


Alfio Quarteroni
Andrea Manzoni
Federico Negri



Reduced Basis Methods for Partial Differential Equations

An Introduction



Springer

TEXT
UN

UNITEXT – La Matematica per il 3+2

Volume 92

Editor-in-chief

A. Quarteroni

Series editors

L. Ambrosio

P. Biscari

C. Ciliberto

M. Ledoux

W.J. Runggaldier

More information about this series at <http://www.springer.com/series/5418>

Alfio Quarteroni · Andrea Manzoni · Federico Negri

Reduced Basis Methods for Partial Differential Equations

An Introduction

 Springer

Alfio Quarteroni
Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland

Andrea Manzoni
Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland

Federico Negri
Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland

ISSN 2038-5722 ISSN 2038-5757 (electronic)
UNITEXT – La Matematica per il 3+2
ISBN 978-3-319-15430-5 ISBN 978-3-319-15431-2 (eBook)
DOI 10.1007/978-3-319-15431-2

Library of Congress Control Number: 2015930287

Springer Cham Heidelberg New York Dordrecht London

© Springer International Publishing Switzerland 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Cover illustration: The cover figure displays a set of reduced basis functions for an advection-diffusion-reaction boundary value problem in a rectangular computational domain.

Cover Design: Simona Colombo, Giochi di Grafica, Milano, Italy

Springer International Publishing AG Switzerland is part of Springer Science+Business Media
(www.springer.com)

Preface

Reduced basis (RB) methods represent a very efficient approach for the numerical approximation of problems involving the repeated solution of differential equations arising from engineering and applied sciences. Noteworthy examples include partial differential equations (PDEs) depending on several parameters, PDE-constrained optimization, and optimal control and inverse problems.

In all these cases, reducing the severe computational complexity is crucial. With this in mind, over the past four decades, reduced-order models (ROMs) have been developed aiming at replacing the original large-dimension numerical problem (typically called high-fidelity approximation) by a reduced problem of substantially smaller dimension. Strategies to generate the reduced problem from the high-fidelity one can be manifold, depending on the context.

The strategy adopted in RB methods consists in the projection of the high-fidelity problem upon a subspace made of specially selected basis functions, representing a set of high-fidelity solutions corresponding to suitably chosen parameters. Pioneering works in this area date back to the late 1970s (e.g., B.O. Almroth et al. [5, 6], D. Nagy [193], A.K. Noor and J.M. Peters [201, 202, 203, 204] and address linear and nonlinear structural analysis problems. The first theoretical analysis of RB methods in connection with the use of the continuation method for parametrized equations was presented by J.P. Fink and W.C. Rheinboldt [109, 110] in the mid 1980s. Extensions to problems in fluid dynamics are primarily due to the contributions of Peterson [210] and Gunzburger [124] in the late 1980s.

The method was set on a more general and sound mathematical ground in the early 2000s thanks to the seminal work of A.T. Patera, Y. Maday and coauthors [214, 255]. Their work has led to a decisive improvement in the computational aspects of RB methods owing to an efficient criterion for the selection of the basis functions, a systematic splitting of the computational procedure into an offline (parameter-independent) and an online (parameter-dependent) phase, and the use of a posteriori error estimates that guarantee certified numerical solutions for the reduced problem. These have become the essential constituents of the RB methods now most widely used. Often, they are also embedded into more general reduced-order models.

RB methods have witnessed a spectacular effervescence in the past decade. Additional achievements during that time relate to the treatment of nonlinear and/or parametrically nonaffine problems by the so-called empirical interpolation method and its several extensions. This has substantially improved RB methods, making possible their application to a broad variety of complex problems such as time-dependent problems, optimal control and design problems, and real-time computing.

This is the first textbook to provide a basic mathematical introduction to RB methods. We present a general formulation of RB methods, analyze their fundamental theoretical properties, and discuss their algorithmic and implementation aspects, highlighting their built-in algebraic and geometric structures. More specifically, we carry out both a priori and a posteriori error analysis, formulate strategies for the construction of accurate reduced basis spaces, and analyze offline-online decomposition strategies to ensure the reduction of computational complexity. The entire mathematical discussion is made more stimulating by the use of several representative examples of applicative interest, in the context of both linear and nonlinear PDEs.

The authors are grateful to Charbel Farhat, Yvon Maday, and Anthony Patera for being source of inspiration and for many fruitful discussions on reduced-order models. We also acknowledge David Amsallem, Luca Dedè, Simone Deparis and Toni Lassila for the great amount of time that they have spent with the authors talking about different subjects covered in this book and, last but not least, Gianluigi Rozza for having introduced the last two authors to the subject. In addition, special thanks are due to Francesca Bonadei and Francesca Ferrari of Springer Italia for their invaluable help in the preparation of the manuscript.

Lausanne, Switzerland
May 2015

Alfio Quarteroni
Andrea Manzoni
Federico Negri

Contents

1	Introduction	1
1.1	Numerical Simulation and Beyond	2
1.2	The Need for Reduction	4
1.3	Reduced Basis Methods for PDEs at a Glance	5
1.4	Accuracy and Computational Efficiency of RB Methods	7
1.5	Content of the Book	8
2	Representative Problems: Analysis and (High-Fidelity)	
	Approximation	11
2.1	Four Problems	11
2.1.1	Advection-Diffusion-Reaction Equation	12
2.1.2	Linear Elasticity Equations	12
2.1.3	Stokes Equations	13
2.1.4	Navier-Stokes Equations	13
2.2	Formulation and Analysis of Variational Problems	14
2.2.1	Strongly Coercive Problems	14
2.2.2	Weakly Coercive (or Inf-Sup Stable) Problems	16
2.2.3	Saddle-Point Problems	17
2.3	Analysis of Three (out of Four) Problems	20
2.3.1	Advection-Diffusion-Reaction Equation	20
2.3.2	Linear Elasticity Equations	22
2.3.3	Stokes Equations	22
2.4	On the Numerical Approximation of Variational Problems	23
2.4.1	Strongly Coercive Problems	23
2.4.2	Algebraic Form of (P_1^h)	25
2.4.3	Computation of the Discrete Coercivity Constant	26
2.4.4	Weakly Coercive Problems	27
2.4.5	Algebraic Form of (P_2^h)	29
2.4.6	Computation of the Discrete Inf-Sup Constant	30
2.4.7	Saddle-Point Problems	31
2.4.8	Algebraic Form of (P_3^h)	33

2.5	Finite Element Spaces	33
2.6	Exercises	35
3	RB Methods: Basic Principles, Basic Properties	39
3.1	Parametrized PDEs: Formulation and Assumptions	39
3.2	High-Fidelity Discretization Techniques	41
3.3	Reduced Basis Methods	43
3.3.1	Galerkin RB Method	45
3.3.2	Least-Squares RB Method	48
3.4	Algebraic Form of Galerkin and Least-Squares RB Problems	51
3.4.1	Galerkin RB Case	51
3.4.2	Least-Squares RB Case	54
3.5	Reduction of Computational Complexity: Offline/Online Decomposition	55
3.6	A Posteriori Error Estimation	56
3.6.1	A Relationship between Error and Residual	57
3.6.2	Error Bound	59
3.7	Practical (and Efficient) Computation of Error Bounds	60
3.7.1	Computing the Norm of the Residual	61
3.7.2	Computing the Stability Factor by the Successive Constraint Method	62
3.7.3	Computing the Stability Factor by Interpolatory Radial Basis Functions	65
3.8	An Illustrative Numerical Example	67
3.9	Exercises	71
4	On the Algebraic and Geometric Structure of RB Methods	73
4.1	Algebraic Construction and Interpretation	73
4.1.1	Algebraic Interpretation of the G-RB Problem	74
4.1.2	Algebraic properties of the G-RB Problem	75
4.1.3	Least-Squares and Petrov-Galerkin RB Problems	77
4.2	Geometric Interpretation	79
4.2.1	Projection and Bases	79
4.2.2	Matrix Characterization of Projection Operators	80
4.2.3	Orthogonal and Oblique Projection Operators	81
4.2.4	The Galerkin Case	82
4.2.5	The Petrov-Galerkin Case	84
4.3	Exercises	86
5	The Theoretical Rationale Behind	87
5.1	The Solution Manifold	87
5.2	When is a Problem Reducible?	89
5.3	Smoothness of the Solution Set	90
5.3.1	Continuity and Compactness	90

5.3.2	Differentiability of the Solution Map and Sensitivity Equations	93
5.4	Dimensionality of the Solution Set	96
5.5	Dimensionality and Analiticity	98
5.5.1	Analiticity of the Solution Map: an Instance	98
5.5.2	Kolmogorov n -width and Analiticity	101
5.6	Kolmogorov n -width and Parametric Complexity	104
5.7	Lagrange, Taylor and Hermite RB Spaces	109
5.8	Exercises	111
6	Construction of RB Spaces by SVD-POD	115
6.1	Basic Notions on Singular Value Decomposition	115
6.1.1	SVD and Low-Rank Approximations	117
6.2	Interlude	119
6.2.1	Image Compression	119
6.2.2	Principal Component Analysis	121
6.3	Proper Orthogonal Decomposition	123
6.3.1	POD for Parametrized Problems	124
6.3.2	POD with Energy Inner Product	127
6.4	\mathcal{P} -continuous Analogue of POD	128
6.5	Back to the Discrete Setting	133
6.6	Our Illustrative Numerical Example Revisited	135
6.7	More on Reducibility	136
6.8	Exercises	139
7	Construction of RB Spaces by the Greedy Algorithm	141
7.1	Greedy Algorithm: an Algebraic Perspective	141
7.1.1	The Idea Behind Greedy Algorithms	142
7.1.2	The Weak Greedy Algorithm	142
7.2	Our Illustrative Numerical Example Revisited	145
7.3	An Abstract Formulation of the Greedy Algorithm	147
7.4	A Priori Error Analysis	150
7.5	Numerical Assessment of a Priori Convergence Results	151
7.6	Exercises	154
8	RB Methods in Action: Setting up the Problem	155
8.1	Going from the Original to the Reference Domain	155
8.2	Change of Variables Formulas	156
8.2.1	Extension to the Vector Case	159
8.3	Advection-Diffusion-Reaction, Case I: Heat Transfer	159
8.3.1	Reference Configuration and Affine Transformations	161
8.3.2	Weak Formulation on the Reference Domain	162
8.3.3	Dealing with Nonhomogeneous Boundary Conditions	164
8.4	Advection-Diffusion-Reaction, Case II: Mass Transfer with Parametrized Source	166

8.5	Advection-Diffusion-Reaction, Case III: Mass Transfer in a Parametrized Domain	167
8.5.1	More on the Transformation of Vector Fields	170
8.6	Linear Elasticity: An Elastic Beam	172
8.7	Fluid Flows, Case I: Backward-Facing Step Channel	173
8.7.1	Reference Domain and Affine Transformation	175
8.7.2	Weak Formulation on the Reference Domain	176
8.8	Fluid Flows, Case II: Sudden Expansion Channel	177
8.9	Problems' Features at a Glance	178
8.10	Exercises	179
9	RB Methods in Action: Computing the Solution	181
9.1	Heat Transfer: Results	181
9.2	An Elastic Beam: Results	184
9.3	Backward-Facing Step Channel, Stokes Flow: Results	186
9.3.1	RB Approximation of Parametrized Stokes Equations	187
9.3.2	A Posteriori Error Estimation	190
9.3.3	Numerical Results: Backward-Facing Step Channel	191
10	Extension to Nonaffine Problems	193
10.1	Empirical Interpolation Method	193
10.1.1	Polynomial Interpolation vs. Empirical Interpolation	194
10.1.2	Empirical Interpolation	195
10.1.3	EIM Algorithm	196
10.2	Error Analysis for the Empirical Interpolation	199
10.2.1	Practical Implementation	201
10.3	Discrete Empirical Interpolation	203
10.4	EIM-G-RB Approximation of Nonaffine Problems	205
10.5	Mass Transfer with Parametrized Source: Results	208
10.5.1	Comparison of EIM and DEIM	209
10.5.2	(D)EIM-G-RB Approximation	210
10.6	Mass Transfer in a Parametrized Domain: Results	212
10.7	Exercises	213
11	Extension to Nonlinear Problems	215
11.1	Parametrized Nonlinear PDEs	215
11.1.1	Navier-Stokes Equations	217
11.1.2	A Semilinear Elliptic PDE	218
11.2	High-Fidelity Approximation	219
11.2.1	Newton's Method	220
11.2.2	Algebraic Formulation	221
11.3	Reduced Basis Approximation	222
11.3.1	Algebraic Formulation	223
11.3.2	Galerkin Projection	224
11.3.3	LS-RB: Newton then Least-Squares	224

11.3.4 LS-RB Revisited: Least-Squares then Gauss-Newton	225
11.4 Reduction of Computational Complexity	226
11.5 A Posteriori Error Estimation for Nonlinear Problems	228
11.6 Application to the Steady Navier-Stokes Equations	232
11.6.1 RB Approximation of the Navier-Stokes Equations	233
11.6.2 A posteriori Error Estimation	235
11.7 Numerical Results: Backward-Facing Step Channel	235
11.8 Numerical Results: Sudden Expansion Channel	237
11.9 Numerical results: a Simplified Bypass Graft	239
11.10 Exercises	242
12 Reduction and Control	245
12.1 Parameter-Dependent PDE-Constrained Optimization	245
12.2 Parametric Optimization Problems	248
12.2.1 Reduction Strategies	251
12.3 Application to an Optimal Flow Control Problem	253
12.4 Parametrized Optimal Control Problems	256
12.4.1 Reduction Strategies	257
12.4.2 A Posteriori Error Estimation	260
12.5 Application to an Optimal Heat Transfer Problem	261
Appendix A Basic Theoretical Tools	265
A.1 Linear Maps, Functionals and Bilinear Forms	265
A.2 Hilbert Spaces	268
A.3 Adjoint Operators	269
A.4 Compact Operators	270
A.5 Differentiation in Linear Spaces	272
A.6 Sobolev Spaces	273
A.6.1 Square-Integrable Functions	274
A.6.2 The Spaces $H^1(\Omega)$ and $H_0^1(\Omega)$	274
A.7 Bochner Spaces	277
A.8 Polynomial Interpolation and Orthogonal Polynomials	277
References	281
Index	293

Chapter 1

Introduction

Thanks to the achievements of numerical analysis and scientific computing in the last decades, numerical simulations in engineering and applied sciences have gained an ever increasing importance. In several fields, from aerospace and mechanical engineering to life sciences, numerical simulations of partial differential equations (PDEs) currently provide a virtual platform ancillary to material/mechanics testing or *in vitro* experiments. These are in turn useful either for (i) the prediction of input/output response or (ii) the design and optimization of a system [237, 238, 216]. The constant increase of available computational power, accompanied by the progressive improvement of algorithms for solving large linear systems, make nowadays possible the numerical simulation of complex, multiscale and multiphysics phenomena by means of *high-fidelity* (or *full-order*) approximation techniques such as the finite element method, finite volumes, finite differences or spectral methods. However, this might be quite demanding, because it involves up to $O(10^6 - 10^9)$ degrees of freedom and several hours (or even days) of CPU time, also on powerful hardware parallel architectures.

High-fidelity approximation techniques can become prohibitive when we expect them to deal quickly and efficiently with the repetitive solution of PDEs. This is, e.g., the case of PDEs depending on parameters, that from now on we will call *parametrized PDEs*. Input parameters of relevant interest in fluid and solid mechanics problems are, e.g., (i) the Reynolds number in nonlinear viscous flows governed by Navier-Stokes equations, (ii) the Grashof or Prandtl numbers in unsteady natural convection problems governed by Boussinesq equations, or (iii) the elastic moduli of a solid structure. In these three cases, evaluating the behavior of the system by means of a high-fidelity technique, such as the finite element (FE) method, is computationally expensive because it entails the solution of very large (nonlinear) algebraic systems, arising from the discretization of the underpinning PDE.

For this class of problems, *reduced-order modeling* – alternatively named *model order reduction* in the literature – is a generic expression used to identify any approach aimed at replacing the high-fidelity problem by one featuring a much lower numerical complexity. Being able to evaluate the solution of this latter problem, for any new parameter instance, at a cost that is independent of the dimension of

the original high-fidelity problem, is the key to the (computational) success of any *reduced-order model* (ROM).

Reduced basis (RB) methods represent a remarkable instance of reduced-order modeling techniques. They exploit the parametric dependence of the PDE solution by combining a handful of high-fidelity solutions (or *snapshots*) computed for a (possibly small) set of parameter values. By this approach, a very large algebraic system is replaced by a much smaller one, whose dimension is related to the number of snapshots.

The strong computational speedup achievable today by RB methods allows to tackle a wide range of problems, due to very short CPU times and limited storage capacities demanded. A short *aperçu* of this class of problems is offered in the next section.

1.1 Numerical Simulation and Beyond

Solving a *forward problem* for a given PDE consists in computing an approximate solution of the PDE and, possibly, some output of interest corresponding to a specific set of input data. In several applications, it might be necessary to compute such solution a number of times (whenever some *input* parameters change). Throughout this book, the input parameter vector will be denoted by $\boldsymbol{\mu} \in \mathcal{P} \subset \mathbb{R}^P$; the parameter space \mathcal{P} represents a closed and bounded subset of the Euclidean space \mathbb{R}^P , $P \geq 1$.

The field variable given by the exact solution of a parametrized PDE can be seen as a *map* $u : \mathcal{P} \rightarrow V$ that to any $\boldsymbol{\mu} \in \mathcal{P}$ associates the solution $u(\boldsymbol{\mu})$ belonging to a suitable functional space V . The field variable may represent a physical quantity such as, e.g., a distribution function, temperature or concentration, a fluid potential, pressure or velocity, a deformation of a solid structure, etc.

In addition, the parameters may characterize geometric features of the computational domain, or some physical or material properties of the model at hand, or else initial and boundary conditions and source terms.

A further relevant quantity may be represented by some *output* of interest, which will be denoted by $z = z(\boldsymbol{\mu})$. The output depends on the input parameter vector through the field variable, namely $z(\boldsymbol{\mu}) = J(u(\boldsymbol{\mu}); \boldsymbol{\mu})$, where $J : V \times \mathcal{P} \rightarrow \mathbb{R}^m$, $m \geq 1$ is a (either linear or nonlinear) *functional*; if $m = 1$ the output is a scalar quantity, otherwise it is a vector. Very often, evaluating the output $z(\boldsymbol{\mu})$, rather than computing the whole field variable $u(\boldsymbol{\mu})$, represents the goal of a parametric analysis.

With the only exception of a few classes of problems, even when $J : V \rightarrow \mathbb{R}^m$ is a linear map, that is, J depends linearly on the field variable u , $z(\boldsymbol{\mu})$ might depend nonlinearly on $\boldsymbol{\mu}$ because of a possible nonlinear dependence of u on $\boldsymbol{\mu}$.

An example of parametrized PDE is provided by the following advection-diffusion equation, modeling a heat transfer problem: given $\boldsymbol{\mu} \in \mathcal{P} \subset \mathbb{R}^P$, solve

$$\begin{cases} -\operatorname{div}(k\nabla u) + \mathbf{b} \cdot \nabla u = s(\boldsymbol{\mu}) & \text{in } \Omega \\ u = 0 & \text{on } \Gamma_D \\ k\nabla u \cdot \mathbf{n} = h(\boldsymbol{\mu}) & \text{on } \Gamma_N, \end{cases} \quad (1.1)$$

where $\Omega \subset \mathbb{R}^d$, $d = 2, 3$ denotes a spatial domain (by that we mean an open bounded connected region with Lipschitz boundary), $\partial\Omega = \Gamma_D \cup \Gamma_N$ its boundary, and \mathbf{n} the outward unit normal on $\partial\Omega$. Here $k > 0$ is the diffusion coefficient, while \mathbf{b} is a transport field (also called the advection). The source term $s = s(\boldsymbol{\mu})$ and the flux $h = h(\boldsymbol{\mu})$ across the boundary Γ_N are given by two parameter-dependent functions $f, h : \mathcal{P} \times \mathbb{R}^d \rightarrow \mathbb{R}$. In such a situation, we might want to, for instance:

- (I1) compute the temperature field in a given subregion of Ω by varying, e.g., the source intensity or heat flux across Γ_N ;
- (I2) evaluate outputs such as the average of the field variable over a given portion Ω_{obs} of the domain or through the boundary

$$z_1(\boldsymbol{\mu}) = \int_{\Omega_{obs}} u(\boldsymbol{\mu}) d\Omega, \quad z_2(\boldsymbol{\mu}) = \int_{\Gamma} u(\boldsymbol{\mu}) d\Gamma,$$

or otherwise first-order quantities such as the flux across a portion of the boundary, or the energy

$$z_3(\boldsymbol{\mu}) = \int_{\Gamma_D} k \frac{\partial u(\boldsymbol{\mu})}{\partial \mathbf{n}} d\Gamma, \quad z_4(\boldsymbol{\mu}) = \int_{\Omega} |\nabla u(\boldsymbol{\mu})|^2 d\Omega,$$

respectively. Other relevant cases when dealing with fluid flows are, e.g., vorticity, drag and lift coefficients;

- (I3) reach a specified target, such as a temperature distribution u_d over the domain, by controlling the system, e.g. through a distributed source term playing the role of a *control* function. This can be achieved by solving a minimization problem,

$$\hat{\boldsymbol{\mu}} = \arg \min_{\mathcal{P}_{ad} \subseteq \mathcal{P}} \int_{\Omega_{obs}} |u(\boldsymbol{\mu}) - u_d|^2 d\Omega$$

subject to the constraint (1.1); $\mathcal{P}_{ad} \subseteq \mathcal{P}$ denotes a subset of the parameter space, referred to as the *admissible control* space.

In several contexts we might be interested in *controlling* a system, that is, either minimize or maximize a physical quantity expressing some desired properties and/or performances of the underlying PDE system by acting on some *control variables* (such as sources, boundary conditions, etc.) or on the shape of the domain itself. In the former case, we deal with *optimal control* problems, while we refer to the latter as *shape optimization* or *optimal design* problems.

Other possible problems of interest are, for instance, *identification or inverse problems*. Whenever some parameters characterizing a system are unknown or uncertain, their values (and/or distributions) may be inferred from indirect observa-

tions or measures by solving an *inverse problem*: given an *observed* or *measured* output, the values of the input resulting in that observation can be found by driving the solution of the PDE – and the corresponding *computed* output – as near as possible to the observed output, e.g. by minimizing a suitable measure of their distance.

Since the most common numerical strategies for the previous problems are based on the use of suitable iterative procedures, optimization and inverse problems under PDE constraints can be recast in the so-called *many-query context*. This involves several input/output evaluations as well as many repeated solutions of the given PDE. Other remarkable instances of many-query problems arise when dealing, e.g., with sensitivity analysis of PDE solutions with respect to input data, parametric studies and statistical analyses employed in the design of experiments.

The message that we would like to convey is that, despite the massive computer resources currently available, problems involving the repeated solution of PDEs on different data settings (many-query context) or requiring a numerical solution within a *real time* context – or at least very rapidly – still represent a challenge for classical numerical techniques.

1.2 The Need for Reduction

The accurate high-fidelity approximation of a PDE entails, from an algebraic standpoint, the solution of a large linear system, whose dimension is given by the number N_h of degrees of freedom required to represent the solution over a suitable finite dimensional space.

Given $\boldsymbol{\mu} \in \mathcal{P}$, we represent this high-fidelity system under the form

$$\mathbb{A}_h(\boldsymbol{\mu})\mathbf{u}_h(\boldsymbol{\mu}) = \mathbf{f}_h(\boldsymbol{\mu}) \quad (1.2)$$

where $\mathbb{A}_h(\boldsymbol{\mu}) \in \mathbb{R}^{N_h \times N_h}$ and $\mathbf{f}_h(\boldsymbol{\mu}) \in \mathbb{R}^{N_h}$ are a $\boldsymbol{\mu}$ -dependent matrix and vector, respectively. Problem (1.2) may for instance arise from the discretization of a second-order, time-independent, linear PDE like (1.1): in this case, the *stiffness* matrix $\mathbb{A}_h(\boldsymbol{\mu})$ encodes the differential operator, $\mathbf{f}_h(\boldsymbol{\mu})$ the data, while $\mathbf{u}_h(\boldsymbol{\mu}) \in \mathbb{R}^{N_h}$ is the vector of degrees of freedom representing the high-fidelity solution.

On such a problem, the interest may be to compute, for a multiple set of parameters, its solution (as in (I1), Sect. 1.1), or any related output (as in (I2), Sect. 1.1). In other cases, we might be interested in solving an optimal control problem (as in (I3), Sect. 1.1) whose state equation is represented by (1.2).

Besides these cases, solving problem (1.2) efficiently for any given $\boldsymbol{\mu} \in \mathcal{P}$ is relevant also when facing a nonlinear and/or a time-dependent problem:

1. for instance, the Newton method for a nonlinear PDE yields a sequence of linear systems of the following form

$$\mathbb{J}_h(\mathbf{u}_h^k(\boldsymbol{\mu}); \boldsymbol{\mu}) \delta \mathbf{u}_h(\boldsymbol{\mu}) = \mathbf{f}_h(\boldsymbol{\mu}) - \mathbf{G}_h(\mathbf{u}_h^k(\boldsymbol{\mu}); \boldsymbol{\mu}), \quad k \geq 1, \quad (1.3)$$

where the vector $\mathbf{G}_h(\cdot; \boldsymbol{\mu})$ denotes a nonlinear operator and the matrix $\mathbb{J}_h(\cdot; \boldsymbol{\mu})$ its Jacobian;

2. in the same way, an implicit scheme for a linear time-dependent (parabolic) PDE leads at any time level $t^{(n)} = n\Delta t$ ($\Delta t > 0$ being the time-step) to the solution of a system which may read as

$$\left(\frac{1}{\Delta t} \mathbb{M}_h(\boldsymbol{\mu}) + \mathbb{A}_h(\boldsymbol{\mu}) \right) \mathbf{u}_h^{(n)}(\boldsymbol{\mu}) = \mathbf{f}_h^{(n)}(\boldsymbol{\mu}) + \frac{1}{\Delta t} \mathbb{M}_h(\boldsymbol{\mu}) \mathbf{u}_h^{(n-1)}(\boldsymbol{\mu}), \quad n \geq 1 \quad (1.4)$$

where $\mathbb{M}_h(\boldsymbol{\mu})$ is a *mass* matrix, for simplicity $\mathbb{A}_h(\boldsymbol{\mu})$ is not varying over the time interval, and $\mathbf{f}_h^{(n)}(\boldsymbol{\mu})$ represents time-dependent data.

For any given $\boldsymbol{\mu} \in \mathcal{P}$, problems (1.3) or (1.4) requires one to solve as many linear systems as the number of Newton iterations or time steps, respectively, thus involving an even higher computational cost in a parametrized context.

Reduced-order modeling tackle this difficulty by:

- (i) replacing (1.2) with a reduced problem featuring a dramatically lower computational complexity, yet
- (ii) retaining the essential features of the map $\boldsymbol{\mu} \mapsto \mathbf{u}_h(\boldsymbol{\mu})$, and
- (iii) guaranteeing that the error between the solution of the reduced problem and the high-fidelity one stays below a desired threshold.

Among the wide range of reduced-order modeling approaches available, in this book we shall focus on the so-called *projection-based methods*, in particular the Galerkin and Petrov-Galerkin *reduced basis* methods. As we will see in the next section, these methods rely on a suitable orthogonality criterion that is applied in step (i) to generate the reduced problem from the high-fidelity one.

The reduced basis methods considered in this book represent a notable instance of projection-based reduced-order modeling. RB methods that are not projection-based can be constructed as well; an instance is given by RB collocation methods [103].

A first intuitive algebraic representation of RB methods, as well as some general remarks concerning their accuracy and computational efficiency, will be briefly discussed in the following sections.

1.3 Reduced Basis Methods for PDEs at a Glance

The key idea of a RB method is to seek, for any new parameter value $\boldsymbol{\mu} \in \mathcal{P}$, in a subspace of lower dimension $N \ll N_h$ an approximate solution of (1.2), expressed as a linear combination of suitable, problem-dependent, basis functions. The latter are generated from a given set of solutions to the high-fidelity problem, called *snapshots*, corresponding to a suitably chosen set of parameter values.

Formally speaking, for any given $\boldsymbol{\mu} \in \mathcal{P}$, (1.2) is replaced by the following *RB problem* (also called *reduced problem*)

$$\mathbb{A}_N(\boldsymbol{\mu})\mathbf{u}_N(\boldsymbol{\mu}) = \mathbf{f}_N(\boldsymbol{\mu}), \quad (1.5)$$

where $\mathbb{A}_N(\boldsymbol{\mu}) \in \mathbb{R}^{N \times N}$, $\mathbf{f}_N(\boldsymbol{\mu}) \in \mathbb{R}^{N \times N}$ and $\mathbf{u}_N(\boldsymbol{\mu}) \in \mathbb{R}^N$ is the *reduced* vector of degrees of freedom. More specifically, we approximate $\mathbf{u}_h(\boldsymbol{\mu})$ by solutions of the form $\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu})$, where $\mathbb{V} \in \mathbb{R}^{N_h \times N}$ is a $\boldsymbol{\mu}$ -independent *transformation matrix*, whose columns collect the (degrees of freedom of the) basis functions, called *reduced basis functions*. Note that $\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu})$ belongs to \mathbb{R}^{N_h} , the same space $\mathbf{u}_h(\boldsymbol{\mu})$ belongs to.

The new low-dimensional unknown $\mathbf{u}_N(\boldsymbol{\mu})$, called *reduced basis (RB) solution*, is determined by requiring that a suitable geometric orthogonality criterion be fulfilled. To see that, let us define the residual of the high-fidelity problem

$$\mathbf{r}_h(\mathbf{v}_h; \boldsymbol{\mu}) = \mathbf{f}_h(\boldsymbol{\mu}) - \mathbb{A}_h(\boldsymbol{\mu})\mathbf{v}_h \quad \forall \mathbf{v}_h \in \mathbb{R}^{N_h}.$$

The vector

$$\mathbf{r}_h^N = \mathbf{r}_h(\mathbb{V}\mathbf{u}_N; \boldsymbol{\mu}) \quad (1.6)$$

can be regarded as the residual of the high-fidelity problem computed on the RB solution. A classical criterion to obtain the RB problem (1.5) from the high-fidelity one (1.2) is to force the residual (1.6) to be orthogonal to the subspace \mathbf{V}_N generated by the columns of \mathbb{V}

$$\mathbb{V}^T(\mathbf{f}_h(\boldsymbol{\mu}) - \mathbb{A}_h(\boldsymbol{\mu})\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu})) = \mathbf{0}, \quad (1.7)$$

that is, the *orthogonal projection* of (1.6) onto \mathbf{V}_N is zero. This is the reason why RB methods can be regarded as projection-based methods. More on projections will be discussed in Chap. 4.

Problem (1.5) is related to problem (1.2) through the following identities

$$\mathbb{A}_N(\boldsymbol{\mu}) = \mathbb{V}^T \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V}, \quad \mathbf{f}_N(\boldsymbol{\mu}) = \mathbb{V}^T \mathbf{f}_h(\boldsymbol{\mu}). \quad (1.8)$$

We will see that problem (1.5) is in fact the algebraic form of a *Galerkin method* over a subspace of dimension N of the high-fidelity (e.g. finite element) space, see Sect. 3.4.1. More general strategies (like Petrov-Galerkin methods) are based on requiring

$$\mathbb{W}^T(\mathbf{f}_h(\boldsymbol{\mu}) - \mathbb{A}_h(\boldsymbol{\mu})\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu})) = \mathbf{0}$$

instead of (1.7), yielding

$$\mathbb{A}_N(\boldsymbol{\mu}) = \mathbb{W}^T \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V}, \quad \mathbf{f}_N(\boldsymbol{\mu}) = \mathbb{W}^T \mathbf{f}_h(\boldsymbol{\mu})$$

instead of (1.8), and are obtained by imposing that the residual (1.6) is orthogonal to the subspace \mathbf{W}_N generated by the columns of a matrix $\mathbb{W} \neq \mathbb{V}$. This can still be interpreted as a projection-based method, see Sect. 3.4.2.

The $\boldsymbol{\mu}$ -independent matrix \mathbb{V} can be efficiently computed *offline*, while for every new instance of the input parameter $\boldsymbol{\mu}$, the reduced problem (1.5) will be solved *online* in a very inexpensive way.

This *offline/online* decoupling strategy is another distinguishing feature of RB methods for parametrized PDEs. The way to make this strategy efficient is a further relevant challenge to face in this context, which will be addressed in the following section.

1.4 Accuracy and Computational Efficiency of RB Methods

Although the dimension of the RB problem (1.5) is very small if compared to the one of problem (1.2), according to relations (1.8) the assembling of system (1.5) would still depend in general on the dimension N_h of the high-fidelity system *for any* $\boldsymbol{\mu} \in \mathcal{P}$. A convenient situation arises when the high-fidelity arrays in (1.2) can be written as

$$\mathbb{A}_h(\boldsymbol{\mu}) = \sum_{q=1}^{Q_a} \theta_a^q(\boldsymbol{\mu}) \mathbb{A}_h^q, \quad \mathbf{f}_h(\boldsymbol{\mu}) = \sum_{q'=1}^{Q_f} \theta_f^{q'}(\boldsymbol{\mu}) \mathbf{f}_h^{q'} \quad (1.9)$$

where $\{\theta_a^q(\boldsymbol{\mu})\}_{q=1}^{Q_a}, \{\theta_f^{q'}(\boldsymbol{\mu})\}_{q'=1}^{Q_f}$ are two sets of Q_a (respectively Q_f) scalar functions $\theta_a^q, \theta_f^{q'} : \mathcal{P} \subset \mathbb{R}^p \rightarrow \mathbb{R}$ and $\{\mathbb{A}_h^q\}_{q=1}^{Q_a}, \{\mathbf{f}_h^{q'}\}_{q'=1}^{Q_f}$ are two sets of $\boldsymbol{\mu}$ -independent matrices (respectively, vectors).

By virtue of this property – which we will refer to as *affine parametric dependence* of $\mathbb{A}_h(\boldsymbol{\mu})$ and $\mathbf{f}_h(\boldsymbol{\mu})$ – by inserting (1.9) in (1.8) we obtain

$$\begin{aligned} \mathbb{A}_N(\boldsymbol{\mu}) &= \sum_{q=1}^{Q_a} \theta_a^q(\boldsymbol{\mu}) \mathbb{V}^T \mathbb{A}_h^q \mathbb{V} = \sum_{q=1}^{Q_a} \theta_a^q(\boldsymbol{\mu}) \mathbb{A}_N^q, \\ \mathbf{f}_N(\boldsymbol{\mu}) &= \sum_{q'=1}^{Q_f} \theta_f^{q'}(\boldsymbol{\mu}) \mathbb{V}^T \mathbf{f}_h^{q'} = \sum_{q'=1}^{Q_f} \theta_f^{q'}(\boldsymbol{\mu}) \mathbf{f}_N^{q'}. \end{aligned}$$

In this way, the arrays

$$\mathbb{A}_N^q = \mathbb{V}^T \mathbb{A}_h^q \mathbb{V} \in \mathbb{R}^{N \times N}, \quad \mathbf{f}_N^{q'} = \mathbb{V}^T \mathbf{f}_h^{q'} \in \mathbb{R}^N,$$

can be computed and stored once (and for all) during a possibly expensive *offline* stage, thus enabling a very rapid (and N_h -independent) assembling of the system (1.8) during the *online* stage, for any given $\boldsymbol{\mu} \in \mathcal{P}$.

Remark 1.1. The strength of RB methods stems from the property that the set of (high-fidelity) solutions to the given parametrized PDE (i.e. the functions corresponding to the vectors $\mathbf{u}_h(\boldsymbol{\mu}), \boldsymbol{\mu} \in \mathcal{P}$, see Chap. 3) can be approximated by a linear combination of very few elements extracted from itself. The possibility to achieve

a very accurate approximation of the whole set by means of very low-dimensional subspaces arises from (i) its regularity (compactness, smoothness, analyticity) and (ii) the parametric complexity of the original problem, more precisely, the number of terms showing up in (1.9) and their regularity. These facts will be elucidated in Chap. 5. •

So far, we have highlighted those features that better characterize RB methods (and make them attractive). However, several challenges hide behind the above procedure, namely:

1. how to construct a reduced basis (that is, a transformation matrix \mathbb{V}) efficiently;
2. how to derive a priori/a posteriori bounds for the error $\|\mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{V}\mathbf{u}_N(\boldsymbol{\mu})\|$, which could also serve for point 1. above;
3. how to treat those cases where the affine parametric dependence exploited in (1.9) does not hold.

We shall address these aspects in the following chapters of the book, with the help of a wealth of numerical examples. After presenting the case of linear elliptic PDEs, we shall return to the algebraic (and geometric) interpretation of RB methods. Then we will present the most common and efficient strategies available to build a reduced basis space, namely *proper orthogonal decomposition* (POD) and *greedy* algorithms. Figure 1.1 encapsulates the workflow involved in RB approximation. The reader should be aware that a more detailed explanation will be provided after we have discussed these two strategies in Chaps. 6 and 7.

We should point out that the aim of the text is not that of providing an exhaustive survey of all existing reduced-order modeling approaches. Several other examples of computational reduction techniques for PDEs (and other relevant problems, such as dynamical systems, large-scale algebraic systems) can be found in other books, e.g., [12, 24, 212, 218]. Nevertheless, several key computational tools developed in this book within the context of RB methods represent general reduction *paradigms*, that turn out to be useful also for many other reduced-order modeling techniques.

1.5 Content of the Book

The overall goal of the textbook is to furnish a general mathematical formulation apt to embrace several (apparently different) RB approaches, investigate their algorithmic aspects and efficient solution strategies, and set up a suitable framework for the analysis of stability and convergence properties related to reduced basis methods.

We will overview results that have been extensively applied in the last decade, mainly related to the Galerkin RB method, and at the same time investigate their extension to the case of Petrov-Galerkin RB methods. We will mainly focus on the case of linear elliptic PDEs: this class of problems – relatively simple, yet relevant to many important applications – proves to be a convenient vehicle for setting up the mathematical formulation on one hand, and illustrating the results of this methodology, which can be extended to other more general problems, on the other.

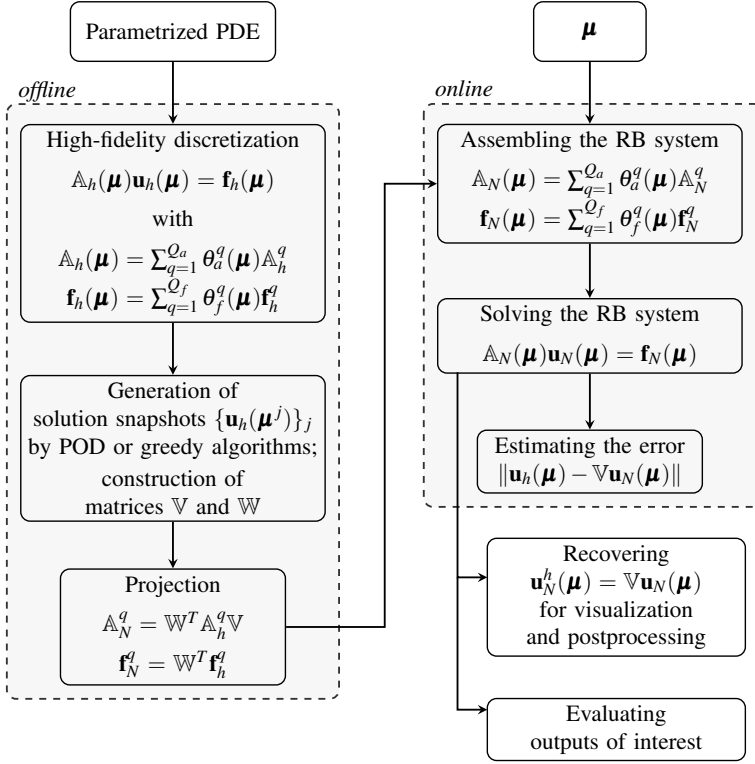


Fig. 1.1 The RB workflow at a glance

Chapter 2 surveys the formulation, analysis and approximation of three important classes of variational problems, namely: strongly coercive, weakly coercive and saddle-point problems. We discuss four examples of interest, which will be exploited throughout the book to address both mathematical and algorithmic aspects of RB methods.

The book's core matter begins to unveil in Chapter 3, where we introduce the multi-faceted aspects of RB methods for parametrized linear elliptic PDEs. We discuss projection methods and indicate the way to obtain a suitable offline/online decomposition in order to reduce the computational complexity. We show how to derive a posteriori error estimates in order to bound the error of the RB approximation with respect to the underpinning high-fidelity approximation.

The algebraic and geometric interpretations of RB methods are discussed in Chapter 4. In Chapter 5 we explore the properties of the solution set, its smoothness, dimension and approximability by lower dimensional spaces. These properties constitute the basis of the mathematical appeal of RB methods (i.e. their accuracy and their computational efficiency).

Chapters 6 and 7 are devoted to the description of two major techniques to build a reduced space: the proper orthogonal decomposition and the greedy algorithm.

In Chapter 8 we elaborate further on the role of parameters and the way they affect the RB construction. In particular, we explain how to map the original, parametrized boundary-value problem into a transformed problem set in a reference domain that is parameter independent. The transformed problem is the cornerstone the RB approximation is built upon. Numerical results obtained on the four leading examples are presented in Chapter 9 in order to show RB methods in action, to discuss their computational performance and to collect several numerical recipes that are useful also when dealing with more complex applications.

We make an excursion into the world of nonaffine problems and nonlinear problems in Chapters 10 and 11, and provide key tools necessary to extend efficiently the RB method to these problems. The Empirical Interpolation Method (EIM) and its discrete counterpart (DEIM) will be discussed in the context of nonaffine problems. Moreover, we will present a general approach for constructing reduced basis spaces and obtaining a posteriori error estimates in the case of nonlinear problems.

Finally, in Chapter 12 we employ RB methods for the efficient solution of parametric optimization problems and optimal control problems governed by elliptic PDEs. This is actually a natural playground where reduced basis methods provide a very effective computational speedup. We address the interplay between optimization and reduction by means of relevant examples.

Many important aspects that would have deserved a deeper discussion will be touched upon (and referred to) only superficially or, in some cases, completely ignored. For lack of space nothing will be said about reduction of time-dependent problems, which have received a great deal of attention in the reduced-order modeling community. For this important topic we refer to the papers [156, 122, 120, 129, 119] and references therein.

Several packages implementing RB techniques have been developed in the past decade and are freely available, such as:

- the `rbMIT` package [144] implements in MATLAB[®] the main algorithms used in the RB framework;
- an implementation of the RB framework is available as part of the open source C++ parallel finite element library `libmesh`, see [152, 153];
- `pymor` [190] is a library for building model order reduction applications with the Python programming language;
- `Dune-RB` [93], a module of the Dune library [81, 82], realizes C++ template classes for generating RB spaces dealing with many high-fidelity discretizations;
- `RBmatlab` [89, 93], a MATLAB[®] library containing reduced-order modeling approaches for linear and nonlinear, affine or arbitrary parameter-dependent evolution problems with finite element, finite volume or local discontinuous Galerkin discretizations;
- `KerMor` [259], a MATLAB[®] library providing routines for model order reduction of dynamical systems using subspace projection and kernels methods.

Many of the algorithms – as well as some of the problems – considered in this book are implemented in the MATLAB[®] package `redbKIT`, freely available at <http://redbkit.github.io/redbKIT>.

Chapter 2

Representative Problems: Analysis and (High-Fidelity) Approximation

Partial differential equations (PDEs) represent the foundation upon which many mathematical models for real-life applications are erected. In order to solve these equations one almost invariably has to resort to efficient approximation techniques (such as the finite element method, for example). These are also called high-fidelity approximations, and represent the building blocks of any kind of reduced-order model, such as the reduced basis (RB) method for parametrized PDEs presented in this book. We review the formulation, analysis and approximation of three important classes of variational problems, namely strongly coercive, weakly coercive (also called noncoercive) and saddle-point (also called mixed variational) problems. Each case is accompanied by some examples of interest.

2.1 Four Problems

In this section we introduce four examples of PDEs which are relevant in several applicative contexts, such as heat transfer, fluid dynamics, computational mechanics. Through these examples we will introduce several notations and basic notions (related to e.g. boundary conditions, functional spaces, etc.) that will be used throughout the whole book. In particular, we will consider a *parametrized version* of these examples in Chap. 8, opening the way to the reduced-order modeling world. We do not provide a detailed derivation of the models we describe, the interested reader can however refer e.g. to [236, 77].

Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$ be a regular domain, and denote by $\partial\Omega$ its boundary. Moreover, let Γ_D and Γ_N indicate a partition of the boundary, i.e. $\partial\Omega = \Gamma_D \cup \Gamma_N$, $\Gamma_D \cap \Gamma_N = \emptyset$, and Γ_D (resp. Γ_N) represents the portion of the boundary where Dirichlet (resp. Neumann) boundary conditions are enforced. For the sake of simplicity, we do not consider Robin boundary conditions, although what is presented here can be easily extended to that case as well.

2.1.1 Advection-Diffusion-Reaction Equation

We first consider the following advection-diffusion-reaction problem

$$\begin{cases} -\operatorname{div}(\mathbb{K}\nabla u) + \mathbf{b} \cdot \nabla u + a_0 u = s & \text{in } \Omega \\ u = g & \text{on } \Gamma_D \\ \mathbb{K}\nabla u \cdot \mathbf{n} = h & \text{on } \Gamma_N \end{cases} \quad (2.1)$$

where $\mathbb{K} : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ is the so-called diffusion (or conductivity) matrix, $\mathbf{b} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a given velocity (convective) field, $a_0(\mathbf{x})$, $s(\mathbf{x})$, $g(\mathbf{x})$, and $h(\mathbf{x})$ are given scalar-valued functions; here \mathbf{n} denotes the outward unit normal on $\partial\Omega$.

Such a model can describe heat conduction and advection in a continuous medium occupying Ω , as well as transport and diffusion of a substance in a region Ω ; u may thus represent the temperature of the medium or the concentration of the substance, respectively. In the former case, s represents the rate of heat per unit mass supplied by an external source. The term $-\operatorname{div}(\mathbb{K}\nabla u)$ is associated with thermal or molecular diffusion, while $\mathbf{b} \cdot \nabla u$ models the transport (or advection) process. The term $a_0 u$ models (linearized) reaction, that is, if u is the concentration of a substance, a_0 represents its rate of decomposition (if $a_0 > 0$) or growth (if $a_0 < 0$).

2.1.2 Linear Elasticity Equations

The linear elastic deformations of an isotropic solid occupying the domain $\Omega \subset \mathbb{R}^d$ is described in terms of the stress tensor $\boldsymbol{\sigma} : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$, the strain tensor $\boldsymbol{\varepsilon} : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$, the body force $\mathbf{f} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ and the displacement field $\mathbf{u} : \mathbb{R}^d \rightarrow \mathbb{R}^d$, this latter being the unknown of the problem. The governing equations consist of an equation stating the equilibrium of forces

$$-\operatorname{div}(\boldsymbol{\sigma}) = \mathbf{f} \quad \text{in } \Omega,$$

the strain-displacement relation

$$\boldsymbol{\varepsilon}(\mathbf{u}) = \frac{1}{2}(\nabla \mathbf{u} + \nabla \mathbf{u}^T)$$

and the constitutive law, which in the linear isotropic case takes the form

$$\boldsymbol{\sigma} = 2\mu\boldsymbol{\varepsilon}(\mathbf{u}) + \lambda(\operatorname{div}(\mathbf{u}))\mathbf{I}.$$

Here μ and λ are the Lamé coefficients, which can be expressed in terms of the Young modulus E and the Poisson coefficient ν as

$$\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}, \quad \mu = \frac{E}{2(1+\nu)}.$$

The equilibrium problem for a linear elastic material can therefore be written as follows:

$$\begin{cases} -\operatorname{div}(\mu(\nabla \mathbf{u} + \nabla \mathbf{u}^T) + \lambda(\operatorname{div} \mathbf{u})\mathbf{I}) = \mathbf{f} & \text{in } \Omega \\ \mathbf{u} = \mathbf{g} & \text{on } \Gamma_D \\ \boldsymbol{\sigma} \mathbf{n} = \mathbf{h} & \text{on } \Gamma_N. \end{cases} \quad (2.2)$$

Note that on the Dirichlet boundary Γ_D we impose a prescribed displacement, whereas on the Neumann boundary Γ_N we impose that the normal stress $\boldsymbol{\sigma} \mathbf{n}$ equals a prescribed traction vector \mathbf{h} .

2.1.3 Stokes Equations

The Stokes equations describe the flow of a Newtonian, incompressible viscous fluid confined in a domain $\Omega \subset \mathbb{R}^d$ when convective forces are negligible with respect to viscous forces:

$$\begin{cases} -\nu \Delta \mathbf{u} + \nabla p = \mathbf{0} & \text{in } \Omega \\ \operatorname{div} \mathbf{u} = 0 & \text{in } \Omega \\ \mathbf{u} = \mathbf{g} & \text{on } \Gamma_D \\ -p \mathbf{n} + \nu(\nabla \mathbf{u}) \mathbf{n} = \mathbf{h} & \text{on } \Gamma_N. \end{cases} \quad (2.3)$$

Here \mathbf{u} and p denote respectively the velocity and pressure fields, while the constant ν denotes the kinematic viscosity. The first equation represents the conservation of the linear momentum of the fluid, while the second equation, called incompressibility condition, enforces the mass conservation. We impose a prescribed velocity \mathbf{g} on the boundary Γ_D , while a nonhomogeneous Neumann condition is imposed on Γ_N .

2.1.4 Navier-Stokes Equations

Upon inserting a convective term in the momentum equation of the Stokes system above we obtain to the steady Navier-Stokes equations:

$$\begin{cases} -\nu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla) \mathbf{u} + \nabla p = \mathbf{0} & \text{in } \Omega \\ \operatorname{div} \mathbf{u} = 0 & \text{in } \Omega \\ \mathbf{u} = \mathbf{g} & \text{on } \Gamma_D \\ -p \mathbf{n} + \nu(\nabla \mathbf{u}) \mathbf{n} = \mathbf{h} & \text{on } \Gamma_N. \end{cases} \quad (2.4)$$

With respect to the models considered so far, equations (2.4) feature a (quadratic) nonlinear term, which prevents casting them in one of the abstract frameworks that will be introduced in Sect. 2.2. Yet, the Navier-Stokes system will be extensively treated in this book since it is an ubiquitous problem in fluid dynamics and serves as a basis for modeling the flow of any Newtonian incompressible fluid.

2.2 Formulation and Analysis of Variational Problems

In this section we introduce some basic results of functional analysis, essential for a correct variational formulation of a broad variety of boundary value problems, encompassing, as special cases, the four problems addressed in Sect. 2.1.

In particular, we briefly review strongly coercive, weakly coercive, and saddle-point problems. For each of them, after introducing their functional setting, we present the main results ensuring their well-posedness: Lax-Milgram lemma, Nečas theorem and Brezzi theorem, respectively. For the proofs of these classical results as well as for a more insightful introduction to PDEs and variational methods, the reader is referred, e.g., to [222, 65, 126, 236, 216, 32]. For the unfamiliar reader, basic notions of functional analysis are instead recalled in Appendix A.

2.2.1 Strongly Coercive Problems

Let us denote by V a Hilbert space on \mathbb{R} and by V' its dual space, i.e. the space of linear and continuous functionals over V . We denote by $(\cdot, \cdot)_V$ the inner product defined over V , and by $\|\cdot\|_V$ the norm induced by $(\cdot, \cdot)_V$, $\|v\|_V = (v, v)_V^{1/2}$ for any $v \in V$. Moreover,

$$\langle G, v \rangle = {}_{V'}\langle G, v \rangle_V \quad \forall G \in V', v \in V,$$

denotes the duality pairing between V' and V .

By introducing a bilinear form $a: V \times V \rightarrow \mathbb{R}$ and a linear functional $f \in V'$, we consider the following abstract variational problem:

find $u \in V$ such that

$$a(u, v) = f(v) \quad \forall v \in V. \quad (P_1)$$

The bilinear form $a(\cdot, \cdot)$ is said to be continuous on $V \times V$ if there exists a constant $\gamma > 0$ such that

$$|a(u, v)| \leq \gamma \|u\|_V \|v\|_V \quad \forall u, v \in V.$$

Moreover, it is said to be strongly coercive (or coercive, or V -elliptic) if there exists a constant $\alpha > 0$ such that

$$a(v, v) \geq \alpha \|v\|_V^2 \quad \forall v \in V. \quad (2.5)$$

γ and α are named the continuity constant and the coercivity constant of $a(\cdot, \cdot)$, respectively. The following fundamental result states that strongly coercive problems are well-posed, that is, they admit a unique solution which depends continuously on the data (see for instance [106, 236] for its proof).

Lemma 2.1 (Lax-Milgram). *Let V be a Hilbert space, $a : V \times V \rightarrow \mathbb{R}$ a continuous, strongly coercive bilinear form on $V \times V$, and $f : V \rightarrow \mathbb{R}$ a bounded linear functional on V . Then, the abstract variational problem (P_1) has a unique solution and it satisfies the stability estimate*

$$\|u\|_V \leq \frac{1}{\alpha} \|f\|_{V'}. \quad (2.6)$$

By considering the differential operator $L \in \mathcal{L}(V, V')$, $v \in V \mapsto Lv \in V'$ defined by

$${}_{V'}\langle Lv, w \rangle_V = a(v, w) \quad \forall v, w \in V,$$

problem (P_1) can be equivalently written as: find $u \in V$ such that

$$Lu = f \quad \text{in } V'. \quad (2.7)$$

The conditions required by Lax-Milgram lemma ensure that L is a bijective mapping, so that (2.7) admits a unique solution.

When the bilinear form $a(\cdot, \cdot)$ is symmetric – that is, $a(u, v) = a(v, u)$ for all $u, v \in V$ – it defines an inner product over V

$$(u, v)_a = a(u, v)$$

and the corresponding norm $\|\cdot\|_a = \sqrt{a(\cdot, \cdot)}$ induced by this scalar product is called *energy norm*. Under these conditions, problem (P_1) is equivalent to a minimization problem for the quadratic functional

$$J(v) = \frac{1}{2}a(v, v) - f(v), \quad (2.8)$$

according to the following

Proposition 2.1. *Let V be a Hilbert space, $a : V \times V \rightarrow \mathbb{R}$ a continuous, symmetric and positive bilinear form on $V \times V$, and $f : V \rightarrow \mathbb{R}$ a bounded linear functional on V . Then, u is the (unique) solution of (P_1) if and only if*

$$J(u) = \min_{v \in V} J(v), \quad (2.9)$$

that is, u minimizes over V the functional $J : V \rightarrow \mathbb{R}$ defined by (2.8).

See Exercise 1 for the proof. (P_1) actually represents the first-order necessary optimality condition for the minimization problem (2.9). Several problems in the field of calculus of variations are formulated in this way; for this reason problem (P_1) is referred to as *variational problem*.

2.2.2 Weakly Coercive (or Inf-Sup Stable) Problems

Let us now consider a more general variational problem. Given two Hilbert spaces V and W along with their dual V' and W' , respectively, the bilinear form $a: V \times W \rightarrow \mathbb{R}$ and the linear functional $f \in W'$, we consider the following abstract variational problem:

$$\begin{aligned} \text{find } u \in V \text{ such that} \\ a(u, w) = f(w) \quad \forall w \in W. \end{aligned} \quad (P_2)$$

The bilinear form $a(\cdot, \cdot)$ is said to be continuous on $V \times W$ if there exists $\gamma > 0$ such that

$$|a(v, w)| \leq \gamma \|v\|_V \|w\|_W \quad \forall v \in V, w \in W, \quad (2.10)$$

and is said to be weakly coercive (or inf-sup stable) if there exists a constant $\beta > 0$ such that:

$$\inf_{v \in V} \sup_{w \in W} \frac{a(v, w)}{\|v\|_V \|w\|_W} \geq \beta, \quad (2.11)$$

and

$$\inf_{w \in W} \sup_{v \in V} \frac{a(v, w)}{\|v\|_V \|w\|_W} > 0. \quad (2.12)$$

As noted in [84], the infimum and supremum above are actually attained, and can be replaced with minimum and maximum.

The Nečas theorem [194] shows that, under these assumptions, weakly coercive problems are well posed.

Theorem 2.1 (Nečas). *Let V and W be two Hilbert spaces, $a: V \times W \rightarrow \mathbb{R}$ a continuous, weakly coercive bilinear form on $V \times W$, and $f: W \rightarrow \mathbb{R}$ a bounded linear functional on W . Then, the variational problem (P_2) has a unique solution which satisfies the stability estimate*

$$\|u\|_V \leq \frac{1}{\beta} \|f\|_{W'}. \quad (2.13)$$

Remark 2.1. Since strong coercivity implies weak coercivity, Lax-Milgram lemma is a special case of Nečas theorem. •

Condition (2.11) – also referred to as *inf-sup condition* – can be reformulated as

$$\exists \beta > 0 : \sup_{w \in W} \frac{a(v, w)}{\|w\|_W} \geq \beta \|v\|_V \quad \forall v \in V,$$

whereas condition (2.12) can be equivalently restated as

$$\text{if } w \in W \text{ is such that } a(v, w) = 0 \quad \forall v \in V, \quad \text{then } w = 0.$$

(2.11) and (2.12) are also necessary conditions for problem (P_2) to be well-posed. Moreover, by considering the differential operator $L \in \mathcal{L}(V, W')$ defined by

$$v' \langle Lv, w \rangle_V = a(v, w) \quad \forall v \in V, w \in W,$$

condition (2.11) is equivalent to

$$\text{Ker}(L) = \{0\}, \quad \text{Range}(L) \text{ is closed,}$$

whereas condition (2.12) is equivalent to

$$\text{Ker}(L^*) = \text{Range}(L)^\perp = \{0\},$$

where $L^* \in \mathcal{L}(W, V')$ denotes the adjoint of the operator L (see Appendix A).

Showing the weakly coercivity property (2.11) and condition (2.12) is in general quite involved. Very often we deal with noncoercive bilinear forms defined on $V \times V$ (i.e. $W = V$) – this is the case, for instance, of several advection-diffusion problems – for which (2.12) can be easily fulfilled, whereas (2.11) is harder to show. The following result (see e.g. [268] for the proof) can be helpful in these cases.

Proposition 2.2. *Let us suppose that $H_0^1(\Omega) \subset V \subset H^1(\Omega)$ and the bilinear form $a : V \times V \rightarrow \mathbb{R}$ fulfills the following conditions:*

1. *Gårding inequality: for some constants $\alpha > 0$ and $\lambda > 0$*

$$a(v, v) \geq \alpha \|v\|_{H^1(\Omega)}^2 - \lambda \|v\|_{L^2(\Omega)}^2 \quad \forall v \in V; \quad (2.14)$$

2. *if $u \in V$ is such that $a(u, v) = 0$ for any $v \in V$, then $u = 0$.*

Then there exists a constant $\beta > 0$ such that

$$\sup_{w \in V} \frac{a(v, w)}{\|w\|_V} \geq \beta \|v\|_V \quad \forall v \in V.$$

Thanks to this result, a bilinear form fulfilling Gårding inequality (2.14) is often referred to as *weakly coercive*.

2.2.3 Saddle-Point Problems

We now turn to a very relevant class of problems, called *mixed variational problems* or *saddle point problems*. Several problems can be formulated and analyzed within this framework – high-order PDEs, Stokes equations, optimal control problems, just to mention a few. For a more in-depth reading we refer e.g. to [32].

Given two Hilbert spaces X and Q along with their dual X' and Q' , respectively, the bilinear forms $d: X \times X \rightarrow \mathbb{R}$, $b: X \times Q \rightarrow \mathbb{R}$ the linear functionals $f_1 \in X'$ and $f_2 \in Q'$, we consider the following mixed variational problem:

find $(x, p) \in X \times Q$ such that

$$\begin{cases} d(x, w) + b(w, p) = f_1(w) & \forall w \in X \\ b(x, q) = f_2(q) & \forall q \in Q. \end{cases} \quad (P_3)$$

Let us define the following subspace of X

$$X_0 = \{w \in X : b(w, q) = 0 \quad \forall q \in Q\} \subset X.$$

The following theorem due to Brezzi [37, 32] establishes sufficient conditions for the saddle-point problem (P_3) to be well posed.

Theorem 2.2 (Brezzi). *Under the following assumptions:*

1. *continuity of the bilinear form $d(\cdot, \cdot)$: there exists a constant $\gamma_d > 0$ such that*

$$|d(x, w)| \leq \gamma_d \|x\|_X \|w\|_X \quad \forall x, w \in X; \quad (2.15)$$

2. *weakly coercivity of the bilinear form $d(\cdot, \cdot)$ on X_0 : there exists a constant $\alpha_0 > 0$ such that*

$$\inf_{x \in X_0} \sup_{w \in X_0} \frac{d(x, w)}{\|x\|_X \|w\|_X} \geq \alpha_0, \quad \inf_{w \in X_0} \sup_{x \in X_0} \frac{d(x, w)}{\|x\|_X \|w\|_X} > 0; \quad (2.16)$$

3. *continuity of the bilinear form $b(\cdot, \cdot)$: there exists a constant $\gamma_b > 0$ such that*

$$|b(w, q)| \leq \gamma_b \|w\|_X \|q\|_Q \quad \forall w \in X, q \in Q; \quad (2.17)$$

4. *the bilinear form $b(\cdot, \cdot)$ satisfies the inf-sup condition*

$$\beta^s = \inf_{q \in Q} \sup_{w \in X} \frac{b(w, q)}{\|w\|_X \|q\|_Q} \geq \beta_0^s > 0, \quad (2.18)$$

there exists a unique solution $(x, p) \in X \times Q$ to the mixed variational problem (P_3) . Moreover, the following stability estimates hold:

$$\|x\|_X \leq \frac{1}{\alpha} \left[\|f_1\|_{X'} + \frac{\alpha_0 + \gamma_d}{\beta_0^s} \|f_2\|_{Q'} \right], \quad (2.19a)$$

$$\|p\|_Q \leq \frac{1}{\beta^s} \left[\left(1 + \frac{\gamma_d}{\alpha_0} \right) \|f_1\|_{X'} + \frac{\gamma_d(\alpha_0 + \gamma_d)}{\alpha_0 + \beta_0^s} \|f_2\|_{Q'} \right]. \quad (2.19b)$$

Remark 2.2. Note that mixed variational problems are a special case of weakly coercive problems. In fact, by setting $V = W = X \times Q$, defining the bilinear form $a : V \times V \rightarrow \mathbb{R}$

$$a((x, p), (w, q)) = d(x, w) + b(w, p) + b(x, q), \quad (2.20)$$

and the functional $f((w, q)) = f_1(w) + f_2(q)$, we can rewrite (P_3) in the form of (P_2) : find $(x, p) \in X$ such that

$$a((x, p), (w, q)) = f((w, q)) \quad \forall (w, q) \in X,$$

which could then be analyzed using Nečas theorem. We remark that if conditions (2.15), (2.17) are fulfilled, then (2.10) is verified, too. Similarly, if conditions (2.16), (2.18) are verified, $a : V \times V \rightarrow \mathbb{R}$ is weakly coercive in the sense of (2.11)–(2.12). For further analysis on the relations between the two theoretical approaches, see e.g. [84, 263]. •

We can also notice that if $d : X \times X \rightarrow \mathbb{R}$ is coercive over X , then conditions (2.16) are automatically fulfilled. Let us recall that the variational problem (P_1) , if $a(\cdot, \cdot)$ is a symmetric and positive bilinear form, is equivalent to a minimization problem for the quadratic functional J defined in (2.8). In the case of a mixed variational problem, if $d(\cdot, \cdot)$ is a symmetric and positive bilinear form, we obtain instead a *constrained minimization problem* for the same quadratic functional, which features a *saddle-point* form.

Given a linear mapping $\mathcal{L} : X \times Q \rightarrow \mathbb{R}$, (x, p) is a *saddle point* of \mathcal{L} if

$$\mathcal{L}(x, q) \leq \mathcal{L}(x, p) \leq \mathcal{L}(y, p) \quad \forall (y, q) \in X \times Q.$$

It is possible to show that (x, p) is a saddle point of L if and only if

$$\inf_{y \in X} \sup_{q \in Q} \mathcal{L}(y, q) = \sup_{q \in Q} \mathcal{L}(x, q) = \mathcal{L}(x, p) = \inf_{y \in X} \mathcal{L}(y, p) = \sup_{q \in Q} \inf_{y \in X} \mathcal{L}(y, q).$$

This characterization allows to interpret (P_3) as a constrained minimization problem.

Proposition 2.3. *Given the Hilbert spaces X and Q , the functionals $f \in X'$ and $g \in Q'$, and the bilinear forms $d(\cdot, \cdot)$ and $b(\cdot, \cdot)$ on $X \times X$ and $X \times Q$, respectively, assume that $a : X \times X \rightarrow \mathbb{R}$ is a symmetric and positive bilinear form on $X \times X$. Then $(x, p) \in X \times Q$ is a solution to problem (P_3) if and only if (x, p) is a saddle point of*

$$\mathcal{L}(w, q) = \frac{1}{2}d(w, w) + b(w, q) - f_1(w) - f_2(q), \quad (2.21)$$

that is, if and only if x is a minimizer of

$$J(w) = \frac{1}{2}d(w, w) - f_1(w)$$

over X under the constraint $b(x, q) = f_2(q) \quad \forall q \in Q$.

Equivalently,

$$x = \arg \min_{w \in X} J(w) \quad \text{s.t.} \quad b(x, q) = f_2(q) \quad \forall q \in Q. \quad (2.22)$$

If, additionally, the assumptions of Theorem 2.2 are fulfilled, problem (P_3) has a unique solution, which is the unique saddle point of the functional \mathcal{L} ; this solution satisfies (P_3) , which can be seen as a first-order necessary (and sufficient) optimality condition for the constrained minimization problem (2.22).

2.3 Analysis of Three (out of Four) Problems

Let us now analyze the well-posedness of the first three problems introduced in Sect. 2.1 by means of the results of the previous section.

2.3.1 Advection-Diffusion-Reaction Equation

The weak formulation of problem (2.1) can be cast in the abstract form (P_1) upon choosing

$$V = H_{\Gamma_D}^1(\Omega) := \{v \in H^1(\Omega) : v|_{\Gamma_D} = 0\}$$

and introducing the following forms (see Exercise 2):

$$a(u, v) = \int_{\Omega} (\mathbb{K} \nabla u \cdot \nabla v + \mathbf{b} \cdot \nabla u v + a_0 u v) d\Omega, \quad (2.23)$$

$$f(v) = \int_{\Omega} s v d\Omega + \int_{\Gamma_N} h v d\Gamma - a(r_g, v). \quad (2.24)$$

Here $r_g \in H^1(\Omega)$ is a lifting function such that $r_g|_{\Gamma_D} = g$; the solution to problem (2.1) is then obtained as $u + r_g$. We assume that there exist $k_m, k_M > 0$ such that the following *uniform ellipticity* condition holds¹

$$k_m |\boldsymbol{\xi}|^2 \leq \mathbb{K}(\mathbf{x}) \boldsymbol{\xi} \cdot \boldsymbol{\xi} \leq k_M |\boldsymbol{\xi}|^2 \quad \forall \boldsymbol{\xi} \in \mathbb{R}^d, \quad \mathbf{x} - \text{a.e. in } \Omega;$$

in this way, \mathbb{K} is a positive definite matrix for every \mathbf{x} in Ω with the minimum (and maximum) eigenvalue bounded from below by k_m (from above by k_M). For this reason, k_m is also called ellipticity constant. Moreover, we assume that $s \in L^2(\Omega)$, $g \in H^{1/2}(\Gamma_D)$, $h \in L^2(\Gamma_N)$.

In order to discuss the well-posedness of problem (2.1) we equip the space V with the norm $\|\cdot\|_V = |\cdot|_{H^1(\Omega)}$, which is equivalent to the $H^1(\Omega)$ norm thanks to

¹ $\mathbf{x} - \text{a.e. in } \Omega$ reads for almost every $\mathbf{x} \in \Omega$, that is, for all $\mathbf{x} \in \Omega$, except for a zero-measure set at most. See Appendix A.6.1.

Poincaré inequality (Proposition A.2 of Appendix A). We consider three relevant cases:

- $\mathbf{b} = \mathbf{0}$ and $a_0 = 0$. Under these conditions, the bilinear form (2.23) is strongly coercive since

$$a(v, v) = \int_{\Omega} \mathbb{K} \nabla v \cdot \nabla v d\Omega \geq k_m \|\nabla v\|_{L^2(\Omega)}^2.$$

Moreover, if $s \in L^2(\Omega)$, $h \in L^2(\Gamma_N)$ and $g \in H^{1/2}(\Omega)$, then $f \in V'$. Therefore, thanks to Lax-Milgram lemma, problem (2.1) is well-posed;

- $\mathbb{K} = \mathbf{I}$ and $\mathbf{b} = \mathbf{0}$. If $a_0 \geq 0$ the problem is still strongly coercive. Before discussing the case $a_0 < 0$, let us denote with (λ, w) the generic pair of eigenvalue-eigenfunction of the Laplacian operator with homogenous boundary conditions, that is the solutions of the following eigenproblem: find $(\lambda, w) \in \mathbb{R} \times V$ such that

$$\int_{\Omega} \nabla w \cdot \nabla v d\Omega = \lambda \int_{\Omega} w v d\Omega \quad \forall v \in V. \quad (2.25)$$

It can be shown that problem (2.25) has infinitely many positive eigenvalues of finite multiplicity (see, e.g., [268]). Moreover the smallest eigenvalue λ_1 can be characterized as

$$\lambda_1 = \inf_{v \in V} \frac{\int_{\Omega} |\nabla v|^2 d\Omega}{\int_{\Omega} v^2 d\Omega}.$$

If $a_0 > -\lambda_1$, then the bilinear form (2.23) is strongly coercive, as thanks to Poincaré inequality

$$a(v, v) = \|\nabla v\|_{L^2(\Omega)}^2 - |a_0| \|v\|_{L^2(\Omega)}^2 \geq \left(1 - \frac{|a_0|}{\lambda_1}\right) \|\nabla v\|_{L^2(\Omega)}^2.$$

Instead, if $a_0 \leq -\lambda_1$ the bilinear form (2.23) is at best only weakly coercive: in fact, we can show that Gårding inequality (2.14) holds with $\lambda = a_0$. However, the condition 2 in Proposition 2.2 holds only if a_0 is not an eigenvalue of the Laplace operator, i.e. $a_0 \neq \lambda_i$. Therefore, if $a_0 \neq \lambda_i$ problem (2.1) admits a unique solution thanks to Nečas theorem;

- $a_0 = 0$. If $\operatorname{div}(\mathbf{b}) \leq 0$ and $\mathbf{b} \cdot \mathbf{n} \geq 0$ almost everywhere on Γ_N , the problem is strongly coercive. Indeed,

$$a(v, v) = \int_{\Omega} \mathbb{K} \nabla u \cdot \nabla v - \frac{1}{2} \int_{\Omega} \operatorname{div}(\mathbf{b}) v^2 + \frac{1}{2} \int_{\Gamma_N} \mathbf{b} \cdot \mathbf{n} v^2 \geq k_m \|\nabla v\|_{L^2(\Omega)}^2.$$

Instead, if $\operatorname{div}(\mathbf{b}) > 0$, or $\mathbf{b} \cdot \mathbf{n} < 0$ over a portion of Γ_N , the problem is at best weakly coercive.

2.3.2 Linear Elasticity Equations

The weak formulation of problem (2.2) reads (see Exercise 3): find $\mathbf{u} \in V = [H_{\Gamma_D}^1(\Omega)]^d$ such that

$$a(\mathbf{u}, \mathbf{v}) = f(\mathbf{v}) \quad \forall \mathbf{v} \in V, \quad (2.26)$$

where we have defined the bilinear form $a: V \times V \rightarrow \mathbb{R}$ as

$$a(\mathbf{u}, \mathbf{v}) = \int_{\Omega} 2\mu \boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{v}) d\mathbf{x} + \int_{\Omega} \lambda \operatorname{div}(\mathbf{u}) \operatorname{div}(\mathbf{v}) d\Omega, \quad (2.27)$$

and the linear form $f: V \rightarrow \mathbb{R}$ as

$$f(\mathbf{v}) = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} d\Omega + \int_{\Gamma_N} \mathbf{h} \cdot \mathbf{v} d\Gamma - a(\mathbf{r}_g, \mathbf{v}). \quad (2.28)$$

Here $\mathbf{r}_g \in [H^1(\Omega)]^d$ is a lifting vector function such that $\mathbf{r}_g|_{\Gamma_D} = \mathbf{g}$; the solution to problem (2.2) is then obtained as $\mathbf{u} + \mathbf{r}_g$. The bilinear form (2.27) is symmetric and strongly coercive owing to Korn's inequality (see e.g. [65]). Therefore problem (2.2) can be cast in the form (P_1) and admits a unique solution $\mathbf{u} \in V$ thanks to Lax-Milgram lemma.

2.3.3 Stokes Equations

We now consider the Stokes equations (2.3). We define the velocity space $X = [H_{\Gamma_D}^1(\Omega)]^d$ and the pressure space $Q = L^2(\Omega)$. The weak formulation of (2.3) can be cast in the mixed form (P_3) upon defining $(x, p) = (\mathbf{u}, p)$, the bilinear forms $d: X \times X \rightarrow \mathbb{R}$ and $b: X \times Q \rightarrow \mathbb{R}$ as

$$d(\mathbf{v}, \mathbf{w}) = \int_{\Omega} \mathbf{v} \nabla \mathbf{v} : \nabla \mathbf{w} d\Omega, \quad b(\mathbf{v}, q) = - \int_{\Omega} q \operatorname{div} \mathbf{v} d\Omega \quad \forall \mathbf{v}, \mathbf{w} \in X, q \in Q;$$

see Exercise 4. Moreover we define the linear functionals $f_1 \in X'$ and $f_2 \in Q'$ as

$$f_1(\mathbf{w}) = \int_{\Gamma_N} \mathbf{h} \cdot \mathbf{w} d\Gamma - d(\mathbf{r}_g, \mathbf{w}), \quad f_2(q) = -b(\mathbf{r}_g, q),$$

where $\mathbf{r}_g \in [H^1(\Omega)]^d$ is a lifting vector function such that $\mathbf{r}_g|_{\Gamma_D} = \mathbf{g}$; the solution to problem (2.3) is then obtained as $\mathbf{u} + \mathbf{r}_g$. The well-posedness of the weak problem can be proved by means of Theorem 2.2; while the continuity properties of the bilinear forms and the coercivity of $d(\cdot, \cdot)$ can be easily verified, more attention has to be paid to the fulfillment of the inf-sup condition (2.18), see Exercises 5, 6. Further details can be found, e.g., in [222, 104].

Remark 2.3. In a similar way, we can also derive the following weak formulation for the Navier-Stokes problem (2.4): find $(\mathbf{u}, p) \in X \times Q$ such that

$$\begin{cases} \bar{d}(\mathbf{u}, \mathbf{v}) + c(\mathbf{u}, \mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) = f_1(\mathbf{v}) & \forall \mathbf{v} \in X \\ b(\mathbf{u}, q) = f_2(q) & \forall q \in Q, \end{cases} \quad (2.29)$$

where the trilinear form $c(\cdot, \cdot, \cdot) : X \times X \times X \rightarrow \mathbb{R}$ is defined as

$$c(\mathbf{u}, \mathbf{w}, \mathbf{v}) = \int_{\Omega} (\mathbf{u} \cdot \nabla) \mathbf{w} \cdot \mathbf{v} d\Omega,$$

the bilinear form $\bar{d} : X \times X \rightarrow \mathbb{R}$ is given by

$$\bar{d}(\mathbf{u}, \mathbf{v}) = d(\mathbf{u}, \mathbf{v}) + c(\mathbf{u}, \mathbf{r}_g, \mathbf{v}) + c(\mathbf{r}_g, \mathbf{u}, \mathbf{v}),$$

and the functional $f_1 \in X'$ is now given by

$$f_1(\mathbf{w}) = \int_{\Gamma_N} \mathbf{h} \cdot \mathbf{w} d\Gamma - d(\mathbf{r}_g, \mathbf{w}) - c(\mathbf{r}_g, \mathbf{r}_g, \mathbf{w}).$$

For the analysis of this problem see Chap. 11. See also [222, 104]. •

2.4 On the Numerical Approximation of Variational Problems

In this section we recall the most important results related to the numerical approximation of the variational problems introduced in Sect. 2.2. Since these approximations are obtained by projecting the original problem upon a finite-dimensional subspace, their well-posedness is subject to assumptions which are similar to the ones required for the analysis of their infinite-dimensional counterpart. Conditions to be fulfilled by discrete problems might be either automatically inherited from the ones on the original problems or not, depending on the different cases.

2.4.1 Strongly Coercive Problems

Let for every $h > 0$ $V_h \subset V$ be a finite-dimensional subspace of V , such that $\dim V_h = N_h$; the subscript h is related to a characteristic discretization parameter (most typically, the grid size). The approximate problem takes the form:

find $u_h \in V_h$ such that

$$a(u_h, v_h) = f(v_h) \quad \forall v_h \in V_h \quad (P_1^h)$$

and is called *Galerkin problem*; the solution u_h of this problem is often known as the *Galerkin approximation* of u . Since $V_h \subset V$, the exact solution u satisfies the weak problem (P_1) for each element $v = v_h \in V_h$, hence we have

$$a(u, v_h) = f(v_h) \quad \forall v_h \in V_h. \quad (2.30)$$

For this property, the Galerkin problem is said *strongly consistent* (see [216]). From (2.30) and the problem statement (P_1^h) , we obtain that u_h satisfies

$$a(u - u_h, v_h) = 0 \quad \forall v_h \in V_h, \quad (2.31)$$

a property which is known as *Galerkin orthogonality*. Indeed, should $a(\cdot, \cdot)$ be symmetric, (2.31) can be interpreted as the orthogonality condition – with respect to the scalar product $a(\cdot, \cdot)$ – between the approximation error, $u - u_h$, and the subspace V_h . Thus, u_h can be seen as the *orthogonal projection* of u onto V_h : among all elements of V_h , u_h is the one minimizing the distance to u in the energy norm,

$$\|u - u_h\|_a \leq \|u - v_h\|_a \quad \forall v_h \in V_h.$$

Remark 2.4. The variational problem (P_1^h) can be equivalently written as: find $u_h \in V_h$ such that

$$Lu_h = f \quad \text{in } V_h'. \quad (2.32)$$

•

With these premises, the following result (of existence, uniqueness, stability and convergence) easily follows (see Exercise 7; further details can be found, e.g., in [65, 222, 126]).

Lemma 2.2. *Let the space V , the bilinear form $a(\cdot, \cdot)$ and the linear functional $f(\cdot)$ satisfy the hypotheses of Lemma 2.1. Let $V_h \subset V$ be a closed subspace. Then $a(\cdot, \cdot)$ is continuous on $V_h \times V_h$ with constant $\gamma_h \leq \gamma$ and coercive on $V_h \times V_h$ with constant $\alpha_h \geq \alpha$. Then, for every $h > 0$, the discretized problem (P_1^h) has a unique solution $u_h \in V_h$, that satisfies the stability estimate*

$$\|u_h\|_V \leq \frac{1}{\alpha_h} \|f\|_{V'}. \quad (2.33)$$

Furthermore, if $u \in V$ denotes the unique solution of (P_1) , the following optimal error inequality is satisfied

$$\|u - u_h\|_V \leq \frac{\gamma}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_V. \quad (2.34)$$

We point out that *optimality* here means that the error in a finite-dimensional approximation $\|u - u_h\|_V$ is bounded from above by the error of the best approximation out of the same finite-dimensional subspace, multiplied by a constant independent of h . However, this is not sufficient to ensure the convergence $\|u - u_h\|_V \rightarrow 0$ when $h \rightarrow 0$: in fact, an additional property related to the approximability of the discrete

spaces is required. Thanks to (2.34) (also known as Céa's Lemma), in order for the method to converge it will be sufficient to require that, for $h \rightarrow 0$, the space V_h tends to "fill" the entire space V . Precisely,

$$\lim_{h \rightarrow 0} \inf_{v_h \in V_h} \|v - v_h\|_V = 0 \quad \forall v \in V. \quad (2.35)$$

In that case, the Galerkin method is convergent and $\|u - u_h\|_V \rightarrow 0$ when $h \rightarrow 0$; V_h must therefore be chosen in order to guarantee the density property (2.35). In this way, by taking a sufficiently small h , it is possible to approximate u by u_h as accurately as desired. The actual convergence rate will depend on the specific choice of the subspace V_h . In the finite element case in which the latter is made of piecewise polynomials of degree $r \geq 1$, $\|u - u_h\|_V$ will tend to zero as $O(h^r)$, see Sect. 2.5.

Remark 2.5. Since $\inf_{v_h \in V_h} \|u - v_h\|_V \leq \|u - u_h\|_V$, if M/α has order 1 the error due to the Galerkin method can be identified with the best approximation error for u in V_h . In any case, both errors have the same infinitesimal order with respect to h . In the case where $a(\cdot, \cdot)$ is a symmetric bilinear form, then (2.34) can be improved as follows (see Exercise 8)

$$\|u - u_h\|_V \leq \left(\frac{\gamma}{\alpha}\right)^{1/2} \inf_{w_h \in V_h} \|u - w_h\|_V. \quad (2.36)$$

•

2.4.2 Algebraic Form of (P_1^h)

The discrete variational problem (P_1^h) is equivalent to the solution of a linear system of equations. Indeed, if we denote by $\{\varphi^j\}_{j=1}^{N_h}$ a basis for the finite-dimensional space V_h , then every $v_h \in V_h$ has a unique representation

$$v_h = \sum_{j=1}^{N_h} v_h^{(j)} \varphi^j, \quad \text{with } \mathbf{v} = (v_h^{(1)}, \dots, v_h^{(N_h)})^T \in \mathbb{R}^{N_h}.$$

By setting $u_h = \sum_{j=1}^{N_h} u_h^{(j)} \varphi^j$, and denoting by \mathbf{u}_h the vector having as components the unknown coefficients $u_h^{(j)}$, (P_1^h) is equivalent to: find $\mathbf{u}_h \in \mathbb{R}^{N_h}$ such that

$$\sum_{j=1}^{N_h} a(\varphi^j, \varphi^i) u_h^{(j)} = f(\varphi^i) \quad \forall i = 1, \dots, N_h, \quad (2.37)$$

that is

$$\mathbb{A}_h \mathbf{u}_h = \mathbf{f}_h, \quad (2.38)$$

where $\mathbb{A}_h \in \mathbb{R}^{N_h \times N_h}$ is the *stiffness* matrix with elements $(\mathbb{A}_h)_{ij} = a(\varphi^j, \varphi^i)$ and $\mathbf{f}_h \in \mathbb{R}^{N_h}$ the vector with components $(\mathbf{f}_h)_i = f(\varphi^i)$.

The coercivity condition (2.5) states that for every $h > 0$ the matrix \mathbb{A}_h in (2.38) is positive definite (and thus nonsingular). Furthermore, is $a(\cdot, \cdot)$ is symmetric so is the matrix \mathbb{A}_h (see Exercise 9). Other properties, such as the condition number or the sparsity structure of \mathbb{A}_h , depend on the chosen basis of V_h ; for instance, basis functions with small support (like those in finite element approximation) make the entries of \mathbb{A}_h related to basis functions having non-intersecting supports to be null.

The numerical solution of (2.38) can be carried out using either direct methods, such as the LU (Cholesky in the symmetric case) factorization, or iterative methods, such as the GMRES or the conjugate gradient method in the symmetric case (see, e.g., [221, 235]).

2.4.3 Computation of the Discrete Coercivity Constant

In the context of a posteriori error estimation for reduced basis approximations addressed in Chap. 3, we will be required to compute the best discrete coercivity constant

$$\alpha_h = \inf_{v_h \in V_h} \frac{a(v_h, v_h)}{\|v_h\|_V^2}. \quad (2.39)$$

From now on, it will be understood that the constant α_h in (2.33) is indeed the one in (2.39). Our goal here is to find an algebraic representation of α_h . To this end, let us denote by \mathbb{X}_h the symmetric positive definite matrix associated to the scalar product in V , i.e.

$$(\mathbb{X}_h)_{ij} = (\varphi^i, \varphi^j)_V, \quad (2.40)$$

so that

$$\|v_h\|_V^2 = \mathbf{v}^T \mathbb{X}_h \mathbf{v} \quad \forall v_h \in V_h. \quad (2.41)$$

We can now rewrite (2.39) as

$$\alpha_h = \inf_{\mathbf{v} \in \mathbb{R}^{N_h}} \frac{\mathbf{v}^T \mathbb{A}_h \mathbf{v}}{\mathbf{v}^T \mathbb{X}_h \mathbf{v}},$$

i.e. α_h is in fact the minimum of a generalized Rayleigh quotient. Since, for any $\mathbf{v} \in \mathbb{R}^{N_h}$, $\mathbf{v}^T \mathbb{A}_h \mathbf{v} = \mathbf{v}^T \mathbb{A}_h^S \mathbf{v}$, where $\mathbb{A}_h^S = \frac{1}{2}(\mathbb{A}_h + \mathbb{A}_h^T)$ denotes the symmetric part of the matrix \mathbb{A}_h , we have that α_h is the smallest eigenvalue λ such that $(\lambda, \mathbf{v}) \in \mathbb{R}_+ \times \mathbb{R}^{N_h}$, $\mathbf{v} \neq \mathbf{0}$, satisfy

$$\mathbb{A}_h^S \mathbf{v} = \lambda \mathbb{X}_h \mathbf{v}. \quad (2.42)$$

By left-multiplying equation (2.42) by $\mathbb{X}_h^{-1/2}$ and operating the change of variable $\mathbf{w} = \mathbb{X}_h^{1/2} \mathbf{v}$, we then obtain

$$\alpha_h = \lambda_{\min}(\mathbb{X}_h^{-1/2} \mathbb{A}_h^S \mathbb{X}_h^{-1/2}). \quad (2.43)$$

2.4.4 Weakly Coercive Problems

So far, we have considered the case in which the trial space (where we seek the solution) and the test space (the one the test functions belong to) coincide. Sometimes, however, trial and test spaces indeed differ. This is the case, for instance, of a weakly coercive problem like (P_2) , where $V \neq W$, which naturally leads to using two different approximation spaces $V_h \subset V$ and $W_h \subset W$. However, in some cases, it may be convenient to introduce different approximation spaces even if $V = W$ and the problem is strongly coercive. For the sake of generality, here we consider the former case, although the formulation is the same also for the latter.

