{\nabla,1}-J}^{j^k_{\nabla,1}+J}$ where $J = \frac{\delta_{visc}}{h}$.

3. finally, we define $c^\star_{av}$, through a simple average, i.e.:

$$c^\star_{av} := \frac{1}{k^\star + 1} \sum_{k=0}^{k^\star} \frac{c_1^{\star,k} + c_2^{\star,k}}{2}$$

Table 3.3 provides evidence to support our strategy.

v	$c^\star_{av}$	$t(v)$ (th.)	$t(v)$ $(c^\star = c^\star_{av})$
0.6	0.6642	1.6667	1.66
1.0	0.6555	1	0.998
1.6	0.6416	0.625	0.624

Table 3.3: comparison between the exact shock starting point and the numerical estimation

3.2.3 Finite volume projection algorithm

Let us consider the equispaced space-time mesh (t^k, x_j) $j = 0, \cdots, \mathcal{N}$, $k = 0, \cdots, \mathbb{K}$ previously defined. Furthermore, let us consider the numerical solution to problem (3.1.1) computed through a finite volume scheme, say $\{u^k_{\delta,j}\}_{j=1}^{\mathcal{N}}$ where $u^k_{\delta,j}$ is a suitable approximation of the solution in the j-th cell at the time t^k.

The problem we are going to focus is the following one: for each time t^k, given the approximate vector $\{u^k_{\delta,j}\}_{j=j_l}^{j_r}$ associated with the function $u(t^k, x)$ with $x \in (x_l, x_r) = (x_{j_l}, x_{j_r})$, we aim at computing the approximate vector $\{\tilde{u}^k_{\delta,j}\}_{j=1}^{\mathcal{N}}$ referred to $\tilde{u}(t^k, x) := u(t^k, x_l + \frac{x_r - x_l}{b-a}(x-a))$ and vice versa. The necessity for this procedure will be clear in section 3.3.1 when we map some parts of the solution in a parameter and time independent domain. Due to the fact that the time does not play any role in this algorithm, in the following the superscript k is omitted.

In order to be clearer about the purposes of the algorithm, we provide a preliminary example. Let us consider the mesh $\{\frac{1}{2}, \frac{3}{2}\}$. The vector $[1, 2]'$ is associated with the piecewise constant function:

$$u(x) = \begin{cases} 1 & x \in [0,1) \\ 2 & x \in [1,2]. \end{cases}$$

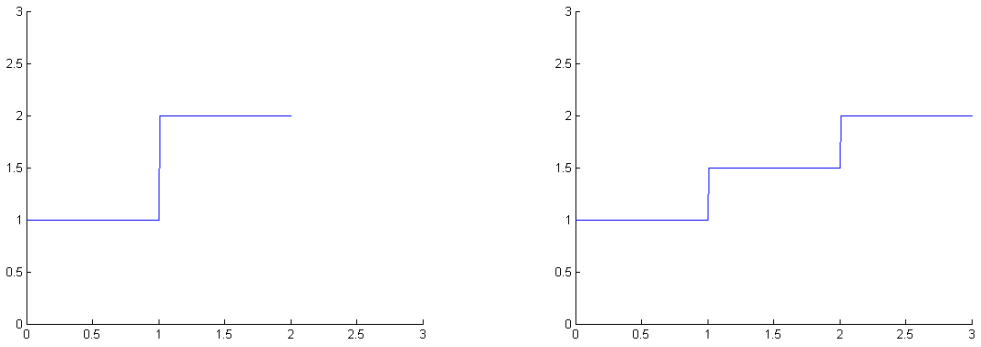
We now want to refer it to the mesh $\{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}\}$. From a mathematical viewpoint this means that we have to define a piecewise constant approximation with respect to the new mesh of the following function:

$$\tilde{u}(x) = u\left(\frac{3}{2}x\right) = \begin{cases} 1 & x \in \left[0, \frac{3}{2}\right) \\ 2 & x \in \left[\frac{3}{2}, 3\right]. \end{cases}$$

Through simple calculations the resulting vector is $[1, \frac{3}{2}, 2]'$. Figure 3.2 plots u and the piecewise approximation of $\tilde{u}$.

Let us extend the previous example to the general case. By taking into account the cell-average interpretation of the approximate vector computed through a conservative method, we associate the following piecewise function with $\{u_{\delta,j}\}_{j=j_l}^{j_r}$:

$$u_\delta(x) := \sum_{x_j \in \mathcal{T}_h} u_{\delta,j} \chi_{[x_j - \frac{h}{2}, x_j + \frac{h}{2})}(x).$$



(a) plot of the piecewise function associated with the initial vector

(b) plot of the piecewise function associated with the mapped vector

Figure 3.2: a simple example of the procedure: given the vector $[1,\ 2]'$ associated with the mesh $\{\frac{1}{2},\ \frac{3}{2}\}$, we want to refer it to the mesh $\{\frac{1}{2},\ \frac{3}{2},\ \frac{5}{2}\}$. Through the algorithm proposed in this section the result is $[1,\ \frac{3}{2},\ 2]'$.

Then we define the new vector $\{\tilde{u}_{\delta,j}\}_{j=j_l}^{j_r}$ as:

$$\tilde{u}_{\delta,j} = \frac{1}{h}\int_{x_j-\frac{h}{2}}^{x_j+\frac{h}{2}} u_\delta\left(x_l + \frac{x_r - x_l}{b-a}(x-a)\right)dx.$$

The opposite case (in which we have an approximate vector defined on all cells and we look for an approximate vector on a given subset) is analogous. In the following the formulas for both cases are derived under the hypothesis that the spatial mesh is fixed[21].

Case 1: $\{u_{\delta,j}\}_{j=j_l}^{j_r} \Rightarrow \{\tilde{u}_{\delta,j}\}_{j=1}^{\mathcal{N}}$

Let $\rho = \frac{x_r-x_l}{b-a} < 1$, $h^\star = h\rho$, $x_j^\star = x_l + h^\star\left(j-\frac{1}{2}\right)$. Furthermore let $x_{j^\star}$ be such that $x_{j^\star} \le x_j^\star \le x_{j^\star+1}$. Using the fact that $(x_j^\star - \frac{h^\star}{2}, x_j^\star + \frac{h^\star}{2}) \subset (x_j^\star - \frac{h}{2}, x_j^\star + \frac{h}{2})$ we can easily verify that:

$$\tilde{u}_{\delta,j} = \frac{1}{h^\star}\left(\mathcal{L}_1 u_{\delta,j^\star} + \mathcal{L}_2 u_{\delta,j^\star+1}\right) \tag{3.2.14}$$

where $\mathcal{L}_1 = \min\left\{h^\star, \left(x_{j^\star} + \frac{h}{2} - x_j^\star + \frac{h^\star}{2}\right)^+\right\}$ and $\mathcal{L}_2 = \min\left\{h^\star, \left(x_j^\star + \frac{h^\star}{2} - x_{j^\star} - \frac{h}{2}\right)^+\right\}$.

Case 2: $\{\tilde{u}_{\delta,j}\}_{j=1}^{\mathcal{N}} \Rightarrow \{u_{\delta,j}\}_{j=j_l}^{j_r}$

Let $\rho = \frac{b-a}{x_r-x_l} > 1$ and $h^\star = \rho h$, $x' = a + \rho\left(x_j - \frac{h}{2} - x_l\right)$, $x'' = a + \rho\left(x_j + \frac{h}{2} - x_l\right)$, we define $j^\star$ such that $x_{j^\star} - \frac{h}{2} < x' < x_{j^\star} + \frac{h}{2}$ and $j^{\star\star}$ such that $x_{j^{\star\star}} - \frac{h}{2} < x'' < x_{j^{\star\star}} + \frac{h}{2}$. Then it is immediate to verify that:

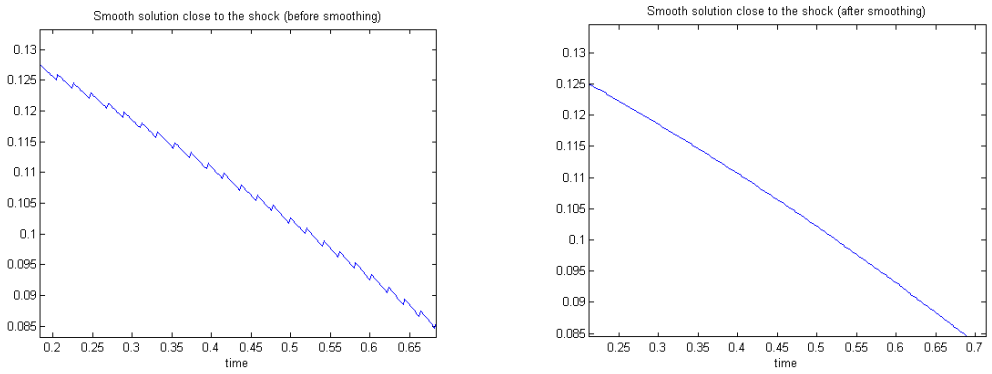
$$u_{\delta,j} = \frac{1}{h^\star}\left((x_{j^\star} + \frac{h}{2} - x')\tilde{u}_{\delta,j^\star} + h\sum_{l=j^\star+1}^{j^{\star\star}-1}\tilde{u}_{\delta,l} + (x_{j^{\star\star}} + \frac{h}{2} - x'')\tilde{u}_{\delta,j^{\star\star}}\right) \tag{3.2.15}$$

3.2.4 Smoothing

Even for an ideal smooth-jump decomposition algorithm, the reconstruction of the solution close to the shock is a difficult task. In particular as Figure 3.3(a) shows, we observe that

[21] The formulas below can be extended to the more general case in which the approximate vector must be defined with respect to a new mesh.

the trend of the solution $u(\mu_j, t, x_{shock}(\mu_j, t)^{\pm})$ may be very irregular. This phenomenon affects the quality of the model order reduction in time because it adds a sort of spurious variability close to the shock. This is why a smooth filter should be applied to reduce the numerical fluctuations. In this work we use the tool `smooth` provided in `Matlab`[22]. Figure 3.3(b) shows the improved result.



(a) Solution before the smoothing

(b) Solution after the smoothing

Figure 3.3: application of the smoothing algorithm

3.3 Main features of the methodology

In this section we introduce the RB method for parametric scalar conservation laws in one space dimension. In order to simplify the methodology we suppose that *the solution presents only one shock that starts at* $t = 0$. In formulas:

$$\begin{cases} \dfrac{\partial}{\partial t}u + \dfrac{\partial}{\partial x}f(u,\boldsymbol{\mu}) = 0 & (t,x) \in Q_{T_{max}} = (0,T_{max}) \times (a,b) \\ u(t,a^-) = \xi_0(\boldsymbol{\mu},t), \quad u(t,b^+) = \xi_1(\boldsymbol{\mu},t) & t \in (0,T_{max}) \\ u(\boldsymbol{\mu},0,x) = u_0(\boldsymbol{\mu},x) & x \in (a,b), \end{cases} \tag{3.3.1a}$$

where

$$\begin{cases} \xi_0 : \mathcal{D} \times (0,T_{max}) \to \mathbb{R} \\ \xi_1 : \mathcal{D} \times (0,T_{max}) \to \mathbb{R} \\ u_0(\boldsymbol{\mu},x) = \begin{cases} u_{0,left}(\boldsymbol{\mu},x) & x \in (a, x_{shock}(\boldsymbol{\mu},0)) \\ u_{0,right}(\boldsymbol{\mu},x) & x \in (x_{shock}(\boldsymbol{\mu},0), b). \end{cases} \end{cases} \tag{3.3.1b}$$

We have supposed that both $x = a$ and $x = b$ are inflow boundaries in the sense of (3.2.5). For this reason we have two boundary conditions. A motivation for this choice is provided in the following remark.

Remark 3.1. *We motivate the case study by considering the following example. Let us consider a highway modelled by a scalar conservation law where the solution* u *represents*

[22] This algorithm is simply based on a moving average smoothing, i.e., given $\mathbf{y} \in \mathbb{R}^N$, $\mathbf{y}_{smooth} \in \mathbb{R}^N$ is defined as:

$$y_{smooth,j} = \frac{y_{j-2} + y_{j-1} + y_j + y_{j+1} + y_{j+2}}{5}.$$

the density of the cars. The highway starts at $x = 0$ and ends at $x = 1$ where it enters the town. If the city is congested (i.e., the density of cars in the city is higher than the density of cars in the highway), we may have that $x = 1$ is an inflow boundary: in practice it determines the generation of a backward wave that propagates inside the highway (the traffic congestion originally confined into the urban area starts to influence also the highway).

In order to follow the strategy outlined in the introduction, we aim at providing:

1. a suitable formulation that splits the initial problem into two subproblems that provide the two smooth components of the solution;

2. an inexpensive (i.e., independent of the underlined spatial mesh) algorithm able to compute the shock equation and the value of the solution on both the sides of the shock curve.

3. an inexpensive algorithm to compute the two smooth solutions.

The section is organized as follows: first we introduce the so-called special formulation; then we propose an algorithm to capture the shock. After that we focus on the offline-online decomposition and on the sampling strategy. Finally we briefly describe how to efficiently solve an input-output relationship based on the solution of a scalar conservation law.

3.3.1 Rankine-Hugoniot condition and the "special" formulation

In this subsection we introduce a new formulation- here called "special" formulation in order to highlight the link with the special BV functions- that is the basis for the RB approach here presented. Starting from the definition of integral solution for the problem (3.1.1)and by making some assumption on the structure of the solution we derive a stronger piecewise formulation. Then, through a simple change of variable, we obtain the final formulation that is ideally suited to the RB context.

Function $u \in C(0, T_{max}; SBV(a,b))$ is defined as the entropic weak solution to (3.1.1) if for all $v \in C_0^\infty(\mathbb{R} \times (a,b))$, it holds:

$$\begin{cases} \int_{Q_{T_{max}}} \left[u \dfrac{\partial v}{\partial t} + f(u) \dfrac{\partial v}{\partial x} \right] dxdt + \int_a^b \left[u(0,x) - u_0(0,x) \right] v(0,x)\, dx = 0, \\ u(t,a) \text{ satisfies } \max_{k \in [u(t,a), \xi_0(t)]} \left\{ \text{sign}\left(u(t,a) - \xi_0(t) \right) \left[f(u(t,a)) - k \right] \right\} = 0, \\ u(t,b) \text{ satisfies } \max_{k \in [u(t,b), \xi_1(t)]} \left\{ \text{sign}\left(u(t,b) - \xi_1(t) \right) \left[f(u(t,b)) - k \right] \right\} = 0, \\ \exists\, E > 0 \text{ such that } \forall\, x, y \in \mathbb{R}\, (x, x+y) \in (a,b)^2 \text{ we have: } u(t, x+y) - u(t,x) \le \frac{E}{t} y. \end{cases} \tag{3.3.2}$$

Let us assume that u has only one shock described by the curve $x_{shock}(t)$ that starts at time $t = 0$. Moreover, we assume that $x_{shock} \in Lip(0, T_{max})$. In our applications the hypothesis is not particularly strict and in Appendix A we discuss how the hypothesis can be relaxed. Therefore the solution is smooth in $\Omega_1 = \{(t,x) \in (0, T_{max}) \times (a,b) : x < x_{shock}(t)\}$ and

$\Omega_2 = \{(t,x) \in (0,T_{max}) \times (a,b) : x > x_{shock}(t)\}$ and solves the following problems:

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial}{\partial x} f(u) = 0 & \text{q.o. in } \Omega_1 \\ \dfrac{\partial u}{\partial t} + \dfrac{\partial}{\partial x} f(u) = 0 & \text{q.o. in } \Omega_2 \\ u(t,a) = \xi_0(t) \quad u(t,b) = \xi_1(t) & t \in (0,T_{max}) \\ u(0,x) = u_0(x) & x \in (a, x_{shock}(0)) \cup (x_{shock}(0), b) \\ \dfrac{f(u(t,x_{shock}(t)^+)) - f(u(t,x_{shock}(t)^-))}{u(t,x_{shock}(t)^+) - u(t,x_{shock}(t)^-)} = \dot{x}_{shock}(t) & t \in (0,T_{max}), \end{cases} \tag{3.3.3}$$

where the last equation coincides with the Rankine-Hugoniot condition. Thanks to the regularity of the solution in each subdomain, by defining

$$\begin{cases} u_1(t,x) := u\left(t, a + \dfrac{x_{shock}(t) - a}{b-a}(x-a)\right) \\ u_2(t,x) := u\left(t, x_{shock}(t) + \dfrac{b - x_{shock}(t)}{b-a}(x-a)\right), \end{cases} \tag{3.3.4}$$

we have that the couple (u_1, u_2) is the only solution in $\left[W^{1,1}(Q_{T_{max}})\right]^2$ to the following problem

$$\begin{cases} \displaystyle\int_{Q_{T_{max}}} \left(u_1 \frac{\partial v}{\partial t} + \frac{b-a}{x_{shock}(t) - a} f(u_1) \frac{\partial v}{\partial y}\right) \frac{x_{shock}(t) - a}{b-a}\, dydt = 0 \quad \forall\, v \in C_0^\infty(Q_{T_{max}}) \\ \dfrac{f(u_2(t,a^-)) - f(u_1(t,b^+))}{u_2(t,a^-) - u_1(t,b^+)} = \dot{x}_{shock}(t) \quad \text{q.o. in } (0,T_{max}) \\ \displaystyle\int_{Q_{T_{max}}} \left(u_2 \frac{\partial v}{\partial t} + \frac{b-a}{b - x_{shock}(t)} f(u_2) \frac{\partial v}{\partial y}\right) \frac{b - x_{shock}(t)}{b-a}\, dydt = 0 \quad \forall\, v \in C_0^\infty(Q_{T_{max}}) \\ u_1(t,a) = \xi_0(t) \quad u_2(b,t) = \xi_1(t) \\ u_1(0,y) = u_0\left(a + \frac{x_{shock}(0)-a}{b-a}(y-a)\right) \quad u_2(0,y) := u_0\left(x_{shock}(0) + \frac{b - x_{shock}(0)}{b-a}(y-a)\right). \end{cases} \tag{3.3.5}$$

Figure 3.4 shows the domain decomposition induced by the given shock $x_{shock}(t)$.