Let $V_h \subset V$ and $W_h \subset W$ be two nontrivial subspaces of V and W , respectively, with $\dim V_h = \dim W_h = N_h < \infty$. We consider the following variational problem:

find $u_h \in V_h$ such that

$$a(u_h, w_h) = f(w_h) \quad \forall w_h \in W_h. \quad (P_2^h)$$

The solution u_h of this problem is known as the *Petrov-Galerkin approximation* of u . As in the case of strongly coercive problems, we can equivalently write problem (P_2^h) in the form: find $u_h \in V_h$ such that

$$Lu_h = f \quad \text{in } W_h'. \quad (2.44)$$

The well-posedness of (P_2^h) is guaranteed by the following theorem [19].

Theorem 2.3 (Babuška). *Let the space V and W , the bilinear form $a(\cdot, \cdot)$ and the functional $f(\cdot)$ satisfy the hypotheses of Theorem 2.1. Let $V_h \subset V$ and $W_h \subset W$ be two closed subspaces. Then $a(\cdot, \cdot)$ is continuous on $V_h \times W_h$. Assume also that the bilinear form $a(\cdot, \cdot)$ satisfies the discrete inf-sup condition*

$$\beta_h = \inf_{v_h \in V_h} \sup_{w_h \in W_h} \frac{a(v_h, w_h)}{\|v_h\|_V \|w_h\|_W} > 0. \quad (2.45)$$

Then, for every $h > 0$, problem (P_2^h) has a unique solution $u_h \in V_h$. Moreover, that solution satisfies the stability estimate

$$\|u_h\|_V \leq \frac{1}{\beta_h} \|f\|_{W'}$$

and, if $u \in V$ denotes the unique solution of (P_2) , the optimal error inequality

$$\|u - u_h\|_V \leq \frac{\gamma}{\beta_h} \inf_{v_h \in V_h} \|u - v_h\|_V. \quad (2.46)$$

Remark 2.6. In the original formulation of this theorem [19] the constant γ/β_h is replaced by $1 + \gamma/\beta_h$. However, in the Hilbert space setting we can remove the “1” in the constant (see, e.g., [263]). •

While the continuity of the bilinear form $a(\cdot, \cdot)$ over $V_h \times V_h$ (respectively, $V_h \times W_h$) is automatically inherited from the continuity property over $V \times V$ (respectively, $V \times W$), we note that:

- in the coercive case, coercivity of $a(\cdot, \cdot)$ over V_h automatically follows from the coercivity property fulfilled by a over V , so that the conformity property $V_h \subset V$ is the only property which is needed to ensure the stability of the discrete variational problem;
- in the weakly coercive case, the inclusions $V_h \subset V$ and $W_h \subset W$ are not sufficient, by themselves, to ensure the fulfillment of the discrete inf-sup condition (2.45). In other words, the conformity of the finite-dimensional subspaces is a necessary but not a sufficient condition to ensure the stability, even if the inf-sup conditions (2.11)–(2.12) hold at the continuous level.

Consequently, the discrete weak coercivity (or inf-sup) assumption (2.45) *must* be explicitly required (as additional hypothesis) on the bilinear form acting on the finite-dimensional subspaces.

In other words, V_h and W_h must be built in such a way that the discrete inf-sup condition (2.45) holds. This restricts the choice of the admissible discretization spaces wherein the problem can be approximated.

Remark 2.7. Given a trial space V_h , a general strategy to obtain a well-posed Petrov-Galerkin approximation of a weakly coercive problem is to define [34, 85] the following *optimal* test space

$$W_h = \text{span}\{T\varphi_i : 1 \leq i \leq N_h\}, \quad (2.47)$$

where the operator $T : V \rightarrow W$ – called *supremizer* operator – is defined by

$$(Tv, w)_W = a(v, w) \quad \forall w \in W. \quad (2.48)$$

Optimality should be intended in the sense that it generates the best possible ratio between the continuity and the stability constant in estimate (2.46), when V is endowed with the energy norm $\|v\|_E = \sqrt{(v, v)_E}$, generated by the following inner product

$$(v, z)_E = (Tv, Tz)_W \quad \forall v, z \in V.$$

Indeed, the error in the Petrov-Galerkin approximation (P_2^h) with W_h as in (2.47) equals the best approximation error in the energy norm [85]

$$\|u - u_h\|_E = \inf_{v \in V_h} \|u - v\|_E. \quad (2.49)$$

We just remark that, even though inexpensive practical realizations of optimal test spaces are available (see e.g. the Discontinuous Petrov-Galerkin method [85]), in general the construction of the space W_h is rather involved.

In this respect, we anticipate that, when dealing with reduced spaces of very small dimension $N \ll N_h$, the construction of optimal test spaces is more straightforward, as it will be detailed in Sect. 3.3. •

2.4.5 Algebraic Form of (P_2^h)

As done for problem (P_1^h) , we can derive an equivalent algebraic formulation of problem (P_2^h) . We remember that $\{\varphi^j\}_{j=1}^{N_h}$ denotes a basis for the space V_h and indicate by $\{\phi^j\}_{j=1}^{N_h}$ a basis for the space W_h , so that the functions of V_h and W_h have a unique representation:

$$\begin{aligned} v_h &= \sum_{j=1}^{N_h} v_h^{(j)} \varphi^j, \quad \text{with } \mathbf{v} = (v_h^{(1)}, \dots, v_h^{(N_h)})^T \in \mathbb{R}^{N_h} \quad \forall v_h \in V_h, \\ w_h &= \sum_{j=1}^{N_h} w_h^{(j)} \phi^j, \quad \text{with } \mathbf{w} = (w_h^{(1)}, \dots, w_h^{(N_h)})^T \in \mathbb{R}^{N_h} \quad \forall w_h \in W_h. \end{aligned} \quad (2.50)$$

By setting $u_h = \sum_{j=1}^{N_h} u_h^{(j)} \varphi^j$ and denoting by \mathbf{u}_h the vector having as components the unknown coefficients $u_h^{(j)}$, (P_2^h) is equivalent to: find $\mathbf{u}_h \in \mathbb{R}^{N_h}$ such that

$$\sum_{j=1}^{N_h} a(\varphi^j, \phi^i) u_h^{(j)} = f(\phi^i) \quad \forall i = 1, \dots, N_h, \quad (2.51)$$

that is

$$\mathbb{A}_h \mathbf{u}_h = \mathbf{f}_h, \quad (2.52)$$

where $(\mathbb{A}_h)_{ij} = a(\varphi^j, \phi^i)$ and $(\mathbf{f}_h)_i = F(\phi^i)$, for $1 \leq i, j \leq N_h$.

In this case, the inf-sup condition (2.45) states that the matrix \mathbb{A}_h is nonsingular $\forall h > 0$, without however any further implication on the sign of its eigenvalues.

Remark 2.8. While the discrete variant (2.45) of the inf-sup condition (2.11) is equivalent to $\text{Ker}(\mathbb{A}_h) = \{\mathbf{0}\}$, i.e. to the nonsingularity of the matrix \mathbb{A}_h , the discrete variant of condition (2.12), that is

$$\inf_{w_h \in W_h} \sup_{v_h \in V_h} \frac{a(v_h, w_h)}{\|v_h\|_V \|w_h\|_W} > 0$$

is equivalent to $\text{Ker}(\mathbb{A}_h^T) = \{\mathbf{0}\}$, i.e. to the nonsingularity of \mathbb{A}_h^T ; see Exercise 10. For a square matrix these two conditions are equivalent. This is why Theorem 2.3 requires only one inf-sup condition. •

2.4.6 Computation of the Discrete Inf-Sup Constant

As we did for the coercivity constant, we want to find an algebraic formula suitable to compute the discrete inf-sup constant β_h defined in (2.45). To this end, we first introduce the discrete supremizer operator $T_h : V_h \rightarrow W_h$ defined as

$$(T_h v_h, w_h)_W = a(v_h, w_h) \quad \forall w_h \in W_h. \quad (2.53)$$

Since

$$a(v_h, w_h) = {}_{V'} \langle L v_h, w_h \rangle_V,$$

$T_h v_h \in W_h$ is in fact the Riesz representative of $L v_h \in W'_h$, so that (see Appendix A)

$$\|T_h v_h\|_W = \|L v_h\|_{W'_h} = \sup_{w_h \in W_h} \frac{a(v_h, w_h)}{\|w_h\|_W}. \quad (2.54)$$

Then, by the definition of supremizer and using (2.54), we obtain the following characterization of the inf-sup constant β_h

$$\beta_h = \inf_{v_h \in V_h} \frac{\|T_h v_h\|_W}{\|v_h\|_V},$$

or, equivalently,

$$\beta_h^2 = \inf_{v_h \in V_h} \frac{\|T_h v_h\|_W^2}{\|v_h\|_V^2}. \quad (2.55)$$

Let us now denote by \mathbb{Y}_h the matrix associated with the scalar product in W , that is

$$(\mathbb{Y}_h)_{ij} = (\phi^j, \phi^i)_W,$$

so that

$$\|w_h\|_W = \mathbf{w}^T \mathbb{Y}_h \mathbf{w} \quad \forall v_h \in V_h, w_h \in W_h.$$

If we denote by \mathbf{t} the vector of coefficients in the expansion of $T_h v_h \in W_h$ with respect to the basis of W_h , by the definition of supremizer we obtain

$$\mathbf{w}^T \mathbb{Y}_h \mathbf{t} = \mathbf{w}^T \mathbb{A}_h^T \mathbf{v} \quad \forall \mathbf{w} \in \mathbb{R}^{N_h},$$

whence $\mathbf{t} = \mathbb{Y}_h^{-1} \mathbb{A}_h \mathbf{v}$ since \mathbf{w} is arbitrary. Now (2.55) gives (using (2.40)–(2.41))

$$\begin{aligned} \beta_h^2 &= \inf_{\mathbf{v} \in \mathbb{R}^{N_h}} \frac{\mathbf{t}^T \mathbb{Y}_h \mathbf{t}}{\mathbf{v}^T \mathbb{X}_h \mathbf{v}} \\ &= \inf_{\mathbf{v} \in \mathbb{R}^{N_h}} \frac{\mathbf{v}^T \mathbb{A}_h^T \mathbb{Y}_h^{-T} \mathbb{Y}_h \mathbb{Y}_h^{-1} \mathbb{A}_h \mathbf{v}}{\mathbf{v}^T \mathbb{X}_h \mathbf{v}} = \inf_{\mathbf{v} \in \mathbb{R}^{N_h}} \frac{\mathbf{v}^T \mathbb{A}_h^T \mathbb{Y}_h^{-1} \mathbb{A}_h \mathbf{v}}{\mathbf{v}^T \mathbb{X}_h \mathbf{v}}. \end{aligned} \quad (2.56)$$

We have thus expressed the square of the inf-sup constant β_h^2 as the minimum of the generalized Rayleigh quotient of the symmetric matrix $\mathbb{A}_h^T \mathbb{Y}_h^{-1} \mathbb{A}_h$.

Then

$$\beta_h = \sqrt{\lambda_{\min}}$$

where λ_{\min} is the smallest eigenvalue λ such that $(\lambda, \mathbf{v}) \in \mathbb{R}_+ \times \mathbb{R}^{N_h}$, $\mathbf{v} \neq \mathbf{0}$, satisfy

$$\mathbb{A}_h^T \mathbb{Y}_h^{-1} \mathbb{A}_h \mathbf{v} = \lambda \mathbb{X}_h \mathbf{v}. \quad (2.57)$$

Remark 2.9. The singular values of a generic matrix $\mathbb{B} \in \mathbb{R}^{n \times n}$ are defined as $\sigma_i(\mathbb{B}) = \sqrt{\lambda_i(\mathbb{B}^T \mathbb{B})}$, $1 \leq i \leq n$ (see Sect. 6.1). By left-multiplying equation (2.57) by $\mathbb{X}_h^{-1/2}$ and operating the change of variable $\mathbf{w} = \mathbb{X}_h^{1/2} \mathbf{v}$, we then obtain

$$\beta_h = \sigma_{\min}(\mathbb{Y}_h^{-1/2} \mathbb{A}_h \mathbb{X}_h^{-1/2}). \quad (2.58)$$

In particular, if $V_h = W_h$ and the matrix \mathbb{A}_h is symmetric, the inf-sup constant β_h becomes $\beta_h = \lambda_{\min}(\mathbb{X}_h^{-1/2} \mathbb{A}_h \mathbb{X}_h^{-1/2})$, that is (2.43). •

Remark 2.10. In the same way, we can also characterize the discrete continuity constant as

$$\gamma_h = \sup_{v_h \in V_h} \sup_{w_h \in W_h} \frac{a(v_h, w_h)}{\|v_h\|_V \|w_h\|_W} = \sigma_{\max}(\mathbb{Y}_h^{-1/2} \mathbb{A}_h \mathbb{X}_h^{-1/2}). \quad (2.59)$$

•

2.4.7 Saddle-Point Problems

Let $X_h \subset X$ and $Q_h \subset Q$ be two subspaces of X and Q , respectively. We consider the following variational problem:

find $(x_h, p_h) \in X_h \times Q_h$ such that

$$\begin{cases} d(x_h, w_h) + b(w_h, p_h) = f_1(w_h) & \forall w_h \in X_h \\ b(x_h, q_h) = f_2(q_h) & \forall q_h \in Q_h \end{cases} \quad (P_3^h)$$

that represents a Galerkin approximation of the mixed variational problem (P_3) . Let

$$X_0^h = \{w_h \in X_h : b(w_h, q_h) = 0 \quad \forall q_h \in Q_h\} \subset X_h.$$

The well-posedness of (P_3^h) follows by the discrete counterpart of Brezzi theorem [32]; see also Exercise 11.

Theorem 2.4 (Brezzi). *Let the space X and Q , the bilinear forms $d(\cdot, \cdot)$, $b(\cdot, \cdot)$ and the functionals $f_1(\cdot)$ and $f_2(\cdot)$ satisfy the hypotheses of Theorem 2.2. Let $X_h \subset X$ and $Q_h \subset Q$ be two finite-dimensional subspaces. Then $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ are continuous on $X_h \times X_h$ and $X_h \times Q_h$, respectively. Assume that the bilinear form $d(\cdot, \cdot)$ is weakly coercive on X_0^h , i.e.*

$$\alpha_h = \inf_{x_h \in X_0^h} \sup_{w_h \in X_0^h} \frac{d(x_h, w_h)}{\|x_h\|_X \|w_h\|_X} \geq \hat{\alpha} > 0,$$

$$\inf_{w_h \in X_0^h} \sup_{x_h \in X_0^h} \frac{d(x_h, w_h)}{\|x_h\|_X \|w_h\|_X} > 0.$$

Moreover suppose that the bilinear form $b(\cdot, \cdot)$ verifies the discrete inf-sup condition

$$\beta_h^s = \inf_{q_h \in Q_h} \sup_{w_h \in X_h} \frac{b(w_h, q_h)}{\|w_h\|_X \|q_h\|_Q} \geq \hat{\beta}^s, \quad (2.60)$$

for a suitable constant $\hat{\beta}^s > 0$ independent of h . Then, for every $h > 0$, problem (P_3^h) has a unique solution $(x_h, p_h) \in X_h \times Q_h$, which satisfies the stability estimates

$$\|x_h\|_X \leq \frac{1}{\hat{\alpha}} \left[\|f_1\|_{X'} + \frac{\hat{\alpha} + \gamma_d}{\hat{\beta}^s} \|f_2\|_{Q'} \right], \quad (2.61a)$$

$$\|p_h\|_Q \leq \frac{1}{\hat{\beta}^s} \left[\left(1 + \frac{\gamma_d}{\hat{\alpha}} \right) \|f_1\|_{X'} + \frac{\gamma_d(\hat{\alpha} + \gamma_d)}{\hat{\alpha} + \hat{\beta}^s} \|f_2\|_{Q'} \right]. \quad (2.61b)$$

If $(x, p) \in X \times Q$ denotes the unique solution of (P_3) , the following optimal error inequality holds

$$\|x - x_h\|_X + \|p - p_h\|_Q \leq C \left(\inf_{w_h \in X_h} \|x - w_h\|_X + \inf_{q_h \in Q_h} \|p - q_h\|_Q \right),$$

where $C = C(\hat{\alpha}, \hat{\beta}^s, \gamma_d, \gamma_b)$ is independent of h .

Condition (2.60) is also called Ladyzhenskaia-Babuška-Brezzi (LBB) inf-sup condition. See, e.g., [32]. In the case of mixed variational problems, well-posedness of the discrete approximation (P_3^h) is ensured provided the inf-sup conditions on both bilinear forms a and b are fulfilled over $X_h \subset X$ and $Q_h \subset Q$, respectively. As in the case of weakly coercive problems, these assumptions are not automatically verified if they are valid on X and Q . This is the reason why they should be explicitly made on the spaces X_h and Q_h .

2.4.8 Algebraic Form of (P_3^h)

If $\{\varphi^i\}_{i=1}^{N_h}$ and $\{\eta^i\}_{i=1}^{M_h}$ denote two bases for X_h and Q_h respectively, being $N_h^u = \dim V_h$ and $N_h^p = \dim Q_h$, then (P_3^h) is equivalent to the linear system

$$\begin{pmatrix} \mathbb{D}_h & \mathbb{B}_h^T \\ \mathbb{B}_h & 0 \end{pmatrix} \begin{pmatrix} \mathbf{x}_h \\ \mathbf{p}_h \end{pmatrix} = \begin{pmatrix} \mathbf{f}_{1h} \\ \mathbf{f}_{2h} \end{pmatrix}, \quad (2.62)$$

where $x_h = \sum_{j=1}^{N_h^u} x_h^{(j)} \varphi^j$, $p_h = \sum_{k=1}^{N_h^p} p_h^{(k)} \eta^k$, \mathbf{x}_h and \mathbf{p}_h are the vectors having as components the unknown coefficients $x_h^{(j)}$, $p_h^{(k)}$,

$$(\mathbb{D}_h)_{ij} = d(\varphi^j, \varphi^i), \quad (\mathbb{B}_h)_{kj} = b(\varphi^j, \eta^k), \quad i, j = 1, \dots, N_h^u, \quad k = 1, \dots, N_h^p$$

are the elements of the matrices \mathbb{D}_h and \mathbb{B}_h , whereas \mathbf{f}_{1h} and \mathbf{f}_{2h} denote the vectors with components $(\mathbf{f}_{1h})_i = f_1(\varphi^i)$ and $(\mathbf{f}_{2h})_k = f_2(\eta^k)$, $i = 1, \dots, N_h^u$, $k = 1, \dots, N_h^p$.

From an algebraic standpoint, Theorem 2.4 ensures that the matrix appearing in the linear system (2.62) is nonsingular; in particular the inf-sup condition (2.60) is equivalent to the condition $\text{Ker}(\mathbb{B}_h^T) = \{\mathbf{0}\}$. Moreover, the matrix in (2.62) is always indefinite, see, e.g., [104, Chap. 6], for the proof.

2.5 Finite Element Spaces

In this book we will consider the finite element (FE) method as high-fidelity approximation technique. Here we recall the most essential ingredients of this technique, limiting ourselves, for the sake of brevity, to the case of scalar coercive second-order elliptic PDEs. For a more comprehensive presentation on this subject, including also the case of vector problems and mixed formulations, we refer the interested reader to, e.g., [38, 222, 105, 35, 32]. We also remark that other approximation techniques such as the discontinuous Galerkin method (see, e.g., [134, 227, 14]), the spectral element method (see, e.g., [50, 51, 52]) or isogeometric analysis (see [74]) could serve as high-fidelity approximations as well.

To define approximations of the space $H^1(\Omega)$ that depend on a parameter h , let us first introduce a *triangulation* of the domain Ω . For the sake of simplicity, most often we will consider domains $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$) with polygonal shape, and meshes (or grids) \mathcal{T}_h which represent their cover with non-overlapping triangles ($d = 2$) or tetrahedra ($d = 3$) K . We denote by h_K the diameter of K and define h to be the maximum value of h_K , $K \in \mathcal{T}_h$.

The most natural (and well-known) strategy to define a finite element space is to consider globally continuous functions that are polynomials of degree r on the single triangles (elements) of the triangulation \mathcal{T}_h , that is, to define

$$X_h^r = \{v_h \in C^0(\overline{\Omega}) : v_h|_K \in \mathbb{P}_r \quad \forall K \in \mathcal{T}_h\}, \quad r = 1, 2, \dots \quad (2.63)$$

X_h^r is the space of globally continuous functions that are polynomials of degree at most r on the single elements of the triangulation \mathcal{T}_h . Moreover, we can define

$$\overset{\circ}{X}_h^r = \{v_h \in X_h^r : v_h|_{\partial\Omega} = 0\} \quad (2.64)$$

and, more in general,

$$V_h = X_h^r \cap V.$$

The spaces X_h^r and $\overset{\circ}{X}_h^r$ represent suitable approximations of $H^1(\Omega)$ and $H_0^1(\Omega)$, respectively. In fact, the spaces X_h^r are all subspaces of $H^1(\Omega)$, as they are made of differentiable functions except for at most a finite number of points (the vertices of the partition \mathcal{T}_h). Moreover, dealing with element-wise polynomials makes the stiffness matrix (2.38) easy to compute. Concerning the expression of an arbitrary element of V_h , we choose a basis $\{\varphi_i\}_{i=1}^{N_h}$ such that:

- the support of the generic basis function φ_i has non-empty intersection only with the support of a small number of other functions of the basis. In this way, many elements of the stiffness matrix will be null;
- the coefficients of the expansion of a generic function $v_h \in X_h^r$ in the basis itself will be the values taken by v_h at carefully chosen points \mathbf{N}_i , with $i = 1, \dots, N_h$ of the grid \mathcal{T}_h , which we call *nodes* (and which form, in general, a superset of the vertices of \mathcal{T}_h). We call such a basis a *Lagrangian* basis.

Finally, we can state a convergence result which holds for the FE approximation (of degree r) to the solution of a general elliptic problem (like (2.1)) under the form (P_1) , showing that the finer the grid and the higher the FE polynomial degree, the faster the convergence of u_h to u (see, e.g., [216]).

Theorem 2.5. *Let $u \in V$ be the exact solution of the variational problem (P_1) and u_h its FE approximation of degree r , i.e. the solution of problem (P_1^h) where $V_h = X_h^r \cap V$. If $u \in V \cap H^{r+1}(\Omega)$, then the following a priori error estimates hold:*

$$\|u - u_h\|_{H^1(\Omega)} \leq \frac{M}{\alpha} C \left(\sum_{K \in \mathcal{T}_h} h_K^{2r} |u|_{H^{r+1}(K)}^2 \right)^{1/2}, \quad (2.65)$$

$$\|u - u_h\|_{H^1(\Omega)} \leq \frac{M}{\alpha} C h^r |u|_{H^{r+1}(\Omega)}, \quad (2.66)$$

C being a constant independent of h and u .

Note that (2.66) follows from (2.65).

In order to increase the accuracy, two different strategies can therefore be pursued: (i) decreasing h , i.e. refining the grid; (ii) increasing r , i.e. using finite elements of higher degree. However, the latter approach can only be pursued if the solution u is regular enough. In general, we can say that if $u \in C^0(\bar{\Omega}) \cap H^{p+1}(\Omega)$ for some $p > 0$, then

$$\|u - u_h\|_{H^1(\Omega)} \leq C h^s |u|_{H^{s+1}(\Omega)}, \quad s = \min\{r, p\}. \quad (2.67)$$

2.6 Exercises

1. By exploiting the bilinearity and the symmetry of the form $a : V \times V \rightarrow \mathbb{R}$, show that the quadratic functional (2.8) is such that

$$J(u + w) = J(u) + \frac{1}{2}a(w, w).$$

Then, exploiting the coercivity of a , show that $J(v) > J(u)$ for any $v \in V$, $v \neq u$. Conversely, write the extremality condition for the minimum u of J , that is, $\lim_{\varepsilon \rightarrow 0} (J(u + \varepsilon v) - J(u)) / \varepsilon = 0$ and recover the expression of (P_1) .

2. Consider the advection-diffusion-reaction equation (2.1). By integrating by parts and exploiting the Green identity formula

$$\int_{\Omega} v \operatorname{div}(\mathbb{K} \nabla w) d\Omega = \int_{\partial\Omega} v \mathbb{K} \nabla w \cdot \mathbf{n} d\Gamma - \int_{\Omega} \mathbb{K} \nabla v \cdot \nabla w d\Omega,$$

show that the weak formulation of this problem is given by (2.23)-(2.24).

3. Consider the linear elasticity equations (2.2). By using the following Green formula

$$\sum_{i,j=1}^d \int_{\Omega} \sigma_{ij}(\mathbf{u}) \varepsilon_{ij}(\mathbf{v}) d\Omega = \sum_{i,j=1}^d \int_{\partial\Omega} \sigma_{ij}(\mathbf{u}) n_j v_i d\Gamma - \sum_{i,j=1}^d \int_{\Omega} \frac{\partial \sigma_{ij}(\mathbf{u})}{\partial x_j} v_i d\Omega$$

show that the weak formulation of this problem is given by (2.26)-(2.28).

Then, prove the coercivity of the corresponding bilinear form, using the following *Korn inequality*

$$\exists C_0 > 0 : \sum_{i,j=1}^d \int_{\Omega} \varepsilon_{ij}(\mathbf{v}) \varepsilon_{ij}(\mathbf{v}) d\Omega \geq C_0 \|\mathbf{v}\|_V^2 \quad \forall \mathbf{v} \in V,$$

for the case $V = [H_{\Gamma_D}^1(\Omega)]^d$.

Finally, precise under which conditions on the regularity of the data the solution of the weak formulation (2.26) exists and is unique.

4. Prove the following identities:

$$\begin{aligned} \int_{\Omega} -v \Delta \mathbf{v} \cdot \mathbf{w} d\Omega &= \int_{\Omega} v \nabla \mathbf{v} : \nabla \mathbf{w} - \int_{\partial\Omega} v \frac{\partial \mathbf{v}}{\partial \mathbf{n}} \cdot \mathbf{w} d\Gamma \quad \forall \mathbf{v}, \mathbf{w} \in [H^1(\Omega)]^d, \\ \int_{\Omega} \nabla p \cdot \mathbf{v} d\Omega &= - \int_{\Omega} p \operatorname{div} \mathbf{v} d\Omega + \int_{\partial\Omega} p \mathbf{v} \cdot \mathbf{n} d\Gamma \quad \forall \mathbf{v} \in [H^1(\Omega)]^d, p \in L^2(\Omega). \end{aligned}$$

5. Show that:

- the bilinear form $d : X \times X \rightarrow \mathbb{R}$ defined by

$$d(\mathbf{v}, \mathbf{w}) = \int_{\Omega} \mathbf{v} \nabla \mathbf{v} : \nabla \mathbf{w} d\Omega \quad \forall \mathbf{v}, \mathbf{w} \in X$$

is continuous and coercive over $X = [H_{\Gamma_D}^1(\Omega)]^d$, using the Poincaré inequality (A.22);

- the bilinear form $b : X \times Q \rightarrow \mathbb{R}$ defined by

$$b(\mathbf{v}, q) = - \int_{\Omega} q \operatorname{div} \mathbf{v} d\Omega \quad \forall \mathbf{v}, \mathbf{w} \in X, q \in Q$$

is continuous over $X \times Q = [H_{\Gamma_D}^1(\Omega)]^d \times L^2(\Omega)$;

- provided that for any $q \in Q$ there exists $\mathbf{v} \in X$ such that $\operatorname{div} \mathbf{v} = q$ and $\|\mathbf{v}\|_X \leq C\|q\|_Q$ for $C > 0$, show that

$$\inf_{q \in Q} \sup_{\mathbf{v} \in X} \frac{b(\mathbf{v}, q)}{\|q\|_Q \|\mathbf{v}\|_X} \geq \frac{1}{C},$$

that is, $b : X \times Q \rightarrow \mathbb{R}$ is inf-sup stable over $X \times Q$, with $\beta^s = 1/C$.

6. Prove Proposition 2.3 in the case of the Stokes equations, that is, show that \mathbf{u} satisfies the weak formulation of the Stokes problem if and only if it is a minimizer of the following functional

$$\mathbf{u} = \arg \min_{\substack{\mathbf{v} \in X_0 \\ \mathbf{v} = \mathbf{g} \text{ over } \Gamma_D}} E(\mathbf{v}),$$

where $X_0 = \{\mathbf{v} \in X = [H^1(\Omega)]^d : \operatorname{div} \mathbf{v} = 0\}$ and $E : X \rightarrow \mathbb{R}$ is given by

$$E(\mathbf{v}) = \frac{1}{2} \int_{\Omega} \nabla \mathbf{v} : \nabla \mathbf{v} d\Omega - \int_{\Omega} \mathbf{f} \cdot \mathbf{v} d\Omega - \int_{\Gamma_N} \mathbf{h} \cdot \mathbf{v} d\Gamma.$$

To do this, let us introduce a multiplier $q \in Q$ related to the constraint $\operatorname{div} \mathbf{v} = 0$ and show that (\mathbf{u}, p) solves the Stokes problem (P_3) if and only if it is a saddle point of the Lagrangian functional

$$L(\mathbf{v}, q) = E(\mathbf{v}) - \int_{\Omega} q \operatorname{div} \mathbf{v} d\Omega;$$

equivalently,

$$L(\mathbf{u}, p) = \min_{\mathbf{v} \in X} \max_{q \in Q} L(\mathbf{v}, q). \quad (2.68)$$

Let us proceed as follows:

- evaluate the Gâteaux derivatives of L with respect to both \mathbf{v} and q and show that the weak formulation of the Stokes problem corresponds to the Euler-

Lagrange equations (also called first-order necessary optimality conditions) related to problem (2.68), that is,

$$\frac{\partial L}{\partial \mathbf{v}}(\mathbf{u}, \delta \mathbf{u}) = 0 \quad \forall \delta \mathbf{u} \in V,$$

$$\frac{\partial L}{\partial q}(p, \delta p) = 0 \quad \forall \delta p \in Q;$$

- show that

$$L(\mathbf{u} + \mathbf{v}, p + q) - L(\mathbf{u}, p) = \frac{1}{2} \int_{\Omega} \nabla \mathbf{u} : \nabla \mathbf{v} d\Omega - \int_{\Omega} q \operatorname{div} \mathbf{v} d\Omega;$$

- show that (\mathbf{u}, p) is a saddle point of L by considering the cases $q = 0$, $\mathbf{v} = \mathbf{0}$ and $q = \frac{1}{\varepsilon} \operatorname{div} \mathbf{v}$ and concluding about the behavior of L in a neighborhood of (\mathbf{u}, p) .
7. Prove that the solution of the Galerkin problem (P_1^h) exists and is unique, provided $a : V \times V \rightarrow \mathbb{R}$ and $f : V \rightarrow \mathbb{R}$ fulfill the hypotheses of the Lax-Milgram Lemma 2.1. Moreover, show the stability estimate (2.33). Furthermore, if $u \in V$ denotes the unique solution of (P_1) , derive (2.34) by taking advantage of the Cauchy-Schwarz inequality and the coercivity of a .
 8. Prove that u_h satisfies

$$a(u_h, v_h) = a(u, v_h) \quad \forall v_h \in V_h$$

and deduce that, in the case the bilinear form $a : V \times V \rightarrow \mathbb{R}$ is coercive, u_h minimizes $J(v_h) = a(v_h, v_h) - 2a(u, v_h)$ and therefore also $J^*(v_h) = J(v_h) + a(u, u) = a(u - v_h, u - v_h)$, the last equality being made possible since the bilinear form is symmetric.

Prove that

$$\sqrt{\alpha} \|u - v_h\|_V \leq \sqrt{a(u - v_h, u - v_h)} \leq \sqrt{\gamma} \|u - v_h\|_V$$

so that, in the case of a symmetric bilinear form, (2.34) can be improved to obtain (2.36).

9. Show that the matrix \mathbb{A}_h in (2.38) is positive definite. Then, show that if the bilinear form $a(\cdot, \cdot)$ is symmetric, so is the matrix \mathbb{A}_h .
10. Prove that the discrete variant (2.45) of the inf-sup condition (2.11) is equivalent to require that

$$\operatorname{Ker}(\mathbb{A}_h) = \{\mathbf{0}\},$$

i.e. to the nonsingularity of the matrix \mathbb{A}_h . Then, show that the discrete variant of condition (2.12), that is

$$\inf_{w_h \in W_h} \sup_{v_h \in V_h} \frac{a(v_h, w_h)}{\|v_h\|_V \|w_h\|_W} > 0$$

is equivalent to require that

$$\text{Ker}(\mathbb{A}_h^T) = \{\mathbf{0}\},$$

i.e. to the nonsingularity of \mathbb{A}_h^T . Finally, show that for a square matrix these two conditions are equivalent.

11. Prove that if the discrete inf-sup condition (2.60) is not fulfilled, the solution of problem (P_3^h) is not unique. Hence, supposing that

$$\inf_{q_h \in Q_h} \sup_{w_h \in X_h} \frac{b(w_h, q_h)}{\|w_h\|_X \|q_h\|_Q} = 0$$

show that there exists (at least one) $q_h^* \in Q_h$ such that $b(x_h, q_h^*) = 0$ for any $x_h \in X_h$; then, conclude that if $(x_h, p_h) \in X_h \times Q_h$ is solution to (P_3^h) , the same happens for $(x_h, p_h + q_h^*)$.

Chapter 3

RB Methods: Basic Principles, Basic Properties

Reduced basis methods are introduced for elliptic linear parametrized PDEs. Any reduced basis (RB) approximation is, in a nutshell, a (Petrov-)Galerkin projection onto an N -dimensional space V_N (the RB space) that approximates the high-fidelity (say, finite element) solution of the given PDE, for any choice of the parameter within a prescribed parameter set. We illustrate the main steps needed to set up such methods efficiently. We discuss in detail projection methods, which represent the main feature of these techniques, and highlight the difference between Galerkin and least-squares RB methods. We show how to obtain a suitable offline/online decomposition meant to lower the computational complexity and then derive a posteriori error estimates for bounding the error of the RB solution with respect to the underlying high-fidelity solution. We consider the rather general case of inf-sup stable operators, of which coercive operators can be regarded as a particular – yet very relevant – instance. Proper orthogonal decomposition (POD) and greedy algorithms, two major techniques employed to build reduced spaces, are described thoroughly in Chaps. 6 and 7.

3.1 Parametrized PDEs: Formulation and Assumptions

As already pointed out in Sect. 1.1, parametrized PDEs are partial differential equations that depend on a set of parameters. The latter may represent material coefficients (e.g. the Lamé coefficients in the elasticity equations, or the conductivity matrix in the advection-diffusion-reaction equation), boundary conditions and source terms. We will denote them by $\boldsymbol{\mu}_{ph}$ (the subscript standing for “physical”). Sometimes the computational domain itself can be represented in terms of (additional) parameters, say $\boldsymbol{\mu}_g$ (the subscript standing for “geometric”).

From now on, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_P)^T$ will denote an *input-parameter* vector collecting both $\boldsymbol{\mu}_{ph}$ and $\boldsymbol{\mu}_g$, whereas the input parameter set \mathcal{P} – i.e. the set of all possible inputs – is assumed to be a compact subset of \mathbb{R}^P . In those cases where the computational domain depends on a set of (geometric) parameters, it is convenient

to reformulate (and solve) the original parametrized PDE onto a new, fixed (parameter independent), *reference domain*. This in particular will allow to combine, high-fidelity solutions that would be otherwise computed on different domains and with different grids. The reference domain can be seen as a particular instance of the parametrized ones, $\Omega = \tilde{\Omega}(\boldsymbol{\mu}^{ref})$, corresponding to a specific choice $\boldsymbol{\mu}_{ref} \in \mathcal{P}$ of the parameter. In this chapter we assume that our parametrized problem, that we are about to introduce, is already set on the reference domain. In the case of a parameter-dependent *original domain*, this problem will be obtained by a suitable transformation of the original one. In this respect, a general procedure to recover the formulation of the problem on the reference domain will be illustrated in Sect. 8.5.

Let us denote by $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$ the reference domain, $V = V(\Omega)$ a suitable Hilbert space, V' its dual. For every $\boldsymbol{\mu} \in \mathcal{P}$, $L(\boldsymbol{\mu}) : V \rightarrow V'$ denotes a second-order differential operator and $f(\boldsymbol{\mu}) : V \rightarrow \mathbb{R}$ a linear and continuous form on V , that is an element of V' . In abstract form, the parametrized problem we focus on can be written as follows:

given $\boldsymbol{\mu} \in \mathcal{P}$, find the solution $u(\boldsymbol{\mu}) \in V$ of

$$L(\boldsymbol{\mu})u(\boldsymbol{\mu}) = f(\boldsymbol{\mu}) \quad \text{in } V'. \quad (3.1)$$

The strong formulation (3.1) will be useful in the remainder to point out several properties of a RB method from both an algebraic and a geometric standpoint. For the sake of construction and numerical approximation, we rely instead on the weak formulation of problem (3.1), which reads as:

given $\boldsymbol{\mu} \in \mathcal{P}$, find $u(\boldsymbol{\mu}) \in V$ such that

$$a(u(\boldsymbol{\mu}), v; \boldsymbol{\mu}) = f(v; \boldsymbol{\mu}) \quad \forall v \in V, \quad (3.2)$$

where the *parametrized* bilinear form $a(\cdot, \cdot; \boldsymbol{\mu}) : V \times V \rightarrow \mathbb{R}$ is obtained from $L(\boldsymbol{\mu})$,

$$a(u, v; \boldsymbol{\mu}) = {}_{V'}\langle L(\boldsymbol{\mu})u, v \rangle_V \quad \forall u, v \in V \quad (3.3)$$

and encodes the differential operator, while the linear form $f(\cdot; \boldsymbol{\mu}) : V \rightarrow \mathbb{R}$ denotes

$$f(v; \boldsymbol{\mu}) = {}_{V'}\langle f(\boldsymbol{\mu}), v \rangle_V. \quad (3.4)$$

In the case of second-order elliptic PDEs, $[H_0^1(\Omega)]^d \subseteq V \subseteq [H^1(\Omega)]^d$.

Let us now state some fundamental assumptions for the well-posedness of the parametrized problem (3.2); they mimic those made in Sect. 2.2 for the non-parametric problem (P_2). We assume that $a(\cdot, \cdot; \boldsymbol{\mu}) : V \times V \rightarrow \mathbb{R}$ is continuous over $V \times V$ for any $\boldsymbol{\mu} \in \mathcal{P}$, i.e there exists a constant $\bar{\gamma} > 0$ such that

$$\gamma(\boldsymbol{\mu}) = \sup_{v \in V} \sup_{w \in V} \frac{a(v, w; \boldsymbol{\mu})}{\|v\|_V \|w\|_V} < \bar{\gamma} \quad \forall \boldsymbol{\mu} \in \mathcal{P}. \quad (3.5)$$

Since $f(\boldsymbol{\mu}) \in V'$ for any $\boldsymbol{\mu} \in \mathcal{P}$, also $f(\cdot; \boldsymbol{\mu})$ is a continuous linear form, i.e there exists a constant $\tilde{\gamma}_F > 0$ such that

$$\gamma_F(\boldsymbol{\mu}) = \sup_{w \in W} \frac{f(w; \boldsymbol{\mu})}{\|w\|_W} < \tilde{\gamma}_F \quad \forall \boldsymbol{\mu} \in \mathcal{P}. \quad (3.6)$$

Here $\gamma(\boldsymbol{\mu})$ and $\gamma_F(\boldsymbol{\mu})$ represent the continuity factors of $a(\cdot, \cdot; \boldsymbol{\mu})$ and of $f(\cdot; \boldsymbol{\mu})$, respectively. We name them *factors* instead of *constants* (as usually done) because in this case they are indeed $\boldsymbol{\mu}$ -dependent functions. Regarding stability, we assume that there exists a constant $\beta_0 > 0$ such that, for each $\boldsymbol{\mu} \in \mathcal{P}$,

$$\beta(\boldsymbol{\mu}) = \inf_{v \in V} \sup_{w \in V} \frac{a(v, w; \boldsymbol{\mu})}{\|v\|_V \|w\|_V} \geq \beta_0, \quad (3.7)$$

$$\inf_{v \in V} \sup_{w \in V} \frac{a(v, w; \boldsymbol{\mu})}{\|v\|_V \|w\|_V} > 0. \quad (3.8)$$

We call $\beta(\boldsymbol{\mu})$ the inf-sup stability factor and we say that $a(\cdot, \cdot; \boldsymbol{\mu})$ is *inf-sup* stable.

Remark 3.1. A particular (and remarkable) case where assumptions (3.7)–(3.8) are verified is when, for each $\boldsymbol{\mu} \in \mathcal{P}$, there exists $\alpha_0 > 0$ such that

$$\alpha(\boldsymbol{\mu}) = \inf_{v \in V} \frac{a(v, v; \boldsymbol{\mu})}{\|v\|_V^2} \geq \alpha_0. \quad (3.9)$$

In this case $a(\cdot, \cdot; \boldsymbol{\mu})$ is coercive (see Sect. 2.2.1) and $\alpha(\boldsymbol{\mu})$ is the coercivity factor. In the context of RB methods several papers – see e.g. [208, 231, 219] and references therein – have focused on either coercive or non-coercive problems. Here we prefer to directly treat the more general case of inf-sup stable problems. •

Provided the continuity properties (3.5)–(3.6) and the stability assumptions are verified, problem (3.2) admits a unique solution, thanks to Nečas theorem (see Theorem 2.1). Moreover, the following stability estimate holds for all $\boldsymbol{\mu} \in \mathcal{P}$

$$\|u(\boldsymbol{\mu})\|_V \leq \frac{1}{\beta(\boldsymbol{\mu})} \|f(\cdot; \boldsymbol{\mu})\|_{V'} \leq \frac{1}{\beta_0} \|f(\cdot; \boldsymbol{\mu})\|_{V'}. \quad (3.10)$$

3.2 High-Fidelity Discretization Techniques

The Galerkin high-fidelity approximation of an inf-sup stable problem under the form (3.2) reads:

find $u_h(\boldsymbol{\mu}) \in V_h$ such that

$$a(u_h(\boldsymbol{\mu}), v_h; \boldsymbol{\mu}) = f(v_h; \boldsymbol{\mu}) \quad \forall v_h \in V_h. \quad (3.11)$$

Following what done in Sect. 2.4.4, problem (3.11) can be equivalently written as:

find $u_h(\boldsymbol{\mu}) \in V_h$ such that

$$L(\boldsymbol{\mu})u_h(\boldsymbol{\mu}) = f(\boldsymbol{\mu}) \quad \text{in } V_h'. \quad (3.12)$$

For the discrete problem (3.11) to be well-posed we assume that

$$\exists \beta_{0,h} > 0 : \beta_h(\boldsymbol{\mu}) = \inf_{v_h \in V_h} \sup_{w_h \in V_h} \frac{a(v_h, w_h; \boldsymbol{\mu})}{\|v_h\|_V \|w_h\|_V} \geq \beta_{0,h} \quad \forall \boldsymbol{\mu} \in \mathcal{P}. \quad (3.13)$$

For the sake of the analysis, we also define the discrete continuity factor

$$\gamma_h(\boldsymbol{\mu}) = \sup_{v_h \in V_h} \sup_{w_h \in V_h} \frac{a(v_h, w_h; \boldsymbol{\mu})}{\|v_h\|_V \|w_h\|_V} \quad (3.14)$$

and note that $\gamma_h(\boldsymbol{\mu}) \leq \gamma(\boldsymbol{\mu})$.

Under the above assumption, problem (3.11) admits a unique solution thanks to Babuška theorem (see Theorem 2.3). Moreover, the following stability estimate

$$\|u_h(\boldsymbol{\mu})\|_V \leq \frac{1}{\beta_{0,h}} \|f(\cdot; \boldsymbol{\mu})\|_{V'}$$

holds. Furthermore, for all $\boldsymbol{\mu} \in \mathcal{P}$

$$\|u(\boldsymbol{\mu}) - u_h(\boldsymbol{\mu})\|_V \leq \frac{\bar{\gamma}}{\beta_{0,h}} \min_{z_h \in V_h} \|u(\boldsymbol{\mu}) - z_h\|_V. \quad (3.15)$$

As we have seen in Sect. 2.4.2, the Galerkin high-fidelity approximation (3.11) is equivalent to the solution of the following linear system

$$\mathbb{A}_h(\boldsymbol{\mu}) \mathbf{u}_h(\boldsymbol{\mu}) = \mathbf{f}_h(\boldsymbol{\mu}) \quad (3.16)$$

where (see Sect. 2.4.5) $\{\varphi^j\}_{j=1}^{N_h}$ denotes a basis for V_h and $\mathbb{A}_h(\boldsymbol{\mu}) \in \mathbb{R}^{N_h \times N_h}$, $\mathbf{f}_h(\boldsymbol{\mu}) \in \mathbb{R}^{N_h}$ are the $\boldsymbol{\mu}$ -dependent *stiffness* matrix and the right-hand side vector with elements

$$(\mathbb{A}_h(\boldsymbol{\mu}))_{ij} = a(\varphi^j, \varphi^i; \boldsymbol{\mu}), \quad (\mathbf{f}_h(\boldsymbol{\mu}))_i = f(\varphi^i; \boldsymbol{\mu}), \quad 1 \leq i, j \leq N_h.$$

Unless otherwise stated, we endow the finite-dimensional space $V_h \subset V$ with the norm $\|\cdot\|_V$ inherited from that of the Hilbert spaces V . We will provide further details on discrete scalar products and norms, and related formulations, in Sect. 4.1.

When $\mathbb{A}_h(\boldsymbol{\mu})$ and $\mathbf{f}_h(\boldsymbol{\mu})$ both depend affinely on the parameters, thanks to (1.9) the high-fidelity problem (3.16) can be solved using the Algorithm 3.1. We will come back on the affine parametric dependence property several times in this chapter and throughout the book.

Algorithm 3.1 High-fidelity system assembling and solving

```

1: function  $\mathbf{u}_h(\boldsymbol{\mu}) = \text{SOLVEHFSYSTEM}(\mathbb{A}_h^q, \mathbf{f}_h^q, \theta_a^q, \theta_f^q, \boldsymbol{\mu})$ 
2:    $\mathbb{A}_h(\boldsymbol{\mu}) = 0, \mathbf{f}_h(\boldsymbol{\mu}) = \mathbf{0}$ 
3:   for  $q = 1 : Q_a$ 
4:      $\mathbb{A}_h(\boldsymbol{\mu}) \leftarrow \mathbb{A}_h(\boldsymbol{\mu}) + \theta_a^q(\boldsymbol{\mu})\mathbb{A}_h^q$ 
5:   end for
6:   for  $q = 1 : Q_f$ 
7:      $\mathbf{f}_h(\boldsymbol{\mu}) \leftarrow \mathbf{f}_h(\boldsymbol{\mu}) + \theta_f^q(\boldsymbol{\mu})\mathbf{f}_h^q$ 
8:   end for
9:   solve linear system  $\mathbb{A}_h(\boldsymbol{\mu})\mathbf{u}_h(\boldsymbol{\mu}) = \mathbf{f}_h(\boldsymbol{\mu})$ 
10: end function

```

3.3 Reduced Basis Methods

Solving the high-fidelity problem (3.11) for any value $\boldsymbol{\mu} \in \mathcal{P}$ entails severe computational costs, which can be mitigated by introducing a suitable reduced-order approximation. With the aim of exploiting – rather than ignoring – the $\boldsymbol{\mu}$ -dependence of the solution, we start with the following observation.

Given the solution set (see Fig. 3.1 for a graphical sketch)

$$\mathcal{M}_h = \{u_h(\boldsymbol{\mu}) \in V_h : \boldsymbol{\mu} \in \mathcal{P}\} \subset V_h \quad (3.17)$$

of the high-fidelity solutions generated as $\boldsymbol{\mu}$ varies over the parameter domain \mathcal{P} , we expect that any $u_h(\boldsymbol{\mu})$ could be well approximated by linearly combining few elements of \mathcal{M}_h . This is true especially when \mathcal{M}_h is low-dimensional and smooth; we will further discuss these issues in Chap. 5. The idea behind RB methods is to generate an approximate solution to problem (3.11) belonging to a low-dimensional subspace $V_N \subset V_h$ of dimension $N \ll N_h$. The smaller N , the cheaper the reduced problem to solve. Precisely, setting a RB method entails:

1. the construction of a basis of V_N . We start from a set of high-fidelity solutions

$$\{u_h(\boldsymbol{\mu}^1), \dots, u_h(\boldsymbol{\mu}^N)\}, \quad (3.18)$$

that we call *snapshots*, corresponding to a set of N selected parameters

$$S_N = \{\boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^N\} \subset \mathcal{P}. \quad (3.19)$$

Then, we generate a set of N functions

$$\{\zeta_1, \dots, \zeta_N\}, \quad (3.20)$$

called the *reduced basis*, by orthonormalization of the snapshots with respect to a suitable scalar product $(\cdot, \cdot)_N$, that is

$$(\zeta_m, \zeta_k)_N = \delta_{km}, \quad 1 \leq k, m \leq N. \quad (3.21)$$

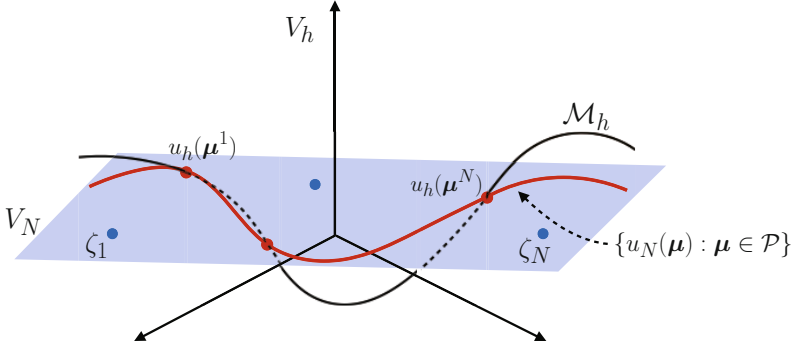


Fig. 3.1 The “snapshots” $u_h(\boldsymbol{\mu}^n)$, $1 \leq n \leq N$ on the parametric manifold \mathcal{M}_h , the RB space $V_N = \text{span}\{\zeta_1, \dots, \zeta_N\} = \text{span}\{u_h(\boldsymbol{\mu}^1), \dots, u_h(\boldsymbol{\mu}^N)\}$ here indicated by the hyperplane and the RB solutions $u_N(\boldsymbol{\mu}) \in V_N$, $\boldsymbol{\mu} \in \mathcal{P}$, represented by the red curve. Case of $P = 1$ parameter

Typically, $(\cdot, \cdot)_N = (\cdot, \cdot)_V$, the V -scalar product. The functions $\{\zeta_1, \dots, \zeta_N\}$ are therefore called *reduced basis functions*, and generate the *reduced basis space*

$$V_N = \text{span}\{\zeta_1, \dots, \zeta_N\}. \quad (3.22)$$

The spaces $\{V_N, N \geq 1\}$ are therefore nested, that is $V_N \supset V_{N-1}$, $N \geq 2$. By construction, the reduced basis functions are no longer¹ solutions of the high-fidelity problem. However,

$$V_N = \text{span}\{\zeta_1, \dots, \zeta_N\} = \text{span}\{u_h(\boldsymbol{\mu}^1), \dots, u_h(\boldsymbol{\mu}^N)\}.$$

The construction of the reduced basis addressed so far is typical of the greedy algorithm, where the snapshots are selected according to a suitable optimality criterion (see Chap. 7). Other approaches are however possible for the generation of the reduced basis: a remarkable instance is the one based on the proper orthogonal decomposition technique, see Chap. 6;

2. the computation of the RB solution $u_N(\boldsymbol{\mu}) \in V_N$, expressed as a linear combination of the reduced basis functions,

$$u_N(\boldsymbol{\mu}) = \sum_{m=1}^N u_N^{(m)}(\boldsymbol{\mu}) \zeta_m \quad (3.23)$$

where $\mathbf{u}_N(\boldsymbol{\mu}) = (u_N^{(1)}(\boldsymbol{\mu}), \dots, u_N^{(N)}(\boldsymbol{\mu})) \in \mathbb{R}^N$ denotes the RB coefficients, also called the generalized coordinates, of $u_N(\boldsymbol{\mu})$ in the reduced basis;

3. the setup of a reduced problem for determining the unknown coefficients $\mathbf{u}_N(\boldsymbol{\mu}) \in \mathbb{R}^N$.

¹ That is, there is no $\boldsymbol{\mu}^j \in \mathcal{P}$ such that, in general, $\zeta_j = u_h(\boldsymbol{\mu}^j)$.

As anticipated in the Introduction, we generate the reduced problem via a projection approach. More precisely, the reduced problem will consist of a set of N equations that are obtained by imposing N (independent) conditions.

The latter enforce the orthogonality of the residual of the high-fidelity problem (3.12) computed on the RB solution

$$r(\boldsymbol{\mu}) = f(\boldsymbol{\mu}) - L(\boldsymbol{\mu})u_N(\boldsymbol{\mu}) \quad (3.24)$$

to the functions of a subspace $W_N \subset V_h$ (of dimension N). This yields the following *Petrov-Galerkin reduced basis* (PG-RB) problem

find $u_N(\boldsymbol{\mu}) \in V_N$ such that

$$\langle L(\boldsymbol{\mu})u_N(\boldsymbol{\mu}) - f(\boldsymbol{\mu}), w_N \rangle = 0 \quad \forall w_N \in W_N, \quad (3.25)$$

or, equivalently,

find $u_N(\boldsymbol{\mu}) \in V_N$ such that

$$a(u_N(\boldsymbol{\mu}), w_N; \boldsymbol{\mu}) = f(w_N; \boldsymbol{\mu}) \quad \forall w_N \in W_N. \quad (3.26)$$

The *Galerkin reduced basis* (G-RB) problem corresponds to the case $W_N = V_N$. For its role in (3.25)–(3.26), W_N is also called *test subspace*.

In this chapter we focus on points 2 and 3, deferring to Chaps. 6 and 7 the discussion of point 1, that is how to construct the RB space V_N . We deal now with the well-posedness of the RB problem (3.26), distinguishing between the Galerkin and the Petrov-Galerkin cases.

3.3.1 Galerkin RB Method

Given $\boldsymbol{\mu} \in \mathcal{P}$, the Galerkin reduced basis (G-RB) approximation of problem (3.2) reads:

find $u_N(\boldsymbol{\mu}) \in V_N$ such that

$$a(u_N(\boldsymbol{\mu}), v_N; \boldsymbol{\mu}) = f(v_N; \boldsymbol{\mu}) \quad \forall v_N \in V_N. \quad (3.27)$$

An easy case occurs when $a(\cdot, \cdot; \boldsymbol{\mu})$ is coercive for any $\boldsymbol{\mu} \in \mathcal{P}$, that is, it fulfills (3.9). A straightforward application of the Lax-Milgram lemma provides the following

Lemma 3.1. *Under the assumptions of Lemma 2.2, for any $\boldsymbol{\mu} \in \mathcal{P}$ the G-RB problem (3.27) has a unique solution $u_N(\boldsymbol{\mu}) \in V_N$, which satisfies the stability estimate*

$$\|u_N(\boldsymbol{\mu})\|_V \leq \frac{1}{\alpha_N(\boldsymbol{\mu})} \|f(\cdot; \boldsymbol{\mu})\|_{V'} \quad (3.28)$$

where

$$\alpha_N(\boldsymbol{\mu}) = \inf_{v \in V_N} \frac{a(v, v; \boldsymbol{\mu})}{\|v\|_V^2} \quad (3.29)$$

is the stability factor of the G-RB problem.

Note that since $V_N \subset V_h$

$$\alpha_N(\boldsymbol{\mu}) \geq \alpha_h(\boldsymbol{\mu}) \quad \forall \boldsymbol{\mu} \in \mathcal{P}$$

where

$$\alpha_h(\boldsymbol{\mu}) = \inf_{v \in V_h} \frac{a(v, v; \boldsymbol{\mu})}{\|v\|_V^2} \quad (3.30)$$

is the stability factor of the high-fidelity problem. The well-posedness of the G-RB problem is therefore inherited from that of the high-fidelity problem: if $a(\cdot, \cdot; \boldsymbol{\mu})$ is coercive for any $\boldsymbol{\mu} \in \mathcal{P}$ over $V_h \times V_h$, then it is coercive over $V_N \times V_N$, too.

We denote by $(v, w)_{\boldsymbol{\mu}} = a(v, w; \boldsymbol{\mu})$ and $\|v\|_{\boldsymbol{\mu}} = \sqrt{(v, v)_{\boldsymbol{\mu}}}$, $\forall v, w \in V$, the inner product and the energy norm induced by the bilinear form $a(\cdot, \cdot; \boldsymbol{\mu})$, provided this latter is symmetric for any $\boldsymbol{\mu} \in \mathcal{P}$. By subtraction of (3.11) from (3.27) we obtain

$$a(u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu}), v_N; \boldsymbol{\mu}) = 0 \quad \forall v_N \in V_N. \quad (3.31)$$

In the symmetric coercive case, this is a *Galerkin orthogonality property* for the reduced problem, as it expresses the orthogonality of the error $u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})$ to the subspace V_N , according to the scalar product $(\cdot, \cdot)_{\boldsymbol{\mu}}$. The RB solution $u_N(\boldsymbol{\mu})$ is therefore the projection of $u_h(\boldsymbol{\mu})$ onto V_N , according to the scalar product $(\cdot, \cdot)_{\boldsymbol{\mu}}$. This proves why the G-RB method is a projection-based method. Indeed, the orthogonal projection theorem (see Appendix A) yields the following property.

Proposition 3.1. *If $a(\cdot, \cdot; \boldsymbol{\mu})$ is symmetric and coercive, then the solution $u_N(\boldsymbol{\mu}) \in V_N$ to (3.27) satisfies the following optimality property*

$$u_N(\boldsymbol{\mu}) = \arg \min_{v \in V_N} \|u_h(\boldsymbol{\mu}) - v\|_{\boldsymbol{\mu}}^2. \quad (3.32)$$

Note that also the converse is true, that is, if $u_N(\boldsymbol{\mu}) \in V_N$ fulfills (3.32), then it solves (3.27); see Exercise 1. In other words, in the energy norm the Galerkin procedure automatically selects the *best* combination of snapshots.

Let us remark that, by choosing the V -norm instead of the energy norm, we would find (see Exercise 2)

$$\|u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_V \leq \left(\frac{\gamma_h(\boldsymbol{\mu})}{\alpha_N(\boldsymbol{\mu})} \right)^{1/2} \inf_{w \in V_N} \|u_h(\boldsymbol{\mu}) - w\|_V, \quad (3.33)$$

where $\gamma_h(\boldsymbol{\mu})$ and $\alpha_N(\boldsymbol{\mu})$ are defined in (3.14), (3.29), respectively. In this case, it is also possible to show (see Exercise 3) that the linear output $z(\boldsymbol{\mu}) = f(u(\boldsymbol{\mu}))$ – called *compliance* – converges as the “square” of the solution error in the energy norm.

If $a(\cdot, \cdot; \boldsymbol{\mu})$ is non coercive over $V_h \times V_h$, the well-posedness of the G-RB problem should be investigated according to the following result, which can be obtained as a direct consequence of the Babuška Theorem 2.3 in the case where the trial space and the test space do coincide.

Theorem 3.1. *Assume that conditions (3.6), (3.14) hold. Assume moreover that the bilinear form $a(\cdot, \cdot; \boldsymbol{\mu})$ satisfies the following inf-sup condition*

$$\exists \beta_{0,N} > 0 \quad \text{s.t.} \quad \beta_N(\boldsymbol{\mu}) = \inf_{v_N \in V_N} \sup_{w_N \in V_N} \frac{a(v_N, w_N; \boldsymbol{\mu})}{\|v_N\|_V \|w_N\|_V} \geq \beta_{0,N}. \quad (3.34)$$

Then, for any $\boldsymbol{\mu} \in \mathcal{P}$ the G-RB problem (3.27) has a unique solution $u_N(\boldsymbol{\mu}) \in V_N$, which satisfies the stability estimate

$$\|u_N(\boldsymbol{\mu})\|_V \leq \frac{1}{\beta_N(\boldsymbol{\mu})} \|f(\cdot; \boldsymbol{\mu})\|_{V'}.$$

A remarkable case of G-RB method for a non-coercive problem is that of Stokes equations (2.3). Another non-coercive case is that of the optimality system resulting from a PDE-constrained optimization problem, that will be addressed in Chap. 12. In these cases the well-posedness of problem (3.27) does not follow from that of the high-fidelity one; indeed, we need to prove the inf-sup condition (3.34).

Remark 3.2. Even though parametrized advection-diffusion problems with dominating advection satisfy the coercivity assumption, their approximation poses serious issues unless suitable stabilization techniques, such as the artificial viscosity or the Streamline-Upwind Petrov Galerkin (SUPG) methods, are employed – see, e.g., [216, Chap. 12]. If such a technique is considered at the reduced-order level, too – that is, the bilinear form appearing in both the high-fidelity problem and the RB problem is the same – the resulting RB approximation will be stable as well. In other words, $\alpha_N(\boldsymbol{\mu}) \geq \alpha_h(\boldsymbol{\mu})$ for any $\boldsymbol{\mu} \in \mathcal{P}$ so that the RB problem does not suffer from a lack of stability if compared to the high-fidelity problem, and (3.28) yields

$$\|u_N(\boldsymbol{\mu})\|_V \leq \frac{1}{\alpha_N(\boldsymbol{\mu})} \|f(\cdot; \boldsymbol{\mu})\|_{V'} \leq \frac{1}{\alpha_h(\boldsymbol{\mu})} \|f(\cdot; \boldsymbol{\mu})\|_{V'}, \quad \forall \boldsymbol{\mu} \in \mathcal{P}.$$

However, one might be tempted to construct a RB approximation by considering different operators in the high-fidelity and the RB problem. This occurs when a stabilization technique is considered at the high-fidelity level (e.g., snapshots are computed by considering a SUPG method) but not at the reduced level (where, for instance, one could use a pure Galerkin method in V_N , by considering the original bilinear form without any stabilization term). In such a case, we warn the reader that the resulting RB approximation might not preserve stability. See, e.g., [76, 206] for further details. •

3.3.2 Least-Squares RB Method

A special instance of PG-RB problem is the *least-squares reduced basis* (LS-RB) method which corresponds to choosing the test space as

$$W_N = R_{V_h}^{-1} L(\boldsymbol{\mu}) V_N.$$

Here $R_{V_h}^{-1} : V'_h \rightarrow V_h$, defined by

$$(R_{V_h}^{-1} f, y)_V =_{V'} \langle f, y \rangle_V \quad \forall f \in V'_h, y \in V_h \quad (3.35)$$

is the inverse of the Riesz map (see also Appendix A) $R_{V_h} : V_h \rightarrow V'_h$, given by

$$_{V'} \langle R_{V_h} x, y \rangle_V = (x, y)_V \quad \forall x, y \in V_h.$$

Remark 3.3. For a generic Hilbert space V and for any $f \in V'$, $w = R_{V_h}^{-1} f$ is the Riesz representative of $f \in V'$ (see Appendix A.2) and is obtained by solving the variational problem

$$(w, v)_V =_{V'} \langle f, v \rangle_V \quad \forall v \in V.$$

Indeed, the Riesz representative of f depends on the choice of the inner product over V . In the case of an inner product induced by a symmetric, coercive bilinear form $a(\cdot, \cdot)$, this yields the variational problem

$$a(w, v) =_{V'} \langle f, v \rangle_V \quad \forall v \in V. \quad •$$

Given $\boldsymbol{\mu} \in \mathcal{P}$, the LS-RB approximation of problem (3.2) reads therefore:

find $u_N(\boldsymbol{\mu}) \in V_N$ such that

$$a(u_N(\boldsymbol{\mu}), w_N; \boldsymbol{\mu}) = f(w_N; \boldsymbol{\mu}) \quad \forall w_N \in W_N = R_{V_h}^{-1} L(\boldsymbol{\mu}) V_N. \quad (3.36)$$

The well-posedness of problem (3.36) is analyzed in the following

Proposition 3.2. *Assume that conditions (3.6), (3.14) hold. Assume moreover that the bilinear form $a(\cdot, \cdot; \boldsymbol{\mu})$ is inf-sup stable over $V_h \times V_h$ in the sense of (3.13). If $W_N = R_{V_h}^{-1} L(\boldsymbol{\mu}) V_N$, then*

$$\beta_N(\boldsymbol{\mu}) = \inf_{v_N \in V_N} \sup_{w_N \in W_N} \frac{a(v_N, w_N; \boldsymbol{\mu})}{\|v_N\|_V \|w_N\|_V} \geq \beta_h(\boldsymbol{\mu}) > 0 \quad \forall \boldsymbol{\mu} \in \mathcal{P} \quad (3.37)$$

and the LS-RB problem (3.36) has a unique solution $u_N(\boldsymbol{\mu}) \in V_N$ for any $\boldsymbol{\mu} \in \mathcal{P}$, which satisfies the stability estimate

$$\|u_N(\boldsymbol{\mu})\|_V \leq \frac{1}{\beta_N(\boldsymbol{\mu})} \|f(\cdot; \boldsymbol{\mu})\|_{V'}. \quad (3.38)$$

Proof. We have

$$\begin{aligned} \beta_N(\boldsymbol{\mu}) &= \inf_{v \in V_N} \sup_{w \in W_N} \frac{v' \langle L(\boldsymbol{\mu})v, w \rangle_V}{\|v\|_V \|w\|_V} \geq \inf_{v \in V_N} \frac{v' \langle L(\boldsymbol{\mu})v, R_{V_h}^{-1} L(\boldsymbol{\mu})v \rangle_V}{\|v\|_V \|R_{V_h}^{-1} L(\boldsymbol{\mu})v\|_V} \\ &= \inf_{v \in V_N} \frac{(R_{V_h}^{-1} L(\boldsymbol{\mu})v, R_{V_h}^{-1} L(\boldsymbol{\mu})v)_V}{\|v\|_V \|R_{V_h}^{-1} L(\boldsymbol{\mu})v\|_V} \geq \inf_{v \in V_h} \frac{\|R_{V_h}^{-1} L(\boldsymbol{\mu})v\|_V}{\|v\|_V} \\ &= \inf_{v \in V_h} \sup_{w \in V_h} \frac{v' \langle L(\boldsymbol{\mu})v, w \rangle_V}{\|v\|_V \|w\|_V} = \beta_h(\boldsymbol{\mu}) > 0, \end{aligned}$$

since (see (A.14))

$$\|R_{V_h}^{-1} L(\boldsymbol{\mu})v\|_V = \|L(\boldsymbol{\mu})v\|_{V'_h} = \sup_{w \in V_h} \frac{v' \langle L(\boldsymbol{\mu})v, w \rangle_V}{\|w\|_V}.$$

By applying Theorem 2.3 we conclude our proof. \square

Note that in this case the test subspace W_N depends on $\boldsymbol{\mu}$, i.e. $W_N = W_N^\boldsymbol{\mu}$. Note also that the operator $R_{V_h}^{-1} L(\boldsymbol{\mu})$ represents the parametrized version of the supremizer operator $T_h^\boldsymbol{\mu} : V_h \rightarrow V_h$ defined in Sect. 2.4.4; in this case (see Exercise 4),

$$T_h^\boldsymbol{\mu} v = \arg \sup_{w \in V_h} \frac{a(v, w; \boldsymbol{\mu})}{\|w\|_V}. \quad (3.39)$$

Indeed, for any $v, w \in V$,

$$(R_{V_h}^{-1} L(\boldsymbol{\mu})v, w)_V =_{V'} \langle L(\boldsymbol{\mu})v, w \rangle_V = a(v, w; \boldsymbol{\mu}) = (T_h^\boldsymbol{\mu} v, w)_V.$$

This property provides an indication on how to generate the basis functions of W_N^μ from those of V_N : for any basis function ζ_i of V_N and any $\mu \in \mathcal{P}$, a corresponding basis function is obtained by solving the following variational problem in V_h :

$$\begin{aligned} &\text{find } \eta_i(\mu) \in W_N^\mu \text{ such that} \\ &(\eta_i(\mu), z)_V = a(\zeta_i, z; \mu) \quad \forall z \in V_h. \end{aligned} \quad (3.40)$$

Therefore, W_N^μ is generated as

$$W_N^\mu = \text{span}\{\eta_i(\mu) \in V_h, i = 1, \dots, N\}. \quad (3.41)$$

The LS-RB approximation (3.36) enjoys some remarkable optimality properties, as shown in the following

Proposition 3.3. *If $a(\cdot, \cdot; \mu)$ is inf-sup stable and $W_N^\mu = R_{V_h}^{-1} L(\mu) V_N$, then the solution $u_N(\mu) \in V_N$ to (3.26) satisfies the following optimality property*

$$u_N(\mu) = \arg \min_{v \in V_N} \|L(\mu)v - f(\mu)\|_{V_h'}^2. \quad (3.42)$$

Moreover, the following best approximation property holds

$$u_N(\mu) = \arg \min_{v \in V_N} \| |u_h(\mu) - v(\mu)| \|_\mu^2, \quad (3.43)$$

where $((v, w))_\mu = (T_h^\mu v, T_h^\mu w)_V = (R_{V_h}^{-1} L(\mu)v, R_{V_h}^{-1} L(\mu)w)_V$ and $\|v\|_\mu^2 = ((v, v))_\mu$ is the induced norm.

Proof. From Theorem A.1 we know that, for u_N to be the minimizer of $\|L(\mu)v - f(\mu)\|_{V_h'}^2$ it is sufficient that

$$L(\mu)u_N(\mu) - f(\mu) \perp_{V_h'} L(\mu)v \quad \forall v \in V_N,$$

that is²

$$(L(\mu)u_N(\mu) - f(\mu), L(\mu)v)_{V_h'} = 0 \quad \forall v \in V_N. \quad (3.44)$$

Thanks to Riesz representation theorem, this is equivalent to the LS-RB approximation (3.36) written in operator form

$$\langle L(\mu)u_N(\mu) - f(\mu), R_{V_h}^{-1} L(\mu)v \rangle = 0 \quad \forall v \in V_N. \quad (3.45)$$

To prove the second part of the proposition, we use again the orthogonal projection theorem. Starting from (3.45), using the problem statement (3.12) and the Riesz

² From (3.44) we can interpret the LS-RB method as a projection method, where the projection over $L(\mu)V_N$ of the residual of the high-fidelity problem computed on the RB solution vanishes.

representation theorem, we obtain

$$(R_{V_h}^{-1}L(\boldsymbol{\mu})(u_N(\boldsymbol{\mu}) - u_h(\boldsymbol{\mu})), R_{V_h}^{-1}L(\boldsymbol{\mu})v)_V = 0 \quad \forall v \in V_N,$$

which implies $u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu}) \perp_{\boldsymbol{\mu}} V_N$; (3.43) then follows from Theorem A.1. \square

By choosing the V -norm instead of the energy norm, we would find the following optimal error inequality (thanks to Babuška theorem, see Exercise 5)

$$\|u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_V \leq \frac{\gamma_h(\boldsymbol{\mu})}{\beta_N(\boldsymbol{\mu})} \inf_{v \in V_N} \|u_h(\boldsymbol{\mu}) - v\|_V. \quad (3.46)$$

Remark 3.4. Because of the optimality property (3.42), the LS-RB method is also referred to as *minimum residual*. See, e.g., [228, 255, 44] for further details. \bullet

For the more general PG-RB problem (3.26), the well-posedness analysis can be carried out by invoking the following general theorem, which encompasses all the previous cases as special cases.

Theorem 3.2. Assume that conditions (3.6), (3.14) hold. Assume moreover that the bilinear form $a(\cdot, \cdot; \boldsymbol{\mu})$ satisfies the following inf-sup condition

$$\exists \beta_{0,N} > 0 \quad \text{s.t.} \quad \beta_N(\boldsymbol{\mu}) = \inf_{v_N \in V_N} \sup_{w_N \in W_N} \frac{a(v_N, w_N; \boldsymbol{\mu})}{\|v_N\|_V \|w_N\|_V} \geq \beta_{0,N}. \quad (3.47)$$

Then, for any $\boldsymbol{\mu} \in \mathcal{P}$ the PG-RB problem (3.27) has a unique solution $u_N(\boldsymbol{\mu}) \in V_N$, which satisfies the stability estimate

$$\|u_N(\boldsymbol{\mu})\|_V \leq \frac{1}{\beta_N(\boldsymbol{\mu})} \|f(\cdot; \boldsymbol{\mu})\|_{V'}$$

and the optimal error inequality (3.46).

3.4 Algebraic Form of Galerkin and Least-Squares RB Problems

In this section we derive the algebraic form of the RB problem (3.26), for both the G-RB and LS-RB approximations.

3.4.1 Galerkin RB Case

We first consider the Galerkin case. Inserting (3.23) into (3.27) and then choosing $v_N = \zeta_n$, $1 \leq n \leq N$, we obtain a set of N linear algebraic equations

$$\sum_{m=1}^N a(\zeta_m, \zeta_n; \boldsymbol{\mu}) u_N^{(m)}(\boldsymbol{\mu}) = f(\zeta_n; \boldsymbol{\mu}), \quad 1 \leq n \leq N. \quad (3.48)$$

We denote by $\mathbb{A}_N(\boldsymbol{\mu}) \in \mathbb{R}^{N \times N}$ the matrix with elements $(\mathbb{A}_N(\boldsymbol{\mu}))_{nm} = a(\zeta_m, \zeta_n; \boldsymbol{\mu})$ and by $\mathbf{f}_N(\boldsymbol{\mu}) \in \mathbb{R}^N$ the vector with components $(\mathbf{f}_N(\boldsymbol{\mu}))_n = f(\zeta_n; \boldsymbol{\mu})$. Then, (3.48) is equivalent to the linear system

$$\mathbb{A}_N(\boldsymbol{\mu}) \mathbf{u}_N(\boldsymbol{\mu}) = \mathbf{f}_N(\boldsymbol{\mu}). \quad (3.49)$$

Remark 3.5. The RB matrix \mathbb{A}_N inherits the properties of symmetry and positivity of \mathbb{A}_h . As of its spectral properties, if the bilinear form $a(\cdot, \cdot; \boldsymbol{\mu})$ is symmetric and coercive and the basis functions ζ_m are V-orthogonal, the spectral condition number

$$\kappa(\mathbb{A}_N(\boldsymbol{\mu})) = \frac{\lambda_{\max}(\mathbb{A}_N(\boldsymbol{\mu}))}{\lambda_{\min}(\mathbb{A}_N(\boldsymbol{\mu}))}$$

of $\mathbb{A}_N(\boldsymbol{\mu})$ can be bounded uniformly with respect to N , since

$$\kappa(\mathbb{A}_N(\boldsymbol{\mu})) \leq \frac{\gamma_h(\boldsymbol{\mu})}{\beta_h(\boldsymbol{\mu})}. \quad (3.50)$$

Instead, if the basis functions ζ_m are not V-orthogonal, we have (see Exercise 6)

$$\kappa(\mathbb{A}_N(\boldsymbol{\mu})) \leq \frac{\gamma_h(\boldsymbol{\mu})}{\beta_h(\boldsymbol{\mu})} \kappa(\mathbb{V}^T \mathbb{X}_h \mathbb{V}), \quad (3.51)$$

the matrix \mathbb{X}_h being defined in (2.40). •

Matrix \mathbb{A}_N is full, whereas the high-fidelity matrix \mathbb{A}_h is (in general) sparse. However, since typically $N \ll N_h$, (3.49) is (in principle) much faster and less computationally intensive to solve than the original high-fidelity linear system (2.38). Unfortunately, the assembly of the reduced matrix $\mathbb{A}_N(\boldsymbol{\mu})$ and vector $\mathbf{f}_N(\boldsymbol{\mu})$ still involves computations whose complexity depends on N_h .

A key ingredient to overcome this drawback is to make the *affine parametric dependence* assumption. As anticipated in Chap. 1, in this case we require both the parametric bilinear form a and the parametric linear form f to be *affine* (or *separable*) with respect to the parameter $\boldsymbol{\mu}$, that is

$$a(w, v; \boldsymbol{\mu}) = \sum_{q=1}^{Q_a} \theta_a^q(\boldsymbol{\mu}) a_q(w, v) \quad \forall v, w \in V, \boldsymbol{\mu} \in \mathcal{P}, \quad (3.52)$$

$$f(v; \boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \theta_f^q(\boldsymbol{\mu}) f_q(v) \quad \forall v \in V, \boldsymbol{\mu} \in \mathcal{P}. \quad (3.53)$$

Here $\theta_a^q: \mathcal{P} \rightarrow \mathbb{R}$, $q = 1, \dots, Q_a$ and $\theta_f^q: \mathcal{P} \rightarrow \mathbb{R}$, $q = 1, \dots, Q_f$, are $\boldsymbol{\mu}$ -dependent functions, whereas $a_q: V \times V \rightarrow \mathbb{R}$, $f_q: V \rightarrow \mathbb{R}$ are $\boldsymbol{\mu}$ -independent forms.

Remark 3.6. Often, the affine parametric dependence automatically follows by the definition of the problem. For instance, in the case of a Poisson problem defined in a domain Ω made by P blocks, each one representing a subregion with (a priori different) constant thermal conductivity μ_i , $i = 1, \dots, P$ – that is, $\bar{\Omega} = \cup_{i=1}^P \bar{\Omega}_i$, the bilinear form associated to the Laplace operator with nonhomogeneous Dirichlet and Neumann conditions (as well as internal flux continuity conditions) is given by

$$a(u, v; \mu) = \sum_{i=1}^P \mu_i \int_{\Omega_i} \nabla u \cdot \nabla v d\Omega \quad (3.54)$$

while

$$f(v; \mu) = \int_{\Gamma_N} h v d\Gamma - \sum_{i=1}^P \mu_i \int_{\Omega_i} \nabla r_g \cdot \nabla v d\Omega \quad (3.55)$$

where g and h denotes Dirichlet and Neumann data, set over Γ_D and Γ_N , respectively. Here we denote by $r_g \in H^1(\Omega)$ the lifting function such that $r_g|_{\Gamma_D} = g$. The affine parametric dependence under the form (3.52)–(3.53) is thus recovered by taking

$$\begin{aligned} a_q(u, v) &= \int_{\Omega_i} \nabla u \cdot \nabla v d\Omega, \quad \theta_a^q(\mu) = \mu_q, \quad 1 \leq q \leq P = Q_a \\ f_1(v) &= \int_{\Gamma_N} h v d\Gamma, \quad \theta_f^1(\mu) = 1, \\ f_q(v) &= - \int_{\Omega_i} \nabla r_g \cdot \nabla v d\Omega, \quad \theta_f^q(\mu) = \mu_q, \quad 2 \leq q \leq P+1 = Q_f. \end{aligned}$$

See Exercise 7 for further details. •

The affine parametric dependence is inherited by the algebraic problem, yielding the following expressions for $\mathbb{A}_N(\mu)$ and $\mathbf{f}_N(\mu)$:

$$\mathbb{A}_N(\mu) = \sum_{q=1}^{Q_a} \theta_a^q(\mu) \mathbb{A}_N^q, \quad \mathbf{f}_N(\mu) = \sum_{q=1}^{Q_f} \theta_f^q(\mu) \mathbf{f}_N^q, \quad (3.56)$$

where the parameter independent matrices \mathbb{A}_N^q and vectors \mathbf{f}_N^q are given by

$$(\mathbb{A}_N^q)_{nm} = a_q(\zeta_m, \zeta_n), \quad (\mathbf{f}_N^q)_m = f_q(\zeta_m), \quad 1 \leq m, n \leq N_h.$$

Finally, since the basis functions ζ_m belong to V_h , we can compute the RB matrices and vectors from the corresponding high-fidelity ones. Indeed, expanding each RB basis function with respect to the basis functions $\{\varphi^i\}_{i=1}^{N_h}$ of V_h ,

$$\zeta_m = \sum_{i=1}^{N_h} \zeta_m^{(i)} \varphi^i, \quad 1 \leq m \leq N, \quad (3.57)$$

we can define the *transformation matrix* $\mathbb{V} \in \mathbb{R}^{N_h \times N}$ whose columns contain the coefficients of the RB basis functions in (3.57), that is $\mathbb{V} = [\boldsymbol{\zeta}_1 \cdots \boldsymbol{\zeta}_N]$, or equivalently

$$(\mathbb{V})_{im} = \zeta_m^{(i)}, \quad 1 \leq m \leq N, \quad 1 \leq i \leq N_h. \quad (3.58)$$

It follows that, for $1 \leq n, m \leq N$,

$$a_q(\zeta_m, \zeta_n) = \sum_{i=1}^{N_h} \sum_{j=1}^{N_h} \zeta_m^{(j)} a_q(\boldsymbol{\varphi}^j, \boldsymbol{\varphi}^i) \zeta_n^{(i)}, \quad f_q(\zeta_n) = \sum_{i=1}^{N_h} f_q(\boldsymbol{\varphi}^i) \zeta_n^{(i)}.$$

Equivalently, in matrix form

$$\mathbb{A}_N^q = \mathbb{V}^T \mathbb{A}_h^q \mathbb{V}, \quad \mathbf{f}_N^q = \mathbb{V}^T \mathbf{f}_h^q, \quad (3.59)$$

where

$$(\mathbb{A}_h^q)_{ij} = a_q(\boldsymbol{\varphi}^j, \boldsymbol{\varphi}^i), \quad (\mathbf{f}_h^q)_i = f_q(\boldsymbol{\varphi}^i), \quad 1 \leq i, j \leq N.$$

3.4.2 Least-Squares RB Case

Since W_N^μ is spanned by the basis $\{\boldsymbol{\eta}_i(\boldsymbol{\mu})\}_{i=1}^N$, see (3.41), it is easily seen that problem (3.36) is equivalent to the following one: find $\mathbf{u}_N(\boldsymbol{\mu}) \in \mathbb{R}^N$ such that

$$\sum_{m=1}^N a(\zeta_m, \boldsymbol{\eta}_n(\boldsymbol{\mu}); \boldsymbol{\mu}) u_N^{(m)}(\boldsymbol{\mu}) = f(\boldsymbol{\eta}_n(\boldsymbol{\mu}); \boldsymbol{\mu}), \quad 1 \leq n \leq N. \quad (3.60)$$

This time we denote by $\mathbb{A}_N(\boldsymbol{\mu}) \in \mathbb{R}^{N \times N}$ the matrix with elements

$$(\mathbb{A}_N(\boldsymbol{\mu}))_{nm} = a(\zeta_m, \boldsymbol{\eta}_n(\boldsymbol{\mu}); \boldsymbol{\mu})$$

and by $\mathbf{f}_N(\boldsymbol{\mu}) \in \mathbb{R}^N$ the vector with components $(\mathbf{f}_N(\boldsymbol{\mu}))_m = f(\boldsymbol{\eta}_m(\boldsymbol{\mu}); \boldsymbol{\mu})$. Then, (3.48) is equivalent to the linear system

$$\mathbb{A}_N(\boldsymbol{\mu}) \mathbf{u}_N(\boldsymbol{\mu}) = \mathbf{f}_N(\boldsymbol{\mu}). \quad (3.61)$$

In order to provide an explicit expression of $\mathbb{A}_N(\boldsymbol{\mu})$, we first observe that, by the definition of the $\boldsymbol{\eta}_n(\boldsymbol{\mu})$'s, we have

$$(\mathbb{A}_N(\boldsymbol{\mu}))_{nm} = (\boldsymbol{\eta}_m(\boldsymbol{\mu}), \boldsymbol{\eta}_n(\boldsymbol{\mu}))_V. \quad (3.62)$$

Let us introduce the vector representation $\boldsymbol{\eta}_n(\boldsymbol{\mu}) \in \mathbb{R}^{N_h}$ of the functions $\boldsymbol{\eta}_n(\boldsymbol{\mu}) \in V_h$. Thanks to (3.40), $\boldsymbol{\eta}_n(\boldsymbol{\mu})$ is the solution of the following N_h -dimensional linear system (with \mathbb{X}_h defined in (2.40)),

$$\mathbb{X}_h \boldsymbol{\eta}_n = \mathbb{A}_h(\boldsymbol{\mu}) \boldsymbol{\zeta}_n \quad \text{so that} \quad \boldsymbol{\eta}_n(\boldsymbol{\mu}) = \mathbb{X}_h^{-1} \mathbb{A}_h(\boldsymbol{\mu}) \boldsymbol{\zeta}_n. \quad (3.63)$$

As a result, (3.62) becomes

$$(\mathbb{A}_N(\boldsymbol{\mu}))_{nm} = \boldsymbol{\eta}_n(\boldsymbol{\mu})^T \mathbb{X}_h \boldsymbol{\eta}_m = \boldsymbol{\zeta}_n^T(\boldsymbol{\mu}) \mathbb{A}_h^T(\boldsymbol{\mu}) \mathbb{X}_h^{-1} \mathbb{A}_h(\boldsymbol{\mu}) \boldsymbol{\zeta}_m,$$

that is

$$\mathbb{A}_N(\boldsymbol{\mu}) = \mathbb{V}^T \mathbb{A}_h^T(\boldsymbol{\mu}) \mathbb{X}_h^{-1} \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V}. \quad (3.64)$$

Similarly, we obtain the following expression for the right-hand side vector

$$\mathbf{f}_N(\boldsymbol{\mu}) = \mathbb{V}^T \mathbb{A}_h^T(\boldsymbol{\mu}) \mathbb{X}_h^{-1} \mathbf{f}_h(\boldsymbol{\mu}). \quad (3.65)$$

Remark 3.7. Equations (3.64) and (3.65) highlight that an LS-RB approximation built on top of a high-fidelity Galerkin approximation is equivalent to a G-RB approximation of the following least-squares high-fidelity problem:

$$\mathbb{A}_h^T(\boldsymbol{\mu}) \mathbb{X}_h^{-1} \mathbb{A}_h(\boldsymbol{\mu}) \mathbf{u}_h(\boldsymbol{\mu}) = \mathbb{A}_h^T(\boldsymbol{\mu}) \mathbb{X}_h^{-1} \mathbf{f}_h(\boldsymbol{\mu}). \quad (3.66)$$

•

Finally, in this case the affine parametric dependence (3.52)-(3.53) yields the following decomposition for $\mathbb{A}_N(\boldsymbol{\mu})$ and $\mathbf{f}_N(\boldsymbol{\mu})$:

$$\begin{aligned} \mathbb{A}_N(\boldsymbol{\mu}) &= \sum_{q_1=1}^{Q_a} \sum_{q_2=1}^{Q_a} \theta_a^{q_1}(\boldsymbol{\mu}) \theta_a^{q_2}(\boldsymbol{\mu}) \mathbb{A}_N^{q_1, q_2}, \\ \mathbf{f}_N(\boldsymbol{\mu}) &= \sum_{q_1=1}^{Q_f} \sum_{q_2=1}^{Q_a} \theta_f^{q_1}(\boldsymbol{\mu}) \theta_a^{q_2}(\boldsymbol{\mu}) \mathbf{f}_N^{q_1, q_2}, \end{aligned} \quad (3.67)$$

where the Q_a^2 parameter independent matrices $\mathbb{A}_N^{q_1, q_2}$ and the $Q_a Q_f$ vectors $\mathbf{f}_N^{q_1, q_2}$ are given by

$$\mathbb{A}_N^{q_1, q_2} = \mathbb{V}^T \mathbb{A}_h^{q_2 T} \mathbb{X}_h^{-1} \mathbb{A}_h^{q_1} \mathbb{V}, \quad \mathbf{f}_N^{q_1, q_2} = \mathbb{V}^T \mathbb{A}_h^{q_2 T} \mathbb{X}_h^{-1} \mathbf{f}_h^{q_1}. \quad (3.68)$$

3.5 Reduction of Computational Complexity: Offline/Online Decomposition

From the computational standpoint, we can take advantage of the affine parametric dependence property by splitting the assembly of the reduced matrices and vectors in two different phases. The former, to be performed offline once and for all, entails the computation of all the N_h -dependent ($\boldsymbol{\mu}$ -independent) structures. In the latter, to be performed online for any given value of $\boldsymbol{\mu} \in \mathcal{P}$, we assemble and solve the RB system with a cost depending only on N . In more detail:

- a) in the offline phase, we assemble and store the reduced matrices and vectors (see Algorithm 3.2). To analyze the computational complexity of these operations, let us denote by n_{vv} the number of operations required to compute a scalar product between two N_h -dimensional vectors, by n_{mv} the number of operations required for a matrix-vector product, and finally by n_{ls} the number of operations required to solve a linear system of the form $\mathbb{X}_h \mathbf{x} = \mathbf{y}$. In the G-RB case we assemble and store the Q_a matrices \mathbb{A}_N^q , for $1 \leq q \leq Q_a$, and the Q_f vectors \mathbf{f}_N^q , with

$$O(Q_f N n_{vv} + Q_a (N n_{mv} + N^2 n_{vv})) \text{ operations.}$$

In the LS-RB case we assemble and store the Q_a^2 matrices $\mathbb{A}_N^{q_1, q_2}$, and the $Q_f Q_a$ vectors $\mathbf{f}_N^{q_1, q_2}$, with

$$O(Q_a (N n_{mv} + N n_{ls}) + Q_f Q_a N n_{vv} + Q_a^2 (N n_{mv} + N^2 n_{vv})) \text{ operations.}$$

- Note that the LS-RB method, in addition of being more expensive than the G-RB, requires more storage, since Q_a^2 (rather than Q_a) RB matrices need to be saved;
- b) in the online phase, given $\boldsymbol{\mu} \in \mathcal{P}$, we first form the RB matrix $\mathbb{A}_N(\boldsymbol{\mu})$ and vector $\mathbf{f}_N(\boldsymbol{\mu})$ by computing the sum (3.56) in the G-RB case, (3.67) in the LS-RB case. This requires $O(Q_a N^2 + Q_f N)$ operations in the former case, $O(Q_a^2 N^2 + Q_f Q_a N)$ in the latter, so that the G-RB method is always faster than the LS-RB (see also Algorithm 3.3). Then, in both cases we solve the dense, N -dimensional RB system with complexity $O(N^3)$. As a result, if N and Q_a are small enough, we can achieve very fast response both for real-time problems and many-query contexts.