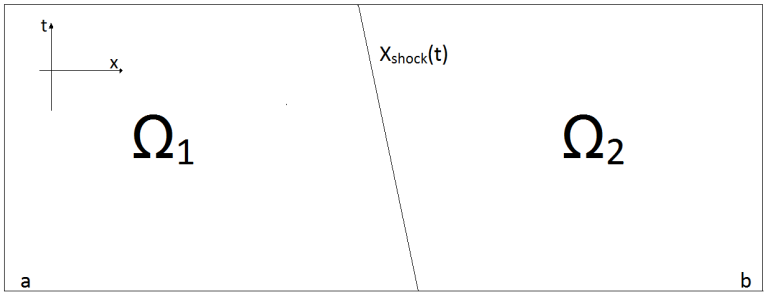


Figure 3.4: domain decomposition induced by the shock.

The fact that (u_1, u_2) solves (3.3.5) descends directly from the hypothesis on the structure of u; concerning the unicity of the solution in the space $\left[W^{1,1}(Q_{T_{max}})\right]^2$, it is a

straightforward application of the characteristic method [109]. We have to prove that the maps

$$\boldsymbol{\tau}_1(t,x) = \begin{bmatrix} t \\ a + \dfrac{x_{shock}(t)-a}{b-a}(x-a) \end{bmatrix} \qquad \boldsymbol{\tau}_2(t,x) = \begin{bmatrix} t \\ x_{shock}(t) + \dfrac{b-x_{shock}(t)}{b-a}(x-a) \end{bmatrix}$$

are regular enough to be a change of coordinates. But it is an obvious implication of the fact that $x_{shock} \in Lip(0,T_{max})$.

Furthermore, we observe that the formulation (3.3.5) is redundant: in fact the shock curve is an outflow boundary for both problems (for entropic solutions the characteristics cannot hit a shock backward in time); so it is possible to consider the two following, completely decoupled, problems:

$$\begin{cases} \displaystyle\int_{Q_{T_{max}}} \left(u_1 \frac{\partial v}{\partial t} + \frac{b-a}{x_{shock}(t)-a} f(u_1) \frac{\partial v}{\partial y} \right) \frac{x_{shock}(t)-a}{b-a}\, dydt = 0 \quad \forall\, v \in C_0^\infty(Q_{T_{max}}) \\ u_1(t,a) = \xi_0(t) \\ u_1(0,y) = u_0(a + \frac{x_{shock}(0)-a}{b-a} y) \end{cases} \tag{3.3.6a}$$

$$\begin{cases} \displaystyle\int_{Q_{T_{max}}} \left(u_2 \frac{\partial v}{\partial t} + \frac{b-a}{b-x_{shock}(t)} f(u_2) \frac{\partial v}{\partial y} \right) \frac{b-x_{shock}(t)}{b-a}\, dydt = 0 \quad \forall\, v \in C_0^\infty(Q_{T_{max}}) \\ u_2(t,b) = \xi_1(t) \\ u_2(0,y) = u_0(x_{shock}(0) + \frac{b-x_{shock}(0)}{b-a}(y-a)). \end{cases} \tag{3.3.6b}$$

Remark 3.2. *This formulation guarantees the splitting of the two smooth problems in an effective way. Due to the fact that the solutions to the new problems are smooth for construction, we expect that the entire manifold could be approximated through linear combinations of a small number of its properly selected members.*

Remark 3.3. *This formulation simplifies the a posteriori error estimation. In fact if we assume that the shock equation is reconstructed correctly, then we can link the error to the solutions of the smooth problems and so taking advantage of the additional regularity.*

3.3.2 Shock capturing algorithm

Let us consider problem (3.3.1). Denoting with $x(t)$ the equation of the characteristic starting from $\xi \in (a,b)$, $z(t) = u(t,x(t))$ and assuming enough regularity, we have that:

$$\begin{cases} \dot{z}(t) = \frac{\partial u}{\partial t}(t,x(t)) + \frac{\partial u}{\partial x}(t,x(t))\dot{x}(t) \\ \frac{\partial u}{\partial t}(t,x(t)) + f'(z(t))\frac{\partial u}{\partial x}(t,x(t)) = 0. \end{cases}$$

From the equations above it is immediate to deduce that:

$$\begin{cases} \dot{x}(t) = f'(z(t)) \quad t>0 \\ x(0) = \xi \end{cases} \Rightarrow x(t) = \xi + \int_0^t f'(z(\tau))\, d\tau, \tag{3.3.7}$$

and

$$\begin{cases} \dot{z}(t) = 0 \quad t > 0 \\ z(0) = u_0(\xi). \end{cases} \Rightarrow z(t) = u_0\left(x(t) - \int_0^t f'(z(\tau))\, d\tau\right) = u_0\left(x(t) - f'(z(t))t\right) \tag{3.3.8}$$

Let us consider the shock curve $x_{shock}(t)$. Suppose that we know $x_{shock}(t^k)$, $u^-(t^k)$ and $u^+(t^k)$. First, we can approximate $x_{shock}(t^{k+1})$ through an explicit discretization of the Rankine-Hugoniot condition:

$$x_{shock}(t^{k+1}) = x_{shock}(t^k) + \Delta t \frac{f(u^+(t^k)) - f(u^-(t^k))}{u^+(t^k) - u^-(t^k)}.$$

Then, we can take advantage of formula (3.3.8) to compute $u^+(t^{k+1})$ and $u^-(t^{k+1})$. Thanks to the smoothness of the solution on both sides of the shock curve and thanks to the fact that $u^+(t^{k+1})$ and $u^-(t^{k+1})$ are close to $u^+(t^k)$ and $u^-(t^k)$, respectively, we expect that the Newton Raphson method[23] is very efficient to solve the equations.

The shock decomposition procedure is listed in Algorithm 10. In order to simplify the notation we denote with $u_{0,left}$ and $u_{0,right}$ the two smooth components of the initial condition.

Algorithm 10 Shock capturing procedure

Given $x_{shock}(t^0)$, $u^+(t^0)$ and $u^-(t^0)$.

$f^{left}_{x,t}(u) := u - u_{0,left}(x - f'(u)t)$

$f^{right}_{x,t}(u) := u - u_{0,right}(x - f'(u)t)$

for $k = 0, \cdots, \mathbb{K} - 1$ **do**

Rankine-Hugoniot condition

$$x_{shock}(t^{k+1}) = x_{shock}(t^k) + \Delta t \frac{f(u^+(t^k)) - f(u^-(t^k))}{f(u^+(t^k)) - f(u^-(t^k))}$$

Newton-Raphson method

Iterations for the left value

$u^-(t^{k+1})$= `NetwonRaphson Algorithm`$(u^-(t^k), f^{left}_{x_{shock}(t^{k+1}),t^{k+1}})$

Iterations for the right value

$u^+(t^{k+1})$= `NetwonRaphson Algorithm`$(u^-(t^k), f^{right}_{x_{shock}(t^{k+1}),t^{k+1}})$

end for

The extension to the general case in which we have the source term is much more difficult. In fact following the same procedure as before, we obtain (let g be the source

[23] The iterative method, see [98], is based on the following iterative assignment:

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)} \quad f \in C^1 \quad f'(x^k) \neq 0$$

As far as the termination condition is concerned, several options are available. In our code we always consider $|f(x^k)| < \epsilon$, with $\epsilon > 0$ a given tolerance.

term):

$$\begin{cases} x_{shock}(t^{k+1}) = \xi + \int_0^{t^{k+1}} f'(z(\tau))\,d\tau \\ z(t^{k+1}) = u_0\left(x_{shock}(t^{k+1}) - \int_0^t f'(z(\tau))\,d\tau\right) + \int_0^{t^{k+1}} g(z(\tau), x(\tau), \tau)\,d\tau \\ \dot{x}_{shock}(t) = \dfrac{f(u^+(t)) - f(u^-(t))}{u^+(t) - u^-(t)} \end{cases} \tag{3.3.9}$$

The formula is fully implicit and difficult to be treated in a offline-online framework. For this reason, in the present work no strategy has been developed to deal with this general case.

3.3.3 A RB approach for the smooth problems: offline-online decomposition

Given an at least continuous function $f : Q_{T_{max}} \to \mathbb{R}$, two are the main ingredients required by an interpolation procedure that aims at reconstructing an approximation of f starting from its knowledge in some points in $Q_{T_{max}}$:

- a suitable interpolatory basis $\{q_k\}_{k=1}^{N_{RB}}$, $q_k : Q_{T_{max}} \to \mathbb{R}$, $k = 1, \cdots, N_{RB}$ and the corresponding interpolation points (t_k, x_k);
- a (possibly efficient) methodology to compute $f(t_k, x_k)$ in the preselected points.

In our particular context the function to be interpolated is $u(\boldsymbol{\mu})$, where $\boldsymbol{\mu} \in \mathcal{D}$ is a given parameter.

The so-called Empirical Interpolation Method described in section 1.8 ([7, 32]) provides the tools to define the interpolatory basis and the points where the solution has to be precomputed.

In order to compute the solution in (t_k, x_k), we can take advantage of the notable characteristic formula that, in the setting of problem (3.3.1), is

$$u(\boldsymbol{\mu}, t, x) = \Lambda(\boldsymbol{\mu}, x - f'(\boldsymbol{\mu}, u(\boldsymbol{\mu}, t, x))t) \tag{3.3.10}$$

where -depending on (t, x)- Λ could be u_0, ξ_0 or ξ_1. Thanks to the above equation, we have the possibility to compute the solution in a single point without knowing the global solution.

We observe that the interpolation procedure must be applied to at least continuous functions defined onto a $\boldsymbol{\mu}$-independent domain. So it is first necessary to split the solution into its two smooth components, by reporting them to a $\boldsymbol{\mu}$-independent configuration and then to apply the interpolation method. On the other hand, the characteristic equation is referred to the original solution.

Let us describe the entire method.

Offline stage: Let us suppose that $\{u(\boldsymbol{\mu}_j)\}_{j=1}^{N_{RB}}$- solution to problem (3.3.1) for $\boldsymbol{\mu} = \boldsymbol{\mu}_j$, $j = 1, \cdots, N_{RB}$ - are given[24].

First of all we decompose the solution in the smooth left and right parts - by employing Algorithm 9 - and we build u_1 and u_2 as defined in (3.3.4), i.e.:

$$u_1(\boldsymbol{\mu}, t, x) = u\left(\boldsymbol{\mu}, t, a + \frac{x_{shock}(\boldsymbol{\mu}, t) - a}{b - a}(x - a)\right) \quad u_2(\boldsymbol{\mu}, t, x) = u\left(\boldsymbol{\mu}, t, x_{shock} + \frac{b - x_{shock}}{b - a}(x - a)\right).$$

[24] In the next subsection we will deal with the sampling strategy.

Due to our choice to use conservative methods to solve the equations, we have that $u_j^n \simeq \frac{1}{h}\int_{x_j-\frac{h}{2}}^{x_j+\frac{h}{2}} u(t^n,x)\,dx$, for this reason $\{u_{1,j}^k\}$ and $\{u_{2,j}^k\}$ must be defined through the formula explained in section 3.2.3.

At this point we can obtain the empirical basis and the interpolation points -the so-called *magic points*- using the procedure proposed in [7] as second step of the above mentioned EIM:

Algorithm 11 Empirical Interpolation Method Part II: $\{\{q_k(\cdot) : 1 \le k \le M_{Max}\}, \{(x_j,t_j) : 1 \le j \le M_{Max}\}, \{B_{i,j} : 1 \le i,j \le M_{Max}\}\} = \texttt{EIM}\{u_l(\boldsymbol{\mu},x_j,t_j) : 1 \le j \le M_{Max}\}$

$(t_1,x_1) = \arg\operatorname{ess\,sup}_{(t,x)\in(0,T_{max})\times(a,b)} \|\xi_1(t,x)\|$, $q_1 = \frac{\xi_1}{\xi_1(t_1,x_1)}$, $B_{11} = 1$
for $M = 2 : M_{max}$ **do**
  find $\sigma \in \mathbb{R}^{M-1}$: $\sum_{j=1}^{M-1} \sigma_j q_j(t_i,x_i) = \xi_M(t_i,x_i)$ for $1 \le i \le M-1$
  $r_M(t,x) = \xi_M(t,x) - \sum_{j=1}^{M-1} \sigma_j q_j(t,x)$
  $(t_M,x_M) = \arg\operatorname{ess\,sup}_{(t,x)\in(0,T_{max})\times(a,b)} \|r_M(t,x)\|$,
  $q_M(t,x) = \frac{r_M(t,x)}{r_M(t_M,x_M)}$, $B_{i,j}^M = q_j(t_i,x_i)$
end for

In the following we refer to (t_j^l,x_j^l) , $\{q_j^l(\cdot,\cdot)\}_j$ and B^l, for $l = 1,2$, to indicate the magic points, the empirical bases and the interpolation matrices computed for u_1 and u_2, respectively.

Online stage: First, we define $x_{shock}(t^k)$, for $k = 1,\cdots,\mathbb{K}$, through the shock capturing Algorithm 10.

Then we refer the magic points to the actual configuration, i.e.:

$$\begin{cases} (t_j^1,\tilde{x}_j^1) = \left(t_j^1, a + \frac{x_{shock}(\boldsymbol{\mu},t_j^1)-a}{b-a}(x_j^1-a)\right) & \text{for } u_1 \\ (t_j^2,\tilde{x}_j^2) = \left(t_j^2, x_{shock}(\boldsymbol{\mu},t_j^2) + \frac{b-x_{shock}(\boldsymbol{\mu},t_j^2)}{b-a}(x_j^2-a)\right) & \text{for } u_2. \end{cases} \tag{3.3.11}$$

Then we compute the solution in these points through (3.3.10). In order to do that, it is necessary to properly define Λ in (3.3.10). It is straightforward to observe that for u_1:

$$\Lambda(\boldsymbol{\mu},\cdot) = \begin{cases} \xi_0(\boldsymbol{\mu},\cdot) & x < a + f'(\boldsymbol{\mu},u_0(\boldsymbol{\mu},a))t \\ u_0(\boldsymbol{\mu},\cdot) & x > a + f'(\boldsymbol{\mu},u_0(\boldsymbol{\mu},a))t \end{cases} \tag{3.3.12}$$

while for u_2:

$$\Lambda(\boldsymbol{\mu},\cdot) = \begin{cases} u_0(\boldsymbol{\mu},\cdot) & x < b + f'(\boldsymbol{\mu},u_0(\boldsymbol{\mu},b))t \\ \xi_1(\boldsymbol{\mu},\cdot) & x > b + f'(\boldsymbol{\mu},u_0(\boldsymbol{\mu},b))t. \end{cases} \tag{3.3.13}$$

The shock data does not enter in the formulas due to the well-known fact that, for entropic solutions, the characteristics cannot hit a shock backward in time.

In order to apply the Newton-Raphson algorithm, we need to properly initialize it: an a priori estimate of the solution can be obtained through the simple strategy explained in Algorithm 12.

After these steps, we can compute $U_j^1(\boldsymbol{\mu}) = u_1^{RB}(\boldsymbol{\mu},x_j^1,t_j^1)$ and $U_j^2(\boldsymbol{\mu}) = u_2^{RB}(\boldsymbol{\mu},x_j^2,t_j^2)$ through the Newton Raphson algorithm and finally the interpolation coefficients by solving the linear systems:

$$B^1\Theta_{u^1}(\boldsymbol{\mu}) = U^1(\boldsymbol{\mu}), \quad B^2\Theta_{u^2}(\boldsymbol{\mu}) = U^2(\boldsymbol{\mu}).$$

In conclusion, the approximation for u_1 and u_2 are built as:

$$\begin{cases} u_1^{RB}(\boldsymbol{\mu}) = \sum_{j=1}^{N_{RB}} \left(\Theta_{u^1}(\boldsymbol{\mu})\right)_j q_j^1, \\ u_2^{RB}(\boldsymbol{\mu}) = \sum_{j=1}^{N_{RB}} \left(\Theta_{u^2}(\boldsymbol{\mu})\right)_j q_j^2, \end{cases} \tag{3.3.14}$$

respectively.

In the next subsection we discuss how to choose the sample set $\{u(\boldsymbol{\mu}_j) : j = 1, \cdots, N\}$.

Algorithm 12 Algorithm for the definition of the initial guess for the Newton Raphson iterative method

---

Offline stage

Compute $\bar{x}_{shock} = \frac{1}{|\mathcal{D}|} \int_{\mathcal{D}} x_{shock}(\boldsymbol{\mu}, 0)\, d\boldsymbol{\mu}$

For each $\boldsymbol{\mu}$-independent term of the expansion of $\xi_0(\boldsymbol{\mu}, \cdot)$ compute $\bar{\xi}_0^q = \dfrac{1}{T} \displaystyle\int_0^T \xi_0^q(t)\, dt$.

For each $\boldsymbol{\mu}$-independent term of the expansion of $\xi_1(\boldsymbol{\mu}, \cdot)$ compute $\bar{\xi}_1^q = \dfrac{1}{T} \displaystyle\int_0^T \xi_1^q(t)\, dt$.

For each $\boldsymbol{\mu}$-independent term of the expansion of $u_{0,left}(\boldsymbol{\mu}, \cdot)$ compute $\bar{u}_{0,left}^q = \dfrac{1}{\bar{x}_{shock} - a} \displaystyle\int_a^{\bar{x}_{shock}} u_{0,left}^q(x)\, dx$.

For each $\boldsymbol{\mu}$-independent term of the expansion of $u_{0,right}(\boldsymbol{\mu}, \cdot)$ compute $\bar{u}_{0,right}^q = \dfrac{1}{b - \bar{x}_{shock}} \displaystyle\int_{\bar{x}_{shock}}^b u_{0,right}^q(x)\, dx$.

Online stage

For each $(t_j^1, \tilde{x}_j^1)$, $j = 1, \cdots, N$

if $\tilde{x}_j^1 < a + f'(\boldsymbol{\mu}, u_0(\boldsymbol{\mu}, a))t_j^1$ **then**

$$u_{start,NR} = \xi_0\left(\boldsymbol{\mu}, \tilde{x}_j^1 - f'\left(\boldsymbol{\mu}, \sum_{q=1}^{Q_{\xi,0}} \Theta_{\xi,0}^q(\boldsymbol{\mu}) \bar{\xi}_0^q\right) t_j^1\right)$$

else

$$u_{start,NR} = u_0\left(\boldsymbol{\mu}, \tilde{x}_j^1 - f'\left(\boldsymbol{\mu}, \sum_{q=1}^{Q_{u,left}} \Theta_{u,left}^q(\boldsymbol{\mu}) u_{0,left}^q\right) t_j^1\right)$$

end if

For each $(t_j^2, \tilde{x}_j^2)$, $j = 1, \cdots, N$

if $\tilde{x}_j^2 < b + f'(\boldsymbol{\mu}, u_0(\boldsymbol{\mu}, b))t_j^2$ **then**

$$u_{start,NR} = u_0\left(\boldsymbol{\mu}, \tilde{x}_j^2 - f'\left(\boldsymbol{\mu}, \sum_{q=1}^{Q_{u,right}} \Theta_{u,right}^q(\boldsymbol{\mu}) u_{0,right}^q\right) t_j^2\right)$$

else

$$u_{start,NR} = \xi_1\left(\boldsymbol{\mu}, \tilde{x}_j^2 - f'\left(\boldsymbol{\mu}, \sum_{q=1}^{Q_{\xi,1}} \Theta_{\xi,1}^q(\boldsymbol{\mu}) \bar{\xi}_1^q\right) t_j^2\right)$$

end if

3.3.4 Sampling strategy

In the Empirical Interpolation method, after defining a suitable fine parameter sample $\Xi_g \subset \mathcal{D}$, the parameters $\boldsymbol{\mu}_j$ -and thus the functions $g(\boldsymbol{\mu}_j, \cdot)$- are chosen through a Greedy algorithm. In our case, if we dispose of a rigorous and reliable a posteriori error estimator such that $\Delta_N(\boldsymbol{\mu}) \geq \|u_{RB,N}(\boldsymbol{\mu}) - u(\boldsymbol{\mu})\|_\star$, where $\|\cdot\|_\star$ is a suitable norm, it is possible to choose the next $\boldsymbol{\mu}_{N+1}$ such that:

$$\boldsymbol{\mu}_{N+1} := \arg\max_{\boldsymbol{\mu}\in\Xi_g} \Delta_N(\boldsymbol{\mu}). \tag{3.3.15}$$

In section 3.4 we discuss about the definition of an a posteriori error estimator for linear (and nonlinear) hyperbolic problems. At the present time as far as we know, no inexpensive and rigorous a posteriori error estimators have been developed for nonlinear hyperbolic equations. For this reason, in all our numerical simulations, we choose equispaced $\boldsymbol{\mu}_j$.

3.3.5 Input-output relationships

As explained in the first chapter, Reduced Bases can provide significant speed-ups in the computation of input-output relationships depending on the solution of a parametrized equation.

In order to explain the methodology, let us consider the following example:

$$s(\boldsymbol{\mu}) = Lu(\boldsymbol{\mu}) = \int_0^T \int_{(a,b)} w(t,x)u(\boldsymbol{\mu},t,x)\,dxdt, \tag{3.3.16}$$

where $u(\boldsymbol{\mu})$ is the solution to (3.3.1).

Some work is necessary to make the computation of $s(\boldsymbol{\mu})$ independent of the spatial mesh. By recalling the definition of $u_1(\boldsymbol{\mu})$ and $u_2(\boldsymbol{\mu})$ in (3.3.4), we have:

$$\begin{aligned}
Lu(\boldsymbol{\mu}) &= \int_0^T \int_{(a,b)} w(t,x)u(\boldsymbol{\mu},t,x)\,dxdt \\
&= \int_0^T \left(\int_{(a,x_{shock}(\boldsymbol{\mu},t))} w(t,x)u(\boldsymbol{\mu},t,x)\,dx + \int_{(x_{shock}(\boldsymbol{\mu},t),b)} w(t,x)u(\boldsymbol{\mu},t,x)\,dx \right) dt \\
&= \int_0^T \left(\int_{(a,b)} \frac{x_{shock}(\boldsymbol{\mu},t)-a}{b-a} w\left(t, a + \frac{x_{shock}(\boldsymbol{\mu},t)-a}{b-a}(y-a)\right) u_1(\boldsymbol{\mu},t,y)\,dy \right. \\
&\quad \left. + \int_{(a,b)} \frac{b-x_{shock}(\boldsymbol{\mu},t)}{b-a} w\left(x_{shock}(\boldsymbol{\mu},t) + \frac{b-x_{shock}(\boldsymbol{\mu},t)}{b-a}(y-a), t\right) u_2(\boldsymbol{\mu},t,x)\,dy \right) dt.
\end{aligned}$$

By applying the EIM, we obtain:

$$\frac{x_{shock}(\boldsymbol{\mu},t)-a}{b-a} w\left(t, a + \frac{x_{shock}(\boldsymbol{\mu},t)-a}{b-a}(y-a)\right) \simeq \sum_{q=1}^{Q_{w,1}} \Theta_{w,1}^q(\boldsymbol{\mu}) w_q^1(t,y) \tag{3.3.17a}$$

and

$$\frac{b-x_{shock}(\boldsymbol{\mu},t)}{b-a} w\left(t, x_{shock}(\boldsymbol{\mu},t) + \frac{b-x_{shock}(\boldsymbol{\mu},t)}{b-a}(y-a)\right) \simeq \sum_{q=1}^{Q_{w,2}} \Theta_{w,2}^q(\boldsymbol{\mu}) w_q^2(t,y). \tag{3.3.17b}$$

Therefore

$$s_{RB}(\boldsymbol{\mu}) = \sum_{q=1}^{Q_{w,1}} \sum_{i=1}^{N_{RB}} \Theta_{w,1}^q(\boldsymbol{\mu}) \Theta_{u,1}^i(\boldsymbol{\mu}) \left(w_q^1, q_i^1\right) + \sum_{q=1}^{Q_{w,2}} \sum_{i=1}^{N_{RB}} \Theta_{w,2}^q(\boldsymbol{\mu}) \Theta_{u,2}^i(\boldsymbol{\mu}) \left(w_q^2, q_i^2\right) \tag{3.3.18a}$$

where:

$$\left(w_q^l, q_i^l\right) = \int_0^T \int_{(a,b)} w_q^l(t,x) q_i^l(t,x)\, dxdt \quad l = 1,2 \tag{3.3.18b}$$

are computable offline.

This guarantees an efficient Offline-Online decomposition algorithm: the online computation is independent of the spatial mesh[25].

3.3.6 The whole algorithm

In order to clarify how the different steps are linked, we conclude the section by describing the entire procedure in a compact form. For the sake of simplicity, we not introduce the Greedy sampling strategy in the following algorithm.

Algorithm 13 Offline-Online decomposition

Offline stage

1: Compute $u(\boldsymbol{\mu}_j)$, $j = 1,\cdots,N_1$ through a truth method (section 3.2.1) and build $u_{smooth,l}(\boldsymbol{\mu}_j)$, $l = 1,2$ through Algorithm 8 (`smooth jump decomposition algorithm`).

2: Through Algorithm 11 compute the magic points (x_j^l, t_j^l), the empirical bases $\{q_j^l(\cdot,\cdot)\}_j$ and the interpolatory matrices $B_{i,j}^l$, with $i,j = 1,\cdots,N$ and $l = 1,2$.

3: Perform the offline part of Algorithm 12 to compute the initial starting points of the Newton Raphson algorithm.

Online stage

1: Compute $x_{shock}(t^k)$, $k = 1,\cdots,\mathbb{K}$.

2: Define the magic points with respect to the actual configuration through formula (3.3.11).

3: Compute the initial starting points for the Newton Raphson method through Algorithm 12.

4: Apply the Newton Raphson method.

5: Compute the interpolation coefficients.

[25] The temporal mesh influences the algorithm only during the shock capturing algorithm.

3.4 A posteriori error estimation: some preliminary comments

In this section we briefly address the issue of the definition of an inexpensive and reliable a posteriori estimator -as the one presented in the introductive chapter - for the problem[26]

$$\begin{cases} \frac{\partial}{\partial t}u + \frac{\partial}{\partial x}f(u) = 0 & (t,x) \in (0,T)\times\mathbb{R} \\ u(0) = u_0 & x \in \mathbb{R}. \end{cases} \tag{3.4.1}$$

As initial step, we briefly recall the most important results from the literature[27].

Even if our focus is the parametric context, in this section we omit the dependence on the parameters in order to simplify the notation.

Historically, a posteriori error estimators for hyperbolic problems have been developed and deeply analysed in the context of adaptive mesh refinement (AMR). Due to the complexity of the problem, several estimators are motivated only by some numerical evidence but they do not guarantee any rigorous bound. For this reason we distinguish between *error indicators* (i.e., quantities for which no formal result exists) and *error estimators* (i.e., quantities which rigorously bound the approximation error). In the first category, we recall (see [67]) the so-called feature indicators ([27]), the reconstruction based indicators, ([60]) and the residual based indicators ([52, 53, 61]).

As far as rigorous a posteriori error bounds are concerned, we recall the work by Gosse and Makridakis, ([44]) - that, starting from Kruzkov-type estimates ([9]), provides a result for scalar conservation laws in one space dimension for E-schemes[28]- and the work by Kröner and Ohlberger, [64] - that, taking advantage of Kuznetsov's theory, provides an a posteriori estimate for multidimensional scalar conservation laws. Finally, we refer to [54] for an error estimator based on a duality technique.

In our opinion, it is very difficult to apply these estimators to our framework[29].

- Error indicators would not guarantee the reliability of the Reduced Basis approximation.
- The error estimate proposed by Gosse and Makridakis is based on an entropy inequality of the form: for each $k \in \mathbb{R}$, the numerical solution u_δ ($\delta = (\Delta t, h)$) must satisfy the following estimate:

$$\frac{\partial}{\partial t}|u_\delta - k| + \frac{\partial}{\partial x}\left(\text{sign}(u_\delta - k)\right)\left[f(u_\delta) - f(k)\right] \leq \frac{\partial}{\partial t}G_k + \frac{\partial}{\partial x}H_k + K_k + \frac{\partial^2}{\partial x^2}L_k \quad \text{in } \mathcal{D}'(\mathbb{R}), \tag{3.4.2a}$$

[26] For simplicity we consider the Cauchy problem.

[27] We refer to [67] for a survey on a posteriori error estimators and indicators for hyperbolic problems.

[28] A numerical method of the form (3.2.1) with $p = 0$, $q = 1$ is called E-scheme if it satisfies the inequality:

$$\text{sign}(w - v)\left(F(v,w) - f(q)\right) \leq 0$$

for all q between w and v. In particular Godunov's method, Lax Friedrichs method and Engquist Osher method are all E-scheme. Osher ([90]) also shows that E-schemes are at most first order accurate.

[29] This is not surprising: error estimators in the context of adaptive mesh refinement are used as criterion to refine the mesh or not. For this reason they must be based on local estimates. On the other hand, in the context of RB we aim at inexpensive a posteriori estimators and so we prefer estimators that do not link the error to a specific region of the domain but are potentially computable in a offline-online framework.

where there exist α_G, α_H, α_K and α_L positive local Radon measures[30] such that:

$$\begin{array}{ll} |G_k| \leq \alpha_G, & |H_k| \leq \alpha_H, \\ |K_k| \leq \alpha_K, & |L_k| \leq \alpha_L, \end{array} \tag{3.4.2b}$$

uniformly with respect to k.

In the Reduced basis context it seems to be extremely difficult to guarantee, on one hand, some independence of the estimator from the underlined mesh and, on the other hand, to prove the cell-based estimate (3.4.2a). Also in [64] a discrete entropy inequality is demanded, thus, even in this case, the application to the RB framework seems to be extremely problematic.

- In order to apply the approach proposed in [54], it is necessary to solve a dual problem that depends, formally, on the exact solution and so must be approximated[31]. A natural choice is to replace the real solution u by our primal-based approximation, say u_{RB}; however, the effect of this choice depends on the stability of the dual problem and seems to be extremely difficult to assess in a offline-online framework.

In the context of a Reduced Basis method, a posteriori error estimators for elliptic non-linear PDEs[32] have been developed by applying the Brezzi-Rappaz-Raviart (BRR) theory [14, 11]. In Appendix B the theory is briefly reviewed. However, the approach requires a number of hypotheses extremely difficult to assess in the context of hyperbolic problems. In the rest of the section we propose a different approach.

3.4.1 Error estimation for strong solutions

In this subsection we derive two error estimators: the first one for a general linear hyperbolic problem and the second one for non-linear conservation laws.

Before starting, we make a preliminary observation. Let us assume that the error introduced by the application of the shock capturing algorithm is negligible with respect to the one introduced by the resolution of the smooth problems (3.3.6). Then, under this assumption, we just need an a posteriori error estimator for the smooth components of the solution. This is the reason why both results below introduce the hypothesis that the solution is regular.

The linear case

Let us consider the following problem:

$$\begin{cases} \frac{\partial}{\partial t}u + a\frac{\partial}{\partial x}u + a_0 u = g & (t,x) \in (0,T)\times\mathbb{R} \\ u(0) = u_0 & x \in \mathbb{R} \end{cases} \tag{3.4.3}$$

where $u_0 \in L^2(\mathbb{R})$, $g \in L^2((0,T)\times\mathbb{R})$, $a \in Lip((0,T)\times\mathbb{R})$ and $a_0 \in L^\infty((0,T)\times\mathbb{R})$.

We first state an important result taken from [99].

[30] See Appendix A for the definition of a Radon measure.

[31] There is another tricky problem not directly connected to the RB method to deal with: the dual equation is a linear hyperbolic equation with discontinuous coefficients for which no wellposedness results have been proven.

[32] See [122] for the derivation of an a posteriori estimator for the steady Navier-Stokes equation using the BRR theory and the more recent paper [94], for the derivation of an a posteriori error estimator for the viscous Burgers equation.

Lemma 3.1. *Let us consider problem (3.4.3) and let*

$$\mathcal{C}(t) := \max\{0, -\min_{\{x\in\mathbb{R}\}}\left(a_0(t,x) - \frac{1}{2}\frac{\partial}{\partial x}a(t,x)\right)\}.$$

Then, the following estimate holds:

$$\|u(t)\|^2_{L^2(\mathbb{R})} \le e^{\lambda(t)}\left(\int_0^t \|g(\tau)\|^2_{L^2(\mathbb{R})}\,d\tau + \|u_0\|^2_{L^2(\mathbb{R})}\right), \tag{3.4.4}$$

where $\lambda(t) = \int_0^t (1 + 2\mathcal{C}(\tau))\,d\tau$.

Proof. By multiplying (3.4.3) by u and integrating in space we obtain:

$$\frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\mathbb{R})} + \int_{\mathbb{R}}\left(\frac{1}{2}a\frac{\partial}{\partial x}u^2(t) + a_0u^2(t)\right)dx = \int_{\mathbb{R}} g(t)u(t)\,dx.$$

Integrating by parts and applying the Young inequality, we obtain:

$$\frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\mathbb{R})} + \int_{\mathbb{R}}\left(a_0 - \frac{1}{2}\frac{\partial}{\partial x}a\right)u^2(t)\,dx \le \frac{1}{2}\|g(t)\|^2_{L^2(\mathbb{R})} + \frac{1}{2}\|u(t)\|^2_{L^2(\mathbb{R})}.$$

Then, by splitting $a_0 - \frac{1}{2}\frac{\partial}{\partial x}a$ in its positive and negative part and by remembering the definition of $\mathcal{C}$:

$$\frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\mathbb{R})} + \int_{\mathbb{R}}\underbrace{\left(a_0 - \frac{1}{2}\frac{\partial}{\partial x}a\right)^+ u^2(t)}_{\ge 0}\,dx \le \frac{1}{2}\|g(t)\|^2_{L^2(\mathbb{R})} + \left(\frac{1}{2} + \mathcal{C}(t)\right)\|u(t)\|^2_{L^2(\mathbb{R})}.$$

Finally, (3.4.4) follows by integrating in time and by applying Gronwall Lemma, see Appendix A. □

We have now the elements to state the following result[33].