3.6 A Posteriori Error Estimation

A posteriori error estimators are computable indicators which employ the residual of the approximate RB solution to derive estimates of the actual solution error.

In the context of high-fidelity techniques, these indicators are largely employed for instance to steer adaptive schemes where either the mesh is locally refined (h -version) or the polynomial degree is increased (p -method), or a combination of both. For more details see, e.g., [3, 253, 192, 118] and references therein.

In the context of RB methods, a posteriori error estimators play an essential role at two distinct stages: they not only guarantee the *reliability* of the reduction process, but also its *efficiency*. Regarding the former aspect, in the online stage the a posteriori estimator allows to bound (from above) the error between the RB solution $u_N(\boldsymbol{\mu})$ and the underlying high-fidelity solution $u_h(\boldsymbol{\mu})$, for each new value $\boldsymbol{\mu} \in \mathcal{P}$.

In this context, an error estimator is required to be:

- *sharp*, that is, it should be as close as possible to the actual (unknown) error;
- *asymptotically correct*, that is when increasing the dimension N of the RB space, the error estimate should tend to zero with the same rate as the actual error;

Algorithm 3.2 Offline computation of reduced matrices and vectors

```

1: function  $[\mathbb{A}_N^q, \mathbf{f}_N^q] = \text{PROJECTSYSTEM}(\mathbb{A}_h^q, \mathbf{f}_h^q, \mathbb{V}, \mathbb{X}, \text{method})$ 
2:   switch method
3:     case G-RB
4:       for  $q = 1 : Q_a$ 
5:          $\mathbb{A}_N^q = \mathbb{V}^T \mathbb{A}_h^q \mathbb{V}$   $\triangleright O(Nn_{mv} + N^2n_{vv})$ 
6:       end for
7:       for  $q = 1 : Q_f$ 
8:          $\mathbf{f}_N^q = \mathbb{V}^T \mathbf{f}_h^q$   $\triangleright O(Nn_{vv})$ 
9:       end for
10:      case LS-RB
11:        for  $q_1 = 1 : Q_a$ 
12:          compute  $\mathbb{Z} = \mathbb{X}_h^{-1} \mathbb{A}_h^{q_1} \mathbb{V}$   $\triangleright O(Nn_{mv} + Nn_{ls})$ 
13:          for  $q_2 = 1 : Q_a$ 
14:             $\mathbb{A}_N^{q_1, q_2} = \mathbb{Z}^T \mathbb{A}_h^{q_2} \mathbb{V}$   $\triangleright O(Nn_{mv} + N^2n_{vv})$ 
15:          end for
16:          for  $q_2 = 1 : Q_f$ 
17:             $\mathbf{f}_N^{q_1, q_2} = \mathbb{Z}^T \mathbf{f}_h^{q_2}$   $\triangleright O(Nn_{vv})$ 
18:          end for
19:        end for
20:      end switch
21: end function

```

- *computationally cheap*, i.e it should be inexpensive to compute with respect to the total computational costs entailed by the solution of the RB problem.

To derive such an estimator, an equivalence will be established between the norm of the error and a corresponding dual norm of the residual. The latter only involves the arrays of the high-fidelity problem (i.e. system matrix and vector) together with the computed RB solution, but not the high-fidelity solution. This equivalence is a direct consequence of the stability of the high-fidelity problem.

3.6.1 A Relationship between Error and Residual

Establishing an error-residual relationship is crucial to derive a posteriori error estimates. Denote by $e_h(\boldsymbol{\mu}) = u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu}) \in V_h$ the error between the high-fidelity and reduced solutions. From (3.11) and (3.27) we get the error representation

$$a(e_h(\boldsymbol{\mu}), v) = f(v; \boldsymbol{\mu}) - a(u_N(\boldsymbol{\mu}), v; \boldsymbol{\mu}) \quad \forall v \in V_h. \quad (3.69)$$

Algorithm 3.3 Online RB system assembling and solving

```

1: function  $\mathbf{u}_N(\boldsymbol{\mu}) = \text{SOLVERBSYSTEM}(\mathbb{A}_N^q, \mathbf{f}_N^q, \boldsymbol{\theta}_a^q, \boldsymbol{\theta}_f^q, \boldsymbol{\mu}, \text{method})$ 
2:   switch method
3:     case G-RB
4:        $\mathbb{A}_N(\boldsymbol{\mu}) = \mathbb{O}, \mathbf{f}_N(\boldsymbol{\mu}) = \mathbf{0}$ 
5:       for  $q = 1 : Q_a$ 
6:          $\mathbb{A}_N(\boldsymbol{\mu}) \leftarrow \mathbb{A}_N(\boldsymbol{\mu}) + \boldsymbol{\theta}_a^q(\boldsymbol{\mu}) \mathbb{A}_N^q$   $\triangleright O(N^2)$ 
7:       end for
8:       for  $q = 1 : Q_f$ 
9:          $\mathbf{f}_N(\boldsymbol{\mu}) \leftarrow \mathbf{f}_N(\boldsymbol{\mu}) + \boldsymbol{\theta}_f^q(\boldsymbol{\mu}) \mathbf{f}_N^q$   $\triangleright O(N)$ 
10:      end for
11:     case LS-RB
12:        $\mathbb{A}_N(\boldsymbol{\mu}) = \mathbb{O}, \mathbf{f}_N(\boldsymbol{\mu}) = \mathbf{0}$ 
13:       for  $q_1 = 1 : Q_a$ 
14:         for  $q_2 = 1 : Q_a$ 
15:            $\mathbb{A}_N(\boldsymbol{\mu}) \leftarrow \mathbb{A}_N(\boldsymbol{\mu}) + \boldsymbol{\theta}_a^{q_1}(\boldsymbol{\mu}) \boldsymbol{\theta}_a^{q_2}(\boldsymbol{\mu}) \mathbb{A}_N^{q_1, q_2}$   $\triangleright O(N^2)$ 
16:         end for
17:       end for
18:       for  $q_1 = 1 : Q_a$ 
19:         for  $q_2 = 1 : Q_f$ 
20:            $\mathbf{f}_N(\boldsymbol{\mu}) \leftarrow \mathbf{f}_N(\boldsymbol{\mu}) + \boldsymbol{\theta}_f^{q_1}(\boldsymbol{\mu}) \boldsymbol{\theta}_a^{q_2}(\boldsymbol{\mu}) \mathbf{f}_N^{q_1, q_2}$   $\triangleright O(N)$ 
21:         end for
22:       end for
23:     end switch
24:     solve linear system  $\mathbb{A}_N(\boldsymbol{\mu}) \mathbf{u}_N(\boldsymbol{\mu}) = \mathbf{f}_N(\boldsymbol{\mu})$   $\triangleright O(N^3)$ 
25: end function

```

By setting

$$r(v; \boldsymbol{\mu}) = f(v; \boldsymbol{\mu}) - a(u_N(\boldsymbol{\mu}), v; \boldsymbol{\mu}) \quad \forall v \in V_h \quad (3.70)$$

we note that

$$\langle r(\boldsymbol{\mu}), v \rangle = r(v; \boldsymbol{\mu})$$

where $r(\boldsymbol{\mu}) \in V'_h$ is the residual of the high-fidelity problem computed on the RB solution, introduced in (3.24). Then, thanks to the continuity of $a(\cdot, \cdot; \boldsymbol{\mu})$ we get

$$|r(v; \boldsymbol{\mu})| \leq \gamma_h(\boldsymbol{\mu}) \|e_h(\boldsymbol{\mu})\|_V \|v\|_V \quad \forall v \in V_h.$$

This estimate and the definition of dual norm imply that

$$\|r(\cdot; \boldsymbol{\mu})\|_{V'_h} \leq \gamma_h(\boldsymbol{\mu}) \|e_h(\boldsymbol{\mu})\|_V.$$

On the other hand, from the error representation and the stability estimate (3.10) we obtain

$$\beta_h(\boldsymbol{\mu}) \|e_h(\boldsymbol{\mu})\|_V \leq \|r(\cdot; \boldsymbol{\mu})\|_{V'_h}.$$

In conclusion

$$\frac{1}{\gamma_h(\boldsymbol{\mu})} \|r(\cdot; \boldsymbol{\mu})\|_{V'_h} \leq \|e_h(\boldsymbol{\mu})\|_V \leq \frac{1}{\beta_h(\boldsymbol{\mu})} \|r(\cdot; \boldsymbol{\mu})\|_{V'_h} \quad (3.71)$$

i.e. the norm of the error is bounded from below and from above by the dual norm of the residual. Since $r(\cdot; \boldsymbol{\mu})$ only involves the high-fidelity arrays and the computed reduced solution $u_N(\boldsymbol{\mu})$, but not $u_h(\boldsymbol{\mu})$, its norm well serves as an a posteriori error estimator.

3.6.2 Error Bound

The quantity

$$\Delta_N(\boldsymbol{\mu}) = \frac{\|r(\cdot; \boldsymbol{\mu})\|_{V'_h}}{\beta_h(\boldsymbol{\mu})} \quad (3.72)$$

can play the role of *error estimator* thanks to the bound (3.71). The associated *effectivity factor* (or, simply, the effectivity) is given by

$$\eta_N(\boldsymbol{\mu}) = \frac{\Delta_N(\boldsymbol{\mu})}{\|e_h(\boldsymbol{\mu})\|_V}.$$

The latter is a measure of the quality of the proposed estimator: for sharpness, we pretend it to be as close to 1 as possible.

Equivalence (3.71) directly implies that the effectivity factor satisfies

$$1 \leq \eta_N(\boldsymbol{\mu}) \leq \frac{\gamma_h(\boldsymbol{\mu})}{\beta_h(\boldsymbol{\mu})} \quad \forall \boldsymbol{\mu} \in \mathcal{P}. \quad (3.73)$$

Since the stability factors $\beta_h(\boldsymbol{\mu})$ and $\gamma_h(\boldsymbol{\mu})$ are the minimum and maximum (generalized) singular values of the operator matrix $\mathbb{A}_h(\boldsymbol{\mu})$, the effectivity upper bound $\kappa_h(\boldsymbol{\mu}) = \gamma_h(\boldsymbol{\mu})/\beta_h(\boldsymbol{\mu})$ is in fact the condition number of the high-fidelity problem, that is it measures the sensitivity of the latter with respect to small perturbations.

Therefore, the effectivity upper bound (3.73) is independent of N , and hence stable with respect to *N-refinement*. At the same time however, independently of the reduced approximation, we can expect large effectivities when the underlying high-fidelity problem is ill-conditioned.

Remark 3.8. A similar result on the a posteriori error estimation for a linear output can be obtained in the so-called compliant case – that is, when $a(\cdot, \cdot; \boldsymbol{\mu})$ is a symmetric bilinear form and $z(\boldsymbol{\mu}) = f(u(\boldsymbol{\mu}))$; see Exercise 8. The generalization to the case of non-compliant outputs is also possible, see, e.g., [219, 231] for further details. •

3.7 Practical (and Efficient) Computation of Error Bounds

In order to obtain an algebraic equivalent of the error bound (3.72), we start by deriving the algebraic counterpart of the error representation (3.69). To this end, let us define the discrete error between the the RB and high-fidelity solutions

$$\mathbf{e}_h(\boldsymbol{\mu}) = \mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}),$$

and the discrete residual (the algebraic counterpart of (3.70))

$$\mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu}) = \mathbf{f}_h(\boldsymbol{\mu}) - \mathbb{A}_h(\boldsymbol{\mu})\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}).$$

Recalling that $\mathbb{A}_h(\boldsymbol{\mu})\mathbf{u}_h(\boldsymbol{\mu}) = \mathbf{f}_h(\boldsymbol{\mu})$, we obtain the following algebraic equivalent of the error representation

$$\mathbb{A}_h(\boldsymbol{\mu})\mathbf{e}_h(\boldsymbol{\mu}) = \mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu}). \quad (3.74)$$

Since the matrix $\mathbb{A}_h(\boldsymbol{\mu})$ is non-singular,

$$\mathbf{e}_h(\boldsymbol{\mu}) = \mathbb{A}_h^{-1}(\boldsymbol{\mu})\mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu}). \quad (3.75)$$

Then, taking the 2-norm on both sides and by the properties of the matrix 2-norm (see e.g. [221, Chap. 1]) we obtain the following upper bound

$$\|\mathbf{e}_h(\boldsymbol{\mu})\|_2 \leq \|\mathbb{A}_h^{-1}(\boldsymbol{\mu})\|_2 \|\mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu})\|_2 = \frac{1}{\sigma_{\min}(\mathbb{A}_h(\boldsymbol{\mu}))} \|\mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu})\|_2$$

for the 2-norm of the error, where $\sigma_{\min}(\mathbb{A}_h(\boldsymbol{\mu}))$ denotes the smallest singular value of $\mathbb{A}_h(\boldsymbol{\mu})$.

Similarly, we can derive a bound for the error in the V -norm. We first left multiply (3.75) by $\mathbb{X}_h^{1/2}$ and use $\mathbb{I} = \mathbb{X}_h^{1/2}\mathbb{X}_h^{-1/2}$ to get

$$\mathbb{X}_h^{1/2}\mathbf{e}_h(\boldsymbol{\mu}) = \mathbb{X}_h^{1/2}\mathbb{A}_h^{-1}(\boldsymbol{\mu})\mathbb{X}_h^{1/2}\mathbb{X}_h^{-1/2}\mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu}).$$

Then we proceed as before to obtain

$$\|\mathbf{e}_h(\boldsymbol{\mu})\|_{\mathbb{X}_h} \leq \|\mathbb{X}_h^{1/2}\mathbb{A}_h^{-1}(\boldsymbol{\mu})\mathbb{X}_h^{1/2}\|_2 \|\mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu})\|_{\mathbb{X}_h^{-1}},$$

that is

$$\|\mathbf{e}_h(\boldsymbol{\mu})\|_{\mathbb{X}_h} \leq \frac{1}{\sigma_{\min}(\mathbb{X}_h^{-1/2} \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{X}_h^{-1/2})} \|\mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu})\|_{\mathbb{X}_h^{-1}} \quad (3.76)$$

where we recall from Sect. 2.4.6 that

$$\sigma_{\min}(\mathbb{X}_h^{-1/2} \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{X}_h^{-1/2}) = \sqrt{\lambda_{\min}(\mathbb{X}_h^{-1/2} \mathbb{A}_h(\boldsymbol{\mu})^T \mathbb{X}_h^{-1} \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{X}_h^{-1/2})} = \beta_h(\boldsymbol{\mu}).$$

The right-hand side of (3.76) provides the algebraic form of the error bound $\Delta_N(\boldsymbol{\mu})$ defined in (3.72). We remark that the definition (3.72) of the error bound $\Delta_N(\boldsymbol{\mu})$ only depends on the basis of V_N , i.e. its definition does not depend on whether we use Galerkin or least-squares projections.

3.7.1 Computing the Norm of the Residual

For an efficient use of the error bound established in the previous section we take advantage of a suitable offline-online computational splitting. To evaluate the dual norm of the residual, we exploit the affine decomposition (3.56).

By the definition of \mathbb{X}_h -scalar product,

$$\begin{aligned} \|\mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu})\|_{\mathbb{X}_h^{-1}}^2 &= \mathbf{f}_h(\boldsymbol{\mu})^T \mathbb{X}_h^{-1} \mathbf{f}_h(\boldsymbol{\mu}) - 2\mathbf{f}_h(\boldsymbol{\mu})^T \mathbb{X}_h^{-1} \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}) \\ &\quad + \mathbf{u}_N^T(\boldsymbol{\mu}) \mathbb{V}^T \mathbb{A}_h^T(\boldsymbol{\mu}) \mathbb{X}_h^{-1} \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}). \end{aligned}$$

Then, by inserting expressions (3.56) we get

$$\begin{aligned} \|\mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu})\|_{\mathbb{X}_h^{-1}}^2 &= \sum_{q_1, q_2=1}^{Q_f} \theta_{q_1}^f(\boldsymbol{\mu}) \theta_{q_2}^f(\boldsymbol{\mu}) \underbrace{\mathbf{f}_h^{q_1 T} \mathbb{X}_h^{-1} \mathbf{f}_h^{q_2}}_{C_{q_1, q_2}} \\ &\quad - 2 \sum_{q_1=1}^{Q_a} \sum_{q_2=1}^{Q_f} \theta_{q_2}^f(\boldsymbol{\mu}) \theta_{q_1}^a(\boldsymbol{\mu}) \mathbf{u}_N(\boldsymbol{\mu})^T \underbrace{\mathbb{V}^T \mathbb{A}_h^{q_1 T} \mathbb{X}_h^{-1} \mathbf{f}_h^{q_2}}_{\mathbf{d}_{q_1, q_2}} \\ &\quad + \sum_{q_1, q_2=1}^{Q_a} \theta_{q_1}^a(\boldsymbol{\mu}) \theta_{q_2}^a(\boldsymbol{\mu}) \mathbf{u}_N(\boldsymbol{\mu})^T \underbrace{\mathbb{V}^T \mathbb{A}_h^{q_1 T} \mathbb{X}_h^{-1} \mathbb{A}_h^{q_2} \mathbb{V}}_{\mathbb{E}_{q_1, q_2}} \mathbf{u}_N(\boldsymbol{\mu}). \end{aligned} \quad (3.77)$$

The $\boldsymbol{\mu}$ -independent quantities $C_{q_1, q_2} \in \mathbb{R}$, $\mathbf{d}_{q_1, q_2} \in \mathbb{R}^N$ and $\mathbb{E}_{q_1, q_2} \in \mathbb{R}^{N \times N}$ can be pre-computed and stored offline (see Algorithm 3.4), while, for any new value of $\boldsymbol{\mu}$, only the parameter-dependent quantities need to be evaluated online.

Algorithm 3.4 Offline computation of $\boldsymbol{\mu}$ -independent terms of the dual norm of residual (3.77)

```

1: function  $[C_{q_1,q_2}, \mathbf{d}_{q_1,q_2}, \mathbb{E}_{q_1,q_2}] = \text{OFFLINERESIDUAL}(\mathbb{A}_h^q, \mathbf{f}_h^q, \mathbb{X}_h, \mathbb{V})$ 
2:   for  $q_1 = 1 : Q_f$ 
3:      $\mathbf{t} = \mathbb{X}_h^{-1} \mathbf{f}_h^{q_1}$   $\triangleright O(n_{ls})$ 
4:     for  $q_2 = 1 : Q_f$ 
5:        $C_{q_1,q_2} = \mathbf{t}^T \mathbf{f}_h^{q_2}$   $\triangleright O(n_{vv})$ 
6:     end for
7:   end for
8:   for  $q_1 = 1 : Q_a$ 
9:      $\mathbb{Z} = \mathbb{X}_h^{-1} \mathbb{A}_h^{q_1} \mathbb{V}$   $\triangleright O(Nn_{mv} + Nn_{ls})$ 
10:    for  $q_2 = 1 : Q_a$ 
11:       $\mathbb{E}_{q_1,q_2} = \mathbb{Z}^T \mathbb{A}_h^{q_2} \mathbb{V}$   $\triangleright O(Nn_{mv} + N^2 n_{vv})$ 
12:    end for
13:    for  $q_2 = 1 : Q_f$ 
14:       $\mathbf{d}_{q_1,q_2} = \mathbb{Z}^T \mathbf{f}_h^{q_2}$   $\triangleright O(Nn_{vv})$ 
15:    end for
16:  end for
17: end function

```

Note that in the LS-RB case only lines 2-7 have to be executed, since $\mathbb{E}_{q_1,q_2} = \mathbb{A}_N^{q_1,q_2}$ and $\mathbf{d}_{q_1,q_2} = \mathbf{f}_N^{q_1,q_2}$. The offline computational complexity scales as

$$O(Q_f n_{ls} + Q_f^2 n_{vv} + Q_a N(n_{ls} + n_{mv}) + Q_f Q_a n_{vv} N + Q_a^2 (N n_{mv} + N^2 n_{vv})),$$

while the online operation count yields

$$O(Q_f^2 + Q_f Q_a N + Q_a^2 N^2).$$

3.7.2 Computing the Stability Factor by the Successive Constraint Method

For any given $\boldsymbol{\mu} \in \mathcal{P}$, the computation of the stability factor

$$\beta_h(\boldsymbol{\mu}) = \sigma_{\min}(\mathbb{X}_h^{-1/2} \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{X}_h^{-1/2}) \quad (3.78)$$

requires to solve the generalized eigenvalue problem (2.57) introduced in Sect. 2.4.6. This would require $O(N_h^\alpha)$ operations, with $\alpha \in [1, 3]$, an unaffordable task in a real-time context. Different strategies have been developed to get rid of this N_h -dependence and enable a fast evaluation of the error bound.

One strategy consists in computing a parameter-dependent lower bound $\beta_h^{\text{LB}}(\boldsymbol{\mu})$ to $\beta_h(\boldsymbol{\mu})$ by means of the *successive constraint method* (SCM). The latter is an iterative procedure based on the successive solution of suitable linear optimization problems, which was introduced in [140] to deal with coercive problems; see also

[59, 60]. An improved version – suitable for more general weakly coercive problems – using the so-called *natural norm* [241] has been analyzed in [143], while a recent application to Stokes equations is given in [230].

For both strongly and weakly coercive problems, the SCM – based on a suitable offline-online strategy – returns a lower bound $\beta_h^{\text{LB}} : \mathcal{P} \rightarrow \mathbb{R}$ such that

$$0 < \beta_h^{\text{LB}}(\boldsymbol{\mu}) \leq \beta_h(\boldsymbol{\mu}) \quad \forall \boldsymbol{\mu} \in \mathcal{P}, \quad (3.79)$$

whose online evaluation requires the solution of a small linear program with computational complexity *independent* of N_h ; see for instance [140, 143] for more details. The resulting error bound (3.72) is thus efficiently computable and rigorous, thanks to (3.79). However, the offline-online strategy developed to build the lower bound is applicable only in case of affine parametric dependence.

The SCM algorithm is rather involved to implement and requires a considerable computational effort. The precise computational cost will be carried out at the end of this section. For the sake of simplicity, we present the algorithm in the case of a coercive problem, following [140, 231], and still denoting by $\beta_h^{\text{LB}}(\boldsymbol{\mu})$ the lower bound to a parametrized coercivity factor $\beta_h(\boldsymbol{\mu})$.

We first introduce an objective function $J^{\text{obj}} : \mathcal{P} \times \mathbb{R}^{Q_a} \rightarrow \mathbb{R}$ given by

$$J^{\text{obj}}(\boldsymbol{\mu}; \mathbf{y}) = \sum_{q=1}^{Q_a} \theta_a^q(\boldsymbol{\mu}) y_q, \quad (3.80)$$

where $\mathbf{y} = (y_1, \dots, y_{Q_a})^T \in \mathbb{R}^{Q_a}$. Thanks to the affine dependence assumption, the coercivity factor may be expressed as

$$\beta_h(\boldsymbol{\mu}) = \inf_{\mathbf{y} \in Y} J^{\text{obj}}(\boldsymbol{\mu}; \mathbf{y}), \quad (3.81)$$

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

$$Y = \left\{ \mathbf{y} \in \mathbb{R}^{Q_a} \mid \exists w_y \in V_h \text{ s.t. } y_q = \frac{a_q(w_y, w_y)}{\|w_y\|_V^2}, 1 \leq q \leq Q_a \right\}.$$

We next introduce the continuity constraint box

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

that is bounded provided that the forms $a_q(\cdot, \cdot)$ are continuous, for any $q = 1, \dots, Q_a$. Finally, we define the coercivity constraint sample

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

We are now ready to construct the lower bound.

For given C_J , and any $\boldsymbol{\mu} \in \mathcal{P}$, we define the lower bound set

$$Y_{\text{LB}}(\boldsymbol{\mu}; C_J) = \left\{ \mathbf{y} \in B : \sum_{q=1}^{Q_a} \theta_a^q(\boldsymbol{\mu}') y_q \geq \beta_h(\boldsymbol{\mu}') \quad \forall \boldsymbol{\mu}' \in C_J \right\}.$$

Since $Y \subset Y_{\text{LB}}(\boldsymbol{\mu}; C_J)$, we can define the lower bound as

$$\beta_h^{\text{LB}}(\boldsymbol{\mu}) = \inf_{\mathbf{y} \in Y_{\text{LB}}(\boldsymbol{\mu}; C_J)} J^{\text{obj}}(\boldsymbol{\mu}; \mathbf{y}). \quad (3.82)$$

Hence, for given $C_J \subset \mathcal{P}$, (3.79) readily follows, see [140, 231]. The lower bound (3.82) is in fact the solution of a linear optimization problem (or Linear Program (LP)) – that is, a problem consisting in the minimization (or maximization) of a linear objective function, subject to linear equality and linear inequality constraints – with Q_a design variables and $2Q_a + J$ inequality constraints. Moreover, given B and the set $\{\beta_h(\boldsymbol{\mu}') | \boldsymbol{\mu}' \in C_J\}$, the operation count to evaluate $\boldsymbol{\mu} \rightarrow \beta_h^{\text{LB}}(\boldsymbol{\mu})$ is N_h -independent.

In order to construct a suitable coercivity constraint sample C_J , we also require an upper bound for the coercivity constant. For given C_J and any $\boldsymbol{\mu} \in \mathcal{P}$, we introduce the upper bound set $Y_{\text{UB}}(\boldsymbol{\mu}; C_J) \subset \mathbb{R}^{Q_a}$ as

$$Y_{\text{UB}}(\boldsymbol{\mu}; C_J) = \{\mathbf{y}^*(\boldsymbol{\mu}') | \boldsymbol{\mu}' \in C_J\}, \quad (3.83)$$

where

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

We define the upper bound as

$$\beta_h^{\text{UB}}(\boldsymbol{\mu}) = \min_{\mathbf{y} \in Y_{\text{UB}}(\boldsymbol{\mu}; C_J)} J^{\text{obj}}(\boldsymbol{\mu}; \mathbf{y}). \quad (3.84)$$

Since $Y_{\text{UB}}(\boldsymbol{\mu}; C_J) \subset Y$, for given $C_J \subset \mathcal{P}$, we have that $\beta_h^{\text{UB}}(\boldsymbol{\mu}) \geq \beta_h(\boldsymbol{\mu})$ for any $\boldsymbol{\mu} \in \mathcal{P}$. Note that, given the set $\{\mathbf{y}^*(\boldsymbol{\mu}') | \boldsymbol{\mu}' \in C_J\}$, the operation count for the online stage to evaluate $\boldsymbol{\mu} \rightarrow \beta_h^{\text{UB}}(\boldsymbol{\mu})$ is independent of N_h .

We now present a greedy algorithm (to be performed offline) for the construction of the set C_J . We require a *training* sample

$$\Xi_{\text{train}}^{\text{SCM}} = \{\boldsymbol{\mu}_{\text{train}, \text{SCM}}^1, \dots, \boldsymbol{\mu}_{\text{train}, \text{SCM}}^{n_{\text{train}}}\} \subset \mathcal{P},$$

of n_{train} parameters point, and a tolerance ε_{SCM} which shall control the error in the lower bound prediction. The greedy procedure is given in Algorithm 3.5.

As already mentioned, the choice of the stopping criterion ε_{SCM} permits to bound the ratio between the coercivity constant and corresponding lower bound as [231]:

$$\frac{\beta_h(\boldsymbol{\mu})}{\beta_h^{\text{LB}}(\boldsymbol{\mu})} \leq \frac{1}{1 - \varepsilon_{\text{SCM}}}, \quad \forall \boldsymbol{\mu} \in \Xi_{\text{train}}^{\text{SCM}}.$$

Algorithm 3.5 SCM algorithm

Input: $\Xi_{\text{train}}^{\text{SCM}} \subset \mathcal{P}$, tolerance $\varepsilon_{\text{SCM}} \in (0, 1)$, $\mu_{\text{SCM}}^1 \in \mathcal{P}$

- 1: set $J = 1$ and $C_1 = \{\mu_{\text{SCM}}^1\}$
- 2: compute $\eta_J(\mu) = \frac{\beta_h^{\text{UB}}(\mu) - \beta_h^{\text{LB}}(\mu)}{\beta_h^{\text{UB}}(\mu)}$
- 3: **while** $\max_{\mu \in \Xi_{\text{train}}^{\text{SCM}}} \eta_J(\mu) > \varepsilon_{\text{SCM}}$
- 4: $\mu_{\text{SCM}}^{J+1} = \arg \max_{\mu \in \Xi_{\text{train}}^{\text{SCM}}} \eta_J(\mu)$
- 5: $C_{J+1} = C_J \cup \mu_{\text{SCM}}^{J+1}$
- 6: $J \leftarrow J + 1$
- 7: $\eta_J(\mu) = \frac{\beta_h^{\text{UB}}(\mu) - \beta_h^{\text{LB}}(\mu)}{\beta_h^{\text{UB}}(\mu)}$
- 8: **end while**
- 9: set $J_{\text{max}} = J$

We briefly summarize the offline and online computational costs:

1. in the offline stage we have to solve $2Q_a$ eigenproblems over V_h to form B and J_{max} eigenproblems over V_h to form $\{\beta_h(\mu') \mid \mu' \in C_{J_{\text{max}}}\}$, to compute $J_{\text{max}}Q_a$ inner products over V_h to form $\{y^*(\mu') \mid \mu' \in C_{J_{\text{max}}}\}$ and finally to solve $n_{\text{train}}J_{\text{max}}$ linear programs of size $Q_a + J$;
2. in the online stage, given a new value of μ , we have to solve a linear program of size $Q_a + J$ to evaluate $\beta_h^{\text{LB}}(\mu)$. The online operation count is thus N_h -independent.

3.7.3 Computing the Stability Factor by Interpolatory Radial Basis Functions

An alternative strategy – targeted to computational efficiency – relies instead on computing an interpolatory approximation $\beta_I(\mu)$ of $\beta_h(\mu)$. Let us denote by $\Xi_{\text{fine}} \subset \mathcal{P}$ a sample set whose dimension $n_{\text{fine}} = |\Xi_{\text{fine}}|$ is sufficiently large. We select a set of parameters (called interpolation points) $\Xi_I = \{\mu^j\}_{j=1}^{n_I} \subset \Xi_{\text{fine}}$ and compute the stability factor $\beta_h(\mu)$ for each $\mu \in \Xi_I$. Then, we compute a suitable interpolant $\beta_I(\mu)$ such that

$$\beta_I(\mu) = \beta_h(\mu) \quad \forall \mu \in \Xi_I \quad \text{and} \quad \beta_I(\mu) > 0 \quad \forall \mu \in \Xi_{\text{fine}}.$$

Depending on the number of parameters and their range of variation, different interpolation methods can be employed. In [196] a simple linear interpolant and an equally spaced grid of interpolation points were used, thanks to the fact that the parameter space was simply two dimensional.

When the parameter space has higher dimension, using uniform grids immediately suffers the *curse of dimensionality*, because of the need to compute $\beta_h(\mu)$ in

Algorithm 3.6 Construction of a positive RBF interpolant of the stability factor**Input:** fine grid Ξ_{fine} , a set Ξ_I of n_I samples

- 1: **for** $j = 1 : n_I$
- 2: set $\boldsymbol{\mu}^j = \Xi_I(j)$ and assemble $\mathbb{A}(\boldsymbol{\mu}^j)$
- 3: compute $\beta_h(\boldsymbol{\mu}^j) = \sigma_{\min}(\mathbb{X}_h^{-1/2} \mathbb{A}_h(\boldsymbol{\mu}^j) \mathbb{X}_h^{-1/2})$
- 4: **end for**
- 5: build the RBF interpolant $\beta_I(\boldsymbol{\mu})$ by solving (3.87)

a huge number of interpolation points. In [181, 183] an interpolatory *radial basis function* (RBF) technique was used, thanks to its suitability to interpolate scattered data in high-dimensional spaces. In order to guarantee the positivity of the interpolant, we interpolate the logarithm of $\beta_h(\boldsymbol{\mu})$ rather than $\beta_h(\boldsymbol{\mu})$ itself. We then define the RBF interpolant of $\log \beta_h(\boldsymbol{\mu})$ as

$$\log \beta_I(\boldsymbol{\mu}) = \omega_0 + \boldsymbol{\omega}^T \boldsymbol{\mu} + \sum_{j=1}^{n_I} \gamma_j \phi(|\boldsymbol{\mu} - \boldsymbol{\mu}^j|), \quad (3.85)$$

where ϕ is a radial basis function³, while the $1 + p + n_I$ interpolation weights $\{\omega_i\}_{i=0}^p, \{\gamma_j\}_{j=1}^{n_I}$ are determined by requiring the following conditions to hold:

$$\log \beta_I(\boldsymbol{\mu}^j) = \log \beta_h(\boldsymbol{\mu}^j), \quad j = 1, \dots, n_I, \quad (3.86a)$$

$$\sum_{j=1}^{n_I} \gamma_j = 0, \quad \sum_{j=1}^{n_I} \gamma_j \boldsymbol{\mu}_p^j = 0, \quad p = 1, \dots, P. \quad (3.86b)$$

Equations (3.86a)-(3.86b) lead to the following symmetric linear system of dimension $1 + P + n_I$

$$\begin{pmatrix} \mathbb{M} & \mathbb{P}^T & \mathbf{1}^T \\ \mathbb{P} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\omega} \\ \omega_0 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{0} \\ 0 \end{pmatrix} \quad (3.87)$$

where $\mathbf{1} = [1, \dots, 1] \in \mathbb{R}^{n_I}$, $\boldsymbol{\beta} = [\log \beta_h(\boldsymbol{\mu}^1), \dots, \log \beta_h(\boldsymbol{\mu}^{n_I})] \in \mathbb{R}^{n_I}$ and

$$(\mathbb{M})_{ij} = \phi(|\boldsymbol{\mu}^i - \boldsymbol{\mu}^j|), \quad (\mathbb{P})_{pj} = \boldsymbol{\mu}_p^j, \quad i, j = 1, \dots, n_I, \quad p = 1, \dots, P.$$

System (3.87) is solved in the offline phase to yield the interpolation weights $\omega_0, \boldsymbol{\omega}, \boldsymbol{\gamma}$. The output of this procedure (reported in Algorithm 3.6) is a positive interpolant $\beta_I(\boldsymbol{\mu})$ which interpolates $\beta_h(\boldsymbol{\mu})$ in each $\boldsymbol{\mu} \in \Xi_I$, and whose online evaluation for a given $\boldsymbol{\mu}$ requires only $O(n_I)$ operations – a number independent of N_h . For further details and improvements based on a suitable adaptive strategy for the selection of the interpolation points, we refer to [183].

The entire procedure for the online evaluation of the error estimate $\Delta_N(\boldsymbol{\mu})$ based on RBF interpolation is summarized in Algorithm 3.7.

³ Examples of commonly used radial basis functions are the Gaussian $\phi(r) = e^{-r^2}$ and thin plate splines $\phi(r) = r^2 \log(r)$. For a general introduction to RBF methods see, e.g., [41].

Algorithm 3.7 Evaluation of the error estimator $\Delta_N(\boldsymbol{\mu})$

```

1: function  $\Delta_N(\boldsymbol{\mu}) = \text{ERROR\_ESTIMATE}(C_{q_1, q_2}, \mathbf{d}_{q_1, q_2}, \mathbb{E}_{q_1, q_2}, \theta_a^q, \theta_f^q, \mathbf{u}_N(\boldsymbol{\mu}), \boldsymbol{\mu})$ 
2:   compute  $\beta_I(\boldsymbol{\mu})$  by evaluating (3.85)
3:   set  $\varepsilon = 0$ 
4:   for  $q_1 = 1 : Q_f$ 
5:     for  $q_2 = 1 : Q_f$ 
6:        $\varepsilon \leftarrow \varepsilon + \theta_f^{q_1}(\boldsymbol{\mu}) \theta_f^{q_2}(\boldsymbol{\mu}) C_{q_1, q_2}$   $\triangleright O(1)$ 
7:     end for
8:   end for
9:   for  $q_1 = 1 : Q_a$ 
10:    for  $q_2 = 1 : Q_a$ 
11:       $\varepsilon \leftarrow \varepsilon + \theta_a^{q_1}(\boldsymbol{\mu}) \theta_a^{q_2}(\boldsymbol{\mu}) \mathbf{u}_N(\boldsymbol{\mu})^T \mathbb{E}_{q_1, q_2} \mathbf{u}_N(\boldsymbol{\mu})$   $\triangleright O(N^2)$ 
12:    end for
13:    for  $q_2 = 1 : Q_f$ 
14:       $\varepsilon \leftarrow \varepsilon + \theta_a^{q_1}(\boldsymbol{\mu}) \theta_f^{q_2}(\boldsymbol{\mu}) \mathbf{u}_N(\boldsymbol{\mu})^T \mathbf{d}_{q_1, q_2}$   $\triangleright O(N)$ 
15:    end for
16:  end for
17:   $\Delta_N(\boldsymbol{\mu}) = \sqrt{\varepsilon} / \beta_I(\boldsymbol{\mu})$ 
18: end function

```

3.8 An Illustrative Numerical Example

We illustrate the computational performance of the RB method introduced in this chapter on a steady heat conduction-convection problem which models the temperature of a fluid flowing into the a heat exchanger device, like the one shown in Fig. 3.2. Several other examples, dealing with the problems introduced in Chap. 2, will be addressed in Chap. 9.

The (non-dimensional) temperature u satisfies the following steady advection-diffusion problem:

$$\left\{ \begin{array}{ll} -v\Delta u + \mathbf{b} \cdot \nabla u = 0 & \text{in } \Omega \\ u = g_p & \text{on } \Gamma_p (p = 1, 2, 3) \\ u = 0 & \text{on } \Gamma_w \cup \Gamma_{in} \\ v\nabla u \cdot \mathbf{n} = 0 & \text{on } \Gamma_n, \end{array} \right. \quad (3.88)$$

where the domain $\Omega \subset \mathbb{R}^3$ and its boundaries are displayed in Fig. 3.2, v is the thermal diffusivity, while the convective field \mathbf{b} represents the (prescribed) velocity of the flow field across the exchanger. For low Reynolds numbers, the latter can be obtained as the solution of the stationary Navier-Stokes equations (see Fig. 3.2).

The fluid enters the channel with a reference temperature $u = 0$, then the heating process is regulated by the temperature values g_1, g_2, g_3 imposed on the baffles Γ_1, Γ_2 and Γ_3 , respectively.

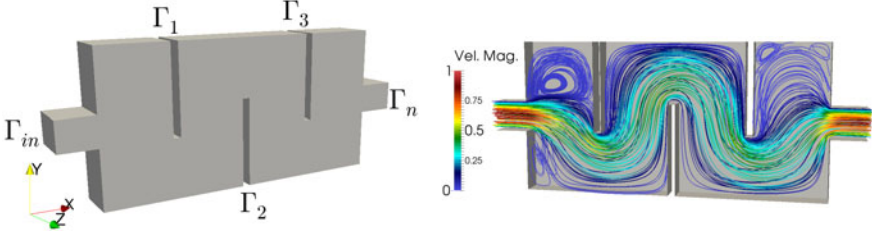


Fig. 3.2 *Left*: computational domain Ω and its boundaries; the flow enters from Γ_{in} ; $\Gamma_1, \Gamma_2, \Gamma_3$ denotes the three baffles through which we regulate the heating process, the rest of the boundary is denoted by Γ_w . *Right*: streamlines of the convective field \mathbf{b} , which is obtained from solving a stationary Navier-Stokes problem for a Reynolds number equal to 100

Problem (3.88) depends on $P = 4$ physical parameters: $\mu_p = g_p$, $p = 1, 2, 3$ are the constant temperatures imposed on the baffles Γ_p , $p = 1, 2, 3$, while $\mu_4 = VL/\nu$ is the Péclet number⁴; here the characteristic length and velocity are set to $L = 1$, $V = 1$, respectively. The parameter domain \mathcal{P} is given by $\mathcal{P} = [0, 12]^3 \times [1, 600] \subset \mathbb{R}^4$.

We highlight that, for the case at hand, the computational domain Ω is fixed (i.e., its shape does not depend on parameters); we are therefore dealing with physical parameters only. The case of geometric parameters affecting the shape of the computational domain will be treated when considering different problems in Chap. 9.

The weak formulation of (3.88) reads: find $u = u(\boldsymbol{\mu}) \in V = H_{\Gamma_D}^1(\Omega)$ such that

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

where $\Gamma_D = \partial\Omega \setminus \Gamma_n$ and

$$a(u, v; \boldsymbol{\mu}) = \frac{1}{\mu_4} \int_{\Omega} \nabla u \cdot \nabla v \, d\Omega + \int_{\Omega} \mathbf{b} \cdot \nabla u v \, d\Omega, \quad f(v; \boldsymbol{\mu}) = - \sum_{p=1}^3 a(r_p(\boldsymbol{\mu}), v; \boldsymbol{\mu})$$

are the parametrized bilinear and linear forms of the problem. Here we denote by $r_p(\boldsymbol{\mu}) \in H^1(\Omega)$ a lifting function such that

$$r_p(\boldsymbol{\mu})|_{\Gamma_p} = g_p(\boldsymbol{\mu}), \quad p = 1, \dots, 3,$$

that is, $r_p(\mathbf{x}; \boldsymbol{\mu}) = \mu_p \psi_p(\mathbf{x})$, being $\psi_p(\mathbf{x}) = 1$ over Γ_p . In this case, obtaining the weak formulation of the parametrized problem is straightforward, since the domain is $\boldsymbol{\mu}$ -independent.

We also remark that the problem admits a simple affine decomposition under the form (3.52)–(3.53) with $Q_a = 2$ and $Q_f = 6$ affine terms.

⁴ We recall that the Péclet number for an advection-diffusion problem is defined by $Pe = VL/\nu$, where V is a characteristic velocity, L is a characteristic length and ν is the diffusion coefficient. It provides a measure of the dominance of the convective term over the diffusion; as such, it plays the same role as the Reynolds number in the Navier-Stokes equations.

In fact,

$$a(u, v; \boldsymbol{\mu}) = \theta_a^1(\boldsymbol{\mu})a_1(u, v) + \theta_a^2(\boldsymbol{\mu})a_2(u, v), \quad f(v; \boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \theta_f^q(\boldsymbol{\mu})f_q(v)$$

where

$$\theta_a^1(\boldsymbol{\mu}) = 1/\mu_4, \quad \theta_a^2(\boldsymbol{\mu}) = 1$$

$$a_1(u, v) = \int_{\Omega} \nabla u \cdot \nabla v d\Omega, \quad a_2(u, v) = \int_{\Omega} \mathbf{b} \cdot \nabla uv d\Omega$$

and, similarly for the right-hand side,

$$\begin{aligned} \theta_f^1(\boldsymbol{\mu}) &= \mu_1/\mu_4, & \theta_f^2(\boldsymbol{\mu}) &= \mu_1, & \theta_f^3(\boldsymbol{\mu}) &= \mu_2/\mu_4, \\ \theta_f^4(\boldsymbol{\mu}) &= \mu_2, & \theta_f^5(\boldsymbol{\mu}) &= \mu_3/\mu_4, & \theta_f^6(\boldsymbol{\mu}) &= \mu_3, \\ f_1(v) &= -a_1(\boldsymbol{\psi}_1, v), & f_2(v) &= -a_2(\boldsymbol{\psi}_1, v), & f_3(v) &= -a_1(\boldsymbol{\psi}_2, v) \\ f_4(v) &= -a_2(\boldsymbol{\psi}_2, v), & f_5(v) &= -a_1(\boldsymbol{\psi}_3, v), & f_6(v) &= -a_2(\boldsymbol{\psi}_3, v). \end{aligned}$$

The problem is then discretized by piecewise linear finite elements: using a mesh made of approximately $2.2 \cdot 10^5$ tetrahedral elements (we exploit the symmetry along the z -axis so that only half of the domain needs to be meshed), we obtain a high-fidelity space V_h of dimension $N_h = 44\,171$. We then construct iteratively a reduced basis space V_N . By imposing a tolerance $\varepsilon = 10^{-4}$ on the (relative) a posteriori error estimate evaluated over a rich training sample $\mathcal{E}_{train} \subset \mathcal{P}$, that is, by requiring that

$$\max_{\boldsymbol{\mu} \in \mathcal{E}_{train}} \frac{\Delta_N(\boldsymbol{\mu})}{\|u_N(\boldsymbol{\mu})\|_V} \leq \varepsilon$$

we obtain a RB space V_N of dimension $N = 29$. Here \mathcal{E}_{train} is a uniform sample extracted from \mathcal{P} of size $|\mathcal{E}_{train}| = 1000$. In Fig. 3.4 we report the average over a test sample $\mathcal{E}_{test} \subset \mathcal{P}$ of both the a posteriori error bound Δ_N and the error $\|u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_V$ as functions of N , evaluated during the online stage. Here \mathcal{E}_{test} is a uniform sample extracted from \mathcal{P} of size $|\mathcal{E}_{test}| = 350$, which does not include any point of \mathcal{E}_{train} . See Sect. 3.7 for the evaluation of the a posteriori error estimate. We will show in Chaps. 6 and 7 how to construct RB spaces; here we just focus on the numerical performance of the reduction procedure.

In the offline stage, the computation of the terms involved by the dual norm of the residual is done in parallel; using 4 cores, the offline stage is executed in about 5 minutes⁵. In Fig. 3.3 some representative examples of temperature distribution obtained through the RB approximation are given. The average time for a single solution of the reduced-order model is about 1 ms, providing a speedup of three orders of magnitude with respect to the high-fidelity model.

Thanks to the offline/online decomposition, for any given parameter value $\boldsymbol{\mu}$ at the online stage the solution of the problem at hand can be obtained very rapidly.

⁵ Numerical results were obtained on a workstation with a Intel Core i5-2400S CPU and 16 GB of RAM.

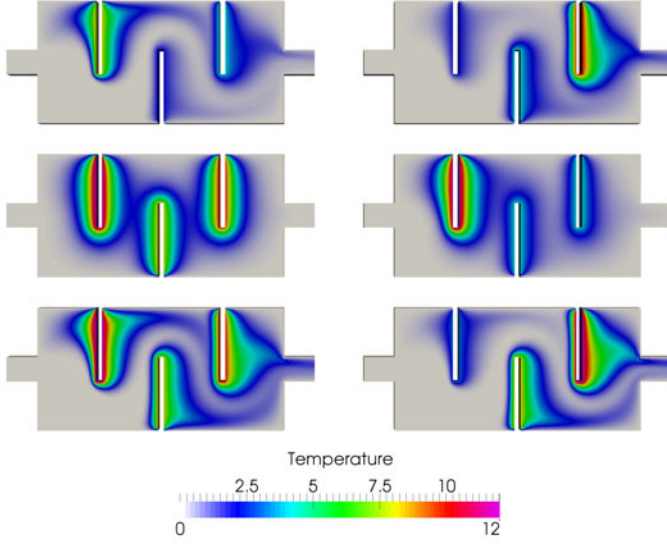


Fig. 3.3 RB solutions of (3.88) corresponding to different sets of parameter values

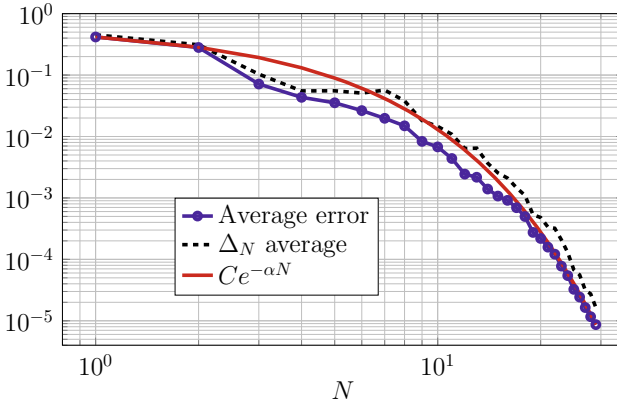


Fig. 3.4 Comparison of the average error $\|u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_V$ and bound $\Delta_N(\boldsymbol{\mu})$ computed on a set of 350 random parameter values

The offline/online decomposition is a direct consequence of the affine parametric dependence property, which allows to split the assembly of the reduced matrices and vectors.

Several comments are in order:

1. the solution $u_h(\boldsymbol{\mu})$ of the problem highly depends on the variation of $\boldsymbol{\mu}$ over the parameter space. Nevertheless, a relatively small number of basis functions is sufficient to capture this variability;

Table 3.1 Computational details for the high-fidelity and reduced-order models of (3.88)

High-fidelity model		Reduced-order model	
Number of FE dofs N_h	44 171	Number of RB dofs	29
Affine operator components Q_a	2	Dofs reduction	1520:1
Affine rhs components Q_f	6	Offline CPU time	≈ 5 min
FE solution time (assembly + solution) ≈ 3.5 s		Online CPU time	1 ms

2. as we can see from the graph reported in Fig. 3.4, the estimator $\Delta_N(\boldsymbol{\mu})$ is very sharp, its effectivity $\eta_N(\boldsymbol{\mu})$ being very close to 1;
3. the error (and its upper bound) show an exponential decay with respect to the dimension N of the RB subspace (where $\alpha \approx 0.38$ in the convergence plot of Fig. 3.4). This is a remarkable property of RB approximations, making them very accurate and efficient.

As we will see in Chap. 5, under suitable assumptions on the parametrized linear and bilinear forms of an elliptic problem (which for the case at hand are indeed fulfilled), an exponential convergence result can be mathematically proven.

Further details related with the computational performances of the RB method for the problem at hand are reported in Table 3.1. We will come back to the heat conduction-convection problem discussed so far for comparing the two strategies exploited to construct RB spaces (see Sect. 7.2) and, later, for solving an associated optimal control problem by the RB method (see Sect. 12.5).

3.9 Exercises

1. Prove the result of Proposition 3.1 by means of the Galerkin orthogonality property and Theorem A.1. Show that also the converse is true.
2. Similarly to the case of (2.36) (see Exercise 8, Chap. 2), show that if $a(\cdot, \cdot; \boldsymbol{\mu})$ is symmetric and coercive, the estimate (3.33) holds.
3. Assume that a linear output of the solution, $z(\boldsymbol{\mu}) = l(u(\boldsymbol{\mu}))$, $l \in V'$, is given by $l = f$, and that $a(\cdot, \cdot; \boldsymbol{\mu}) : V \times V \rightarrow \mathbb{R}$ for any $\boldsymbol{\mu} \in \mathcal{P}$ is a symmetric, continuous and coercive bilinear form. Show that in this case,

$$z_h(\boldsymbol{\mu}) - z_N(\boldsymbol{\mu}) = \|u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_{\boldsymbol{\mu}}^2, \quad (3.89)$$

being $z_h(\boldsymbol{\mu}) = l(u_h(\boldsymbol{\mu}))$, $z_N(\boldsymbol{\mu}) = l(u_N(\boldsymbol{\mu}))$.

Hint: using the compliance assumption, write $z_h(\boldsymbol{\mu}) - z_N(\boldsymbol{\mu}) = a(u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu}), u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu}); \boldsymbol{\mu}) + a(u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu}), u_N(\boldsymbol{\mu}); \boldsymbol{\mu})$, then exploit (3.31).

4. Prove relation (3.39).

5. Show that for the LS-RB problem

$$\|u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_V \leq \frac{\gamma_h(\boldsymbol{\mu})}{\beta_h(\boldsymbol{\mu})} \inf_{v \in V_N} \|u_h(\boldsymbol{\mu}) - v\|_V \quad \forall \boldsymbol{\mu} \in \mathcal{P}.$$

Hint: recall from the proof of Proposition 3.2 that $\beta_N(\boldsymbol{\mu}) \geq \beta_h(\boldsymbol{\mu})$.

6. Prove the inequalities (3.50) and (3.51). *Hint:* to prove the second inequality it is useful to recall that if $f, g : A \rightarrow \mathbb{R}$ are two bounded functions, then $\inf_A (fg) \geq \inf_A f \inf_A g$ and $\sup_A (fg) \leq \sup_A f \sup_A g$.
7. Let us consider a heat conduction problem in a square domain Ω made of $P = B_1 \times B_2$ blocks, each one representing a subregion with (a priori different) constant thermal conductivity. Denote by $\bar{\Omega} = \bigcup_{i=1}^P \bar{\Omega}_i$, where Ω_i , $i = 1 \dots, P$, is the subregion featuring a conductivity $\mu_i > 0$. The strong form reads:

$$\begin{cases} -\operatorname{div}(k(\boldsymbol{\mu}) \nabla u) = 0 & \text{in } \Omega \\ u = 1 & \text{on } \Gamma_{top} \\ k(\boldsymbol{\mu}) \frac{\partial u(\boldsymbol{\mu})}{\partial \mathbf{n}} = 1 & \text{on } \Gamma_{base} \\ k(\boldsymbol{\mu}) \frac{\partial u(\boldsymbol{\mu})}{\partial \mathbf{n}} = 0 & \text{on } \partial\Omega \setminus (\Gamma_{base} \cup \Gamma_{top}) \end{cases}$$

whereas the conductivity field is given by $k(\mathbf{x}; \boldsymbol{\mu}) = \sum_{i=1}^P \mu_i \chi_{\Omega_i}(\mathbf{x})$.

- Recover the weak formulation (3.2) of the problem, expliciting the space V , the bilinear form $a(\cdot, \cdot; \boldsymbol{\mu})$ and the linear form $f(\cdot; \boldsymbol{\mu})$;
 - show that $a(\cdot, \cdot; \boldsymbol{\mu})$ and $f(\cdot; \boldsymbol{\mu})$ fulfill the affine parametric dependence property (3.52)–(3.53);
 - consider the more general case of a diffusion coefficient that is not constant over each block Ω_i . Which assumption has to be fulfilled by the function $k(\mathbf{x}; \boldsymbol{\mu})$ in order to obtain a bilinear form satisfying property (3.52)?
8. Show that, in the energy norm, the following a posteriori error estimate holds for the RB approximation of a strongly coercive problem

$$\|u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_{\boldsymbol{\mu}} \leq \frac{\|r(\cdot; \boldsymbol{\mu})\|_{V_h'}}{(\alpha_h(\boldsymbol{\mu}))^{1/2}}.$$

Then, by exploiting the result of Exercise 3, show that the following a posteriori error estimate holds for a compliant output:

$$\|z_h(\boldsymbol{\mu}) - z_N(\boldsymbol{\mu})\|_{\boldsymbol{\mu}} \leq \frac{\|r(\cdot; \boldsymbol{\mu})\|_{V_h'}^2}{\alpha_h(\boldsymbol{\mu})}.$$

Chapter 4

On the Algebraic and Geometric Structure of RB Methods

RB methods are revisited from both an algebraic and a geometric standpoint. A number of relationships between the Galerkin RB approximation (as well as least-squares RB approximation) and the Galerkin high-fidelity approximation (3.11) are highlighted, for the purpose of illustrating, in a more fitting way and from a different perspective, the mathematical structure underpinning RB methods. The key role played by the transformation matrix in defining orthogonal and oblique projections is emphasized.

4.1 Algebraic Construction and Interpretation

Let, for every $\boldsymbol{\mu} \in \mathcal{P}$, $u_h(\boldsymbol{\mu}) \in V_h$ and $u_N(\boldsymbol{\mu}) \in V_N$ represent¹ the high-fidelity and the RB solutions of problem (3.11) and (3.26), respectively. Then we denote by $\mathbf{u}_h(\boldsymbol{\mu}) \in \mathbb{R}^{N_h}$ and $\mathbf{u}_N(\boldsymbol{\mu}) \in \mathbb{R}^N$ the vectors of degrees of freedom associated to the functions $u_h(\boldsymbol{\mu})$ and $u_N(\boldsymbol{\mu})$, respectively, which are given by (see (2.50) and (3.23))

$$\mathbf{u}_h(\boldsymbol{\mu}) = (u_h^{(1)}(\boldsymbol{\mu}), \dots, u_h^{(N_h)}(\boldsymbol{\mu}))^T, \quad \mathbf{u}_N(\boldsymbol{\mu}) = (u_N^{(1)}(\boldsymbol{\mu}), \dots, u_N^{(N)}(\boldsymbol{\mu}))^T.$$

Let $\{\boldsymbol{\varphi}^r\}_{r=1}^{N_h}$ denote a basis for the high-fidelity space V_h that is orthogonal with respect to a *discrete* scalar product $(\cdot, \cdot)_h$. In the case of a Lagrangian FE basis (see Sect. 2.5), the discrete scalar product that ensures orthogonality is

$$(u_h, v_h)_h = \sum_{r=1}^{N_h} u_h(\mathbf{N}_r) v_h(\mathbf{N}_r),$$

where $\{\mathbf{N}_r\}_{r=1}^{N_h}$ is the set of FE nodes such that $\boldsymbol{\varphi}^r(\mathbf{N}_s) = \delta_{rs}$, $r, s = 1, \dots, N_h$; note that other choices of scalar products can be made. We recall that $\{\boldsymbol{\zeta}_k\}_{k=1}^N$ denotes a basis of the reduced space V_N , orthonormal with respect to a *discrete* scalar product that we shall denote by $(\cdot, \cdot)_N$, see (3.21), and that we can choose, a priori, different

¹ We remark that here $V_N \subset V_h$ denotes a function space, not a subset of \mathbb{R}^{N_h} .

from $(\cdot, \cdot)_h$. Indeed, as we have seen in Sect. 3.4, we usually orthonormalize the basis functions $\{\zeta_k\}$ with respect to the V -scalar product, i.e. we choose $(\cdot, \cdot)_N = (\cdot, \cdot)_V$.

As anticipated in Sect. 2.4 (see (2.50)) we can consider the following bijection between the spaces \mathbb{R}^{N_h} and V_h :

$$\begin{cases} \mathbf{v}_h \in \mathbb{R}^{N_h} \leftrightarrow v_h \in V_h, \\ \mathbf{v}_h = (v_h^{(1)}, \dots, v_h^{(N_h)})^T \leftrightarrow v_h = \sum_{r=1}^{N_h} v_h^{(r)} \varphi^r \end{cases} \quad (4.1)$$

and, similarly, between the spaces \mathbb{R}^N and V_N :

$$\begin{cases} \mathbf{v}_N \in \mathbb{R}^N \leftrightarrow v_N \in V_N, \\ \mathbf{v}_N = (v_N^{(1)}, \dots, v_N^{(N)})^T \leftrightarrow v_N = \sum_{k=1}^N v_N^{(k)} \zeta_k. \end{cases} \quad (4.2)$$

Thanks to the orthonormality of the basis functions, the coefficients of the representation provided by (4.1) and (4.2) are given by

$$v_h^{(r)} = (v_h, \varphi^r)_h, \quad v_N^{(k)} = (v_N, \zeta_k)_N, \quad r = 1, \dots, N_h, \quad k = 1, \dots, N.$$

Using bijection (4.1), we can easily prove that

$$(\mathbf{u}_h, \mathbf{v}_h)_2 = (u_h, v_h)_h \quad \forall \mathbf{u}_h, \mathbf{v}_h \in \mathbb{R}^{N_h} \quad (\text{equivalently, } \forall u_h, v_h \in V_h). \quad (4.3)$$

Indeed

$$\begin{aligned} (u_h, v_h)_h &= \left(\sum_{r=1}^{N_h} u_h^{(r)} \varphi^r, \sum_{s=1}^{N_h} v_h^{(s)} \varphi^s \right)_h \\ &= \sum_{r,s=1}^{N_h} u_h^{(r)} v_h^{(s)} (\varphi^r, \varphi^s)_h = \sum_{r,s=1}^{N_h} u_h^{(r)} v_h^{(r)} = (\mathbf{u}_h, \mathbf{v}_h)_2. \end{aligned}$$

Similarly, using bijection (4.2) we can show that (see Exercise 1)

$$(\mathbf{u}_N, \mathbf{v}_N)_2 = (u_N, v_N)_N \quad \forall \mathbf{u}_N, \mathbf{v}_N \in \mathbb{R}^N \quad (\text{equivalently, } \forall u_N, v_N \in V_N). \quad (4.4)$$

4.1.1 Algebraic Interpretation of the G-RB Problem

We first discuss the algebraic connection between the G-RB problem (3.27) and the Galerkin high-fidelity approximation (3.11). We recall that the algebraic form of the G-RB problem (3.27) reads (see (3.49))

$$\mathbb{A}_N(\boldsymbol{\mu}) \mathbf{u}_N(\boldsymbol{\mu}) = \mathbf{f}_N(\boldsymbol{\mu}), \quad (4.5)$$

with $\mathbf{f}_N = (f_N^{(1)}, \dots, f_N^{(N)})^T$, $f_N^{(k)} = f(\zeta_k)$, $(\mathbb{A}_N(\boldsymbol{\mu}))_{km} = a(\zeta_m, \zeta_k; \boldsymbol{\mu})$, for $k, m = 1, \dots, N$. On the other hand, the Galerkin high-fidelity approximation (3.11) reads in matrix form as (see (3.16))

$$\mathbb{A}_h(\boldsymbol{\mu})\mathbf{u}_h(\boldsymbol{\mu}) = \mathbf{f}_h(\boldsymbol{\mu}), \quad (4.6)$$

with

$$\begin{aligned} \mathbf{f}_h(\boldsymbol{\mu}) &= (f_h^{(1)}(\boldsymbol{\mu}), \dots, f_h^{(N_h)}(\boldsymbol{\mu}))^T, \quad f_h^{(r)}(\boldsymbol{\mu}) = f(\varphi^r; \boldsymbol{\mu}), \quad r = 1, \dots, N_h, \\ (\mathbb{A}_h(\boldsymbol{\mu}))_{rs} &= a(\varphi^s, \varphi^r; \boldsymbol{\mu}), \quad r, s = 1, \dots, N_h. \end{aligned}$$

As seen in Chap. 3, an algebraic relation between the RB problem (4.5) and the high-fidelity problem (4.6) can be established by means of the transformation matrix $\mathbb{V} \in \mathbb{R}^{N_h \times N}$ introduced in (3.58),

$$(\mathbb{V})_{rk} = \zeta_k^{(r)} = (\zeta_k, \varphi^r)_h, \quad r = 1, \dots, N_h, \quad k = 1, \dots, N. \quad (4.7)$$

Using this matrix, we can easily obtain the following algebraic identities:

$$\mathbb{A}_N(\boldsymbol{\mu}) = \mathbb{V}^T \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V}, \quad \mathbf{f}_N(\boldsymbol{\mu}) = \mathbb{V}^T \mathbf{f}_h(\boldsymbol{\mu}). \quad (4.8)$$

Indeed,

$$\begin{aligned} (\mathbb{V}^T \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V})_{km} &= \sum_{r,s=1}^{N_h} (\mathbb{V})_{kr}^T (\mathbb{A}_h(\boldsymbol{\mu}))_{rs} (\mathbb{V})_{sm} = \sum_{r,s=1}^{N_h} \zeta_k^{(r)} a(\varphi^s, \varphi^r; \boldsymbol{\mu}) \zeta_m^{(s)} \\ &= a \left(\sum_{s=1}^{N_h} \zeta_m^{(s)} \varphi^s, \sum_{r=1}^{N_h} \zeta_k^{(r)} \varphi^r; \boldsymbol{\mu} \right) = a(\zeta_m, \zeta_k; \boldsymbol{\mu}) = (\mathbb{A}_N(\boldsymbol{\mu}))_{km}, \\ (\mathbb{V}^T \mathbf{f}_h(\boldsymbol{\mu}))_k &= \sum_{r=1}^{N_h} \mathbb{V}_{rk} f_h^{(r)}(\boldsymbol{\mu}) = \sum_{r=1}^{N_h} \zeta_k^{(r)} f(\varphi^r; \boldsymbol{\mu}) \\ &= f \left(\sum_{r=1}^{N_h} \zeta_k^{(r)} \varphi^r; \boldsymbol{\mu} \right) = f(\zeta_k; \boldsymbol{\mu}) = (\mathbf{f}_N(\boldsymbol{\mu}))_k. \end{aligned} \quad (4.9)$$

4.1.2 Algebraic properties of the G-RB Problem

We can now characterize the (vector representation of the) error

$$\mathbf{e}_h(\boldsymbol{\mu}) = \mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}) \quad (4.10)$$

between the solution of the G-RB problem and the Galerkin high-fidelity solution in terms of the (vector representation of the) high-fidelity residual of the G-RB so-

lution, expressed by

$$\mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu}) = \mathbf{f}_h(\boldsymbol{\mu}) - \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}). \quad (4.11)$$

The following lemma (see Exercise 2 for its proof) provides the main algebraic connection between G-RB and Galerkin high-fidelity approximations.

Lemma 4.1. *The following algebraic relations hold:*

$$\mathbb{A}_h(\boldsymbol{\mu}) \mathbf{e}_h(\boldsymbol{\mu}) = \mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu}) \quad (4.12)$$

$$\mathbb{V}^T \mathbb{A}_h(\boldsymbol{\mu}) \mathbf{u}_h(\boldsymbol{\mu}) = \mathbf{f}_N(\boldsymbol{\mu}) \quad (4.13)$$

$$\mathbb{V}^T \mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu}) = \mathbf{0}. \quad (4.14)$$

Note that condition (4.12) is the algebraic counterpart of the Galerkin orthogonality property (3.31) valid for the G-RB problem.

In summary, for a given matrix \mathbb{V} of reduced bases, the Galerkin reduced basis G-RB problem (4.5) can be formally obtained as follows:

Galerkin Reduced Basis (G-RB) problem

1. consider the Galerkin high-fidelity problem (4.6);
2. set $\mathbf{u}_h(\boldsymbol{\mu}) = \mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}) + \mathbf{e}_h(\boldsymbol{\mu})$, where $\mathbf{u}_N \in \mathbb{R}^N$ has to be determined and the error \mathbf{e}_h is the difference between \mathbf{u}_h and $\mathbb{V} \mathbf{u}_N$;
3. left multiply (4.6) by \mathbb{V}^T to obtain

$$\mathbb{V}^T \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}) - \mathbb{V}^T \mathbf{f}_h(\boldsymbol{\mu}) = -\mathbb{V}^T \mathbb{A}_h(\boldsymbol{\mu}) \mathbf{e}_h(\boldsymbol{\mu}),$$

that is

$$\mathbb{V}^T \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}) - \mathbb{V}^T \mathbf{f}_h(\boldsymbol{\mu}) = -\mathbb{V}^T \mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu});$$

4. require \mathbf{u}_N to satisfy $\mathbb{V}^T \mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu}) = \mathbf{0}$, or equivalently

$$\mathbb{V}^T \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}) = \mathbb{V}^T \mathbf{f}_h(\boldsymbol{\mu}). \quad (4.15)$$

If $\mathbb{A}_h(\boldsymbol{\mu})$ is symmetric and positive definite, then the G-RB solution satisfies the following residual *minimization property*

$$\mathbf{u}_N(\boldsymbol{\mu}) = \arg \min_{\mathbf{v}_N \in \mathbb{R}^N} \|\mathbf{r}_h(\mathbf{v}_N; \boldsymbol{\mu})\|_{\mathbb{A}_h^{-1}(\boldsymbol{\mu})}^2. \quad (4.16)$$

As a matter of fact, we have

$$\begin{aligned} \|\mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu})\|_{\mathbb{A}_h^{-1}(\boldsymbol{\mu})}^2 &= (\mathbf{f}_h(\boldsymbol{\mu}) - \mathbb{A}_h(\boldsymbol{\mu})\nabla \mathbf{u}_N, \mathbf{f}_h(\boldsymbol{\mu}) - \mathbb{A}_h(\boldsymbol{\mu})\nabla \mathbf{u}_N)_{\mathbb{A}_h^{-1}(\boldsymbol{\mu})} = \\ &= (\mathbb{A}_h^{-1/2}(\boldsymbol{\mu})\mathbf{f}_h(\boldsymbol{\mu}) - \mathbb{A}_h^{1/2}(\boldsymbol{\mu})\nabla \mathbf{u}_N, \mathbb{A}_h^{-1/2}(\boldsymbol{\mu})\mathbf{f}_h(\boldsymbol{\mu}) - \mathbb{A}_h^{1/2}(\boldsymbol{\mu})\nabla \mathbf{u}_N). \end{aligned}$$

This can be regarded as the least-squares solution of the system

$$\mathbb{A}_h^{1/2}(\boldsymbol{\mu})\nabla \mathbf{u}_N = \mathbb{A}_h^{-1/2}(\boldsymbol{\mu})\mathbf{f}_h(\boldsymbol{\mu}),$$

whose corresponding *normal equations*² are

$$\nabla^T \mathbb{A}_h^{1/2}(\boldsymbol{\mu}) \mathbb{A}_h^{1/2}(\boldsymbol{\mu}) \nabla \mathbf{u}_N = \nabla^T \mathbb{A}_h^{1/2}(\boldsymbol{\mu}) \mathbb{A}_h^{-1/2}(\boldsymbol{\mu}) \mathbf{f}_h(\boldsymbol{\mu}) = \nabla^T \mathbf{f}_h(\boldsymbol{\mu}).$$

Note that the latter coincide with the G-RB problem (4.5).

4.1.3 Least-Squares and Petrov-Galerkin RB Problems

The Galerkin projection, which leads to the G-RB problem discussed so far, is the most common strategy to build a RB method. In this case, the trial space (namely, the space where we seek the solution) and the test space are equal; from an algebraic standpoint this is reflected by the identities (4.8), where the matrix by which we pre- and post-multiply the high-fidelity stiffness matrix is indeed the same. However, the trial and the test space may be chosen in a different way, giving rise to what we have called a Petrov-Galerkin formulation. In this section we provide some insights about this approach.

The Least-Squares Reduced Basis method

As seen in Chap. 3 an alternative approach to G-RB is the least-squares reduced basis LS-RB – sometimes also called *minimum residual* – method. The LS-RB solution satisfies (see Proposition 3.3)

$$\mathbf{u}_N(\boldsymbol{\mu}) = \arg \min_{\mathbf{v}_N \in \mathbb{R}^N} \|\mathbf{r}_h(\mathbf{v}_N; \boldsymbol{\mu})\|_{\mathbb{X}_h}^2. \quad (4.17)$$

Note that (4.17) is nothing but the discrete equivalent of relation (3.42). Moreover, also note that the minimization criterion (4.17) applies for any matrix \mathbb{A}_h , whereas

² We recall that, given $\mathbf{c} \in \mathbb{R}^{N_h}$ and $\mathbb{B} \in \mathbb{R}^{N_h \times N}$, the overdetermined system $\mathbb{B}\mathbf{v} = \mathbf{c}$ can be solved in the least-squares sense, by seeking $\mathbf{u} = \arg \min_{\mathbf{v} \in \mathbb{R}^N} \|\mathbf{c} - \mathbb{B}\mathbf{v}\|_2^2$. The solution is unique provided that the N columns of \mathbb{B} are linearly independent, and can be obtained by solving the following system of normal equations

$$(\mathbb{B}^T \mathbb{B})\mathbf{u} = \mathbb{B}^T \mathbf{c}.$$

(4.16), relative to the G-RB method, requires \mathbb{A}_h to be symmetric and positive definite. The solution to (4.17) coincides with the solution of the *normal equations*

$$(\mathbb{X}_h^{-1} \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V})^T \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V} \mathbf{u}_N = (\mathbb{X}_h^{-1} \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V})^T \mathbf{f}_h(\boldsymbol{\mu}),$$

that is

$$(\mathbb{X}_h^{-1} \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V})^T \mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu}) = \mathbf{0}. \quad (4.18)$$

For a given matrix \mathbb{V} , the LS-RB problem can therefore be obtained as follows:

Least-Squares Reduced Basis (LS-RB) problem

1. consider the Galerkin high-fidelity problem (4.6);
2. set $\mathbf{u}_h(\boldsymbol{\mu}) = \mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}) + \mathbf{e}_h(\boldsymbol{\mu})$, where $\mathbf{u}_N \in \mathbb{R}^N$ has to be determined and the error \mathbf{e}_h is the difference between \mathbf{u}_h and $\mathbb{V} \mathbf{u}_N$;
3. left multiply (4.6) by $(\mathbb{X}_h^{-1} \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V})^T$ to obtain

$$(\mathbb{X}_h^{-1} \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V})^T \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V} \mathbf{u}_N = (\mathbb{X}_h^{-1} \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V})^T (\mathbf{f}_h(\boldsymbol{\mu}) - \mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu}));$$

4. require $\mathbf{u}_N(\boldsymbol{\mu})$ to satisfy $(\mathbb{X}_h^{-1} \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V})^T \mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu}) = \mathbf{0}$, that is, equivalently,

$$\mathbb{V}^T \mathbb{A}_h^T(\boldsymbol{\mu}) \mathbb{X}_h^{-1} \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}) = \mathbb{V}^T \mathbb{A}_h^T(\boldsymbol{\mu}) \mathbb{X}_h^{-1} \mathbf{f}_h(\boldsymbol{\mu}). \quad (4.19)$$

Note that (4.19) can be rewritten in the form (4.5) provided we set

$$\mathbb{A}_N(\boldsymbol{\mu}) = \mathbb{V}^T \mathbb{A}_h^T(\boldsymbol{\mu}) \mathbb{X}_h^{-1} \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V}, \quad \mathbf{f}_N(\boldsymbol{\mu}) = \mathbb{V}^T \mathbb{A}_h^T(\boldsymbol{\mu}) \mathbb{X}_h^{-1} \mathbf{f}_h(\boldsymbol{\mu}).$$

The Petrov-Galerkin Reduced Basis method

Problem (4.19) can be regarded as a special instance of the following Petrov-Galerkin (rather than Galerkin) method: find $u_N(\boldsymbol{\mu}) \in V_N$ such that

$$a(u_N(\boldsymbol{\mu}), w_N; \boldsymbol{\mu}) = f(w_N; \boldsymbol{\mu}) \quad \forall w_N \in W_N, \quad (4.20)$$

where $W_N \subset V_h$ is a subspace of dimension N , different from V_N . If we denote by $\{\eta_k, k = 1, \dots, N\}$ a basis for W_N , and by $\mathbb{W} \in \mathbb{R}^{N_h \times N}$ the matrix whose entries are

$$(\mathbb{W})_{rk} = (\eta_k, \boldsymbol{\varphi}^r)_h, \quad r = 1, \dots, N_h, \quad k = 1, \dots, N, \quad (4.21)$$

we can still express (4.20) in the algebraic form (4.5); this time, however, instead of (4.8) we have

$$\mathbf{f}_N(\boldsymbol{\mu}) = \mathbb{W}^T \mathbf{f}_h(\boldsymbol{\mu}), \quad \mathbb{A}_N(\boldsymbol{\mu}) = \mathbb{W}^T \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V}. \quad (4.22)$$

For any two given matrices \mathbb{V} and \mathbb{W} , the PG-RB method can be obtained as follows:

Petrov-Galerkin Reduced Basis (PG-RB) problem

1. consider the Galerkin high-fidelity problem (4.6);
2. set $\mathbf{u}_h(\boldsymbol{\mu}) = \mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}) + \mathbf{e}_h(\boldsymbol{\mu})$, where $\mathbf{u}_N \in \mathbb{R}^N$ has to be determined and the error \mathbf{e}_h is the difference between \mathbf{u}_h and $\mathbb{V}\mathbf{u}_N$;
3. left multiply (4.6) by \mathbb{W}^T to obtain

$$\mathbb{W}^T \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}) = \mathbb{W}^T \mathbf{f}_h - \mathbb{W}^T \mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu});$$

4. require $\mathbf{u}_N(\boldsymbol{\mu})$ to satisfy $\mathbb{W}^T \mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu}) = \mathbf{0}$, that is, equivalently,

$$\mathbb{W}^T \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}) = \mathbb{W}^T \mathbf{f}_h(\boldsymbol{\mu}). \quad (4.23)$$

As already anticipated, the LS-RB problem (4.19) is a special case of the PG-RB problem (4.23) corresponding to the choice $\mathbb{W} = \mathbb{X}_h^{-1} \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V}$.

4.2 Geometric Interpretation

We can also characterize the RB approximation \mathbf{u}_N , as well as the error $\mathbf{e}_h = \mathbf{u}_h - \mathbb{V}\mathbf{u}_N$ between the solution of the reduced and the high-fidelity approximations, from a geometric standpoint. To this end, we first recall the definition of projection operators in finite-dimensional spaces and their matrix representation.

4.2.1 Projection and Bases

Subspace projection and basis construction are two central concepts in reduced-order modeling for parametrized PDEs. We first introduce projection operators over a subspace of a finite-dimensional space and then focus on their matrix representation. See e.g. [235] for further details and proofs of the mentioned results.

A (discrete) *projection operator* (or, simply, *projector*) is any linear map $\mathbb{P} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ (that is a $n \times n$ square matrix) such that $\mathbb{P}^2 = \mathbb{P}$. If \mathbb{P} is a projector, then $\mathbb{I} - \mathbb{P}$ is a projector too. Moreover, since every $\mathbf{x} \in \mathbb{R}^n$ can be written as $\mathbf{x} = \mathbb{P}\mathbf{x} + (\mathbb{I} - \mathbb{P})\mathbf{x}$, the entire space \mathbb{R}^n can be decomposed as a *direct sum*

$$\mathbb{R}^n = \text{Ker}(\mathbb{P}) \oplus \text{Range}(\mathbb{P}).$$

In fact, if $\mathbf{x} \in \text{Range}(\mathbb{P})$, then $\mathbb{P}\mathbf{x} = \mathbf{x}$, whereas if $\mathbf{x} \in \text{Ker}(\mathbb{P})$, then $\mathbb{P}\mathbf{x} = \mathbf{0}$. On the other hand, for each pair of subspaces M and S such that $M \oplus S = \mathbb{R}^n$, it is possible

to define a unique projector \mathbb{P} such that $\text{Range}(\mathbb{P}) = M$ and $\text{Ker}(\mathbb{P}) = S$, so that each element of \mathbb{R}^n can be written as $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$, $\mathbf{x}_1 \in M$, $\mathbf{x}_2 \in S$.

For any $\mathbf{x} \in \mathbb{R}^n$, the projection $\mathbf{v} = \mathbb{P}\mathbf{x}$ is such that $\mathbf{v} \in M$, $\mathbf{x} - \mathbf{v} \in S$ or, equivalently,

$$\mathbf{v} \in M, \quad \mathbf{x} - \mathbf{v} \perp L, \quad (4.24)$$

where $L = S^\perp$ is the orthogonal complement (see (A.10), Appendix A) of S . We say that \mathbb{P} projects \mathbf{x} onto the subspace M orthogonally to L , i.e. \mathbf{v} is the projection of \mathbf{x} onto M , orthogonal to L . If \mathbb{P} has rank m , the range of $\mathbb{I} - \mathbb{P}$ has dimension $n - m$, and $L = S^\perp$ has dimension m as well. Here we consider orthogonality with respect to the Euclidean scalar product, although any scalar product can be used as well, see, e.g. the proof of Proposition 3.3.

If $\dim(M) = \dim(L) = m$, for any $\mathbf{x} \in \mathbb{R}^n$, the projection $\mathbf{v} = \mathbb{P}\mathbf{x}$ is univocally defined through conditions (4.24).

4.2.2 Matrix Characterization of Projection Operators

It is useful to derive a matrix characterization of projection operators in terms of the bases chosen to represent the subspaces M and L , say $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ for the subspace $M = \text{Range}(\mathbb{P})$ and $\{\mathbf{w}_1, \dots, \mathbf{w}_m\}$ for the subspace $L = (\text{Ker}(\mathbb{P}))^\perp$. Let us assume that the basis functions of the two spaces are mutually orthonormal, that is $\mathbf{v}_i^T \mathbf{w}_j = \delta_{ij}$, $i, j = 1, \dots, m$. If we consider the following matrices

$$\mathbb{V} = [\mathbf{v}_1 \mid \dots \mid \mathbf{v}_m] \in \mathbb{R}^{n \times m}, \quad \mathbb{W} = [\mathbf{w}_1 \mid \dots \mid \mathbf{w}_m] \in \mathbb{R}^{n \times m},$$

the orthonormality condition translates into

$$\mathbb{W}^T \mathbb{V} = \mathbb{I}_m, \quad (4.25)$$

where \mathbb{I}_m represents the identity matrix of size $m \times m$. Since the projection $\mathbf{v} = \mathbb{P}\mathbf{x} \in M$, we can express it as a linear combination of the basis vectors of M , that is there exists $\mathbf{y} = [y_1 \dots y_m]^T \in \mathbb{R}^m$ such that

$$\mathbf{v} = \sum_{i=1}^m y_i \mathbf{v}_i, \quad (4.26)$$

or, equivalently, $\mathbb{P}\mathbf{x} = \mathbb{V}\mathbf{y}$. Hence, the condition $\mathbf{x} - \mathbb{P}\mathbf{x} \perp L$ can be equivalently written as $\mathbf{w}_j^T (\mathbf{x} - \mathbb{V}\mathbf{y}) = 0 \forall j = 1, \dots, m$, that is, in matrix form,

$$\mathbb{W}^T (\mathbf{x} - \mathbb{V}\mathbf{y}) = \mathbf{0}. \quad (4.27)$$

In particular, (4.26) is the algebraic analogue of the first relation in (4.24), which establishes the m degrees of freedom yielding $\mathbf{v} = \mathbb{P}\mathbf{x}$, whereas (4.27) provides the matrix form of the second relation in (4.24), which gives the m conditions that define $\mathbb{P}\mathbf{x}$ from the former degrees of freedom.