Lemma 3.2. *Let u be the solution to (3.4.3) and let u_{RB} be a given function in $H^1((0,T)\times\mathbb{R})$. We define the strong residual*

$$r_{RB} := g - \frac{\partial}{\partial t}u_{RB} - a\frac{\partial}{\partial x}u_{RB} - a_0u_{RB}, \qquad \bar{r}_{RB} := u_{RB}(0) - u_0.$$

Then, $r_{RB} \in L^2((0,T)\times\mathbb{R})$ and $\bar{r}_{RB} \in L^2(\mathbb{R})$ and the following estimate holds:

$$\|u(t) - u_{RB}(t)\|^2_{L^2(\mathbb{R})} \le e^{\lambda(t)}\left(\int_0^t \|r_{RB}(\tau)\|^2_{L^2(\mathbb{R})}\,d\tau + \|\bar{r}_{RB}\|^2_{L^2(\mathbb{R})}\right) \tag{3.4.5}$$

where λ is defined as in Lemma 3.1.

Let us gather some comments about Lemma 3.2.

- Equation (3.4.5) provides an a posteriori error estimator for problem (3.4.3). Thanks to the formulation we derived in this chapter, the regularity assumption is not particularly restrictive in many situations. In addition, we point out that no regularity assumption about the exact solution is made.
- In order to extend estimate (3.4.5) to the multidimensional case, it is sufficient to generalize Lemma 3.1 (see [99], section 14.3.1, for all the details).

[33] The proof is a straightforward application of Lemma 3.1 and it is consequently here omitted.

- Estimate (3.4.5) is formally equivalent to the ones proposed in the first chapter in the elliptic and parabolic equations. For this reason the offline-online decomposition for the rapid evaluation of the term in brackets can be performed by following the same steps as in section 1.5.2. However, this estimation is not based on a variational approach; in contrast to (1.5.6) where the solution u is the truth solution to the discretized problem, in (3.4.5) u is the exact solution to the original equation.
- The quality of the estimator depends on $\mathcal{C}$. In the parameter dependent context, it is necessary to define a suitable offline-online strategy to estimate it. In this work we do not deal with this issue.

The nonlinear case

In this subsection, we try to extend the result in Lemma 3.2 to the nonlinear case.

Let us consider the following two problems:

$$\begin{cases} \frac{\partial}{\partial t}u + af'(u)\frac{\partial}{\partial x}u = 0 & (t,x) \in (0,T)\times\mathbb{R} \\ u(0) = u_0 & x \in \mathbb{R}, \end{cases} \tag{3.4.6a}$$

$$\begin{cases} \frac{\partial}{\partial t}u_{RB} + af'(u_{RB})\frac{\partial}{\partial x}u_{RB} = -r_{RB} & (t,x) \in (0,T)\times\mathbb{R} \\ u_{RB}(0) = u_0 & x \in \mathbb{R}, \end{cases} \tag{3.4.6b}$$

where $a \in Lip((0,T)\times\mathbb{R})$ and $u_0 \in W^{1,1}(\mathbb{R})$. We make the following assumptions:

$$\begin{aligned} &r_{RB} \in L^2((0,T)\times\mathbb{R}) \quad f \in Lip_K \cap C^2(\mathbb{R}) \\ &u, u_{RB} \in W^{1,1}((0,T)\times\mathbb{R}) \end{aligned} \tag{3.4.7}$$

Under hypotheses (3.4.7), problem (3.4.6a) is formally equivalent to the smooth problems (3.3.6). By subtracting (3.4.6b) from (3.4.6a), we obtain

$$\begin{cases} \frac{\partial}{\partial t}(u-u_{RB}) + f'(u)\frac{\partial}{\partial x}(u-u_{RB}) + a(f'(u)-f'(u_{RB}))\frac{\partial}{\partial x}u_{RB} = r_{RB} & (t,x) \in (0,T)\times\mathbb{R} \\ u(0) - u_{RB}(0) = 0 & x \in \mathbb{R}, \end{cases}$$

and, by remembering the notable Lagrange theorem ([91]), we have

$$\begin{cases} \frac{\partial}{\partial t}(u-u_{RB}) + af'(u)\frac{\partial}{\partial x}(u-u_{RB}) + \left[af'(\xi_{(t,x)})\frac{\partial}{\partial x}u_{RB}\right](u-u_{RB}) = r_{RB} & (t,x) \in (0,T)\times\mathbb{R} \\ u(0) - u_{RB}(0) = 0 & x \in \mathbb{R}, \end{cases}$$

where $\xi_{(t,x)}$ is such that $f'(\xi_{(t,x)}) = f'(u(t,x)) - f'(u_{RB}(t,x))$. Now we have the ingredients to state the following.

Lemma 3.3. *Let us suppose that hypotheses (3.4.7) hold. Then, let us define*

$$\mathcal{C}(t) := \max\left\{0, -\min_{(t,x)\in(0,T)\times\mathbb{R}}\left(af'(\xi_{(t,x)})\frac{\partial}{\partial x}u_{RB} - \frac{\partial}{\partial x}(af'(u))\right)\right\}.$$

Then the following estimate holds:

$$\|u(t) - u_{RB}(t)\|^2_{L^2(\mathbb{R})} \le e^{\lambda(t)}\left(\int_0^t \|r_{RB}(\tau)\|^2_{L^2(\mathbb{R})}\,d\tau + \|\bar{r}_{RB}\|^2_{L^2(\mathbb{R})}\right) \tag{3.4.8}$$

where λ is defined as in (3.4.5).

In order to use (3.4.8) to define an a posteriori error estimator, it is necessary to define a procedure to estimate $\mathcal{C}$. In this work no strategy to handle this problem is presented. However, Lemma 3.2 provides evidence that the quantity:

$$\int_0^T \|r_{RB}(\tau)\|^2_{L^2(\mathbb{R})}\, d\tau + \|r_{\bar{R}B}\|^2_{L^2(\mathbb{R})}$$

can be assumed as a reasonably good indicator at least during the sampling strategy.

3.5 Numerical simulations

After presenting the theoretical elements behind the methodology, we motivate it by providing some numerical examples.

Before starting, we discuss an important preliminary point. In order to assess the convergence of the reduced solution, we compare it with the corresponding truth solution. When Galerkin projection is applied, we observe that the convergence is independent of the underlined mesh. On the other hand, in our case *the underlined mesh influences the convergence of the RB solution with respect to the truth one*. The reason is that in our method we use two different strategies in the offline and online stage instead of simply reducing the number of test functions as in Galerkin projection-based methods. As a result, the convergence of the reduced solution to the truth one is the consequence of the convergence of the reduced solution to the exact solution. Therefore, due to the fact that the difference between the exact and the truth solution depends on the mesh we have that the convergence of the RB solution with respect to the truth one is limited by the accuracy of the truth approximation.

Moreover, as explained in section 3.2.1, conservative methods produce cell-average approximations of the solution: if we use these approximations as pointwise values of the approximate solutions we intrinsically introduce a $\Theta(h^2)$ error.

3.5.1 First example: convergence study with respect to the number of basis functions

Let us consider the following problem where $\mu \in (-0.5, 0.5)$.

$$\begin{cases} \dfrac{\partial}{\partial t}u(t) + \dfrac{1+\mu}{6}\dfrac{\partial}{\partial x}\left(u^2\right) = 0 & (t,x) \in (0,1] \times (-5,5), \\ u(0,x) = sin(x) + \dfrac{\mu}{10}(x^2 - 25) & \\ u(-5,t) = -sin(5) \quad u(5,t) = sin(5). & \end{cases} \tag{3.5.1}$$

We consider two different truth approximations, the first one is built upon an equispaced mesh with $(\Delta t, h) = (0.01, 0.002)$; the second one is computed on a finer equispaced mesh with $(\Delta t, h) = (0.005, 0.001)$. The conservative method is based on the Lax-Friedrichs flux for both the meshes.

In the tables below the errors with respect to $\|\cdot\|_{L^2(0,1,L^2(-5,5))}$ norm are gathered. We also collect the shock starting time[34]- that happens after the end of the considered time window and the solution norm $\|u(\mu)\|_{L^2(0,1;L^2(-5,5))}$ in order to give the possibility to the reader to compute the relative error.

We observe that for the former mesh, see Table 3.4, the error saturates at $\mathcal{O}$ (0.01) whereas for the latter mesh, see Table 3.5, the error saturates at $\mathcal{O}$ (0.005). Due to the fact that Lax Friedrichs is first-order accurate, the result is in good agreement with the prior observation.

[34]The shock is computed with the smooth-jump decomposition algorithm.

	$\mu = -0.35$	$\mu = -0.15$	$\mu = 0.15$	$\mu = 0.35$
t_{shock}	4.234	3.762	2.91	2.236
$\|u(\mu)\|_{L^2(0,1;L^2(-5,5))}$	3.0807	2.4874	2.4976	3.0991
$N = 2$	0.1139	0.2157	0.2429	0.1498
$N = 4$	0.0104	0.011	0.0049	0.0415
$N = 8$	0.0050	0.0069	0.0094	0.0116
$N = 16$	0.0053	0.0066	0.0087	0.0128

Table 3.4: Approximation error $\|u_{RB}(\mu) - u(\mu)\|_{L^2(0,1;L^2(-5,5))}$ for the coarse mesh.

	$\mu = -0.35$	$\mu = -0.15$	$\mu = 0.15$	$\mu = 0.35$
t_{shock}	4.234	3.762	2.91	2.236
$\|u(\mu)\|_{L^2(0,1;L^2(-5,5))}$	3.0807	2.4874	2.4976	3.0991
$N = 2$	0.1143	0.2168	0.2447	0.1512
$N = 4$	0.0102	0.0007	0.0014	0.0360
$N = 8$	0.0020	0.0025	0.0042	0.0042
$N = 16$	0.0036	0.0033	0.0048	0.0058

Table 3.5: approximation error $\|u_{RB}(\mu) - u(\mu)\|_{L^2(0,1;L^2(-5,5))}$ for the fine mesh.

3.5.2 Second example: analysis of the convergence in the presence of a shock

Let us consider the following problem where $\mu \in (0.3, 1.7)$:

$$\begin{cases} \frac{\partial u}{\partial t}(t) + \mu \frac{\partial}{\partial x}(u(1-u)) = 0 & (t,x) \in (0,2] \times (-5,5), \\ u(x,0) = g(x) = \begin{cases} \frac{1}{10} + \frac{1}{10} sin(x) & x < 0.5 \\ \frac{1}{2} + \frac{1}{10} sin(x) & x > 0.5. \end{cases} \\ u(-5,t) = \frac{1}{10} - \frac{1}{10} sin(5) \quad u(5,t) = \frac{1}{2} + \frac{1}{10} sin(5) \end{cases} \tag{3.5.2}$$

The truth approximation is based on a equispaced mesh with $(\Delta t, h) = (0.002, 0.01)$. The conservative method uses the Lax-Friedrichs flux. The drift in the smooth jump algorithm is $\delta = 10h = 0.1$. As it is easy to observe, the shock propagates from $(0.5, 0)$. First of all, we compare the smooth jump decomposition algorithm and the shock capturing algorithm. By remembering that the smooth jump decomposition algorithm approximates the position of the shock with the nearest grid node, the errors less than $h = 0.01$ are considered negligible. As shown in Table 3.6 the $L^\infty(\Omega)$-norm of the corresponding error is below 0.01 for each choice of μ and less sensitive to such a choice.

μ	$\|x^\star_{SJ}(\cdot) - x^\star_{SC}(\cdot)\|_\infty$
0.5	0.0064
1	0.0063
1.5	0.0062

Table 3.6: differences in the approximation of the shock between the smooth jump decomposition algorithm ($x^\star_{SJ}$) and the shock capturing algorithm ($x^\star_{SC}$).

Now we turn to the RB approximation. In Table 3.7 the error of the RB method with respect to the truth approximation is listed for two different choices of the reduced basis and three different values of the parameter μ.

	$\mu = 0.5$	$\mu = 1$	$\mu = 1.5$
$N = 2$	0.0496	0.0591	0.0525
$N = 4$	0.0481	0.0490	0.0492

Table 3.7: values of the error $\|u_{RB}(\mu) - u(\mu)\|_{L^2(0,2;L^2(-5,5))}$ between the RB solution and the truth approximation for three different values of the parameter and two different bases.

In order to explain the results we observe that the overall error is composed by two distinct components:

1. the error associated with the smooth-jump algorithm (i.e., the error related to the reconstruction of the shock and to the mapping);
2. the error related to the approximation of the smooth problems (3.3.6) (i.e., the error linked to the interpolation procedure).

Figure 3.5 -in which the error function $\|u_{RB}(\mu)(t^k) - u(\mu)(t^k)\|_{L^2(-5,5)}$ is plotted - shows that for $N = 2$ the error depends on time, on the contrary for $N = 4$ the error is not influenced by the time. This provides evidence to associate the first error component (the one related to the smooth-jump algorithm) with the oscillatory in time behaviour of the error function and the second component (the one related to the interpolation error) with the monotone increasing in time behaviour of the error function. According to this interpretation, we have that for $N > 2$ the error related to the smooth jump algorithm dominates over the error associated with the interpolation procedure.

3.5.3 Third example: the input-output relation

In this last simulation we consider the following input-output relation:

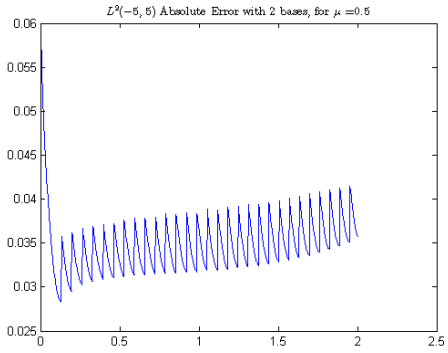
$$s(\mu) = Lu(\mu) = \int_0^1 \int_{-5}^5 \frac{1}{1+x^2} u(\mu)\, dxdt, \tag{3.5.3a}$$

where $u(\mu)$ is the solution to the following conservation law[35]

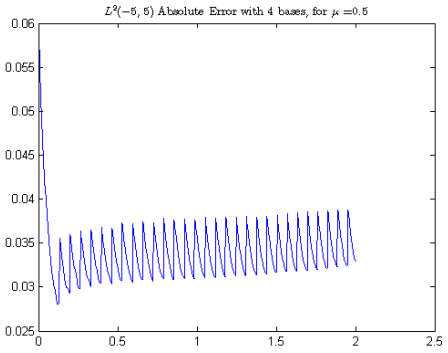
$$\begin{cases} \dfrac{\partial}{\partial t}u(t) + \mu \dfrac{\partial}{\partial x}\left(u \log \dfrac{1}{u}\right) = 0 & (t,x) \in (0,2] \times (-5,5), \\ u(x,0) = g(x) = \begin{cases} \frac{1}{5} + \frac{1}{10} sin(x) & x < 0.5 \\ \frac{1}{2} + \frac{1}{10} sin(x) & x > 0.5, \end{cases} & \\ u(-5,t) = \frac{1}{5} - \frac{1}{10} sin(5) \quad u(5,t) = \frac{1}{2} + \frac{1}{10} sin(5). & \end{cases} \tag{3.5.3b}$$

We consider an equispaced mesh $h = 0.01$, $\Delta t = 0.002$, $\mathcal{D} = [0.3, 1.7]$. The numerical flux is the Godunov flux. The drift in the smooth jump algorithm is $\delta = 5h = 0.1$. In Table 3.8 below the errors for some values of the parameter in the state equation are listed; in Table 3.9 the resulting outputs and the computational time needed to obtain it are listed.

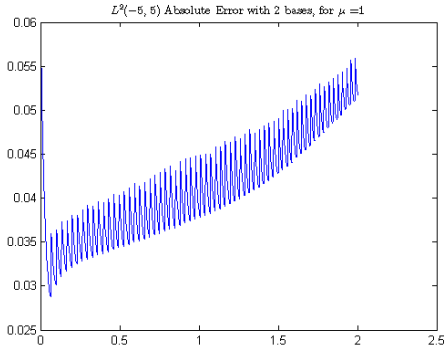
[35] This flux is used in hyperbolic traffic models: it was proposed by Greenbery and supported by experimental data from the Lincoln tunnel in New York (see [40] for further details).