Thanks to (4.25), we obtain $\mathbf{y} = \mathbb{W}^T \mathbf{x}$, that is, $\mathbb{P}\mathbf{x} = \mathbb{V}\mathbb{W}^T \mathbf{x}$. This yields the following matrix representation of \mathbb{P}

$$\mathbb{P} = \mathbb{V}\mathbb{W}^T \in \mathbb{R}^{n \times n}. \quad (4.28)$$

When the orthogonality condition (4.25) does not hold, we obtain the more involved relation

$$\mathbb{P} = \mathbb{V}(\mathbb{W}^T \mathbb{V})^{-1} \mathbb{W}^T. \quad (4.29)$$

4.2.3 Orthogonal and Oblique Projection Operators

Among projection operators, a relevant class is that of *orthogonal projectors*. If $S^\perp = M$, that is, when $\text{Ker}(\mathbb{P}) = \text{Range}(\mathbb{P})^\perp$, \mathbb{P} is said to be the *orthogonal projector* onto M . A projector which is not orthogonal is said to be *oblique*.

A simple characterization of orthogonal projectors can be provided through the following requirements to be satisfied for any $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbb{P}\mathbf{x} \in M \quad \text{and} \quad (\mathbb{I} - \mathbb{P})\mathbf{x} \perp M \quad (4.30)$$

or, equivalently,

$$\mathbb{P}\mathbf{x} \in M \quad \text{and} \quad \mathbf{y}^T ((\mathbb{I} - \mathbb{P})\mathbf{x}) = 0 \quad \forall \mathbf{y} \in M. \quad (4.31)$$

It is possible to show (see Exercise 3) that a projector is orthogonal if and only if it is symmetric, that is, $\mathbb{P} = \mathbb{P}^T$, where \mathbb{P}^T denotes the transpose of \mathbb{P} defined by

$$(\mathbb{P}^T \mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbb{P}\mathbf{y}) \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \quad (4.32)$$

Here (\cdot, \cdot) denotes the Euclidean inner product in \mathbb{R}^n .

Let us now provide a relevant property of orthogonal projectors that will be exploited in the following to characterize a (Galerkin) RB problem. Given an orthogonal matrix $\mathbb{V} \in \mathbb{R}^{n \times m}$ whose columns form an orthonormal basis of $M = \text{span}(\mathbb{V}) \subset \mathbb{R}^n$, we can define an orthogonal projection operator \mathbb{P} by the matrix

$$\mathbb{P} = \mathbb{V}\mathbb{V}^T, \quad (4.33)$$

such that $\text{Ker}(\mathbb{P}) = \text{Range}(\mathbb{P})^\perp = M^\perp$; (4.33) is nothing but a particular case of relationship (4.28). It is straightforward to prove that \mathbb{P} is a projection – that is, $\mathbb{P}^2 = \mathbb{P}$ – and that it is symmetric. We underline that this representation of the orthogonal projector \mathbb{P} is not unique, since any orthonormal basis \mathbb{V} will provide a different representation.

On the other hand, the projection $\mathbb{P}\mathbf{x}$ can be characterized through the following property, which does not involve a basis (see Exercise 4 for its proof): $\mathbb{P}\mathbf{x}$ is the point in M that minimizes the distance from \mathbf{x} , that is, for any vector $\mathbf{x} \in \mathbb{R}^n$,

$$\|\mathbf{x} - \mathbb{P}\mathbf{x}\|_2 = \min_{\mathbf{y} \in M} \|\mathbf{x} - \mathbf{y}\|_2.$$

4.2.4 The Galerkin Case

We now provide a geometric interpretation of the G-RB problem in terms of orthogonal projections. To this end, we exploit the fact that the transformation matrix \mathbb{V} defined by (4.7) allows to define an orthogonal projection on the reduced subspace $\mathbf{V}_N = \text{span}\{\mathbf{v}_1, \dots, \mathbf{v}_N\}$ of \mathbb{R}^{N_h} generated by the column vectors of the matrix \mathbb{V} . Then $\dim(\mathbf{V}_N) = N$ because of the linear independence of the columns of \mathbb{V} .

For the sake of simplicity, we assume the basis functions to be orthonormal with respect to the euclidean scalar product, i.e.

$$\mathbb{V}^T \mathbb{V} = \mathbb{I}_N. \quad (4.34)$$

Proposition 4.1. *The following results hold:*

1. the matrix $\mathbb{P} = \mathbb{V}\mathbb{V}^T \in \mathbb{R}^{N_h \times N_h}$ is an orthogonal projector from the whole space \mathbb{R}^{N_h} onto the subspace \mathbf{V}_N ;
2. the matrix $\mathbb{I}_{N_h} - \mathbb{P} = \mathbb{I}_{N_h} - \mathbb{V}\mathbb{V}^T \in \mathbb{R}^{N_h \times N_h}$ is a projector from the whole space \mathbb{R}^{N_h} onto \mathbf{V}_N^\perp , the subspace of \mathbb{R}^{N_h} orthogonal to \mathbf{V}_N ;
3. the residual $\mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu})$ satisfies

$$\mathbb{P}\mathbf{r}_h(\mathbf{u}_N; \boldsymbol{\mu}) = \mathbf{0}, \quad (4.35)$$

that is, it belongs to the orthogonal space \mathbf{V}_N^\perp .

Proof. The necessary and sufficient condition for \mathbb{P} to be an orthogonal projector is that $\mathbb{P}^2 = \mathbb{P}$ and $\mathbb{P}^T = \mathbb{P}$. The latter is trivially verified, while the former is a direct consequence of the orthogonality property (4.34), as $\mathbb{P}^2 = \mathbb{V}\mathbb{V}^T\mathbb{V}\mathbb{V}^T = \mathbb{V}\mathbb{V}^T$. Moreover, for each $\mathbf{v}_h \in \mathbb{R}^{N_h}$, $\mathbb{V}^T \mathbf{v}_h$ defines a vector $\mathbf{w}_N \in \mathbb{R}^N$, so that $\mathbb{V}\mathbf{w}_N \in \mathbf{V}_N$ and therefore $\text{Range}(\mathbb{P}) = \mathbf{V}_N$. Property 2 follows from property 1, while (4.35) follows from (4.14). \square

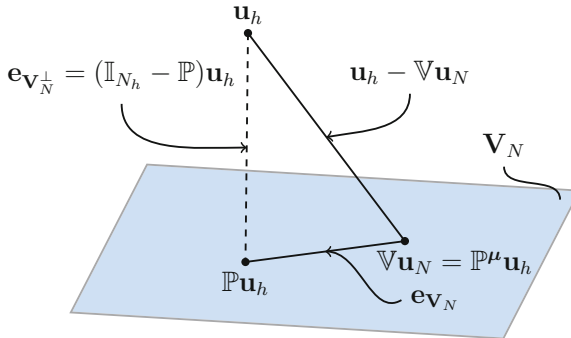


Fig. 4.1 The subspace \mathbf{V}_N of \mathbb{R}^{N_h} and the vectors $\mathbf{u}_N \in \mathbb{R}^N$, $\mathbb{V}\mathbf{u}_N \in \mathbf{V}_N$ and $\mathbf{u}_h \in \mathbb{R}^{N_h}$

If we assume the basis to be orthonormal, the error (4.10) can be additively split into two orthogonal terms (see also Fig. 4.1)

$$\begin{aligned} \mathbf{e}_h(\boldsymbol{\mu}) &= \mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}) = (\mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{P}\mathbf{u}_h(\boldsymbol{\mu})) + (\mathbb{P}\mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{V}\mathbf{u}_N(\boldsymbol{\mu})) \\ &= (\mathbb{I}_{N_h} - \mathbb{P})\mathbf{u}_h(\boldsymbol{\mu}) + \mathbb{V}(\mathbb{V}^T\mathbf{u}_h(\boldsymbol{\mu}) - \mathbf{u}_N(\boldsymbol{\mu})) = \mathbf{e}_{\mathbf{V}_N^\perp}(\boldsymbol{\mu}) + \mathbf{e}_{\mathbf{V}_N}(\boldsymbol{\mu}). \end{aligned}$$

The first term, orthogonal to \mathbf{V}_N , accounts for the fact that the high-fidelity solution does not strictly belong to the reduced subspace \mathbf{V}_N (see Fig. 4.1), whereas the second one, parallel to \mathbf{V}_N , accounts for the fact that a different problem from the original one is solved. Indeed, we have the following result.

Proposition 4.2. *The full order representation $\tilde{\mathbf{u}}_h(\boldsymbol{\mu}) = \mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}) \in \mathbb{R}^{N_h}$ of $\mathbf{u}_N \in \mathbb{R}^N$, solves the following “equivalent” high-fidelity problem,*

$$\mathbb{V}\mathbb{V}^T\mathbb{A}_h(\boldsymbol{\mu})\mathbb{V}\mathbb{V}^T\tilde{\mathbf{u}}_h = \mathbb{V}\mathbb{V}^T\mathbf{f}_h(\boldsymbol{\mu}). \quad (4.36)$$

The matrix $\mathbb{V}\mathbb{V}^T\mathbb{A}_h(\boldsymbol{\mu})\mathbb{V}\mathbb{V}^T$ has rank N and its Moore-Penrose pseudoinverse is given by

$$(\mathbb{V}\mathbb{V}^T\mathbb{A}_h(\boldsymbol{\mu})\mathbb{V}\mathbb{V}^T)^\dagger = \mathbb{V}\mathbb{A}_N^{-1}(\boldsymbol{\mu})\mathbb{V}^T. \quad (4.37)$$

Proof. Problem (4.36) can be obtained by left-multiplying equation (4.15) by \mathbb{V} and then exploiting the identity $\mathbb{I}_N = \mathbb{V}^T\mathbb{V}$, which yields

$$\mathbb{V}\mathbb{V}^T\mathbb{A}_h\mathbb{V}\mathbb{V}^T\mathbb{V}\mathbf{u}_N = \mathbb{V}\mathbb{V}^T\mathbf{f}_h.$$

Since $\text{Range}(\mathbb{P}) = \mathbf{V}_N$ and $\mathbb{V}\mathbb{V}^T\mathbb{A}_h\mathbb{V}\mathbb{V}^T = \mathbb{P}\mathbb{A}_h\mathbb{P}$, we conclude that

$$\text{Range}(\mathbb{P}\mathbb{A}_h\mathbb{P}) = \mathbf{V}_N$$

and its rank is equal to N . Finally, defining $\mathbb{B} = \mathbb{V}\mathbb{A}_N\mathbb{V}^T$ and $\mathbb{B}^\dagger = \mathbb{V}\mathbb{A}_N^{-1}\mathbb{V}^T$, we can easily check that

$$\mathbb{B}\mathbb{B}^\dagger\mathbb{B} = \mathbb{B}, \quad \mathbb{B}^\dagger\mathbb{B}\mathbb{B}^\dagger = \mathbb{B}^\dagger, \quad (\mathbb{B}\mathbb{B}^\dagger)^T = \mathbb{B}\mathbb{B}^\dagger, \quad (\mathbb{B}^\dagger\mathbb{B})^T = \mathbb{B}^\dagger\mathbb{B}, \quad (4.38)$$

i.e. \mathbb{B}^\dagger is the Moore-Penrose pseudoinverse of \mathbb{B} . \square

Remark 4.1. We remark that the results previously obtained follow from assumption (4.34) on the 2-orthogonality of the reduced basis functions. Actually, in practice the orthonormalization is usually performed with respect to the V -scalar product, realized by the symmetric positive definite matrix $\mathbb{X}_h \in \mathbb{R}^{N_h \times N_h}$ defined in (2.40). Consequently, instead of (4.34) we obtain the new orthonormality relation

$$\mathbb{Y}^T\mathbb{Y} = \mathbb{V}^T\mathbb{X}_h\mathbb{V} = \mathbb{I}_N,$$

where $\mathbb{Y}^T = \mathbb{V}^T\mathbb{X}_h^{1/2}$, $\mathbb{Y} = \mathbb{X}_h^{1/2}\mathbb{V}$. In the same way, the projection matrix is $\mathbb{P} = \mathbb{Y}\mathbb{Y}^T$. The results proven above still hold, provided \mathbb{V} is replaced by \mathbb{Y} . \bullet

Let us introduce the matrix representation of the Galerkin projector \mathbb{P}^μ defined in Sect. 7.1:

$$\mathbb{P}^\mu = \mathbb{V} \mathbb{A}_N^{-1}(\mu) \mathbb{V}^T \mathbb{A}_h(\mu) = \mathbb{V}(\mathbb{V}^T \mathbb{A}_h(\mu) \mathbb{V})^{-1} \mathbb{V}^T \mathbb{A}_h(\mu).$$

A geometric interpretation of the Galerkin orthogonality and optimal approximation properties satisfied by the G-RB problem is given by the following Proposition.

Proposition 4.3. *If the matrix $\mathbb{A}_h(\mu)$ is symmetric positive definite, then:*

1. *the matrix $\mathbb{P}^\mu \in \mathbb{R}^{N_h \times N_h}$ is an $\mathbb{A}_h(\mu)$ -orthogonal projector from the whole space \mathbb{R}^{N_h} onto the subspace \mathbf{V}_N ;*
2. *the G-RB solution \mathbf{u}_N is the best approximation in \mathbf{V}_N to \mathbf{u}_h with respect to the $\mathbb{A}_h(\mu)$ -scalar product. Moreover, the error \mathbf{e}_h satisfies*

$$\mathbf{e}_h(\mu) = (\mathbb{I}_{N_h} - \mathbb{P}^\mu) \mathbf{u}_h(\mu). \quad (4.39)$$

Proof. We first prove that \mathbb{P}^μ is a projection. We have

$$(\mathbb{P}^\mu)^2 = \mathbb{V} \mathbb{A}_N^{-1} \mathbb{V}^T \mathbb{A}_h \mathbb{V} \mathbb{A}_N^{-1} \mathbb{V}^T \mathbb{A}_h = \mathbb{V} \mathbb{A}_N^{-1} \mathbb{A}_N \mathbb{A}_N^{-1} \mathbb{V}^T \mathbb{A}_h = \mathbb{P}^\mu.$$

Then, since

$$(\mathbb{P}^\mu)^T \mathbb{A}_h = \mathbb{A}_h^T \mathbb{V} \mathbb{A}_N^{-T} \mathbb{V}^T \mathbb{A}_h = \mathbb{A}_h \mathbb{V} \mathbb{A}_N^{-1} \mathbb{V}^T \mathbb{A}_h = \mathbb{A}_h \mathbb{P}^\mu,$$

\mathbb{P}^μ is an \mathbb{A}_h -orthogonal projector. To prove Property 2, it suffices to note that

$$\mathbb{P}^\mu \mathbf{u}_h(\mu) = \mathbb{V} \mathbb{A}_N^{-1}(\mu) \mathbb{V}^T \mathbb{A}_h(\mu) \mathbf{u}_h(\mu) = \mathbb{V} \mathbb{A}_N^{-1}(\mu) \mathbb{V}^T \mathbf{f}_h(\mu) = \mathbb{V} \mathbf{u}_N(\mu). \quad \square$$

4.2.5 The Petrov-Galerkin Case

We now consider the more general case of PG-RB approximation. Let \mathbf{W}_N be the subspace of \mathbb{R}^{N_h} generated by the column vectors of the orthogonal basis \mathbb{W} .

Proposition 4.4. *If the matrix $\mathbb{W}^T \mathbb{V}$ is nonsingular (i.e. $\mathbf{W}_N \cap \mathbf{V}_N^\perp = \{\mathbf{0}\}$), the following results hold:*

1. *the matrix $\mathbb{P}_{V,W} = \mathbb{V}(\mathbb{W}^T \mathbb{V})^{-1} \mathbb{W}^T \in \mathbb{R}^{N_h \times N_h}$ is an oblique projection matrix from \mathbb{R}^{N_h} onto \mathbf{V}_N , orthogonally to \mathbf{W}_N ;*
2. *the matrix $\mathbb{I}_{N_h} - \mathbb{P}_{V,W}$ is an oblique projection matrix from the whole space \mathbb{R}^{N_h} onto the space \mathbf{W}_N^\perp , which is the subspace of \mathbb{R}^{N_h} orthogonal to \mathbf{W}_N ;*
3. *the residual $\mathbf{r}_h(\mathbf{u}_N; \mu)$ belongs to \mathbf{W}_N^\perp , that is, it satisfies*

$$\mathbb{P}_{V,W} \mathbf{r}_h(\mathbf{u}_N; \mu) = \mathbf{0}. \quad (4.40)$$

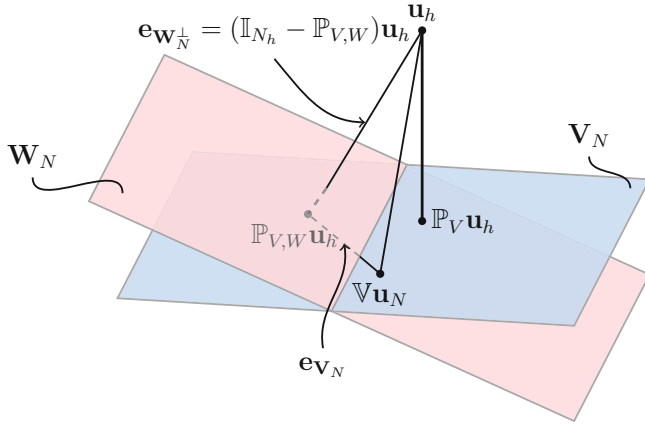


Fig. 4.2 The subspaces \mathbf{V}_N and \mathbf{W}_N of \mathbb{R}^{N_h} and the vectors $\mathbf{u}_N \in \mathbb{R}^N$, $\mathbb{V}\mathbf{u}_N \in \mathbf{V}_N$ and $\mathbf{u}_h \in \mathbb{R}^{N_h}$

In this case, the error $\mathbf{e}_h = \mathbf{u}_h - \mathbb{V}\mathbf{u}_N$ can be split as (see also Fig. 4.2)

$$\begin{aligned} \mathbf{e}_h &= \mathbf{u}_h - \mathbb{V}\mathbf{u}_N = (\mathbf{u}_h - \mathbb{P}_{V,W}\mathbf{u}_h) + (\mathbb{P}_{V,W}\mathbf{u}_h - \mathbb{V}\mathbf{u}_N) \\ &= (\mathbb{I}_{N_h} - \mathbb{P}_{V,W})\mathbf{u}_h + \mathbb{V}((\mathbb{W}^T\mathbb{V})^{-1}\mathbb{W}^T\mathbf{u}_h - \mathbf{u}_N) = \mathbf{e}_{\mathbf{W}_N^\perp} + \mathbf{e}_{\mathbf{V}_N}. \end{aligned}$$

As in the case of Galerkin projection, the first term accounts for the fact that \mathbf{u}_h does not strictly belong to \mathbf{V}_N , whereas the second one accounts for the fact that $\tilde{\mathbf{u}}_h = \mathbb{V}\mathbf{u}_N$ solves the following “equivalent” high-fidelity problem (see Exercise 7)

$$\mathbb{W}\mathbb{W}^T\mathbb{A}_h(\boldsymbol{\mu})\mathbb{V}\mathbb{V}^T\tilde{\mathbf{u}}_h = \mathbb{W}\mathbb{W}^T\mathbf{f}_h(\boldsymbol{\mu}). \quad (4.41)$$

Remark 4.2. Orthogonal and oblique projections have both played a role in the previous construction of RB methods. Orthogonal (Galerkin) projection methods correspond to the case when the subspace \mathbf{W}_N is the same as \mathbf{V}_N , while oblique (Petrov-Galerkin) projections to the one when \mathbf{W}_N is different from \mathbf{V}_N . Galerkin and least-squares projections can be encountered in many other areas of mathematics. For example, in the context of Krylov subspace methods, Galerkin projection leads to the conjugate gradient method, which is guaranteed to converge for symmetric positive definite matrices. On the other hand, the least-squares projection leads to the generalized minimum residual (GMRES) method, which is suited for more general nonsingular matrices. For a discussion of projection processes and their connection with Krylov subspace methods see [235, 26]. •

4.3 Exercises

1. Using bijection (4.2), show that

$$(\mathbf{u}_N, \mathbf{v}_N)_2 = (u_h, v_h)_N \quad \forall \mathbf{u}_N, \mathbf{v}_N \in \mathbb{R}^N \quad (\text{equivalently, } \forall u_N, v_N \in V_N).$$

2. Prove Lemma 4.1.
3. Prove that a projector is orthogonal if and only if it is symmetric, that is, if and only if $P = P^T$ in (4.32).
4. Prove that if P is an orthogonal projector onto a subspace $M \subset \mathbb{R}^n$, then for any vector $\mathbf{x} \in \mathbb{R}^n$ the following relationship holds

$$\min_{\mathbf{y} \in M} \|\mathbf{x} - \mathbf{y}\|_2 = \|\mathbf{x} - P\mathbf{x}\|_2.$$

5. Show that the Moore-Penrose pseudoinverse of $\mathbb{V}\mathbb{V}^T \mathbb{A}_h(\boldsymbol{\mu})\mathbb{V}\mathbb{V}^T$ is given by $\mathbb{V}\mathbb{A}_N^{-1}(\boldsymbol{\mu})\mathbb{V}^T$ (i.e. verify conditions (4.38) in the proof of Proposition 4.2).
6. Prove Proposition 4.4.
7. Prove that the full-order representation $\tilde{\mathbf{u}}_h = \mathbb{V}\mathbf{u}_N$ of the PG-RB problem solves the “equivalent” high-fidelity problem (4.41).

Chapter 5

The Theoretical Rationale Behind

We discuss relevant theoretical features of the RB methods seen in the previous chapters. We specifically highlight those properties of the solution manifold that are directly inherited from the parametrized differential operators. We define the Kolmogorov n -width to measure how well suited n -dimensional subspaces are to approximate the solution manifold. At the end we show that a wise selection of snapshots yields exponential convergence when approximating the solution manifold. This key property, that warrants the computational efficiency of RB methods, is provided by greedy algorithms and POD techniques used to construct RB spaces that are introduced in the subsequent chapters.

5.1 The Solution Manifold

Throughout this section we consider the problem (3.2)

$$\forall \boldsymbol{\mu} \in \mathcal{P}, \text{ find } u(\boldsymbol{\mu}) \in V : a(u(\boldsymbol{\mu}), v; \boldsymbol{\mu}) = f(v; \boldsymbol{\mu}) \quad \forall v \in V, \quad (5.1)$$

and we assume $a(\cdot, \cdot; \boldsymbol{\mu})$ to be continuous and inf-sup stable over $V \times V$ for all $\boldsymbol{\mu} \in \mathcal{P}$, and $f(\cdot; \boldsymbol{\mu}) \in V'$ to be continuous for all $\boldsymbol{\mu} \in \mathcal{P}$. We denote by

$$\varphi : \mathcal{P} \rightarrow V, \quad \boldsymbol{\mu} \mapsto u(\boldsymbol{\mu})$$

the (exact) *solution map*, and by

$$\mathcal{M} = \varphi(\mathcal{P}) = \{u(\boldsymbol{\mu}) \in V : \boldsymbol{\mu} \in \mathcal{P} \subset \mathbb{R}^P\} \quad (5.2)$$

the solution manifold (or solution set) of (5.1). For ease of notation, we will very often use u to denote the map φ ; u will therefore indicate at the same time the solution map and the solution $u(\boldsymbol{\mu})$ of problem (5.1).

As seen in the previous chapters, the goal of RB methods is to provide a numerical approximation $u_N(\boldsymbol{\mu})$ for $\boldsymbol{\mu} \in \mathcal{P}$ that is *uniform* over the entire set \mathcal{M} . More

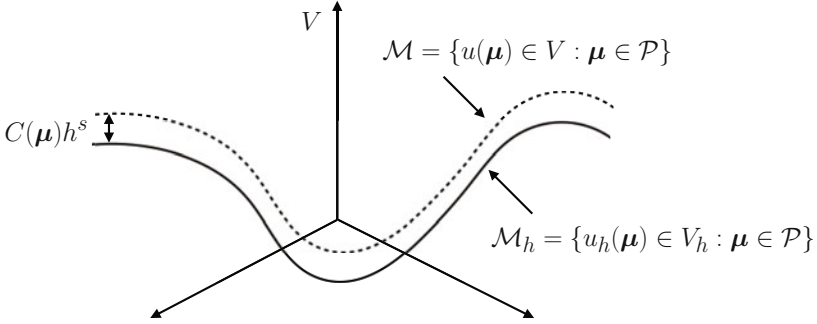


Fig. 5.1 The solution set \mathcal{M} and its high-fidelity approximation $\mathcal{M}_h \subset V_h$. Case of $P = 1$ parameter

precisely, we look for a N -dimensional subspace $V_N \subset V$ of dimension N (with N as small as possible) such that

$$\inf_{v \in V_N} \|u(\boldsymbol{\mu}) - v\|_V < \varepsilon \quad \text{for all } \boldsymbol{\mu} \in \mathcal{P}. \quad (5.3)$$

A natural question that arises is how small can we expect N to be? The answer depends, among other factors, on the smoothness of the solution map and the number Q_a, Q_f of terms appearing in the affine expansions (3.52)–(3.53) of both $a(\cdot, \cdot; \boldsymbol{\mu})$ and $f(\cdot; \boldsymbol{\mu})$. Depending on the context, in this chapter we will derive some theoretical results for either $\boldsymbol{\varphi}$ or the discrete solution map

$$\boldsymbol{\varphi}_h : \mathcal{P} \rightarrow V_h, \quad \boldsymbol{\mu} \mapsto u_h(\boldsymbol{\mu}),$$

being $u_h(\boldsymbol{\mu})$ the solution of the high-fidelity problem corresponding to (5.1), and the corresponding *discrete* manifold

$$\mathcal{M}_h = \boldsymbol{\varphi}_h(\mathcal{P}) = \{u_h(\boldsymbol{\mu}) \in V_h : \boldsymbol{\mu} \in \mathcal{P}\} \subset V_h, \quad (5.4)$$

that is the set of high-fidelity solutions generated as the input varies over the whole parameter domain \mathcal{P} . In fact, RB methods seek a low-dimensional approximation of this latter, rather than of \mathcal{M} (see Fig. 5.1 for a graphical sketch). However, several results holding for \mathcal{M} can be easily extended to \mathcal{M}_h .

We assume that \mathcal{M}_h can be made as close as desired to \mathcal{M} by choosing a suitable high-fidelity discretization technique. For instance, if (5.1) represents the weak formulation of a parametrized second-order elliptic PDE and $V_h \subset V = H^1(\Omega)$ is a conforming finite elements subspace spanned by piecewise polynomial shape functions of degree $r \geq 1$ defined on a mesh of maximum element size h , the a priori error estimate (2.67) yields

$$\|u(\boldsymbol{\mu}) - u_h(\boldsymbol{\mu})\|_V \leq C(\boldsymbol{\mu}) h^s, \quad s = \min\{r, p\}, \quad (5.5)$$

provided $u(\boldsymbol{\mu}) \in H^{p+1}(\Omega) \cap V$ for some $p \geq 1$. Thus, \mathcal{M}_h provides a good approximation to \mathcal{M} as $N_h = \dim(V_h)$ increases; however, only for quite large N_h we can expect a uniformly small (with respect to $\boldsymbol{\mu}$) error in the approximation.

5.2 When is a Problem Reducible?

We aim at stating some rigorous results about smoothness and dimensionality of the solution set, and the way these properties impact on (i) reducibility of the high-fidelity problem and (ii) approximability of the solution manifold. Let us recall that RB methods attempt at approximating the elements of the high-fidelity solution set \mathcal{M}_h by a *linear, global approximation* under the separable form

$$u_N(\mathbf{x}; \boldsymbol{\mu}) = \sum_{j=1}^N u_N^{(j)}(\boldsymbol{\mu}) \zeta_j(\mathbf{x}). \quad (5.6)$$

A similar separable form has been made when we have assumed the *affine parametric dependence* of both $a(\cdot, \cdot; \boldsymbol{\mu})$ and $f(\cdot; \boldsymbol{\mu})$, see (3.52)–(3.53).

The goal is to keep the dimension N of the RB space as small as possible. Not only, we also would like the rate at which N decreases – i.e. the minimum dimension $\bar{N} = \bar{N}(\varepsilon)$ required to obtain (5.3), as a function of ε – to be as fast as possible, and to be able to *predict* this decay rate.

Several features playing a potential role in the problem reducibility are highlighted below:

1. *local vs. global dimension*. For any given parameter value $\boldsymbol{\mu} \in \mathcal{P} \subset \mathbb{R}^P$, the *local* dimension of the solution set corresponds to the number P of parameters [108, 109]. As a matter of fact, for any $\boldsymbol{\mu}_0 \in \mathcal{P}$, the solution set can be approximated around $\boldsymbol{\mu}_0$ through the tangent hyperplane, which is spanned by P linearly independent vectors, each one corresponding to the first partial derivatives of the discrete solution map $\boldsymbol{\varphi}_h$ with respect to μ_1, \dots, μ_P . Since our goal is to approximate the whole solution set (and not just one of its elements) through a linear expansion like (5.6), a practical indication of the *global* dimension of the solution set is given by $\bar{N} = \bar{N}(\varepsilon)$, provided a sufficiently small tolerance ε has been prescribed and the approximation (5.6) satisfies (5.3) with $N = \bar{N}$;
2. *approximability* of \mathcal{M}_h by N -dimensional subspaces. This property will be expressed by the so-called Kolmogorov n -width of \mathcal{M}_h , see Sect. 5.4;
3. *differentiability and regularity of the solution map* with respect to the parameters. A remarkable case is the one occurring when the solution map is *analytic*;
4. *parametric complexity*, meaning the number of terms characterizing the affine parametric expansions (3.52)–(3.53). Parametric complexity plays an important role in those cases where the problem is nonaffine, that is, it does not fulfill the affinity assumptions (3.52)–(3.53). In these cases, the affine expansion is in principle made by an infinite number of terms; an approximated expansion

made by a finite number of terms can be computed according to suitable iterative procedures, as shown in Chap. 10. In those cases, a rapid decay of the $\boldsymbol{\mu}$ -dependent coefficients appearing in the (approximate) affine expansion or, equivalently, the existence of few dominant terms, yields a small number of basis functions in the expansion (5.6). This is true at least in the case of linear elliptic problems showing strong regularity properties on the solution map. At some extent, a separability property of the parameter-dependent coefficients of our problem is crucial to ensure the solution separability, and ultimately the reducibility of the problem. In the case of linear problems, this feature shares some analogies with the well-known linear *superposition principle*;

5. *nature of the problem.* Depending on the problem at hand and the nature of its parameters, we might deal with RB approximations which can be either quite hard to construct even if the problem depends on just $P = 1$ parameter, such as in the case of Navier-Stokes equations and other nonlinear problems, or simpler to construct even if depending on a much larger number of parameters, for instance in the pure diffusive case. In general, reduction of parametrized PDEs becomes much more difficult when passing from elliptic (or dissipative time-dependent) to purely hyperbolic problems (for instance pure transport equations).

5.3 Smoothness of the Solution Set

In this section we focus on the interplay between the smoothness of the $\boldsymbol{\mu}$ -dependent bilinear form a and linear form f and two properties of the solution manifold \mathcal{M} , namely its smoothness and compactness.

5.3.1 Continuity and Compactness

Let us assume that $a(\cdot, \cdot; \boldsymbol{\mu})$ and $f(\cdot; \boldsymbol{\mu})$ are Lipschitz-continuous with respect to $\boldsymbol{\mu}$, according to the following

Definition 5.1. The parametrized bilinear form $a : V \times V \times \mathcal{P} \rightarrow \mathbb{R}$ is Lipschitz-continuous with respect to $\boldsymbol{\mu}$ (uniformly with respect to u and v) if there exists $L_a > 0$ such that

$$|a(u, v; \boldsymbol{\mu}) - a(u, v; \boldsymbol{\mu}')| \leq L_a \|u\|_V \|v\|_V \|\boldsymbol{\mu} - \boldsymbol{\mu}'\| \quad \forall \boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathcal{P}, \forall u, v \in V.$$

Similarly, the parametrized linear form $f : V \times \mathcal{P} \rightarrow \mathbb{R}$ is Lipschitz-continuous with respect to $\boldsymbol{\mu}$ (uniformly with respect to v) if there exists $L_f > 0$ such that

$$|f(v; \boldsymbol{\mu}) - f(v; \boldsymbol{\mu}')| \leq L_f \|v\|_V \|\boldsymbol{\mu} - \boldsymbol{\mu}'\| \quad \forall \boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathcal{P}, \forall v \in V. \quad \diamond$$

Remark 5.1. Under the affine parametric dependence assumption (3.52)–(3.53), to guarantee the Lipschitz continuity with respect to $\boldsymbol{\mu}$ of $a(\cdot, \cdot; \boldsymbol{\mu})$ and $f(\cdot; \boldsymbol{\mu})$, it is sufficient to require the functions $\theta_a^q(\boldsymbol{\mu})$, $\theta_f^q(\boldsymbol{\mu})$ to be Lipschitz continuous:

$$\begin{aligned} |\theta_a^q(\boldsymbol{\mu}) - \theta_a^q(\boldsymbol{\mu}')| &\leq L_a^q \|\boldsymbol{\mu} - \boldsymbol{\mu}'\| \quad \forall q = 1, \dots, Q_a \quad \forall \boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathcal{P}, \\ |\theta_f^q(\boldsymbol{\mu}) - \theta_f^q(\boldsymbol{\mu}')| &\leq L_f^q \|\boldsymbol{\mu} - \boldsymbol{\mu}'\| \quad \forall q = 1, \dots, Q_f \quad \forall \boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathcal{P}. \end{aligned}$$

As a matter of fact, we obtain

$$\begin{aligned} |a(u, v; \boldsymbol{\mu}) - a(u, v; \boldsymbol{\mu}')| &= \left| \sum_{q=1}^{Q_a} \theta_a^q(\boldsymbol{\mu}) a_q(u, v) - \sum_{q=1}^{Q_a} \theta_a^q(\boldsymbol{\mu}') a_q(u, v) \right| \\ &= \left| \sum_{q=1}^{Q_a} [\theta_a^q(\boldsymbol{\mu}) - \theta_a^q(\boldsymbol{\mu}')] a_q(u, v) \right| \leq \underbrace{\sum_{q=1}^{Q_a} \gamma^q L_a^q}_{L_a} \|\boldsymbol{\mu} - \boldsymbol{\mu}'\| \|u\|_V \|v\|_V, \\ |f(v; \boldsymbol{\mu}) - f(v; \boldsymbol{\mu}')| &= \left| \sum_{q=1}^{Q_f} \theta_f^q(\boldsymbol{\mu}) f_q(v) - \sum_{q=1}^{Q_f} \theta_f^q(\boldsymbol{\mu}') f_q(v) \right| \\ &= \left| \sum_{q=1}^{Q_f} [\theta_f^q(\boldsymbol{\mu}) - \theta_f^q(\boldsymbol{\mu}')] f_q(v) \right| \leq \underbrace{\sum_{q=1}^{Q_f} \gamma_F^q L_f^q}_{L_f} \|\boldsymbol{\mu} - \boldsymbol{\mu}'\| \|v\|_V \end{aligned}$$

where

$$\gamma^q = \sup_{v \in V} \sup_{w \in V} \frac{a_q(v, w)}{\|v\|_V \|w\|_V}, \quad \gamma_F^q = \sup_{v \in V} \frac{f_q(v)}{\|v\|_V},$$

represent the continuity constants of the $\boldsymbol{\mu}$ -independent forms. •

Let us also assume that $a(\cdot, \cdot; \boldsymbol{\mu})$ is coercive over $V \times V$, for any $\boldsymbol{\mu} \in \mathcal{P}$. Then, we can prove the following

Proposition 5.1. *Let the bilinear form $a(\cdot, \cdot; \boldsymbol{\mu})$ and the linear form $f(\cdot; \boldsymbol{\mu})$ be Lipschitz-continuous with respect to $\boldsymbol{\mu}$. Then the solution $u(\boldsymbol{\mu})$ of problem (5.1) is Lipschitz-continuous with respect to $\boldsymbol{\mu}$, i.e., there exists $L_u > 0$ such that*

$$\|u(\boldsymbol{\mu}) - u(\boldsymbol{\mu}')\|_V \leq L_u \|\boldsymbol{\mu} - \boldsymbol{\mu}'\| \quad \forall \boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathcal{P}. \quad (5.7)$$

Proof. Let $u = u(\boldsymbol{\mu})$, $u' = u(\boldsymbol{\mu}')$ and note that

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

Subtracting these two equations and rearranging the terms we obtain

$$a(u, v; \boldsymbol{\mu}) - a(u', v; \boldsymbol{\mu}) + a(u', v; \boldsymbol{\mu}) - a(u', v; \boldsymbol{\mu}') = f(v; \boldsymbol{\mu}) - f(v; \boldsymbol{\mu}').$$

By exploiting the Lipschitz continuity of $a(\cdot, \cdot; \boldsymbol{\mu})$ and $f(\cdot; \boldsymbol{\mu})$, we have

$$\begin{aligned} a(u - u', v; \boldsymbol{\mu}) &= f(v; \boldsymbol{\mu}) - f(v; \boldsymbol{\mu}') - a(u', v; \boldsymbol{\mu}) + a(u', v; \boldsymbol{\mu}') \\ &\leq L_f \|v\|_V \|\boldsymbol{\mu} - \boldsymbol{\mu}'\| + L_a \|u'\|_V \|v\|_V \|\boldsymbol{\mu} - \boldsymbol{\mu}'\|. \end{aligned}$$

By choosing $v = u - u'$, using the coercivity property and invoking the stability estimate (3.10), we obtain

$$\beta(\boldsymbol{\mu}) \|u - u'\|_V \leq L_f(\boldsymbol{\mu}) \|\boldsymbol{\mu} - \boldsymbol{\mu}'\| + L_a \frac{\|f(\boldsymbol{\mu}')\|_{V'}}{\beta(\boldsymbol{\mu}')} \|\boldsymbol{\mu} - \boldsymbol{\mu}'\|$$

that is (5.7), with

$$L_u = \frac{1}{\beta_0} \left(L_f + L_a \frac{\tilde{\gamma}_F}{\beta_0} \right). \quad \square$$

The generalization of (5.7) to the case where $a(\cdot, \cdot; \boldsymbol{\mu})$ is inf-sup stable rather than coercive is straightforward, see Exercise 1.

Thanks to the stability estimate (2.13), we can easily see that the solution set \mathcal{M} is bounded. Moreover, provided the parametrized forms are Lipschitz-continuous with respect to $\boldsymbol{\mu}$, $\boldsymbol{\mu} \mapsto u(\boldsymbol{\mu})$ is a continuous map, that is, $u \in C^0(\mathcal{P}; V)$, being

$$C^0(\mathcal{P}; V) = \left\{ v : \mathcal{P} \rightarrow V : \boldsymbol{\mu} \mapsto v(\boldsymbol{\mu}) \text{ is continuous and } \max_{\boldsymbol{\mu} \in \mathcal{P}} \|v(\boldsymbol{\mu})\|_V < +\infty \right\}.$$

A further consequence which easily follows is the compactness of \mathcal{M} .

Corollary 5.1. *The solution set \mathcal{M} is compact in V .*

Proof. If $u : X \rightarrow Y$ is a continuous function from a compact (metric) space X to a (metric) space Y , then $f(X)$ is a compact set (see, e.g. [234, Theorem 4.14]). Since $\mathcal{P} \subset \mathbb{R}^P$ is a compact subset of \mathbb{R}^P and $\boldsymbol{\mu} \mapsto u(\boldsymbol{\mu})$ is a continuous map for any $\boldsymbol{\mu} \in \mathcal{P}$, we conclude that \mathcal{M} is a compact set in V . \square

Remark 5.2. We also allow the case of a solution set that is not locally smooth at some isolated points; this is, e.g., the case of the parametrized Helmholtz equation $\Delta u + \mu u = 0$, which has a smooth solution manifold except for μ equal to λ , an eigenvalue of the Laplace operator. In general, a reasonable requirement for the solution manifold to be fulfilled is that \mathcal{M} is *piecewise* smooth – intuitively, we admit a finite number of discontinuities. These latter might be caused by discontinuities of the parametric functions $\theta_a^q, \theta_f^q : \mathcal{P} \rightarrow \mathbb{R}$, as well as by the presence of parameter values $\hat{\boldsymbol{\mu}}$ such that $\beta(\hat{\boldsymbol{\mu}}) = 0$. In this case, we can introduce a partition $\mathcal{P} = \bigcup_{k=1}^K \mathcal{P}^k$ of the parameter space such that the parametric functions are continuous and the problem is well-posed on each \mathcal{P}^k . Then, $u \in C^0(\mathcal{P}^k; V)$ for any $k = 1, \dots, K$. See e.g. [99, 143, 162, 241] for further details about the RB approximation of parametrized Helmholtz problems. \bullet

5.3.2 Differentiability of the Solution Map and Sensitivity Equations

An additional property which can be easily shown is related with the differentiability – and, more in general, on the C^k regularity – of the solution map φ . Such a property is ensured provided the forms appearing in (5.1) enjoy similar properties with respect to $\boldsymbol{\mu}$, according to the following

Definition 5.2. The parametrized continuous bilinear form $a : V \times V \times \mathcal{P} \rightarrow \mathbb{R}$ is differentiable with respect to μ_i at the point $\boldsymbol{\mu} \in \mathcal{P}$ if, for any $u, v \in V$, the limit

$$\frac{\partial a}{\partial \mu_i}(u, v; \boldsymbol{\mu}) = \lim_{h \rightarrow 0} \frac{1}{h} (a(u, v; \boldsymbol{\mu} + h\mathbf{e}_i) - a(u, v; \boldsymbol{\mu})) \quad (5.8)$$

exists. Here $\frac{\partial a}{\partial \mu_i} : V \times V \times \mathcal{P} \rightarrow \mathbb{R}$ denotes the parametrized bilinear form corresponding to the partial derivative of $a(\cdot, \cdot; \boldsymbol{\mu})$ with respect to μ_i , evaluated at $\boldsymbol{\mu}$. \diamond

Definition 5.3. The parametrized continuous linear form $f : V \times \mathcal{P} \rightarrow \mathbb{R}$ is differentiable with respect to μ_i at the point $\boldsymbol{\mu} \in \mathcal{P}$ if, for any $v \in V$, the limit

$$\frac{\partial f}{\partial \mu_i}(v; \boldsymbol{\mu}) = \lim_{h \rightarrow 0} \frac{1}{h} (f(v; \boldsymbol{\mu} + h\mathbf{e}_i) - f(v; \boldsymbol{\mu})) \quad (5.9)$$

exists. As before, $\frac{\partial f}{\partial \mu_i} : V \times \mathcal{P} \rightarrow \mathbb{R}$ denotes the parametrized linear form corresponding to the partial derivative of $f(\cdot; \boldsymbol{\mu})$ with respect to μ_i , evaluated at $\boldsymbol{\mu}$. \diamond

Higher-order derivatives are defined by re-iterating the previous definitions.

Remark 5.3. The partial derivatives (5.8)–(5.9) are directional derivatives, to be intended in the Gâteaux sense (see Sect. A.5). In the case of affinely parametrized forms, the differentiability of $a : V \times V \times \mathcal{P} \rightarrow \mathbb{R}$ and $f : V \times \mathcal{P} \rightarrow \mathbb{R}$ with respect to μ_i , at the point $\boldsymbol{\mu}$, is automatically ensured if the functions $\theta_a^q, \theta_f^{q'} : \mathcal{P} \rightarrow \mathbb{R}$ are differentiable, for any $q = 1, \dots, Q_a, q' = 1, \dots, Q_f$, respectively. \bullet

A straightforward application of the *Implicit Function Theorem* (see e.g. [267, 66]) allows to prove the following result on the solution of problem (5.1).

Proposition 5.2. Let $a(\cdot, \cdot; \cdot) : V \times V \times \mathcal{P} \rightarrow \mathbb{R}$ and $f(\cdot; \cdot) : V \times \mathcal{P} \rightarrow \mathbb{R}$ be C^k maps with respect to $\boldsymbol{\mu}$, for some $k \geq 0$. Moreover, suppose that $a(\cdot, \cdot; \boldsymbol{\mu})$ is continuous and inf-sup stable for all $\boldsymbol{\mu} \in \mathcal{P}$, and $f(\cdot; \boldsymbol{\mu})$ is continuous for all $\boldsymbol{\mu} \in \mathcal{P}$. Then, the solution map φ is of class C^k .

Remark 5.4. If we introduce the graph $\mathcal{G}(\varphi)$ of the solution map φ , $\mathcal{G}(\varphi) = \{(\boldsymbol{\mu}, v) \in \mathbb{R}^P \times V : \boldsymbol{\mu} \in \mathcal{P} \text{ and } v = u(\boldsymbol{\mu})\}$, it can be proved (see, e.g. [108, 110, 164]) that $\mathcal{G}(\varphi)$ is a P -dimensional manifold of class C^k . \bullet

A practical way to evaluate the first-order derivative of the solution map φ is based on the use of the (first-order) sensitivity equations, according to the following result.

Proposition 5.3. *Let us assume that $a : V \times V \times \mathcal{P} \rightarrow \mathbb{R}$ and $f : V \times \mathcal{P} \rightarrow \mathbb{R}$ are differentiable with respect to μ_i at the point $\boldsymbol{\mu} \in \mathcal{P}$. Then, the solution $u(\boldsymbol{\mu})$ to problem (5.1) is differentiable with respect to μ_i . Moreover, if a and f are affinely decomposable as in (3.52)–(3.53), the partial derivatives $\partial u(\boldsymbol{\mu})/\partial \mu_i$, $i = 1, \dots, P$ of $u(\boldsymbol{\mu})$ satisfy the following sensitivity equations: for all $\boldsymbol{\mu} \in \mathcal{P}$,*

$$a\left(\frac{\partial u(\boldsymbol{\mu})}{\partial \mu_i}, v; \boldsymbol{\mu}\right) = \sum_{q'=1}^{Q_f} \frac{\partial \theta_f^{q'}(\boldsymbol{\mu})}{\partial \mu_i} f_{q'}(v) - \sum_{q=1}^{Q_a} \frac{\partial \theta_a^q(\boldsymbol{\mu})}{\partial \mu_i} a_q(u(\boldsymbol{\mu}), v) \quad \forall v \in V. \quad (5.10)$$

Proof. The differentiability of the map ϕ directly follows from Proposition 5.2. We show that in the case of affinely parametrized forms, the partial derivatives $\partial u(\boldsymbol{\mu})/\partial \mu_i$, $i = 1, \dots, P$ – to which we refer to as *parametric sensitivities* – satisfy equation (5.10), by formally deriving problem (5.1) with respect to $\boldsymbol{\mu}$. See Exercise 3 for further details in a more general case.

Let us consider problem (5.1) for $\boldsymbol{\mu} + h\mathbf{e}_i = [\mu_1, \dots, \mu_{i-1}, \mu_i + h, \mu_{i+1}, \dots, \mu_P]^T$ and assume that h is small enough to guarantee that $\boldsymbol{\mu} + h\mathbf{e}_i \in \mathcal{P}$. Then

$$\sum_{q=1}^{Q_a} \theta_a^q(\boldsymbol{\mu} + h\mathbf{e}_i) a_q(u(\boldsymbol{\mu} + h\mathbf{e}_i), v) = \sum_{q'=1}^{Q_f} \theta_f^{q'}(\boldsymbol{\mu} + h\mathbf{e}_i) f_{q'}(v) \quad \forall v \in V. \quad (5.11)$$

By subtracting problem (5.1) from (5.11) we obtain

$$\begin{aligned} \sum_{q=1}^{Q_a} \theta_a^q(\boldsymbol{\mu} + h\mathbf{e}_i) a_q(u(\boldsymbol{\mu} + h\mathbf{e}_i), v) - \sum_{q=1}^{Q_a} \theta_a^q(\boldsymbol{\mu}) a_q(u(\boldsymbol{\mu}), v) \\ = \sum_{q'=1}^{Q_f} \left(\theta_f^{q'}(\boldsymbol{\mu} + h\mathbf{e}_i) - \theta_f^{q'}(\boldsymbol{\mu}) \right) f_{q'}(v) \quad \forall v \in V. \end{aligned} \quad (5.12)$$

Dividing by h and letting $h \rightarrow 0$, we obtain at the right-hand side

$$\frac{\partial \theta_f^{q'}(\boldsymbol{\mu})}{\partial \mu_i} = \lim_{h \rightarrow 0} \frac{\theta_f^{q'}(\boldsymbol{\mu} + h\mathbf{e}_i) - \theta_f^{q'}(\boldsymbol{\mu})}{h}. \quad (5.13)$$

Similarly, by summing and subtracting $\sum_{q=1}^{Q_a} \theta_a^q(\boldsymbol{\mu}) a_q(u(\boldsymbol{\mu} + h\mathbf{e}_i), v)$ at the left-hand side and dividing by h , we obtain

$$\begin{aligned} \sum_{q=1}^{Q_a} \left(\frac{\theta_a^q(\boldsymbol{\mu} + h\mathbf{e}_i) - \theta_a^q(\boldsymbol{\mu})}{h} \right) a_q(u(\boldsymbol{\mu} + h\mathbf{e}_i), v) \\ + \sum_{q=1}^{Q_a} \theta_a^q(\boldsymbol{\mu}) a_q\left(\frac{u(\boldsymbol{\mu} + h\mathbf{e}_i) - u(\boldsymbol{\mu})}{h}, v\right). \end{aligned}$$

By taking the limit for $h \rightarrow 0$ for all $v \in V$, similarly to (5.13) we obtain in the first term

$$\lim_{h \rightarrow 0} \left(\frac{\theta_a^q(\boldsymbol{\mu} + h\mathbf{e}_i) - \theta_a^q(\boldsymbol{\mu})}{h} \right) a_q(u(\boldsymbol{\mu} + h\mathbf{e}_i), v) = \frac{\partial \theta_a^q(\boldsymbol{\mu})}{\partial \mu_i} a_q(u(\boldsymbol{\mu}), v)$$

while for the second term

$$\lim_{h \rightarrow 0} a_q \left(\frac{u(\boldsymbol{\mu} + h\mathbf{e}_i) - u(\boldsymbol{\mu})}{h}, v \right) = a_q \left(\frac{\partial u(\boldsymbol{\mu})}{\partial \mu_i}, v \right).$$

Then (5.10) easily follows. \square

In particular, we obtain that $u \in C^1(\mathcal{P}; V)$ provided that the functions θ_a^q, θ_f^q are of class $C^1(\mathcal{P})$. We remark that the parametric sensitivities of $u(\boldsymbol{\mu})$ are themselves the solution of a parametrized PDE, which shares the same bilinear form of the problem for $u(\boldsymbol{\mu})$; its right-hand side involves instead the partial derivatives of the parametrized operators, as well as the solution $u(\boldsymbol{\mu})$.

A similar result can be shown for higher order derivatives as well; the higher-order sensitivity equations still involve the bilinear form $a(\cdot, \cdot; \boldsymbol{\mu})$ at the left-hand side, whereas lower-order derivatives of both $a(\cdot, \cdot; \boldsymbol{\mu})$ and $f(\cdot; \boldsymbol{\mu})$ would appear at the right-hand side.

The sensitivity equations enable to quantify the rate of variation of φ in a neighborhood of a point $\boldsymbol{\mu}_0 \in \mathcal{P}$ in a more direct way than the Lipschitz estimate (5.7). In fact, the stability estimate (3.10) directly yields

$$\begin{aligned} \left| \frac{\partial u(\boldsymbol{\mu}_0)}{\partial \mu_i} \right|_V &\leq \frac{1}{\beta(\boldsymbol{\mu}_0)} \left(\sum_{i=1}^P \sum_{q=1}^{Q_f} \gamma_F^q \left| \frac{\partial \theta_f^q(\boldsymbol{\mu}_0)}{\partial \mu_i} \right| \right. \\ &\quad \left. + \sum_{i=1}^P \sum_{q=1}^{Q_a} \gamma^q \left| \frac{\partial \theta_a^q(\boldsymbol{\mu}_0)}{\partial \mu_i} \right| \frac{1}{\beta(\boldsymbol{\mu}_0)} \sum_{q=1}^{Q_f} \gamma_F^q \theta_f^q(\boldsymbol{\mu}_0) \right). \end{aligned} \quad (5.14)$$

A large variability of φ in a neighborhood of $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ can be due to linear and bilinear forms with parametric functions θ_f^q, θ_a^q featuring large partial derivatives, but also to a possible ill-conditioning of the problem, that is, when $\beta(\boldsymbol{\mu}_0) \approx 0$.

Generalizing the result of Proposition 5.3, we obtain that, for every $\boldsymbol{\mu} \in \mathcal{P}$, the k -th derivative of u with respect to μ_i , $k \geq 1$, denoted by $w_i^k(\boldsymbol{\mu}) = \partial^k u / \partial \mu_i^k(\boldsymbol{\mu}) \in V$, whether exists, satisfies the problem

$$a(w_i^k(\boldsymbol{\mu}), v; \boldsymbol{\mu}) = \frac{\partial^k f}{\partial \mu_i^k}(v; \boldsymbol{\mu}) - \sum_{l=1}^k \binom{k}{l} \frac{\partial^l a}{\partial \mu_i^l}(w_i^{k-l}(\boldsymbol{\mu}), v; \boldsymbol{\mu}) \quad \forall v \in V \quad (5.15)$$

for all $k = 1, \dots, K$.

Under the assumption of affinely parametrized forms, (5.15) can be more easily rewritten as

$$a\left(w_i^k(\boldsymbol{\mu}), v; \boldsymbol{\mu}\right) = \frac{\partial^k \theta_f^q(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}_i^k} f_q(v) - \sum_{l=1}^k \binom{k}{l} \frac{\partial^l \theta_a^q(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}_i^l} a_q\left(w_i^{k-l}(\boldsymbol{\mu}), v\right) \quad \forall v \in V \quad (5.16)$$

provided the functions θ_a^q, θ_f^q are of class $C^k(\mathcal{D})$.

Remark 5.5. The smoothness, compactness and differentiability results presented so far are valid for the high-fidelity solution $u_h(\boldsymbol{\mu})$ as well. •

5.4 Dimensionality of the Solution Set

Concerning the *approximability* of the solution set, we start by asking how well \mathcal{M}_h can be approximated (uniformly with respect to $\boldsymbol{\mu}$) by a finite-dimensional subspace of prescribed dimension. To answer this question, we recall the important notion of *Kolmogorov n -width* [211, 188].

Let K be a compact set of a generic Hilbert space X , and consider a generic n -dimensional subspace $X_n \subset X$. If we define the distance between an element $x \in X$ and X_n as

$$d(x; X_n) = \inf_{x_n \in X_n} \|x - x_n\|_X \quad (5.17)$$

any element $\hat{x}_n \in X_n$ which realizes the infimum, that is

$$\|x - \hat{x}_n\|_X = d(x, X_n), \quad (5.18)$$

is called the *best approximation* of x in X_n . A very natural question is whether the n -dimensional subspace is suitable to approximate all the elements $x \in K$.

To be precise, we quantify the *worst* possible best approximation as the angle between the subspace X_n and the set K , defined¹ by

$$d(K; X_n) = \sup_{x \in K} d(x; X_n). \quad (5.19)$$

The distance between a subspace X_n and K is determined by the worst-case scenario. Finding the best n -dimensional subspace of X for approximating K determines the minimum, over all possible n -dimensional subspaces of X , of the deviation (5.19), that is,

$$d_n(K; X) = \inf_{\substack{X_n \subset X \\ \dim(X_n)=n}} d(K; X_n) = \inf_{\substack{X_n \subset X \\ \dim(X_n)=n}} \sup_{x \in K} \inf_{x_n \in X_n} \|x - x_n\|_V. \quad (5.20)$$

¹ We can refer to $d(K; X_n)$ as to the *discrepancy* or *deviation* between X_n and K as well.

The number $d_n(K; X)$ is called the *Kolmogorov n -width* of K , first introduced by Kolmogorov [154]. It represents the best achievable accuracy in the V -norm when all possible elements of K are approximated by elements belonging to a linear n -dimensional subspace $X_n \subset X$. A subspace \hat{X}_n of dimension at most n such that

$$d(K; \hat{X}_n) = d_n(K; X)$$

is called an *optimal n -dimensional subspace* for $d_n(K; X)$.

Replacing X by V_h and K by \mathcal{M}_h , we can now define the Kolmogorov n -width of the solution set \mathcal{M}_h as

$$d_n(\mathcal{M}_h; V_h) = \inf_{\substack{V_n \subset V_h \\ \dim(V_n)=n}} d(\mathcal{M}_h; V_n) = \inf_{\substack{V_n \subset V_h \\ \dim(V_n)=n}} \sup_{\boldsymbol{\mu} \in \mathcal{P}} \inf_{v_n \in V_n} \|u_h(\boldsymbol{\mu}) - v_n\|_V. \quad (5.21)$$

Since V_h is a Hilbert space, there exists an orthogonal projection operator $\Pi_{V_n} : V \rightarrow V_n$ such that

$$\|v - \Pi_{V_n} v\|_V = \min_{v_n \in V_n} \|v - v_n\|_V \quad \forall v \in V_h.$$

The Kolmogorov n -width of \mathcal{M}_h can thus be expressed as

$$d_n(\mathcal{M}_h; V_h) = \inf_{\substack{V_n \subset V_h \\ \dim(V_n)=n}} \|u_h - \Pi_{V_n} u_h\|_{L^\infty(\mathcal{P}; V)}. \quad (5.22)$$

For $n = N$, (5.21) corresponds to the best achievable error in a uniform sense when approximating the solution manifold \mathcal{M}_h by elements of the RB space V_N . In this regard, the Kolmogorov n -width is relevant for deciding whether or not a given parametrized problem can be efficiently reduced. Evaluating this quantity is a hard task from a theoretical standpoint.

A possible approach to measure (5.20) would consist in embedding K into appropriate Sobolev spaces (e.g. the space $V = H_0^1(\Omega)$ in case of a Dirichlet problem for the Laplace operator). This would lead us to consider the $H^s(\Omega)$ Sobolev spatial regularity of the individual elements of \mathcal{M} (or \mathcal{M}_h) for $s > 1$. However, due to possible lack of regularity of the solution $u(\boldsymbol{\mu})$ (or $u_h(\boldsymbol{\mu})$), this can be rather small, and that would reflect into a slow decay of $d_N(\mathcal{M}; V_N)$ (or $d_N(\mathcal{M}_h; V_N)$).

A more convenient strategy is to adopt a different perspective: instead of exploiting the (low) spatial smoothness of the individual solutions we should rather exploit the fact that \mathcal{M} (or \mathcal{M}_h) is the image of the solution map, which is smooth and depends anisotropically from the parametric variables. The latter property is concerned with the spectral behavior of the affine parametric expansion of both linear and bilinear forms of our parametrized PDE.

In some cases, the n -width of the solution manifold can be directly deduced from that of the space of the parametric coefficients (say, $\theta_q^a(\boldsymbol{\mu})$, $q = 1, \dots, Q_a$ and $\theta_q^f(\boldsymbol{\mu})$, $q = 1, \dots, Q_f$) of the affine expansions (3.52)–(3.53). See Sect. 5.6 for a detailed analysis and [70] for further details.

In our analysis, we will look for upper bounds of the n -width, whose characterization might require either an estimate of the interpolation error in the parameter space, or the expansion of the solution on a fundamental basis. In the following sections we analyze these two options and the way the properties of the solution map enhance RB accuracy and computational efficiency.

5.5 Dimensionality and Analiticity

In this section we examine the interplay between the analiticity of the map φ and the dimensionality of the solution set \mathcal{M} , by considering a particular – yet relevant – case, a diffusion equation whose diffusion coefficient and source term are parameter-dependent spatial fields.

5.5.1 Analiticity of the Solution Map: an Instance

In this section we show that the solution map $\varphi : \boldsymbol{\mu} \mapsto u(\boldsymbol{\mu})$ of a parametrized diffusion problem is analytic with respect to the parameter $\boldsymbol{\mu}$, provided that the input data, regarded as $\boldsymbol{\mu}$ -dependent functions, have infinite derivatives that do not grow too fast. We recall that a function $u : X \rightarrow \mathbb{R}$ is real-analytic in an open set X if, for every $x \in X$, its power-series expansion converges to u in an open ball centered around x .

We consider a problem which can be cast under the form (5.1): for any $\boldsymbol{\mu} \in \mathcal{P}$, find $u \in V = H_0^1(\Omega)$ such that

$$\begin{cases} -\operatorname{div}(k(\mathbf{x}; \boldsymbol{\mu}) \nabla u) = s(\mathbf{x}; \boldsymbol{\mu}) & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega. \end{cases} \quad (5.23)$$

The associated parametrized forms are

$$\begin{aligned} a(u, v; \boldsymbol{\mu}) &= \int_{\Omega} k(\mathbf{x}; \boldsymbol{\mu}) \nabla u \cdot \nabla v \, d\Omega, \\ f(v; \boldsymbol{\mu}) &= \int_{\Omega} s(\mathbf{x}; \boldsymbol{\mu}) v \, d\Omega. \end{aligned} \quad (5.24)$$

If we assume that $k(\mathbf{x}; \boldsymbol{\mu})$ is positive and bounded from above, and that $k(\mathbf{x}; \boldsymbol{\mu}) \geq \bar{k}(\mathbf{x}) > 0$ a.e. in Ω , for any $\boldsymbol{\mu} \in \mathcal{P}$, then $a(\cdot, \cdot; \boldsymbol{\mu})$ is continuous and strongly coercive. Moreover, if for every $\boldsymbol{\mu} \in \mathcal{P}$, $s(\cdot; \boldsymbol{\mu}) \in L^2(\Omega)$, then $f(\cdot; \boldsymbol{\mu})$ is continuous.

We express the parameter space \mathcal{P} as the cartesian product of the one dimensional parameter spaces $\mathcal{P}_i = [\mu_i^{\min}, \mu_i^{\max}] \subset \mathbb{R}$, $i = 1, \dots, P$, that is

$$\mathcal{P} = \prod_{i=1}^P \mathcal{P}_i.$$

Moreover, we denote with \mathcal{P}_i^* the set

$$\mathcal{P}_i^* = \prod_{\substack{q=1 \\ q \neq i}}^P \mathcal{P}_q.$$

Following [20], we can show that under mild assumptions on the growth of the derivatives of the parametrized forms, the solution to problem (5.23) is analytic.

Proposition 5.4. *If for any $\mu \in \mathcal{P}$ and for any $i = 1, \dots, P$ there exists $0 < \gamma_i < +\infty$ such that*

$$\left\| \frac{1}{k(\cdot; \mu)} \frac{\partial^m k(\cdot; \mu)}{\partial \mu_i^m} \right\|_{L^\infty(\Omega)} \leq \gamma_i^m m!$$

and

$$\frac{1}{1 + \|s(\cdot; \mu)\|_{L^2(\Omega)}} \left\| \frac{\partial^m s(\cdot; \mu)}{\partial \mu_i^m} \right\|_{L^2(\Omega)} \leq \gamma_i^m m!,$$

then for every $\mu_i^* \in \mathcal{P}_i^*$ the solution $u = u(\mu_i, \mu_i^*)$, as a function of μ_i , admits an analytic extension $u(z, \mu_i^*)$ in the region of the complex plane

$$\Sigma(\mathcal{P}_i; \tau_i) = \{z \in \mathbb{C}, \|z - \mu_i\| \leq \tau_i \forall \mu_i \in \mathcal{P}_i\}$$

with $0 < \tau_i < 2\gamma_i$. Moreover, for all $z \in \Sigma(\mathcal{P}_i; \tau_i)$

$$\|u(z)\|_{C^0(\mathcal{P}_i^*, V)} \leq \frac{C_P}{k(1 - 2\tau_i \gamma_i)} (2\|f\|_{C^0(\mathcal{P}; V)} + 1) \quad (5.25)$$

being C_P the Poincaré constant.

Proof. We perform a one-dimensional analysis in each direction μ_i , $i = 1, \dots, P$. For every $\mu \in \mathcal{P}$, $w_i^m(\mu)$, the m -th partial derivative of $u(\mu)$ with respect to μ_i , satisfies the problem

$$a(w_i^m(\mu), v; \mu) = \int_{\Omega} \frac{\partial^m s(\cdot; \mu)}{\partial \mu_i^m} v d\Omega - \sum_{l=1}^m \binom{m}{l} \frac{\partial^l a}{\partial \mu_i^l} (w_i^{m-l}(\mu), v; \mu) \quad \forall v \in V \quad (5.26)$$

which directly follows from (5.15).

Thanks to the coercivity of $a(\cdot, \cdot; \boldsymbol{\mu})$ and Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} \|\sqrt{k(\cdot; \boldsymbol{\mu})} \nabla w_i^m(\boldsymbol{\mu})\|_{L^2(\Omega)} &\leq \frac{C_P}{\sqrt{k}} \left\| \frac{\partial^m f}{\partial \boldsymbol{\mu}_i^m} \right\|_{L^2(\Omega)} \\ &\quad + \sum_{l=1}^m \binom{m}{l} \left\| \frac{\partial^l k(\cdot; \boldsymbol{\mu})}{\partial \boldsymbol{\mu}_i^l} \frac{1}{k(\cdot; \boldsymbol{\mu})} \right\|_{L^\infty(\Omega)} \|\sqrt{k(\cdot; \boldsymbol{\mu})} w_i^{m-l}(\boldsymbol{\mu})\|_{L^2(\Omega)}. \end{aligned}$$

By defining

$$R_m = \frac{1}{m!} \|\sqrt{k(\cdot; \boldsymbol{\mu})} \nabla w_i^m(\boldsymbol{\mu})\|_{L^2(\Omega)}$$

and taking advantage of the bounds on the derivatives of k and s , we obtain the following (recursive) inequalities

$$\begin{aligned} R_m &\leq \sum_{l=1}^m \gamma_l' R_{m-1} + \frac{C_P}{\sqrt{k}} \gamma_l' (1 + \|s(\cdot; \boldsymbol{\mu})\|_{L^2(\Omega)}) \\ &\leq \frac{1}{2} (2\gamma_l)^l \left(R_0 + \frac{C_P}{\sqrt{k}} (1 + \|s(\cdot; \boldsymbol{\mu})\|_{L^2(\Omega)}) \right). \end{aligned}$$

By definition of R_k and the stability estimate for $u(\boldsymbol{\mu})$ provided by the Lax-Milgram lemma, we obtain

$$R_0 = \|\sqrt{k(\cdot; \boldsymbol{\mu})} \nabla u(\boldsymbol{\mu})\|_{L^2(\Omega)} \leq \frac{C_P}{\sqrt{k}} \|s(\cdot; \boldsymbol{\mu})\|_{L^2(\Omega)}$$

so that

$$\frac{1}{m!} \|\nabla w_i^m(\boldsymbol{\mu})\|_{L^2(\Omega)} \leq \frac{R_m}{\sqrt{k}}$$

and finally

$$\frac{1}{m!} \|\nabla w_i^m(\boldsymbol{\mu})\|_{L^2(\Omega)} \leq \frac{C_P}{k} \left(2\|s(\cdot; \boldsymbol{\mu})\|_{L^2(\Omega)} + 1 \right) (2\gamma_l)^k. \quad (5.27)$$

Let us now define, for any $\boldsymbol{\mu}_i^* \in \mathcal{P}_i^*$, the power series $u : \mathbb{C} \rightarrow C^0(\mathcal{P}_i^*, V)$ given by

$$u(\mathbf{x}; z, \boldsymbol{\mu}_i^*) = \sum_{m=0}^{\infty} \frac{(z - \mu_i)^m}{m!} \frac{\partial^m}{\partial \mu_i^m} u(\mathbf{x}; \mu_i, \boldsymbol{\mu}_i^*).$$

Thanks to the bound (5.27) we have

$$\begin{aligned} \|u(z)\|_{C^0(\mathcal{P}_i^*; V)} &\leq \sum_{m=0}^{\infty} \frac{|z - \mu_i|^m}{m!} \left\| \frac{\partial^m}{\partial \mu_i^m} u(\cdot; \mu_i, \cdot) \right\|_{C^0(\mathcal{P}_i^*; V)} \\ &\leq \frac{C_P}{k} \max_{\mu_i \in \mathcal{P}_i} \left(2\|s(\cdot; \mu_i, \boldsymbol{\mu}_i^*)\|_{C^0(\mathcal{P}_i^*; L^2(\Omega))} + 1 \right) \sum_{m=0}^{\infty} (|z - \mu_i| 2\gamma_l)^m \\ &\leq \frac{C_P}{k} \left(2\|s\|_{C^0(\Gamma; L^2(\Omega))} + 1 \right) \sum_{m=0}^{\infty} (|z - \mu_i| 2\gamma_l)^m. \end{aligned}$$

For all $z \in \mathbb{C}$ such that $\|z - \mu_i\| \leq \tau_i < 1/(2\gamma_i)$ the series converges and

$$\|u(z)\|_{C^0(\mathcal{P}_i^*; V)} \leq \frac{C_P}{\bar{k}(1 - 2\tau_i\gamma_i)} (2\|f\|_{C^0(\mathcal{P}; V)} + 1).$$

The power series converges for every $\mu_i \in \mathcal{P}_i$; by a continuation argument, we can extend u analytically on the whole disc $\Sigma(\mathcal{P}_i; \tau_i)$, and (5.25) holds. \square

Remark 5.6. This analiticity result holds for the high-fidelity solution $u_h(\mu)$, too. \bullet

5.5.2 Kolmogorov n -width and Analiticity

We investigate the relation between the analiticity of the solution manifold and its Kolmogorov n -width. The connection between the (rapid) convergence of RB methods and smoothness in parameters has been first discussed in [6], and stands at the basis of *a priori* convergence of RB methods. Here we provide a general result on this connection, thanks to a suitable interpolation procedure relying on orthogonal polynomials over the parameter space \mathcal{P} . Similarly to what is done to obtain *a priori* error estimates for the finite element method, we bound the approximation error in terms of the interpolation error. Then, an estimation of this latter enables to obtain the desired *a priori* result; see e.g. [216, Chap. 4] for further details.

To cast the parametrized problem (5.23) in this framework, we first need to introduce a polynomial space and a set of interpolation nodes in every parameter direction. Let us consider the space

$$Y_h^m = Q_m(\mathcal{P}) \times V_h$$

where $Q_m(\mathcal{P}) \subset L^2(\mathcal{P})$ is the span of tensor product polynomials with degree at most $m = (m_1, \dots, m_P)$, being m_j the degree with respect to the j -th parameter component. More specifically,

$$Q_m(\mathcal{P}) = \bigotimes_{j=1}^P Q_{m_j}(\mathcal{P}_j)$$

where for any component $j = 1, \dots, P$,

$$Q_{m_j}(\mathcal{P}_j) = \text{span}\{\mu_j^k, k = 0, \dots, m_j\}$$

is the space of polynomials of degree up to m_j (with respect to the parameter μ_j). The dimension of $Q_m(\mathcal{P})$ is

$$\dim(Q_m(\mathcal{P})) = \prod_{j=1}^P (m_j + 1) = n.$$

For each dimension $j = 1, \dots, P$, let us denote by $\{\mu_{j,k_j}, k_j = 1, \dots, m_j + 1\}$ the $m_j + 1$ roots of the polynomial q_{m_j+1} belonging to the family of orthogonal polynomials with respect to the weight ρ_j . See Sect. A.8 for further details about relevant families of orthogonal polynomials in approximation theory. Note that

$$\int_{\mathcal{P}_j} q_{m_j+1}(\mu) w(\mu) \rho_j(\mu) d\mu = 0 \quad \forall w \in \mathcal{Q}_{m_j}(\mathcal{P}_j).$$

For each $j = 1, \dots, P$ we also denote by $\{l_{j,i}\}_{i=1}^{m_j+1}$ the Lagrange basis of the space \mathcal{Q}_{m_j} ,

$$l_{j,i} \in \mathcal{Q}_{m_j}(\mathcal{P}_j), \quad l_{j,i}(\mu_{j,k}) = \delta_{jk}, \quad j, k = 1, \dots, m_j + 1.$$

Then, we set

$$l_k(\mu) = \prod_{j=1}^P l_{j,k_j}(\mu_j), \quad k = 1, \dots, n$$

where to any set of indices $[k_1, \dots, k_P]$ we can associate the global index

$$k = k_1 + m_1(k_2 - 1) + m_1 m_2(k_3 - 1) + \dots$$

and we denote by $\mu_k = (\mu_{1,k_1}, \mu_{2,k_2}, \dots, \mu_{P,k_P}) \in \mathcal{P}$. It is now possible to define the n -dimensional approximation space

$$V_h^m = \text{span}\{u_h(\mu_k), k = 1, \dots, n\}, \quad (5.28)$$

where we seek an approximate solution $u_{h,m}(\mu)$ to the following high-fidelity problem (notations from (5.24)): find $u_h(\mu) \in V_h \subset V$ such that

$$a(u_h(\mu), v_h; \mu) = f(v_h; \mu) \quad \forall v_h \in V_h. \quad (5.29)$$

To this end, we collocate (5.29) at the zeros – here denoted for simplicity by $\{\mu_1, \dots, \mu_n\}$ – of orthogonal polynomials (such as Chebyshev or Legendre polynomials) and build the reduced solution $u_{h,m}(\mu) \in V_h^m$ by interpolating the collocated solutions at μ . We thus look for an approximation to $u_h(\mu)$ in the form (see (A.29))

$$u_{h,m}(\mu) = \sum_{k=1}^n u_h(\mu_k) l_k(\mu)$$

where $u_h(\mu_k) \in V_h$ is the high-fidelity solution obtained for $\mu = \mu_k$. Equivalently, we can express the solution $u_{h,m}(\mu)$ by introducing the Lagrange interpolation operator $\mathcal{I}_m^\mu : C^0(\mathcal{P}; V_h) \rightarrow \mathcal{Q}_m(\mathcal{P}; V_h)$, so that

$$\mathcal{I}_m^\mu v(\mathbf{x}; \mu) = \sum_{k=1}^n v(\mathbf{x}; \mu_k) l_k(\mu).$$

Hence, $u_{h,m}(\mathbf{x}; \mu) = \mathcal{I}_m^\mu u_h(\mathbf{x}; \mu)$. We use this notation to highlight the fact that interpolation applies to the \mathbf{x} variable and the result is then regarded as a function of

μ . If $\mathcal{P} \subset \mathbb{R}$, it is possible to prove (see, e.g. [20, Lemma 4.4]) the following result concerning the best approximation error in $C^0(\mathcal{P}; V_h)$.

Lemma 5.1. *Given a function $v \in C^0(\mathcal{P}; V_h)$ which admits an analytic extension in the region $\Sigma(\mathcal{P}; \tau) = \{z \in \mathbb{C} : \|z - \mu\| \leq \tau \ \forall \mu \in \mathcal{P}\}$ for some $\tau > 0$, it holds that*

$$\min_{w \in \mathcal{Q}_m \otimes V_h} \|v - w\|_{C^0(\mathcal{P}; V)} \leq \frac{2}{\rho - 1} e^{-m \ln(\rho)} \max_{\mu \in \Sigma(\mathcal{P}; \tau)} \|v(\mu)\|_V$$

where

$$1 < \rho = 2 \frac{\tau}{|\mathcal{P}|} + \sqrt{1 + 4 \frac{\tau^2}{|\mathcal{P}|^2}}.$$

Then, by employing a one-dimensional argument (similarly to what we have done in Sect. 5.5.1), it is possible to prove the following convergence result, showing that the error decays exponentially fast with respect to m under the assumption of analiticity of the solution manifold \mathcal{M}_h stated in Sect. 5.5.1.

The interested reader can refer to [20, Theorem 4.1] for the proof; a similar result can also be found in [71].

Theorem 5.1. *Under the assumptions of Proposition (5.4), there exist positive constants*

$$r_j = \ln \left(2 \frac{\tau_j}{|\mathcal{P}_j|} + \sqrt{1 + 4 \frac{\tau_j^2}{|\mathcal{P}_j|^2}} \right), \quad j = 1, \dots, P$$

and a further constant C , independent of h and m , such that the error between the solutions $u_{h,m}(\mu) \in V_h^m$ and $u_h(\mu) \in V_h$ to problem (5.29) is bounded as follows

$$\|u_h - u_{h,m}\|_{C^0(\mathcal{P}; V)} = \|u_h - \mathcal{I}_m^\mu u_h\|_{C^0(\mathcal{P}; V)} \leq C \sum_{j=1}^P e^{-r_j m_j}. \quad (5.30)$$

Here τ_j is smaller than the distance between \mathcal{P}_j and the nearest singularity in the complex plane, as defined in Proposition 5.4.

It is now possible to exploit the results derived so far to provide a bound to the Kolmogorov n -width of the solution manifold \mathcal{M}_h when this latter is analytic. Starting from relation (5.21) and choosing the space V_h^m as particular instance of n -dimensional space, thanks to Theorem 5.1 we obtain (see (5.21)):

$$\begin{aligned} d_n(\mathcal{M}_h; V_h) &= \inf_{\substack{V_n \subset V_h \\ \dim(V_n)=n}} d(\mathcal{M}_h; V_n) \leq \sup_{\mu \in \mathcal{P}} \|u_h(\cdot; \mu) - \mathcal{I}_m^\mu u_h(\cdot; \mu)\|_V \\ &\leq \|u_h - \mathcal{I}_m^\mu u_h\|_{C^0(\mathcal{P}; V)} \leq C \sum_{j=1}^P e^{-r_j m_j}. \end{aligned} \quad (5.31)$$

We conclude by noticing that if the solution $u_h(\boldsymbol{\mu})$ depends analytically on the parameter $\boldsymbol{\mu}$ – that is, the solution map φ_h is analytic – the Kolmogorov n -width of the solution manifold decays exponentially, following a similar decay of the interpolation error obtained when dealing with families of orthogonal polynomials.

Indeed, we highlight that the exponential convergence of numerical approximations is often linked to a spectral argument. In this respect, RB methods can be understood as spectral methods, where instead of generic global polynomial functions, we use problem-dependent global smooth approximation basis functions².

In general, the RB approximation of solutions of elliptic equations with regular coefficients has indeed been very successful. We warn however the reader that the analytic regularity of the solution manifold \mathcal{M}_h is *not* necessary in order to successfully apply the RB method.

Remark 5.7. The previous result extends to the case of a more general parametrized elliptic problem an *a priori* exponential convergence result established in [175, 174] for the case of elliptic PDEs (3.2) that satisfy the following assumptions: $f(\cdot) \in V'$ is $\boldsymbol{\mu}$ -independent,

$$a(w, v; \boldsymbol{\mu}) = a_0(w, v) + \boldsymbol{\mu} a_1(w, v) \quad \forall w, v \in V, \quad (5.32)$$

$\boldsymbol{\mu} \in \mathcal{P} = [0, \mu_M]$, $\mu_M > 0$, the bilinear forms $a_0 : V \times V \rightarrow \mathbb{R}$ and $a_1 : V \times V \rightarrow \mathbb{R}$ are symmetric, continuous and positive semi-definite, and a_0 is coercive. •

A question which naturally arises is about the possibility to construct RB spaces for a large class of problems through general algorithms rather than on *ad hoc* sampling strategies like the one addressed in this section. This generalization, however, should not jeopardize the achievement of an exponential convergence with respect to the dimension of the corresponding RB space when approximating the high-fidelity manifold \mathcal{M}_h . As we will see in the two following chapters, two very general techniques – namely, proper orthogonal decomposition and greedy algorithms – meet this goal.

5.6 Kolmogorov n -width and Parametric Complexity

We now investigate the link between the Kolmogorov n -width and the *parametric complexity*, which, for affinely parametrized problems, is defined by the couple $\{Q_a, Q_f\}$, being Q_a and Q_f the numbers of terms appearing in the affine expansions (3.52)–(3.53), respectively. Together with parametric regularity of operators, parametric complexity is indeed the other feature enabling a reduced-order approximation of the parametrized problem.

² It is worthy to highlight that a few cornerstone results for the analysis of spectral element methods in the Eighties [207, 180, 176, 29] were made by scientists who then gave fundamental contributions to the early development of RB methods in the early 2000s.

Let us consider the $\boldsymbol{\mu}$ -dependent problem (5.1) where (see Exercise 4 for the more general case $Q_a > 2$)

$$\begin{aligned} a(u, v; \boldsymbol{\mu}) &= \theta_a^1(\boldsymbol{\mu})a_1(u, v) + \theta_a^2(\boldsymbol{\mu})a_2(u, v), \\ f(v; \boldsymbol{\mu}) &= \sum_{q=1}^{Q_f} \theta_f^q(\boldsymbol{\mu})f_q(v), \end{aligned} \quad (5.33)$$

and seek a solution for any $\boldsymbol{\mu} \in \mathcal{P}$ under a parametrically separable form

$$u_h(\boldsymbol{\mu}) = \sum_{k=0}^{\infty} \theta_k(\boldsymbol{\mu})l_k, \quad (5.34)$$

where the functions l_k do not depend on $\boldsymbol{\mu}$. The expansion (5.34), together with standard estimates for convergent power series, enables to find an abstract approximation space which is uniformly exponentially convergent over the entire parameter range \mathcal{P} .

For the sake of simplicity, we express the high-fidelity approximation of problem (5.1), (5.33) in the following algebraic form³

$$\mathbb{A}(\boldsymbol{\mu})\mathbf{u}_h(\boldsymbol{\mu}) = (\theta_a^1(\boldsymbol{\mu})\mathbb{A}_1 + \theta_a^2(\boldsymbol{\mu})\mathbb{A}_2)\mathbf{u}_h(\boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \theta_f^q(\boldsymbol{\mu})\mathbf{f}_q. \quad (5.35)$$

We introduce the following $(\mathbb{X}_h, \mathbb{X}_h^{-1})$ matrix norm (which realizes the $\mathcal{L}(V_h, V_h')$ norm)

$$\|\mathbb{B}\|_{\mathbb{X}_h, \mathbb{X}_h^{-1}} = \sup_{\mathbf{v} \in \mathbb{R}^{N_h}} \frac{\|\mathbb{B}\mathbf{v}\|_{\mathbb{X}_h^{-1}}}{\|\mathbf{v}\|_{\mathbb{X}_h}} = \sup_{\mathbf{v} \in \mathbb{R}^{N_h}} \frac{\|\mathbb{X}_h^{-1/2}\mathbb{B}\mathbb{X}_h^{-1/2}\mathbf{v}\|_2}{\|\mathbf{v}\|_2} \quad \forall \mathbb{B} \in \mathbb{R}^{N_h \times N_h}. \quad (5.36)$$

We also recall that the spectral radius $\rho(\mathbb{B})$ of a matrix \mathbb{B} is defined as the maximum among the moduli of its eigenvalues. If $\rho(\mathbb{B}) < 1$, then

$$\sum_{k=0}^{\infty} \mathbb{B}^k = (\mathbb{I} - \mathbb{B})^{-1}. \quad (5.37)$$

Let us now assume that (i) the matrix \mathbb{A}_1 is invertible and (ii) the problem satisfies a global condition for the spectral radius ρ being

$$\rho\left(\frac{\theta_a^2(\boldsymbol{\mu})}{\theta_a^1(\boldsymbol{\mu})}\mathbb{A}_1^{-1}\mathbb{A}_2\right) < 1 \quad \forall \boldsymbol{\mu} \in \mathcal{P}, \quad (5.38)$$

which we interpret by saying that the term $\theta_a^1(\boldsymbol{\mu})\mathbb{A}_1$ dominates the complete matrix $\mathbb{A}(\boldsymbol{\mu})$. Let us remark that a problem such as (5.35) can arise for example from the

³ For the sake of readability, throughout this section we omit the subscript h to high-fidelity matrices and right-hand side vectors.

approximation of advection-diffusion or reaction-diffusion problems, as those we introduced in Sect. 2.1, where \mathbb{A}_1 contains the (dominant) diffusion operator and \mathbb{A}_2 contains all the other terms.

Formally, the solution of this problem can be written as

$$\mathbf{u}_h(\boldsymbol{\mu}) = \left(\mathbb{I} + \frac{\theta_a^2(\boldsymbol{\mu})}{\theta_a^1(\boldsymbol{\mu})} \mathbb{A}_1^{-1} \mathbb{A}_2 \right)^{-1} (\theta_a^1(\boldsymbol{\mu}) \mathbb{A}_1)^{-1} \left(\sum_{q=1}^{Q_f} \theta_f^q(\boldsymbol{\mu}) \mathbf{f}_q \right),$$

from which, thanks to (5.38) and (5.37), we obtain the series expansion for the solution

$$\mathbf{u}_h(\boldsymbol{\mu}) = \sum_{k=0}^{\infty} \sum_{q=1}^{Q_f} \frac{(-1)^k (\theta_a^2(\boldsymbol{\mu}))^k \theta_f^q(\boldsymbol{\mu})}{(\theta_a^1(\boldsymbol{\mu}))^{k+1}} (\mathbb{A}_1^{-1} \mathbb{A}_2)^k \mathbb{A}_1^{-1} \mathbf{f}_q.$$

Let us define the fundamental basis vectors

$$\boldsymbol{\psi}_{k,q} = (\mathbb{A}_1^{-1} \mathbb{A}_2)^k \mathbb{A}_1^{-1} \mathbf{f}_q \in \mathbb{R}^{N_h}, \quad q = 1, \dots, Q_f, \quad k = 0, 1, \dots,$$

which can be computed by the following iterative procedure

$$\boldsymbol{\psi}_{0,q} = \mathbb{A}_1^{-1} \mathbf{f}_q, \quad \boldsymbol{\psi}_{k+1,q} = \mathbb{A}_1^{-1} \mathbb{A}_2 \boldsymbol{\psi}_{k,q}, \quad k \geq 0, \quad \text{for all } q = 1, \dots, Q_f.$$

Accordingly, we denote by

$$V_n^\psi = \text{span} \left\{ \boldsymbol{\psi}_{k,q} : k = 0, \dots, m-1, q = 1, \dots, Q_f \right\} \subset \mathbb{R}^{N_h}$$

the n -dimensional subspace given by the first n fundamental basis vectors, being $n = Q_f \cdot m$. Hence, we can express the solution of problem (5.35) through the following series

$$\mathbf{u}_h(\boldsymbol{\mu}) = \sum_{k=0}^{\infty} \sum_{q=1}^{Q_f} \frac{(-1)^k (\theta_a^2(\boldsymbol{\mu}))^k \theta_f^q(\boldsymbol{\mu})}{(\theta_a^1(\boldsymbol{\mu}))^{k+1}} \boldsymbol{\psi}_{k,q}. \quad (5.39)$$

We can draw the following conclusions from formula (5.39):

1. in the special case $\mathbb{A}_2 = 0$ the parametric dependence enters only through the right hand side. Hence the series (5.39) is in fact a finite sum

$$\mathbf{u}_h(\boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \frac{\theta_f^q(\boldsymbol{\mu})}{\theta_a^1(\boldsymbol{\mu})} \boldsymbol{\psi}_{0,q}. \quad (5.40)$$

The Q_f dimensional subspace $\text{span}\{\boldsymbol{\psi}_{0,1}, \dots, \boldsymbol{\psi}_{0,Q_f}\}$ will provide the exact solution to problem (5.35) for any $\boldsymbol{\mu} \in \mathcal{P}$. Thus, the dimension of the high-fidelity solution set \mathcal{M}_h is bounded by the number Q_f of terms giving the affine representation of the problem data.

Hence, a finite-dimensional RB space of dimension $N \approx Q_f$ obtained by sampling the high-fidelity manifold will provide a very good approximation in this case; however, a slightly larger number of basis functions is needed in princi-

ple since we rely on the solution of problem (5.35) at selected parameter points instead than computing the fundamental basis vectors $\{\boldsymbol{\Psi}_{0,1}, \dots, \boldsymbol{\Psi}_{0,Q_f}\}$;

2. if the decay of the series coefficients in (5.39) is fast, the solutions $\mathbf{u}_h(\boldsymbol{\mu})$ can be well approximated by a handful of fundamental basis functions $\boldsymbol{\Psi}_{k,q}$, $k = 0, 1, \dots$ and $q = 1, \dots, Q_f$.

In general, the fundamental basis functions $\boldsymbol{\Psi}_{k,q}$ are obtained by a nonlinear combination of solutions to problem (5.35), hence they do not yield a RB space according to the procedure illustrated in Sect. 3.3. They are, however, useful for estimating the m -width of the solution set, according to the following

Lemma 5.2. *Let us assume that, for any $q = 1, \dots, Q_f$, there exists a positive sequence $\{\gamma_{k,q}\}_{k=1}^{\infty}$ such that*

$$\|\boldsymbol{\Psi}_{k,q}\|_{\mathbb{X}_h} \leq \gamma_{k,q} \quad \forall q = 1, \dots, Q_f.$$

Then, the n -width of the high-fidelity solution set \mathcal{M}_h is bounded as follows

$$d_n(\mathcal{M}_h; V_h) \leq \sup_{\boldsymbol{\mu} \in \mathcal{P}} \sum_{k=n}^{\infty} \sum_{q=1}^{Q_f} \left| \frac{(\theta_a^2(\boldsymbol{\mu}))^k \theta_f^q(\boldsymbol{\mu})}{(\theta_a^1(\boldsymbol{\mu}))^{k+1}} \right| \gamma_{k,q}. \quad (5.41)$$

Proof. By using the definition (5.21) and formula (5.39), we obtain

$$\begin{aligned} d_n(\mathcal{M}_h; V_h) &= \inf_{V_n \subset V_h} \sup_{\boldsymbol{\mu} \in \mathcal{P}} \inf_{\mathbf{u}_n \in V_n} \|\mathbf{u}_h(\boldsymbol{\mu}) - \mathbf{u}_n\|_{\mathbb{X}_h} \leq \sup_{\boldsymbol{\mu} \in \mathcal{P}} \inf_{\mathbf{u}_n \in V_n^\Psi} \|\mathbf{u}_h(\boldsymbol{\mu}) - \mathbf{u}_n\|_{\mathbb{X}_h} \\ &\leq \sup_{\boldsymbol{\mu} \in \mathcal{P}} \left\| \mathbf{u}_h(\boldsymbol{\mu}) - \sum_{k=0}^{n-1} \sum_{q=1}^{Q_f} \frac{(-1)^k (\theta_a^2(\boldsymbol{\mu}))^k \theta_f^q(\boldsymbol{\mu})}{(\theta_a^1(\boldsymbol{\mu}))^{k+1}} \boldsymbol{\Psi}_{k,q} \right\|_{\mathbb{X}_h} \\ &= \sup_{\boldsymbol{\mu} \in \mathcal{P}} \left\| \sum_{k=n}^{\infty} \sum_{q=1}^{Q_f} \frac{(-1)^k (\theta_a^2(\boldsymbol{\mu}))^k \theta_f^q(\boldsymbol{\mu})}{(\theta_a^1(\boldsymbol{\mu}))^{k+1}} \boldsymbol{\Psi}_{k,q} \right\|_{\mathbb{X}_h} \\ &\leq \sup_{\boldsymbol{\mu} \in \mathcal{P}} \sum_{k=n}^{\infty} \sum_{q=1}^{Q_f} \left| \frac{(\theta_a^2(\boldsymbol{\mu}))^k \theta_f^q(\boldsymbol{\mu})}{(\theta_a^1(\boldsymbol{\mu}))^{k+1}} \right| \|\boldsymbol{\Psi}_{k,q}\|_{\mathbb{X}_h}. \end{aligned}$$

□

Remark 5.8. The estimate (5.41) is independent of the number of parameters P and the regularity of the coefficient functions $\theta_a^1, \theta_a^2: \mathcal{P} \rightarrow \mathbb{R}$. The latter can be taken, for the sake of this result, in $L^\infty(\mathcal{P})$. •

Thanks to estimate (5.41) we can show an exponential convergence result for problem (5.35).

Theorem 5.2. Assume that

$$\exists \varepsilon \in (0, 1) \quad \text{such that} \quad \left| \frac{\theta_a^2(\boldsymbol{\mu})}{\theta_a^1(\boldsymbol{\mu})} \right| \leq \frac{1 - \varepsilon}{\|\mathbb{A}_1^{-1} \mathbb{A}_2\|_{\mathbb{X}_h, \mathbb{X}_h^{-1}}} \quad \forall \boldsymbol{\mu} \in \mathcal{P}. \quad (5.42)$$

Then the n -width of the solution set \mathcal{M}_h of (5.35) converges exponentially, i.e.

$$\exists C, \alpha > 0 \quad \text{such that} \quad d_n(\mathcal{M}_h; V_h) \leq C e^{-\alpha n}. \quad (5.43)$$

Proof. By using the upper bound (5.41), for $n = m \cdot Q_f$, we obtain

$$\begin{aligned} d_n(\mathcal{M}_h; V_h) &\leq \sup_{\boldsymbol{\mu} \in \mathcal{P}} \sum_{k=n}^{\infty} \sum_{q=1}^{Q_f} \left| \frac{(\theta_a^2(\boldsymbol{\mu}))^k \theta_f^q(\boldsymbol{\mu})}{(\theta_a^1(\boldsymbol{\mu}))^{k+1}} \right| \|(\mathbb{A}_1^{-1} \mathbb{A}_2)^k \mathbb{A}_1^{-1} \mathbf{f}_q\|_{\mathbb{X}_h} \\ &\leq Q_f \cdot \sup_{\boldsymbol{\mu}, q} \left\{ \left| \frac{\theta_f^q(\boldsymbol{\mu})}{\theta_a^1(\boldsymbol{\mu})} \right| \|\mathbb{A}_1^{-1} \mathbf{f}_q\|_{\mathbb{X}_h} \right\} \cdot \sum_{k=n}^{\infty} \left| \frac{(\theta_a^2(\boldsymbol{\mu}))^k}{(\theta_a^1(\boldsymbol{\mu}))^k} \right| \|\mathbb{A}_1^{-1} \mathbb{A}_2\|_{\mathbb{X}_h, \mathbb{X}_h^{-1}}^k \\ &= Q_f \cdot \sup_{\boldsymbol{\mu}, q} \left\{ \left| \frac{\theta_f^q(\boldsymbol{\mu})}{\theta_a^1(\boldsymbol{\mu})} \right| \|\mathbb{A}_1^{-1} \mathbf{f}_q\|_{\mathbb{X}_h} \right\} \cdot (1 - \varepsilon)^m \sum_{k=0}^{\infty} (1 - \varepsilon)^k \\ &= \frac{Q_f}{\varepsilon} \cdot \sup_{\boldsymbol{\mu}, q} \left\{ \left| \frac{\theta_f^q(\boldsymbol{\mu})}{\theta_a^1(\boldsymbol{\mu})} \right| \|\mathbb{A}_1^{-1} \mathbf{f}_q\|_{\mathbb{X}_h} \right\} \cdot \exp\left(\frac{\log(1 - \varepsilon)}{Q_f} n\right). \end{aligned}$$

We can now obtain the estimate (5.43) by setting $\alpha = -\log(1 - \varepsilon)/Q_f$ and

$$C = \frac{Q_f}{\varepsilon} \cdot \sup_{\boldsymbol{\mu}, q} \left\{ \left| \frac{\theta_f^q(\boldsymbol{\mu})}{\theta_a^1(\boldsymbol{\mu})} \right| \|\mathbb{A}_1^{-1} \mathbf{f}_q\|_{\mathbb{X}_h} \right\}. \quad \square$$

An RB Petrov-Galerkin approximation of problem (5.35), reads (see (4.23))

$$\mathbb{W}^T (\theta_a^1(\boldsymbol{\mu}) \mathbb{A}_1 + \theta_a^2(\boldsymbol{\mu}) \mathbb{A}_2) \mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \theta_f^q(\boldsymbol{\mu}) \mathbb{W}^T \mathbf{f}_q.$$

A representation formula similar to (5.39) holds for \mathbf{u}_N too

$$\mathbf{u}_N(\boldsymbol{\mu}) = \sum_{k=0}^{\infty} \sum_{q=1}^{Q_f} \frac{(-1)^k (\theta_a^2(\boldsymbol{\mu}))^k \theta_f^q(\boldsymbol{\mu})}{(\theta_a^1(\boldsymbol{\mu}))^{k+1}} \boldsymbol{\psi}_{k,q}^N.$$

The reduced fundamental basis vectors $\boldsymbol{\psi}_{k,q}^N$ are now given by

$$\boldsymbol{\psi}_{0,q}^N = (\mathbb{W}^T \mathbb{A}_1 \mathbb{V})^{-1} \mathbb{W}^T \mathbf{f}_q, \quad \boldsymbol{\psi}_{k+1,q}^N = (\mathbb{W}^T \mathbb{A}_1 \mathbb{V})^{-1} \mathbb{W}^T \mathbb{A}_2 \mathbb{V} \boldsymbol{\psi}_{k,q}^N, \quad k \geq 0.$$

Correspondingly, we obtain the following error representation formula

$$\mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}) = \sum_{k=0}^{\infty} \sum_{q=1}^{Q_f} \frac{(-1)^k (\theta_a^2(\boldsymbol{\mu}))^k \theta_f^q(\boldsymbol{\mu})}{(\theta_a^1(\boldsymbol{\mu}))^{k+1}} \left(\boldsymbol{\psi}_{k,q} - \mathbb{V}\boldsymbol{\psi}_{k,q}^N \right).$$

To estimate the error we need to establish how fast $\|\boldsymbol{\psi}_{k,q} - \mathbb{V}\boldsymbol{\psi}_{k,q}^N\|_{\mathbb{X}_h}$ tend to zero with respect to N , for all k and q .

Remark 5.9. Even if the global spectral condition (5.38) does not hold, we can try to expand the solution locally, i.e. in different subsets⁴ of the parameter domain around different $\boldsymbol{\mu}^*$ and obtain *local* (with respect to $\boldsymbol{\mu}$) RB basis functions. This would hopefully allow the solution to be expanded over the whole spatial domain, but for a subset of parameter values only. In this case, if $\mathcal{P}^1, \dots, \mathcal{P}^M$ provide a partition of the original parameter domain \mathcal{P} into M subsets, the local spectral condition would require that in each subdomain \mathcal{P}^m

$$\exists i(m) : \quad \rho \left(\frac{\theta_a^{j(m)}(\boldsymbol{\mu})}{\theta_a^{i(m)}(\boldsymbol{\mu})} \mathbb{A}_{i(m)}^{-1} \mathbb{A}_{j(m)} \right) < 1 \quad \text{for all } \boldsymbol{\mu} \in \mathcal{P}^m, \quad \text{for } j(m) \neq i(m),$$

i.e., in each parameter subset \mathcal{P}^m there is a dominant term \mathbb{A}_q (even if the dominant part of the operator can change from subset to subset). If such a local spectral condition holds, the previous results can be extended to prove the existence of local exponentially convergent approximation spaces also in this case. •

5.7 Lagrange, Taylor and Hermite RB Spaces

Before closing this chapter, we comment on possible alternative strategies to generate a RB space once a given sample of values has been selected from the parameter space. In the previous section we have considered a first example of realization of RB space: starting from a set (3.19) of N selected parameter values, the RB space is obtained by evaluating the corresponding high-fidelity solutions $\{u_h(\boldsymbol{\mu}^i)\}_{i=1}^N$, or *snapshots*. Provided they are linearly independent, we define the RB space $V_N \subset V_h$ as in (3.22). Very often, the Gram-Schmidt orthonormalization procedure is then applied to the set of snapshots in order to end up with an orthonormal set of basis functions, see Sect. 7.1.2 for a detailed description.

This is by far the most common way of constructing reduced subspaces, and we call $\{u_h(\boldsymbol{\mu}^1), \dots, u_h(\boldsymbol{\mu}^N)\}$ a *Lagrangian reduced basis*. As we have seen, in this case we approximate $u_N(\boldsymbol{\mu})$ by a linear combination of N snapshots. In the following chapters, we will present two powerful techniques to generate such a set of reduced basis functions.

⁴ This leads to the so-called *hp* reduced basis method [101], where different reduced bases (analogous to *p*-refinement in FE methods) are constructed in different subsets of the parameter domain (analogous to *h*-refinement in FE methods).

A possible alternative consists in evaluating both the field and the first-order parametric sensitivities at a set of parameter points in \mathcal{P} by solving (5.10), and define

$$V_{NP}^{Hermite} = \text{span} \left\{ u_h(\boldsymbol{\mu}^i), \frac{\partial u_h}{\partial \mu_j}(\boldsymbol{\mu}^i), \quad j = 1, \dots, P, \boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^N \in S_N \right\},$$

yielding a RB space of dimension NP . This is the so-called *Hermite* RB space.

A third option is given by the *Taylor* RB space, which is defined as the span of the field $u_h(\bar{\boldsymbol{\mu}})$ and the parametric sensitivities up to a given order K , evaluated at a chosen parameter point $\boldsymbol{\mu} = \bar{\boldsymbol{\mu}}$. For the case of $P = 1$ parameters, this becomes

$$V_{N,\bar{\boldsymbol{\mu}}}^{Taylor} = \text{span} \left\{ \frac{\partial^k u_h}{\partial \mu^k}(\bar{\boldsymbol{\mu}}), \quad k = 0, \dots, N-1 \right\},$$

yielding a RB space of dimension N ; a similar definition can be given in the case of $P > 1$ parameters, by warning that its dimension is growing fast as soon as the number P increases.

Smoothness with respect to parameters is thus essential to define a Taylor RB space, which approximates the elements of the manifold \mathcal{M}_h under the form

$$u_N(\boldsymbol{\mu}) = u_h(\boldsymbol{\mu}_0) + \nabla_{\boldsymbol{\mu}} u_h(\boldsymbol{\mu}_0)^T (\boldsymbol{\mu} - \boldsymbol{\mu}_0) + \nabla_{\boldsymbol{\mu}}^2 u_h(\boldsymbol{\mu}_0)^T (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^2 + \dots$$

A Taylor RB space provides a locally good approximation around the chosen parameter value $\boldsymbol{\mu} = \bar{\boldsymbol{\mu}}$; the smaller the parametric sensitivities, the higher the accuracy. The magnitude of parametric sensitivities can in principle be controlled thanks to estimates under the form (5.14). Whenever the parameter space is suitably explored for the sake of snapshots' selection, a Lagrangian reduced basis should provide a better approximation to the solution manifold.

Pioneering theoretical works in the Eighties about RB methods – see, e.g., [203, 109, 213] – focused on Taylor RB spaces, involving higher-order derivatives of the solution with respect to parameters, evaluated around a parameter value $\boldsymbol{\mu} = \bar{\boldsymbol{\mu}}$. Noor [201] used a Taylor basis to build a local reduced space for tracing the post-buckling behavior of a nonlinear structure.

The continuation idea was used also by Peterson [210] to compute Navier-Stokes solutions with increasing Reynolds number in the case of a flow over a forward facing step. Again a Taylor basis was constructed and used to extrapolate an initial guess for the Newton method at a slightly higher Reynolds number.

Ito and Ravindran [147] were apparently the first ones to use a Hermite basis in a uniform approximation context, rather than for a pure continuation method. The Lagrange and Hermite bases were compared on a driven cavity problem, where the Hermite approach was somewhat superior. No stability problems were reported and the Hermite basis with only two basis functions was able to extrapolate solutions to much larger Reynolds numbers.

In the works of Hay et al. [130, 131] sensitivity information was introduced into the proper orthogonal decomposition framework. The parametric sensitivities of the

POD modes were derived and computed. The test problems were related with channel flow around a cylindrical obstacle, either by using a simple parametrization as the Reynolds number, or a more involved geometric parametrization of the obstacle. The use of a Hermite ROM considerably improved the validity of the reduced solutions away from the parametric snapshots. However, in the more involved geometrical parametrization case the Hermite basis completely failed, as it did not converge to the exact solution even when the number of POD modes was increased.

Carlberg and Farhat [56] proposed an approach called “compact POD”, based on goal-oriented Petrov-Galerkin projection to minimize the approximation error subject to a chosen output criteria. They included sensitivity information with proper weighting coming from the Taylor-expansion and “mollification” of basis functions far away from the snapshot parameter. This strategy was then successfully applied to the optimization of an aeroelastic wing configuration by building local basis functions along the path to the optimal wing configuration.

5.8 Exercises

1. Show Proposition 5.1 in the case $a(\cdot, \cdot; \boldsymbol{\mu})$ is inf-sup stable over $V \times V$, for any $\boldsymbol{\mu} \in \mathcal{P}$.
2. Let $u \in V$ be the solution of the problem

$$\int_{\Omega} \mathbf{v} \nabla u \cdot \nabla v d\Omega = \int_{\Omega} f v d\Omega \quad \forall v \in V \quad (5.44)$$

where $\mathbf{v} = \mathbf{v}(\mathbf{x}) \in L^{\infty}(\Omega)$ is a prescribed conductivity field.

- a. Under the assumption that $0 < m \leq \mathbf{v}(\mathbf{x}) \leq M < \infty$ for any $\mathbf{x} \in \Omega$, show that problem(5.44) admits a unique solution;
- b. if u_0 and u_1 are solutions of (5.44) with the same right-hand side f and with coefficients $\mathbf{v} = \mathbf{v}_0$ and $\mathbf{v} = \mathbf{v}_1$, respectively, and if both these coefficients satisfy the assumption made at the previous point, show that

$$\|u_1 - u_0\|_V \leq \frac{\|f\|_{V'}}{m^2} \|\mathbf{v}_1 - \mathbf{v}_0\|_{L^{\infty}(\Omega)}.$$

3. Let $u = u(\boldsymbol{\mu}) \in V$ be the solution of problem (5.44) where now $v(\mathbf{x}, \boldsymbol{\mu})$ is a parametrized diffusion coefficient, given by

$$v(x; \boldsymbol{\mu}) = \sum_{j=1}^P \mu_j \psi_j(\mathbf{x}); \quad (5.45)$$

moreover, define the set (for $\delta, R : 0 < \delta < 2R$)

$$\mathcal{P}_\delta = \{\boldsymbol{\mu} \in \mathcal{R}^P : \delta \leq |v(\mathbf{x}; \boldsymbol{\mu})| \leq 2R \ \forall \mathbf{x} \in \Omega\} \subset \mathcal{P}.$$

We want to show that the solution map admits partial derivatives $\partial u(\boldsymbol{\mu})/\partial \mu_j \in V$ with respect to each variable μ_1, \dots, μ_P , and that this derivative is the solution of the following problem: for any $\boldsymbol{\mu} \in \mathcal{P}_\delta$, find $\partial u(\boldsymbol{\mu})/\partial \mu_j \in V$ such that

$$\int_{\Omega} v(\mathbf{x}; \boldsymbol{\mu}) \nabla \frac{\partial u(\boldsymbol{\mu})}{\partial \mu_j} \cdot \nabla v d\Omega = - \int_{\Omega} \psi_j \nabla u(\boldsymbol{\mu}) \cdot \nabla v d\Omega \quad \forall v \in V; \quad (5.46)$$

for the sake of simplicity, let us fix $j \geq 1$ from now on. To prove this result, let us consider the following steps.

- a. Show that

$$\frac{\delta}{2} \leq |v(\mathbf{x}; \boldsymbol{\mu} + h\mathbf{e}_i)| \leq 2R + \frac{\delta}{2} \quad \forall \mathbf{x} \in \Omega;$$

- b. show that, if $\boldsymbol{\mu} \in \mathcal{P}_\delta$,

$$\begin{aligned} \|u(\boldsymbol{\mu} + h\mathbf{e}_i) - u(\boldsymbol{\mu})\|_V &= \|\nabla u(\boldsymbol{\mu} + h\mathbf{e}_i) - \nabla u(\boldsymbol{\mu})\|_{L^2(\Omega)} \\ &\leq |h| \|\psi_j\|_{L^\infty(\Omega)} \frac{4\|f\|_{V'}}{\delta^2}; \end{aligned}$$

- c. show that, for $h \in \mathbb{R} \setminus \{0\}$,

$$w_h(\boldsymbol{\mu}) = \frac{u(\boldsymbol{\mu} + h\mathbf{e}_i) - u(\boldsymbol{\mu})}{h}$$

is the unique solution to

$$\int_{\Omega} v(\mathbf{x}; \boldsymbol{\mu}) \nabla w_h(\boldsymbol{\mu}) \cdot \nabla v d\Omega = L_h(v) \quad \forall v \in V$$

where $L_h : v \rightarrow L_h(v) = - \int_{\Omega} \nabla u(\boldsymbol{\mu} + h\mathbf{e}_i) \cdot \nabla v d\Omega$. Prove that L_h is a continuous linear functional on V and that L_h converges towards L_0 in V' as $h \rightarrow 0$;

- d. show that $w_h \rightarrow w_0$ in V , where $w_0 \in V$ is the solution to the problem

$$\int_{\Omega} v(\mathbf{x}; \boldsymbol{\mu}) \nabla w_0(\boldsymbol{\mu}) \cdot \nabla v d\Omega = L_0(v) \quad \forall v \in V.$$

Then, conclude that $w_0 = \partial u(\boldsymbol{\mu})/\partial \mu_j$ is the unique solution to (5.46);

- e. generalize the result of points b., c. and d. to the case where also the right-hand side $f = f(\mathbf{x}; \boldsymbol{\mu})$ depends on the parameter vector $\boldsymbol{\mu} \in \mathcal{P}$;
- f. the assumption (5.45) guarantees that the bilinear form $a(\cdot, \cdot; \boldsymbol{\mu})$ features an affine dependence. Generalize the results of points b., c. and d. to the case of a general affine expansion of the form

$$a(u, v; \boldsymbol{\mu}) = \sum_{q=1}^{Q_a} \theta_q^a(\boldsymbol{\mu}) a_q(u, v).$$

4. Referring to the analysis of Sect. 5.6, consider the more general case of a parametrized bilinear form consisting of $Q_a > 2$ terms, yielding the following problem

$$\left(\theta_a^1(\boldsymbol{\mu}) \mathbb{A}_1 + \sum_{r=2}^{Q_a} \theta_a^r(\boldsymbol{\mu}) \mathbb{A}_r \right) \mathbf{u}_h(\boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \theta_f^q(\boldsymbol{\mu}) \mathbf{f}_q. \quad (5.47)$$

- a. Show that, under the global spectral condition

$$\rho \left(\sum_{r=2}^{Q_a} \frac{\theta_a^r(\boldsymbol{\mu})}{\theta_a^1(\boldsymbol{\mu})} \mathbb{A}_1^{-1} \mathbb{A}_r \right) < 1, \quad (5.48)$$

and using (5.48), the solution to (5.47) can be expressed as

$$\mathbf{u}_h(\boldsymbol{\mu}) = \sum_{k=0}^{\infty} \sum_{q=1}^{Q_f} \frac{(-1)^k \theta_f^q(\boldsymbol{\mu})}{[\theta_a^1(\boldsymbol{\mu})]^{k+1}} \boldsymbol{\psi}_{k,q}(\boldsymbol{\mu}); \quad (5.49)$$

- b. derive the following recursive definition of the fundamental basis vectors (which depend explicitly on the parameter $\boldsymbol{\mu}$):

$$\boldsymbol{\psi}_{0,q} = \mathbb{A}_1^{-1} \mathbf{f}_q, \quad \boldsymbol{\psi}_{k+1,q}(\boldsymbol{\mu}) = \left[\sum_{r=2}^{Q_a} \theta_a^r(\boldsymbol{\mu}) \mathbb{A}_1^{-1} \mathbb{A}_r \right] \boldsymbol{\psi}_{k,q}(\boldsymbol{\mu}) \quad (5.50)$$

for $k \geq 0$, for all $q = 1, \dots, Q_f$;

- c. by introducing a multi-index $\boldsymbol{\rho}^{(k)} = (\rho_1, \rho_2, \dots, \rho_k)$ of dimension k and letting $\boldsymbol{\rho}^{(0)} = \emptyset$, define a set of parameter-free basis functions $\boldsymbol{\phi}_{k,q,\boldsymbol{\rho}}$ according to the following recursive relations:

$$\boldsymbol{\phi}_{0,q,\boldsymbol{\rho}^{(0)}} = \mathbb{A}_1^{-1} \mathbf{f}_q, \quad \boldsymbol{\phi}_{k+1,q,\boldsymbol{\rho}^{(k+1)}} = \mathbb{A}_1^{-1} \mathbb{A}_{\rho_{k+1}} \boldsymbol{\phi}_{k,q,\boldsymbol{\rho}^{(k)}};$$

Using the parameter-free basis, rewrite the fundamental basis vectors defined in (5.50) in a $\boldsymbol{\mu}$ -independent form.

- d. how many terms does the k -th level expansion for $\boldsymbol{\psi}_{k,q}$ contain? Under which assumption on the (magnitude of the) coefficients $\theta_a^q(\boldsymbol{\mu})$, $\theta_f^q(\boldsymbol{\mu})$ is it possible to derive exponentially decaying n -width estimates in the case $Q_a \gg 1$, otherwise infeasible?

Chapter 6

Construction of RB Spaces by SVD-POD

After recalling basic notions about the singular value decomposition of a matrix, we address the so-called *proper orthogonal decomposition* (POD), a classical, global approach of matrix analysis used in a broad variety of mathematical contexts. We elaborate on the use of POD to generate a RB space and, finally, we make an excursion on the continuous analogue of POD.

6.1 Basic Notions on Singular Value Decomposition

The *singular value decomposition* (SVD) of a matrix $\mathbb{A} \in \mathbb{R}^{m \times n}$ plays a prominent role in numerical linear algebra and data analysis. After reviewing its most relevant properties, we emphasize its connection with low-rank approximations and data compression, focusing on those results which are useful for the derivation and analysis of the POD algorithm. For more on SVD see, e.g., [117, 250, 68].

SVD is a diagonalization process involving left and right multiplication by orthogonal matrices. More precisely (see Fig. 6.1 and 6.2):

if $\mathbb{A} \in \mathbb{R}^{m \times n}$ is a real matrix, there exist two orthogonal matrices

$$\mathbb{U} = [\boldsymbol{\zeta}_1 \mid \dots \mid \boldsymbol{\zeta}_m] \in \mathbb{R}^{m \times m}, \quad \mathbb{Z} = [\boldsymbol{\psi}_1 \mid \dots \mid \boldsymbol{\psi}_n] \in \mathbb{R}^{n \times n}$$

such that

$$\mathbb{A} = \mathbb{U} \boldsymbol{\Sigma} \mathbb{Z}^T, \quad \text{with } \boldsymbol{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_p) \in \mathbb{R}^{m \times n} \quad (6.1)$$

and $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p \geq 0$, for $p = \min(m, n)$.

The matrix factorization (6.1) is called singular value decomposition (SVD) of \mathbb{A} and the numbers $\sigma_i = \sigma_i(\mathbb{A})$ are called *singular values* of \mathbb{A} . $\boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_m$ are called *left singular vectors* of \mathbb{A} , whereas $\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_n$ *right singular vectors* of \mathbb{A} , as

$$\mathbb{A} \boldsymbol{\psi}_j = \sigma_j \boldsymbol{\zeta}_j, \quad \mathbb{A}^T \boldsymbol{\zeta}_j = \sigma_j \boldsymbol{\psi}_j, \quad i, j = 1, \dots, n.$$

$$\boxed{\mathbb{A}} = \boxed{\mathbb{U}} \boxed{\Sigma} \boxed{\mathbb{Z}^T}$$

Fig. 6.1 SVD decomposition of a matrix $\mathbb{A} \in \mathbb{R}^{m \times n}$ with $m \geq n$

$$\boxed{\mathbb{A}} = \boxed{\mathbb{U}} \boxed{\Sigma} \boxed{\mathbb{Z}^T}$$

Fig. 6.2 SVD decomposition of a matrix $\mathbb{A} \in \mathbb{R}^{m \times n}$ with $m \leq n$

Not only, (6.1) implies the spectral decompositions

$$\mathbb{A}\mathbb{A}^T = \mathbb{U}\Sigma\Sigma^T\mathbb{U}^T, \quad \mathbb{A}^T\mathbb{A} = \mathbb{Z}\Sigma^T\Sigma\mathbb{Z}^T$$

with

$$\Sigma\Sigma^T = \text{diag}(\sigma_1^2, \dots, \sigma_p^2, \underbrace{0, \dots, 0}_{m-p \text{ times}}), \quad \Sigma^T\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_p^2, \underbrace{0, \dots, 0}_{n-p \text{ times}}).$$

Since $\mathbb{A}\mathbb{A}^T$ and $\mathbb{A}^T\mathbb{A}$ are symmetric matrices, the left (resp. right) singular vectors of \mathbb{A} turn out to be the eigenvectors of $\mathbb{A}\mathbb{A}^T$ (resp. $\mathbb{A}^T\mathbb{A}$), see Exercise 1. Indeed, there is a very close relationship between the SVD of \mathbb{A} and the eigenvalue problems for $\mathbb{A}^T\mathbb{A}$ and $\mathbb{A}\mathbb{A}^T$, since

$$\sigma_i(\mathbb{A}) = \sqrt{\lambda_i(\mathbb{A}^T\mathbb{A})}, \quad i = 1, \dots, p.$$

The largest and the smallest singular values are also denoted by

$$\sigma_{\max} = \max_{i=1, \dots, p} \sigma_i, \quad \sigma_{\min} = \min_{i=1, \dots, p} \sigma_i.$$

Remark 6.1. If $A \in \mathbb{R}^{n \times n}$ is a symmetric matrix, then $\sigma_i(\mathbb{A}) = |\lambda_i(\mathbb{A})|$, with $\lambda_1(\mathbb{A}) \geq \lambda_2(\mathbb{A}) \geq \dots \geq \lambda_n(\mathbb{A})$ being the eigenvalues of \mathbb{A} (see Exercise 2). •

The singular values of a matrix are related to both its norm and its condition number. In fact, for any matrix $\mathbb{A} \in \mathbb{R}^{m \times n}$,

$$\|\mathbb{A}\|_2 = \sigma_{\max}, \quad \|\mathbb{A}\|_F = \sqrt{\sum_{i=1}^p \sigma_i^2}. \quad (6.2)$$

where the *Frobenius norm* is defined as

$$\|\mathbb{A}\|_F = \left(\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2 \right)^{1/2}. \quad (6.3)$$

Moreover, if $\mathbb{A} \in \mathbb{R}^{n \times n}$ is nonsingular, by inverting relation (6.1) we obtain

$$\mathbb{A}^{-1} = \mathbb{Z}\Sigma^{-1}\mathbb{U}^T,$$

with $\Sigma^{-1} = \text{diag}(\sigma_1^{-1}, \dots, \sigma_n^{-1})$. This shows that σ_n^{-1} is the largest singular value of \mathbb{A}^{-1} , whence

$$\|\mathbb{A}^{-1}\|_2 = \frac{1}{\sigma_n}, \quad \kappa(\mathbb{A}) = \|\mathbb{A}\|_2 \|\mathbb{A}^{-1}\|_2 = \frac{\sigma_1}{\sigma_n}.$$

Remark 6.2. SVD features a remarkable geometric interpretation. Let us consider the hyper-ellipsoid $E = \{\mathbb{A}\mathbf{x} : \mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_2 = 1\} \subset \mathbb{R}^m$ obtained as the image of the unitary sphere in \mathbb{R}^n . Then its semiaxes have lengths equal to the singular values of \mathbb{A} and directions equal to those of the singular vectors of \mathbb{A} . •

Let r denote the rank of a matrix $\mathbb{A} \in \mathbb{R}^{m \times n}$. The matrix

$$\mathbb{A}^\dagger = \mathbb{Z}\Sigma^\dagger\mathbb{U}^T, \quad \text{where} \quad \Sigma^\dagger = \text{diag}(\sigma_1^{-1}, \dots, \sigma_r^{-1}, 0, \dots, 0)$$

is called the *Moore-Penrose pseudo-inverse*, or *generalized inverse*, of \mathbb{A} . We point out that

$$\mathbb{A}^\dagger = (\mathbb{A}^T \mathbb{A})^{-1} \mathbb{A}^T \quad \text{if } \text{rank}(\mathbb{A}) = n < m$$

and that

$$\mathbb{A}^\dagger = \mathbb{A}^{-1} \quad \text{if } \text{rank}(\mathbb{A}) = n = m.$$

6.1.1 SVD and Low-Rank Approximations

A compact characterization of the rank of a matrix is possible thanks to SVD. Since $\text{rank}(\mathbb{A}) = \text{rank}(\Sigma)$ and the rank of a diagonal matrix is equal to the number of its nonzero diagonal entries, if $\mathbb{A} \in \mathbb{R}^{m \times n}$ has r positive singular values, then $\text{rank}(\mathbb{A}) = r$. Not only, we can provide an orthonormal basis for both the kernel and the range of \mathbb{A} as follows:

$$\text{Ker}(\mathbb{A}) = \text{span}\{\boldsymbol{\psi}_{r+1}, \dots, \boldsymbol{\psi}_n\}, \quad \text{Range}(\mathbb{A}) = \text{span}\{\boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_r\}.$$

Another peculiar feature of SVD is that, if $\mathbb{A} \in \mathbb{R}^{m \times n}$ has rank equal to r , then it can be written as the sum of r rank-1 matrices

$$\mathbb{A} = \sum_{i=1}^r \sigma_i \boldsymbol{\zeta}_i \boldsymbol{\psi}_i^T.$$

This formula is very useful to compute low-rank approximations of a matrix because, thanks to properties (6.2), the partial sum of $k \leq r$ terms captures as much of the *energy* of the matrix \mathbb{A} as possible. Here by *energy* we mean either the 2-norm or the Frobenius norm of a matrix.

The precise formulation of this property is provided in the next theorem. This is indeed a cornerstone in the computation of the *best approximation* of a given matrix \mathbb{A} by matrices of lower rank, and stands at the basis of the POD algorithm for the construction of a reduced basis space.

Theorem 6.1 (Schmidt-Eckart-Young). *Given a matrix $\mathbb{A} \in \mathbb{R}^{m \times n}$ of rank r , the matrix*

$$\mathbb{A}_k = \sum_{i=1}^k \sigma_i \boldsymbol{\zeta}_i \boldsymbol{\psi}_i^T, \quad 0 \leq k \leq r, \quad (6.4)$$

satisfies the optimality property

$$\|\mathbb{A} - \mathbb{A}_k\|_F = \min_{\substack{\mathbb{B} \in \mathbb{R}^{m \times n} \\ \text{rank}(\mathbb{B}) \leq k}} \|\mathbb{A} - \mathbb{B}\|_F = \sqrt{\sum_{i=k+1}^r \sigma_i^2}. \quad (6.5)$$

If $k = r$ the sum in (6.5) is null.

A similar result holds by considering the 2-norm instead of the Frobenius norm: for any $0 < k \leq r$, the matrix \mathbb{A}_k defined in (6.4) is also such that

$$\|\mathbb{A} - \mathbb{A}_k\|_2 = \min_{\substack{\mathbb{B} \in \mathbb{R}^{m \times n} \\ \text{rank}(\mathbb{B}) \leq k}} \|\mathbb{A} - \mathbb{B}\|_2 = \sigma_{k+1}. \quad (6.6)$$

As noted in [244, 245], Theorem (6.1) can be regarded as the finite-dimensional version of a more general approximation theorem of 1907 for integral operators due to E. Schmidt (see [239]), which was later rediscovered by C. Eckart and G. Young in 1936 [96] (and for this reason it is sometimes improperly called the Eckart-Young theorem). In 1960, L. Mirsky generalized this theorem to all unitarily invariant norms [191]. For a proof of (6.5) we refer to [244], while a proof of the optimality with respect to the euclidean norm can be found, e.g., in [117, 250].

In RB construction the SVD of a matrix $\mathbb{A} \in \mathbb{R}^{m \times n}$, $m \geq n$, is often realized by picking a rectangular left matrix $\mathbb{U}_1 \in \mathbb{R}^{m \times n}$ instead of a squared matrix $\mathbb{U} \in \mathbb{R}^{m \times m}$. In such a case, instead of (6.1) we obtain

$$\mathbb{A} = \mathbb{U}_1 \boldsymbol{\Sigma}_1 \boldsymbol{\mathbb{Z}}^T,$$

with $\boldsymbol{\Sigma}_1 \in \mathbb{R}^{n \times n}$. This represents the so-called *thin* SVD. In this case

$$\mathbb{U}_1 = \mathbb{U}(:, 1:n) = [\boldsymbol{\zeta}_1 \mid \dots \mid \boldsymbol{\zeta}_n] \in \mathbb{R}^{m \times n}$$

(see also Fig. 6.3) and

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}(1:n, 1:n) = \text{diag}(\sigma_1, \dots, \sigma_n) \in \mathbb{R}^{n \times n}.$$

$$\mathbb{A} = \begin{bmatrix} \text{shaded } U_1 & \text{white} \end{bmatrix} \begin{bmatrix} \text{shaded } \Sigma_1 & \text{white} \end{bmatrix} \begin{bmatrix} \text{shaded } Z^T \end{bmatrix}$$

Fig. 6.3 Thin SVD

In the remainder, the use of the thin (rather than the classical) SVD will result from the context, depending upon the dimensions of the matrices appearing in the factorization.

6.2 Interlude

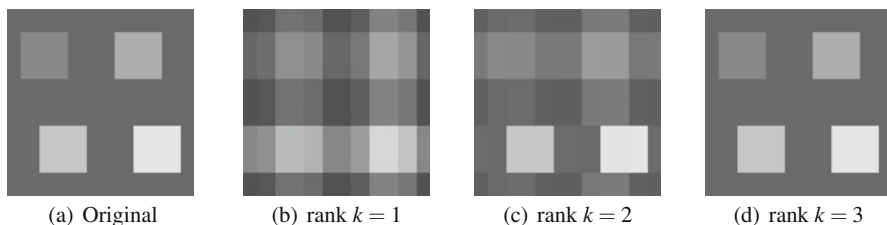
In this section we briefly discuss the application of SVD to two fields – apparently very different, indeed very close – namely image compression and data analysis.

6.2.1 Image Compression

Let us consider a grayscale image having $m \times n$ pixels: each pixel is an integer number $a_{ij} \in [0, 255]$ representing the intensity varying from white (255) to black (0). The image thus requires the storage of nm integers.

Performing the SVD of the intensity matrix $\mathbb{A} = (a_{ij})$ allows to compress the original image. Indeed, a compressed image (at level k) corresponds to the best approximation of rank k to the original matrix, represented by matrix \mathbb{A}_k of Theorem 6.1. Moreover, the rank k compressed image requires to store only $(m + n + 1)k$ integers.

Let us consider for example the grayscale image represented in Fig. 6.4(a).

**Fig. 6.4** Example of a grayscale image and its low-rank approximations of rank $r = 1, 2, 3$

The corresponding matrix $\mathbb{A} \in \mathbb{R}^{400 \times 400}$ is defined as

$$\mathbb{A}_{ij} = \begin{cases} 110, & 50 \leq i \leq 150, 30 \leq j \leq 130 \\ 150, & 50 \leq i \leq 150, 230 \leq j \leq 330 \\ 180, & 250 \leq i \leq 350, 70 \leq j \leq 170 \\ 220, & 250 \leq i \leq 350, 270 \leq j \leq 370 \\ 80, & \text{elsewhere.} \end{cases}$$

By computing its SVD for $k = 1, 2, 3$, we obtain the compressed images shown in Fig. 6.4(b),(c),(d). Surprisingly, the low-rank approximation of rank $k = 3$ of the matrix yields an image which is exactly equal to the original picture. The reason is that the matrix \mathbb{A} has rank $k = 3$, as shown by the plot of the singular values reported in Fig. 6.6(a)). For the picture in Fig. 6.5 the compression is more difficult than in the former case. In fact, as we can see in Fig. 6.6(b), the singular values decrease much slower than before.

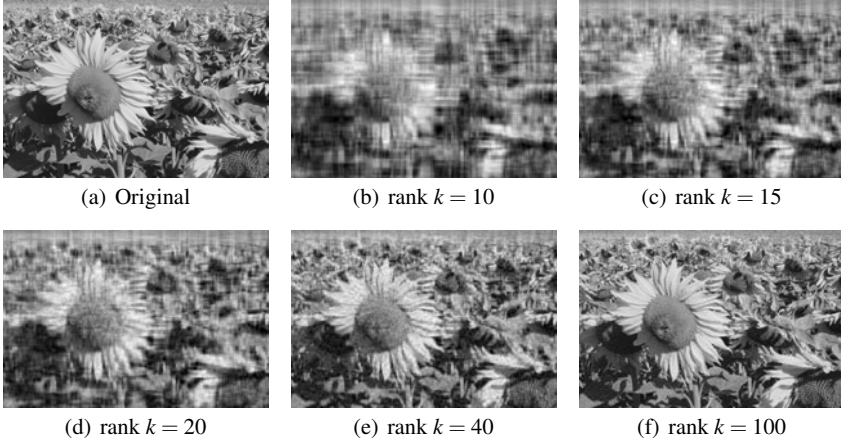


Fig. 6.5 Compression of a grayscale image by low-rank approximation

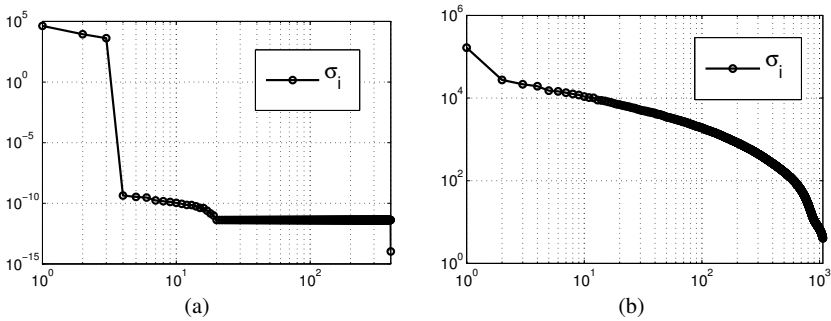


Fig. 6.6 Singular values of the images reported in Fig. 6.4 (a) and Fig. 6.5 (b)

6.2.2 Principal Component Analysis

Singular value decomposition, as well as principal orthogonal decomposition, are closely related to *principal component analysis* (PCA). Firstly introduced by Pearson in 1901 [209], then developed independently by Hotelling in 1936 [139], PCA is nowadays recognized as one of the most powerful tools for data mining. The goal of PCA is to reduce the dimensionality of multivariate data, by expressing the original variables in terms of fewer uncorrelated components, these latter being obtained by suitable linear combinations of the former.

Suppose the data $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{R}^n$ represent m independent observations of the variables X_1, \dots, X_n , $m > n$; the matrix representation of the dataset is given by

$$\mathbb{A} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_m^T \end{bmatrix} \in \mathbb{R}^{m \times n}.$$

The (sample) mean is

$$\bar{\mathbf{x}} = \frac{1}{m} \sum_{j=1}^m \mathbf{x}_j \in \mathbb{R}^n,$$

whereas the (sample) covariance matrix \mathbb{C} is

$$\mathbb{C} = \mathbb{A}^T \mathbb{A} \in \mathbb{R}^{n \times n}.$$

We denote by $\lambda_1(\mathbb{C}) \geq \lambda_2(\mathbb{C}) \geq \dots \geq \lambda_n(\mathbb{C})$ the eigenvalues of \mathbb{C} and by $\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_n$ the corresponding eigenvectors. Any linear combination $\mathbb{A}\mathbf{w} \in \mathbb{R}^m$ of the columns of \mathbb{A} has sample mean $\mathbf{w}^T \bar{\mathbf{x}}$ and sample variance $\mathbf{w}^T \mathbb{C} \mathbf{w}$; for any $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^n$, the sample covariance of $(\mathbf{x}_i^T \mathbf{w}_1, \mathbf{x}_i^T \mathbf{w}_2)$ is $\mathbf{w}_1^T \mathbb{C} \mathbf{w}_2$. The objective of PCA is to construct *uncorrelated* linear combinations of the measured variables that account for a given amount of the variation in the sample: the uncorrelated combinations with the largest variances will be called (sample) *principal components*.

More specifically, principal components are such that:

- their number is smaller than (or equal to) the number n of original variables;
- each subsequent component has the highest variance possible, being at the same time uncorrelated with (i.e., orthogonal to) the preceding components.

The first principal component $\zeta_1 = \mathbf{w}_1^T \mathbf{x}$ thus has to satisfy

$$\mathbf{w}_1 = \arg \max_{\mathbf{w}} \mathbf{w}^T \mathbb{C} \mathbf{w} = \arg \max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbb{C} \mathbf{w}}{\mathbf{w}^T \mathbf{w}};$$

here we restrict the coefficient vectors \mathbf{w} to satisfy $\|\mathbf{w}\|_2 = (\mathbf{w}^T \mathbf{w})^{1/2} = 1$. The maximum is the largest eigenvalue $\lambda_1(\mathbb{C})$ and it is attained for the choice $\mathbf{w}_1 = \boldsymbol{\psi}_1$, that is, the first principal component is given by $\zeta_1 = \boldsymbol{\psi}_1^T \mathbf{x}$. Usually, the observations \mathbf{x}_i , $i = 1, \dots, m$ are centered by subtracting the sample mean.

Since this operation has no effect on the sample covariance matrix \mathbb{C} , we often denote by

$$\zeta_1 = \boldsymbol{\psi}_1^T (\mathbf{x} - \bar{\mathbf{x}})$$

the first principal component. At the k -th step, the k -th (sample) principal component is the linear combination $\zeta_k = \mathbf{w}_k^T \mathbf{x}$ such that

$$\mathbf{w}_k = \arg \max_{\mathbf{w}} \frac{\mathbf{w}^T \hat{\mathbb{A}}_{k-1}^T \hat{\mathbb{A}}_{k-1} \mathbf{w}}{\mathbf{w}^T \mathbf{w}}, \quad \hat{\mathbb{A}}_{k-1} = \mathbb{A} - \sum_{s=1}^{k-1} \mathbb{A} \mathbf{w}_s \mathbf{w}_s^T$$

that is, it maximizes the sample variance of $\mathbf{w}_k^T \mathbf{x}$, subject to $\mathbf{w}_k^T \mathbf{w}_k = 1$ and zero sample covariance for all pairs $(\mathbf{w}_s^T \mathbf{x}, \mathbf{w}_k^T \mathbf{x})$, $0 \leq s < k$. For $k = 2, \dots, n$, this provides the eigenvectors of \mathbb{C} , with the maximum values for the quantity in brackets given by $\lambda_k(\mathbb{C})$. The principal components are thus given by

$$\zeta_k = \boldsymbol{\psi}_k^T (\mathbf{x} - \bar{\mathbf{x}}), \quad k = 1, \dots, n.$$

Since

$$\begin{aligned} \text{sample variance}(\zeta_k) &= \lambda_k(\mathbb{C}), & k = 1, \dots, n \\ \text{sample covariance}(\zeta_k, \zeta_l) &= 0, & k \neq l, \end{aligned}$$

PCA diagonalizes the (sample) variance-covariance matrix \mathbb{C} . Moreover,

$$\bar{\zeta}_k = \frac{1}{m} \sum_{j=1}^m \boldsymbol{\psi}_k^T (\mathbf{x}_j - \bar{\mathbf{x}}) = \frac{1}{m} \boldsymbol{\psi}_k^T \left(\sum_{j=1}^m (\mathbf{x}_j - \bar{\mathbf{x}}) \right) = 0$$

that is, the sample mean of each principal component is zero.

The geometric interpretation of Theorem 6.1 (see Remark 6.2) can be exploited to highlight the connection between SVD and data analysis provided by PCA. If we consider an hyper-ellipsoid in \mathbb{R}^n , its best one-dimensional approximation is the line segment corresponding to its longest axis. More in general, its best k -dimensional approximation is the k -dimensional ellipsoid spanned by its k longest axes. PCA acts on a set of m points in \mathbb{R}^n , the data observations, and describes their variability in terms of a new set of orthogonal directions, each one accounting for the largest variability in the data.

Geometrically, the initial data can be plotted as m points in the n -dimensional space (see Fig. 6.7). The same data can now be expressed in the new coordinate system, whose center is the sample mean $\bar{\mathbf{x}}$ and whose axes are given by the eigenvectors of \mathbb{C} . As $\|\boldsymbol{\psi}_k\| = 1$, $|\boldsymbol{\psi}_k^T (\mathbf{x} - \bar{\mathbf{x}})|$ represents the length of the projection of $\mathbf{x} - \bar{\mathbf{x}}$ on the unit vector $\boldsymbol{\psi}_k$. As a matter of fact, (sample) principal components result from translating the origin of the original coordinate system to $\bar{\mathbf{x}}$ and then rotating the coordinate axes until they pass through the set of points in the directions of maximum variance (see Fig. 6.7(a)).

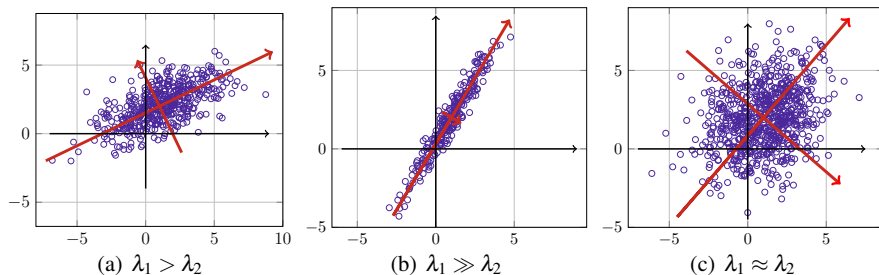


Fig. 6.7 $m = 500$ observations of $n = 2$ correlated variables, generated by three different bivariate Gaussian distributions

On the other hand, if data show the same variability along each direction (see Fig. 6.7(c)), the axes can coincide with any two perpendicular directions; this happens if the eigenvalues of the sample variance-covariance matrix \mathbb{C} are nearly equal, and the sample variation is homogeneous in all directions. Hence, the less *anisotropy* is shown by the original data, the most difficult is data representation in fewer than n directions. Instead, if the last eigenvalues λ_k are sufficiently small so that the variation along the corresponding axis $\boldsymbol{\psi}_k$ is negligible, the last few (sample) principal components can be ignored, and the data can be suitably approximated by their representation in the space of the retained components (see Fig. 6.7(b): in this case the first principal component accounts for most of the data variability). As we will see in the following sections, this represents a crucial factor for the *reducibility* of a PDE problem through proper orthogonal decomposition.

6.3 Proper Orthogonal Decomposition

Proper orthogonal decomposition (POD) is, in a very broad sense, a technique for reducing the dimensionality of a given dataset (and, generally speaking, of a system) by representing it onto an orthonormal basis which is optimal in a least-squares sense. The original variables are transformed into a new set of uncorrelated variables (called POD modes, or principal components), the first few modes ideally retaining most of the *energy* present in all of the original variables. A lower dimensional representation of the data is thus obtained by truncating the new basis to retain the first few modes. In the theory of stochastic processes this procedure is known as Karhunen-Loève (KL) decomposition, whereas in multivariate statistics it is precisely the Principal Component Analysis introduced in the previous section.

The first applications of POD in scientific computing were concerned with the simulation of turbulent flows, and the extraction of (both spatial and temporal) coherent structures appearing in fully developed turbulent flows [16, 17, 78, 138, 243, 242] and date back to the early '90s, after the pioneering work by Lumley in the late '60s [170]. In its early applications, POD was meant to represent and analyze complex data; the interested reader can find further details for instance in [28, 137, 151].

In the context of reduced-order modeling, the POD method has been mostly used – in conjunction with (Petrov-)Galerkin projection methods – to build reduced-order models of time-dependent problems [156, 157, 189, 223] and, more recently, in the context of parametrized systems [64, 42, 45, 148, 247, 149, 18]. See, e.g., [212, 256] for further details.

6.3.1 POD for Parametrized Problems

Nowadays, POD is successfully applied in the context of reduced-order modelling for parametrized PDEs. Consider a set $\Xi_s = \{\boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^{n_s}\}$ of n_s parameter samples and the corresponding set of snapshots $\{u_h(\boldsymbol{\mu}^1), \dots, u_h(\boldsymbol{\mu}^{n_s})\}$, that is the solutions of the high-fidelity problem (3.11). We define the snapshot matrix $\mathbb{S} \in \mathbb{R}^{N_h \times n_s}$ as

$$\mathbb{S} = [\mathbf{u}_1 \mid \dots \mid \mathbf{u}_{n_s}],$$

where the vectors $\mathbf{u}_i \in \mathbb{R}^{N_h}$, $1 \leq i \leq n_s$, represent the degrees of freedom of the functions $u_h(\boldsymbol{\mu}^i) \in V_h$ (i.e. $\mathbf{u}_i^{(j)} = u_h^{(j)}(\boldsymbol{\mu}^i)$ for $1 \leq i \leq n_s$ and $1 \leq j \leq N_h$), see (2.50).

According to (6.1), the SVD decomposition of \mathbb{S} reads

$$\mathbb{S} = \mathbb{U} \boldsymbol{\Sigma} \mathbb{Z}^T, \quad (6.7)$$

where $\mathbb{U} = [\boldsymbol{\zeta}_1 \mid \dots \mid \boldsymbol{\zeta}_{N_h}] \in \mathbb{R}^{N_h \times N_h}$ and $\mathbb{Z} = [\boldsymbol{\psi}_1 \mid \dots \mid \boldsymbol{\psi}_{n_s}] \in \mathbb{R}^{n_s \times n_s}$ are orthogonal matrices, and $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_r) \in \mathbb{R}^{N_h \times n_s}$ with $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$. Here $r \leq \min(N_h, n_s)$ denotes the rank of \mathbb{S} , which is strictly smaller than n_s if the snapshot vectors are not all linearly independent. Then, we can write

$$\mathbb{S} \boldsymbol{\psi}_i = \sigma_i \boldsymbol{\zeta}_i \quad \text{and} \quad \mathbb{S}^T \boldsymbol{\zeta}_i = \sigma_i \boldsymbol{\psi}_i, \quad i = 1, \dots, r \quad (6.8)$$

or, equivalently,

$$\mathbb{S}^T \mathbb{S} \boldsymbol{\psi}_i = \sigma_i^2 \boldsymbol{\psi}_i \quad \text{and} \quad \mathbb{S} \mathbb{S}^T \boldsymbol{\zeta}_i = \sigma_i^2 \boldsymbol{\zeta}_i, \quad i = 1, \dots, r \quad (6.9)$$

i.e. $\sigma_i^2, i = 1, \dots, r$, are the nonzero eigenvalues of the matrix $\mathbb{S}^T \mathbb{S}$ (and also of $\mathbb{S} \mathbb{S}^T$), listed in nondecreasing order. The matrix $\mathbb{C} = \mathbb{S}^T \mathbb{S} \in \mathbb{R}^{n_s \times n_s}$ is called *correlation matrix*; its elements are given by

$$\mathbb{C}_{ij} = \mathbf{u}_i^T \mathbf{u}_j, \quad 1 \leq i, j \leq n_s.$$

For any $N \leq n_s$, the *POD basis* $\mathbb{V} \in \mathbb{R}^{N_h \times N}$ of dimension N is defined as the set of the first N left singular vectors $\boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_N$ of \mathbb{U} or, equivalently, the set of vectors

$$\boldsymbol{\zeta}_j = \frac{1}{\sigma_j} \mathbb{S} \boldsymbol{\psi}_j, \quad 1 \leq j \leq N \quad (6.10)$$

obtained from the first N eigenvectors $\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_N$ of the correlation matrix \mathbb{C} .

By construction, the POD basis is orthonormal. Moreover, it minimizes, over all possible N -dimensional orthonormal bases $\mathbb{W} = [\mathbf{w}_1 \mid \dots \mid \mathbf{w}_N] \in \mathbb{R}^{N_h \times N}$, the sum of the squares of the errors between each snapshot vector \mathbf{u}_i and its projection onto the subspace spanned by \mathbb{W} . More precisely, recalling that the projection $\Pi_{\mathbb{W}} \mathbf{x}$ of a vector $\mathbf{x} \in \mathbb{R}^{N_h}$ onto $\text{span}(\mathbb{W})$ is given by

$$\Pi_{\mathbb{W}} \mathbf{x} = \sum_{j=1}^N (\mathbf{x}, \mathbf{w}_j)_2 \mathbf{w}_j = \mathbb{W} \mathbb{W}^T \mathbf{x},$$

the following property holds.

Proposition 6.1. *Let $\mathcal{V}_N = \{\mathbb{W} \in \mathbb{R}^{N_h \times N} : \mathbb{W}^T \mathbb{W} = \mathbb{I}_N\}$ be the set of all N -dimensional orthonormal bases. Then,*

$$\sum_{i=1}^{n_s} \|\mathbf{u}_i - \mathbb{V} \mathbb{V}^T \mathbf{u}_i\|_2^2 = \min_{\mathbb{W} \in \mathcal{V}_N} \sum_{i=1}^{n_s} \|\mathbf{u}_i - \mathbb{W} \mathbb{W}^T \mathbf{u}_i\|_2^2 = \sum_{i=N+1}^r \sigma_i^2. \quad (6.11)$$

Proof. By Theorem 6.1, the best rank N approximation of \mathbb{S} is given by

$$\mathbb{S}_N = \sum_{i=1}^N \sigma_i \boldsymbol{\zeta}_i \boldsymbol{\psi}_i^T.$$

From the second relation in (6.8), we have that

$$\boldsymbol{\psi}_i = \frac{1}{\sigma_i} \mathbb{S}^T \boldsymbol{\zeta}_i,$$

and therefore

$$\mathbb{S}_N = \sum_{i=1}^N \sigma_i \boldsymbol{\zeta}_i \frac{1}{\sigma_i} (\mathbb{S}^T \boldsymbol{\zeta}_i)^T = \sum_{i=1}^N \boldsymbol{\zeta}_i (\boldsymbol{\zeta}_i^T \mathbb{S}) = \mathbb{V} \mathbb{V}^T \mathbb{S}.$$

For a matrix $\mathbb{A} \in \mathbb{R}^{m \times n}$ we have (see (6.3))

$$\|\mathbb{A}\|_F^2 = \sum_{i=1}^n \|\mathbf{a}_i\|_2^2,$$

where \mathbf{a}_i denotes the i -th column of \mathbb{A} . Thus, for $\mathbb{W} \in \mathcal{V}_N$ we can write

$$\sum_{i=1}^{n_s} \|\mathbf{u}_i - \mathbb{W} \mathbb{W}^T \mathbf{u}_i\|_2^2 = \|\mathbb{S} - \mathbb{W} \mathbb{W}^T \mathbb{S}\|_F^2.$$

Since, by (6.5),

$$\|\mathbb{S} - \mathbb{S}_N\|_F^2 = \min_{\substack{\mathbb{B} \in \mathbb{R}^{N_h \times n_s} \\ \text{rank}(\mathbb{B}) \leq N}} \|\mathbb{S} - \mathbb{B}\|_F^2 \leq \min_{\mathbb{W} \in \mathcal{V}_N} \|\mathbb{S} - \mathbb{W} \mathbb{W}^T \mathbb{S}\|_F^2$$

and $\mathbb{S}_N = \mathbb{V}\mathbb{V}^T\mathbb{S}$, we conclude that

$$\|\mathbb{S} - \mathbb{V}\mathbb{V}^T\mathbb{S}\|_F^2 = \min_{\mathbb{W} \in \mathcal{V}_N} \|\mathbb{S} - \mathbb{W}\mathbb{W}^T\mathbb{S}\|_F^2. \quad \square$$

From Proposition 6.1 it follows that the error in the POD basis is equal to the sum of the squares of the singular values corresponding to the neglected POD modes. This result suggests a suitable criterion to select the minimal POD dimension $N \leq r$ such that the projection error is smaller than a desired tolerance ε_{POD} . Indeed, it is sufficient to choose N as the smallest integer such that

$$I(N) = \frac{\sum_{i=1}^N \sigma_i^2}{\sum_{i=1}^r \sigma_i^2} \geq 1 - \varepsilon_{\text{POD}}^2, \quad (6.12)$$

that is the energy retained by the last $r - N$ modes is equal or smaller than $\varepsilon_{\text{POD}}^2$. $I(N)$ represents the percentage of energy of the snapshots captured by the first N POD modes, and it is referred to as the *relative information content* of the POD basis [2]. Equivalently, the criterion (6.12) ensures that the relative error between \mathbb{S} and its N -rank approximation \mathbb{S}_N is smaller than ε_{POD} , i.e.

$$\frac{\|\mathbb{S} - \mathbb{S}_N\|_F}{\|\mathbb{S}\|_F} \leq \varepsilon_{\text{POD}}.$$

The procedure summarized in Algorithm 6.1 combines the definition of POD basis – provided by either (6.7) or (6.10) – together with the optimality criterion (6.12). We remark that computing the POD basis by solving an eigenvalue problem for the correlation matrix \mathbb{C} yields inaccurate results for the modes associated to small singular values. This is due to the roundoff errors introduced while constructing \mathbb{C} and the fact that $\kappa(\mathbb{C}) = (\kappa(\mathbb{S}))^2$. In such cases it is recommended to construct the POD basis by means of stable algorithms for the computation of the SVD, see e.g. [117, 68].

Algorithm 6.1 POD algorithm

```

1: function  $\mathbb{V} = \text{POD}(\mathbb{S}, \varepsilon_{\text{POD}})$ 
2:   if  $n_s \leq N_h$  then
3:     form the correlation matrix  $\mathbb{C} = \mathbb{S}^T\mathbb{S}$ 
4:     solve the eigenvalue problem  $\mathbb{C}\boldsymbol{\psi}_i = \sigma_i^2\boldsymbol{\psi}_i$ ,  $i = 1, \dots, r$ 
5:     set  $\boldsymbol{\zeta}_i = \frac{1}{\sigma_i}\mathbb{S}\boldsymbol{\psi}_i$ 
6:   else
7:     form the matrix  $\mathbb{K} = \mathbb{S}\mathbb{S}^T$ 
8:     solve the eigenvalue problem  $\mathbb{K}\boldsymbol{\zeta}_i = \sigma_i^2\boldsymbol{\zeta}_i$ ,  $i = 1, \dots, r$ 
9:   end if
10:  define  $N$  as the minimum integer such that  $I(N) \geq 1 - \varepsilon_{\text{POD}}$ 
11:   $\mathbb{V} = [\boldsymbol{\zeta}_1 \mid \dots \mid \boldsymbol{\zeta}_N]$ 
12: end function

```

6.3.2 POD with Energy Inner Product

Since the snapshots functions $u_h(\boldsymbol{\mu}^i)$ belong to $V_h \subset V$, it is natural to seek an alternative POD basis which minimizes the \mathbb{X}_h -norm – rather than the $\|\cdot\|_2$ norm – of the projection error of the snapshots vectors \mathbf{u}_i . In particular, we seek a basis $\mathbb{W} \in \mathcal{V}_N^{\mathbb{X}_h}$, with

$$\mathcal{V}_N^{\mathbb{X}_h} = \{\mathbb{W} \in \mathbb{R}^{N_h \times N} : \mathbb{W}^T \mathbb{X}_h \mathbb{W} = \mathbb{I}_N\},$$

which minimizes the squares of the \mathbb{X}_h -norm of the error between each snapshot vector \mathbf{u}_i and its \mathbb{X}_h -orthogonal projection onto the subspace spanned by \mathbb{W} , i.e.

$$\min_{\mathbb{W} \in \mathcal{V}_N^{\mathbb{X}_h}} \sum_{i=1}^{n_s} \|\mathbf{u}_i - \mathbb{P}_W^{\mathbb{X}_h} \mathbf{u}_i\|_{\mathbb{X}_h}^2. \quad (6.13)$$

Here

$$\mathbb{P}_W^{\mathbb{X}_h} \mathbf{x} = \sum_{j=1}^N (\mathbf{x}, \mathbf{w}_j)_{\mathbb{X}_h} \mathbf{w}_j = \mathbb{W} \mathbb{W}^T \mathbb{X}_h \mathbf{x} \quad (6.14)$$

is the \mathbb{X}_h -orthogonal projection of $\mathbf{x} \in \mathbb{R}^{N_h}$ onto $\text{span}(\mathbb{W})$. By (6.14) we have

$$\begin{aligned} \sum_{i=1}^{n_s} \|\mathbf{u}_i - \mathbb{P}_W^{\mathbb{X}_h} \mathbf{u}_i\|_{\mathbb{X}_h}^2 &= \sum_{i=1}^{n_s} \|\mathbf{u}_i - \mathbb{W} \mathbb{W}^T \mathbb{X}_h \mathbf{u}_i\|_{\mathbb{X}_h}^2 \\ &= \sum_{i=1}^{n_s} \|\mathbb{X}_h^{1/2} \mathbf{u}_i - \mathbb{X}_h^{1/2} \mathbb{W} \mathbb{W}^T \mathbb{X}_h \mathbf{u}_i\|_2^2 = \|\mathbb{X}_h^{1/2} \mathbb{S} - \mathbb{X}_h^{1/2} \mathbb{W} \mathbb{W}^T \mathbb{X}_h \mathbb{S}\|_F^2. \end{aligned} \quad (6.15)$$

Substituting into (6.13) and setting $\tilde{\mathbb{S}} = \mathbb{X}_h^{1/2} \mathbb{S}$ and $\tilde{\mathbb{W}} = \mathbb{X}_h^{1/2} \mathbb{W}$, we finally obtain

$$\min_{\mathbb{W} \in \mathcal{V}_N^{\mathbb{X}_h}} \sum_{i=1}^{n_s} \|\mathbf{u}_i - \mathbb{P}_W^{\mathbb{X}_h} \mathbf{u}_i\|_{\mathbb{X}_h}^2 = \min_{\tilde{\mathbb{W}} \in \mathcal{V}_N} \|\tilde{\mathbb{S}} - \tilde{\mathbb{W}} \tilde{\mathbb{W}}^T \tilde{\mathbb{S}}\|_F^2.$$

At this stage, Schmidt-Eckart-Young Theorem 6.1 and Proposition 6.1 yield the following result.

Proposition 6.2. Let $\mathbb{S} = [\mathbf{u}_1 \mid \dots \mid \mathbf{u}_{n_s}] \in \mathbb{R}^{N_h \times n_s}$ be a given matrix of rank $r \leq \min(N_h, n_s)$, $\mathbb{X}_h \in \mathbb{R}^{N_h \times N_h}$ a symmetric positive definite matrix, $\tilde{\mathbb{S}} = \mathbb{X}_h^{1/2} \mathbb{S}$ and $\tilde{\mathbb{S}} = \tilde{\mathbb{U}} \tilde{\Sigma} \tilde{\mathbb{Z}}^T$ its singular value decomposition, where

$$\tilde{\mathbb{U}} = [\tilde{\boldsymbol{\zeta}}_1 \mid \dots \mid \tilde{\boldsymbol{\zeta}}_{N_h}] \in \mathbb{R}^{N_h \times N_h}, \quad \tilde{\mathbb{Z}} = [\tilde{\boldsymbol{\psi}}_1 \mid \dots \mid \tilde{\boldsymbol{\psi}}_{n_s}] \in \mathbb{R}^{n_s \times n_s}$$

are orthogonal matrices and $\tilde{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_r) \in \mathbb{R}^{N_h \times n_s}$ with $\sigma_1 \geq \dots \geq \sigma_r$. Then, for $N \leq r$, the POD basis $\mathbb{V} = [\mathbb{X}_h^{-1/2} \tilde{\boldsymbol{\zeta}}_1 \mid \dots \mid \mathbb{X}_h^{-1/2} \tilde{\boldsymbol{\zeta}}_N]$ is such that

$$\sum_{i=1}^{n_s} \|\mathbf{u}_i - \mathbb{V} \mathbb{V}^T \mathbb{X}_h \mathbf{u}_i\|_{\mathbb{X}_h}^2 = \min_{\mathbb{W} \in \mathcal{V}_N^{\mathbb{X}_h}} \sum_{i=1}^{n_s} \|\mathbf{u}_i - \mathbb{W} \mathbb{W}^T \mathbb{X}_h \mathbf{u}_i\|_{\mathbb{X}_h}^2 = \sum_{i=N+1}^r \sigma_i^2.$$

From a computational standpoint, since

$$\tilde{\mathbf{S}}^T \tilde{\mathbf{S}} \tilde{\boldsymbol{\Psi}}_i = \sigma_i^2 \tilde{\boldsymbol{\Psi}}_i, \quad i = 1, \dots, r, \quad (6.16)$$

if $n_s < N_h$ we can conveniently obtain the POD basis without forming the matrix $\mathbb{X}_h^{1/2}$. Indeed, we first compute the correlation matrix $\tilde{\mathbf{C}} = \tilde{\mathbf{S}}^T \tilde{\mathbf{S}} = \mathbf{S}^T \mathbb{X}_h \mathbf{S}$ and its first N eigenvectors $\tilde{\boldsymbol{\Psi}}_1, \dots, \tilde{\boldsymbol{\Psi}}_N$. Then, we define the POD basis as $\mathbb{V} = [\boldsymbol{\zeta}_1 \mid \dots \mid \boldsymbol{\zeta}_N]$, where

$$\boldsymbol{\zeta}_i = \mathbb{X}_h^{-1/2} \tilde{\boldsymbol{\zeta}}_i = \mathbb{X}_h^{-1/2} \frac{1}{\sigma_i} \tilde{\mathbf{S}} \tilde{\boldsymbol{\Psi}}_i = \frac{1}{\sigma_i} \mathbf{S} \tilde{\boldsymbol{\Psi}}_i.$$

The complete procedure is summarized in Algorithm 6.2.

Algorithm 6.2 POD algorithm with respect to the \mathbb{X}_h norm

```

1: function  $\mathbb{V} = \text{POD}(\mathbf{S}, \mathbb{X}_h, \varepsilon_{\text{POD}})$ 
2:   if  $n_s \leq N_h$  then
3:     form the correlation matrix  $\tilde{\mathbf{C}} = \mathbf{S}^T \mathbb{X}_h \mathbf{S}$ 
4:     solve the eigenvalue problem  $\tilde{\mathbf{C}} \tilde{\boldsymbol{\Psi}}_i = \sigma_i^2 \tilde{\boldsymbol{\Psi}}_i, \quad i = 1, \dots, r$ 
5:     set  $\tilde{\boldsymbol{\zeta}}_i = \frac{1}{\sigma_i} \tilde{\mathbf{S}} \tilde{\boldsymbol{\Psi}}_i$ 
6:   else
7:     form the matrix  $\tilde{\mathbf{K}} = \mathbb{X}_h^{1/2} \mathbf{S} \mathbf{S}^T \mathbb{X}_h^{1/2}$ 
8:     solve the eigenvalue problem  $\tilde{\mathbf{K}} \tilde{\boldsymbol{\zeta}}_i = \sigma_i^2 \tilde{\boldsymbol{\zeta}}_i, \quad i = 1, \dots, r$ 
9:     set  $\tilde{\boldsymbol{\zeta}}_i = \mathbb{X}_h^{-1/2} \tilde{\boldsymbol{\zeta}}_i$ 
10:  end if
11:  define  $N$  as the minimum integer such that  $I(N) \geq 1 - \varepsilon_{\text{POD}}$ 
12:   $\mathbb{V} = [\boldsymbol{\zeta}_1 \mid \dots \mid \boldsymbol{\zeta}_N]$ 
13: end function

```

6.4 \mathcal{P} -continuous Analogue of POD

The POD basis is, among all basis of dimension $N < n_s$, the one which best approximates the set of solution snapshots $\mathcal{M}_h^s = \{\mathbf{u}_h(\boldsymbol{\mu}^1), \dots, \mathbf{u}_h(\boldsymbol{\mu}^{n_s})\}$. However, we still have to investigate:

1. the approximation property of the POD basis with respect to the entire solutions set $\mathcal{M}_h = \{\mathbf{u}_h(\boldsymbol{\mu}) : \boldsymbol{\mu} \in \mathcal{P}\}$;
2. how to select the parameter samples $\Xi_s = \{\boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^{n_s}\}$ so that the corresponding snapshots set is sufficiently representative of the solutions set.

The former issue is strictly related to the analysis we carried out in Chap. 5, whereas the latter calls for the setting of the POD algorithm. With this in mind, it is useful to introduce a \mathcal{P} -continuous version of the POD technique. Indeed, if we are

interested to find a POD basis of dimension N that approximates the entire solutions set \mathcal{M}_h , we have to consider the following minimization problem

$$\min_{\mathbb{W} \in \mathcal{V}_N} \int_{\mathcal{P}} \|\mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{W}\mathbb{W}^T \mathbf{u}_h(\boldsymbol{\mu})\|_2^2 d\boldsymbol{\mu}. \quad (6.17)$$

(We keep using the same notations for \mathbb{W} and \mathbb{V} used in the finite-dimensional case.) The solution of this *nonconvex* optimization problem is due to E. Schmidt [239] (see also [244, 245]), who first introduced an infinite-dimensional analogue of the SVD and showed how to use it to obtain an optimal low-rank approximation to a (compact) integral operator.

Let us suppose that $\mathbf{u}_h(\boldsymbol{\mu}) \in L^2(\mathcal{P}; \mathbb{R}^{N_h})$, i.e.

$$\int_{\mathcal{P}} \|\mathbf{u}_h(\boldsymbol{\mu})\|_2^2 d\boldsymbol{\mu} < \infty;$$

$\mathbf{u}_h(\boldsymbol{\mu})$ is called a Hilbert-Schmidt kernel. A continuous analogue of the snapshots matrix \mathbb{S} is given by the operator $T : L^2(\mathcal{P}) \rightarrow \mathbb{R}^{N_h}$ such that

$$Tg = \int_{\mathcal{P}} \mathbf{u}_h(\boldsymbol{\mu}) g(\boldsymbol{\mu}) d\boldsymbol{\mu} \quad \forall g \in L^2(\mathcal{P}). \quad (6.18)$$

Its adjoint $T^* : \mathbb{R}^{N_h} \rightarrow L^2(\mathcal{P})$ is defined as

$$(g, T^* \mathbf{w})_{L^2(\mathcal{P})} = (Tg, \mathbf{w})_2 \quad \forall g \in L^2(\mathcal{P}), \mathbf{w} \in \mathbb{R}^{N_h}.$$

As a result

$$T^* \mathbf{w} = (\mathbf{u}_h(\boldsymbol{\mu}), \mathbf{w})_2 \quad \forall \mathbf{w} \in \mathbb{R}^{N_h}, \quad (6.19)$$

that is, T^* is the continuous analogue of the matrix \mathbb{S}^T . Note that $Tg \in \mathbb{R}^{N_h}$ and $\text{rank}(T) = r \leq N_h$, while $T^* \mathbf{w} = (T^* \mathbf{w})(\boldsymbol{\mu}) \in L^2(\mathcal{P})$. Since $\mathbf{u}_h(\boldsymbol{\mu})$ is a Hilbert-Schmidt kernel, T is a Hilbert-Schmidt, and thus compact, operator (see Sect. A.4 in Appendix A). Moreover, the Hilbert-Schmidt norm of T (see (A.21)) coincides with the norm of its kernel (see e.g. [225, 116]), i.e.

$$\|T\|_{HS}^2 = \|\mathbf{u}_h(\boldsymbol{\mu})\|_{L^2(\mathcal{P}; \mathbb{R}^{N_h})}^2.$$

Since T is compact, $K = TT^* : \mathbb{R}^{N_h} \rightarrow \mathbb{R}^{N_h}$ and $C = T^*T : L^2(\mathcal{P}) \rightarrow L^2(\mathcal{P})$ are self-adjoint, non-negative, compact operators, given by

$$Cg = \int_{\mathcal{P}} (\mathbf{u}_h(\boldsymbol{\mu}), \mathbf{u}_h(\boldsymbol{\mu}'))_2 g(\boldsymbol{\mu}') d\boldsymbol{\mu}' \quad \forall g \in L^2(\mathcal{P}),$$

$$K\mathbf{w} = \int_{\mathcal{P}} \mathbf{u}_h(\boldsymbol{\mu}) (\mathbf{u}_h(\boldsymbol{\mu}), \mathbf{w})_2 d\boldsymbol{\mu} \quad \forall \mathbf{w} \in \mathbb{R}^{N_h}.$$

Note that the operators K and C are nothing but the \mathcal{P} -continuous analogue of the correlation matrices $\mathbb{K} = \mathbb{S}\mathbb{S}^T$ and $\mathbb{C} = \mathbb{S}^T\mathbb{S}$.

Actually, since K is a linear map from \mathbb{R}^{N_h} to \mathbb{R}^{N_h} , it is represented by the $N_h \times N_h$ symmetric positive definite matrix

$$K = \int_{\mathcal{P}} \mathbf{u}_h(\boldsymbol{\mu}) \mathbf{u}_h^T(\boldsymbol{\mu}) d\boldsymbol{\mu}$$

whose eigenvalues $\sigma_1^2 \geq \dots \geq \sigma_r^2 \geq 0$ and associated orthonormal eigenvectors $\boldsymbol{\zeta}_i \in \mathbb{R}^{N_h}$ satisfy

$$K \boldsymbol{\zeta}_i = \sigma_i^2 \boldsymbol{\zeta}_i, \quad i = 1, \dots, r. \quad (6.20)$$

Moreover, the functions $\psi_i \in L^2(\mathcal{P})$ defined by

$$\psi_i = \frac{1}{\sigma_i} T^* \boldsymbol{\zeta}_i, \quad i = 1, \dots, r,$$

are the eigenvectors of C . Finally, as $\mathbf{u}_h(\boldsymbol{\mu})$ admits the expansion

$$\mathbf{u}_h(\boldsymbol{\mu}) = \sum_{i=1}^r \sigma_i \boldsymbol{\zeta}_i \psi_i(\boldsymbol{\mu}) = \sum_{i=1}^r \boldsymbol{\zeta}_i (\mathbf{u}_h(\boldsymbol{\mu}), \boldsymbol{\zeta}_i)_2, \quad (6.21)$$

the following decomposition holds for T

$$T = \sum_{i=1}^r \sigma_i \boldsymbol{\zeta}_i (\psi_i(\boldsymbol{\mu}), \cdot)_{L^2(\mathcal{P})}. \quad (6.22)$$

As in the matrix context, it can be easily proved that

$$\|T\|_{\mathcal{L}(L^2(\mathcal{P}), \mathbb{R}^{N_h})} = \sigma_1, \quad \|T\|_{HS} = \sqrt{\sum_{i=1}^r \sigma_i^2}. \quad (6.23)$$

Moreover, truncating the sum (6.22) to the first N terms, we obtain the best rank N approximation to the operator T .

Theorem 6.2 (Schmidt). *The operator $T_N : L^2(\mathcal{P}) \rightarrow \mathbb{R}^{N_h}$ defined by*

$$T_N = \sum_{i=1}^N \sigma_i \boldsymbol{\zeta}_i (\psi_i(\boldsymbol{\mu}), \cdot)_{L^2(\mathcal{P})}, \quad 0 \leq N < r, \quad (6.24)$$

satisfies the following optimality property

$$\|T - T_N\|_{HS} = \min_{B \in \mathcal{B}_N} \|T - B\|_{HS} = \sqrt{\sum_{i=N+1}^r \sigma_i^2}, \quad (6.25)$$

where $\mathcal{B}_N = \{B \in \mathcal{L}(L^2(\mathcal{P}), \mathbb{R}^{N_h}) : \text{rank}(B) \leq N\}$.

For the proof of this fundamental theorem we follow the approach of Weyl [258] as reported in [244], rather than relying on the original proof by Schmidt [239]. We first need the following two lemmas.

Lemma 6.1. *Let $T : L^2(\mathcal{P}) \rightarrow \mathbb{R}^{N_h}$ be the rank- r operator defined in (6.18) and $B_N \in \mathcal{B}_N$ be a linear operator with finite rank $N < r$. Then*

$$\sigma_1(T - B_N) \geq \sigma_{N+1}(T). \quad (6.26)$$

Proof. It is well known that B_N has finite rank if and only if there exist two orthonormal sets $\{\mathbf{u}_i\}_{i=1}^N$ and $\{v_i(\boldsymbol{\mu})\}_{i=1}^N$ such that

$$B_N = \sum_{i=1}^N \mathbf{u}_i(v_i, \cdot)_{L^2(\mathcal{P})}.$$

The restriction of B_N to $\text{span}(\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_{N+1})$ has a non-trivial kernel. Thus there exists $\boldsymbol{\gamma} \in \mathbb{R}^{N+1}$ with $\|\boldsymbol{\gamma}\|_2 = 1$ such that $B_N \mathbf{y} = 0$, with $\mathbf{y} = \sum_{i=1}^{N+1} \gamma_i \boldsymbol{\psi}_i \in L^2(\mathcal{P})$. Denoting by $\sigma_i = \sigma_i(T)$ the singular values of T , we have

$$\begin{aligned} (\sigma_1(T - B_N))^2 &= \|T - B_N\|_{\mathcal{L}(L^2(\mathcal{P}), \mathbb{R}^{N_h})}^2 = \sup_{g \in L^2(\mathcal{P})} \|(T - B_N)g\|_2^2 \\ &\geq \|(T - B_N)\mathbf{y}\|_2^2 = \|\mathbf{T}\mathbf{y}\|_2^2 = \left\| \sum_{i=1}^r \sum_{j=1}^{N+1} \sigma_i \zeta_i(\boldsymbol{\psi}_i, \gamma_j \boldsymbol{\psi}_j)_{L^2(\mathcal{P})} \right\|_2^2 \\ &= \sum_{i=1}^{N+1} \sigma_i^2 \gamma_i^2 \geq \sigma_{N+1}^2. \end{aligned} \quad \square$$

Lemma 6.2. *Let $T', T'' \in \mathcal{B}_r$ be such that $T = T' + T''$. Then, for $i, j = 1, \dots, r$*

$$\sigma_{i+j-1}(T) \leq \sigma_i(T') + \sigma_j(T''). \quad (6.27)$$

Proof. We denote by $\sigma'_i = \sigma_i(T')$ and $\sigma''_i = \sigma_i(T'')$ the eigenvalues of T' and T'' , respectively. We first prove (6.27) for $i = j = 1$,

$$\begin{aligned} \sigma_1 &= \|T\|_{\mathcal{L}(L^2(\mathcal{P}), \mathbb{R}^{N_h})} = \|T' + T''\|_{\mathcal{L}(L^2(\mathcal{P}), \mathbb{R}^{N_h})} \\ &\leq \|T'\|_{\mathcal{L}(L^2(\mathcal{P}), \mathbb{R}^{N_h})} + \|T''\|_{\mathcal{L}(L^2(\mathcal{P}), \mathbb{R}^{N_h})} = \sigma'_1 + \sigma''_1. \end{aligned} \quad (6.28)$$

Then, we note that for any $i, j \geq 2$, $\sigma_1(T' - T'_{i-1}) = \sigma_i(T')$, $\sigma_1(T'' - T''_{j-1}) = \sigma_j(T'')$ where the operators T'_{i-1} and T''_{j-1} are formed in analogy with (6.24). Thus, by (6.28) we find

$$\sigma'_i + \sigma''_j = \sigma_1(T' - T'_{i-1}) + \sigma_1(T'' - T''_{j-1}) \geq \sigma_1(T - T'_{i-1} - T''_{j-1}).$$

Since

$$\text{rank}(T'_{i-1} + T''_{j-1}) \leq \text{rank}(T'_{i-1}) + \text{rank}(T''_{j-1}) = i + j - 2,$$

Lemma 6.1 implies that $\sigma_1(T - T'_{i-1} - T''_{j-1}) \geq \sigma_{i+j-1}$. \square

Proof of Theorem 6.2. We first note that, thanks to the orthonormality of the $\{\boldsymbol{\zeta}_i, i = 1, \dots, r\}$ and $\{\boldsymbol{\psi}_i, i = 1, \dots, r\}$,

$$\|T - T_N\|_{HS}^2 = \left\| \sum_{i=N+1}^r \sigma_i \boldsymbol{\zeta}_i \boldsymbol{\psi}_i \right\|_{L^2(\mathcal{P})}^2 = \sum_{i=N+1}^r \sigma_i^2.$$

Then, we prove

$$\|T - B_N\|_{HS}^2 \geq \sum_{i=N+1}^r \sigma_i^2 \quad \forall B_N \in \mathcal{B}_N,$$

thanks to Lemma 6.2. We set $T' = T - B_N$, $T'' = B_N$ and $j = N + 1$; since $\sigma_{N+1}(B_N) = 0$, it follows

$$\sigma_{i+N}(T) \leq \sigma_i(T - B_N).$$

Summing up from $i = 1$ to $i = r$ we finally obtain

$$\sum_{i=1}^r (\sigma_i(T - B_N))^2 = \|T - B_N\|_{HS}^2 \geq \sum_{i=N+1}^r \sigma_i^2 \quad \forall B_N \in \mathcal{B}_N. \quad \square$$

Theorem 6.2 provides a solution to the minimization problem (6.17), as stated in the following Proposition.