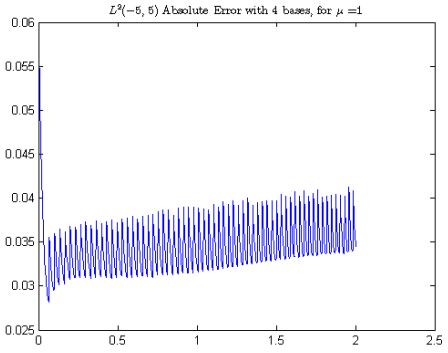
(a) $\mu = 0.5$ $N = 2$



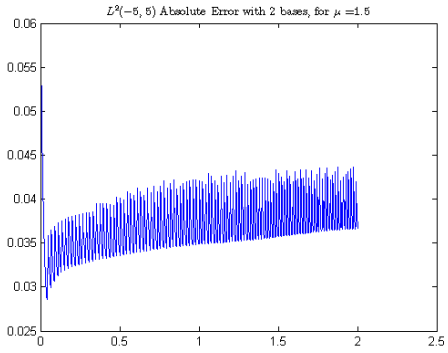
(b) $\mu = 0.5$ $N = 4$



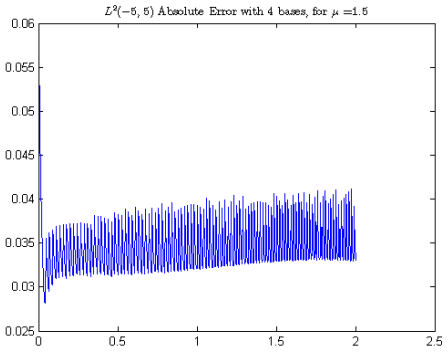
(c) $\mu = 1$ $N = 2$



(d) $\mu = 1$ $N = 4$



(e) $\mu = 1.5$ $N = 2$



(f) $\mu = 1.5$ $N = 4$

Figure 3.5: L^2-error vs time: $\|u_{RB}(\mu)(t^k) - u(\mu)(t^k)\|_{L^2(-5,5)}$ plotted with respect to time.

	$\mu = 0.5$	$\mu = 1$	$\mu = 1.5$
$N_{RB} = 16$	0.0570	0.0682	0.0682

Table 3.8: $\|u_{RB}(\mu)(t^k) - u(\mu)(t^k)\|_{L^2(-5,5)}$ for three different values of parameter μ.

We think that the results concerning the speed-up are extremely positive. The truth method used is an explicit scheme so the computational effort is proportional to $\mathcal{O}(C_1 \mathcal{N} \mathbb{K})$, where C_1 is the cost associated with the evaluation of the Godunov flux while $\mathcal{N}$ is the spatial mesh dimension and $\mathbb{K}$ is the temporal mesh dimension. On the other hand, the

	$\mu = 0.5$	$\mu = 1$	$\mu = 1.5$
$N_{w,1} = N_{w,2} = N_{RB} = 16$	0.8098	0.7955	0.7832
Truth Results	0.8114	0.7970	0.7850
Speed-up	24.5978	25.0350	25.2591

Table 3.9: error and speed-up in the output evaluation.

reduced method is dominated by the cost associated with the shock capturing algorithm that is $\mathcal{O}(C_2\mathbb{K})$, where C_2 is the cost related to the application of the Newton Raphson algorithm for finding the roots of a nonlinear equation.

Due to the fact that C_1 is significantly less than C_2, we cannot expect significantly higher computational savings. On the other hand, we think that in higher dimensions, we can obtain larger speed-ups.

3.6 Conclusions

The method proposed in this chapter can be viewed as the combination of three different steps:

1. the shock detection;
2. the definition of a surrogate problem for the smooth components of the solution based on a domain decomposition approach;
3. the online interpolation.

The idea of applying a preliminary domain decomposition and then using the standard RB method on the single components of the solution is very close to the Reduced Basis Element Method (RBEM) first proposed in [75] for the Stokes problem as well as to the Reduced Basis Hybrid Method (RBHM) proposed in [56]; but in our case the domain decomposition is induced by the parameter both directly and through the solution of the problem.

Concerning to the method used to solve the smooth problems , the approach we employ is -with respect to our knowledge- new and quite far from the techniques used in the RB method[36].

The numerical simulations show that the method is able to reconstruct the solution in an efficient and reasonable sharp way. The main limitation of this approach is linked to the reconstruction of the shock: due to the fact that the online and the offline methods are based on completely different formulation it is not reasonable to expect errors less than $\mathcal{O}(h)$. It could be interesting to solve offline the problems (3.3.6) and then using directly the solutions to these problems- instead of the results of the smooth jump decomposition algorithm- to build the empirical basis.

As future steps, we aim at extending the proposed approach to the bidimensional case. As explained in the conclusions of the second chapter, in order to extend this methodology it is necessary to define a suitable geometric reduction technique to deal with the domain decomposition induced by the shock.

In [29] the so called *Empirical Operator Interpolation method* (EOI) has been proposed to deal with nonlinear time-dependent PDEs: given the discretized problem, EOI is first applied to provide a surrogate problem; then a projection-based algorithm is used to solve the problem. Even though the article provides some numerical evidence to support the approach, we think that the methodology has some drawbacks that limit its range of application.

First of all, the approximation of the differential operator is a critical aspect. In order to show why, we remember a classical result from [19].

Theorem 3.1. *Let us consider the problems*

$$\begin{cases} \dfrac{\partial}{\partial t}u + \dfrac{\partial}{\partial x}f(u) = 0 \quad in\ (t,x) \in (0,T]\times\mathbb{R} \\ u(0) = u_0 \quad in\ x \in \mathbb{R}, \end{cases} \qquad \begin{cases} \dfrac{\partial}{\partial t}u_\epsilon + \dfrac{\partial}{\partial x}f_\epsilon(u_\epsilon) = 0 \quad in\ (t,x) \in (0,T]\times\mathbb{R} \\ u_\epsilon(0) = u_0 \quad in\ x \in \mathbb{R}. \end{cases} \tag{3.6.1}$$

Then, the following estimate holds:

$$\|u(t,\cdot) - u_\epsilon(t,\cdot)\|_{L^1(\mathbb{R})} \leq \|u_0\|_{BV(\mathbb{R})}\|f' - f'_\epsilon\|_{L^\infty(\mathbb{R})}t \tag{3.6.2}$$

where $\|u_0\|_{BV(\mathbb{R})} = |Du_0|(\mathbb{R})$ *(see Appendix A).*

[36] In [94] the online solution is computed through an interpolation process; however, such an interpolation is based only on pre-computed data.

Therefore, while for parabolic and elliptic problems the continuous dependence is assessed by controlling the L^∞-norm of the approximation, for hyperbolic problems we need a control in the stronger $W^{1,\infty}$-norm.

On the other hand, in our method we do not need any approximation of the differential operator.

Even starting from a stable truth approximation, a given model order reduction process can generate instability. Therefore, it could be necessary to introduce some forms of "reduced viscosity"[37], i.e., forms of stabilization that are not based on the local properties[38] of the solution such as the so-called flux-limiters used to obtain high-order schemes for hyperbolic equations. In our opinion it would be very difficult to find an effective stabilization method for the RB approximation.

As shown in several numerical test-cases in the literature, EIM is a stable procedure so starting from a stable truth approximation, no instability is generated.

For these reasons we are of the opinion that methods based on empirical interpolation could have more potentiality than projection-based methods.

[37] For instance, in [37] a Petrov-Galerkin approach is used to obtain a stable POD reduced model. With respect to our knowledge, in the context of Reduced Basis methods, there are no significant examples of such techniques.

[38] This point is crucial in order to reach the online independence on the spatial mesh.

Conclusions and Perspectives

In this work starting from the analysis of the state of the art, we focused on two open issues of interest for an efficient and reliable resolution of parametrized partial differential equations.

Concerning the *piecewise transfinite map for small deformation*, we have first motivated the approach from a theoretical viewpoint (through Lemmas 2.1 and 2.2) and then we have numerically assessed it on some test cases.

In our opinion the next step should be to compare the approach here proposed with the other techniques (e.g., already well established FFD, RBF and traditional transfinite maps) on some relevant test cases. It could be interesting not only to assess whether each mapping strategy is able to correctly describe a given deformation but also to analyze how the application of a certain map influences the differential problem in terms of FE approximation property (see indicators (2.5.2) (2.5.4)) as well as of convergence of the RB approximation.

Concerning the method proposed to deal with conservation laws, we have first discussed about the difficulties intrinsic to tackle discontinuous solutions; then we have proposed a new strategy based on the so-called *characteristic method*. Numerical simulations provide some evidence about the good performances of our new approach.

The next evident (but not straightforward) steps are the extension of this methodology to bidimensional problems and the development of a rigorous a posteriori error estimator. Concerning the first issue, we think that the transfinite maps developed in this thesis could be well suited to this purpose (as explained in the conclusions of the second chapter). As far as the error estimator is concerned, we think that Lemmas 3.2 and 3.3 might represent the starting point for a future work on a new a posteriori error estimator.

In perspective, we think that it could be also interesting to test this characteristic-based method on other classes of time-dependent problems, such as advection dominated nonlinear parabolic equations. The fact that no approximation of the differential operator is required in our approach could represent a great advantage for these problems.

Appendix A

Some theoretical results

In this appendix we report some important results that have been extensively used throughout the work.

A.1 Existence and uniqueness for linear variational problems

A.1.1 The Lax-Milgram and Babuska theorems

Let U and V be two real Hilbert spaces. In the following we denote with $(\cdot,\cdot)_U$, $\|\cdot\|_U$ and $(\cdot,\cdot)_V$, $\|\cdot\|_V$ the inner products and the corresponding norm associated with U and V, respectively. Let us consider the following bilinear form:

$$\Phi : U \times V \to \mathbb{R}. \tag{A.1.1}$$

We say that Φ is γ-*continuous* if:

$$\Phi(u,v) \leq \gamma \|u\|_U \|v\|_V, \quad \forall\, u \in U\ \forall\, v \in V. \tag{A.1.2}$$

If $U = V$, we say that Φ is α-*coercive* if:

$$\Phi(u,u) \geq \alpha \|u\|_U^2, \quad \forall\, u \in U \text{ with } \alpha > 0. \tag{A.1.3}$$

Otherwise, if U and V are different, we say that Φ is β-*inf-sup stable* if:

$$\beta = \inf_{u \in U} \sup_{v \in V} \frac{\Phi(u,v)}{\|u\|_U \|v\|_V} > 0. \tag{A.1.4}$$

We have now the elements to state the main results. For the proofs we refer for Lax-Milgram lemma to [109] theorem 6.5 and for the Babuska theorem[1] to [4] .

Theorem A.1. ***(Lax-Milgram Lemma)*** *Let V be a real Hilbert space and let $\Phi : V \times V \to \mathbb{R}$ be a γ-continuous and α-coercive bilinear form. Then for all $F \in V'$ the problem*

$$\text{Find } u \in V \text{ such that } \Phi(u,v) = F(v) \quad \forall\, v \in V, \tag{A.1.5}$$

admits one and only one solution $u \in V$ such that

$$\|u\|_V \leq \frac{1}{\alpha} \|F\|_{V'}. \tag{A.1.6}$$

[1] Babuska theorem is also known as Necas theorem.

Theorem A.2. ***(Babuska inf-sup condition)*** *Let U, V be two real Hilbert spaces and let $\Phi : U \times V \to \mathbb{R}$ be a γ-continuous bilinear form. Then the problem*

$$\text{Find } u \in U \text{ such that } \Phi(u, v) = F(v) \quad \forall\, v \in V \tag{A.1.7}$$

admits one and only one solution for all $F \in V'$ if and only if Φ is β-inf-sup stable. Furthermore, the following stability estimation holds:

$$\|u\|_U \leq \frac{1}{\beta}\|F\|_{V'}. \tag{A.1.8}$$

A.1.2 Elliptic problems with L^2 boundary data: the transposition method

Let us consider the following problem

$$\begin{cases} \mathcal{L}w = -\mathrm{div}(A\nabla w) + \mathbf{b}\cdot\nabla w + cw = f & \text{in } \Omega \\ w = g & \text{on } \partial\Omega \end{cases} \tag{A.1.9}$$

where $g \in L^2(\partial\Omega)$, $f \in L^2(\Omega)$, Ω is a domain of class C^2.

Due to the fact that $g \notin H^{\frac{1}{2}}(\partial\Omega)$, it is not possible to set the problem in the standard framework. Therefore, we aim at introducing a weaker formulation for (A.1.9). Let us consider $\psi \in H^2(\Omega) \cap H_0^1(\Omega)$; by integrating by part twice we obtain:

$$\int_\Omega w\mathcal{L}^\star\psi = \int_\Omega f\psi - \int_{\partial\Omega} g\partial_{\mathcal{L}^\star}\psi\, d\sigma \qquad \text{where } \partial_{\mathcal{L}^\star}\psi := \left(A^T\nabla\psi + \psi\mathbf{b}\right)\cdot\mathbf{n} \tag{A.1.10}$$

Definition A.1. *The function $w \in L^2(\Omega)$ is said to be weak solution of (A.1.9)-(A.1.10) if it satisfies (A.1.10) for all $\psi \in H^2(\Omega) \cap H_0^1(\Omega)$.*

The following theorem, from [73], guarantees the wellposedness of the problem defined above.

Theorem A.3. ***(Transposition method)*** *Let X, Y be Hilbert spaces, $T : X \to Y$ be a continuous isomorphism and $F \in X'$. Then the equation:*

$$(Tx, w) = Fx \quad \forall\, x \in X \tag{A.1.11}$$

admits one and only one solution $\bar{w} \in Y$. Moreover,

$$\|\bar{w}\|_Y \leq \|T^{-1}\|_{\mathcal{L}(Y,X)}\|F\|_{X'} \tag{A.1.12}$$

Corollary A.1. *Problem (A.1.9) admits one and only one weak solution $w \in L^2(\Omega)$. Moreover,*

$$\|w\|_{L^2(\Omega)} \leq c(\Omega)\left\{\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}\right\}. \tag{A.1.13}$$

A.1.3 A useful comparison result

In this subsection we present a useful comparison result, in this thesis used in Lemma 3.1. For the proof, we refer, for instance, to [99], Lemma 1.4.1.

Lemma A.1. ***(Gronwall Lemma)*** *Let $f \in L^1(0,T)$ be a non-negative function, g and ϕ be continuous functions on $[0,T]$. If ϕ satisfies:*

$$\phi(t) \leq g(t) + \int_0^t f(\tau)\phi(\tau)\, d\tau \quad \forall\, t \in [0,T], \tag{A.1.14}$$

then

$$\phi(t) \le g(t) + \int_0^t f(s)g(s)exp\left(\int_s^t f(\tau)\,d\tau\right) ds \quad \forall\, t \in [0,T]. \tag{A.1.15}$$

Furthermore, if g is nondecreasing, then:

$$\phi(t) \le g(t)exp\left(\int_0^t f(\tau)\,d\tau\right) \quad \forall\, t \in [0,T]. \tag{A.1.16}$$

A.2 BV and SBV spaces

This section introduces the functional spaces $BV(\Omega)$ and $SBV(\Omega)$, where $\Omega \subset \mathbb{R}^d$. The following presentation is finalized to give a structure theorem that motivates our approach in the treatment of hyperbolic conservation laws. For further discussions on the topic, we refer to [2]. In this section $\mathcal{B}$ always denotes the Borel σ-algebra.

Let m be a measure on the σ-algebra $\mathcal{B}$ of Borel sets of $\mathbb{R}^d$.

Definition A.2. *The measure m is said to be a* Radon measure *if:*

- *the measure m is* inner regular *i.e.*

$$m(B) = \sup_{\{K\in\mathcal{B}:\, K\subset\subset B\}} m(K);$$

- *the measure m is* locally finite*, i.e., if every point has a neighborhood of finite measure.*

We have now the elements to introduce the space of the bounded variation functions.

Definition A.3. *Let $u \in L^1(\Omega)$; we say that u is a* a function of bounded variation in Ω *if the distributional derivative of u is representable by a finite Radon measure in Ω, i.e.:*

$$\int_\Omega u\frac{\partial\phi}{\partial x_i}\,dx = -\int_\Omega \phi\, dD_i u \quad \forall\, \phi \in C_0^\infty(\Omega) \tag{A.2.1}$$

for some $\mathbb{R}^d$-valued measure $Du = (D_1u, \cdots, D_du)$ in Ω. The vector space of all functions of bounded variation in Ω is denoted by $BV(\Omega)$.

Definition A.4. *Let $u \in L^1_{loc}(\Omega)$. The variation $V(u,\Omega)$ of u in Ω is defined by:*

$$V(u,\Omega) := \sup\left\{\int_\Omega u\,div\,\phi\,dx :\, \phi \in C_c^1(\Omega),\ \|\phi\|_{L^\infty} \le 1\right\}. \tag{A.2.2}$$

The following theorem clarifies the relationship between $BV(\Omega)$ and $V(\cdot,\Omega)$

Theorem A.4. *Let $u \in L^1(\Omega)$. Then $u \in BV(\Omega)$ if and only if $V(u,\Omega) < \infty$. In addition $V(u,\Omega)$ coincides with $|Du|(\Omega)$, for any $u \in BV(\Omega)$, and $u \mapsto |Du|(\Omega)$ is lower semicontinuous in $BV(\Omega)$ with respect to the $L^1_{loc}(\Omega)$ topology.*

If $\Omega \subset \mathbb{R}$, it is possible to introduce a simpler definition of variation.