Proposition 6.3. *The POD basis $\mathbb{V} = [\boldsymbol{\zeta}_1 \mid \dots \mid \boldsymbol{\zeta}_N] \in \mathbb{R}^{N_h \times N}$ is such that*

$$\int_{\mathcal{P}} \|\mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{V}\mathbb{V}^T \mathbf{u}_h(\boldsymbol{\mu})\|_2^2 d\boldsymbol{\mu} = \min_{\mathbb{W} \in \mathcal{V}_N} \int_{\mathcal{P}} \|\mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{W}\mathbb{W}^T \mathbf{u}_h(\boldsymbol{\mu})\|_2^2 d\boldsymbol{\mu}. \quad (6.29)$$

Moreover

$$\int_{\mathcal{P}} \|\mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{V}\mathbb{V}^T \mathbf{u}_h(\boldsymbol{\mu})\|_2^2 d\boldsymbol{\mu} = \sum_{i=N+1}^r \sigma_i^2. \quad (6.30)$$

Proof. Since $\boldsymbol{\psi}_i = \frac{1}{\sigma_i} T^* \boldsymbol{\zeta}_i$, thanks to the definition of T^* we can express T_N as

$$\begin{aligned} T_N g &= \sum_{i=1}^N \sigma_i \boldsymbol{\zeta}_i (\boldsymbol{\psi}_i(\boldsymbol{\mu}), g(\boldsymbol{\mu}))_{L^2(\mathcal{P})} = \int_{\mathcal{P}} (\mathbf{u}_h(\boldsymbol{\mu}), \boldsymbol{\zeta}_i)_2 \boldsymbol{\zeta}_i g(\boldsymbol{\mu}) d\boldsymbol{\mu} \\ &= \int_{\mathcal{P}} \mathbb{V}\mathbb{V}^T \mathbf{u}_h(\boldsymbol{\mu}) g(\boldsymbol{\mu}) d\boldsymbol{\mu} \quad \forall g \in L^2(\mathcal{P}). \end{aligned}$$

Then

$$\|T - T_N\|_{HS}^2 = \|T - \mathbb{V}\mathbb{V}^T T\|_{HS}^2 = \int_{\mathcal{P}} \|\mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{V}\mathbb{V}^T \mathbf{u}_h(\boldsymbol{\mu})\|_2^2 d\boldsymbol{\mu}.$$

Since

$$\|T - T_N\|_{HS}^2 = \min_{B \in \mathcal{B}_N} \|T - B\|_{HS}^2 \leq \min_{\mathbb{W} \in \mathcal{V}_N} \|T - \mathbb{W}\mathbb{W}^T T\|_{HS}^2,$$

and $T_N = \mathbb{V}\mathbb{V}^T T$, we conclude that

$$\|T - \mathbb{V}\mathbb{V}^T T\|_{HS}^2 = \min_{\mathbb{W} \in \mathcal{Y}_N} \|T - \mathbb{W}\mathbb{W}^T T\|_{HS}^2. \quad \square$$

Proceeding as in Sect. 6.3.2, we can generalize Proposition 6.3 to the case where the \mathbb{X}_h norm, rather than the Euclidean norm, is considered.

Remark 6.3 (Relation between Kolmogorov N -width and POD). As already noticed, the Kolmogorov N -width defined by (5.21) measures the extent at which \mathcal{M}_h can be approximated – with respect to the $L^\infty(\mathcal{P}; V_h)$ norm – by N -dimensional subspaces of V_h . On the other hand, the POD basis \mathbb{V} of Proposition 6.3 enjoys a $L^2(\mathcal{P}; V_h)$ optimality property. Denoting by V_N the subspace generated by the columns of $\mathbb{W} \in \mathcal{Y}_N^{\mathbb{X}_h}$, the quantity

$$\delta_2(\mathcal{M}_h; V_N) = \left(\int_{\mathcal{P}} \|\mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{P}_W^{\mathbb{X}_h} \mathbf{u}_h(\boldsymbol{\mu})\|_{\mathbb{X}_h}^2 d\boldsymbol{\mu} \right)^{1/2} = \|\mathbf{u}_h - \mathbb{P}_W^{\mathbb{X}_h} \mathbf{u}_h\|_{L^2(\mathcal{P}; V_h)}$$

can be regarded as an $L^2(\mathcal{P}; V_h)$ -analogue of the deviation $d(\mathcal{M}_h; V_N)$ of \mathcal{M}_h from V_N (see Sect. 5.4). Consequently, the POD basis \mathbb{V} generates an optimal N -dimensional subspace for $\delta_{N,2}(\mathcal{M}_h; V_h)$, where

$$\delta_{N,2}(\mathcal{M}_h; V_h) = \inf_{\substack{V_N = \text{span}(\mathbb{W}) \\ \mathbb{W} \in \mathcal{Y}_N^{\mathbb{X}_h}}} \delta_2(\mathcal{M}_h; V_N).$$

Moreover,

$$\begin{aligned} \delta_{N,2}(\mathcal{M}_h; V_h) &\leq \inf_{\substack{V_N = \text{span}(\mathbb{W}) \\ \mathbb{W} \in \mathcal{Y}_N^{\mathbb{X}_h}}} \left(\int_{\mathcal{P}} d\boldsymbol{\mu} \right)^{1/2} \|\mathbf{u}_h - \mathbb{P}_W^{\mathbb{X}_h} \mathbf{u}_h\|_{L^\infty(\mathcal{P}; V_h)} \\ &= |\mathcal{P}|^{1/2} d_N(\mathcal{M}_h; V_h), \end{aligned} \quad (6.31)$$

where $|\mathcal{P}|$ represents the Lebesgue measure of \mathcal{P} . •

6.5 Back to the Discrete Setting

Thanks to the previous analysis of the \mathcal{P} -continuous version of the POD, we can now answer the questions raised at the beginning of Sect. 6.4 concerning approximation and sampling properties of POD basis functions. The key is to regard the discrete minimization problem (6.11) as an approximation of the continuous minimization problem (6.17). To this end, we introduce a suitable quadrature formula

$$\int_{\mathcal{P}} f(\boldsymbol{\mu}) d\boldsymbol{\mu} \approx \sum_{i=1}^{n_s} w_i f(\boldsymbol{\mu}^i), \quad (6.32)$$

to approximate the integrals over \mathcal{P} , for any continuous function $f: \mathcal{P} \rightarrow \mathbb{R}$, where $w_i > 0$ and $\boldsymbol{\mu}^i$ are suitable quadrature weights and quadrature points, respectively. Then,

$$\int_{\mathcal{P}} \|\mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{W}\mathbb{W}^T \mathbf{u}_h(\boldsymbol{\mu})\|_2^2 d\boldsymbol{\mu} \approx \sum_{i=1}^{n_s} w_i \|\mathbf{u}_h(\boldsymbol{\mu}^i) - \mathbb{W}\mathbb{W}^T \mathbf{u}_h(\boldsymbol{\mu}^i)\|_2^2. \quad (6.33)$$

The parameter samples location are thus defined by a suitable quadrature formula in the parameter space. Indeed, the right-hand side of (6.33) differs from (6.11) only by the weighting coefficients w_i . However, by defining $\mathbb{D} = \text{diag}(w_1, \dots, w_{n_s}) \in \mathbb{R}^{n_s \times n_s}$,

$$\sum_{i=1}^{n_s} w_i \|\mathbf{u}_h(\boldsymbol{\mu}^i) - \mathbb{W}\mathbb{W}^T \mathbf{u}_h(\boldsymbol{\mu}^i)\|_2^2 = \|\mathbb{S}\mathbb{D}^{1/2} - \mathbb{W}\mathbb{W}^T \mathbb{S}\mathbb{D}^{1/2}\|_F^2. \quad (6.34)$$

The POD basis associated to the snapshots set $\mathcal{M}_h^s = \{\mathbf{u}_h(\boldsymbol{\mu}^1), \dots, \mathbf{u}_h(\boldsymbol{\mu}^{n_s})\}$ is given by $\mathbb{V} = [\boldsymbol{\zeta}_1 \mid \dots \mid \boldsymbol{\zeta}_N]$, with $\boldsymbol{\zeta}_i = \frac{1}{\sigma_i} \mathbb{S}\mathbb{D}^{1/2} \tilde{\boldsymbol{\psi}}_i$ and

$$\tilde{\mathbb{S}}^T \tilde{\mathbb{S}} \tilde{\boldsymbol{\psi}}_i = \sigma_i^s \tilde{\boldsymbol{\psi}}_i, \quad \text{with} \quad \tilde{\mathbb{S}} = \mathbb{S}\mathbb{D}^{1/2}.$$

The complete procedure to compute \mathbb{V} is summarized in Algorithm 6.3.

Algorithm 6.3 POD algorithm with energy norm and quadratures weights

```

1: function  $\mathbb{V} = \text{POD}(\mathbb{S}, \mathbb{X}_h, \mathbb{D}, \varepsilon_{\text{POD}})$ 
2:   if  $n_s \leq N_h$  then
3:     set  $\tilde{\mathbb{S}} = \mathbb{S}\mathbb{D}^{1/2}$ 
4:     form the correlation matrix  $\tilde{\mathbb{C}} = \tilde{\mathbb{S}}^T \mathbb{X}_h \tilde{\mathbb{S}}$ 
5:     solve the eigenvalue problem  $\tilde{\mathbb{C}} \tilde{\boldsymbol{\psi}}_i = \sigma_i^2 \tilde{\boldsymbol{\psi}}_i, \quad i = 1, \dots, r$ 
6:     set  $\boldsymbol{\zeta}_i = \frac{1}{\sigma_i} \tilde{\mathbb{S}} \tilde{\boldsymbol{\psi}}_i$ 
7:   else
8:     form the matrix  $\tilde{\mathbb{K}} = \mathbb{X}_h^{1/2} \mathbb{S} \mathbb{D} \mathbb{S}^T \mathbb{X}_h^{1/2}$ 
9:     solve the eigenvalue problem  $\tilde{\mathbb{K}} \tilde{\boldsymbol{\zeta}}_i = \sigma_i^2 \tilde{\boldsymbol{\zeta}}_i, \quad i = 1, \dots, r$ 
10:    set  $\boldsymbol{\zeta}_i = \mathbb{X}_h^{-1/2} \tilde{\boldsymbol{\zeta}}_i$ 
11:   end if
12:   define  $N$  as the minimum integer such that  $I(N) \geq 1 - \varepsilon_{\text{POD}}$ 
13:    $\mathbb{V} = [\boldsymbol{\zeta}_1 \mid \dots \mid \boldsymbol{\zeta}_N]$ 
14: end function
```

Let us now denote by \mathbb{V}^∞ the POD basis solution of the *continuous* minimization problem (6.29) and by \mathbb{V}^{n_s} the POD basis minimizing (6.34); moreover we define

$$\mathcal{E}(\mathbb{W}) = \int_{\mathcal{P}} \|\mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{W}\mathbb{W}^T \mathbf{u}_h(\boldsymbol{\mu})\|_2^2 d\boldsymbol{\mu},$$

$$\mathcal{E}^s(\mathbb{W}) = \sum_{i=1}^{n_s} w_i \|\mathbf{u}_h(\boldsymbol{\mu}^i) - \mathbb{W}\mathbb{W}^T \mathbf{u}_h(\boldsymbol{\mu}^i)\|_2^2.$$

Thanks to the optimality result (6.29) it follows that $\mathcal{E}(\mathbb{V}^\infty) \leq \mathcal{E}(\mathbb{V}^{n_s})$. Furthermore,

$$\mathcal{E}(\mathbb{V}^{n_s}) \leq |\mathcal{E}(\mathbb{V}^{n_s}) - \mathcal{E}^S(\mathbb{V}^{n_s})| + \mathcal{E}^S(\mathbb{V}^{n_s}) \leq E_s + \sum_{i=N+1}^r (\sigma_i^S)^2, \quad (6.35)$$

where E_s denotes the quadrature (or sampling) error. Some remarks are in order:

- the retained energy criterion (6.12) can serve as a reliable estimate of the projection error $\mathcal{E}(\mathbb{V}^{n_s})$, provided an appropriate sampling of the parameter space is performed. In fact, if E_s is smaller than the truncation error (i.e. the second term) in (6.35), then

$$\mathcal{E}(\mathbb{V}^{n_s}) \lesssim \sum_{i=N+1}^r (\sigma_i^S)^2$$

and therefore

$$\frac{\mathcal{E}(\mathbb{V}^{n_s})}{\|\mathbf{u}_h\|_{L^2(\mathcal{P}; \mathbb{R}^{N_h})}^2} \lesssim \varepsilon_{\text{POD}}^2;$$

- the quadrature error E_s depends on: (i) the quadrature formula chosen, (ii) the number of quadrature points, (iii) the smoothness of the integrand, (iv) the dimension P of the parameter space. Moreover, suitable estimates of E_s are available depending on the quadrature formula used;
- different sampling strategies can be employed depending on the dimension of the parameter space. While tensorial (also called full factorial) sampling is suitable for low-dimensional problems (typically for $P \leq 3$), statistical methods like random (Monte Carlo) or latin hypercube (LHS) sampling (see, e.g., [69, 167]) and sparse grids (see, e.g., [114, 46]) are recommended as soon as the dimension of the parameter space gets large. A comparison between different types of sampling strategies for a two dimensional parameters space is reported in Fig. 6.8.

6.6 Our Illustrative Numerical Example Revisited

We now use POD to construct a RB approximation to the steady heat conduction-convection problem introduced in Sect. 3.8. We recall that $P = 4$ and $\mathcal{P} = [0, 12]^3 \times [1, 600]$. To begin with, we compute a set of snapshots by solving the high-fidelity approximation of problem (3.88) for $n_s \approx 400$ parameter configurations, selected by means of different sampling strategies. We start by considering four different tensorial grids: the first one is made of $4^3 \times 6$ equally spaced points, the second one of $3^3 \times 15$ equally spaced points, the third one of $3^3 \times 15$ points equally spaced in the first three directions and log spaced in the last one, and finally the fourth grid is made of $3^3 \times 15$ Clenshaw-Curtis points [67, 114]. As shown in Fig. 6.9(a), using a larger number of points in the direction of μ_4 is crucial in order to better capture the variability of the solution over \mathcal{P} .

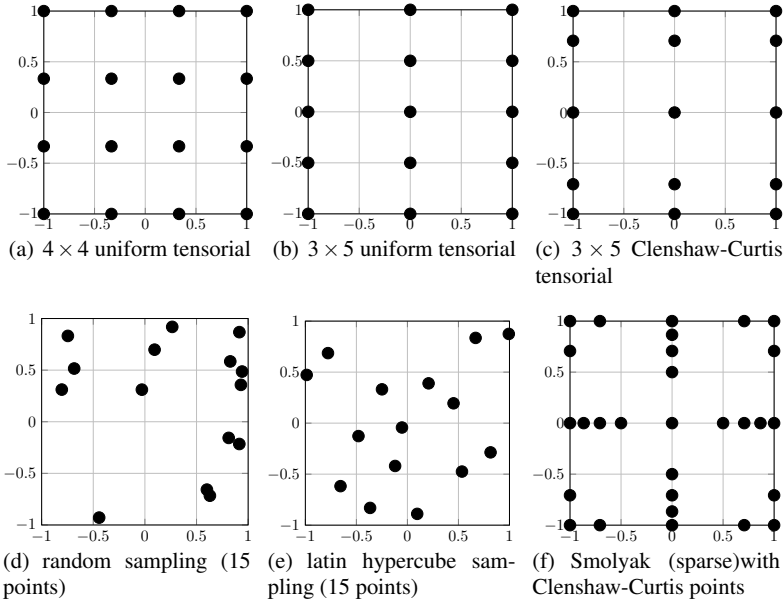


Fig. 6.8 Tensorial, sparse and statistical strategies to sample the parameter domain $\mathcal{P} = [-1, 1]^2$

We then consider a (Smolyak) sparse grid made of 389 Clenshaw-Curtis points and two additional sample sets of 405 points generated by random and latin hypercube sampling. For the problem at hand, these two latter yield the most accurate approximation of the spectrum of the solution set (see Fig. 6.9(b)).

To investigate the influence of the number of samples n_s on the singular values of the correlation matrix, we report in Fig. 6.10 the spectrum obtained using latin hypercube sampling with $n_s = 25, 50, 100, 200, 5000$. Since the spectrum obtained for $n_s = 100$ is already sufficiently accurate, we use the corresponding snapshot set to extract a reduced basis. With a tolerance $\varepsilon_{\text{POD}} = 10^{-5}$, the relative information content criterion yields a basis \mathbb{V} of dimension $N = 26$. Once the reduced model is built, we compute the ingredients required by the online evaluation of the a posteriori error estimate, as described in Algorithm 6.4.

6.7 More on Reducibility

POD is a valuable computational tool to understand if a problem is reducible or not. In Sect. 5.2 we have discussed some key aspects which play a potential role in the problem reducibility and can be evaluated without relying on the construction of a set of snapshots of the high-fidelity problem. Here we complement our discussion with some observations which can be performed once a high-fidelity approximation has been constructed.

Algorithm 6.4 RB approximation construction by POD algorithm**Input:** Tolerance ε_{POD} , train sample $\Xi_s \subset \mathcal{P}$

- 1: **for** $\mu \in \Xi_s$
- 2: $\mathbf{u}_h(\mu) = \text{SOLVEHFSYSTEM}(\mathbb{A}_h^q, \mathbf{f}_h^q, \theta_a^q, \theta_f^q, \mu)$
- 3: $\mathbb{S} \leftarrow [\mathbb{S} \ \mathbf{u}_h(\mu)]$
- 4: **end for**
- 5: $\mathbb{V} = \text{POD}(\mathbb{S}, \mathbb{X}_h, \varepsilon_{\text{POD}}, \mathbb{D})$
- 6: $[\mathbb{A}_N^q, \mathbf{f}_N^q] = \text{PROJECTSYSTEM}(\mathbb{A}_h^q, \mathbf{f}_h^q, \mathbb{V}, \mathbb{X}_h, \text{method})$
- 7: $[C_{q_1, q_2}, \mathbf{d}_{q_1, q_2}, \mathbb{E}_{q_1, q_2}] = \text{OFFLINERESIDUAL}(\mathbb{A}_h^q, \mathbf{f}_h^q, \mathbb{X}_h, \mathbb{V})$

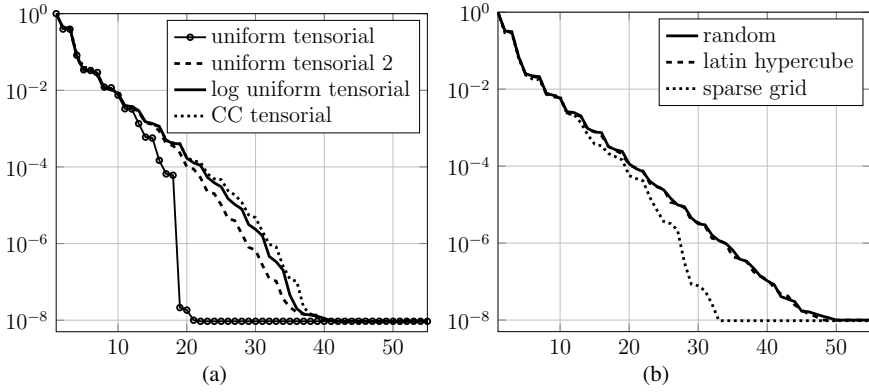


Fig. 6.9 First 55 singular values of the correlation matrix corresponding to different sampling strategies (the singular values are normalized with respect to σ_1)

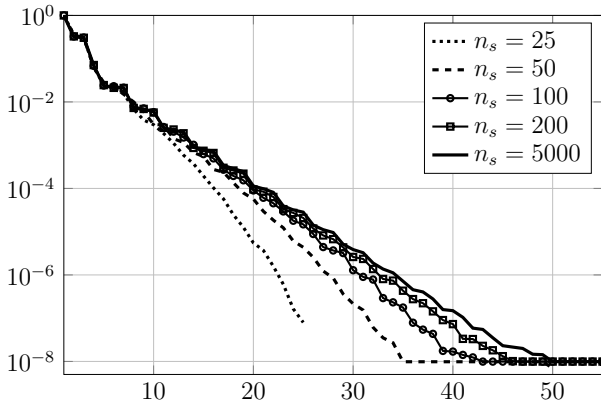


Fig. 6.10 First 55 singular values of the correlation matrix obtained by latin hypercube sampling with $n_s = 25, 50, 100, 200, 5000$

This can also be considered as a preprocessing step in the entire workflow of ROM construction, to better understand if a problem is reducible or not. Inspired by the POD construction of a RB space, to have a hint on whether the solution set \mathcal{M}_h is a low-dimensional manifold, we proceed as follows:

1. we collect (several) snapshots of the high-fidelity system into a matrix \mathbb{S} ;
2. we perform POD by using the SVD of \mathbb{S} (or the eigenpairs of the correlation \mathbb{C});
3. we check if the decay of the singular values is sufficiently rapid.

In the affirmative case, a limited number of modes will potentially suffice to represent the solution set; building a ROM is strongly advised. We can gain more insight by considering a problem for which very few basis functions suffice to recover the solution, and then a problem for which reduction cannot be performed.

First, we consider the one-dimensional elliptic equation

$$\begin{aligned} -(1+\mu)u''(x) &= 1, & x \in (0, 1) \\ u(0) &= 0, \quad u(1) = 1, \end{aligned} \tag{6.36}$$

whose exact solution is $u(x) = ((3+2\mu)x - x^2)/2(1+\mu)$; we compute $n_s = 500$ snapshots with $\mu \in [10^{-3}, 10]$. The μ -dependent solution can be represented up to roundoff precision with $N = 2$ basis functions, see Fig. 6.11(a); indeed, \mathcal{M} is a manifold of dimension 2, since all its elements can be obtained as linear combinations of any two snapshots u_0, u_1 (see Exercise 6). Then, we consider the one-dimensional linear transport equation

$$\begin{aligned} \partial_t u(x, t) + c \partial_x u(x, t) &= 0, & (x, t) \in \mathbb{R} \times (0, T) \\ u(x, 0) &= u_0(x), & x \in \mathbb{R} \end{aligned} \tag{6.37}$$

whose exact solution is $u(x, t) = u_0(x - ct)$; here we take $c = T = 1$ and $u_0(x) = (1/\sqrt{2\pi\sigma})e^{-x^2/2\sigma}$. We compute $n_s = 100$ snapshots in *time*, $u(x, t^1), \dots, u(x, t^{n_s})$ being $\{t^1, \dots, t^{n_s}\}$ a random sample. In Fig. 6.11(b) we plot the eigenvalues of the correlation matrix of snapshots for different values of $\sigma = 10^{-\beta}$, $\beta = 1, 2, \dots, 6$. Depending on the value of σ , eigenvalues show a fast decay ($\sigma = 10^{-1}, 10^{-2}$), a slow decay ($\sigma = 10^{-3}, 10^{-4}$) or even no decay at all ($\sigma = 10^{-5}, 10^{-6}$).

Thanks to the relation between Kolmogorov N -width and POD (see Remark 6.3), checking the decay of the eigenvalues of the correlation matrix \mathbb{C} built from a set of snapshots (equivalently, the singular values of \mathbb{S}) is a way to assess the reducibility of a given problem: the larger the information provided by the sampled snapshots to represent the whole solution set (in this case the eigenvalues of \mathbb{C} decay fast), the more feasible the reduction process.

However, it is easy to construct examples where the singular values of computed snapshots do not decay. For instance, consider again problem (6.37); take n_s snapshots of this solution at times $t = 0, \Delta t, 2\Delta t, \dots, (N_s - 2)\Delta t, T$. Assume that $u_0 \in L^2(\mathbb{R})$ and such that the measure of its support is $\text{supp } u_0 < |c|\Delta t/2$.

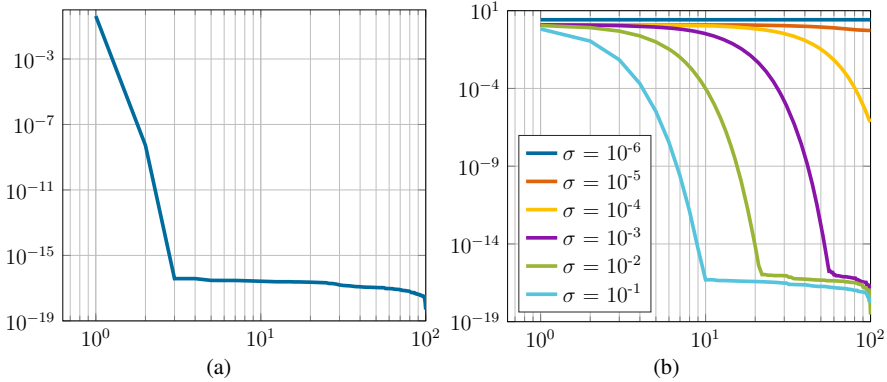


Fig. 6.11 First 100 eigenvalues of the correlation matrix obtained from $n_s = 100$ snapshots of the elliptic equation (6.36) (left) and from $n_s = 500$ time snapshots of the linear transport equation (6.37) with $\sigma = 10^{-\beta}$, $\beta = 1, 2, \dots, 6$ (right)

Thus,

$$\int_{\mathbb{R}} u_j(\xi) u_k(\xi) d\xi = \int_{\mathbb{R}} u_0(\xi - c j \Delta t) u_0(\xi - c k \Delta t) d\xi = \|u_0\|_{L^2(\mathbb{R})} \delta_{jk}. \quad (6.38)$$

The correlation matrix is diagonal with all eigenvalues equal – hence, they do not decay at all. This provides an example of the reason why reduced-order modeling for (several) highly advection dominated problems (or problems featuring traveling waves and more in general hyperbolic problems) is still a remarkably challenging task.

6.8 Exercises

1. Show that $\mathbb{A}^T \mathbb{A} \mathbf{v}_i = \sigma_i^2 \mathbf{v}_i$, $\mathbb{A} \mathbb{A}^T \mathbf{u}_i = \sigma_i^2 \mathbf{u}_i$ for any $i = 1, \dots, n$, whence $\lambda_i(\mathbb{A}^T \mathbb{A}) = \lambda_i(\Sigma^2) = (\sigma_i(\mathbb{A}))^2$.
2. Show that if $\mathbb{A} \in \mathbb{R}^{n \times n}$ is symmetric with eigenvalues given by $\lambda_1(\mathbb{A}) \geq \lambda_2(\mathbb{A}) \geq \dots \geq \lambda_n(\mathbb{A})$, then $\sigma_i(\mathbb{A}) = |\lambda_i(\mathbb{A})|$.
3. Show that if $\mathbb{A} \in \mathbb{R}^{m \times n}$, then

$$\sigma_{\max}(\mathbb{A}) = \max_{\mathbf{y} \in \mathbb{R}^m, \mathbf{x} \in \mathbb{R}^n} \frac{\mathbf{y}^T \mathbb{A} \mathbf{x}}{\|\mathbf{x}\| \|\mathbf{y}\|}.$$

4. Show that for any matrix $\mathbb{A} \in \mathbb{R}^{m \times n}$ of rank n , $\|\mathbb{A}(\mathbb{A}^T \mathbb{A})^{-1} \mathbb{A}^T\| = 1$.

5. Prove that the Moore-Penrose pseudo-inverse matrix \mathbb{A}^\dagger is the unique minimizer of the functional $\|\mathbb{A}\mathbb{X} - \mathbb{I}_m\|$ where $\|\cdot\|_F$ is the Frobenius norm (6.3), that is

$$\mathbb{A}^\dagger = \arg \min_{\mathbb{X} \in \mathbb{R}^{n \times n}} \|\mathbb{A}\mathbb{X} - \mathbb{I}_m\|_F.$$

6. Show that the solution to (6.36) is $u(\mu) = ((3 + 2\mu)x - x^2)/2(1 + \mu)$. Then, prove that $\mathcal{M} = \{u(\mu) : \mu \in \mathcal{P}\}$ is a manifold of dimension 2, by showing that $u(\mu) = \omega_0(\mu)u(0) + \omega_1(\mu)u(1)$ for any $\mu \in \mathcal{P}$, and compute $\omega_0(\mu)$, $\omega_1(\mu)$.

Chapter 7

Construction of RB Spaces by the Greedy Algorithm

We describe another, indeed very popular, approach to construct a reduced basis space in the context of parametrized PDEs: the so-called greedy algorithm. This consists in an iterative sampling from the parameter space fulfilling at each step a suitable optimality criterion that relies on the a posteriori error estimate. We illustrate the main features of this procedure in the algebraic RB framework, and then address its continuous counterpart, discuss some a priori convergence results and verify them using numerical tests.

7.1 Greedy Algorithm: an Algebraic Perspective

In the case of parametrized PDEs, the POD method for the construction of a RB space might entail a severe computational cost, due to the evaluation of a large number n_s of snapshots of the high-fidelity problem. In successful cases, the number N of modes extracted from the POD algorithm to form the RB space will be much smaller than n_s . If the solution of the high-fidelity problem is computationally demanding, performing POD can be excessively expensive.

Greedy algorithms represent an efficient alternative to POD, as they allow the construction of the reduced space by minimizing the amount of snapshots to be evaluated. The goal of such a procedure is to evaluate N snapshots to construct a RB space of dimension N , by seeking at each step the local optimum. We warn however the reader about two potentially critical aspects:

1. to be efficient, a greedy algorithm must be supported by an a posteriori error estimate for the error $\|u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_V$, whose evaluation must be performed in a very inexpensive way for any $\boldsymbol{\mu} \in \mathcal{P}$;
2. even in presence of such an error bound, a greedy algorithm is not necessarily cheaper than POD, since at each step a (possibly demanding) maximization problem has to be solved. Lowering the computational complexity in this respect is thus essential to make greedy algorithms competitive with POD.

7.1.1 The Idea Behind Greedy Algorithms

In numerical optimization, the term *greedy algorithm* is associated to any kind of technique which achieves an optimal solution to a problem by choosing at every iteration an element fulfilling a local optimality criterion. Since their introduction in the early 70s [97], greedy algorithms have been extensively analyzed and applied for tackling combinatorial optimization problems, as well as problems arising in operations research and decision theory; see, e.g., [73, Chap. 16].

In the context of RB methods, a greedy algorithm is a procedure for the construction of a subspace by iteratively adding a new basis vector at each step, instead than optimizing over all possible N -dimensional subspaces. More precisely, at the generic iteration $1 \leq n \leq N-1$ we assume that we are given a sample set

$$S_n = \{\boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^n\},$$

the corresponding subspace

$$V_n = \text{span}\{u_h(\boldsymbol{\mu}^1), \dots, u_h(\boldsymbol{\mu}^n)\} \quad (7.1)$$

and an orthonormal basis

$$\mathbb{V} = [\boldsymbol{\zeta}_1 \mid \dots \mid \boldsymbol{\zeta}_n] \in \mathbb{R}^{N_h \times n} \quad (7.2)$$

for V_n . The latter is obtained by orthonormalization of the snapshot set

$$\mathbb{S} = [\mathbf{u}_h(\boldsymbol{\mu}^1) \mid \dots \mid \mathbf{u}_h(\boldsymbol{\mu}^n)] \in \mathbb{R}^{N_h \times n},$$

where $(\mathbf{u}_h(\boldsymbol{\mu}^i))_j = u_h^{(j)}(\boldsymbol{\mu}^i)$, for $1 \leq j \leq N_h$, $1 \leq i \leq n$. Then we set

$$\boldsymbol{\mu}^{n+1} = \arg \max_{\boldsymbol{\mu} \in \mathcal{P}} \|\mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{V}\mathbf{u}_n(\boldsymbol{\mu})\|_{\mathbb{X}_h}, \quad (7.3)$$

where $\mathbf{u}_n(\boldsymbol{\mu}) \in \mathbb{R}^n$ denotes the solution of the RB problem (3.26). In other words, at each step the retained snapshot is the element of the solution set which is worst approximated by the current RB approximation.

As we will see in the following section, even though a greedy strategy is very effective, the cost entailed by solving problem (7.3) is actually very high, thus making the procedure outlined so far computationally prohibitive.

7.1.2 The Weak Greedy Algorithm

Originally introduced in [214, 215], the weak greedy algorithm represents nowadays a standard RB technique for parametrized PDEs.

This algorithm is obtained by replacing, in (7.3):

- the parameter set \mathcal{P} over which we should in principle find the supremum with a very fine sample $\mathcal{E}_{\text{train}} \subset \mathcal{P}$, of cardinality $|\mathcal{E}_{\text{train}}| = n_{\text{train}}$. Such a training sample serves to select our RB space – or train our RB approximation. This nevertheless still requires solving several high-fidelity approximation problems;
- the approximation error $\|\mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{V}\mathbf{u}_n(\boldsymbol{\mu})\|_{\mathbb{X}_h}$ with the a posteriori error estimator $\Delta_n(\boldsymbol{\mu})$ built in Sect. 3.7 such that

$$\|\mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{V}\mathbf{u}_n(\boldsymbol{\mu})\|_{\mathbb{X}_h} \leq \Delta_n(\boldsymbol{\mu}) \quad \forall \boldsymbol{\mu} \in \mathcal{P}.$$

Introducing the training sample enables to turn the problem of finding the maximum of the error bound over \mathcal{P} into a simpler enumeration problem – that is, by listing the values of the error bounds in decreasing order and selecting the largest one. In fact, performing a numerical optimization at each step of the procedure could be quite expensive because we can only determine local but not global solutions.

Hence, at each step $n = 1, \dots, N-1$ of the (weak) greedy algorithm, we need to:

1. evaluate the a posteriori error bound $\Delta_n(\boldsymbol{\mu})$ (referred to the n -dimensional RB approximation built over V_n) for any $\boldsymbol{\mu} \in \mathcal{E}_{\text{train}}$;
2. find, by solving an enumeration problem,

$$\boldsymbol{\mu}^{n+1} = \arg \max_{\boldsymbol{\mu} \in \mathcal{E}_{\text{train}}} \Delta_n(\boldsymbol{\mu}).$$

In other words, at the n -th iteration of this algorithm to the *retained* snapshots, over all possible candidate $\mathbf{u}_h(\boldsymbol{\mu})$, $\boldsymbol{\mu} \in \mathcal{E}_{\text{train}}$, we append the particular candidate snapshot that the a posteriori error bound predicts to be the worst approximated by the RB prediction associated to V_n . Then, the final size N of the RB space V_N is such that

$$\max_{\boldsymbol{\mu} \in \mathcal{E}_{\text{train}}} \Delta_N(\boldsymbol{\mu}) \leq \varepsilon_g$$

where $\varepsilon_g > 0$ is a prescribed, sufficiently small, stopping tolerance.

It is now clear why the a posteriori error bound must be computable in a very inexpensive way: the (weak) greedy algorithm requires only $O(N)$ calls to the high-fidelity solver, but yields $O(Nn_{\text{train}})$ evaluations of the a posteriori error bound – each one requiring the solution of a RB problem. The algorithmic details are reported in Algorithm 7.1 (see also Fig. 7.1). The basis \mathbb{V} is kept orthonormal (with respect to $(\cdot, \cdot)_{V_h}$) by iteratively orthonormalizing the new element appended to the existing basis through a Gram-Schmidt procedure (see Algorithm 7.2). Further details can be found in [231], where a similar procedure is developed also with respect to the energy norm. This is particularly relevant in the compliant case, since the error in the energy norm is directly related to the error in the output (see (3.89)).

Remark 7.1. Other estimators, such as the relative error bound $\Delta_N(\boldsymbol{\mu})/\|\mathbf{u}_N(\boldsymbol{\mu})\|_V$, or error bounds for some outputs of interest, can be used to evaluate the accuracy of the RB space. In any case, evaluating such a criterion must be inexpensive. •

Algorithm 7.1 (Weak) greedy algorithm

Input: Maximum number of iterations N_{\max} , stopping tolerance ε_g , train sample $\mathcal{E}_{\text{train}} \subset \mathcal{P}$, starting point $\boldsymbol{\mu}^1 \in \mathcal{P}$

Output: Basis $\mathbb{V} \in \mathbb{R}^{N_h \times N}$

```

1:  $\mathcal{E}_g = [], \mathbb{V} = []$ 
2:  $N = 0, \delta_0 = \varepsilon_g + 1$ 
3: while  $N < N_{\max}$  and  $\delta_N > \varepsilon_g$ 
4:    $N \leftarrow N + 1$ 
5:   compute  $\mathbf{u}_h(\boldsymbol{\mu}^N)$ 
6:    $\boldsymbol{\zeta}_N = \text{GRAMSCHMIDT}(\mathbb{V}, \mathbf{u}_h(\boldsymbol{\mu}^N), \mathbb{X}_h)$ 
7:    $\mathbb{V} \leftarrow [\mathbb{V} \boldsymbol{\zeta}_N]$ 
8:    $\mathcal{E}_g \leftarrow \mathcal{E}_g \cup \{\boldsymbol{\mu}^N\}$ 
9:    $[\delta_N, \boldsymbol{\mu}^{N+1}] = \max_{\boldsymbol{\mu} \in \mathcal{E}_{\text{train}}} \Delta_N(\boldsymbol{\mu})$ 
10: end while

```

Algorithm 7.2 Gram-Schmidt orthonormalization

```

1: function  $\mathbf{z} = \text{GRAMSCHMIDT}(\mathbb{V}, \mathbf{u}, \mathbb{X})$ 
2:   if  $\mathbb{V} = []$  then
3:      $\mathbf{z} = \mathbf{u}$ 
4:   else
5:      $\mathbf{z} = \mathbf{u} - \mathbb{V}\mathbb{V}^T \mathbb{X}\mathbf{u}$ 
6:   end if
7:    $\mathbf{z} \leftarrow \mathbf{z} / \|\mathbf{z}\|_{\mathbb{X}}$ 
8: end function

```

Remark 7.2. The choice of a good training sample is a delicate issue. In fact, $\mathcal{E}_{\text{train}}$ should be (i) small for efficiency reasons, but (ii) sufficiently large in order to represent the parameter set as well as possible. As a matter of fact, the performance of the greedy algorithm – and, at a larger extent, that of the RB method – crucially depends on how well the training sample is chosen. Typically, these samples are chosen by Monte Carlo methods with respect to a uniform or log-uniform density. A generalization of the (weak) greedy algorithm relying on local approximation spaces (over \mathcal{P}) is proposed in [179]. This local adaptive greedy algorithm enables to account for solutions’ anisotropy (with respect to parameters) through adaptive training sets and an empirically built, problem dependent distance function which can be evaluated on the fly. Another approach to tackle greedy sampling of higher-dimensional functions is proposed in [133]. First ideas in local *hp* RB methods featuring basis functions defined over the subsets of a partition of \mathcal{P} can be found in [101, 100]; see also [99, 102, 107, 128, 10] for further details. •

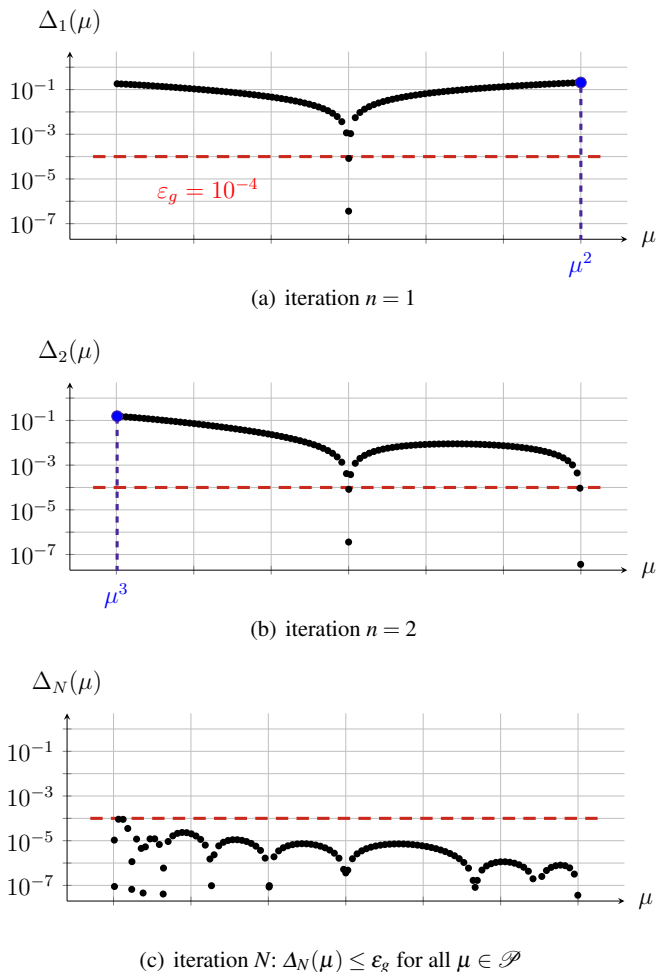


Fig. 7.1 The (weak) greedy algorithm convergence history. At the iteration N the N -th sampled parameter μ^N is selected

7.2 Our Illustrative Numerical Example Revisited

We use the greedy algorithm to construct a RB approximation to the steady heat conduction-convection problem introduced in Sect. 3.8. We first perform a preprocessing step to build the RBF interpolant (3.85) of the stability factor $\beta_h(\mu)$; since the problem is coercive, we use formula (2.43) to compute the stability factor at each interpolation point. As regards these latter, since $\beta_h(\mu)$ depends only on μ_4 , we choose equidistributed interpolation points along this direction.

Algorithm 7.3 ROM construction by greedy algorithm**Input:** Maximum number of iterations N_{\max} , stopping tolerance ε_g , train sample

 $\mathcal{E}_{\text{train}} \subset \mathcal{P}$, starting point $\boldsymbol{\mu}^1 \in \mathcal{P}$

- 1: $\mathcal{E}_g = []$, $\mathbb{V} = []$
- 2: $N = 0$, $\delta_0 = \varepsilon_g + 1$
- 3: **while** $N < N_{\max}$ and $\delta_N > \varepsilon_g$
- 4: $N \leftarrow N + 1$
- 5: $\mathbf{u}_h(\boldsymbol{\mu}^N) = \text{SOLVEHFSYSTEM}(\mathbb{A}_h^q, \mathbf{f}_h^q, \theta_a^q, \theta_f^q, \boldsymbol{\mu})$
- 6: $\boldsymbol{\zeta}_N = \text{GRAMSCHMIDT}(\mathbb{V}, \mathbf{u}_h(\boldsymbol{\mu}^N), \mathbb{X}_h)$
- 7: $\mathbb{V} \leftarrow [\mathbb{V} \ \boldsymbol{\zeta}_N]$
- 8: $\mathcal{E}_g \leftarrow \mathcal{E}_g \cup \{\boldsymbol{\mu}^N\}$
- 9: $[\mathbb{A}_N^q, \mathbf{f}_N^q] = \text{PROJECTSYSTEM}(\mathbb{A}_h^q, \mathbf{f}_h^q, \mathbb{V}, \mathbb{X}_h, \text{method})$
- 10: $[C_{q_1, q_2}, \mathbf{d}_{q_1, q_2}, \mathbb{E}_{q_1, q_2}] = \text{OFFLINERESIDUAL}(\mathbb{A}_h^q, \mathbf{f}_h^q, \mathbb{X}_h, \mathbb{V})$
- 11: **for** $\boldsymbol{\mu} \in \mathcal{E}_{\text{train}}$
- 12: $\mathbf{u}_N(\boldsymbol{\mu}) = \text{SOLVERBSYSTEM}(\mathbb{A}_N^q, \mathbf{f}_N^q, \theta_a^q, \theta_f^q, \boldsymbol{\mu}, \text{method})$
- 13: $\Delta_N(\boldsymbol{\mu}) = \text{ERRORESTIMATE}(C_{q_1, q_2}, \mathbf{d}_{q_1, q_2}, \mathbb{E}_{q_1, q_2}, \theta_a^q, \theta_f^q, \mathbf{u}_N(\boldsymbol{\mu}), \boldsymbol{\mu})$
- 14: **end for**
- 15: $[\delta_N, \boldsymbol{\mu}^{N+1}] = \max \Delta_N(\boldsymbol{\mu})$
- 16: **end while**

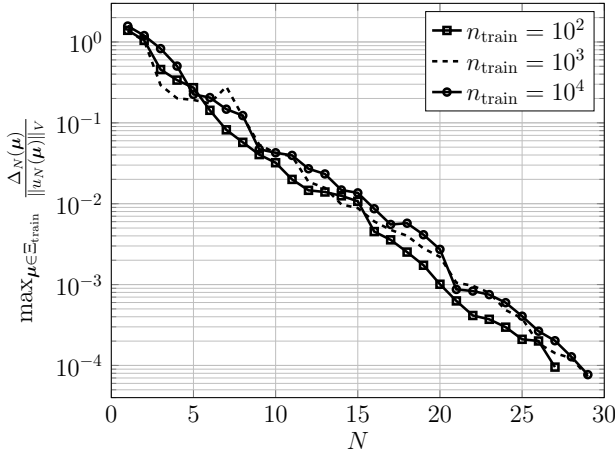


Fig. 7.2 Convergence history of the greedy algorithm: we report $\max_{\boldsymbol{\mu} \in \mathcal{E}_{\text{train}}} (\Delta_N(\boldsymbol{\mu}) / \|u_N(\boldsymbol{\mu})\|_V)$, $N = 1, \dots, 31$ for different training samples $\mathcal{E}_{\text{train}}$

We then build the G-RB approximation by means of the greedy algorithm 7.1 as detailed¹ in Algorithm 7.3, with $N_{\max} = 50$, $\varepsilon_g = 10^{-4}$, $\boldsymbol{\mu}^1 = [6 \ 6 \ 6 \ 300]$ and different training samples $\mathcal{E}_{\text{train}}$. In Fig. 7.2 we report the convergence history of the

¹ The notation \max at line 15 is the same as the function \max in MATLAB: $[M, \hat{x}] = \max f(x)$ implies $\hat{x} = \arg \max f(x)$ and $M = f(\hat{x})$.

greedy algorithm with different Ξ_{train} made of $n_{\text{train}} = 10^2, 10^3, 10^4$ points selected by latin hypercube sampling. The algorithm stops when $N = 27$ in the first case, while two additional iterations are required to achieve the tolerance ε_g in the second and third case. As regards the computational times, the algorithm takes 217, 239 and 275 seconds to run using the different grids, respectively².

7.3 An Abstract Formulation of the Greedy Algorithm

In order to carry out a convergence analysis of the greedy algorithm, it is useful to formulate this latter in a more general and abstract setting, as follows. Given a compact set K in a Hilbert space V , we seek functions $\{x_1, x_2, \dots, x_N\}$ such that each $x \in K$ is well approximated by the elements of the subspace $K_N = \text{span}\{x_1, \dots, x_N\}$. Given a tolerance $\varepsilon_g > 0$, the greedy algorithm for the selection of $\{x_1, \dots, x_N\}$ proceeds as follows:

$$\begin{aligned}
 & x_1 = \arg \max_{x \in K} \|x\|_V \\
 & \text{assume } x_1, \dots, x_{N-1} \text{ are defined} \\
 & \text{consider } K_{N-1} = \text{span}\{x_1, \dots, x_{N-1}\} \\
 & \text{define } x_N = \arg \max_{x \in K} d(x, K_{N-1}) \\
 & \text{iterate until } \max_{x \in K} d(x, K_N) < \varepsilon_g.
 \end{aligned} \tag{7.4}$$

Here $d(x, K_N)$ denotes the distance between an element $x \in K$ and the subspace K_N (see (5.17)). Since V is a Hilbert space, we have

$$d(x, K_N) = \|x - \Pi_{K_N} x\|_V, \tag{7.5}$$

where Π_{K_N} is the orthogonal projection operator on K_N with respect to the scalar product $(\cdot, \cdot)_V$. At each step, the elements provided by the previous algorithm are orthonormalized by the following Gram-Schmidt procedure:

$$\zeta_1 = \frac{x_1}{\|x_1\|_V}, \quad \zeta_n = \frac{x_n - \Pi_{K_{n-1}} x_n}{\|x_n - \Pi_{K_{n-1}} x_n\|_V}, \quad n = 2, \dots, N,$$

so that $K_N = \text{span}\{\zeta_1, \dots, \zeta_N\}$. In particular, for any $x \in V$,

$$\Pi_{K_N} x = \sum_{n=1}^N (x, \zeta_n)_V \zeta_n.$$

When K is the manifold \mathcal{M}_h given by (5.4), the algorithm (7.4) becomes:

² The loop over Ξ_{train} at lines 11-14 of Algorithm 7.3 was performed in parallel using 4 cores on a workstation with an Intel Core i5-2400S CPU and 16 GB of RAM.

$$\begin{aligned}
& \boldsymbol{\mu}^1 = \arg \max_{\boldsymbol{\mu} \in \mathcal{P}} \|u_h(\boldsymbol{\mu})\|_V \\
& \text{given the samples } \boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^{N-1} \\
& \text{consider } V_{N-1} = \text{span}\{u_h(\boldsymbol{\mu}^1), \dots, u_h(\boldsymbol{\mu}^{N-1})\} \\
& \text{define } \boldsymbol{\mu}^N = \arg \max_{\boldsymbol{\mu} \in \mathcal{P}} d(u_h(\boldsymbol{\mu}), V_{N-1}) \\
& \text{iterate until } \max_{\boldsymbol{\mu} \in \mathcal{P}} d(u_h(\boldsymbol{\mu}), V_N) < \varepsilon_g.
\end{aligned} \tag{7.6}$$

Since the distance between $u_h(\boldsymbol{\mu})$ and V_{N-1} is given by

$$d(u_h(\boldsymbol{\mu}), V_{N-1}) = \|u_h(\boldsymbol{\mu}) - \Pi_{V_{N-1}} u_h(\boldsymbol{\mu})\|_V, \tag{7.7}$$

at each step the retained snapshot $u_h(\boldsymbol{\mu}^N)$ is the element of the solution set which is worst approximated by its orthogonal projection onto V_{N-1} , rather than by the current RB approximation as in (7.3).

Remark 7.3. The algorithm (7.6) can be viewed as a practical way to tackle the problem (5.22) of determining the optimal N -dimensional subspace for the Kolmogorov N -width $d_N(\mathcal{M}_h; V_h)$. Indeed, we solve the minimization problem (5.22) through a greedy procedure, building iteratively a N -dimensional subspace. At each step $n = 1, \dots, N$ we select a single function to be included in the basis of the reduced space – which corresponds to the locally optimal choice – rather than performing the optimization over all possible N -dimensional subspaces. •

As already mentioned, algorithm (7.6) is computationally expensive: at each step, seeking the best snapshot entails solving an optimization problem, where computing the distance $d(u_h(\boldsymbol{\mu}), V_N)$ for any $\boldsymbol{\mu} \in \mathcal{P}$ requires many expensive evaluations of the high-fidelity solution $u_h(\boldsymbol{\mu})$. A variant of this greedy strategy is obtained by replacing the distance $d(u_h(\boldsymbol{\mu}), V_N)$ with a surrogate $s_N(\boldsymbol{\mu})$ such that

$$c_s s_N(\boldsymbol{\mu}) \leq d(u_h(\boldsymbol{\mu}), V_N) \leq C_s s_N(\boldsymbol{\mu}) \quad \forall \boldsymbol{\mu} \in \mathcal{P}, \tag{7.8}$$

for some positive constants $0 < c_s \leq C_s$. Then, the new sample is defined as

$$\boldsymbol{\mu}^N = \arg \max_{\boldsymbol{\mu} \in \mathcal{P}} s_{N-1}(\boldsymbol{\mu}), \tag{7.9}$$

yielding the following algorithm

$$\begin{aligned}
& \boldsymbol{\mu}^1 = \arg \max_{\boldsymbol{\mu} \in \mathcal{P}} \|u_h(\boldsymbol{\mu})\|_V \\
& \text{given the samples } \boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^{N-1} \\
& \text{consider } V_{N-1} = \text{span}\{u_h(\boldsymbol{\mu}^1), \dots, u_h(\boldsymbol{\mu}^{N-1})\} \\
& \text{define } \boldsymbol{\mu}^N = \arg \max_{\boldsymbol{\mu} \in \mathcal{P}} s_{N-1}(\boldsymbol{\mu}) \\
& \text{iterate until } \max_{\boldsymbol{\mu} \in \mathcal{P}} s_N(\boldsymbol{\mu}) < \varepsilon_g.
\end{aligned} \tag{7.10}$$

For instance, by defining

$$s_N(\boldsymbol{\mu}) = \|u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_V \quad (7.11)$$

we obtain the algorithm described in Sect. 7.1.1; recall that $\|u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_V$ is related to the distance $d(u_h(\boldsymbol{\mu}), V_N)$ through the a priori error estimate (3.33) provided by the Céa's Lemma. In this case, if $a(\cdot, \cdot; \boldsymbol{\mu})$ is coercive, we find (see Exercise 1)

$$c_s = \alpha_0 / \bar{\gamma}, \quad C_s = 1,$$

where we have assumed that $\gamma(\boldsymbol{\mu}) \leq \bar{\gamma}$ and $\alpha(\boldsymbol{\mu}) \geq \alpha_0$ for all $\boldsymbol{\mu} \in \mathcal{P}$. The computationally feasible variant provided by Algorithm 7.1 is instead obtained by defining the inexpensive surrogate

$$s_N(\boldsymbol{\mu}) = \Delta_N(\boldsymbol{\mu}). \quad (7.12)$$

If $a(\cdot, \cdot; \boldsymbol{\mu})$ is coercive, combining Céa's Lemma and (3.73) we find (see Exercise 1)

$$c_s = \alpha_0^2 / \bar{\gamma}^2, \quad C_s = 1.$$

Remark 7.4. Algorithm (7.10) finds efficiently an approximation to the optimal N -dimensional subspace for the Kolmogorov N -width $d_N(\mathcal{M}_h; V_h)$, as stated in problem (5.22). Indeed, we can derive the following relation between the Kolmogorov N -width and the criterion (7.9). By replacing in the definition (5.22) the deviation $d(\mathcal{M}_h; V_N)$ between \mathcal{M}_h and V_N with

$$\delta_s(\mathcal{M}_h; V_N) = \sup_{\boldsymbol{\mu} \in \mathcal{P}} s_N(\boldsymbol{\mu}) \quad (7.13)$$

we obtain

$$\delta_{N,s}(\mathcal{M}_h; V_h) = \inf_{\substack{V_N \subset V_h \\ \dim(V_N)=N}} \delta_s(\mathcal{M}_h; V_N) \quad (7.14)$$

as the counterpart of the Kolmogorov N -width $d_N(\mathcal{M}_h; V_h)$. Note that

$$d_N(\mathcal{M}_h; V_h) \leq \delta_{N,s}(\mathcal{M}_h; V_h) \leq \delta_s(\mathcal{M}_h; V_N). \quad \bullet$$

Algorithm (7.10) can be stated in a more general form by noting that, if $\boldsymbol{\mu}^N$ is chosen accordingly to (7.9), then

$$d(u_h(\boldsymbol{\mu}^N), V_{N-1}) \geq \rho \max_{\boldsymbol{\mu} \in \mathcal{P}} d(u_h(\boldsymbol{\mu}), V_{N-1}), \quad (7.15)$$

with $\rho = c_s / C_s \in (0, 1]$. In fact, by (7.8)

$$d(u_h(\boldsymbol{\mu}^N), V_{N-1}) \geq c_s s_{N-1}(\boldsymbol{\mu}^N) = c_s \max_{\boldsymbol{\mu} \in \mathcal{P}} s_{N-1}(\boldsymbol{\mu}) \geq \frac{c_s}{C_s} \max_{\boldsymbol{\mu} \in \mathcal{P}} d(u_h(\boldsymbol{\mu}), V_{N-1}).$$

Inequality (7.15) motivates the definition of an *abstract weak greedy algorithm* [30] that takes the following form (and reduces to (7.6) for $\rho = 1$):

$$\begin{aligned}
 &\text{choose } \boldsymbol{\mu}^1 \text{ s.t. } \|u_h(\boldsymbol{\mu}^1)\|_V \geq \rho \max_{\boldsymbol{\mu} \in \mathcal{P}} \|u_h(\boldsymbol{\mu})\|_V \\
 &\text{given the samples } \boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^{N-1} \\
 &\quad \text{consider } V_{N-1} = \text{span}\{u_h(\boldsymbol{\mu}^1), \dots, u_h(\boldsymbol{\mu}^{N-1})\} \\
 &\quad \text{choose } \boldsymbol{\mu}^N \text{ s.t. } d(u_h(\boldsymbol{\mu}^N), V_{N-1}) \geq \rho \max_{\boldsymbol{\mu} \in \mathcal{P}} d(u_h(\boldsymbol{\mu}), V_{N-1}) \\
 &\text{iterate until } \max_{\boldsymbol{\mu} \in \mathcal{P}} d(u_h(\boldsymbol{\mu}), V_N) < \varepsilon_g.
 \end{aligned} \tag{7.16}$$

7.4 A Priori Error Analysis

A priori error analysis aims at providing upper bounds for $d(\mathcal{M}_h; V_N)$ by comparison with the Kolmogorov N -width. The latter minimizes, among all N -dimensional subspaces, the projection error for the whole set \mathcal{M}_h . If $d(\mathcal{M}_h; V_N)$ decayed at a rate comparable to $d_N(\mathcal{M}_h; V_h)$, this would implicitly assure that the greedy algorithm provides the (asymptotically) best possible accuracy attainable by N -dimensional subspaces.

In the following we assume that the bilinear form $a(\cdot, \cdot; \boldsymbol{\mu})$ is continuous, symmetric and coercive for all $\boldsymbol{\mu} \in \mathcal{P}$, so that there exists $\tilde{\gamma}, \alpha_0 > 0$ such that $\gamma(\boldsymbol{\mu}) \leq \tilde{\gamma}$ and $\alpha(\boldsymbol{\mu}) \geq \alpha_0$ for all $\boldsymbol{\mu} \in \mathcal{P}$. In general the optimal subspace with respect to the Kolmogorov N -width is not spanned by elements of the set \mathcal{M}_h being approximated, thus we possibly have that $d_N(\mathcal{M}_h; V_h) \ll d(\mathcal{M}_h; V_N)$. The following result, shown in [40], provides an estimate for $d(\mathcal{M}_h; V_N)$ in terms of $d_N(\mathcal{M}_h; V_h)$.

Theorem 7.1. *If V_N is built using the greedy algorithm (7.6), then there exists $C > 0$ independent of N and $\boldsymbol{\mu}$ such that*

$$d(\mathcal{M}_h; V_N) = \sup_{\boldsymbol{\mu} \in \mathcal{P}} \|u_h(\boldsymbol{\mu}) - \Pi_{V_N} u_h(\boldsymbol{\mu})\|_V \leq C(N+1) \delta_0^{N+1} d_N(\mathcal{M}_h; V_h) \tag{7.17}$$

with $\delta_0 = 2$.

Thanks to the a priori error estimate (3.33), from (7.17) we obtain

$$\sup_{\boldsymbol{\mu} \in \mathcal{P}} \|u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_V \leq C \sqrt{\tilde{\gamma}/\alpha_0} (N+1) \delta_0^{N+1} d_N(\mathcal{M}_h; V_h).$$

Therefore, if the N -width converges at exponential rate, then also the error of the best approximation in V_N does, as stated in the following result [40].

Corollary 7.1. *If \mathcal{M}_h has an exponentially small Kolmogorov N -width,*

$$d_N(\mathcal{M}_h; V_h) \leq ce^{-\delta N} \quad \text{with} \quad \delta > \log \delta_0, \quad (7.18)$$

then the reduced basis method built using the greedy algorithm (7.6) converges exponentially with respect to N , i.e. there exists $\eta > 0$ such that

$$\|u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_V \leq Ce^{-\eta N} \quad \forall \boldsymbol{\mu} \in \mathcal{P}.$$

The same result holds if we consider algorithm (7.10) with either $s_N(\boldsymbol{\mu}) = \|u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_V$ or $s_N(\boldsymbol{\mu}) = \Delta_N(\boldsymbol{\mu})$, provided that $\delta > \log \delta_0$, with $\delta_0 = 1 + (\frac{\tilde{\gamma}}{\alpha_0})^{1/2}$ in the first case and $\delta_0 = 1 + \frac{\tilde{\gamma}}{\alpha_0} (\frac{\tilde{\gamma}}{\alpha_0})^{1/2}$ in the second case.

Although interesting from a theoretical viewpoint, the comparison result (7.17) is only useful if $d_N(\mathcal{M}_h; V_h)$ decays to zero faster than $N^{-1}\delta_0^{-N}$, with $\delta_0 \geq 2$. This result has been further improved in [30], where it was shown that for the greedy algorithm (7.6) (or, equivalently, (7.16) with $\rho = 1$)

$$d(\mathcal{M}_h; V_N) \leq \frac{2}{\sqrt{3}} 2^N d_N(\mathcal{M}_h; V_h).$$

Moreover, if there exist $C, c > 0$ such that $d_N(\mathcal{M}_h; V_h) \leq Ce^{-cN^\beta}$, then, for all $\rho \in (0, 1]$ there exists $\tilde{C}, \tilde{c} > 0$ independent of N such that

$$\sup_{\boldsymbol{\mu} \in \mathcal{P}} \|u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_V \leq C\sqrt{\tilde{\gamma}/\alpha_0} d(\mathcal{M}_h; V_N) \leq \tilde{C}\sqrt{\tilde{\gamma}/\alpha_0} e^{-\tilde{c}N^{\beta/(\beta+1)}}. \quad (7.19)$$

In the case of algebraic convergence, that is if $d_N(\mathcal{M}_h; V_h) \leq MN^{-\beta}$ for some $M, \beta > 0$, there exists $C = C(\rho, \beta)$ independent of N such that

$$\sup_{\boldsymbol{\mu} \in \mathcal{P}} \|u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_V \leq C\sqrt{\tilde{\gamma}/\alpha_0} d(\mathcal{M}_h; V_N) \leq C\sqrt{\tilde{\gamma}/\alpha_0} MN^{-\beta}. \quad (7.20)$$

Hence, provided the Kolmogorov N -width of \mathcal{M}_h – an *intrinsic* property of the problem – decays exponentially with respect to N , the greedy algorithm is able to provide (about) the same decay in the error between $u_h(\boldsymbol{\mu})$ and $\Pi_{V_N} u_h(\boldsymbol{\mu})$ and henceforth between $u_h(\boldsymbol{\mu})$ and $u_N(\boldsymbol{\mu})$, thanks to the optimality of the Galerkin projection.

7.5 Numerical Assessment of a Priori Convergence Results

We have seen in the previous section that the convergence rate of RB approximations built by means of the greedy algorithm is linked to the Kolmogorov N -width of the solution manifold. We now introduce a model problem where the exact parameter-dependent solution can be expanded as a Neumann series, leading to a constructive

proof that the N -width of the solution set in this case converges exponentially, according to the result contained in Theorem 5.2. We will show that the rate of convergence of the greedy algorithm is indeed comparable to the one predicted by our N -width upper bound (5.43).

We consider a diffusion problem in a square $\Omega = (-1, 1)^2$ with four circular subregions $\Omega_1, \dots, \Omega_4$ of radius 0.2 as depicted in Fig. 7.3: given $\boldsymbol{\mu} \in \mathcal{P} = [-1 + \varepsilon, 1 - \varepsilon] \times [-1, 1]^2 \times [0, 10]^5$, find $u = u(\boldsymbol{\mu})$ such that:

$$\begin{aligned} -(1 + \mu_1 \chi_\omega) \Delta u &= \mu_4 \chi_{\Omega \setminus \omega} + \sum_{q=1}^4 \mu_{q+4} \chi_{\Omega_q} \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Gamma_1 \cup \Gamma_3, \quad \frac{\partial u}{\partial n} = \mu_2 \quad \text{on } \Gamma_4, \quad \frac{\partial u}{\partial n} = \mu_3 \quad \text{on } \Gamma_2, \end{aligned} \quad (7.21)$$

where the Γ_k 's denote the four sides of the square and $\omega = \bigcup_{q=1}^4 \Omega_q$ is the union of the disks. The first parameter μ_1 controls the difference between the isotropic diffusion coefficient inside the disks versus the background conductivity, while μ_2, \dots, μ_8 characterize boundary and source terms.

This problem admits an affine parametric expansion as in (5.33) with $Q_f = 7$ and

$$\begin{aligned} \theta_1^a(\boldsymbol{\mu}) &= 1, \quad \theta_2^a(\boldsymbol{\mu}) = \mu_1, \quad \theta_i^f(\boldsymbol{\mu}) = \mu_{i+1}, \quad i = 1, \dots, 7 \\ a_1(u, v) &= \int_{\Omega} \nabla u \cdot \nabla v \, d\Omega, \quad a_2(u, v) = \int_{\omega} \nabla u \cdot \nabla v \, d\Omega, \\ f_i(v) &= \int_{\Gamma_{i+1}} v \, d\Omega, \quad i = 1, 2, \quad f_3(v) = \int_{\Omega \setminus \omega} v \, d\Omega, \quad f_{3+j}(v) = \int_{\Omega_j} v \, d\Omega, \quad j = 1, \dots, 4. \end{aligned}$$

Moreover, its high-fidelity approximation features the algebraic form (5.35) and satisfies the global spectral condition (5.38) provided that $\mu_1 \in [-(1 - \varepsilon), 1 - \varepsilon]$ for some $\varepsilon > 0$, see Exercise 2. Therefore, problem (7.21) fits into the framework of Sect. 5.6 and its solution can be written as a combination of the fundamental basis vectors thanks to the formula (5.39).

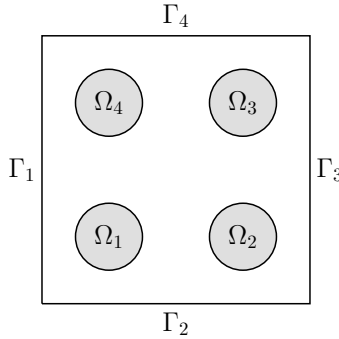


Fig. 7.3 Geometry, subdomains and boundaries for problem (7.21)

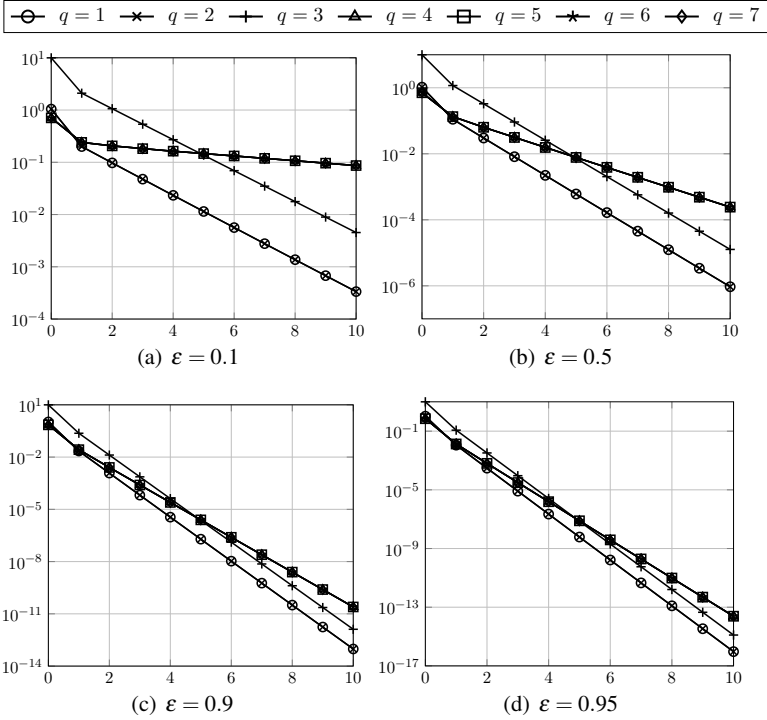


Fig. 7.4 Convergence of fundamental series (5.39) coefficients for different values of ε in (5.42)

We consider four different cases: $\varepsilon = 0.1$, $\varepsilon = 0.5$, $\varepsilon = 0.9$, and $\varepsilon = 0.95$. If $\varepsilon = 1$, the solution manifold is a Q_f -dimensional subspace of V as indicated by (5.40), hence the greedy algorithm terminates after exactly 7 iterations. In Fig. 7.4 we plot the behavior of the fundamental series terms

$$\max_{\mu \in \mathcal{P}} \left| \frac{(\theta_a^2(\mu))^k \theta_f^q(\mu)}{(\theta_a^1(\mu))^{k+1}} \right| \|\Psi_{k,q}\|_{\mathbb{X}_h}$$

with respect to k ; indeed, this yields the convergence rate of the N -width upper bound (5.43). For $\varepsilon = 0.1$, a very slow convergence of the fundamental series is observed for some of the terms. To obtain the G-RB approximation the weak greedy algorithm was driven by the a posteriori error estimator $\Delta_N(\mu)$ (with a stopping tolerance $\varepsilon_g = 10^{-3}$). This required $N = 21$ basis functions for the case $\varepsilon = 0.1$, $N = 19$ for $\varepsilon = 0.5$, $N = 14$ for $\varepsilon = 0.9$, and $N = 14$ for $\varepsilon = 0.95$.

In Fig. 7.5 we plot the corresponding convergence rates of the greedy algorithm compared to the N -width upper bound predictions given by (5.43). In each case an exponential convergence of the G-RB approximation is observed. The actual exponential decay rate depends on ε ; for $\varepsilon = 0.1$ the N -width estimate is far too pessimistic when compared to the true rate of convergence, due to the slow conver-

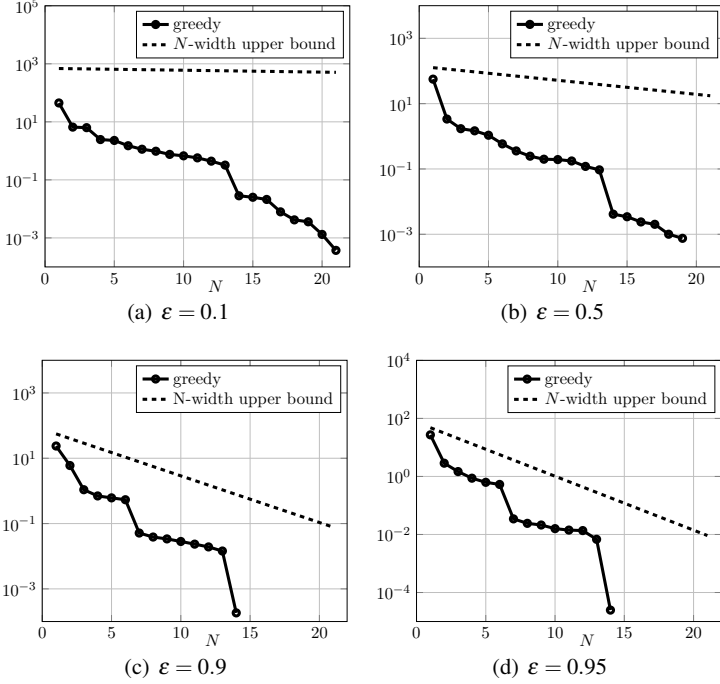


Fig. 7.5 Comparison of the N -width upper bound estimate (5.43) and the greedy algorithm convergence rate

gence of some terms of the fundamental series (see Fig. 7.4(a)). However, as $\varepsilon \rightarrow 1$ the N -width estimate (5.43) becomes more and more indicative of the convergence rate observed during the greedy algorithm. According to Fig. 7.4(c–d)) at the limit all the fundamental series coefficients converge at roughly the same rate, so that the upper bound (5.43) provides a tighter estimate of the N -width.

7.6 Exercises

1. Show that if $a(\cdot, \cdot; \boldsymbol{\mu})$ is coercive and $s_N(\boldsymbol{\mu}) = \|u_h(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})\|_V$, then $c_s = \alpha_0/\bar{\gamma}$ and $C_s = 1$ in the expression (7.8) of the bounds for the projection error. Similarly, show that $c_s = \alpha_0^2/\bar{\gamma}^2$ and $C_s = 1$ if $s_N(\boldsymbol{\mu}) = \Delta_N(\boldsymbol{\mu})$.
2. Prove that the high-fidelity approximation of problem (7.21) features the algebraic form (5.35) and satisfies the global spectral condition (5.38) provided that $\mu_1 \in [-(1 - \varepsilon), 1 - \varepsilon]$ for some $\varepsilon > 0$.

Chapter 8

RB Methods in Action: Setting up the Problem

We show how a parametrized PDE can be transformed into an equivalent problem on a reference (parameter-independent) domain. General mathematical tools for this operation are given, and examples of parametrized PDEs are discussed that are inspired by the four problems of Chap. 2. The primary purpose is to highlight the different role played by physical and geometric parameters. A further critical issue addressed concerns the possible affine parametric dependence of the linear and bilinear forms defined over the reference domain. With the aid of a host of meaningful examples we discuss the way the affine parametric dependence is affected by the problem parametrization.

8.1 Going from the Original to the Reference Domain

By revisiting the four problems introduced in the first chapter, we address some examples of parameters of interest, and analyze how they affect the variational formulation of the problem.

From now on, we will denote by $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)^T \in \mathcal{P}$ an *input-parameter* vector, by $\tilde{\Omega}(\boldsymbol{\mu}) \subset \mathbb{R}^d$ the ($\boldsymbol{\mu}$ -dependent) *original domain* and by $\Omega \subset \mathbb{R}^d$ the ($\boldsymbol{\mu}$ -independent) *reference domain*; see Sect. 3.1. $\tilde{\Omega}(\boldsymbol{\mu})$ is obtained from Ω through a parametric map

$$\boldsymbol{\Phi} : \Omega \times \mathcal{P} \rightarrow \mathbb{R}^d, \quad \mathbf{x} \mapsto \tilde{\mathbf{x}}(\boldsymbol{\mu}), \quad (8.1)$$

that is

$$\tilde{\Omega}(\boldsymbol{\mu}) = \boldsymbol{\Phi}(\Omega; \boldsymbol{\mu}) \quad \forall \boldsymbol{\mu} \in \mathcal{P}. \quad (8.2)$$

The map $\boldsymbol{\Phi}$ will depend in fact only on the geometric parameters $\boldsymbol{\mu}_g$. In those cases in which the original problem is formulated in a fixed ($\boldsymbol{\mu}$ -independent) domain, that is, $\boldsymbol{\mu}$ only consists of physical parameters $\boldsymbol{\mu}_{ph}$, then $\tilde{\Omega} = \Omega$ and $\boldsymbol{\Phi}$ is the identity map.

For the sake of notation, our original problem set in the original domain $\tilde{\Omega}(\boldsymbol{\mu}_g)$ is denoted in compact form as

$$\tilde{P} = \tilde{P}(\tilde{\Omega}(\boldsymbol{\mu}_g); \boldsymbol{\mu}_{ph}).$$

We need to transform this problem into an equivalent one,

$$P = P(\Omega; \boldsymbol{\mu}_{ph}, \boldsymbol{\mu}_g)$$

set in the reference domain Ω . More specifically, in differential form problem \tilde{P} reads

$$\tilde{L}(\boldsymbol{\mu}_{ph})\tilde{u}(\boldsymbol{\mu}) = \tilde{f}(\boldsymbol{\mu}_{ph}) \quad \text{in } \tilde{V}'(\boldsymbol{\mu}_g).$$

Its weak formulation¹ reads: find $\tilde{u} = \tilde{u}(\boldsymbol{\mu}) \in \tilde{V}(\boldsymbol{\mu}_g)$ such that

$$\tilde{a}(u(\boldsymbol{\mu}), v; \boldsymbol{\mu}) = \tilde{f}(v; \boldsymbol{\mu}) \quad \forall v \in \tilde{V}(\boldsymbol{\mu}_g). \quad (8.3)$$

By means of the inverse of the transformation (8.2), we pull back problem (8.3) onto the reference domain Ω , yielding: find $u = u(\boldsymbol{\mu}) \in V$ such that

$$a(u(\boldsymbol{\mu}), v; \boldsymbol{\mu}) = f(v; \boldsymbol{\mu}) \quad \forall v \in V. \quad (8.4)$$

This is precisely the problem (3.2) which stands at the basis of the construction of RB approximations illustrated in Chap. 3.

The bilinear form $a(\cdot, \cdot; \boldsymbol{\mu})$ and the linear form $f(\cdot; \boldsymbol{\mu})$ differ from $\tilde{a}(\cdot, \cdot; \boldsymbol{\mu})$ and $\tilde{f}(\cdot; \boldsymbol{\mu})$ because of the presence of geometric factors induced by the transformation $\Phi(\cdot; \boldsymbol{\mu}_g)$. Section 8.2 collects a few essential tools that are useful to obtain a and f . Several illustrative examples will be provided in Sects. 8.3–8.8. To avoid cumbersome notation, we will no longer distinguish between $\boldsymbol{\mu}_{ph}$ and $\boldsymbol{\mu}_g$, denoting by $\boldsymbol{\mu}$ the vector of all parameters.

8.2 Change of Variables Formulas

Let us first recall some change of variables formulas for multiple integrals which are useful for the derivation of the weak formulation of a parametrized PDE. We denote with x_l and \tilde{x}_k (for $k, l = 1, \dots, d$) the coordinates on the reference domain Ω

¹ With the aim of better highlighting the different role played by the parameters, with a little abuse of notation we can write

$$\tilde{a}(u(\boldsymbol{\mu}), v; \boldsymbol{\mu}) = \int_{\tilde{\Omega}(\boldsymbol{\mu}_g)} \tilde{L}(\boldsymbol{\mu}_{ph})\tilde{u}(\boldsymbol{\mu})v \, d\tilde{\Omega}(\boldsymbol{\mu}_g), \quad \tilde{f}(v; \boldsymbol{\mu}_{ph}) = \int_{\tilde{\Omega}(\boldsymbol{\mu}_g)} \tilde{f}(\boldsymbol{\mu}_{ph})v \, d\tilde{\Omega}(\boldsymbol{\mu}_g).$$

However, we warn the reader that the above identification needs to be understood up to an integration by parts. For a rigorous derivation we refer to the following sections, where several examples will be addressed.

and those on the original domain $\tilde{\Omega}(\boldsymbol{\mu})$, respectively. Let us consider the Jacobian matrix $\mathbb{J}_{\boldsymbol{\Phi}}(\mathbf{x}; \boldsymbol{\mu}) \in \mathbb{R}^{d \times d}$ of the map $\boldsymbol{\Phi}(\cdot; \boldsymbol{\mu})$ introduced in (8.1),

$$(\mathbb{J}_{\boldsymbol{\Phi}}(\mathbf{x}; \boldsymbol{\mu}))_{kl} = \frac{\partial \tilde{x}_k}{\partial x_l}(\mathbf{x}) = \frac{\partial \Phi_k(\mathbf{x}; \boldsymbol{\mu})}{\partial x_l}(\mathbf{x}), \quad k, l = 1, \dots, d.$$

If we assume that, for any $\boldsymbol{\mu} \in \mathcal{P}$, its determinant $|\mathbb{J}_{\boldsymbol{\Phi}}(\mathbf{x}; \boldsymbol{\mu})|$ is different than 0 for any $\mathbf{x} \in \Omega$, then the map $\boldsymbol{\Phi}$ is well-defined.

For any integrable function $\tilde{\psi} : \tilde{\Omega} \rightarrow \mathbb{R}$ the following formula

$$\int_{\tilde{\Omega}(\boldsymbol{\mu})} \tilde{\psi}(\tilde{\mathbf{x}}) d\tilde{\Omega} = \int_{\Omega} \psi(\mathbf{x}) |\mathbb{J}_{\boldsymbol{\Phi}}(\mathbf{x}; \boldsymbol{\mu})| d\Omega \quad (8.5)$$

provides the change of variable, where $\psi = \tilde{\psi} \circ \boldsymbol{\Phi}$. The Jacobian matrix and its determinant depend *a priori* on both the spatial coordinates \mathbf{x} and the parameter vector $\boldsymbol{\mu}$. In case $\boldsymbol{\Phi}$ is an *affine transformation*, that is

$$\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\mu}) = \mathbb{B}(\boldsymbol{\mu})\mathbf{x} + \mathbf{c}(\boldsymbol{\mu}), \quad \text{with } \mathbb{B}(\boldsymbol{\mu}) \in \mathbb{R}^{d \times d}, \mathbf{c}(\boldsymbol{\mu}) \in \mathbb{R}^d,$$

then $\mathbb{J}(\mathbf{x}; \boldsymbol{\mu}) = \mathbb{B}(\boldsymbol{\mu})$, and both the Jacobian matrix and its determinant depend just on $\boldsymbol{\mu} \in \mathcal{P}$. From now on, we omit the dependence on the spatial coordinates of the functions appearing in the integrals whenever it is clear from the context.

In case of integrals involving derivatives, we need to introduce some extra transformations. Let us denote by $\boldsymbol{\Phi}^{-1}(\cdot; \boldsymbol{\mu})$ the inverse of $\boldsymbol{\Phi}(\cdot; \boldsymbol{\mu})$, such that $\Omega = \boldsymbol{\Phi}^{-1}(\tilde{\Omega}(\boldsymbol{\mu}); \boldsymbol{\mu})$, and by

$$(\mathbb{J}_{\boldsymbol{\Phi}^{-1}}(\tilde{\mathbf{x}}; \boldsymbol{\mu}))_{kl} = \frac{\partial x_k}{\partial \tilde{x}_l}(\tilde{\mathbf{x}}) = \frac{\partial \Phi_k^{-1}(\tilde{\mathbf{x}}; \boldsymbol{\mu})}{\partial \tilde{x}_l}(\tilde{\mathbf{x}}), \quad k, l = 1, \dots, d,$$

its Jacobian matrix. Then²

$$\mathbb{J}_{\boldsymbol{\Phi}^{-1}}(\tilde{\mathbf{x}}; \boldsymbol{\mu}) = (\mathbb{J}_{\boldsymbol{\Phi}}(\mathbf{x}; \boldsymbol{\mu}))^{-1}, \quad \tilde{\mathbf{x}} = \boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\mu})$$

so that

$$|\mathbb{J}_{\boldsymbol{\Phi}^{-1}}(\tilde{\mathbf{x}}; \boldsymbol{\mu})| = \frac{1}{|\mathbb{J}_{\boldsymbol{\Phi}}(\mathbf{x}; \boldsymbol{\mu})|}.$$

Thanks to the chain rule,

$$\frac{\partial \tilde{\psi}(\tilde{\mathbf{x}})}{\partial \tilde{x}_i} = \sum_{j=1}^d \frac{\partial \psi(\mathbf{x})}{\partial x_j} \frac{\partial x_j}{\partial \tilde{x}_i}, \quad i = 1, \dots, d,$$

² According to the inverse function theorem (see, e.g., [234, Chap. 9]), if the Jacobian of the map $\boldsymbol{\Phi}$ is continuous and nonsingular at the point \mathbf{x} , then $\boldsymbol{\Phi}$ is invertible when restricted to some neighborhood of \mathbf{x} and $\mathbf{J}_{\boldsymbol{\Phi}^{-1}} \circ \boldsymbol{\Phi} = (\mathbf{J}_{\boldsymbol{\Phi}})^{-1}$.

On the other hand, if the determinant of the Jacobian matrix of $\boldsymbol{\Phi}$ is not zero at a point \mathbf{x} , there is neighbourhood of \mathbf{x} in which $\boldsymbol{\Phi}$ is invertible.

we obtain the compact expression

$$\nabla_{\tilde{\mathbf{x}}} \tilde{\psi}(\tilde{\mathbf{x}}) = [\mathbb{J}_{\Phi^{-1}}(\tilde{\mathbf{x}})]^T \nabla_{\mathbf{x}} \psi(\mathbf{x}) = (\mathbb{J}_{\Phi}(\mathbf{x}; \boldsymbol{\mu}))^{-T} \nabla_{\mathbf{x}} \psi(\mathbf{x}), \quad (8.6)$$

where we have denoted by $\nabla_{\tilde{\mathbf{x}}}$ (resp. $\nabla_{\mathbf{x}}$) the gradient with respect to the coordinates of the original (resp. reference) domain.

Thanks to formula (8.6), we obtain the following relations for the change of variables in integrals involving derivatives, valid for any $\tilde{\psi}, \tilde{\chi} \in H^1(\tilde{\Omega})$:

$$\begin{aligned} \int_{\tilde{\Omega}(\boldsymbol{\mu})} \nabla_{\tilde{\mathbf{x}}} \tilde{\psi} \cdot \nabla_{\tilde{\mathbf{x}}} \tilde{\chi} d\tilde{\Omega} &= \int_{\Omega} (\mathbb{J}_{\Phi}^{-T}(\mathbf{x}; \boldsymbol{\mu}) \nabla_{\mathbf{x}} \psi) \cdot (\mathbb{J}_{\Phi}^{-T}(\tilde{\mathbf{x}}; \boldsymbol{\mu}) \nabla_{\mathbf{x}} \chi) |\mathbb{J}_{\Phi}(\mathbf{x}; \boldsymbol{\mu})| d\Omega \\ \int_{\tilde{\Omega}(\boldsymbol{\mu})} \tilde{\mathbf{b}} \cdot \nabla_{\tilde{\mathbf{x}}} \tilde{\psi} d\tilde{\Omega} &= \int_{\Omega} \chi \mathbf{b} \cdot (\mathbb{J}_{\Phi}^{-T}(\mathbf{x}; \boldsymbol{\mu}) \nabla_{\mathbf{x}} \psi) |\mathbb{J}_{\Phi}(\mathbf{x}; \boldsymbol{\mu})| d\Omega, \end{aligned}$$

where $\psi = \tilde{\psi} \circ \Phi$, $\chi = \tilde{\chi} \circ \Phi$, and $\mathbf{b} = \tilde{\mathbf{b}} \circ \Phi$. See Exercise 1 for the related proofs.

Remark 8.1. We recall that, given a generic mapping $\Phi : \Omega \rightarrow \tilde{\Omega}$ and a vector field $\mathbf{b} : \Omega \rightarrow \mathbb{R}^d$, the transformation yielding

$$\tilde{\mathbf{b}}(\tilde{\mathbf{x}}) = \mathbf{b} \circ \Phi^{-1}(\mathbf{x})$$

does not preserve the divergence of \mathbf{b} . In particular, if $\text{div}_{\mathbf{x}} \mathbf{b} = 0$ over Ω , we do not necessarily have that $\text{div}_{\tilde{\mathbf{x}}} \tilde{\mathbf{b}} = 0$ over $\tilde{\Omega}$. Depending on the problem at hand, a different transformation for vector fields, called *Piola transformation*, will often be used. See Sect. 8.5.1. •

In a more compact form, we can write

$$\int_{\tilde{\Omega}(\boldsymbol{\mu})} \nabla_{\tilde{\mathbf{x}}} \tilde{\psi} \cdot \nabla_{\tilde{\mathbf{x}}} \tilde{\chi} d\tilde{\Omega} = \sum_{k,l=1}^d \int_{\Omega} \frac{\partial \psi}{\partial x_k} v_{kl} \frac{\partial \chi}{\partial x_l} d\Omega \quad (8.7)$$

where for any $\boldsymbol{\mu} \in \mathcal{P}$, $\mathbf{v} : \mathbb{R}^d \times \mathcal{P} \rightarrow \mathbb{R}^{d \times d}$ is given by

$$\mathbf{v}(\mathbf{x}; \boldsymbol{\mu}) = \mathbb{J}_{\Phi}^{-1}(\mathbf{x}; \boldsymbol{\mu}) \mathbb{J}_{\Phi}^{-T}(\mathbf{x}; \boldsymbol{\mu}) |\mathbb{J}_{\Phi}(\mathbf{x}; \boldsymbol{\mu})|. \quad (8.8)$$

In the same way,

$$\int_{\tilde{\Omega}(\boldsymbol{\mu})} \tilde{\mathbf{b}} \cdot \nabla_{\tilde{\mathbf{x}}} \tilde{\psi} d\tilde{\Omega} = \sum_{k,l=1}^d \int_{\Omega} b_k \eta_{kl} \frac{\partial \psi}{\partial x_l} d\Omega \quad (8.9)$$

where $\boldsymbol{\eta} : \mathbb{R}^d \times \mathcal{P} \rightarrow \mathbb{R}^{d \times d}$ is given by

$$\boldsymbol{\eta}(\mathbf{x}; \boldsymbol{\mu}) = \mathbb{J}_{\Phi}^{-T}(\mathbf{x}; \boldsymbol{\mu}) |\mathbb{J}_{\Phi}(\mathbf{x}; \boldsymbol{\mu})|. \quad (8.10)$$

The parametrized tensors $\mathbf{v}(\mathbf{x}; \boldsymbol{\mu})$, $\boldsymbol{\eta}(\mathbf{x}; \boldsymbol{\mu})$ encode all the information concerning the parameters, and allow to derive the weak formulation of a parametrized PDE, which stands at the basis of the implementation of a RB method.

8.2.1 Extension to the Vector Case

The vector counterpart of formula (8.5) for the change of variables under the sign of integral is given by

$$\int_{\tilde{\Omega}(\boldsymbol{\mu})} \tilde{\boldsymbol{\psi}} d\tilde{\Omega} = \int_{\Omega} \boldsymbol{\psi} |\mathbb{J}_{\boldsymbol{\Phi}}| d\Omega \quad (8.11)$$

for any integrable function $\tilde{\boldsymbol{\psi}} : \Omega \rightarrow \mathbb{R}^d$, where $\boldsymbol{\psi} = \tilde{\boldsymbol{\psi}} \circ \boldsymbol{\Phi}$ and $|\mathbb{J}_{\boldsymbol{\Phi}}(\mathbf{x}; \boldsymbol{\mu})|$ denotes the determinant of the Jacobian matrix, defined as in the scalar case.

In case of integrals involving derivatives, formula (8.7) becomes

$$\int_{\tilde{\Omega}(\boldsymbol{\mu})} \nabla_{\tilde{\mathbf{x}}} \tilde{\boldsymbol{\psi}} : \nabla_{\tilde{\mathbf{x}}} \tilde{\boldsymbol{\chi}} d\tilde{\Omega} = \sum_{i,k,l=1}^d \int_{\Omega} \frac{\partial \psi_i}{\partial x_k} \mathbf{v}_{kl} \frac{\partial \chi_i}{\partial x_l} d\Omega \quad (8.12)$$

where $\mathbf{v} : \mathbb{R}^d \times \mathcal{P} \rightarrow \mathbb{R}^{d \times d}$ is the parametrized tensor defined in (8.8). Similarly, we obtain the following formulas involving the derivatives of only one function:

$$\int_{\tilde{\Omega}(\boldsymbol{\mu})} (\tilde{\mathbf{b}} \cdot \nabla_{\tilde{\mathbf{x}}}) \tilde{\boldsymbol{\psi}} \cdot \tilde{\boldsymbol{\chi}} d\tilde{\Omega} = \sum_{i,k,l=1}^d \int_{\Omega} b_k \eta_{kl} \frac{\partial \psi_i}{\partial x_l} \chi_i d\Omega, \quad (8.13)$$

$$\int_{\tilde{\Omega}(\boldsymbol{\mu})} \tilde{\boldsymbol{\psi}} \operatorname{div} \tilde{\boldsymbol{\chi}} d\tilde{\Omega} = \sum_{k,l=1}^d \int_{\Omega} \psi \eta_{kl} \frac{\partial \chi_k}{\partial x_l} d\Omega, \quad (8.14)$$

where $\boldsymbol{\eta} : \mathbb{R}^d \times \mathcal{P} \rightarrow \mathbb{R}^{d \times d}$ is the parametrized tensor defined in (8.10).

8.3 Advection-Diffusion-Reaction, Case I: Heat Transfer

We consider a heat conduction/convection problem occurring, e.g., in electronic devices made by different materials characterized by different thermal conductivities. We suppose that the advection field has a varying intensity, and acts on a portion of the domain whose width is also changing.

We denote by $\tilde{\Omega}(\mu_1) = (0, 1 + \mu_1) \times (0, 1)$ a domain consisting of four subdomains (see Fig. 8.1),

$$\begin{aligned} \tilde{\Omega}_1(\mu_1) &= \left(\frac{2}{3}, 1 + \mu_1\right) \times (0, 1), & \tilde{\Omega}_2 &= \left(\frac{1}{3}, \frac{2}{3}\right) \times \left(0, \frac{1}{6}\right), \\ \tilde{\Omega}_3 &= \left(0, \frac{2}{3}\right) \times (0.5, 0.7), & \tilde{\Omega}_4 &= \left(0, \frac{2}{3}\right) \times (0, 1) \setminus \left(\tilde{\Omega}_2 \cup \tilde{\Omega}_3\right). \end{aligned}$$

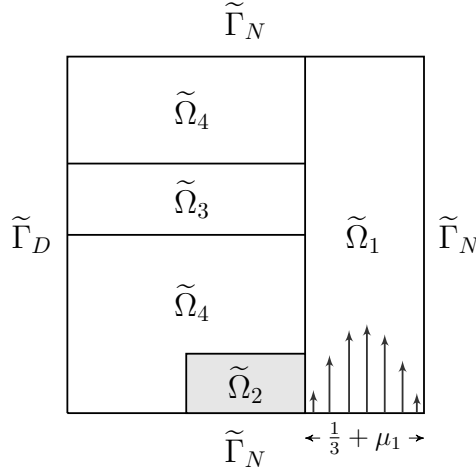


Fig. 8.1 Geometry, subdomains and boundaries for problem (8.15)

The governing parametrized PDE problem for the temperature $\tilde{u} = \tilde{u}(\boldsymbol{\mu})$ reads

$$\begin{cases} -\operatorname{div}(\tilde{k}(\tilde{\mathbf{x}}; \boldsymbol{\mu}) \nabla \tilde{u}) + \tilde{\mathbf{b}}(\tilde{\mathbf{x}}; \boldsymbol{\mu}) \cdot \nabla \tilde{u} = \tilde{s}(\tilde{\mathbf{x}}) & \text{in } \tilde{\Omega}(\mu_1) \\ \tilde{u} = 0 & \text{on } \tilde{\Gamma}_D \\ \tilde{k}(\tilde{\mathbf{x}}; \boldsymbol{\mu}) \nabla \tilde{u} \cdot \tilde{\mathbf{n}} = 0 & \text{on } \tilde{\Gamma}_N \end{cases} \quad (8.15)$$

where $\tilde{\Gamma}_D = \{0\} \times [0, 1]$ and $\tilde{\Gamma}_N = \partial \tilde{\Omega} \setminus \tilde{\Gamma}_D$, while $\tilde{\mathbf{b}}$ is a prescribed advection field. Across all internal interfaces the solution \tilde{u} (the temperature) as well as its flux $\tilde{k}(\tilde{\mathbf{x}}; \boldsymbol{\mu}) \nabla \tilde{u} \cdot \tilde{\mathbf{n}}$ are continuous (in the sense of traces, see A.6.2).

Problem (8.15) can, e.g., model a cooling device for (an array of) electronic components; here the component – which is a poor conductor – is assumed to occupy the subdomain $\tilde{\Omega}_4$, a conducting material occupies the domain $\tilde{\Omega}_3$, whereas $\tilde{\Omega}_1$ is occupied by a fluid whose velocity is described by a fully-developed parabolic field; see [171] for a description of such problem.

Since these subregions are characterized by different thermal conductivities, $\tilde{k}(\mathbf{x}; \boldsymbol{\mu})$ can be expressed by

$$\tilde{k}(\mathbf{x}; \mu_3) = \chi_{\tilde{\Omega}_1 \cup \tilde{\Omega}_4}(\tilde{\mathbf{x}}) + 100 \chi_{\tilde{\Omega}_2}(\tilde{\mathbf{x}}) + \mu_3 \chi_{\tilde{\Omega}_3}(\tilde{\mathbf{x}}).$$

Here χ_A denotes the characteristic function of the subdomain $A \subset \Omega$, whereas μ_3 is the conductivity that varies over $\tilde{\Omega}_3$. The advection field in $\tilde{\Omega}_1$ is given by

$$\tilde{\mathbf{b}}(\mu_2; \tilde{\mathbf{x}}) = [0, \chi_{\tilde{\Omega}_1} 162 \mu_2 / (1 + 3 \mu_1)^3 (\tilde{x}_1 - 2/3)(1 + \mu_1 - \tilde{x}_1)]^T.$$

Ω_2 is the heated element, and we represent the source term as

$$\tilde{s}(\tilde{\mathbf{x}}) = 10 \chi_{\tilde{\Omega}_2}(\tilde{\mathbf{x}}).$$

Indeed, a heat flow occurs from $\tilde{\Omega}_2$ to $\tilde{\Omega}_1$, then from $\tilde{\Omega}_1$ to the higher conductivity material. The maximum amplitude μ_2 of the advection field and the thermal conductivity μ_3 over $\tilde{\Omega}_3$ play the role of input physical parameters, whereas the width μ_1 of the subdomain $\tilde{\Omega}_1$ is a geometric input parameter.

The weak formulation of problem (8.15) over the original domain $\tilde{\Omega}(\mu_1)$ reads: find $\tilde{u}(\boldsymbol{\mu}) \in \tilde{V}(\boldsymbol{\mu}) = H_{\tilde{\Gamma}_D}^1(\tilde{\Omega}(\mu_1))$ such that

$$\tilde{a}(\tilde{u}(\boldsymbol{\mu}), v; \boldsymbol{\mu}) = \tilde{f}(v; \boldsymbol{\mu}) \quad \forall v \in \tilde{V}(\boldsymbol{\mu}), \quad (8.16)$$

where

$$\begin{aligned} \tilde{a}(u, v; \boldsymbol{\mu}) = & \int_{\tilde{\Omega}_1(\mu_1)} \nabla u \cdot \nabla v d\tilde{\Omega} + \frac{162\mu_2}{(1+3\mu_1)^3} \int_{\tilde{\Omega}_1(\mu_1)} (x_1 - \frac{2}{3})(1 + \mu_1 - x_1) \frac{\partial u}{\partial x_2} v d\tilde{\Omega} \\ & + \int_{\tilde{\Omega}_2} 100 \nabla u \cdot \nabla v d\tilde{\Omega} + \mu_3 \int_{\tilde{\Omega}_3} \nabla u \cdot \nabla v d\tilde{\Omega} + \int_{\tilde{\Omega}_4} \nabla u \cdot \nabla v d\tilde{\Omega}, \end{aligned} \quad (8.17)$$

$$\tilde{f}(v; \boldsymbol{\mu}) = \int_{\tilde{\Omega}_2} 10v d\tilde{\Omega}. \quad (8.18)$$

Note that the solution space $\tilde{V}(\boldsymbol{\mu})$ as well as the bilinear form $\tilde{a}(\cdot, \cdot; \boldsymbol{\mu})$ and the linear form $\tilde{f}(\cdot; \boldsymbol{\mu})$ are all parameter dependent.

8.3.1 Reference Configuration and Affine Transformations

As we have seen in Chaps. 3, 6 and 7 the RB solution of a given parametrized PDE is a linear combination of (suitably selected) high-fidelity solutions, corresponding to given parameter values (the so-called *snapshots*).

For the case at hand, we introduce the following reference domain

$$\Omega = \tilde{\Omega}(\mu_1^{ref}) = (0, 1) \times (0, 1),$$

corresponding to the choice $\mu_1^{ref} = 0$; in particular

$$\Omega_1 = \tilde{\Omega}_1(\mu_1^{ref}) = (\frac{2}{3}, 1) \times (0, 1).$$

According with (8.1)-(8.2), the original domain $\tilde{\Omega}(\mu_1)$ can now be obtained as the image of the reference domain Ω through the parametric map

$$\tilde{\mathbf{x}} = \boldsymbol{\Phi}(\mathbf{x}, \mu_1) = \begin{cases} \boldsymbol{\Phi}_{\Omega_1}(\mathbf{x}, \mu_1), & \mathbf{x} \in \Omega_1 \\ \mathbf{x}, & \mathbf{x} \in \Omega_2 \cup \Omega_3 \cup \Omega_4 \end{cases} \quad (8.19)$$

where

$$\Phi_{\Omega_1}(\mathbf{x}, \mu_1) = \begin{bmatrix} 1+3\mu_1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} -2\mu_1 \\ 0 \end{bmatrix}. \quad (8.20)$$

In this (quite simple, indeed) situation, Φ can be expressed through an affine transformation over each subdomain $\Omega_1, \dots, \Omega_{n_{dom}}$ (with $n_{dom} = 4$),

$$\tilde{\mathbf{x}}(\boldsymbol{\mu}) = \mathbb{B}_i(\boldsymbol{\mu})\mathbf{x} + \mathbf{c}_i(\boldsymbol{\mu}), \quad i = 1, \dots, n_{dom}$$

where $\mathbb{B}_i(\boldsymbol{\mu}) \in \mathbb{R}^{d \times d}$ and $\mathbf{c}_i(\boldsymbol{\mu}) \in \mathbb{R}^d$ represent, for any $\boldsymbol{\mu} \in \mathcal{P}$, $i = 1, \dots, n_{dom}$, a matrix and a vector. \mathbb{B}_i encodes geometric features such as rotations and scalings, while \mathbf{c}_i is responsible for translations. Then

$$\mathbb{J}_{\Phi_{\Omega_i}}(\tilde{\mathbf{x}}; \boldsymbol{\mu}) = \mathbb{B}_i(\boldsymbol{\mu}), \quad i = 1, \dots, n_{dom}.$$

In this specific example, $\tilde{\mathbf{x}}$, \mathbb{B}_i and \mathbf{c}_i actually depend only on μ_1 and

$$\mathbb{B}_1(\boldsymbol{\mu}) = \begin{bmatrix} 1+3\mu_1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{c}_1(\boldsymbol{\mu}) = \begin{bmatrix} -2\mu_1 \\ 0 \end{bmatrix}.$$

Moreover, the transformation is the identity map over Ω_i , $i = 2, 3, 4$, that is

$$\mathbb{B}_i(\boldsymbol{\mu}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{c}_i(\boldsymbol{\mu}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad i = 2, 3, 4$$

since the subdomains Ω_i , $i = 2, 3, 4$ are not affected by any parameter dependence. As a matter of fact,

$$|\mathbb{J}_{\Phi_{\Omega_1}}(\tilde{\mathbf{x}}, \mu_1)| = 1 + 3\mu_1$$

whereas

$$|\mathbb{J}_{\Phi_{\Omega_i}}(\tilde{\mathbf{x}}, \mu_1)| = 1, \quad i = 2, 3, 4.$$

Last, but not least, we require that $\Phi_{\Omega_i}(\mathbf{x}, \mu_1) = \mathbf{x}$ for any $\mathbf{x} \in \Omega_1 \cap \Omega_i$, $i = 2, 3, 4$, in order to obtain a map which is globally continuous over Ω .

8.3.2 Weak Formulation on the Reference Domain

To pull back the weak formulation (8.16) onto the reference domain Ω , we need to operate a change of variables in the integrals appearing in the forms (8.17)–(8.18) and then apply the formulas (8.5), (8.7) and (8.9). For the case at hand, since Ω_1 is the only parameter-dependent subdomain, we compute the inverse map

$$\Phi_{\Omega_1}^{-1}(\mathbf{x}, \mu_1) = \begin{bmatrix} \frac{1}{1+3\mu_1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} \frac{-2\mu_1}{1+3\mu_1} \\ 0 \end{bmatrix}$$

and we note that $\Phi_{\Omega_i}^{-1}(\mathbf{x}) = \mathbb{I}$ for $i = 2, 3, 4$; consequently,

$$(\mathbb{J}_{\Phi_{\Omega_1}}(\mathbf{x}; \boldsymbol{\mu}))^{-1} = \begin{bmatrix} \frac{1}{1+3\mu_1} & 0 \\ 0 & 1 \end{bmatrix}.$$

Then, we evaluate the expression of the conductivity coefficient

$$k(\mathbf{x}; \mu_3) = \chi_{\Omega_1 \cup \Omega_4}(\mathbf{x}) + 100\chi_{\Omega_2}(\mathbf{x}) + \mu_3\chi_{\Omega_3}(\mathbf{x}),$$

the advection field $\mathbf{b} = \tilde{\mathbf{b}} \circ \Phi(\mathbf{x}; \boldsymbol{\mu})$ over the reference domain

$$\mathbf{b}(\mu_2; \mathbf{x}) = \left[0, \chi_{\Omega_1} \frac{162\mu_2}{1+3\mu_1} (1-x_1)(x_1 - \frac{2}{3}) \right]^T$$

and the source term

$$s(\mathbf{x}) = 10\chi_{\Omega_2}(\mathbf{x}).$$

The weak formulation of problem (8.15) on the reference domain reads: find $u(\boldsymbol{\mu}) \in V = H_{TD}^1(\Omega)$ such that

$$a(u(\boldsymbol{\mu}), v; \boldsymbol{\mu}) = f(v; \boldsymbol{\mu}) \quad \forall v \in V, \quad (8.21)$$

where (see Exercise 2 for further details)

$$\begin{aligned} a(u, v; \boldsymbol{\mu}) &= \frac{1}{1+3\mu_1} \int_{\Omega_1} \frac{\partial u}{\partial x_1} \frac{\partial v}{\partial x_1} d\Omega + (1+3\mu_1) \int_{\Omega_1} \frac{\partial u}{\partial x_2} \frac{\partial v}{\partial x_2} d\Omega \\ &+ 162\mu_2 \int_{\Omega_1} (1-x_1)(x_1 - \frac{2}{3}) \frac{\partial u}{\partial x_2} v d\Omega \end{aligned} \quad (8.22)$$

$$+ 100 \int_{\Omega_2} \nabla u \cdot \nabla v d\Omega + \mu_3 \int_{\Omega_3} \nabla u \cdot \nabla v d\Omega + \int_{\Omega_4} \nabla u \cdot \nabla v d\Omega,$$

$$f(v; \boldsymbol{\mu}) = 10 \int_{\Omega_2} v d\Omega. \quad (8.23)$$

As we have seen in Chap. 3, the setup, implementation and analysis of RB methods are carried out starting from the weak formulation (8.21). Because in the case at hand the parametrized tensors (8.8)–(8.10) depend only on the parameter $\boldsymbol{\mu}$ (and not on the spatial variables $\mathbf{x} \in \Omega$), the bilinear form $a(u, v; \boldsymbol{\mu})$ admits an affine expansion. More precisely,

$$a(u, v; \boldsymbol{\mu}) = \sum_{q=1}^{Q_a} \theta_a^q(\boldsymbol{\mu}) a_q(u, v) \quad (8.24)$$

where $Q_a = 6$ and

$$\begin{aligned}
 \theta_a^1(\boldsymbol{\mu}) &= \frac{1}{1+3\mu_1}, & a_1(u, v) &= \int_{\Omega_1} \frac{\partial u}{\partial x_1} \frac{\partial v}{\partial x_1} d\Omega \\
 \theta_a^2(\boldsymbol{\mu}) &= 1+3\mu_1, & a_2(u, v) &= \int_{\Omega_1} \frac{\partial u}{\partial x_2} \frac{\partial v}{\partial x_2} d\Omega \\
 \theta_a^3(\boldsymbol{\mu}) &= 162\mu_2, & a_3(u, v) &= \int_{\Omega_1} (1-x_1)(x_1-\frac{2}{3}) \frac{\partial u}{\partial x_2} v d\Omega \\
 \theta_a^4(\boldsymbol{\mu}) &= 100, & a_4(u, v) &= \int_{\Omega_2} \nabla u \cdot \nabla v d\Omega \\
 \theta_a^5(\boldsymbol{\mu}) &= \mu_3, & a_5(u, v) &= \int_{\Omega_3} \nabla u \cdot \nabla v d\Omega \\
 \theta_a^6(\boldsymbol{\mu}) &= 1, & a_6(u, v) &= \int_{\Omega_4} \nabla u \cdot \nabla v d\Omega.
 \end{aligned}$$

In the same way, the linear form $f(v; \boldsymbol{\mu})$ can be expressed as

$$f(v; \boldsymbol{\mu}) = \sum_{q=1}^{Q_f} \theta_f^q(\boldsymbol{\mu}) f_q(v) \quad (8.25)$$

where $Q_f = 1$ and

$$\theta_f^1(\boldsymbol{\mu}) = 10, \quad f_1(v) = \int_{\Omega_2} v d\Omega.$$

We have thus succeeded to write bilinear and linear forms as linear combinations of $\boldsymbol{\mu}$ -independent bilinear or linear forms, respectively, whose coefficients are suitable $\boldsymbol{\mu}$ -dependent scalar, real functions. Such a decomposition, enabled whenever $\boldsymbol{\mu}$ and \mathbf{x} are *separable* variables, is at the basis of the offline/online decoupling strategy, which is another distinguishing feature of RB methods for parametrized PDEs.

8.3.3 Dealing with Nonhomogeneous Boundary Conditions

Thus far we have only considered homogeneous boundary conditions, either of Dirichlet or Neumann kind. In the case of nonhomogeneous (and possibly parametrized) boundary conditions, further terms need to be included in the parametrized linear form $f(\cdot; \boldsymbol{\mu})$. Let us consider the following problem (a generalization of (8.15)):

$$\left\{ \begin{array}{ll} -\operatorname{div}(\tilde{k}(\tilde{\mathbf{x}}; \boldsymbol{\mu}) \nabla \tilde{u}) + \tilde{\mathbf{b}}(\tilde{\mathbf{x}}; \boldsymbol{\mu}) \cdot \nabla \tilde{u} = \tilde{s}(\tilde{\mathbf{x}}) & \text{in } \tilde{\Omega}(\mu_1) \\ \tilde{u} = \tilde{g}(\boldsymbol{\mu}) & \text{on } \tilde{\Gamma}_D \\ \tilde{k}(\tilde{\mathbf{x}}; \boldsymbol{\mu}) \nabla \tilde{u} \cdot \tilde{\mathbf{n}} = \tilde{h}(\boldsymbol{\mu}) & \text{on } \tilde{\Gamma}_N \end{array} \right. \quad (8.26)$$

where $\tilde{g}(\boldsymbol{\mu}) \in L^2(\tilde{\Gamma}_D)$ and $\tilde{h}(\boldsymbol{\mu}) \in L^2(\tilde{\Gamma}_N)$ are given parametrized Dirichlet and Neumann data.

Nonhomogenous Neumann conditions are simple to treat. By integrating by parts the diffusion term, the following integral

$$\int_{\tilde{\Gamma}_N(\mu_1)} \tilde{h} v d\tilde{\Gamma}$$

appears in the linear form at the right-hand side.

Concerning instead Dirichlet boundary conditions, let us introduce a lifting function $\tilde{r}_g \in H^1(\tilde{\Omega}(\mu_1))$ such that

$$\tilde{r}_g(\boldsymbol{\mu})|_{\tilde{\Gamma}_D(\mu_1)} = \tilde{g}(\boldsymbol{\mu}).$$

The weak formulation reads as follows: find $\tilde{u}(\boldsymbol{\mu}) \in \tilde{V}(\boldsymbol{\mu}) = H_{\tilde{\Gamma}_D}^1(\tilde{\Omega}(\mu_1))$ such that

$$\tilde{a}(\tilde{u}(\boldsymbol{\mu}), v; \boldsymbol{\mu}) = \tilde{f}(v; \boldsymbol{\mu}) \quad \forall v \in \tilde{V}(\boldsymbol{\mu}) \quad (8.27)$$

where $\tilde{a}(\cdot, \cdot; \boldsymbol{\mu})$ is defined as in (8.17), whereas

$$\tilde{f}(v; \boldsymbol{\mu}) = \int_{\tilde{\Omega}(\mu_1)} \tilde{s} v d\tilde{\Omega} + \int_{\tilde{\Gamma}_N(\mu_1)} \tilde{h} v d\tilde{\Gamma} - \tilde{a}(\tilde{r}_g, v; \boldsymbol{\mu}). \quad (8.28)$$

The solution of the original problem (8.26) is thus given by $\tilde{u}(\boldsymbol{\mu}) + \tilde{r}_g(\boldsymbol{\mu})$.

To obtain the transformed problem over the reference domain Ω , we denote by $g(\boldsymbol{\mu}) = \tilde{g}(\boldsymbol{\mu}) \circ \boldsymbol{\Phi}(\cdot; \boldsymbol{\mu})$ and $h(\boldsymbol{\mu}) = \tilde{h}(\boldsymbol{\mu}) \circ \boldsymbol{\Phi}(\cdot; \boldsymbol{\mu})$ the transformed Dirichlet and Neumann data, respectively. We point out that, even if the original data \tilde{g} and \tilde{h} are parameter independent, g and h are $\boldsymbol{\mu}$ -dependent because of the Jacobian of the $\boldsymbol{\mu}$ -dependent map $\boldsymbol{\Phi}$. In the same way, let us denote by $r_g \in H^1(\Omega)$ a lifting function such that

$$r_g(\boldsymbol{\mu})|_{\Gamma_D} = g(\boldsymbol{\mu}). \quad (8.29)$$

Moreover, let us recall the following formula for the change of variables in *boundary* integrals

$$\int_{\tilde{\Gamma}(\boldsymbol{\mu})} \tilde{\psi} d\tilde{\Gamma} = \int_{\Gamma} \psi |\mathbb{J}_{\boldsymbol{\Phi}}(\mathbf{x}; \boldsymbol{\mu})| \mathbf{t} d\Gamma$$

where \mathbf{t} is the unit tangent vector on Γ . The weak formulation of (8.26) over the reference domain thus reads: find $u(\boldsymbol{\mu}) \in V = H_{\Gamma_D}^1(\Omega)$ such that

$$a(u(\boldsymbol{\mu}), v; \boldsymbol{\mu}) = f(v; \boldsymbol{\mu}) \quad \forall v \in V \quad (8.30)$$

where $a(\cdot, \cdot; \boldsymbol{\mu})$ is defined as in (8.22) and

$$f(v; \boldsymbol{\mu}) = \int_{\Omega} s v |\mathbb{J}_{\boldsymbol{\Phi}}(\mathbf{x}; \boldsymbol{\mu})| d\Omega + \int_{\Gamma} h v |\mathbb{J}_{\boldsymbol{\Phi}}(\mathbf{x}; \boldsymbol{\mu})| \mathbf{t} d\Gamma - a(r_g(\boldsymbol{\mu}), v; \boldsymbol{\mu}). \quad (8.31)$$

The solution of the problem in Ω is thus given by $u(\boldsymbol{\mu}) + r_g(\boldsymbol{\mu})$.

8.4 Advection-Diffusion-Reaction, Case II: Mass Transfer with Parametrized Source

We now consider a mass transfer problem describing for instance the behavior of pollutant emissions released by industrial chimneys into the atmosphere, or by a plant in a river. In these cases, the evolution of the concentration of the pollutant can be modelled by an advection-diffusion-reaction equation, while the emission is described by a (either distributed or pointwise) parametrized source term, under the following parametrized PDE form:

$$\begin{cases} -\mu_1 \Delta u + \mathbf{b}(\mu_2) \cdot \nabla u + a_0 u = s(\boldsymbol{\mu}) & \text{in } \Omega \\ \mu_1 \nabla u \cdot \mathbf{n} = 0 & \text{on } \Gamma_N = \partial\Omega. \end{cases} \quad (8.32)$$

Here $u = u(\boldsymbol{\mu})$ is the pollutant concentration, a_0 is a positive constant representing the intensity of reaction processes, $\mathbf{b}(\mu_2)$ is a (constant in space) advection field,

$$\mathbf{b}(\mu_2) = [\cos(\mu_2) \ \sin(\mu_2)]^T$$

representing e.g. the wind speed, and

$$s(\mathbf{x}; \boldsymbol{\mu}) = \exp\left(-\frac{(x_1 - \mu_3)^2 + (x_2 - \mu_4)^2}{\mu_5^2}\right)$$

describes the pollutant emission, characterized in terms of its position (μ_3, μ_4) and its spreading μ_5 : for small values of μ_5 the source is much localized around its center (μ_3, μ_4) (with a point source in the limit $\mu_5 \rightarrow 0$), whereas larger values of μ_5 yield a larger spreading of the source. Parameters represent physical properties, such as the molecular diffusivity μ_1 of the chemical species, the direction of the wind speed μ_2 and the location of the source term.

In this case the original domain is fixed (parameter-independent), and thus coincides with the reference one, that is $\tilde{\Omega} = \Omega$. No geometric transformation is therefore required. The weak formulation of problem (8.32) can be directly set over Ω : find $u(\boldsymbol{\mu}) \in V = H^1(\Omega)$ such that

$$a(u(\boldsymbol{\mu}), v; \boldsymbol{\mu}) = f(v; \boldsymbol{\mu}) \quad \forall v \in V \quad (8.33)$$

where

$$a(u, v; \boldsymbol{\mu}) = \int_{\Omega} \left(\mu_1 \nabla u \cdot \nabla v + \cos(\mu_2) \frac{\partial u}{\partial x_1} v + \sin(\mu_2) \frac{\partial u}{\partial x_2} v \right) d\Omega, \quad (8.34)$$

$$f(v; \boldsymbol{\mu}) = \int_{\Omega} e^{-\frac{(x_1 - \mu_3)^2 + (x_2 - \mu_4)^2}{\mu_5^2}} v d\Omega. \quad (8.35)$$

The bilinear form $a(u, v; \boldsymbol{\mu})$ can be expressed through an affine expansion of the form (8.24). Unfortunately, the linear form $f(v; \boldsymbol{\mu})$ cannot be automatically ex-

pressed under the form (8.25), since $\boldsymbol{\mu}$ and \mathbf{x} are not separable variables. In other words, the exponential function yields a *nonaffine parametrization*, since the expression appearing in the integral (8.35) is a nonaffine function of \mathbf{x} and $\boldsymbol{\mu}$.

If the problem is not affinely parametrized (this in general occurs when the geometric transformation (8.1) is not affine, or if some physical coefficients are non-affine functions of \mathbf{x} and $\boldsymbol{\mu}$), the parameter $\boldsymbol{\mu}$ cannot be separated from the spatial coordinates \mathbf{x} in the quantities appearing under the sign of integral. It is however important to notice that whenever the affinity assumption is not naturally induced by the problem, it can be recovered through interpolation, yielding an additional pre-processing phase before assembling the finite element structures. Chapter 10 will be devoted to the analysis of *nonaffine problems*.

8.5 Advection-Diffusion-Reaction, Case III: Mass Transfer in a Parametrized Domain

We now consider a mass transfer problem in a parametrized domain described by a more general (nonaffine) geometric transformation. Deriving the weak formulation on the reference domain in this case is more challenging.

We consider as reference domain a cylinder of radius $R = 0.5$ and length $L = 5$,

$$\Omega = \{\mathbf{x} \in \mathbb{R}^3 : x_1 + x_2 \leq R^2, x_3 \in (0, L)\}$$

where

$$\Gamma_{in} = \{\mathbf{x} \in \mathbb{R}^3 : x_1 + x_2 < R^2, x_3 = 0\}, \quad \Gamma_{out} = \{\mathbf{x} \in \mathbb{R}^3 : x_1 + x_2 < R^2, x_3 = L\}$$

are its inlet and outlet sections, respectively, and

$$\Gamma_w = \{\mathbf{x} \in \mathbb{R}^3 : x_1 + x_2 = R^2, x_3 \in (0, L)\}$$

represents its lateral surface. By means of the following parametric map

$$\tilde{\mathbf{x}} = \boldsymbol{\Phi}(\mathbf{x}; \mu_1), \quad \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \end{bmatrix} = \begin{bmatrix} x_1 \left(1 - \mu_1 \exp \left(-\frac{(x_3 - c)^2}{0.1} \right) \right) \\ x_2 \left(1 - \mu_1 \exp \left(-\frac{(x_3 - c)^2}{0.1} \right) \right) \\ x_3 \end{bmatrix} \quad (8.36)$$

we operate a (regular) radial restriction on the section $x_3 = c$ (with $c = 1.5$), of amplitude $\mu_1 < 1$, thus yielding a lateral surface featuring a bell-shape section. The parametric map (8.36) is nonaffine, thus its Jacobian $\mathbb{J}_{\boldsymbol{\Phi}}(\mathbf{x}; \boldsymbol{\mu}_g)$ depends both on the spatial variables and on the geometric parameter μ_1 . An example of deformation generated by the map (8.36) is shown in Fig. 8.2.

We consider the following mass transfer problem, describing the diffusion-advection-reaction of a chemical species released from the inlet section $\tilde{\Gamma}_{in}$ of the

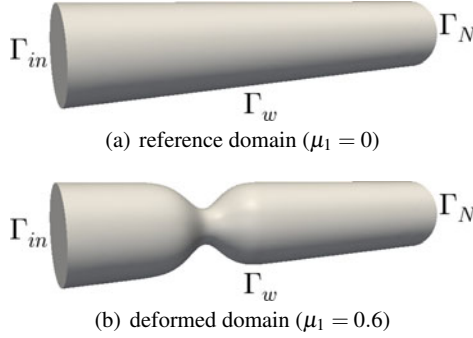


Fig. 8.2 Reference and original domain with boundaries for problem (8.58)

cylinder

$$\left\{ \begin{array}{ll} -\operatorname{div}(\tilde{\mathbf{k}}(\tilde{\mathbf{x}}; \boldsymbol{\mu}_{ph}) \nabla \tilde{u}) + \tilde{\mathbf{b}}(\tilde{\mathbf{x}}; \boldsymbol{\mu}_{ph}) \cdot \nabla \tilde{u} + \tilde{a}_0(\tilde{\mathbf{x}}; \boldsymbol{\mu}_{ph}) \tilde{u} = \tilde{s}(\tilde{\mathbf{x}}; \boldsymbol{\mu}_{ph}) & \text{in } \tilde{\Omega}(\mu_1) \\ \tilde{u} = \tilde{g}(\boldsymbol{\mu}_{ph}) & \text{on } \tilde{\Gamma}_{in} \\ \tilde{u} = 0 & \text{on } \tilde{\Gamma}_w \\ \tilde{\mathbf{k}}(\tilde{\mathbf{x}}; \boldsymbol{\mu}_{ph}) \nabla \tilde{u} \cdot \tilde{\mathbf{n}} = \tilde{h}(\boldsymbol{\mu}_{ph}) & \text{on } \tilde{\Gamma}_{out} \end{array} \right. \quad (8.37)$$

where $\tilde{g}(\boldsymbol{\mu}_{ph}) \in L^2(\partial \tilde{\Gamma}_{in})$ and $\tilde{h}(\boldsymbol{\mu}_{ph}) \in L^2(\tilde{\Gamma}_{out})$ are parametrized Dirichlet and Neumann data prescribed on $\tilde{\Gamma}_D = \tilde{\Gamma}_{in} \cup \tilde{\Gamma}_w$ and $\tilde{\Gamma}_N = \tilde{\Gamma}_{out}$, respectively. In this case we suppose that the parametrized diffusivity of the chemical species is anisotropic in space, and is thus described by a symmetric and positive definite matrix $\tilde{\mathbf{k}}(\tilde{\mathbf{x}}; \boldsymbol{\mu}_{ph}) : \mathbb{R}^d \times \mathcal{P}_p \rightarrow \mathbb{R}^{d \times d}$. Moreover, $\tilde{s}(\tilde{\mathbf{x}}; \boldsymbol{\mu}_{ph})$ represents a distributed source in $\tilde{\Omega}$.

The weak formulation of (8.37) over the reference domain Ω reads: find $u = u(\boldsymbol{\mu}) \in V = H_{\Gamma_D}^1(\Omega)$ such that

$$a(u(\boldsymbol{\mu}), v; \boldsymbol{\mu}) = f(v; \boldsymbol{\mu}) \quad \forall v \in V \quad (8.38)$$

where

$$\begin{aligned} a(u, v; \boldsymbol{\mu}) = & \sum_{i,j=1}^d \int_{\Omega} \frac{\partial u}{\partial x_i} v_{ij}(\mathbf{x}; \boldsymbol{\mu}) \frac{\partial v}{\partial x_j} d\Omega + \sum_{i=1}^d \int_{\Omega} \beta_i(\mathbf{x}; \boldsymbol{\mu}) \frac{\partial u}{\partial x_i} v d\Omega \\ & + \int_{\Omega} a_0(\mathbf{x}; \boldsymbol{\mu}_{ph}) uv |\mathbb{J}\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\mu}_g)| d\Omega \end{aligned} \quad (8.39)$$

and

$$\begin{aligned} f(v; \boldsymbol{\mu}) = & \int_{\Omega} s(\mathbf{x}, \boldsymbol{\mu}_{ph}) v |\mathbb{J}\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\mu}_g)| d\Omega \\ & + \int_{\Gamma} h(\mathbf{x}, \boldsymbol{\mu}_{ph}) v |\mathbb{J}\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\mu}_g)| \mathbf{t} d\Gamma - a(r_g(\boldsymbol{\mu}), v; \boldsymbol{\mu}). \end{aligned} \quad (8.40)$$

Here we have denoted by

$$\mathbf{v}(\mathbf{x}; \boldsymbol{\mu}) = \mathbb{J}_{\boldsymbol{\Phi}}^{-1}(\mathbf{x}; \boldsymbol{\mu}) \widetilde{\boldsymbol{\kappa}}(\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\mu}_g); \boldsymbol{\mu}_{ph}) \mathbb{J}_{\boldsymbol{\Phi}}^{-T}(\mathbf{x}; \boldsymbol{\mu}_g) |\mathbb{J}_{\boldsymbol{\Phi}}(\mathbf{x}; \boldsymbol{\mu}_g)| \quad (8.41)$$

the transformed diffusion tensor in the reference domain, and by

$$\boldsymbol{\beta}(\mathbf{x}; \boldsymbol{\mu}) = \mathbb{J}_{\boldsymbol{\Phi}}^{-1}(\mathbf{x}; \boldsymbol{\mu}_g) \widetilde{\mathbf{b}}(\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\mu}_g); \boldsymbol{\mu}_{ph}) |\mathbb{J}_{\boldsymbol{\Phi}}(\mathbf{x}; \boldsymbol{\mu}_g)| \quad (8.42)$$

the transformed advection field over the reference domain. Moreover, we have denoted by $r_g \in H^1(\Omega)$ a lifting function as in (8.29). The solution in Ω is thus given by $u(\boldsymbol{\mu}) + r_g(\boldsymbol{\mu})$. See Exercise 3 for this derivation. Some remarks are in order:

1. the diffusion tensor (8.41) and the advection field (8.42) depend on $\boldsymbol{\mu}_g$ even though their original counterparts $\widetilde{\boldsymbol{\kappa}}$ and $\widetilde{\mathbf{b}}$ did not;
2. because of the nonaffine parametrization of the original problem, the affine expansions (8.24)-(8.25) do not hold in this case. An approximate affine expansion can however be defined, see Chap. 10;
3. in some cases of interest (see e.g. the problem discussed in Sect. 8.3), we may take advantage of splitting the reference domain Ω into subdomains

$$\overline{\Omega} = \bigcup_{k=1}^{n_{dom}} \overline{\Omega}_k,$$

for, either (i) the sake of geometric representation, or (ii) to better account for different physical properties holding over the different subdomains.

In this case, each original subdomain $\widetilde{\Omega}_k$ can be regarded as the image of a subdomain Ω_k of Ω through a map $\boldsymbol{\Phi}_{\Omega_k} : \Omega^k \times \mathcal{P} \rightarrow \mathbb{R}^d$, that is $\widetilde{\Omega}_k(\boldsymbol{\mu}) = \boldsymbol{\Phi}_{\Omega_k}(\Omega_k; \boldsymbol{\mu})$, $k = 1, \dots, n_{dom}$; these maps must satisfy

$$\boldsymbol{\Phi}_{\Omega_k}(\mathbf{x}; \boldsymbol{\mu}_g) = \boldsymbol{\Phi}_{\Omega_{k'}}(\mathbf{x}; \boldsymbol{\mu}_g) \quad \forall \mathbf{x} \in \Omega_k \cap \Omega_{k'}, \quad 1 \leq k < k' \leq n_{dom}.$$

In the most general case, accounting for different physical coefficients over the subdomains, and different geometric maps $\boldsymbol{\Phi}_{\Omega_k}$ on each subdomain Ω_k , instead of the bilinear form (8.39) we might have

$$\begin{aligned} a(u, v; \boldsymbol{\mu}) &= \sum_{i,j=1}^d \sum_{k=1}^{n_{dom}} \int_{\Omega_k} \frac{\partial u}{\partial x_i} v_{k,ij}(\mathbf{x}; \boldsymbol{\mu}) \frac{\partial v}{\partial x_j} d\Omega \\ &\quad + \sum_{i=1}^d \sum_{k=1}^{n_{dom}} \int_{\Omega_k} \beta_{k,i}(\mathbf{x}; \boldsymbol{\mu}) \frac{\partial u}{\partial x_i} v d\Omega \\ &\quad + \sum_{k=1}^{n_{dom}} \int_{\Omega_k} a_{0,k}(\mathbf{x}; \boldsymbol{\mu}_{ph}) uv |\mathbb{J}_{\boldsymbol{\Phi}_k}(\mathbf{x}; \boldsymbol{\mu}_g)| d\Omega, \end{aligned}$$

where

$$\mathbf{v}_k(\mathbf{x}; \boldsymbol{\mu}) = \mathbb{J}_{\boldsymbol{\Phi}_k}^{-1}(\mathbf{x}; \boldsymbol{\mu}) \widetilde{\boldsymbol{\kappa}}_k(\boldsymbol{\Phi}_k(\mathbf{x}; \boldsymbol{\mu}_g); \boldsymbol{\mu}_{ph}) \mathbb{J}_{\boldsymbol{\Phi}_k}^{-T}(\mathbf{x}; \boldsymbol{\mu}_g) |\mathbb{J}_{\boldsymbol{\Phi}_k}(\mathbf{x}; \boldsymbol{\mu}_g)|$$

and

$$\boldsymbol{\beta}_k(\mathbf{x}; \boldsymbol{\mu}) = \mathbb{J}_{\boldsymbol{\Phi}_k}^{-1}(\mathbf{x}; \boldsymbol{\mu}_g) \tilde{\mathbf{b}}_k(\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\mu}_g); \boldsymbol{\mu}_{ph}) |\mathbb{J}_{\boldsymbol{\Phi}_k}(\mathbf{x}; \boldsymbol{\mu}_g)|$$

represent the diffusion tensor and of the advection field over the subdomain Ω_k , respectively. In the same way, the linear form (8.40) would be replaced by

$$\begin{aligned} f(v; \boldsymbol{\mu}) &= \sum_{k=1}^{n_{dom}} \int_{\Omega_k} s_k(\mathbf{x}; \boldsymbol{\mu}_{ph}) v |\mathbb{J}_{\boldsymbol{\Phi}_k}(\mathbf{x}; \boldsymbol{\mu}_g)| d\Omega \\ &+ \sum_{l=1}^{L_{neum}} \int_{\Gamma_l} h(\mathbf{x}; \boldsymbol{\mu}_{ph}) v |\mathbb{J}_{\boldsymbol{\Phi}_k}(\mathbf{x}; \boldsymbol{\mu}_g)| \mathbf{t} |d\Gamma - a(r_g(\boldsymbol{\mu}), v; \boldsymbol{\mu}) \end{aligned}$$

where the Neumann boundary becomes $\Gamma_N = \cup_{l=1}^{L_{neum}} \Gamma_l$ and

$$\delta_{kl} = \begin{cases} 1 & \text{if } \Gamma_l \subset \partial\Omega_k \\ 0 & \text{otherwise.} \end{cases}$$

8.5.1 More on the Transformation of Vector Fields

So far, when mapping the problem onto the reference domain, for the sake of simplicity we have considered a straightforward *componentwise* transformation of the advection field. By so doing, starting from the vector function $\tilde{\mathbf{b}}(\tilde{\mathbf{x}}; \boldsymbol{\mu}_{ph})$ representing the advection field over the original domain, we obtain the corresponding expression on the reference domain as

$$\mathbf{b}(\mathbf{x}; \boldsymbol{\mu}) = \tilde{\mathbf{b}}(\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\mu}_g); \boldsymbol{\mu}_{ph}) \quad \forall \mathbf{x} \in \Omega. \quad (8.43)$$

Similarly, we would obtain

$$\tilde{\mathbf{b}}(\tilde{\mathbf{x}}; \boldsymbol{\mu}) = \mathbf{b}(\boldsymbol{\Phi}^{-1}(\tilde{\mathbf{x}}; \boldsymbol{\mu}_g); \boldsymbol{\mu}_{ph}) \quad \forall \tilde{\mathbf{x}} \in \tilde{\Omega} \quad (8.44)$$

if we start from the advection field $\mathbf{b}(\mathbf{x}; \boldsymbol{\mu}_{ph})$ defined over the reference domain.

A different transformation should however be considered in order to preserve some important physical features of the advection field over each parametrized configuration. For instance, let us consider the geometric configuration described in the previous section (see Fig. 8.2) and the (parameter independent) advection field

$$\mathbf{b}(x_1, x_2, x_3) = \left[0, 0, 1 - \frac{x_1^2 + x_2^2}{R^2} \right]^T, \quad (x_1, x_2, x_3) \in \Omega \quad (8.45)$$

corresponding to a Poiseuille parabolic flow vanishing on Γ_w (see also Fig. 8.3(a)). The flow develops along the x_3 direction (parallel to the cylinder's axis), hence only the third component is non-null. By transforming the vector field (8.45) under (8.44), we obtain an advection field which remains parallel to the x_3 direction and does not deflect when passing through the cylinder restriction (see Exercise 4).

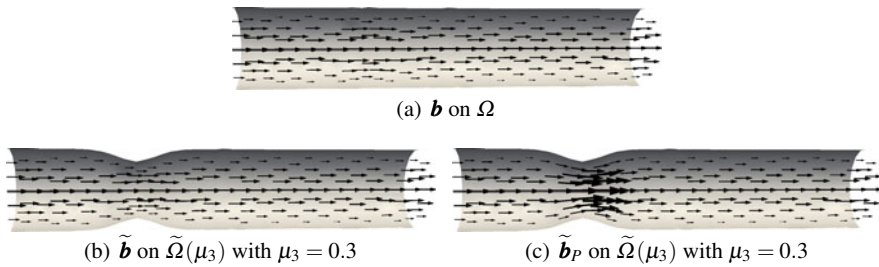


Fig. 8.3 Transformation of the advection field defined in (8.45)

A closer look at Fig. 8.3(b) reveals that the transformed advection field is not physically meaningful: in fact, although it vanishes on the lateral boundary of the cylinder Γ_w and reaches its maximum on the centerline of the cylinder, its flux over each longitudinal section $x_3 = l$, $z \in (0, L)$ of the cylinder is not constant. In other terms, even if $\operatorname{div}_{\mathbf{x}} \mathbf{b} = 0$ for any $\mathbf{x} \in \Omega$, unfortunately $\operatorname{div}_{\tilde{\mathbf{x}}} \tilde{\mathbf{b}} \neq 0$ for any $\tilde{\mathbf{x}} \in \tilde{\Omega}$.

A map of vector fields which is divergence-preserving is the *Piola transformation*: for any $\mathbf{u} : \Omega \rightarrow \mathbb{R}^d$, the Piola transformation $\tilde{\mathbf{u}}_P = \mathbf{P}_{\Phi}(\mathbf{u}) : \tilde{\Omega} \rightarrow \mathbb{R}^d$ is given by

$$\tilde{\mathbf{u}}_P(\tilde{\mathbf{x}}) = \frac{1}{|\mathbb{J}_{\Phi}(\mathbf{x})|} \mathbb{J}_{\Phi}(\mathbf{x}) \mathbf{u}(\mathbf{x}) \quad (8.46)$$

where $\tilde{\mathbf{x}} = \Phi(\mathbf{x})$. If $\tilde{\mathbf{u}}_P = \mathbf{P}_{\Phi}(\mathbf{u})$, $\tilde{q} = q \circ \Phi^{-1}$ for some $q : \Omega \rightarrow \mathbb{R}$, and $\tilde{\mathbf{n}}, \mathbf{n}$ denote the unit outward normals on $\partial \tilde{\Omega}$ and $\partial \Omega$, respectively, then (see Exercise 5)

$$\int_{\tilde{\Omega}} \operatorname{div}_{\tilde{\mathbf{x}}} \tilde{\mathbf{u}}_P \tilde{q} d\tilde{\Omega} = \int_{\Omega} \operatorname{div}_{\mathbf{x}} \mathbf{u} q d\Omega, \quad (8.47)$$

$$\int_{\partial \tilde{\Omega}} \tilde{\mathbf{u}}_P \cdot \tilde{\mathbf{n}} d\tilde{\Omega} = \int_{\partial \Omega} \mathbf{u} \cdot \mathbf{n} d\Omega. \quad (8.48)$$

We report in Fig. 8.3(c) an instance of the advection field $\tilde{\mathbf{b}}_P$ obtained via the Piola transformation.

Operating the Piola transformation of the vector quantities appearing in a PDE problem in general leads to a much more involved formulation. This, however, is not always mandatory. For instance, if the vector field vanishes over the boundaries whose orientation is not preserved by the geometric transformation, the component-wise transformation (8.43)–(8.44) allows to preserve the divergence property of the vector field, that is, (8.47)–(8.48) are still valid by replacing $\tilde{\mathbf{u}}_P$ with $\tilde{\mathbf{u}}$.

In other cases, the Piola transformation may not be meaningful (from a modeling point of view) and should therefore be avoided. This is for instance the case (see Fig. 8.4) where the orientation of the advection field ought to be the same on both the original and the reference domain.

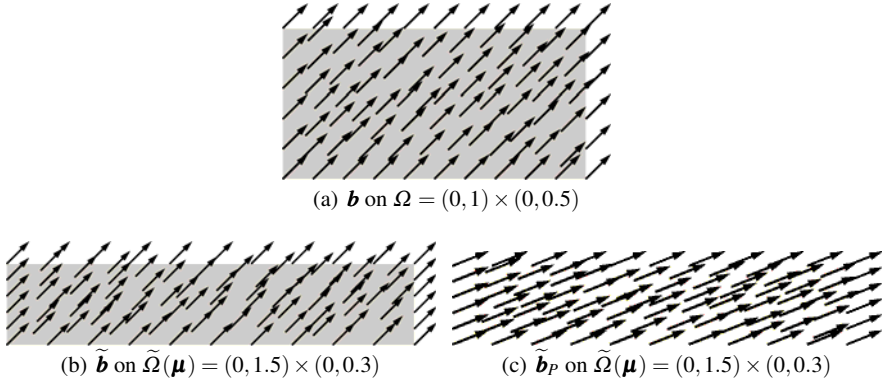


Fig. 8.4 Transformation of the advection field $\mathbf{b} = (1, 1)^T$

8.6 Linear Elasticity: An Elastic Beam

We now consider a problem modeling the displacement of a given elastic structure under prescribed normal stresses (and zero displacement) on its boundaries. Our medium represents a three-dimensional elastic beam with a given shape, made of materials with different properties (such as the Young modulus and the Poisson coefficient).

We are interested, for instance, to characterize different scenarios, such as finding which material allows to keep the displacement of the beam under a suitable target value, once a given traction is applied, or, alternatively, which is the highest traction intensity admissible over a single boundary face by considering different material properties, in order for the displacement to stand below a prescribed threshold.

In this case the original domain is fixed (parameter-independent), and thus it coincides with the reference one, that is $\tilde{\Omega} = \Omega$. The parametrized PDE is therefore given by the linear elasticity equations (2.2), supplemented by parametrized boundary conditions:

$$\left\{ \begin{array}{ll} -\operatorname{div}(\mu(\nabla \mathbf{u} + \nabla \mathbf{u}^T) + \lambda(\operatorname{div} \mathbf{u})\mathbf{I}) = \mathbf{0} & \text{in } \Omega \\ \mathbf{u} = \mathbf{0} & \text{on } \Gamma_D \\ \boldsymbol{\sigma} \mathbf{n} = \mu_3 \mathbf{n} & \text{on } \Gamma_{N_1} \\ \boldsymbol{\sigma} \mathbf{n} = \mu_4 \mathbf{n} & \text{on } \Gamma_{N_2} \\ \boldsymbol{\sigma} \mathbf{n} = \mu_5 \mathbf{n} & \text{on } \Gamma_{N_3}. \end{array} \right. \quad (8.49)$$

The Lamé coefficients μ and λ are expressed in terms of two parameters, the Young modulus $\mu_1 = E$ and the Poisson coefficient $\mu_2 = \nu$,

$$\lambda(\mu_1, \mu_2) = \frac{\mu_1 \mu_2}{(1 + \mu_2)(1 - 2\mu_2)}, \quad \mu(\mu_1, \mu_2) = \frac{\mu_1}{2(1 + \mu_2)}.$$

The weak formulation of problem (8.49) reads: find $\mathbf{u}(\boldsymbol{\mu}) \in V = [H_{D}^1(\Omega)]^3$ such that

$$a(\mathbf{u}, \mathbf{v}; \boldsymbol{\mu}) = f(\mathbf{v}; \boldsymbol{\mu}) \quad \forall \mathbf{v} \in V, \quad (8.50)$$

where

$$a(\mathbf{u}, \mathbf{v}; \boldsymbol{\mu}) = \int_{\Omega} 2\mu(\mu_1, \mu_2) \boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{v}) d\Omega + \int_{\Omega} \lambda(\mu_1, \mu_2) \operatorname{div}(\mathbf{u}) \operatorname{div}(\mathbf{v}) d\Omega \quad (8.51)$$

and

$$f(\mathbf{v}; \boldsymbol{\mu}) = \int_{\Gamma_{N_1}} \mu_3 \mathbf{n} \cdot \mathbf{v} d\Gamma + \int_{\Gamma_{N_2}} \mu_4 \mathbf{n} \cdot \mathbf{v} d\Gamma + \int_{\Gamma_{N_3}} \mu_5 \mathbf{n} \cdot \mathbf{v} d\Gamma. \quad (8.52)$$

The problem is affinely parametrized, and an expansion for both $a(\cdot, \cdot; \boldsymbol{\mu})$ and $f(\cdot; \boldsymbol{\mu})$ under the form (8.24)–(8.25) can be obtained setting $Q_a = 2$ and

$$\begin{aligned} \theta_a^1(\boldsymbol{\mu}) &= \frac{\mu_1}{(1 + \mu_2)}, & a_1(u, v) &= \int_{\Omega} \boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{v}) d\Omega \\ \theta_a^2(\boldsymbol{\mu}) &= \frac{\mu_1 \mu_2}{(1 + \mu_2)(1 - 2\mu_2)}, & a_2(u, v) &= \int_{\Omega} \operatorname{div}(\mathbf{u}) \operatorname{div}(\mathbf{v}) d\Omega \end{aligned}$$

for the bilinear form $a(\cdot, \cdot; \boldsymbol{\mu})$, and $Q_f = 3$,

$$\begin{aligned} \theta_f^1(\boldsymbol{\mu}) &= \mu_3, & f_1(v) &= \int_{\Gamma_{N_1}} \mathbf{n} \cdot \mathbf{v} d\Gamma \\ \theta_f^2(\boldsymbol{\mu}) &= \mu_4, & f_2(v) &= \int_{\Gamma_{N_2}} \mathbf{n} \cdot \mathbf{v} d\Gamma \\ \theta_f^3(\boldsymbol{\mu}) &= \mu_5, & f_3(v) &= \int_{\Gamma_{N_3}} \mathbf{n} \cdot \mathbf{v} d\Gamma \end{aligned}$$

for the linear form $f(\cdot; \boldsymbol{\mu})$.

8.7 Fluid Flows, Case I: Backward-Facing Step Channel

We now consider a fluid flow (modelled by either the Stokes or the Navier-Stokes equations) through a parametrized backward-facing step channel such as the one reported in Fig. 8.7.

We also deal with different flow regimes, by changing the viscosity and the inlet velocity, thus yielding a parametrized problem depending on three parameters: the kinematic viscosity $\mu_1 = \nu$, the amplitude μ_2 of the inlet velocity profile \mathbf{g} , and the step height μ_3 . This is a classical benchmark problem in computational fluid dynamics, which has been intensively investigated also for the sake of hydrodynamic stability theory, see e.g. [31, 123].

The steady Navier-Stokes problem in this case reads as

$$\left\{ \begin{array}{ll} -\frac{1}{\mu_1} \Delta \tilde{\mathbf{u}} + \delta(\tilde{\mathbf{u}} \cdot \nabla) \tilde{\mathbf{u}} + \nabla \tilde{p} = \mathbf{0} & \text{in } \tilde{\Omega}(\mu_3) \\ \operatorname{div} \tilde{\mathbf{u}} = 0 & \text{in } \tilde{\Omega}(\mu_3) \\ \tilde{\mathbf{u}} = \mu_2 \tilde{\mathbf{g}} & \text{on } \tilde{\Gamma}_{in}(\mu_3) \\ \tilde{\mathbf{u}} = \mathbf{0} & \text{on } \tilde{\Gamma}_w(\mu_3) \\ -\tilde{p} \tilde{\mathbf{n}} + \nu(\nabla \tilde{\mathbf{u}}) \tilde{\mathbf{n}} = \mathbf{0} & \text{on } \tilde{\Gamma}_N(\mu_3). \end{array} \right. \quad (8.53)$$

We denote by $(\tilde{\mathbf{u}}, \tilde{p})$ the velocity and the pressure fields defined on the original domain $\tilde{\Omega} \subset \mathbb{R}^2$, while $\tilde{\mathbf{g}} \in [L^2(\tilde{\Gamma})]^d$ is a given parabolic velocity profile. The Reynolds number is

$$\operatorname{Re} = \frac{LU}{\nu},$$

where L is a characteristic length of the domain $\tilde{\Omega}$ and U a characteristic velocity of the flow. Navier-Stokes equations correspond to the case $\delta = 1$; if $\delta = 0$, the (quadratic) convective term is neglected, obtaining the steady Stokes equations.

The weak formulation of problem (8.53) reads (see (2.29)): find $(\tilde{\mathbf{u}}, \tilde{p}) \in X(\boldsymbol{\mu}) \times Q(\boldsymbol{\mu}) = [H_{\tilde{\Gamma}_d}^1(\tilde{\Omega}(\mu_3))]^d \times L^2(\tilde{\Omega}(\mu_3))$ such that

$$\left\{ \begin{array}{ll} \tilde{d}(\tilde{\mathbf{u}}, \mathbf{v}) + \tilde{c}(\tilde{\mathbf{u}}, \tilde{\mathbf{r}}_g, \mathbf{v}) + \tilde{c}(\tilde{\mathbf{r}}_g, \tilde{\mathbf{u}}, \mathbf{v}) + \tilde{c}(\tilde{\mathbf{u}}, \tilde{\mathbf{u}}, \mathbf{v}) + \tilde{b}(\mathbf{v}, p) = \tilde{f}_1(\mathbf{v}) & \forall \mathbf{v} \in X \\ \tilde{b}(\tilde{\mathbf{u}}, q) = \tilde{f}_2(q) & \forall q \in Q, \end{array} \right. \quad (8.54)$$

where

$$\tilde{d}(\mathbf{v}, \mathbf{w}; \boldsymbol{\mu}) = \int_{\tilde{\Omega}(\mu_3)} \frac{1}{\mu_1} \nabla \mathbf{v} : \nabla \mathbf{w} d\Omega, \quad \tilde{b}(\mathbf{v}, q; \boldsymbol{\mu}) = - \int_{\Omega(\mu_3)} q \operatorname{div} \mathbf{v} d\Omega,$$

$$\tilde{c}(\mathbf{u}, \mathbf{w}, \mathbf{v}; \boldsymbol{\mu}) = \delta \int_{\tilde{\Omega}(\mu_3)} (\mathbf{u} \cdot \nabla) \mathbf{w} \cdot \mathbf{v} d\Omega,$$

$$\tilde{f}_1(\mathbf{v}; \boldsymbol{\mu}) = -\mu_2 \tilde{d}(\tilde{\mathbf{r}}_g, \mathbf{v}; \boldsymbol{\mu}) - \mu_2^2 c(\tilde{\mathbf{r}}_g, \tilde{\mathbf{r}}_g, \mathbf{v}; \boldsymbol{\mu}), \quad \tilde{f}_2(q) = -b(\tilde{\mathbf{r}}_g, q; \boldsymbol{\mu}).$$

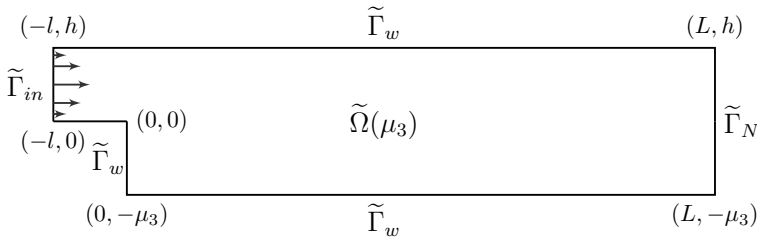


Fig. 8.5 Original domain and boundaries for problem (8.53)

Here $\tilde{\mathbf{r}}_g \in [H^1(\Omega)]^d$ is a lifting vector function such that $\tilde{\mathbf{r}}_g|_{\tilde{\Gamma}_D} = \tilde{\mathbf{g}}$; the solution to problem (8.53) is then obtained as $\mathbf{u} + \mathbf{r}_g$. For the case at hand, we can consider $Re = \mu_1\mu_2\mu_3$ as a representative Reynolds number. Some comments are in order:

1. at the right-hand side of the momentum equation, the lifting of the Dirichlet data yields the term $\mu_2\tilde{d}(\tilde{\mathbf{r}}_g, \mathbf{v}; \boldsymbol{\mu})$, as in the case of a linear problem (see, e.g., Sect. 8.3.3) and to a further term $\mu_2^2 c(\tilde{\mathbf{r}}_g, \tilde{\mathbf{r}}_g, \mathbf{v}; \boldsymbol{\mu})$ which originates from the trilinear form $\tilde{c}(\cdot, \cdot, \cdot; \boldsymbol{\mu})$.
In the same way, at the left-hand side the lifting function produces two additional terms, which are linear in $\tilde{\mathbf{u}}$, arising from the trilinear form;
2. we easily obtain the weak formulation of the corresponding Stokes problem by posing $\delta = 0$: in this case, the lifting of the Dirichlet data only produces the term coming from the bilinear form $\tilde{d}(\cdot, \cdot; \boldsymbol{\mu})$ at the right-hand side;
3. the right-hand side of the continuity equation vanishes depending on the property of the function $\tilde{\mathbf{r}}_g$: should the latter be divergence free, then $\tilde{f}_2(q) = 0$.

The following subsections are devoted to the derivation of the weak formulation of problem (8.53) onto a suitable reference domain.

8.7.1 Reference Domain and Affine Transformation

We define the reference domain as $\Omega = \tilde{\Omega}(\mu_3^{ref})$ and introduce the partition

$$\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2$$

where (see also Fig. 8.7.1)

$$\Omega_1 = (-l, L) \times (0, h), \quad \Omega_2 = (0, L) \times (0, -\mu_3^{ref}).$$

By choosing $\mu_3^{ref} = 1$, the original domain $\tilde{\Omega}(\mu_3)$ can be obtained as the image of the reference domain $\Omega = \tilde{\Omega}(\mu_3^{ref})$ through the following parametric map

$$\tilde{\mathbf{x}} = \boldsymbol{\Phi}(\mathbf{x}, \mu_3) = \begin{cases} \mathbf{x}, & \mathbf{x} \in \Omega_1 \\ \boldsymbol{\Phi}_{\Omega_2}(\mathbf{x}; \mu_3), & \mathbf{x} \in \Omega_2 \end{cases} \quad (8.55)$$

where

$$\boldsymbol{\Phi}_{\Omega_2}(\mathbf{x}; \mu_3) = \begin{bmatrix} 1 & 0 \\ 0 & \mu_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \quad (8.56)$$

In this situation, the parametric map (8.1) between \mathbf{x} and $\tilde{\mathbf{x}}(\boldsymbol{\mu})$ can be expressed through an affine transformation over each subdomain Ω_1, Ω_2 .

8.7.2 Weak Formulation on the Reference Domain

To derive the weak formulation onto the reference domain, we follow the same procedure already considered to transform the advection-diffusion-reaction problem of Sect. 8.3.1. We first characterize the Jacobian of the parametric map Φ , as well as its determinant, which enter in the definition of the parametrized tensors (8.8)–(8.10). For the case at hand, we have

$$\mathbb{J}_{\Phi_{\Omega_1}}(\tilde{\mathbf{x}}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbb{J}_{\Phi_{\Omega_2}}(\tilde{\mathbf{x}}; \mu_3) = \begin{bmatrix} 1 & 0 \\ 0 & \mu_3 \end{bmatrix}$$

and, correspondingly,

$$|\mathbb{J}_{\Phi_{\Omega_1}}(\tilde{\mathbf{x}})| = 1, \quad |\mathbb{J}_{\Phi_{\Omega_2}}(\tilde{\mathbf{x}}; \mu_3)| = \mu_3.$$

We then compute

$$\Phi_{\Omega_2}^{-1}(\mathbf{x}; \mu_3) = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{\mu_3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad (\mathbb{J}_{\Phi_{\Omega_2}}(\mathbf{x}; \mu))^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{\mu_3} \end{bmatrix}.$$

The weak formulation of problem (8.53) over the reference domain Ω reads: find $(\mathbf{u}, p) \in X \times Q = [H_{\Gamma_d}^1(\Omega)]^d \times L^2(\Omega)$ such that

$$\begin{cases} d(\mathbf{u}, \mathbf{v}; \mu) + c(\mathbf{u}, \mathbf{r}_g, \mathbf{v}; \mu) + c(\mathbf{r}_g, \mathbf{u}, \mathbf{v}; \mu) + c(\mathbf{u}, \mathbf{u}, \mathbf{v}; \mu) \\ \quad \quad \quad + b(\mathbf{v}, p; \mu) = f_1(\mathbf{v}; \mu) & \forall \mathbf{v} \in X \\ b(\mathbf{u}, q; \mu) = f_2(q; \mu) & \forall q \in Q, \end{cases} \quad (8.57)$$

where

$$\begin{aligned} d(\mathbf{v}, \mathbf{w}; \mu) &= \frac{1}{\mu_1} \int_{\Omega_1} \nabla \mathbf{v} : \nabla \mathbf{w} d\Omega + \frac{1}{\mu_1 \mu_3} \int_{\Omega_2} \frac{\partial \mathbf{v}}{\partial x_1} \frac{\partial \mathbf{w}}{\partial x_1} d\Omega + \frac{\mu_3}{\mu_1} \int_{\Omega_2} \frac{\partial \mathbf{v}}{\partial x_2} \frac{\partial \mathbf{w}}{\partial x_2} d\Omega \\ b(\mathbf{v}, q; \mu) &= - \int_{\Omega_1} q \operatorname{div} \mathbf{v} d\Omega - \int_{\Omega_2} q \left(\frac{\partial v_1}{\partial x_1} + \mu_3 \frac{\partial v_2}{\partial x_2} \right) d\Omega, \end{aligned}$$

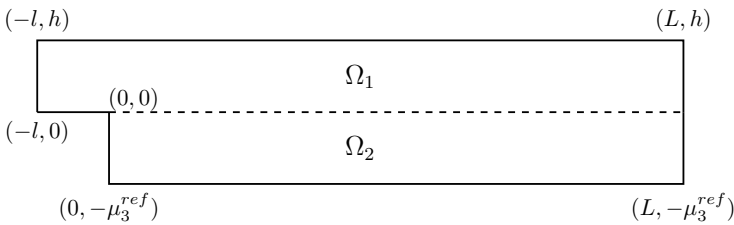


Fig. 8.6 Reference domain for problem (8.53)

$$c(\mathbf{u}, \mathbf{w}, \mathbf{v}; \boldsymbol{\mu}) = \delta \int_{\Omega_1} (\mathbf{u} \cdot \nabla) \mathbf{w} \cdot \mathbf{v} d\Omega + \delta \int_{\Omega_2} \sum_{i=1}^2 \left(u_1 \frac{\partial v_i}{\partial x_1} w_i + \mu_3 u_2 \frac{\partial v_i}{\partial x_2} w_i \right) d\Omega,$$

$$f_1(\mathbf{v}; \boldsymbol{\mu}) = -\mu_2 d(\mathbf{r}_g, \mathbf{v}; \boldsymbol{\mu}) - \mu_2^2 c(\mathbf{r}_g, \mathbf{r}_g, \mathbf{v}; \boldsymbol{\mu}), \quad f_2(q) = -b(\mathbf{r}_g, q; \boldsymbol{\mu}).$$

Here $\mathbf{r}_g \in [H^1(\Omega)]^d$ is a lifting function such that $\mathbf{r}_g|_{\Gamma_D} = \mathbf{g}$; the solution to problem (8.53) is then obtained as $\mathbf{u} + \mathbf{r}_g$.

It is straightforward to show that the forms above can be written as linear combinations of $\boldsymbol{\mu}$ -independent forms multiplied by suitable $\boldsymbol{\mu}$ -dependent scalar, real functions, under the form (8.24)–(8.25); see Exercise 6 for further details.

8.8 Fluid Flows, Case II: Sudden Expansion Channel

We now consider the case of a more involved geometric (nonaffine) parametrization; similarly to Sect. 8.5, the goal of this section is to derive the parametrized weak formulation of Navier-Stokes equations involving a geometrically parametrized domain. In particular, we consider the case of the sudden expansion domain, already introduced in Sect. 8.5. Fluid flows through sudden (rectilinear) expansions have been extensively analyzed – see e.g. [4, 92] – because of their interest for hydrodynamic stability and bifurcating solutions.

Here we consider the following problem:

$$\left\{ \begin{array}{ll} -\tilde{\mathbf{v}}(\boldsymbol{\mu}_{ph}) \Delta \tilde{\mathbf{u}} + (\tilde{\mathbf{u}} \cdot \nabla) \tilde{\mathbf{u}} + \nabla \tilde{p} = \tilde{\mathbf{s}}(\boldsymbol{\mu}_{ph}) & \text{in } \tilde{\Omega}(\boldsymbol{\mu}_g) \\ \operatorname{div} \tilde{\mathbf{u}} = 0 & \text{in } \tilde{\Omega}(\boldsymbol{\mu}_g) \\ \tilde{\mathbf{u}} = \tilde{\mathbf{g}}(\boldsymbol{\mu}_{ph}) & \text{on } \tilde{\Gamma}_{in}(\boldsymbol{\mu}_g) \\ \tilde{\mathbf{u}} = \mathbf{0} & \text{on } \tilde{\Gamma}_w(\boldsymbol{\mu}_g) \\ -\tilde{p} \tilde{\mathbf{n}} + \tilde{\mathbf{v}}(\boldsymbol{\mu}_{ph}) \frac{\partial \tilde{\mathbf{u}}}{\partial \tilde{\mathbf{n}}} = \tilde{\mathbf{h}}(\boldsymbol{\mu}_{ph}) & \text{on } \tilde{\Gamma}_N(\boldsymbol{\mu}_g) \end{array} \right. \quad (8.58)$$

where, referring to the parametric map of Sect. 8.5, we have $\boldsymbol{\mu}_g = \boldsymbol{\mu}_1$. For the sake of generality, we admit that both Dirichlet and Neumann conditions can be parameter-dependent.

The weak formulation over the reference domain Ω reads as in (8.57) with

$$d(\mathbf{v}, \mathbf{w}; \boldsymbol{\mu}) = \sum_{i,j=1}^d \int_{\Omega} \frac{\partial \mathbf{w}}{\partial x_i} v_{ij}(\mathbf{x}; \boldsymbol{\mu}) \frac{\partial \mathbf{v}}{\partial x_j} d\Omega, \quad (8.59)$$

$$b(\mathbf{v}, q; \boldsymbol{\mu}) = - \sum_{i,j=1}^d \int_{\Omega} q \eta_{ij}(\mathbf{x}; \boldsymbol{\mu}) \frac{\partial v_i}{\partial x_j} d\Omega, \quad (8.60)$$

$$c(\mathbf{u}, \mathbf{w}, \mathbf{v}; \boldsymbol{\mu}) = \delta \sum_{i,j,k=1}^d \int_{\Omega} u_i \eta_{ji}(\mathbf{x}; \boldsymbol{\mu}) \frac{\partial w_k}{\partial x_j} v_k d\Omega \quad (8.61)$$

and

$$\bar{d}(\mathbf{v}, \mathbf{w}; \boldsymbol{\mu}) = d(\mathbf{v}, \mathbf{w}; \boldsymbol{\mu}) + c(\mathbf{v}, \mathbf{r}_g, \mathbf{w}; \boldsymbol{\mu}) + c(\mathbf{r}_g, \mathbf{v}, \mathbf{w}; \boldsymbol{\mu}). \quad (8.62)$$

Here we have denoted by

$$\mathbf{v}(\mathbf{x}; \boldsymbol{\mu}) = \mathbb{J}_{\boldsymbol{\Phi}}^{-1}(\mathbf{x}; \boldsymbol{\mu}) \tilde{\mathbf{v}}(\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\mu}_g); \boldsymbol{\mu}_{ph}) \mathbb{J}_{\boldsymbol{\Phi}}^{-T}(\mathbf{x}; \boldsymbol{\mu}_g) |\mathbb{J}_{\boldsymbol{\Phi}}(\mathbf{x}; \boldsymbol{\mu}_g)| \quad (8.63)$$

and

$$\boldsymbol{\eta}(\mathbf{x}; \boldsymbol{\mu}) = \mathbb{J}_{\boldsymbol{\Phi}}^{-T}(\mathbf{x}; \boldsymbol{\mu}_g) |\mathbb{J}_{\boldsymbol{\Phi}}(\mathbf{x}; \boldsymbol{\mu}_g)| \quad (8.64)$$

the tensors appearing in the diffusion term and in both the pressure/divergence and the advection term, respectively, where

$$\tilde{\mathbf{v}}(\tilde{\mathbf{x}}; \boldsymbol{\mu}_{ph}) = \tilde{\mathbf{v}}(\boldsymbol{\mu}_p) \mathbf{I}_d,$$

$\mathbf{I}_d \in \mathbb{R}^{d \times d}$ being the identity map. Concerning the right-hand sides, we have

$$f_1(\mathbf{v}) = \int_{\Omega} \mathbf{s}(\mathbf{x}; \boldsymbol{\mu}_{ph}) \cdot \mathbf{v} |\mathbb{J}_{\boldsymbol{\Phi}}(\mathbf{x}; \boldsymbol{\mu}_g)| d\Omega \quad (8.65)$$

$$-d(\mathbf{r}_g(\boldsymbol{\mu}_{ph}), \mathbf{v}; \boldsymbol{\mu}) - c(\mathbf{r}_g(\boldsymbol{\mu}_{ph}), \mathbf{r}_g(\boldsymbol{\mu}_{ph}), \mathbf{v}; \boldsymbol{\mu}),$$

$$f_2(q) = -b(\mathbf{r}_g, q; \boldsymbol{\mu}), \quad (8.66)$$

where $\mathbf{r}_g \in [H^1(\Omega)]^d$ is a suitable lifting function. See Exercise 7 for the derivation of such a formulation.

We remark that in this case we have $\mathbf{u} = \tilde{\mathbf{u}} \circ \boldsymbol{\Phi}$, $p = \tilde{p} \circ \boldsymbol{\Phi}$. Although we are not considering the Piola transformation (8.46) to map the velocity field onto the reference domain, we consider a geometric transformation that preserves the orientation of the normal vector to those boundary portions where $\mathbf{u} \neq \mathbf{0}$, so that the flow incompressibility is conserved. For more general cases dealing with fluid flows requiring the use of the Piola transformation see, e.g., [87, 145].

8.9 Problems' Features at a Glance

For pedagogical purposes, in Table 8.1 we summarize the main features of the problems addressed in this chapter that are especially relevant for their RB approximation. In particular, we highlight if the problem at hand is either affine or nonaffine, scalar or vector, linear or nonlinear, and if contains physical or geometric parameters (or both). Finally, we indicate in which section of the book the corresponding RB approximation will be addressed.

Although we have limited our discussion to affine and simple nonaffine geometric transformations built by prescribing deformations in explicit forms, more complex geometric parametrizations can be faced. Geometric transformations based on a set of control points – whose positions play indeed the role of geometric parameters – can be efficiently used, as in the case of free-form deformation techniques [163, 186, 160], radial basis functions interpolants [185] and isogeometric analy-

Table 8.1 Main features of the parametrized problems analyzed in this chapter

	ADR I	ADR II	ADR III	ELA I	NS I	NS II
affine	✓			✓	✓	
nonaffine		✓	✓			✓
scalar	✓	✓	✓			
vector				✓	✓	✓
linear	✓	✓	✓	✓		
nonlinear					✓	✓
physical param.	✓	✓	✓	✓	✓	✓
geometric param.	✓		✓		✓	✓
reduced in Sect.	9.1	10.5	10.6	9.2	9.3.3–11.7	11.8

sis [187]. The latter allow a direct interface with computer aided design tools. The use of domain decomposition techniques and conformal mappings in the RB context has led to the so-called *reduced basis element method* and its variants, see e.g. [177, 178, 168, 169, 145] for further details. This represents an active area of investigation in the RB framework where fast progress has to be expected.

8.10 Exercises

1. Using the chain rule and the relation (8.6), derive the formulas

$$\int_{\tilde{\Omega}(\boldsymbol{\mu})} \nabla_{\tilde{\mathbf{x}}} \tilde{\psi} \cdot \nabla_{\tilde{\mathbf{x}}} \tilde{\chi} d\tilde{\Omega} = \int_{\Omega} (\mathbb{J}_{\boldsymbol{\Phi}}^{-T}(\mathbf{x}, \boldsymbol{\mu}) \nabla_{\mathbf{x}} \psi) \cdot (\mathbb{J}_{\boldsymbol{\Phi}}^{-T}(\tilde{\mathbf{x}}, \boldsymbol{\mu}) \nabla_{\tilde{\mathbf{x}}} \chi) |\mathbb{J}_{\boldsymbol{\Phi}}(\mathbf{x}, \boldsymbol{\mu})| d\Omega$$

$$\int_{\tilde{\Omega}(\boldsymbol{\mu})} \tilde{\mathbf{b}} \cdot \nabla_{\tilde{\mathbf{x}}} \tilde{\psi} \tilde{\chi} d\tilde{\Omega} = \int_{\Omega} \mathbf{b} \cdot (\mathbb{J}_{\boldsymbol{\Phi}}^{-1}(\mathbf{x}, \boldsymbol{\mu}) \nabla_{\mathbf{x}} \psi) \chi |\mathbb{J}_{\boldsymbol{\Phi}}(\mathbf{x}, \boldsymbol{\mu})| d\Omega$$

for the change of coordinates in integrals involving derivatives, and show that they can be written in the more compact form (8.7)–(8.9).

2. Derive the weak formulation of problem (8.15) over the reference domain, by proceeding through the following steps:
 - a. derive the expression of the forms (8.17) and (8.18);
 - b. find the expression of the map (8.19)–(8.20) and characterize the expression of the tensors $\mathbf{v}(\mathbf{x}, \boldsymbol{\mu})$, $\boldsymbol{\eta}(\mathbf{x}, \boldsymbol{\mu})$ defined by (8.8)–(8.10);
 - c. finally, derive the expression of the forms (8.22)–(8.23) appearing in the parametrized weak formulation (8.21).

3. Derive the weak formulation of problem (8.37) over the reference domain, by considering the steps of Exercise 2 in order to recover the expression of the bilinear and linear forms (8.39)–(8.40).
4. Consider as reference domain $\tilde{\Omega}$ the cylinder defined in Sect. 8.5 and the vector field (8.45) defined over $\tilde{\Omega}$. Show that under the map (8.44), the resulting advection field $\tilde{b}(\tilde{\mathbf{x}}; \tilde{\boldsymbol{\mu}})$ remains parallel to the x_3 direction.
5. Taking advantage of the relations for the change of variables in integrals involving derivatives, and using the definition (8.46) of Piola transformation, show that if $\tilde{\mathbf{u}}_P = \mathbf{P}_{\Phi}(\mathbf{u})$, $\tilde{q} = q \circ \Phi^{-1}$ for some $q : \Omega \rightarrow \mathbb{R}$, and $\tilde{\mathbf{n}}, \mathbf{n}$ denote the unit outward normals on $\partial\tilde{\Omega}$ and $\partial\Omega$, respectively, then

$$\int_{\tilde{\Omega}} \operatorname{div}_{\tilde{\mathbf{x}}} \tilde{\mathbf{u}}_P q d\tilde{\Omega} = \int_{\Omega} \operatorname{div}_{\mathbf{x}} \mathbf{u} q d\Omega,$$

$$\int_{\partial\tilde{\Omega}} \tilde{\mathbf{u}}_P \cdot \tilde{\mathbf{n}} d\tilde{\Omega} = \int_{\partial\Omega} \mathbf{u} \cdot \mathbf{n} d\Omega.$$

6. Derive the weak formulation of problem (8.53) onto the reference domain $\tilde{\Omega}$ by taking advantage of the change of variables formulas introduced in Sect. 8.2.1. Then, recover an affine expansion for the forms appearing in (8.57).
7. Derive the weak formulation of problem (8.58) by taking advantage of the change of variables formulas introduced in Sect. 8.2.1.

Chapter 9

RB Methods in Action: Computing the Solution

We present a selection of numerical results dealing with the RB approximation of the parametrized problems formulated in the previous chapter. For each problem we highlight the RB method's computational performance, assess its accuracy by means of a posteriori error bounds, and show various options for the construction of the RB space (either via POD or the greedy algorithm) and different projection criteria (G-RB versus LS-RB methods). Here we focus on linear affine PDEs, and defer nonaffine and nonlinear problems to Chaps. 10 and 11 respectively.

9.1 Heat Transfer: Results

We consider the heat transfer problem formulated in Sect. 8.3, modeling a cooling device for (an array of) electronic components occupying the domain $\Omega = (0, 1) \times (0, 1) \subset \mathbb{R}^2$. The problem depends on $P = 3$ parameters:

- $\mu_1 \in [-0.2, 0.6]$ is the width of the subdomain Ω_1 , see Fig. 8.1;
- $\mu_2 \in [1, 15]$ represents the maximum amplitude of the advection field;
- $\mu_3 \in [2, 30]$ is the thermal conductivity over Ω_3 .

For the high-fidelity approximation we use \mathbb{P}_1 finite elements built over a discretization of the domain made by triangular elements, resulting in 13 640 vertices, 26 874 triangles and a high-fidelity space V_h of dimension $N_h = 13\,538$. The problem is only weakly coercive because $\mathbf{b}(\boldsymbol{\mu}) \cdot \mathbf{n} \leq 0$ on $\Gamma_N \cap \{y = 0\}$. We approximate the stability factor (3.78) through the RBF interpolant of Sect. 3.7.3.

Then, we consider a G-RB and an LS-RB method, performing the construction of the RB space through the greedy algorithm for both. Note that the well-posedness of the G-RB problem is not automatically inherited from that of the high-fidelity problem if the problem is only weakly coercive. For the case at hand, condition (3.34) is fulfilled and the G-RB approximation proves to be stable. Instead, the LS-RB method always provides a well-posed problem.