Definition A.5. *Let $a,b \in \mathbb{R}$ with $a < b$ and $I = (a,b)$. For any function $u : I \to \mathbb{R}$ the pointwise variation $pV(u,I)$ is defined by:*

$$pV(u,I) := \sup\left\{\sum_{i=1}^{n-1} |u(t_{i+1}) - u(t_i)| :\ n \ge 2 \quad a < t_1 < \cdots < t_n < b\right\}. \tag{A.2.3}$$

If $\Omega \subset \mathbb{R}$ is an open set, the pointwise variation $pV(u,\Omega)$ is defined by $\sum_j pV(u,I_j)$, where $\{I_j\}_j$ is a set of disjoint intervals such that $\Omega = \bigcup_j I_j$.

The following result provides a second definition for the variation in one-dimensional domains.

Lemma A.2. *For any $u \in L^1_{loc}(\Omega)$ we have that*

$$V(u,\Omega) := \inf\left\{pV(v,\Omega) : v = u\ \mathcal{L}^1 \text{ a.e. in } \Omega\right\}. \tag{A.2.4}$$

Furthermore, the infimum in (A.2.4) is achieved.

In order to state the above mentioned structure result for BV functions, we first introduce some definitions. In what follows $\mathcal{L}^d$ and $\mathcal{H}^n$ denote respectively the Lebesgue measure on $\mathbb{R}^d$ and the n-dimensional Hausdorff measure on Euclidean spaces. A set $J \subset \mathbb{R}^d$ is said *countably $\mathcal{H}^n$-rectifiable* if there exist countably n-dimensional Lipschitz graphs Γ_i such that $\mathcal{H}^n(J \setminus \bigcup \Gamma_i) = 0$. Given a Borel measure μ and a Borel set A we denote by $\mu \llcorner A$ the measure given by $\mu \llcorner A(C) = \mu(A \cap C)$.

The *approximate discontinuity set* $S_w \subset \Omega$ of a locally summable function $w : \Omega \to \mathbb{R}$ and the *approximate limit* are defined as follows: $x \notin S_w$ if and only if there exists $z \in \mathbb{R}$ satisfying

$$\lim_{r \in 0^+} \frac{1}{r^d} \int_{B_r(x)} |w(y) - z|\, dy = 0.$$

If such z exists, it is unique and denoted by $\tilde{w}(x)$, i.e., the approximate limit of w in x. It is possible to prove that S_w is Borel and that $\tilde{w}$ is a Borel function in its domain[2]. Therefore, for the notable Lebesgue differentation theorem, the set S_w is Lebesgue negligible and $\tilde{w} = w$ a.e. in $\Omega \setminus S_w$.

Similarly, it is possible to define the *approximate jump set* $J_w \subset S_w$, by requiring the existence of $a, b \in \mathbb{R}$ and of an unit vector $\boldsymbol{\nu} \in \mathbb{R}^d$ such that

$$\lim_{r \to 0^+} \frac{1}{r^d} \int_{B_r^+(x,\boldsymbol{\nu})} |w(y) - a|\, dy = 0, \quad \lim_{r \to 0^+} \frac{1}{r^d} \int_{B_r^-(x,\boldsymbol{\nu})} |w(y) - b|\, dy = 0$$

where:

$$\begin{cases} B_r^+(x,\boldsymbol{\nu}) := \{y \in B_r(x) : (y-x,\boldsymbol{\nu}) > 0\}, \\ B_r^-(x,\boldsymbol{\nu}) := \{y \in B_r(x) : (y-x,\boldsymbol{\nu}) < 0\}. \end{cases}$$

If the triplet $(a, b, \boldsymbol{\nu})$ exists, it is unique up to a permutation of a and b and a change of sign of $\boldsymbol{\nu}$. We refer to it as $(w^+(x), w^-(x), \boldsymbol{\nu}(x))$, where $w^\pm(x)$ are called *approximate one-sided limits* of w at x. Like in the previous case, it is possible to prove that J_w is a Borel set and that $w^\pm$ and $\boldsymbol{\nu}$ can be chosen as Borel functions in their domain.

The following theorem represents the main result of this section.

Theorem A.5. ***(Federer Vol'pert)*** *Let $w \in BV(\Omega)$. Then $\mathcal{H}^{d-1}(S_w \setminus J_w) = 0$ and J_w is a countably $\mathcal{H}^{d-1}$-rectifiable set. If we denote by $D^a w$ the absolutely continuous part of Dw - with respect to the d-dimensional Lebesgue measure- and by $D^s w$ the singular part, then we have that $D^s w = D^j w + D^c w$, where:*

$$D^j w = (w^+ - w^-)\, \boldsymbol{\nu}_{J_w} \mathcal{H}^{d-1} \llcorner J_w, \tag{A.2.5}$$

$$D^c w(E) = 0 \text{ for any Borel set } E \text{ with } \mathcal{H}^{d-1}(E) < \infty. \tag{A.2.6}$$

Corollary A.2. *Let $w \in BV(a,b)$. Then if $S_w = J_w$, $\tilde{w}$ is continuous on $\Omega \setminus J_w$ and $\tilde{w}$ has classical left and right limits (which coincide with $w^\pm(x)$) at any $x \in J_w$. Therefore*

$$D^j w = \sum_{x \in J_w} (w^+(x) - w^-(x))\, \delta_x. \tag{A.2.7}$$

[2]See section 3.6 from [2] for the proof and further discussions.

Remark A.1. *Theorem A.5 motivates the procedure through we derived in section 3.3.1. Due to the numerable additivity of the Lebesgue integral problems (3.3.6) could be defined even without the hypothesis that* $x_{shock} \in Lip(0, T_{max})$.

The previous theorem motivates the following definition.

Definition A.6. *We say that* $u \in BV(\Omega)$ *is a* special function with bounded variation *and we write* $u \in SBV(\Omega)$, *if the Cantor part of its derivative* $D^c u$ *is zero. Thanks to Theorem A.5 we have that:*

$$Du = D^a u + D^j u = \nabla u \mathcal{L}^d + (u^+ - u^-)\boldsymbol{\nu}_u \mathcal{H}^{d-1} \llcorner J_u \quad \forall\, u \in SBV(\Omega). \tag{A.2.8}$$

This space is particularly relevant in several applications of Calculus of Variations. It is possible to prove closure and compactness properties of the space with respect to a suitable weak-$\star$ convergence in BV. We refer to [2], chapter 4 for the details.

We conclude the section by citing the following L^p embedding result.

Theorem A.6. *We have that:*

$$BV(\mathbb{R}^d) \hookrightarrow L^{\frac{d}{d-1}}(\mathbb{R}^d) \text{ if } d > 1, \quad BV(\mathbb{R}) \hookrightarrow L^\infty(\mathbb{R}) \text{ if } d = 1. \tag{A.2.9}$$

If $\Omega \subset \mathbb{R}^d$ *is Lipschitz, then* $BV(\Omega) \hookrightarrow L^p(\Omega)$ *for* $p \leq \frac{d}{d-1}$ *and the embedding is compact for* $p < \frac{d}{d-1}$.

A.3 Mathematical analysis of scalar conservation laws

In this section we analyse the following Cauchy problem:

$$\begin{cases} \dfrac{\partial}{\partial t}u + \dfrac{\partial}{\partial x}f(u) = 0 & (t,x) \in (0,\infty) \times \mathbb{R} \\ u(0,x) = u_0(x) & x \in \mathbb{R} \end{cases} \tag{A.3.1}$$

We point out that most of the theorems reported below are valid for more general domains; however, for the sake of simplicity, we limit the presentation to problem (A.3.1). We refer to [36] for a general discussion about first order PDEs. On the other hand, the content of this appendix discussion is mainly taken from [1, 36].

Definition A.7. *Let us suppose that* $u_0 \in L^\infty(\mathbb{R})$, *then* $u \in L^\infty((0,\infty) \times \mathbb{R})$ *is said to be* integral solution *to (A.3.1) if*

$$\int_0^\infty \int_{\mathbb{R}} \left(u\frac{\partial v}{\partial t} + f(u)\frac{\partial v}{\partial x} \right) dxdt + \int_{\mathbb{R}} u_0(x)v(0,x)\,dx = 0 \quad \forall\, v \in C_0^\infty(\mathbb{R}^2). \tag{A.3.2}$$

From identity (A.3.2) it is possible to deduce the following property of the integral solution.

Lemma A.3. *Let* $s : (t_0, t_1) \subset (0,\infty) \to \mathbb{R}$ *be the equation of a curve of discontinuity of* u *(shock) and let* $u_l(t)$ *and* $u_r(t)$ *be the limits of the integral solution from the left and from the right. Then the following identity holds*

$$\dot{s}(t) = \frac{f(u_r(t)) - f(u_l(t))}{u_r(t) - u_l(t)}. \tag{A.3.3}$$

In general the integral solution is not unique[3]. In order to avoid non-physical solutions, we look for solutions that satisfy the following *Entropy condition*[4].

Definition A.8. *An integral solution to (A.3.2) is said to be entropic if it satisfies the following E-condition:*

$$f'(u_l(t)) < \dot{s}(t) < f'(u_r(t)). \tag{A.3.4}$$

We observe that, if f is uniformly convex ($F'' \geq \theta > 0$), the E-condition is equivalent to require $u_l > u_r$ along any shock curve.

In order to construct an integral solution for (A.3.2) we apply the so-called *method of characteristics* ([36]).

Definition A.9. *The Legendre transform of $L : \mathbb{R} \to \mathbb{R}$ is:*

$$L^\star(p) = \sup_{q \in \mathbb{R}} \{pq - L(q)\}. \tag{A.3.5}$$

The following theorem is taken from [1].

Theorem A.7. ***(Hopf-Lax formula)*** *Assume $f : \mathbb{R} \to \mathbb{R}$, $f \in C^2(\Omega)$ uniformly convex and $u_0 \in L^1(\mathbb{R})$ and set*

$$v_0(y) = \int_{-\infty}^{y} u_0(y)\, dy.$$

Let $v(t,x) := \min\{t f^\star\left(\frac{x-y}{t}\right) + v_0(y) : \ y \in \mathbb{R}\}$. Then the following statements hold:

1. *for any $t > 0$ there exists a countable set S_t such that the minimum is attained at a unique point $y(t,x)$ for any $x \notin S_t$;*

2. *the map $x \mapsto y(t,x)$ is nondecreasing in its domain, its jump set is S_t and $v(t,\cdot)$ is differentiable at any $x \notin S_t$ with*

$$f'\left(\frac{\partial v}{\partial x}(t,x)\right) = \frac{x - y(t,x)}{t}. \tag{A.3.6}$$

   In particular $\frac{\partial v}{\partial x}(t,\cdot)$ is continuous on $\mathbb{R} \setminus S_t$;

3. *there exists a constant $C > 0$ such that:*

$$\frac{\partial v}{\partial x}(t, x+y) \leq \frac{\partial v}{\partial x}(t,x) + \frac{C}{t} y \quad \forall\, y \geq 0 \text{ and } x, x+y \notin S_t. \tag{A.3.7}$$

   This is called Oleinik E-condition.

4. *v is a Lipschitz map and $u = \frac{\partial v}{\partial x}$ is the unique entropy solution to (A.3.2) with the initial condition $u(0,\cdot) = u_0$*

5. *$u : (0,\infty) \to L^1(\mathbb{R})$ is continuous with respect to the L^1_{loc} topology.*

We can use the Hopf-Lax variational principle to define backward characteristics emanating from points (t,x) with $x \in \mathbb{R} \setminus S_t$.

[3] See [36, 109] for an instructive example for Burgers equation.

[4] The terminology is roughly motivated by a rough analogy with the thermodynamic principle that physical entropy cannot decrease as time goes forward.

Definition A.10. *Let $x \notin S_t$. The segment joining (t,x) with $(0,y(t,x))$ is called (backward minimal) characteristic emanating from (t,x). These segments, when parametrized with constant speed on the interval $[0,t]$, are minimizers of the variational problem related to the Hopf Lax formula:*

$$\min\left\{\int_0^t f^\star(\dot\gamma(s))\,ds + v_0(\gamma(0)) :\ \gamma \in C^1([0,t];\mathbb{R})\ \gamma(t) = x\right\}. \tag{A.3.8}$$

Indeed the strict convexity of $f^\star$ forces the minimizers to be straight lines and forces a constant speed parametrization.

Remark A.2. *Thanks to the monotonicity of $y(t,\cdot)$, characteristics emanating from points $x, y \notin S_t$ with $x \neq y$ do not intersect in the open upper half plane. As a consequence, two different characteristics starting even at different times are either one contained in the other or do not intersect.*

We conclude this section with two important results.

Theorem A.8. *Let us consider problem (A.3.1) and let u, v and w be the entropic integral solutions associated with u_0, v_0 and w_0 respectively. Then:*

1. *(from [40]) the following maximum principle holds:*

$$\|u(t,\cdot)\|_{L^\infty} \le \|u_0(\cdot)\|_{L^\infty}; \tag{A.3.9}$$

2. *(from [65]) the solution operator is a L^1-contraction, i.e.,*

$$\|v(t,\cdot) - w(t,\cdot)\|_{L^1} \le \|v_0(\cdot) - w_0(\cdot)\|_{L^1}; \tag{A.3.10}$$

3. *(from [95]) for every $\Phi : \mathbb{R} \to \mathbb{R}$ Lipschitz continuous monotonous function we have that:*

$$|\Phi(u(t,\cdot))|_{BV} \le |\Phi(u_0(\cdot))|_{BV} \tag{A.3.11}$$

where we remember that $|u|_{BV(\Omega)} = |Du|(\Omega)$.

As we motivated in the chapter related to hyperbolic problems, we aim at setting our problem in SBV, however there are counterexamples that show that even starting from a Lipschitz initial datum it is possible to obtain, for some $t > 0$, $u(t)$ such that $D^c u \neq 0$ (see Remark 2.1 [24]). However, the following result proved in [1] holds:

Theorem A.9. *Let $u \in L^\infty((0,\infty)\times\mathbb{R})$ be the entropy solution to (A.3.2) where $u_0 \in SBV_{loc}(\mathbb{R})$ and $f \in C^2$ is locally uniformly convex. Then, there exists $S \subset \mathbb{R}$ at most countable such that, $\forall\, \tau \in \mathbb{R} \setminus S$, the following holds*

$$u(\tau,\cdot) \in SBV_{loc}(\mathbb{R}). \tag{A.3.12}$$

As a consequence $u \in SBV_{loc}((0,\infty)\times\mathbb{R})$.

Appendix B

Preliminary results for the BRR theory

In the context of the Reduced Basis method, the Brezzi-Rappaz-Raviart (BRR) theory has been extensively used to find a posteriori error estimators for non-linear problems ([122, 15, 94]). In this appendix we present only three preliminary abstract results that constitute the basis of this theory.

The BRR theory was originally proposed in [11]. On the other hand, for this presentation, we mainly refer to the successive review work [14].

B.1 Some notations and basic lemmas

Let $(X, \|\cdot\|_X)$, $(Z, \|\cdot\|_Z)$ be real Banach spaces; we indicate with $B_X(x,\delta)$ and with $\bar{B}_X(x,\delta)$ the open and the closed balls in X[1], respectively. Furthermore, we indicate with

$$\|(x,z)\|_{X\times Z} = \|x\|_X + \|z\|_Z, \quad \|A\|_{\mathcal{L}(X,Z)} = \sup_{x\in\bar{B}_X(x,1)} \|Ax\|_Z.$$

the product space and the $\mathcal{L}(X,Z)$ norms.

We first remember the notable Banach-Caccioppoli theorem[2] that will the basis of the proof of the main theorem below.

Theorem B.1. *(**Banach-Caccioppoli**) Assume $A : X \to X$ be a contractive mapping, i.e.,*

$$\|A[u] - A[v]\| \le \gamma\|u - v\| \quad \forall\, u, v \in X \quad \text{and with } \gamma < 1.$$

Then A has a unique fixed point $A[x] = x$.

Now we present a useful lemma.

Lemma B.1. *Let $A \in \mathcal{L}(X,Z)$ be invertible. Let $B \in \mathcal{L}(X,Z)$ and suppose that $\|A^{-1}B\|_{\mathcal{L}(X,X)} < 1$. Then $A + B \in \mathcal{L}(X;Z)$ is invertible and the following inequality holds:*

$$\|(A+B)^{-1}\|_{\mathcal{L}(Z,X)} \le \frac{1}{1 - \|A^{-1}B\|_{\mathcal{L}(X,X)}}$$

[1] We will omit the subscript X when there is no risk for misunderstandings.

[2] For the proof see for example [36] section 9.2.1. theorem 1.

Proof. We start by proving that the map is bijective. By contradiction let $x \in X$ such that $(A+B)x = 0$, thus:

$$(A+B)x = A(I + A^{-1}B)x \Rightarrow (I + A^{-1}B)x = 0 \Rightarrow \|A^{-1}B\| \geq 1.$$

This is in contradiction with the hypothesis. We have proved the injectivity, concerning the surjectivity we consider the following map:

$$H : X \to X \quad H(x) = A^{-1}(y - Bx).$$

We observe that $\|H(x_1) - H(x_2)\|_X \leq \|A^{-1}B\|_{\mathcal{L}(X,X)}\|x_1 - x_2\|_X$; thus it is a contraction for the hypotheses. The fixed theorem above assures the existence of a fixed point.

We now prove the estimate. Let $C : X \to X$, $C = A^{-1}B$. It is easy to observe that:

$$I = I - C + C \Rightarrow \left(\text{post-multiplying both terms for } (I-C)^{-1}\right) \quad (I-C)^{-1} = I + C(I-C)^{-1}.$$

Thus, thanks to the hypothesis on the norm of C,

$$\|(I-C)^{-1}\|_{\mathcal{L}(X,X)} \leq 1 + \|C\|_{\mathcal{L}(X,X)}\|(I-C)^{-1}\|_{\mathcal{L}(X,X)} \Rightarrow \|(I-C)^{-1}\|_{\mathcal{L}(X,X)} \leq \frac{1}{1 - \|C\|_{\mathcal{L}(X,X)}}.$$

In conclusion:

$$\|(A+B)^{-1}\|_{\mathcal{L}(Z,X)} = \|A^{-1}(I + A^{-1}B)^{-1}\|_{\mathcal{L}(Z,X)} \leq \frac{1}{1 - \|A^{-1}B\|_{\mathcal{L}(X,X)}}\|A^{-1}\|_{\mathcal{L}(Z,X)}.$$

□

We conclude this section by defining the Fréchet derivative and introducing the Taylor formula.

Definition B.1. *$G : U \subset X \to Z$ is said to be Fréchet differentiable in $x_0 \in U$ if there exists $L \in \mathcal{L}(X,Z)$ such that:*

$$\|G(x_0 + h) - G(x_0) - Lh\|_Z = o(\|h\|_X) \quad \forall\, h \text{ such that } x_0 + h \in U. \tag{B.1.1}$$

If G is Fréchet differentiable in $U \subset X$, $DG : U \to \mathcal{L}(X,Z)$ such that:

$$x \mapsto DG(x)$$

is said to be the Fréchet derivative of G. We say that $G \in C^1$ if DG is continuous in U. Assuming that $G \in C^p$, the following Taylor formula holds:

$$G(y) = G(x) + \sum_{k=1}^{p-1} \frac{1}{k!} D^k G(x) \underbrace{(y-x, \cdots, y-x)}_{k \text{ times}} + \frac{1}{(p-1)!} \int_0^1 (1-t)^{p-1} D^p G(x + t(y-x))(y-x, \cdots, y-x)\, dt. \tag{B.1.2}$$

B.1.1 Implicit and inverse function theorems

In this section we focus on the problem:

$$\text{Find } u \in X \text{ such that } G(u) = 0. \tag{B.1.3}$$

First of all, we prove an important result that is fundamental in the error analysis. Then we state an implicit and an inverse function theorem written in a form that is suitable for the application to the parametric framework.

Theorem B.2. *Let $G : X \to Z$ be a C^1-operator. Let $v \in X$ such that $DG(v) \in \mathcal{L}(X, Z)$ is an isomorphism. We consider the following quantities:*

$$\begin{aligned} &\epsilon := \|G(v)\|_Z \\ &\gamma := \|DG(v)^{-1}\|_{\mathcal{L}(Z,X)} \\ &L(\alpha) := \sup_{x \in \bar{B}(v,\alpha)} \|DG(v) - DG(x)\|_{\mathcal{L}(X,Z)}. \end{aligned} \tag{B.1.4}$$

Suppose that $2\gamma L(2\gamma\epsilon) \leq 1$. Thus the problem (B.1.3) has an unique solution $u \in \bar{B}(v, 2\gamma\epsilon)$ and $DG(u) \in \mathcal{L}(X, Z)$ is invertible with $\|DG(u)^{-1}\|_{Z:X} \leq 2\gamma$. Moreover

$$\|y - u\|_X \leq 2\gamma \|G(y)\|_Z \quad \forall y \in \bar{B}(v, 2\gamma\epsilon). \tag{B.1.5}$$

Proof. Let us consider the following operator:

$$H(x) = x - DG(v)^{-1} G(x) \tag{B.1.6}$$

Clearly each fixed point of H is a zero of the mapping G. In order to apply Banach-Caccioppoli theorem, we have to verify the following hypotheses:

- H maps $\bar{B}(v, 2\gamma\epsilon)$ into itself.
- H is contractive.

For any $x \in \bar{B}(v, 2\gamma\epsilon)$ we can write:

$$H(x) - v = DG(v)^{-1} \left[DG(v)(x - v) - (G(x) - G(v)) \right] - DG(v)^{-1} G(v).$$

Applying the Taylor expansion (B.1.2) with $p = 1$, we obtain:

$$\begin{aligned} \|H(x) - v\| \leq & \|DG(v)^{-1}\| \left[\|DG(v)(x - v)\| + \left\| \int_0^1 (DG(v) - DG(v + t(x - v))(x - v)\, dt \right\| \right] \\ \leq & \gamma \left[\epsilon + \underbrace{L(2\gamma\epsilon) 2\gamma}_{\leq 1} \epsilon \right] \\ \leq & 2\gamma\epsilon \end{aligned}$$

Thus $H : \bar{B}(v, 2\gamma\epsilon) \to \bar{B}(v, 2\gamma\epsilon)$. Let us consider $x, y \in \bar{B}(v, 2\gamma\epsilon)$, then:

$$H(x) - H(y) = DG(v)^{-1} \int_0^1 \left[DG(v) - DG(y + t(x - y)) \right] (x - y)\, dt.$$

Thus

$$\begin{aligned} \|H(x) - H(y)\| \leq & \|DG(v)^{-1}\| \left\| \int_0^1 (DG(v) - DG(v + t(x - y))(x - y)\, dt \right\| \\ \leq & \gamma L(2\gamma\epsilon) \|x - y\| \leq \frac{1}{2} \|x - y\| \end{aligned}$$

We have proved that H is a contraction. So there exists a unique fixed point u in the ball $\bar{B}(v, 2\gamma\epsilon)$. Let $A = DG(v)$ and $B = DG(u) - DG(v)$, from the hypothesis we have that:

$$\|A^{-1}B\| = \|DG(v)^{-1}(DG(u) - DG(v))\| \leq \|DG(v)^{-1}\| \|(DG(u) - DG(v))\| \leq \gamma L(2\gamma\epsilon) \leq \frac{1}{2}.$$

Thus for Lemma B.1, $DG(u)$ is invertible and $\|DG(u)\| \leq 2\gamma$. We conclude proving the (B.1.5)[3]:

$$\begin{aligned}\|u-y\| =& \|H(u)-y\| \\ =& \left\| DG(v)^{-1}\left[-G(y)+\int_0^1 (DG(v)-DG(u+t(y-u)))(u-y)\,dt\right]\right\| \\ \leq& \gamma\left[\|G(y)\|+L(\alpha)\|u-y\|\right]\end{aligned}$$

Thus we finally get: $\|u-y\| \leq \frac{\gamma}{1-\gamma L(\alpha)}\|G(y)\|$. □

Remark B.1. *In theorem B.2 the hypothesis on the differentiability of G can be relaxed: the Fréchet derivative could be replaced by an isomorphism $A \in \mathcal{L}(X,Z)$ and $v \in X$ such that $2\gamma L(2\gamma\epsilon) \leq 1$, with:*

$$\epsilon := \|G(v)\|_Z \quad \gamma := \|DG(v)^{-1}\|_{\mathcal{L}(Z,X)}$$
$$L(\alpha) := \sup_{x\in\bar{B}(v,\alpha)} \frac{\|G(x)-G(y)-A(x-y)\|_Z}{\|x-y\|_X}.$$

Then problem (B.1.3) has an unique solution in $\bar{B}(v, 2\gamma\epsilon)$. Moreover, the estimate (B.1.5) still holds.

Remark B.2. *Theorem B.2 is the fundamental result for the error analysis when we consider approximations of nonlinear problems*[4].

Theorem B.3. *For $v \in X$ and the function $G : X \to Z$ of class C^p, $p \geq 1$, we assume $DG(v) \in \mathcal{L}(X,Z)$ to be an isomorphism and that α satisfies $2\gamma L(\alpha) \leq 1$, with $\gamma = \|DG(v)^{-1}\|_{\mathcal{L}(Z,X)}$. Then there exists a C^p mapping $F : B(G(v), \frac{\alpha}{2\gamma}) \to B(v, \frac{\alpha}{2\gamma})$ such that, for all $z \in B(G(v), \frac{\alpha}{2\gamma})$, we have*

$$G(F(z)) = z, \quad DF(z) = [DG(F(z))]^{-1}. \tag{B.1.7}$$

Moreover, for all z_1, z_2 in $B(G(v), \frac{\alpha}{2\gamma})$

$$\|F(z_1) - F(z_2)\|_X \leq 2\gamma\|z_1 - z_2\|_Z. \tag{B.1.8}$$

The proof is based on the same ideas of the preceding result. In this case the mapping H is:

$$H(x) = x + DG(v)^{-1}(z - G(x)).$$

We conclude recalling a version of the implicit function theorem.

Theorem B.4. *Let Λ, X, Z be three Banach spaces and $G : \Lambda \times X \to Z$ be a C^1 mapping. For a given $(\lambda_0, x_0) \in \Lambda \times X$, we assume $D_x G(\lambda_0, x_0) \in \mathcal{L}(X,Z)$ to be an isomorphism:*

$$\epsilon := \|G(\lambda_0, x_0)\|_Z, \quad \gamma_0 := \|D_\lambda G(\lambda_0, x_0)\|_{\Lambda:Z}$$
$$\gamma_1 := \|D_x G(\lambda_0, x_0)^{-1}\|_{\mathcal{L}(Z,X)} \quad L(\alpha) := \sup_{(\lambda,x)\in\bar{B}((\lambda_0,x_0),\alpha)} \|DG(\lambda_0, x_0) - DG(\lambda, x)\|_{\mathcal{L}(\Lambda\times XZ)}. \tag{B.1.9}$$

Let α be such that $2\gamma L(\alpha) \leq 1$ with $\gamma = \max(\gamma_1, 1 + \gamma_0\gamma_1)$. If $\epsilon < \frac{\alpha}{4\gamma}$, then there exists a unique C^p mapping $g : B(\lambda_0, \frac{\alpha}{4\gamma}) \subset \Lambda \to B(x_0, \alpha) \subset X$ satisfying:

$$G(\lambda, g(\lambda)) = 0 \quad \|g(\lambda) - x_0\|_X \leq 2\gamma(\epsilon + \|\lambda - \lambda_0\|_\Lambda), \tag{B.1.10}$$

for all $\lambda \in B(\lambda_0, \frac{\alpha}{4\gamma})$.

[3] Here we use the fact that $y - H(y) = DG(v)^{-1}G(y)$.

[4] In [15, 122] this is the result used to construct the a posteriori error estimator.

Remark B.3. *Theorem B.4 is potentially extremely important in the RB framework to deal with a non-affine parametric dependence: in fact, in order to reach a parametrically affine form, we have to approximate some parameters of the differential operator. This result provides a powerful tool to study how the perturbation on the parameters influences the solution.*

Functional spaces and main symbols

Throughout the work, we used the following notation:

Ω indicates an open set, if not specified otherwise it is a bounded Lipschitz domain;

$\|\cdot\|_{L^p(\Omega)}$ $1 \leq p \leq \infty$ indicates the Banach norm for the space $L^p(\Omega)$;

$\|\cdot\|_{W^{k,p}(\Omega)}$ $k \in \mathbb{N}$, $1 \leq p \leq \infty$ indicates the Banach norm for the Sobolev space $W^{k,p}(\Omega)$;

$C_0^\infty(\Omega)$ is the set of functions belonging to $C^\infty(\Omega)$ compactly supported in Ω;

$Lip(\Omega) := W^{1,\infty}(\Omega)$ is the set of the Lipschitz functions, $Lip_K(\Omega) := \left\{u \in Lip(\Omega) : \ \|Du\|_{L^\infty(\Omega)} \leq K\right\}$;

let X, Z be two Banach spaces, $(\mathcal{L}(X,Z), \|\cdot\|_{\mathcal{L}(X,Z)})$ indicates the Banach space of the linear and continuous operators from X to Z;

let X be a Banach space, X' indicates the dual space of X, $\langle\cdot,\cdot\rangle_{X'\times X}$ indicates the duality product;

$\hookrightarrow$ is the embedding operator, $X \hookrightarrow Y$ if and only if $X \subset Y$, $\|u\|_Y \leq C\|u\|_X$ for all $u \in X$.

Bibliography

[1] L. Ambrosio and C. de Lellis. A note on admissible solutions of 1d scalar conservation laws and 2d Hamilton-Jacobi equations. *J. Hyperbol. Diff. Eq.*, 31(4):813–826, 2004.

[2] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford Mathematical Monographs, Oxford, 2000.

[3] J. Anderson. *Introduction to Flight*. Mc-Graw-Hill Higher Education, New York, IV edition, 2005.

[4] I. Babuska. Error-bounds for finite element method. *Num. Math.*, 16:322–333, 1971.

[5] F. Ballarin. Ottimizzazione di forma per flussi viscosi tridimensionali in geometrie cardiovascolari. Master's thesis, Politecnico di Milano, 2011. available at `https://www.politesi.polimi.it/handle/10589/30601`.

[6] C. Bardos, A.Y. Le Roux, and J.C. Nedélec. First order quasilinear equations with boundary conditions. *Commun. Part Diff Eq*, 4(9):1017–1034, 1979.

[7] M. Barrault, Y. Maday, N.C. Nguyen, and A.T. Patera. An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations. *C. R. Math. Acad. Sci. Paris*, 339(9):667–672, 2004.

[8] P. Binev, A. Cohen, W. Dahmen, C. Petrova, and P. Wojtaszczyk. Convergence rates for the Greedy Algorithms in Reduced Basis Methods. *SIAM J. Math. Anal.*, 43:1457–1472, 2011.

[9] F. Bouchut and B. Perthame. Kruzkov's estimates for scalar conservation laws revisited. *Trans. Amer. Math. Soc.*, 350:2847–2870, 1998.

[10] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, Paris, I edition, 2010.

[11] F. Brezzi, J. Rappaz, and P.A. Raviart. Finite dimensional approximation of nonlinear problems: branches of nonsingular solutions. *Num. Math.*, 36(1–36), 1980.

[12] A. Buffa, Y. Maday, A. Patera, C. Prud'homme, and G. Turinici. A priori convergence of the Greedy Algorithm for the Parametrized Reduced Basis. *Math. Model. Numer. Anal.*, 46(3):595–603, 2012.

[13] M. Buhmann. *Radial Basis Functions*. Cambridge University Press, Cambridge, 2003.

[14] G. Caloz and J. Rappaz. *Numerical Analysis for Nonlinear and Bifurcation Problems*, volume 5 of *Handbook of Numerical Analysis*. 1997.

[15] C. Canuto, T. Tonn, and K. Urban. A posteriori error analysis of the reduced basis method for non affine parametrized nonlinear pde's. *SIAM J. Numer. Anal.*, 47(3):2001–2022, 2009.

[16] J. Castillo. *Mathematical Aspects of Numerical Grid Generation*, volume 8. SIAM Frontiers in Applied Mathematics, Philadelphia, 1991.

[17] S. Chaturantabut and D. Sorensen. Nonlinear model reduction via discrete empirical interpolation. *SIAM Jour. Sci. Comput.*, 32(5):2737–2764, 2010.

[18] P.G. Ciarlet and J.L. Lions. *Handbook of Numerical Analysis: Finite Element Methods (Part 1)*. North-Holland, Amsterdam, 1991.

[19] B. Cockburn and G. Gripenberg. Continuous dependence on the nonlinearities of solutions of degenerate parabolic equations. *J. Diff. Eq.*, 151:231–251, 1999.

[20] A. Crivellaro. Ricostruzione adattiva di dati sparsi mediante funzioni a simmetria radiale. Master's thesis, Politecnico di Milano, 2012. available at `http://mox.polimi.it/it/progetti/pubblicazioni/viewtesi.php?id=532&en=en`.

[21] R. Dautray and J.L. Lions. *Mathematical Analysis and Numerical Methods for Science and Technology*, volume 5 of *Evolution problems I*. Springer-Verlag, Paris, 1992.

[22] M. de Berg, O. Cheong, M. van Kreveld, and M. Overmars. *Computational Geometry: Algorithms and Applications*. Springer-Verlag, Berlin, 2008.

[23] C. de Boor. *A Practical Guide to Splines*. Springer-Verlag, New York, 2001.

[24] C. de Lellis. Hyperbolic equations and sbv functions. page 10. cedram. Journées equations aux dérivées partielles.

[25] L. Dedé. Reduced basis method for parametrized elliptic advection reaction problems. *J. of Comput. Math.*, 28(1):122–148, 2010.

[26] A. Demlow, D. Leykekhman, A.H. Schatz, and L.B. Wahlbin. Best approximation property in the W^1_∞ norm on graded meshes. *Math. Comp.*, 81:743–764, 2012.

[27] B. Devolder, J. Glimm, J.W. Grove, Y. Kang, Y. Lee, K. Pao, D.H. Sharp, and K. Ye. Uncertainty quantification for multiscale equations. *J. Fluids Eng.*, 124(1):29–41, 2002.

[28] E. H. Dowell and K.C. Hall. Modeling of fluid structure interaction. *Annu. Rev. Fluid. Mech.*, 33:445–490, 2001.

[29] M. Drohmann, B. Haasdonk, and M. Ohlberger. Reduced basis approximation for nonlinear parametrized evolution equations based on empirical operator interpolation. *SIAM J. Sci. Comput.*, 34(2):A937–A969, 2012.

[30] Q. Du and M. Gunzburger. Centroidal Voronoi tessellation based proper orthogonal decomposition analysis. *Internat. Ser. Numer. Math.*, 143:137–150, 2002.

[31] R.O. Duda, P.E.Hart, and D.G. Stork. *Pattern Classification*. II edition, 2001.

[32] J.L. Eftang, M. Grepl, and A.T. Patera. A posteriori error bounds for the empirical interpolation method. *C. R. Math. Acad. Sci. Paris, Série I*, 348(9–10):575–579, 2010.

[33] J.L. Eftang, M.A. Grepl, A.T. Patera, and E. Rónquist. Approximation of parametric derivatives by the Empirical Interpolation Method. *Found. Comput. Math.*, 2012.

[34] J.L. Eftang and B. Stamm. Parameter Multi-Domain *hp*-empirical Interpolation. *Int J Numer Meth Eng*, 90(4):412–428, 2012.

[35] K. Eriksson, D. Estep, P. Hansbo, and C.J. Johnson. *Computational Differential Equations*. Studentlitteratur, Lund, 1996.

[36] L.C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, II edition, 2010.

[37] F. Fang, C.C. Pain, I.M. Navon, A.H. Elsheikh, J. Du, and D. Xiao. Nonlinear Petrov -Galerkin methods for reduced order hyperbolic equations and discontinuous finite element methods. *Elsevier*, 2012. accepted.

[38] G. Farin. *Curves and Surfaces for Computer-Aided Geometric Design: a Practical Guide*. Morgan Kaufmann, London, 2001.

[39] R.L. Fox and H. Miura. An approximate analysis technique for design calculations. *AIAA J.*, 9(1):177–179, 1971.

[40] M. Garavello and B. Piccoli. *Traffic Flow on Networks – Conservation Laws Models*, volume 1. American Institute of Mathematical Sciences, New York, 2006.

[41] F. Gelsomino and G. Rozza. Comparison and combination of reduced order modelling techniques in 3d parametrized heat transfer problems. *Math. Comp. Mod. Dyn. Syst.*, 17(4):373–391, 2011.

[42] G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, III edition, 1996.

[43] W. Gordon and C. Hall. Transfinite element method: blending-function interpolation over arbitrary curved element domains. *NUM MATH*, 21:109–129, 1973.

[44] L. Gosse and C. Makridakis. Two a posteriori error estimates for one-dimensional scalar conservation laws. *SIAM J. Numer. Anal.*, 38(3):964–988, 2000.

[45] L. Graetz. Ueber die Wármeleitungsfáhigkeit von Flüssigkeiten. *Annalen der Physik und Chemie*, 18(79), 1883.

[46] C. Gunther. Reduced basis method for the shape optimization of racing car components. Master's thesis, Aachen-EPFL, 2008. available at `http://infoscience.epfl.ch/record/128720`.

[47] M.D. Gunzburger, J. Peterson, and J.N. Shadid. Reduced order modeling of time-dependent PDEs with multiple parameters in the boundary data. *Comp. Meth. App. Mech.*, 196:1030–1047, 2007.

[48] B. Haasdonk. Convergence rates of the POD-Greedy method. *Math. Model. Numer. Anal.*, 2012. accepted.

[49] B. Haasdonk and M. Ohlberger. Reduced basis method for Finite Volume approximations of parametrized evolution equations. *Math. Model. Numer. Anal.*, 42(2):277–302, 2008.

[50] H.Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educational Psychol.*, 1933.

[51] P. Holmes and J. Lumley. *Turbulence, Coherent Structures, Dynamical Systems.* Cambridge University Press, Cambridge, 1998.

[52] P. Houston, J.A. Mackenzie, and E. Süli. A posteriori error analysis for numerical approximations of Friedrichs systems. *Numer. Math.*, 82:433–470, 1999.

[53] P. Houston, C. Schwab, and E. Süli. Stabilized hp-finite element method for hyperbolic problems. *SIAM J. Numer. Anal.*, 37:1618–1643, 2000.

[54] P. Houston and E. Süli. Adaptive Finite Element Approximation of Hyperbolic Problems. In T. Barth and H. Deconinck, editors, *Error Estimation and Adaptive Discretization Methods in Computational Fluid Dynamics*, volume 25, pages 269–344. 2002.

[55] L. Iapichino. *Reduced Basis Methods for the Solution of Parametrized PDEs in Repetive and Complex Networks with Application to CFD.* PhD thesis, École Polytechnique Fédérale de Lausanne, 2012. N. 5529, *http://infoscience.epfl.ch.*

[56] L. Iapichino, A. Quarteroni, and G. Rozza. A reduced basis hybrid method for the coupling of parametrized domains represented by fluidic networks. *Comput. Method Appl. M.*, 221–222:63–82, 2012.

[57] E. Isaacson and H.B. Keller. *The Symmetric Eigenvalue Problem.* Dover, New York, 1994.

[58] K. Ito and S.S. Ravindran. A reduced order method for simulation and control of fluid flows. *J Comput Phys*, 142(2):403–425, 1998.

[59] A. D. Izaak. Kolmogorov widths in finite dimensional spaces with mixed norms. *Math. Notes*, 55(1):43–52, 1994.

[60] C. Johnson. Adaptive finite elements for conservation laws. In A. Quarteroni, editor, *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations*, volume 1697 of *Lecture Notes in Mathematics*. Berlin. Springer-Verlag.

[61] C. Johnson and A. Szepessy. Adaptive finite element methods for conservation laws based on a posteriori error estimates. *Comm. Pure Appl. Math.*, 48:199–234, 1995.

[62] K. Karhunen. Zur spektraltheorie stochastischer prozesse. *Annales Academiae Scientiarum Fennicae*, 37, 1946.

[63] A. Koshakji. Free form deformation techniques for 3d shape optimization problems. Master's thesis, Politecnico di Milano-EPFL, 2009. available at `https://www.politesi.polimi.it/handle/10589/15401`.

[64] D. Kröner and M. Ohlberger. A posteriori error estimates for upwind finite volume schemes for nonlinear conservation laws in multi dimensions. *Math. Comput.*, 69(229):25–39, 1999.

[65] S.N. Kruzkov. First-order Quasilinear Equations in several independent variables. *Mat. Sbornik*, 123:228–255, 1970.

[66] P. Krysl, S. Lall, and J.E. Mardsen. Dimensional model reduction in non-linear finite element dynamics of solids and structures. *Int. J. Num. Meth. Eng.*, 51:479–504, 2001.

[67] M. Laforest, M.A. Christon, and T. E. Voth. Error indicators and estimators for hyperbolic problems. Technical report, Sandia National Laboratory, 2003.

[68] T. Lassila, A. Manzoni, and G. Rozza. Reduction strategies for shape dependent inverse problems in haemodynamics. Technical report, EPFL, 2012. To appear on System Modelling and Optimization, Springer 2012.

[69] T. Lassila and G. Rozza. Parametric Free-Form shape design with PDE models and reduced basis method. *COMPUT METHOD APPL M*, 199(23–24):1583–1592, 2010.

[70] T. M. Lassila, A. Manzoni, and G. Rozza. On the approximation of stability factors for general parametrized partial differential equations with a two-level affine decomposition. *Math. Mod. Num. Anal.*, 46:1555–1576, 2012.

[71] R.J. LeVeque. *Numerical Methods for Conservation Laws*. Birkhauser Verlag, Berlin, 1992.

[72] R.J. LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press, Cambridge, 2002.

[73] J.L. Lions and E. Magenes. *Problémes aux Limites non Homogénes*. Dunod, Paris, 1968.

[74] M. M. Loeve. *Probability Theory*. Van Nostrand, London, 1955.

[75] A.E. Lóvgren, Y. Maday, and E.M. Rónquist. A reduced basis element method for the steady Stokes problem. *ESAIM, Math. Model. Numer. Anal.*, 40(3):529–552, 2006.

[76] J. Lumley and P. Blossey. Control of turbulence. *Annu. Rev. Fluid. Mech.*, 30:311–327, 1998.

[77] Y. Maday. Reduced-basis method for the rapid and reliable solution of partial differential equations. Madrid, 2006.

[78] A. Manzoni. *Reduced Models for Optimal Control, Shape Optimization and Inverse Problems in Haemodynamics*. PhD thesis, École Polytechnique Fédérale de Lausanne, 2012. N. 5402, *http://infoscience.epfl.ch*.

[79] A. Manzoni, D. B. Huinh, and G. Rozza. Reduced basis approximation and a posteriori error estimation for stokes flows in parametrized geometries: roles of the inf-sup stability constants. *Numer. Math.*, 2010. submitted.

[80] A. Manzoni, A. Quarteroni, and G. Rozza. Certified reduced basis approximation for parametrized partial differential equations and applications. *J. Math. Ind.*, 1(3):1–44, 2011.

[81] A. Manzoni, A. Quarteroni, and G. Rozza. Model reduction techniques for fast blood flow simulation in parametrized geometry. *Int. J. Numer. Meth. Bio. Eng.*, 2011.

[82] A. Manzoni, A. Quarteroni, and G. Rozza. Shape optimization for viscous flows by reduced basis methods and free-form deformation. *Int. J. Numer. Meth. Fluids*, 2011. in press.

[83] A. Manzoni, A. Quarteroni, and G. Rozza. Computational Reduction for Parametrized PDEs: Strategies and Applications. *Milan J. Math.*, 2012. EPFL MATHICSE Report 15.2012.

[84] MATLAB. *version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts, 2010.

[85] B Mohammadi and O. Pironneau. *Applied Shape Optimal Design*. Oxford University Press, Oxford, 2001.

[86] A.M. Morris, C. B. Allen, and T.C.S. Rendall. CFD-based optimization of aerofoils using radial basis functions for domain element parametrization and mesh deformation. *Int. J. Numer. Methods Fluids*, 58:827–860, 2008.

[87] N. C. Nguyen, G. Rozza, and A.T. Patera. Reduced basis approximation and a posteriori error estimation for the time-dependent viscous Burgers equation. *Calcolo*, 46(3):157–185, 2009.

[88] N.C. Nguyen, K. Veroy, and A.T. Patera. *Certified real-time solution of parametrized partial differential equations*. Handbook of Materials Modeling, S. Yip Ed., Kluwer Academic Publishing, Springer, Dordrecht, 2005.

[89] A.K. Noor. Recent advances in reduction methods for non-linear problems. *Comput. Struct.*, 13:31–44, 1981.

[90] S. Osher. Riemann solvers, the entropy condition, and difference approximations. *SIAM J. Numer. Anal.*, 22:947–961, 1985.

[91] C. D. Pagani and S. Salsa. *Analisi Matematica*, volume I. Masson, Milan, 1991.

[92] B.N. Parlett. *The Symmetric Eigenvalue Problem*. SIAM, Philadelphia, 1998.

[93] A.T. Patera and G. Rozza. *Reduced Basis Approximation and A Posteriori Error Estimation for Parametrized Partial Differential Equations*. to appear in MIT Pappalardo Graduate Monographs in Mechanical Engineering, 2009. ©Massachusetts Institute of Technology, Version 1.0.

[94] A.T. Patera, K. Urban, and M. Yano. A space-time certified reduced basis method for Burgers' equation. Technical report, Mit, 2012.

[95] B. Perthame and M. Westdickenberg. Total oscillation diminishing property for scalar conservation laws. *Numer. Math.*, 100:331–349, 2005.

[96] P. Pettersson, G. Iaccarino, and J. Nordstrom. Numerical analysis of the Burgers' equation in the presence of uncertainty. *J. Comput. Phys.*, 2009.

[97] A. Quarteroni. *Numerical Models for Differential Problems*. Springer-Verlag, Milan, I edition, 2009.

[98] A. Quarteroni, L. Sacco, and F. Saleri. *Numerical Mathematics*, volume 37 of *Texts in Applied Mathematics*. Springer, II edition, 2007.

[99] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations*. Springer-Verlag, Berlin-Heidelberg, II edition, 1994.

[100] S.S. Ravindran. Reduced-order adaptive controllers for fluid flows using POD. *J. Sci. Comp.*, 15(4):457–478, 2000.

[101] S.S. Ravindran. A reduced order approach to optimal control of fluids flow using proper orthogonal decomposition. *Int. J. Num. Meth. Fluids*, 34(5):425–448, 2002.

[102] rbMIT Library. `http://augustine.mit.edu/methodology/methodology_rbmit_system.htm`. *MIT, Cambridge*, 2007–2010. ©Massachusetts Institute of Technology.

[103] G. Rozza. Optimization, control and shape design of an arterial bypass. *Int. J Numer. Meth. in Fluids*, 47(10–11):1411–1419, 2005.

[104] G. Rozza. *Shape design by optimal flow control and reduced basis techniques: applications to bypass configurations in haemodynamics*. PhD thesis, École Polytechnique Fédérale de Lausanne, 2005. N. 3400, *http://infoscience.epfl.ch*.

[105] G. Rozza. Reduced basis methods for Stokes equations in domains with non-affine parametric dependence. *Comput. Vis. Sci.*, 12(1):23–35, 2009.

[106] 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. Meth. Eng*, 15:229–275, 2008.

[107] G. Rozza, D.B.P. Huynh, A.T. Patera, and S.Sen. A successive constraint linear optimization method for lower bpunds of parametric coercivity and inf-sup stability constants. *C R Acad Sci Paris Ser I*, 345:473–478, 2008.

[108] F. Saleri, P. Gervasio, G. Rozza, and A. Manzoni. `MLife`, a Matlab® library for Finite Elements, tutorial (in progress). 2010-2011. Copyright Politecnico di Milano.

[109] S. Salsa. *Equazioni Differenziali. Metodi, Modelli e Applicazioni*. Springer-Verlag, Milan, II edition, 2010.

[110] S. Salsa and G. Verzini. *Equazioni a derivate parziali. Complementi ed esercizi*. Springer-Verlag, Milan, 2005.

[111] J. A. Samareh. Aerodynamic shape optimization based on Free-Form Deformation. *AIAA 2004-4630*, 2004.

[112] T.W. Sedemberg and S.R. Parry. Free-Form of solid geometric models. *Comput. Graph.*, 20(4):109–129, 1986.

[113] K. Steih and K. Urban. Space-time reduced basis methods for time periodic parabolic evolution problems. Technical report, University of Ulm, 2012.

[114] T.Bui-Thanh, M. Damodaran, and K. Willcox. Proper Orthogonal Decomposition extensions for parametric applications in transonic aerodynamics. In *Proc. 16th AIAA Comput. Fluid Dynamics*, 2003.

[115] T. Tonn. *Reduced Basis Method for non-affine Elliptic parametrized PDEs*. PhD thesis, Universitat Ulm, 2011.

[116] T. Tonn and K. Urban. A reduced-basis method for solving parameter-dependent convection diffusion problems around rigid bodies. ECCOMAS CFD 2006 Proceedings, 2006. TU Delft.

[117] L. Trefethen. *Numerical Linear Algebra.* SIAM, Philapeldhia, 1997.

[118] A.A. Trezzini. Reduced Basis Method for 3D Problems governed by parametrized PDEs and Applications. Master's thesis, Politecnico di Milano, 2009/2010. available at `https://www.politesi.polimi.it/handle/10589/2602`.

[119] G. Turk and J.. O'Brien. Shape trasformation using variational implicit functions. In *ACM SIG-GRAPH 99*, 1999. available at `http://graphics.berkeley.edu/papers/Turk-STU-1999-08/Turk-STU-1999-08.pdf`.

[120] K. Urban and A.T. Patera. An improved error bound for Reduced Basis approximation of linear parabolic problems. Technical report, MIT, 2012. submitted to Math. Comp.

[121] K. Veroy. *Reduced Basis Methods Applied to Problems in Elasticity: Analysis and Applications.* PhD thesis, MIT, 2003.

[122] K. Veroy and A.T. Patera. Certified real time solution of the parametrized steady incompressible navier stokes equations: rigorous reduced basis a posteriori error bounds. *Int.J. Numer. Meth. Fluids*, 47(8–9):773–788, 2005.

[123] K. Veroy, C. Prud'homme, and A.T. Patera. Reduced-basis approximation of the viscous Burgers equation: Rigorous a posteriori error bounds. *C. R. Acad. Sci. Paris*, 337(9):619–624, 2003.

[124] K. Veroy, C. Prud'homme, D.V. Rovas, and A.T. Patera. A posteriori error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations. In *Proc. 16th AIAA Comput. Fluid Dynamics*, 2003.

[125] K. Veroy, D. Rovas, and A. T. Patera. A posteriori error estimation for Reduced-Basis approximation of parametrized elliptic coercive partial differential equations: "convex inverse" bound conditioners. *ESAIM: Control, Optimization and Calculus of Variations*, 8:1007–1028, 2002.

[126] Z. Zhang and A. Naga. A new finite element gradient recovery method: Superconvergence property. *SIAM J. Sci. Comput.*, 26:1192–1213, 2005.

[127] O. C. Zienkiewicz and J. Z. Zhu. The superconvergent patch recovery snd a posteriori error estimates. Part I: the recovery technique. *Int. J. Num. Meth. Eng.*, 33:1331–1364, 1992.