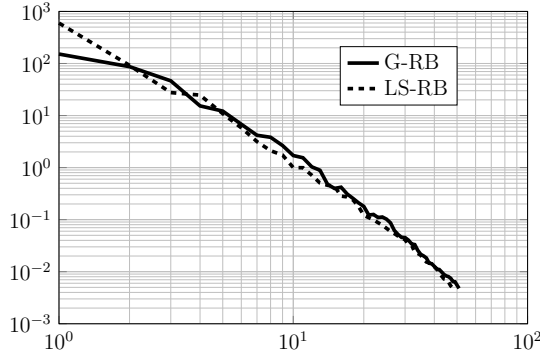


Fig. 9.1 Convergence of the greedy algorithm: maximum relative error bound $\Delta_N(\boldsymbol{\mu})/\|u_N(\boldsymbol{\mu})\|_V$ evaluated over a training sample $\mathcal{E}_{\text{train}}$ as a function of N , in both the G-RB and the LS-RB case

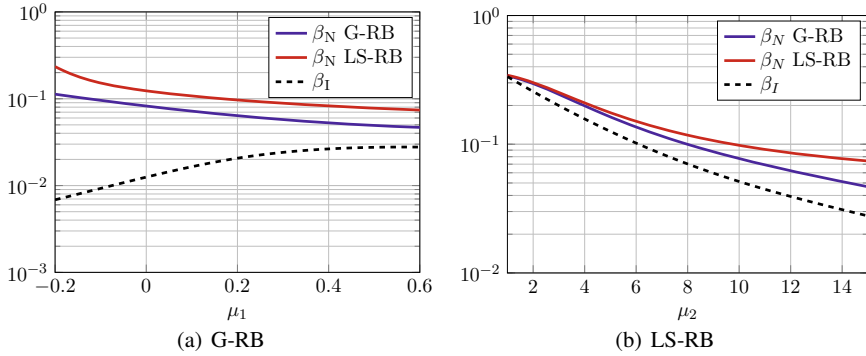


Fig. 9.2 Interpolant of the stability factor $\beta_I(\boldsymbol{\mu})$ as a function of μ_1 with $\mu_2 = 15$, $\mu_3 = 30$ (left) and as a function of μ_2 with $\mu_1 = 0.6$, $\mu_3 = 30$ (right)

By considering a training sample $\mathcal{E}_{\text{train}}$ of size $n_{\text{train}} = |\mathcal{E}_{\text{train}}| = 2000$ obtained by latin hypercube sampling, and prescribing a stopping tolerance $\varepsilon_g = 5 \cdot 10^{-3}$ on the relative error bound $\Delta_N(\boldsymbol{\mu})/\|u_N(\boldsymbol{\mu})\|_V$, we end up with a reduced space made by $N = 51$ (resp., $N = 48$) in the G-RB (resp. LS-RB) case, see Fig. 9.1. Running the greedy algorithm requires¹ 147 seconds in the former case and 222 seconds in the latter; the time gap is due to the larger amount of arrays required by the offline/online decomposition in the LS-RB case with respect to the G-RB case, see Sect. 3.4.2.

In Fig. 9.2 we plot the interpolant β_I of the high-fidelity stability factor, as well as the stability factors of both the G-RB and the LS-RB problem – defined by (3.34) and (3.38), respectively – as functions of μ_1 and μ_2 , keeping the other parameter components equal to their maximum values.

¹ Numerical results were obtained on a workstation with a Intel Core i5-2400S CPU and 16 GB of RAM. Parallelism is exploited to speed up some *embarrassingly* parallel portions of the algorithms we propose, such as the computation of the terms involved by the dual norm of the residual, the construction of the RBF interpolant for evaluating the stability factor, or the computation of POD snapshots. The reported computational times will mainly serve to compare the different strategies.

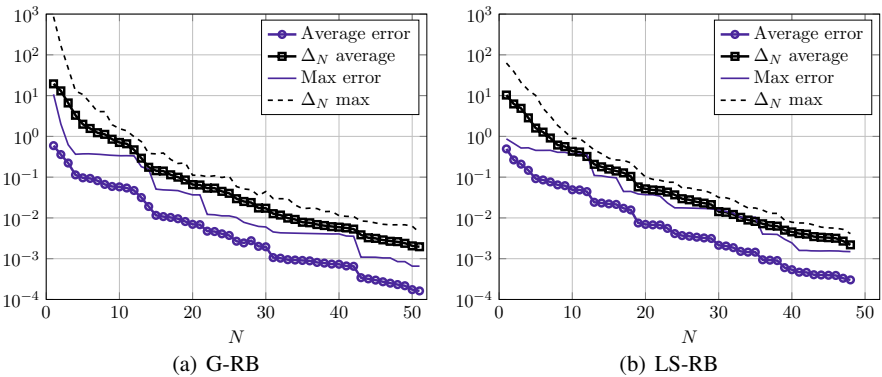


Fig. 9.3 Relative (average and maximum) error and bound over a test sample of 200 parameter values

The online evaluation time is about the same, ranging from 3 ms to 3.5 ms in the two cases; assembling the LS-RB problem requires Q_a^2 , $Q_f Q_a$ terms for the matrix $\mathbb{A}_N(\boldsymbol{\mu})$ and the vector $\mathbf{f}_N(\boldsymbol{\mu})$, respectively, against Q_a , Q_f terms in the G-RB case. The online convergence of the relative error between the RB and the high-fidelity approximation, evaluated over a test sample of 200 parameter values, is reported in Fig. 9.3, along with the corresponding error bounds.

The computational speedup with respect to the high-fidelity solver is of about 66 (resp. 57) in the G-RB (resp. LS-RB) case. This gain is primarily due to the fact that we need to solve a system of dimension $N \times N$ instead of one of dimension $N_h \times N_h$ in the high-fidelity case; here N_h/N ranges between 265 and 282. These comparative data however should not be regarded as carved on the stone. They obviously depend on the way the two codes (the high-fidelity and the RB’s) are actually implemented. Moreover, sometimes for programming convenience some cross-benefits may arise. For instance, for the problem at hand the offline CPU time also includes the construction of the RBF interpolant of $\beta_h(\boldsymbol{\mu})$, whereas the high-fidelity solution relies on high-fidelity arrays already assembled according to the affine decomposition. See Tab. 9.1 for further details. In Fig. 9.4 we display some RB solutions of the problem, corresponding to different parameter values.

Table 9.1 Computational details for the high-fidelity and reduced-order models of (8.15). The reported FE solution time is the one related to the affine model (without FE assembly)

High-fidelity model		G-RB greedy		LS-RB greedy	
FE dofs N_h	13 538	RB dofs	51	RB dofs	48
Q_a	6	Dofs reduction	265:1	Dofs reduction	282:1
Q_f	1	Greedy CPU time	147 s	Greedy CPU time	222 s
FE solution time	0.2 s	Online CPU time	3 ms	Online CPU time	3.5 ms

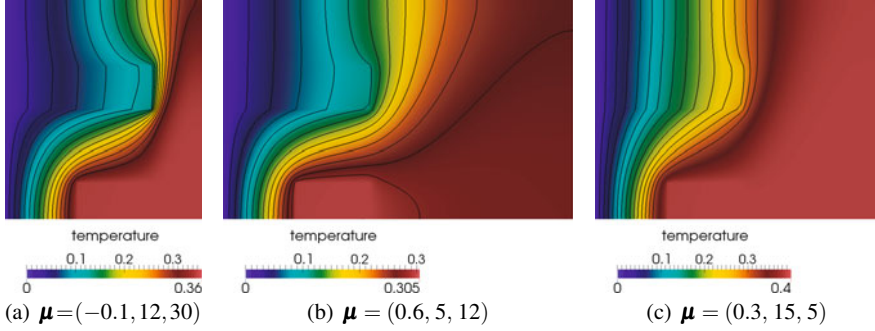


Fig. 9.4 Representative solutions of problem (8.15) for different parameter values

9.2 An Elastic Beam: Results

Let us now consider the RB approximation of a linear elasticity problem to model the displacement of an elastic beam under prescribed normal stresses (and zero displacement) on its boundary (see Sect. 8.6). We consider an elastic beam occupying the domain $\Omega = [0, 0.2] \times [0, 3] \times [0, 0.2] \subset \mathbb{R}^3$ and characterized by:

- the Young modulus $\mu_1 = E \in [10, 90]$ GPa;
- the Poisson coefficient $\mu_2 = \nu \in [0.25, 0.42]$;
- the intensity $\mu_3, \mu_4, \mu_5 \in [-0.4, 0.4]$ MPa of the normal stresses over $\Gamma_{N_1}, \Gamma_{N_2}, \Gamma_{N_3}$.

For the high-fidelity approximation we use \mathbb{P}_1 finite elements built over a discretization of the domain made by tetrahedral elements, resulting in 6091 vertices, 26467 tetrahedra and a high-fidelity space V_h of dimension $N_h = 18078$.

Concerning the evaluation of the stability factor, we compute an interpolatory approximation of $\beta_h(\mu)$ relying on radial basis functions on a cartesian grid of 5×4 points in the (μ_1, μ_2) plane; note that μ_3, μ_4 and μ_5 do not affect the problem matrix. Very few interpolation points in two (parameter) dimensions are enough to approximate the stability factor in an accurate way, above all when it shows a higher degree of regularity on the parameters. The result is reported in Fig. 9.5(a). The stability factor is required to evaluate the a posteriori error bound during the online stage, to check the accuracy of the RB approximation. Concerning the construction of the RB space, we run the POD algorithm: starting from $n_s = 100$ snapshots, we retain the first $N = 15$ POD modes, see Fig. 9.6. The online evaluation time is about 8 ms, whereas the high-fidelity FE solution (relying on high-fidelity arrays already assembled according to the affine decomposition of the problem) takes 0.5 s, thus yielding a computational speedup of about 60. See Tab. 9.2 for further details. The online convergence of the error between the RB and the high-fidelity approximation, evaluated over a test sample of 200 parameter values, is reported in Fig. 9.5(b).

Some RB solutions of the problem, corresponding to different parameter values, are displayed in Fig. 9.7.

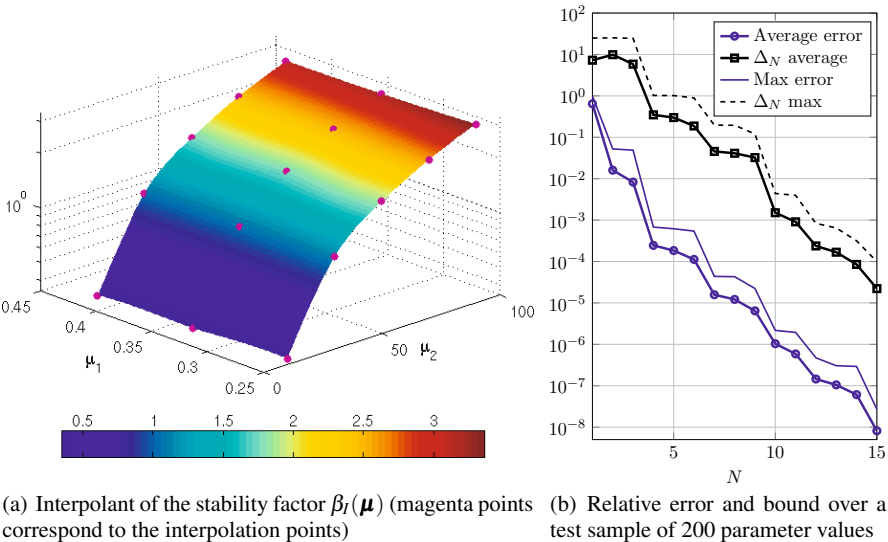


Fig. 9.5 Error estimate: stability factor and comparison between error and error bounds

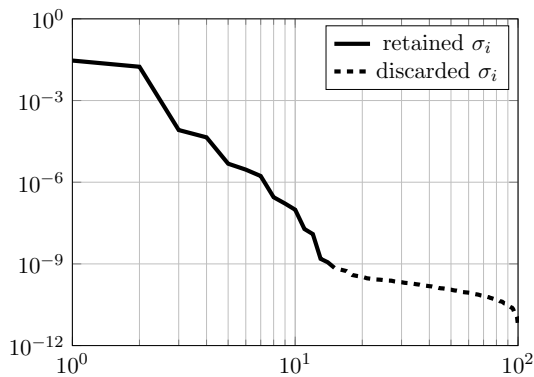


Fig. 9.6 Construction of the RB space through the POD algorithm: decay of the singular values of the snapshot matrix

Table 9.2 Computational details for the high-fidelity and reduced-order models of (8.49); the reported FE solution time is the one related to the affine model (without FE assembly)

High-fidelity model		Reduced-order model	
Number of FE dofs N_h	18 078	Number of RB dofs	15
Affine operator components Q_a	2	Dofs reduction	1205:1
Affine rhs components Q_f	3	Offline CPU time	≈ 68 s
FE solution time	≈ 0.5 s	Online CPU time	8 ms



Fig. 9.7 Representative solutions of problem (8.49) for different parameter values. In each case, we report the displaced (*red*) and initial (*gray*) configuration

9.3 Backward-Facing Step Channel, Stokes Flow: Results

In this section we show how to set a RB method for the approximation of a parametrized Stokes problem, such as the one describing a fluid flow in a backward-facing step channel, focusing on the well-posedness of the reduced problem and the possible choices of the reduced spaces. Since the setting of RB approximation for Stokes equations is more involved than in previous cases, we provide further details about its construction. This framework will also be exploited when dealing with the RB approximation of the Navier-Stokes equations in Chap. 11.

The weak formulation of a parametrized Stokes problem over Ω reads (see Sects. 8.7–8.8): find $(\mathbf{u}(\boldsymbol{\mu}), p(\boldsymbol{\mu})) \in X \times Q = [H_{T_d}^1(\Omega)]^d \times L^2(\Omega)$ such that

$$\begin{cases} d(\mathbf{u}(\boldsymbol{\mu}), \mathbf{v}; \boldsymbol{\mu}) + b(\mathbf{v}, p(\boldsymbol{\mu}); \boldsymbol{\mu}) = f_1(\mathbf{v}; \boldsymbol{\mu}) & \forall \mathbf{v} \in X \\ b(\mathbf{u}(\boldsymbol{\mu}), q; \boldsymbol{\mu}) = f_2(q; \boldsymbol{\mu}) & \forall q \in Q, \end{cases} \quad (9.1)$$

where $d(\cdot, \cdot; \boldsymbol{\mu}) : X \times X \rightarrow \mathbb{R}$ and $b(\cdot, \cdot; \boldsymbol{\mu}) : X \times Q \rightarrow \mathbb{R}$ have been defined in (8.59)–(8.60), respectively. Moreover, $f_1(\cdot; \boldsymbol{\mu}) : X \rightarrow \mathbb{R}$ and $f_2(\cdot; \boldsymbol{\mu}) : Q \rightarrow \mathbb{R}$ collect the contributions from both source terms and essential boundary conditions.

A high-fidelity approximation to problem (9.1) can be constructed by using a stable pair of finite elements spaces, such as Taylor-Hood \mathbb{P}_2 – \mathbb{P}_1 or \mathbb{P}_1^b – \mathbb{P}_1 elements; see, e.g., [104, 222] for further details. Here we denote by $X_h \subset X$, $Q_h \subset Q$ the corresponding velocity and pressure approximation spaces, and by $V_h = X_h \times Q_h$. We denote by $\{\boldsymbol{\phi}^j\}_{j=1}^{N_h^u}$ and by $\{\phi^j\}_{j=1}^{N_h^p}$ a basis for X_h and Q_h , respectively. Because of the vectorial nature of the problem we set $X_h = [Y_h]^d$, being $Y_h \subset H^1(\Omega)$. We also denote by $\mathbb{X}_{h,u}$, $\mathbb{X}_{h,p}$ the matrices associated to the scalar products in X , Q , respectively; see (2.40). The high-fidelity approximation yields the following $\boldsymbol{\mu}$ -dependent linear system (see (2.62))

$$\begin{pmatrix} \mathbb{D}_h(\boldsymbol{\mu}) & \mathbb{B}_h^T(\boldsymbol{\mu}) \\ \mathbb{B}_h(\boldsymbol{\mu}) & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_h(\boldsymbol{\mu}) \\ \mathbf{p}_h(\boldsymbol{\mu}) \end{pmatrix} = \begin{pmatrix} \mathbf{f}_{1h}(\boldsymbol{\mu}) \\ \mathbf{f}_{2h}(\boldsymbol{\mu}) \end{pmatrix}. \quad (9.2)$$

Here $u_h(\boldsymbol{\mu}) = \sum_{j=1}^{N_h} u_h^{(j)}(\boldsymbol{\mu}) \boldsymbol{\varphi}^j$, $p_h(\boldsymbol{\mu}) = \sum_{k=1}^{M_h} p_h^{(k)}(\boldsymbol{\mu}) \phi^k$ are the high-fidelity approximation of the velocity and the pressure fields, $\mathbf{u}_h(\boldsymbol{\mu})$ and $\mathbf{p}_h(\boldsymbol{\mu})$ are the vectors whose components are $u_h^{(j)}$, $p_h^{(k)}$,

$$(\mathbb{D}_h(\boldsymbol{\mu}))_{ij} = d(\boldsymbol{\varphi}^j, \boldsymbol{\varphi}^i; \boldsymbol{\mu}), \quad i, j = 1, \dots, N_h$$

$$(\mathbb{B}_h(\boldsymbol{\mu}))_{kj} = b(\boldsymbol{\varphi}^j, \phi^k; \boldsymbol{\mu}), \quad k = 1, \dots, M_h, j = 1, \dots, N_h$$

$$(\mathbf{f}_{1h}(\boldsymbol{\mu}))_{(i)} = f_1(\boldsymbol{\varphi}^i; \boldsymbol{\mu}), \quad (\mathbf{f}_{2h}(\boldsymbol{\mu}))_{(k)} = f_2(\phi^k; \boldsymbol{\mu}), \quad i = 1, \dots, N_h, k = 1, \dots, M_h.$$

In this case, the following discrete inf-sup condition is verified by the parametrized bilinear form $b(\cdot, \cdot; \boldsymbol{\mu})$, see (2.60):

$$\exists \beta_{0,h}^s > 0 : \beta_h^s(\boldsymbol{\mu}) = \inf_{q_h \in \mathcal{Q}_h} \sup_{\mathbf{w}_h \in X_h} \frac{b(\mathbf{w}_h, q_h; \boldsymbol{\mu})}{\|\mathbf{w}_h\|_X \|q_h\|_Q} \geq \beta_{0,h}^s \quad \forall \boldsymbol{\mu} \in \mathcal{P} \quad (9.3)$$

ensuring the existence and uniqueness of the solution of (9.2). An equivalent way to express the inf-sup stability factor hinges upon the introduction of a (pressure) supremizer operator $T_p^\mu : \mathcal{Q}_h \rightarrow X_h$, similarly to the one in (3.39), defined by

$$(T_p^\mu q, \mathbf{w})_X = b(\mathbf{w}, q; \boldsymbol{\mu}) \quad \forall \mathbf{w} \in X_h. \quad (9.4)$$

Similarly to what we have done in Sect. 3.4.2 – see equation 3.63 – for any $q_h \in \mathcal{Q}_h$, (9.4) turns into the linear system

$$\mathbb{X}_{h,\mathbf{u}} \mathbf{t}_h^\mu = \mathbb{B}_h^T(\boldsymbol{\mu}) \mathbf{q}_h \quad \forall \mathbf{q}_h \in \mathbb{R}^{N_h^p} \quad (9.5)$$

whose solution $\mathbf{t}_h^\mu = \mathbf{t}_h^\mu(\mathbf{q}_h) \in \mathbb{R}^{N_h^u}$ is the corresponding algebraic supremizer solution. Moreover, similarly to what we have done in Sect. 2.4.6, see (2.55), we have

$$\beta_h^s(\boldsymbol{\mu}) = \inf_{q \in \mathcal{Q}_h} \frac{\|T_p^\mu q\|_X}{\|q\|_Q}. \quad (9.6)$$

9.3.1 RB Approximation of Parametrized Stokes Equations

An RB approximation of (9.2) can be obtained by exploiting either a LS-RB method, or a G-RB method – this latter on a suitably chosen couple of RB spaces for velocity and pressure variables. In both cases we end up with a well-posed RB problem. Here we focus on the construction of a RB space through the greedy algorithm (see Sect. 7.1.2), denoting by $S_N = \{\boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^N\}$ the set of selected parameters.

We first consider the G-RB method, and define

$$\mathcal{Q}_N = \text{span}\{p_h(\boldsymbol{\mu}^n), n = 1, \dots, N\}, \quad (9.7)$$

$$X_N^\mu = \text{span}\{\mathbf{u}_h(\boldsymbol{\mu}^n), T_p^\mu p_h(\boldsymbol{\mu}^n), n = 1, \dots, N\}. \quad (9.8)$$

The G-RB problem reads: find $(\mathbf{u}_N(\boldsymbol{\mu}), p_N(\boldsymbol{\mu})) \in X_N^\mu \times Q_N$ such that:

$$\begin{cases} d(\mathbf{u}_N(\boldsymbol{\mu}), \mathbf{v}_N; \boldsymbol{\mu}) + b(\mathbf{v}_N, p_N(\boldsymbol{\mu}); \boldsymbol{\mu}) = f_1(\mathbf{v}_N; \boldsymbol{\mu}) & \forall \mathbf{v}_N \in X_N^\mu \\ b(\mathbf{u}_N(\boldsymbol{\mu}), q_N; \boldsymbol{\mu}) = f_2(q_N; \boldsymbol{\mu}) & \forall q_N \in Q_N. \end{cases} \quad (9.9)$$

This problem is well-posed. In particular, the following result holds (see e.g. [233, 230]).

Proposition 9.1. *For every $\boldsymbol{\mu} \in \mathcal{P}$ we define the inf-sup stability factor*

$$\beta_N^s(\boldsymbol{\mu}) = \inf_{q \in Q_N} \sup_{\mathbf{w} \in X_N^\mu} \frac{b(\mathbf{w}, q; \boldsymbol{\mu})}{\|\mathbf{w}\|_X \|q\|_Q}. \quad (9.10)$$

Then

$$\beta_N^s(\boldsymbol{\mu}) \geq \beta_h^s(\boldsymbol{\mu}) \geq \beta_{0,h}^s > 0 \quad \forall \boldsymbol{\mu} \in \mathcal{P}. \quad (9.11)$$

Proof. Thanks to (9.3), (9.10) and the definitions (9.7)–(9.8), we have

$$\begin{aligned} \beta_N^s(\boldsymbol{\mu}) &= \inf_{q \in Q_N} \sup_{\mathbf{w} \in X_N^\mu} \frac{b(\mathbf{w}, q; \boldsymbol{\mu})}{\|\mathbf{w}\|_X \|q\|_Q} \geq \inf_{q \in Q_N} \frac{b(\mathbf{w}, T_p^\mu q; \boldsymbol{\mu})}{\|q\|_Q} \\ &= \inf_{q \in Q_N} \sup_{\mathbf{w} \in X_h} \frac{b(\mathbf{w}, q; \boldsymbol{\mu})}{\|\mathbf{w}\|_X \|q\|_Q} \geq \inf_{q \in Q_h} \sup_{\mathbf{w} \in X_h} \frac{b(\mathbf{w}, q; \boldsymbol{\mu})}{\|\mathbf{w}\|_X \|q\|_Q} = \beta_h^s(\boldsymbol{\mu}). \end{aligned}$$

Then (9.11) follows thanks to (9.3). Note that $\beta_N^s(\boldsymbol{\mu})$ is therefore bounded from below by a stability constant independent of N and $\boldsymbol{\mu}$. \square

From an algebraic standpoint, the RB spaces (9.7)–(9.8) can be generated by two transformation matrices

$$\mathbb{V}_p \in \mathbb{R}^{N_h^p \times N}, \quad \mathbb{V}_u^\mu = [\tilde{\mathbb{V}}_u \quad \mathbb{V}_s^\mu] \in \mathbb{R}^{N_h^u \times 2N} \quad (9.12)$$

respectively, being

$$\mathbb{V}_s^\mu = \mathbb{X}_u^{-1} \mathbb{B}_h^T(\boldsymbol{\mu}) \mathbb{V}_p. \quad (9.13)$$

Problem (9.9) can be equivalently expressed as the following $\boldsymbol{\mu}$ -dependent linear system of dimension $3N \times 3N$ – note that $\dim(X_N^\mu) = 2N = 2\dim(Q_N)$:

$$\begin{pmatrix} \mathbb{D}_N(\boldsymbol{\mu}) & \mathbb{B}_N^T(\boldsymbol{\mu}) \\ \mathbb{B}_N(\boldsymbol{\mu}) & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_N(\boldsymbol{\mu}) \\ \mathbf{p}_N(\boldsymbol{\mu}) \end{pmatrix} = \begin{pmatrix} \mathbf{f}_{1N}(\boldsymbol{\mu}) \\ \mathbf{f}_{2N}(\boldsymbol{\mu}) \end{pmatrix} \quad (9.14)$$

where we can express the RB matrices as

$$\mathbb{D}_N(\boldsymbol{\mu}) = \begin{pmatrix} \tilde{\mathbb{V}}_u^T \mathbb{D}_h(\boldsymbol{\mu}) \tilde{\mathbb{V}}_u & (\mathbb{V}_s^\mu)^T \mathbb{X}_u^{-1} \mathbb{D}_h(\boldsymbol{\mu}) \tilde{\mathbb{V}}_u \\ \tilde{\mathbb{V}}_u^T \mathbb{D}_h(\boldsymbol{\mu}) \mathbb{V}_s^\mu & (\mathbb{V}_s^\mu)^T \mathbb{D}_h(\boldsymbol{\mu}) \mathbb{V}_s^\mu \end{pmatrix} \in \mathbb{R}^{2N \times 2N} \quad (9.15)$$

$$\mathbb{B}_N(\boldsymbol{\mu}) = (\mathbb{V}_p^T \mathbb{B}_h(\boldsymbol{\mu}) \tilde{\mathbb{V}}_u \quad \mathbb{V}_p^T \mathbb{B}_h(\boldsymbol{\mu}) \mathbb{V}_s^\mu) \in \mathbb{R}^{N \times 2N}. \quad (9.16)$$

Indeed, this can be regarded as a special case of (4.9), provided we identify $\mathbb{A}_h(\boldsymbol{\mu})$ and $\mathbb{A}_N(\boldsymbol{\mu})$ as the matrices appearing at the left hand sides of (9.2) and (9.14), respectively, and

$$\mathbb{V} = \begin{pmatrix} \mathbb{V}_u^{\boldsymbol{\mu}} & 0 \\ 0 & \mathbb{V}_p \end{pmatrix} \in \mathbb{R}^{(N_h^u + N_h^p) \times 3N}. \quad (9.17)$$

The RB velocity space is $\boldsymbol{\mu}$ -dependent because of the presence of the $\boldsymbol{\mu}$ -dependent supremizer operator; under the assumption of affine parametric dependence (3.59), RB arrays can be assembled through the usual offline/online decomposition, although this procedure features a moderately large computational cost. In fact, denoting by Q_d, Q_b the number of separable terms of $d(\cdot, \cdot; \boldsymbol{\mu})$ and $b(\cdot, \cdot; \boldsymbol{\mu})$, respectively, assembling the RB matrices \mathbb{D}_N and \mathbb{B}_N appearing in (9.14) requires the construction of $O(Q_d Q_b^2)$ matrices of size $2N \times 2N$ and $O(Q_b^2)$ matrices of size $2N \times N$, respectively.

This is indeed quite similar to the construction of the test space in the LS-RB method, see Sect. 3.4.2. However, the Galerkin projection over the $\boldsymbol{\mu}$ -dependent RB velocity space forces not only test but also trial functions to be $\boldsymbol{\mu}$ -dependent, thus yielding additional complexity and cost.

To overcome this drawback, a $\boldsymbol{\mu}$ -independent enrichment of the velocity space can be operated, introducing the velocity RB space [233, 230]

$$X_N = \text{span}\{\mathbf{u}_h(\boldsymbol{\mu}^n), T_p^{\boldsymbol{\mu}^n} p_h(\boldsymbol{\mu}^n), n = 1, \dots, N\} \quad (9.18)$$

instead of (9.8). Correspondingly, the velocity RB space is no longer $\boldsymbol{\mu}$ -dependent, and can be generated by a transformation matrix

$$\mathbb{V}_u = [\tilde{\mathbb{V}}_u \quad \mathbb{V}_s] \in \mathbb{R}^{N_h^u \times 2N} \quad (9.19)$$

being

$$\mathbb{V}_s = [\mathbb{X}_u^{-1} \mathbb{B}_h^T(\boldsymbol{\mu}^1) \mathbf{p}_h(\boldsymbol{\mu}^1) \mid \dots \mid \mathbb{X}_u^{-1} \mathbb{B}_h^T(\boldsymbol{\mu}^N) \mathbf{p}_h(\boldsymbol{\mu}^N)]. \quad (9.20)$$

Hence the assembling of the RB matrices \mathbb{D}_N and \mathbb{B}_N requires in this case the construction of $O(Q_d)$ matrices of size $2N \times 2N$ and $O(Q_b)$ matrices of size $2N \times N$, respectively. Note that the condition $\beta_N^s(\boldsymbol{\mu}) \geq \beta_{0,h}^s > 0$ for a constant $\beta_{0,h}^s$ independent of N and $\boldsymbol{\mu}$ (that would guarantee the fulfillment of the inf-sup condition for the velocity space (9.18)) does not follow directly from (9.3) anymore. Nevertheless, from the results that we report below there is numerical evidence that the corresponding solution is stable. See, e.g. [229, 230] for further details.

An LS-RB method can be introduced also in the case of Stokes equations, the derivation of the reduced problem being indeed very similar to that of the G-RB case described above, except for the construction of the RB spaces. In algebraic form, we define

$$\mathbb{V} = \begin{pmatrix} \tilde{\mathbb{V}}_u & 0 \\ 0 & \mathbb{V}_p \end{pmatrix} \in \mathbb{R}^{(N_h^u + N_h^p) \times 2N}$$

and obtain the test space as

$$\mathbb{W} = \mathbb{X}_h^{-1} \mathbb{A}_h(\boldsymbol{\mu}) \mathbb{V} \in \mathbb{R}^{(N_h^u + N_h^p) \times 2N}.$$

Note that the velocity space has dimension N in this case. The LS-RB problem is then obtained following the same procedure of Sect. 3.4.2; the trial space is $\boldsymbol{\mu}$ -dependent, because of the $\boldsymbol{\mu}$ -dependence in $\mathbb{A}_h(\boldsymbol{\mu})$, thus implying a larger number of arrays to store to handle with the offline/online decomposition.

Remark 9.1. If the high-fidelity problem is approximated by a couple of unstable spaces (like $\mathbb{P}_1 - \mathbb{P}_1$ finite elements) and then stabilized by suitable pressure or residual-based stabilizations (see, e.g., [222, 104]), then a G-RB approximation can be set up without the need of supremizer enrichment (see [87] and Sect. 11.8 for some numerical results). •

9.3.2 A Posteriori Error Estimation

In the Stokes case an a posteriori error bound directly follows from the general framework introduced in Sect. 3.6. Indeed, by setting $V = X \times Q$,

$$a((\mathbf{u}, p), (\mathbf{w}, q); \boldsymbol{\mu}) = d(\mathbf{u}; \mathbf{w}; \boldsymbol{\mu}) + b(\mathbf{w}, p; \boldsymbol{\mu}) + b(\mathbf{u}, q; \boldsymbol{\mu})$$

and $f((\mathbf{w}, q); \boldsymbol{\mu}) = f_1(\mathbf{w}; \boldsymbol{\mu}) + f_2(q; \boldsymbol{\mu})$, by invoking the result in (3.71) we obtain

$$(\|\mathbf{u}_h(\boldsymbol{\mu}) - \mathbf{u}_N(\boldsymbol{\mu})\|_X^2 + \|p_h(\boldsymbol{\mu}) - p_N(\boldsymbol{\mu})\|_Q^2)^{1/2} \leq \Delta_N(\boldsymbol{\mu}) = \frac{\|r(\cdot; \boldsymbol{\mu})\|_{V'_h}}{\beta_h(\boldsymbol{\mu})}.$$

Here the residual is given by

$$r((\mathbf{w}, q); \boldsymbol{\mu}) = f((\mathbf{w}, q); \boldsymbol{\mu}) - a((\mathbf{u}, p), (\mathbf{w}, q); \boldsymbol{\mu}) = r_u(\mathbf{w}; \boldsymbol{\mu}) + r_p(q; \boldsymbol{\mu}),$$

being

$$r_u((\mathbf{w}, q); \boldsymbol{\mu}) = f_1(\mathbf{w}; \boldsymbol{\mu}) - d(\mathbf{u}; \mathbf{w}; \boldsymbol{\mu}) - b(\mathbf{w}, p; \boldsymbol{\mu}), \quad r_p(q; \boldsymbol{\mu}) = f_2(q; \boldsymbol{\mu}) - b(\mathbf{u}, q; \boldsymbol{\mu}),$$

whence $\|r(\cdot; \boldsymbol{\mu})\|_{V'_h} = \left(\|r_u(\cdot; \boldsymbol{\mu})\|_{X'_h}^2 + \|r_p(\cdot; \boldsymbol{\mu})\|_{Q'_h}^2 \right)^{1/2}$. The inf-sup stability factor $\beta_h(\boldsymbol{\mu})$ is the one of $a(\cdot, \cdot; \boldsymbol{\mu})$ and is defined in (3.13). See, e.g. [230] for further explanations about the RB approximation of parametrized Stokes equations, as well for numerical test cases. An alternative procedure to obtain a posteriori error estimates for velocity and pressure variables separately is presented in [112, 113].

9.3.3 Numerical Results: Backward-Facing Step Channel

We now come back to the backward-facing step channel problem we introduced in Sect. 8.7. Here we consider $P = 3$ parameters:

- $\mu_1 \in [20, 50]$, being $\nu = 1/\mu_1$ the viscosity of the fluid;
- $\mu_2 = L \in [9, 12]$ is the length of the channel, see Fig. 8.7 (rather than the inlet velocity profile, as in (8.53));
- $\mu_3 \in [0.3, 2]$ is the step height, see Fig. 8.7.

For the high-fidelity approximation we use $\mathbb{P}_2 - \mathbb{P}_1$ finite elements built over a discretization of the domain made by triangular elements, resulting in 5930 vertices, 11493 triangles and a high-fidelity space V_h of dimension $N_h = 51\,268$.

By considering a training sample $\mathcal{E}_{\text{train}}$ of size $n_{\text{train}} = 5000$ obtained by latin hypercube sampling, and a stopping tolerance $\varepsilon_g = 10^{-4}$ on the relative error bound $\Delta_N(\boldsymbol{\mu})/\|u_N(\boldsymbol{\mu})\|_V$, we end up with a reduced space made by $N = 26$ pressure basis functions and $2N = 52$ velocity basis functions in the G-RB case – see (9.18) – thus yielding $3N = 78$ degrees of freedom for the G-RB problem. The offline stage in this case requires a CPU time of 1169 s. In the LS-RB case, through the greedy algorithm (with the same training sample and the same tolerance) we select $N = 27$ pressure basis functions and $N = 27$ velocity basis functions, thus yielding $2N = 54$ degrees of freedom for the LS-RB problem.

The online convergence of the error between the G-RB (resp. LS-RB) and the high-fidelity approximation, evaluated over a test sample of 400 parameter values, is reported in Fig. 9.8.

The online evaluation time is about 1.25 ms in the G-RB case and 5.81 ms in the LS-RB case, whereas the high-fidelity FE solution (relying on high-fidelity arrays already assembled according to the affine decomposition of the problem) takes about 1 s, thus yielding a computational speedup of about 800 (resp. 170) in the G-RB (resp. LS-RB) case.

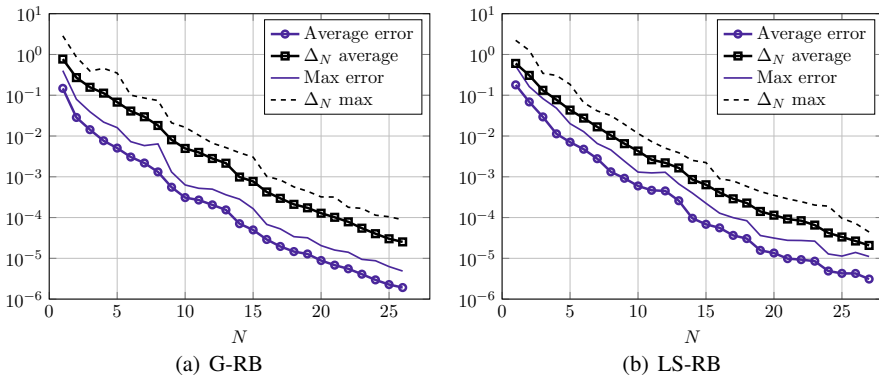


Fig. 9.8 Relative (average and maximum) error and bound over a test sample of 400 parameter values

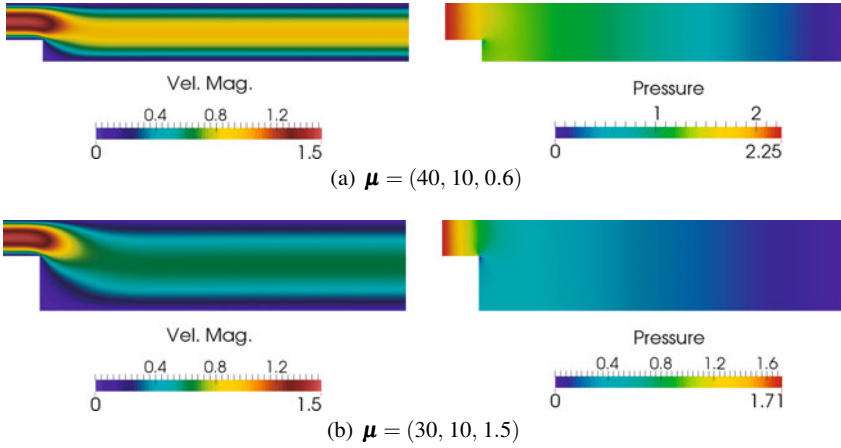


Fig. 9.9 Velocity magnitude and pressure fields obtained by solving the G-RB problem for different values of the parameters

Table 9.3 Computational details for the high-fidelity and reduced-order models for the Stokes problem

High-fidelity model		G-RB greedy		LS-RB greedy	
FE dofs N_h	51 268	RB dofs	78	RB dofs	54
$Q_d + Q_b$	6	Dofs reduction	657:1	Dofs reduction	949:1
$Q_{f1} + Q_{f2}$	6	Greedy CPU time	1 169 s	Greedy CPU time	1 150 s
FE solution time	≈ 1 s	Online CPU time	1.25 ms	Online CPU time	5.81 ms

The higher CPU time required to solve online the LS-RB problem with respect to the G-RB one is due to the larger (order of $(Q_d + Q_b)^2$ vs. $Q_d + Q_b$) number of terms to be summed to assemble the reduced operators. Evaluating the a posteriori error bound at the online stage, for any parameter value, takes 9 ms in the G-RB case and 7 ms in the LS-RB case: the lower estimation time in the LS-RB case is motivated by the fact that evaluating the error bound only depends on the number of RB degrees of freedom, and is independent of the left projection. We point out that in both cases the evaluation of the error bound is more expensive than solving the RB problem itself: this underlines the crucial need to deal with efficiently computable error bounds once again. See Tab. 9.3 for further details. Finally, we show in Fig. 9.9 the RB approximation of the solution obtained for different parameter values.

The examples we have presented in this chapter are just a few *instances* of a long list of linear parametrized PDEs to which the RB method has been successfully applied. Among others, we mention in particular frequency-domain wave phenomena relevant to acoustics (such as the Helmholtz problem [241, 240, 142]), electromagnetics problems (such as the Maxwell equations [61, 62, 107]) and problems in structural mechanics [141].

Chapter 10

Extension to Nonaffine Problems

We explain how to set up a RB method for problems not fulfilling the assumption of affine parametric dependence. Since the possibility to devise an offline/online decomposition relies on that assumption, in case of nonaffine problems we recover an approximate affine expansion by means of the so-called empirical interpolation method (EIM). We provide a detailed description of the EIM, focusing on linear problems for the sake of simplicity. A possible alternative formulation, referred to as discrete empirical interpolation method (DEIM), is also presented. As shown in the following chapter, EIM is an essential tool to ensure an offline/online decomposition, under suitable assumptions, also for nonlinear parametrized PDEs.

10.1 Empirical Interpolation Method

The affine parametric dependence assumption (3.52)–(3.53) is crucial to implement efficient RB methods enabling an offline/online decomposition for both the generation of the RB space (see Sect. 3.5) and the derivation of a posteriori error estimates (see Sect. 3.7). However, often this assumption is not automatically fulfilled by the $\boldsymbol{\mu}$ -dependent linear and bilinear forms. This is e.g. the case of the linear functional (8.35) associated to the right-hand side of problem (8.32). By approximating the function $s(\mathbf{x}; \boldsymbol{\mu})$ with $s_M(\mathbf{x}; \boldsymbol{\mu})$ as

$$s(\mathbf{x}; \boldsymbol{\mu}) = s_M(\mathbf{x}; \boldsymbol{\mu}) + e_{EIM}(\mathbf{x}; \boldsymbol{\mu}) = \sum_{m=1}^M \gamma_m(\boldsymbol{\mu}) \rho_m(\mathbf{x}) + e_{EIM}(\mathbf{x}; \boldsymbol{\mu}),$$

and requiring that, for a given tolerance ε ,

$$\varepsilon_M(\boldsymbol{\mu}) = \|e_{EIM}(\cdot; \boldsymbol{\mu})\|_{L^\infty(\Omega)} \leq \varepsilon \quad \forall \boldsymbol{\mu} \in \mathcal{P},$$

(8.35) can be approximated by

$$f_M(v; \boldsymbol{\mu}) = \int_{\Omega} s_M(\mathbf{x}; \boldsymbol{\mu}) v d\Omega = \sum_{m=1}^M \gamma_m(\boldsymbol{\mu}) \int_{\Omega} \rho_m(\mathbf{x}) v d\Omega. \quad (10.1)$$

This can be cast in the general affine expansion (3.53) provided we set $Q_f = M$ and

$$\Theta_f^m(\boldsymbol{\mu}) = \gamma_m(\boldsymbol{\mu}), \quad f^m(v) = \int_{\Omega} \rho_m(\mathbf{x}) v d\Omega, \quad 1 \leq m \leq M.$$

Implementing an efficient interpolation technique for $\boldsymbol{\mu}$ -dependent functions like $s(\mathbf{x}; \boldsymbol{\mu})$ and analyzing the impact of this further approximation on the RB methods is one of the goals of this chapter.

10.1.1 Polynomial Interpolation vs. Empirical Interpolation

Lagrangian interpolation is a classical approach to approximate general functions on a domain by requiring that the approximation is exact in a finite set of (interpolation) points. The Lagrangian interpolant is a linear combination of pre-selected, linearly independent basis functions, such as polynomial functions – see Sect. A.8. Choosing good interpolation points in an arbitrary domain $\Omega \subset \mathbb{R}^d$ is in general a critical task.

For a family of parameter-dependent functions $\mathcal{G} = \{g(\cdot; \boldsymbol{\mu}), \boldsymbol{\mu} \in \mathcal{P}\} \subset C^0(\overline{\Omega})$ we could in principle use the *collocation method* exploited in Sect. 5.5.2 to build a family of approximants

$$g_M(\mathbf{x}; \boldsymbol{\mu}) = \mathcal{J}_M^{\boldsymbol{\mu}} g(\mathbf{x}; \boldsymbol{\mu}) = \sum_{m=1}^M g(\mathbf{x}; \boldsymbol{\mu}^m) l_m(\boldsymbol{\mu}).$$

Here we denote by the superscript $\boldsymbol{\mu}$ the fact that the interpolation is performed with respect to $\boldsymbol{\mu}$, being $M \geq 1$ a given integer, $\{\boldsymbol{\mu}^m, m = 1, \dots, M\}$ a set of (e.g. Gauss-Lobatto) points in the parameter space $\mathcal{P} \subset \mathbb{R}^P$, also referred to as *collocation points*, $\{l_m(\boldsymbol{\mu}), m = 1, \dots, M\}$ the corresponding characteristic Lagrange polynomials and $\{g(\mathbf{x}; \boldsymbol{\mu}^m), m = 1, \dots, M\}$ a set of coefficients. This is the idea behind the *stochastic collocation method* for PDEs with random inputs: if $\boldsymbol{\mu} \in \mathcal{P}$ is a random input, the stochastic collocation method [20, 262] relies on a Lagrange interpolation of a set of deterministic PDE solutions evaluated at the collocation points.

Alternatively, we could build an approximation

$$g_M(\mathbf{x}; \boldsymbol{\mu}) = \mathcal{J}_M^{\mathbf{x}} g(\mathbf{x}; \boldsymbol{\mu}) = \sum_{m=1}^M g(\mathbf{x}^m; \boldsymbol{\mu}) l_m(\mathbf{x})$$

from an opposite viewpoint, by relying on a set of (e.g. Gauss-Lobatto) points $\{\mathbf{x}^m, m = 1, \dots, M\}$ in the spatial domain Ω . In this case, the superscript \mathbf{x} indicates the fact that the interpolation is performed with respect to \mathbf{x} .

Both approaches suffer from severe limitations: although Gauss-Lobatto interpolation enables to achieve rapid rates of convergence, it is only defined on intervals, rectangles and, more generally, domains that are tensor products of one-dimensional intervals, making therefore its use on general domains somehow impracticable. On the other hand, dealing with the first option can become infeasible because of the curse of dimensionality: usually $P \geq d$ (the spatial dimension) and the use of tensor grids become very soon impracticable as soon as $P = O(10)$. In that case the use of *sparse grids* becomes mandatory. Last, but not least, in both cases the interpolating polynomial would be constructed by relying on pre-defined set of basis functions and without exploiting the *joint* dependence of f on \mathbf{x} and $\boldsymbol{\mu}$.

The *empirical interpolation method* (EIM) relies instead on the use of basis functions built by sampling f at a suitably selected set of points in \mathcal{P} , instead than using predefined basis functions. First introduced in [23] and initially designed for treating nonaffine problems in the RB context – see, e.g. [120] – the empirical interpolation method has a broader scope, see e.g. the analysis reported in [172] related to the approximation properties for sets with small Kolmogorov n -width. Moreover, interpolation points (where the approximation is required to match the function being interpolated) are adaptively, iteratively added to the set without having to recompute all the existing points, so that EIM has a hierarchical nature. Furthermore, EIM achieves exponential convergence rate for analytical functions, as it happens for the Chebyshev or Legendre interpolants, and it is applicable to domains $\Omega \subset \mathbb{R}^d$ of arbitrary shape.

10.1.2 Empirical Interpolation

The purpose of EIM is to find approximations to elements of \mathcal{G} through an operator $\mathcal{I}_M^{\mathbf{x}}$ that interpolates the function $g(\cdot; \boldsymbol{\mu})$ at some carefully selected points in Ω .

Given an interpolatory system defined by a set of basis functions $\{\rho_1, \dots, \rho_M\}$ (linear combination of particular snapshots $g(\cdot; \boldsymbol{\mu}_{EIM}^1), \dots, g(\cdot; \boldsymbol{\mu}_{EIM}^M)$) and interpolation points $T_M = \{\mathbf{t}^1, \dots, \mathbf{t}^M\} \subset \overline{\Omega}$ – commonly referred to as *magic points* – the interpolant $\mathcal{I}_M^{\mathbf{x}} g(\cdot; \boldsymbol{\mu})$ of $g(\cdot; \boldsymbol{\mu})$ with $\boldsymbol{\mu} \in \mathcal{P}$ admits the separable expansion

$$\mathcal{I}_M^{\mathbf{x}} g(\mathbf{x}; \boldsymbol{\mu}) = \sum_{j=1}^M \gamma_j(\boldsymbol{\mu}) \rho_j(\mathbf{x}), \quad \mathbf{x} \in \Omega \quad (10.2)$$

and satisfies the M interpolation constraints

$$\mathcal{I}_M^{\mathbf{x}} g(\mathbf{t}^i; \boldsymbol{\mu}) = g(\mathbf{t}^i; \boldsymbol{\mu}), \quad i = 1, \dots, M. \quad (10.3)$$

Indeed, (10.3) yields the following linear system to solve

$$\sum_{j=1}^M \rho_j(\mathbf{t}^i) \gamma_j(\boldsymbol{\mu}) = g(\mathbf{t}^i; \boldsymbol{\mu}), \quad i = 1, \dots, M$$

that is, in matrix form

$$\mathbb{B}_M \boldsymbol{\gamma}(\boldsymbol{\mu}) = \mathbf{g}_M(\boldsymbol{\mu}) \quad \forall \boldsymbol{\mu} \in \mathcal{P} \quad (10.4)$$

where

$$(\mathbb{B}_M)_{ij} = \rho_j(\mathbf{t}^i), \quad (\boldsymbol{\gamma}(\boldsymbol{\mu}))_j = \gamma_j(\boldsymbol{\mu}), \quad (\mathbf{g}_M(\boldsymbol{\mu}))_i = g(\mathbf{t}^i; \boldsymbol{\mu}), \quad i, j = 1, \dots, M.$$

Conditions ensuring that \mathbb{B}_M is invertible will be given below.

10.1.3 EIM Algorithm

The construction of the basis functions yielding the approximation space $X_M = \text{span}\{\rho_1, \dots, \rho_M\}$ and interpolation points $T_M = \{\mathbf{t}^1, \dots, \mathbf{t}^M\}$ is based on a greedy algorithm [172] – indeed very similar in spirit to the one of Sect. 7.1.1. This procedure provides also a sample of parameter points $S_M = \{\boldsymbol{\mu}_{EIM}^1, \dots, \boldsymbol{\mu}_{EIM}^M\}$, needed to construct the basis functions $\rho_i(\mathbf{x})$, $i = 1, \dots, M$. To start, let us choose our first sample point as

$$\boldsymbol{\mu}_{EIM}^1 = \arg \max_{\boldsymbol{\mu} \in \mathcal{P}} \|g(\cdot; \boldsymbol{\mu})\|_{L^\infty(\Omega)},$$

define $S_1 = \{\boldsymbol{\mu}_{EIM}^1\}$ and the first generating function as

$$\xi_1(\mathbf{x}) = g(\mathbf{x}; \boldsymbol{\mu}_{EIM}^1).$$

Concerning the interpolation nodes, we first set

$$\mathbf{t}^1 = \arg \max_{\mathbf{x} \in \Omega} |\xi_1(\mathbf{x})|, \quad T_1 = \{\mathbf{t}^1\};$$

then, we define the first basis function as

$$\rho_1(\mathbf{x}) = \xi_1(\mathbf{x}) / \xi_1(\mathbf{t}^1),$$

and set $X_1 = \text{span}\{\rho_1\}$. Finally, we set the initial interpolation matrix

$$(\mathbb{B}_M)_{11} = \rho_1(\mathbf{t}^1) = 1.$$

At this stage, the available information allows to define the interpolant as the only function colinear with ρ_1 that coincides with g at \mathbf{t}^1 , that is $\mathcal{S}_1^{\mathbf{x}} g(\mathbf{x}; \boldsymbol{\mu}) = g(\mathbf{t}^1; \boldsymbol{\mu}) \rho_1(\mathbf{x})$. Note that the first interpolation point \mathbf{t}^1 is the point where the first basis function attains its maximum.

At the m -th step, $m = 1, \dots, M-1$, given the (nested) set $T_m = \{\mathbf{t}^1, \dots, \mathbf{t}^m\}$ of interpolation points and the set $\{\rho_1, \dots, \rho_m\}$ of basis functions, we select as $(m+1)$ -th generating function the snapshot which is the worst approximated by the current interpolant.

In other words, we select the snapshot which maximizes the error between g and $\mathcal{J}_m^{\mathbf{x}}g$,

$$\boldsymbol{\mu}_{EIM}^{m+1} = \arg \max_{\boldsymbol{\mu} \in \mathcal{D}} \|g(\cdot; \boldsymbol{\mu}) - \mathcal{J}_m^{\mathbf{x}}g(\cdot; \boldsymbol{\mu})\|_{L^\infty(\Omega)}, \quad (10.5)$$

$$\xi_{m+1}(\mathbf{x}) = g(\mathbf{x}; \boldsymbol{\mu}_{EIM}^{m+1}).$$

Then, we set $S_{m+1} = S_m \cup \{\boldsymbol{\mu}_{EIM}^{m+1}\}$.

To choose the $(m+1)$ -th interpolation point, we first evaluate the residual

$$r_{m+1}(\mathbf{x}) = \xi_{m+1}(\mathbf{x}) - \mathcal{J}_m^{\mathbf{x}}\xi_{m+1}(\mathbf{x})$$

by solving the linear system

$$\sum_{j=1}^m \rho_j(\mathbf{t}^i) \gamma_j = \xi_{m+1}(\mathbf{t}^i), \quad i = 1, \dots, m$$

to characterize the interpolant $\mathcal{J}_m^{\mathbf{x}}\xi_{m+1}$; then, we set

$$\mathbf{t}^{m+1} = \arg \max_{\mathbf{x} \in \bar{\Omega}} |r_{m+1}(\mathbf{x})| \quad (10.6)$$

that is, that point of Ω where ξ_{m+1} is worst approximated. Finally, we define the new basis function as

$$\rho_{m+1}(\mathbf{x}) = \frac{\xi_{m+1}(\mathbf{x}) - \mathcal{J}_m^{\mathbf{x}}\xi_{m+1}(\mathbf{x})}{\xi_{m+1}(\mathbf{t}^{m+1}) - \mathcal{J}_m^{\mathbf{x}}\xi_{m+1}(\mathbf{t}^{m+1})} = \frac{r_{m+1}(\mathbf{x})}{r_{m+1}(\mathbf{t}^{m+1})} \quad (10.7)$$

and we set $X_{m+1} = \text{span}\{\rho_i, i = 1, \dots, m+1\}$. The whole procedure is performed until a given tolerance ε_{EIM} is reached, or a given number M_{\max} of terms is computed; see Algorithm 10.1.

Algorithm 10.1 Empirical interpolation method (continuous version)

Input: max number of iterations M_{\max} , tolerance ε

Output: basis functions $\{\rho_1(\mathbf{x}), \dots, \rho_M(\mathbf{x})\}$, interpolation points $\{\mathbf{t}^1, \dots, \mathbf{t}^M\}$

- 1: $M = 0$, $e_0 = \varepsilon + 1$, $\mathcal{J}_0^{\mathbf{x}}g(\mathbf{x}; \boldsymbol{\mu}) = 0$,
 - 2: $\boldsymbol{\mu}^1 = \arg \max_{\boldsymbol{\mu} \in \mathcal{D}} \|g(\cdot; \boldsymbol{\mu})\|_{L^\infty(\Omega)}$
 - 3: **while** $M < M_{\max}$ and $e_M > \varepsilon$
 - 4: $M \leftarrow M + 1$
 - 5: $r(\mathbf{x}) = g(\mathbf{x}; \boldsymbol{\mu}^M) - \mathcal{J}_{M-1}^{\mathbf{x}}g(\mathbf{x}; \boldsymbol{\mu}^M)$
 - 6: $\mathbf{t}^M = \arg \max_{\mathbf{x} \in \bar{\Omega}} |r(\mathbf{x})|$
 - 7: $\rho_M(\mathbf{x}) = r(\mathbf{x})/r(\mathbf{t}^M)$
 - 8: $[e_M, \boldsymbol{\mu}^{M+1}] = \arg \max_{\boldsymbol{\mu} \in \mathcal{D}} \|g(\cdot; \boldsymbol{\mu}) - \mathcal{J}_M^{\mathbf{x}}g(\cdot; \boldsymbol{\mu})\|_{L^\infty(\Omega)}$
 - 9: **end while**
-

EIM yields a sequence of hierarchical spaces $X_1 \subset X_2 \subset \dots X_M$, such that the interpolation is exact for any $v \in X_M$ – that is,

$$\mathcal{J}_M^{\mathbf{x}}v = v \quad \forall v \in X_M$$

provided that $\dim(X_M) = M$ and that the matrix $\mathbb{B}_M \in \mathbb{R}^{M \times M}$ is invertible (see Exercise 1 for the proof). Moreover, we can show that the construction discussed so far yields indeed a set $\{\rho_1, \dots, \rho_M\}$ of linearly independent basis functions.

Theorem 10.1. *The construction of the interpolation points is well-defined and, for any $M \leq M_{\max} < \dim(\text{span}\{\mathcal{G}\})$, $X_M = \text{span}\{\rho_1, \dots, \rho_M\} = \text{span}\{\xi_1, \dots, \xi_M\}$ is of dimension M . In addition, \mathbb{B}_M is lower triangular with $(\mathbb{B}_M)_{ii} = 1$, $i = 1, \dots, M$.*

Proof. The property $\text{span}\{\rho_1, \dots, \rho_M\} = \text{span}\{\xi_1, \dots, \xi_M\} = X_M$ directly follows from the construction of the normalized ρ_i 's with respect to the ξ_j 's. Then, we proceed by induction. Clearly, $X_1 = \text{span}\{\rho_1\}$ has dimension 1 and $\mathbb{B}_1 = 1$ is invertible. Next, let us assume that $X_{M-1} = \text{span}\{\rho_1, \dots, \rho_{M-1}\}$ is of dimension $M-1$. If (i) \mathbb{B}_{M-1} is invertible and (ii) $|r_M(\mathbf{t}^M)| > 0$, we may form $X_M = \text{span}\{\rho_1, \dots, \rho_M\}$.

To prove that \mathbb{B}_M is invertible, it is enough to observe that

$$(\mathbb{B}_{M-1})_{ij} = \rho_j(\mathbf{t}^i) = r_j(\mathbf{t}^i)/r_j(\mathbf{x}^j), \quad i, j = 1, \dots, M-1. \quad (10.8)$$

Then $(\mathbb{B}_{M-1})_{ij} = 0$ if $i < j$, $(\mathbb{B}_{M-1})_{ij} = 1$ if $i = j$, whereas $|(\mathbb{B}_M)_{ij}| \leq 1$ if $i > j$ since $\mathbf{x}^j = \arg \max_{\mathbf{x} \in \overline{\Omega}} |r_j(\mathbf{x})|$, $j = 1, \dots, M$, according to (10.6). The matrix \mathbb{B}_M is lower triangular and it is such that $(\mathbb{B}_M)_{ii} = 1$, $i = 1, \dots, M$, hence it is invertible.

To prove that $\dim(X_M) = M$ (whence $\mathbf{t}^1, \dots, \mathbf{t}^M$ are distinct) we observe that

$$\begin{aligned} \|r_M\|_{L^\infty(\Omega)} &= \|g(\cdot; \boldsymbol{\mu}_{EIM}^M) - \mathcal{J}_{M-1}^{\mathbf{x}} g(\cdot; \boldsymbol{\mu}_{EIM}^M)\|_{L^\infty(\Omega)} \\ &\geq \max_{\boldsymbol{\mu} \in \mathcal{P}} \|g(\cdot; \boldsymbol{\mu}) - \mathcal{J}_{M-1}^{\mathbf{x}} g(\cdot; \boldsymbol{\mu})\|_{L^\infty(\Omega)} \geq d_{M_{\max}}(\mathcal{G}; X), \end{aligned}$$

with

$$d_{M_{\max}}(\mathcal{G}; X) = \inf_{\substack{\hat{X} \subset X \\ \dim(\hat{X}) = M_{\max}}} \sup_{\boldsymbol{\mu} \in \mathcal{P}} \inf_{z \in \hat{X}} \|g(\cdot; \boldsymbol{\mu}) - z\|_{L^\infty(\Omega)} > 0$$

since $M_{\max} < \dim(\text{span}\{\mathcal{G}\})$. If $\dim(X_M) \neq M$, we have that $\xi_M \in X_{M-1}$ and thus $\|r_M\|_{L^\infty(\Omega)} = 0$, which provides the contradiction and yields $\dim(X_M) = M$. \square

Note that if $\dim(\text{span}\{\mathcal{G}\}) = M^*$, the algorithm stops after $M = M^*$ iterations. As long as $M \leq M^*$, the previous theorem ensures that the basis functions $\{\rho_1, \dots, \rho_M\}$ and the snapshots $\{\xi_1, \dots, \xi_M\}$ span the same space X_M . In particular, it is better to deal with the former since, as resulting from (10.8),

$$\rho_i(\mathbf{t}^i) = 1, \quad i = 1, \dots, M, \quad \rho_j(\mathbf{t}^i) = 0, \quad 1 \leq i < j \leq M.$$

Remark 10.1. EIM is a very general technique, which can be employed in other contexts than nonaffine PDEs, see e.g. [172] for a detailed presentation. For instance, the algorithm can be exploited to generate a set of interpolation points, only, for a given, preexisting family of interpolating functions $\{\xi_1, \dots, \xi_M\}$. In this respect, either a POD strategy for an earlier selection of those functions, or a set featuring a canonical basis and ordering – such as monomials or Legendre polynomials – could be used [172].

In these cases, EIM enables to compute (a nested set of) interpolation points and a corresponding ordering of the basis functions. Rather than approximating a given function, its goal is to devise a generic interpolation scheme based on a set of magic points, adapted to any domain $\Omega \subset \mathbb{R}^d$. •

10.2 Error Analysis for the Empirical Interpolation

In this section we carry out an error analysis of the EIM based on the results of [172, 120]. Let us first introduce a set of characteristic (Lagrangian) functions $\{l_i^M \in X_M\}$ – similarly to the characteristic polynomials (A.25) in the case of Lagrange interpolation – to facilitate the construction of the interpolation operator $\mathcal{J}_M^{\mathbf{x}}$ in X_M over the set of magic points T_M .

For any given M , we can express

$$\mathcal{J}_M^{\mathbf{x}} g(\mathbf{x}; \boldsymbol{\mu}) = \sum_{i=1}^M g(\mathbf{t}^i; \boldsymbol{\mu}) l_i^M(\mathbf{x}), \quad l_i^M(\mathbf{x}) = \sum_{j=1}^M \rho_j(\mathbf{x}) (\mathbb{B}_M^{-1})_{ji}; \quad (10.9)$$

by definition, $l_i^M(\mathbf{x}^j) = \delta_{ij}$, $i, j = 1, \dots, M$. The existence of characteristic functions directly follows from the nonsingularity of the matrix \mathbb{B}_M (see Exercise 2). The a priori error analysis of the EIM involves the Lebesgue constant

$$\Lambda_M = \sup_{\mathbf{x} \in \Omega} \sum_{i=1}^M |l_i^M(\mathbf{x})|;$$

note that Λ_M depends on X_M and the magic points T_M , but is $\boldsymbol{\mu}$ -independent. An upper bound (indeed quite pessimistic) for the Lebesgue constant is $\Lambda_M \leq 2^M - 1$, see Exercise 3. We also recall that the Lebesgue constant affects the a priori estimate for the interpolation error (see (A.26)).

Proposition 10.1. *For any $g \in \mathcal{G}$, the interpolation error satisfies*

$$\varepsilon_M(\boldsymbol{\mu}) := \|g(\cdot; \boldsymbol{\mu}) - \mathcal{J}_M^{\mathbf{x}} g(\cdot; \boldsymbol{\mu})\|_{L^\infty(\Omega)} \leq (1 + \Lambda_M) \inf_{g_M \in X_M} \|g(\cdot; \boldsymbol{\mu}) - g_M\|_{L^\infty(\Omega)}. \quad (10.10)$$

See Exercise 4 for the proof. The estimate (10.10) provides a theoretical basis for the stability of the empirical interpolation method. The last term in the right-hand side of (10.10) is referred to as the *best approximation* of g by elements in X_M in the L^∞ -norm.

Similarly to the convergence result for the greedy RB algorithm provided by Theorem 7.1, also in the case of the empirical interpolation method it is possible to link the convergence rate of EIM approximations to the Kolmogorov M -width of the manifold \mathcal{G} . The following result (see e.g. [172] for its proof) holds.

Theorem 10.2. Assume that $\mathcal{G} \subset X = C^0(\Omega)$, and that there exists a sequence of nested finite-dimensional spaces $\mathcal{Z}_1 \subset \mathcal{Z}_2 \dots$, $\dim(\mathcal{Z}_M) = M$, and $\mathcal{Z}_M \subset \text{span}\{\mathcal{G}\}$ such that there exists $c > 0$ and $\alpha > \log 4$ with

$$d(\mathcal{G}, \mathcal{Z}_M) = \sup_{\mu \in \mathcal{P}} \inf_{v_M \in \mathcal{Z}_M} \|g(\cdot; \mu) - v_M\|_X \leq ce^{-\alpha M}. \quad (10.11)$$

Then

$$\sup_{\mu \in \mathcal{P}} \|g(\cdot; \mu) - \mathcal{I}_M^{\mathbf{x}} g(\cdot; \mu)\|_{L^\infty(\Omega)} \leq ce^{-(\alpha - \log 4)M}.$$

Remark 10.2. Condition (10.11) implies in particular that $d_M(\mathcal{G}; X) \leq ce^{-\alpha M}$, i.e. \mathcal{G} has an exponentially small Kolmogorov M -width. \bullet

To derive an a posteriori error estimate for the EIM error we proceed as follows. Given an approximation $g_M(\cdot; \mu)$, for $M \leq M_{\max} - 1$, let us define

$$E_M(\mathbf{x}; \mu) = \hat{\varepsilon}_M(\mu) \rho_{M+1}(\mathbf{x}), \quad \hat{\varepsilon}_M(\mu) = |g(\mathbf{t}^{M+1}; \mu) - \mathcal{I}_M^{\mathbf{x}} g(\mathbf{t}^{M+1}; \mu)|.$$

In general, $\varepsilon_M(\mu) \geq \hat{\varepsilon}_M(\mu)$. However, it is possible to show that

Proposition 10.2. If $g(\cdot; \mu) \in X_{M+1}$, then

1. $g(\mathbf{x}; \mu) - \mathcal{I}_M^{\mathbf{x}} g(\mathbf{x}; \mu) = \pm E_M(\mathbf{x}; \mu)$;
2. $\varepsilon_M(\mu) = \hat{\varepsilon}_M(\mu)$.

Proof. If $g(\cdot; \mu) \in X_{M+1}$ there exists $\kappa(\mu) \in \mathbb{R}^{M+1}$ such that $g(\mathbf{x}; \mu) - \mathcal{I}_M^{\mathbf{x}} g(\mathbf{x}; \mu) = \sum_{i=1}^{M+1} \kappa_i(\mu) \rho_i(\mathbf{x})$. Taking $\mathbf{x} = \mathbf{t}^j$, $j = 1, \dots, M+1$, we get

$$\sum_{i=1}^{M+1} \kappa_i(\mu) \rho_i(\mathbf{t}^i) = g(\mathbf{t}^j; \mu) - \mathcal{I}_M^{\mathbf{x}} g(\mathbf{t}^j; \mu),$$

from which we obtain that (i) $\kappa_i(\mu) = 0$, $i = 1, \dots, M$ since $g(\mathbf{t}^i; \mu) - \mathcal{I}_M^{\mathbf{x}} g(\mathbf{t}^i; \mu) = 0$, $i = 1, \dots, M$ and \mathbb{B}_M is lower triangular, i.e. $\rho_i(\mathbf{x}^j) = 0$ if $i < j$, and that (ii) $\kappa_{M+1}(\mu) = g(\mathbf{t}^{M+1}; \mu) - \mathcal{I}_M^{\mathbf{x}} g(\mathbf{t}^{M+1}; \mu)$ since $\rho_{M+1}(\mathbf{t}^{M+1}) = 1$. This concludes the proof of point 1. Point 2 directly follows since $\|\rho_{M+1}\|_{L^\infty(\Omega)} = 1$. \square

However, as in general $g(\cdot; \mu) \notin X_{M+1}$, we only have that $\varepsilon_M(\mu) \geq \hat{\varepsilon}_M(\mu)$ for any $\mu \in \mathcal{P}$, i.e., $\hat{\varepsilon}_M(\mu)$ is a lower bound of the interpolation error in L^∞ -norm. Nevertheless, if $\varepsilon_M(\mu) \rightarrow 0$ very fast, we expect the effectivity

$$\eta_M(\mu) = \frac{\hat{\varepsilon}_M(\mu)}{\|g(\cdot; \mu) - \mathcal{I}_M^{\mathbf{x}} g(\cdot; \mu)\|_{L^\infty(\Omega)}}$$

to be close to 1. In any case, evaluating the estimator only requires an additional evaluation of $g(\cdot; \mu)$ at a single point in Ω . For this reason, we define the *one point error estimator* as

$$\Delta_M(\mu) = \hat{\varepsilon}_M(\mu) \quad (10.12)$$

corresponding to the interpolation error at the $(M+1)$ -th *magic point*, the one where the residual $r_M(\mathbf{x})$ attains its maximum.

While not a rigorous a posteriori error bound, this quantity provides a heuristic measure of the EIM error. A rigorous (but very expensive to compute) a posteriori error bound that combines analytical upper bounds for the parametric derivatives of f , the Lebesgue constant and interpolation errors evaluated at a set of points in \mathcal{P} , can be found in [98].

10.2.1 Practical Implementation

Similarly to the (weak) greedy algorithm described in Sect. 7.1.2, finding the supremum in (10.5) and (10.6) – see lines 6 and 8 of Algorithm 10.1, respectively – is not computationally feasible unless an approximation of both Ω and \mathcal{P} is considered. For this reason, we introduce:

1. a fine sample $\Xi_{\text{train}}^{\text{EIM}} \subset \mathcal{P}$ of cardinality $|\Xi_{\text{train}}^{\text{EIM}}| = n_{\text{train}}^{\text{EIM}}$ to train the EIM algorithm – similarly to the training sample introduced in Algorithm 7.1;
2. a discrete approximation $\Omega_h = \{\mathbf{x}^k\}_{k=1}^{N_q}$ of Ω of dimension N_q . In the finite element context, the points \mathbf{x}^i can be for instance the vertices of the computational mesh, or the quadrature points.

By so doing, problems (10.5) and (10.6) are turned into simpler enumeration problems. In this setting, we can also provide an algebraic version of the EIM (see also Algorithm 10.2). We first introduce the vector representation $\mathbf{g} : \mathcal{P} \rightarrow \mathbb{R}^{N_q}$ of $g : \Omega_h \times \mathcal{P} \rightarrow \mathbb{R}$, defined as

$$(\mathbf{g}(\boldsymbol{\mu}))_k = g(\mathbf{x}^k; \boldsymbol{\mu}), \quad k = 1, \dots, N_q$$

obtained by evaluating the function g in Ω_h , for any $\boldsymbol{\mu} \in \mathcal{P}$. Then, we denote by $\mathbb{Q} \in \mathbb{R}^{N_q \times M}$ the matrix

$$\mathbb{Q} = [\boldsymbol{\rho}_1 \mid \dots \mid \boldsymbol{\rho}_M]$$

whose columns are the discrete representation of the basis functions $\{\rho_1, \dots, \rho_M\}$, i.e. $(\mathbb{Q})_{kj} = \rho_j(\mathbf{x}^k)$. Moreover we denote by $\mathcal{J} = \{i_1, \dots, i_M\}$ a set of interpolation indices such that $\{\mathbf{t}^1, \dots, \mathbf{t}^M\} = \{\mathbf{x}^{i_1}, \dots, \mathbf{x}^{i_M}\}$. The discrete representation $\mathbf{g}_M : \mathcal{P} \rightarrow \mathbb{R}^{N_q}$ of the interpolation operator $\mathcal{S}_M^{\mathbf{x}}$ is given by

$$\mathbf{g}_M(\boldsymbol{\mu}) = \mathbb{Q}\boldsymbol{\gamma}(\boldsymbol{\mu}) \in \mathbb{R}^{N_q},$$

where $\boldsymbol{\gamma}(\boldsymbol{\mu}) \in \mathbb{R}^M$ is the solution of the following linear system

$$(\mathbf{g}_M(\boldsymbol{\mu}))_{i_m} = \sum_{j=1}^M \gamma_j(\boldsymbol{\mu}) (\boldsymbol{\rho}_j)_{i_m} = (\mathbf{g}(\boldsymbol{\mu}))_{i_m}, \quad m = 1, \dots, M. \quad (10.13)$$

Denoting by $\mathbf{g}_{\mathcal{J}}(\boldsymbol{\mu}) \in \mathbb{R}^M$ the vector whose components are $(\mathbf{g}_{\mathcal{J}}(\boldsymbol{\mu}))_m = (\mathbf{g}(\boldsymbol{\mu}))_{i_m}$ for $m = 1, \dots, M$, and noting that the $M \times M$ matrix \mathbb{B}_M can be easily formed by restricting the $N_q \times M$ matrix \mathbb{Q} to the rows \mathcal{J} , i.e. $\mathbb{B}_M = \mathbb{Q}_{\mathcal{J}}$, (10.13) can be written

Algorithm 10.2 Empirical interpolation method (computable version): offline and online phases

```

1: function  $[\mathbb{Q}, \mathcal{J}] = \text{EIM\_OFFLINE}(\mathcal{T}_{\text{train}}^{EIM}, \Omega_h, M_{\text{max}}, \varepsilon_{\text{EIM}})$ 
2:    $M = 0, e_0 = \varepsilon_{\text{EIM}} + 1$ 
3:    $\boldsymbol{\mu}^1 = \arg \max_{\boldsymbol{\mu} \in \mathcal{T}_{\text{train}}^{EIM}} \|\mathbf{g}(\boldsymbol{\mu})\|_{\infty}$ 
4:    $\mathbf{r} = \mathbf{g}(\boldsymbol{\mu}^1), \mathbb{Q} = []$ 
5:   while  $M < M_{\text{max}}$  and  $e_M > \varepsilon_{\text{EIM}}$ 
6:      $M \leftarrow M + 1$ 
7:      $i_M = \arg \max_{i=1, \dots, N_q} |\mathbf{r}_i|$ 
8:      $\boldsymbol{\rho}_M = \mathbf{r} / \mathbf{r}_{i_M}$ 
9:      $\mathbb{Q} \leftarrow \mathbb{Q} \cup \boldsymbol{\rho}_M, \mathcal{J} \leftarrow \mathcal{J} \cup i_M,$ 
10:     $[e_M, \boldsymbol{\mu}^{M+1}] = \arg \max_{\boldsymbol{\mu} \in \mathcal{T}_{\text{train}}^{EIM}} \|\mathbf{g}(\boldsymbol{\mu}) - \mathbb{Q}\mathbb{Q}_{\mathcal{J}}^{-1}\mathbf{g}_{\mathcal{J}}(\boldsymbol{\mu})\|_{\infty}$ 
11:     $\mathbf{r} = \mathbf{g}(\boldsymbol{\mu}^{M+1}) - \mathbb{Q}\mathbb{Q}_{\mathcal{J}}^{-1}\mathbf{g}_{\mathcal{J}}(\boldsymbol{\mu}^{M+1})$ 
12:  end while
13: end function

1: function  $\boldsymbol{\gamma}(\boldsymbol{\mu}) = \text{EIM\_ONLINE}(\mathbb{Q}_{\mathcal{J}}, \boldsymbol{\mu}, \{\mathbf{t}^1, \dots, \mathbf{t}^M\})$ 
2:   form  $\mathbf{g}_{\mathcal{J}}(\boldsymbol{\mu})$  by evaluating  $\mathbf{g}(\cdot, \boldsymbol{\mu})$  in the interpolation points  $\{\mathbf{t}^1, \dots, \mathbf{t}^M\}$ 
3:   solve  $\mathbb{Q}_{\mathcal{J}}\boldsymbol{\gamma}(\boldsymbol{\mu}) = \mathbf{g}_{\mathcal{J}}(\boldsymbol{\mu})$ 
4: end function

```

in compact form as

$$\mathbb{Q}_{\mathcal{J}}\boldsymbol{\gamma}(\boldsymbol{\mu}) = \mathbf{g}_{\mathcal{J}}(\boldsymbol{\mu}). \quad (10.14)$$

We finally obtain the following expression for the EIM approximation

$$\mathbf{g}_M(\boldsymbol{\mu}) = \mathbb{Q}\mathbb{Q}_{\mathcal{J}}^{-1}\mathbf{g}_{\mathcal{J}}(\boldsymbol{\mu}) \quad \forall \boldsymbol{\mu} \in \mathcal{P}. \quad (10.15)$$

Note that the solution of the dense linear system (10.14) has complexity $O(M^2)$, since the matrix $\mathbb{Q}_{\mathcal{J}}$ is lower triangular.

Remark 10.3. At each iteration, algorithm 10.2 requires to evaluate $\mathbf{g}(\boldsymbol{\mu})$ for $\boldsymbol{\mu} \in \mathcal{T}_{\text{train}}^{EIM}$. Should this operation be expensive, one may form and store the (possibly dense) matrix

$$\mathbb{S} = [\mathbf{g}(\boldsymbol{\mu}^1) \mid \dots \mid \mathbf{g}(\boldsymbol{\mu}^{n_{\text{train}}^{EIM}})] \in \mathbb{R}^{N_q \times n_{\text{train}}^{EIM}}$$

once and for all before entering the while loop. However, already for moderately large N_q and n_{train}^{EIM} , storing the matrix \mathbb{S} can be quite challenging. For instance, in the setting of Sect. 3.8, approximating a function defined over a set of $N_q = 4 \cdot 2.2 \cdot 10^5$ points (corresponding to 4 quadrature points for each tetrahedron) using a training set of dimension $n_{\text{train}}^{EIM} = 10^3$, requires to store as much as about 7 GB of data. •

10.3 Discrete Empirical Interpolation

An alternative to the EIM for approximating a nonaffinely parametrized function is the so-called discrete empirical interpolation method (DEIM), originally introduced in [58]. To avoid misunderstandings, we underline that the computable version of EIM (see Algorithm 10.2) is discrete too, being implemented on arrays.

Similarly to EIM, DEIM approximates a nonlinear function $\mathbf{g}: \boldsymbol{\mu} \in \mathcal{P} \subset \mathbb{R}^P \rightarrow \mathbf{g}(\boldsymbol{\mu}) \in \mathbb{R}^{N_q}$ by projection onto a low-dimensional subspace spanned by a basis \mathbb{Q} ,

$$\mathbf{g}(\boldsymbol{\mu}) \approx \mathbf{g}_M(\boldsymbol{\mu}) = \mathbb{Q}\boldsymbol{\gamma}(\boldsymbol{\mu}), \quad (10.16)$$

where $\mathbb{Q} = [\boldsymbol{\rho}_1, \dots, \boldsymbol{\rho}_M] \in \mathbb{R}^{N_q \times M}$ and $\boldsymbol{\gamma}(\boldsymbol{\mu}) \in \mathbb{R}^M$ is the corresponding vector of coefficients, with $M \ll N_q$. The difference is on the construction of the basis \mathbb{Q} , that is obtained operating a POD on a set of snapshots

$$\mathbb{S} = [\mathbf{g}(\boldsymbol{\mu}_{DEIM}^1) \mid \dots \mid \mathbf{g}(\boldsymbol{\mu}_{DEIM}^{n_s})], \quad n_s^{DEIM} > M$$

instead than being embedded in the EIM greedy algorithm. Note that for both EIM and DEIM the interpolation points are iteratively selected with the same greedy algorithm. DEIM thus requires to:

- (i) construct a set of snapshots obtained by sampling $\mathbf{g}(\boldsymbol{\mu})$ at values $\boldsymbol{\mu}_{DEIM}^i$, $i = 1, \dots, n_s$ and apply POD to extract the basis

$$[\boldsymbol{\rho}_1, \dots, \boldsymbol{\rho}_M] = \text{POD}([\mathbf{g}(\boldsymbol{\mu}_{DEIM}^1), \dots, \mathbf{g}(\boldsymbol{\mu}_{DEIM}^{n_s})], \varepsilon_{\text{POD}}),$$

where $\varepsilon_{\text{DEIM}}$ is a prescribed tolerance;

- (ii) select iteratively M indices $\mathcal{I} \subset \{1, \dots, N_q\}$, $|\mathcal{I}| = M$ from the basis \mathbb{Q} using a greedy procedure, which minimizes at each step the interpolation error over the snapshots set measured in the maximum norm. This operation is indeed the same as for the selection of the EIM magic points;
- (iii) given a new $\boldsymbol{\mu}$, in order to compute the coefficients vector $\boldsymbol{\gamma}(\boldsymbol{\mu})$, interpolation constraints are imposed at the M points corresponding to the selected indices, thus requiring the solution of the following linear system

$$\mathbb{Q}_{\mathcal{I}} \boldsymbol{\gamma}(\boldsymbol{\mu}) = \mathbf{g}_{\mathcal{I}}(\boldsymbol{\mu}), \quad (10.17)$$

where $\mathbb{Q}_{\mathcal{I}} \in \mathbb{R}^{M \times M}$ is the matrix formed by the \mathcal{I} rows of \mathbb{Q} . As a result,

$$\mathbf{g}_M(\boldsymbol{\mu}) = \mathbb{Q}\mathbb{Q}_{\mathcal{I}}^{-1} \mathbf{g}_{\mathcal{I}}(\boldsymbol{\mu}). \quad (10.18)$$

The construction of the basis and the selection of interpolation points is summarized in Algorithm 10.3. In this case, the solution of the linear system (10.17) has complexity $O(M^3)$. Similar results to those of Sect. 10.1.3 concerning the well-posedness of the DEIM procedure can be proved, see e.g. [58] for further details.

Algorithm 10.3 Discrete empirical interpolation method: offline and online phases

```

1: function  $[\mathbb{Q}, \mathcal{J}] = \text{DEIM\_OFFLINE}(\mathbb{S}, \varepsilon_{\text{DEIM}})$ 
2:    $[\boldsymbol{\rho}_1 \mid \dots \mid \boldsymbol{\rho}_M] = \text{POD}(\mathbb{S}, \varepsilon_{\text{DEIM}})$ 
3:    $i_m = \arg \max_{i=1, \dots, N_q} |(\boldsymbol{\rho}_1)_i|$ 
4:    $\mathbb{Q} = \boldsymbol{\rho}_1, \mathcal{J} = \{i_1\},$ 
5:   for  $m = 2 : M$ 
6:      $\mathbf{r} = \boldsymbol{\rho}_m - \mathbb{Q}\mathbb{Q}_{\mathcal{J}}^{-1}(\boldsymbol{\rho}_m)_{\mathcal{J}}$ 
7:      $i_m = \arg \max_{i=1, \dots, N_q} |\mathbf{r}_i|$ 
8:      $\mathbb{Q} \leftarrow [\mathbb{Q} \mid \boldsymbol{\rho}_m], \mathcal{J} \leftarrow \mathcal{J} \cup i_m$ 
9:   end for
10: end function

1: function  $\boldsymbol{\gamma}(\boldsymbol{\mu}) = \text{DEIM\_ONLINE}(\mathbb{Q}_{\mathcal{J}}, \boldsymbol{\mu}, \{\mathbf{t}^1, \dots, \mathbf{t}^M\})$ 
2:   form  $\mathbf{g}_{\mathcal{J}}(\boldsymbol{\mu})$  by evaluating  $g(\cdot, \boldsymbol{\mu})$  in the interpolation points  $\{\mathbf{t}^1, \dots, \mathbf{t}^M\}$ 
3:   solve  $\mathbb{Q}_{\mathcal{J}} \boldsymbol{\gamma}(\boldsymbol{\mu}) = \mathbf{g}_{\mathcal{J}}(\boldsymbol{\mu})$ 
4: end function

```

Here we only point out that the error between \mathbf{g} and its DEIM approximation \mathbf{g}_M can be bounded as

$$\|\mathbf{g}(\boldsymbol{\mu}) - \mathbf{g}_M(\boldsymbol{\mu})\|_2 \leq \|\mathbb{Q}_{\mathcal{J}}^{-1}\|_2 \|(\mathbb{I} - \mathbb{Q}\mathbb{Q}^T)\mathbf{g}(\boldsymbol{\mu})\|_2, \quad (10.19)$$

with

$$\|(\mathbb{I} - \mathbb{Q}\mathbb{Q}^T)\mathbf{g}(\boldsymbol{\mu})\|_2 \approx \sigma_{M+1}, \quad (10.20)$$

being σ_{M+1} the first discarded singular value of the matrix \mathbb{S} when selecting M basis through the POD procedure, see (6.6). This approximation holds for any $\boldsymbol{\mu} \in \mathcal{P}$ provided a suitable sampling in the parameter space has been carried out to build the snapshot matrix \mathbb{S} . In that case, the predictive projection error (10.20) is comparable to the training projection error σ_{M+1} . Similarly to the *one point* error estimator (10.12) for EIM, (10.19) exploits the information related to the first discarded term and can be seen as a heuristic measure of the DEIM error.

Remark 10.4. The interpolation condition (10.17) can be generalized to the case where more sample indices ($|\mathcal{J}| > M$) than basis functions are considered. This leads to the so-called *gappy POD* reconstruction [42, 57], which provides an approximation under the form (10.16) where instead of the linear system (10.17) we need to solve

$$\boldsymbol{\gamma}(\boldsymbol{\mu}) = \arg \min_{\mathbf{x} \in \mathbb{R}^M} \|\mathbf{g}_{\mathcal{J}}(\boldsymbol{\mu}) - \mathbb{Q}_{\mathcal{J}} \mathbf{x}\|_2.$$

The solution of this least-squares problem yields $\mathbf{g}_M(\boldsymbol{\mu}) = \mathbb{Q}\mathbb{Q}_{\mathcal{J}}^+ \mathbf{g}_{\mathcal{J}}(\boldsymbol{\mu})$, where $\mathbb{Q}_{\mathcal{J}}^+$ is the Moore-Penrose pseudoinverse of the matrix $\mathbb{Q}_{\mathcal{J}}$. •

10.4 EIM-G-RB Approximation of Nonaffine Problems

We show how to construct a RB approximation of the problem of Sect. 10.1 which exploits the empirical interpolation method to recover an affine parametric dependence in the originally nonaffine operators. Although both EIM and DEIM lead to the same computational procedure from a practical standpoint, we exploit the former to develop our analysis in a more straightforward way.

Consider the model problem (5.23) where for the sake of interpolation we assume that both $s(\cdot; \boldsymbol{\mu})$ and $k(\cdot; \boldsymbol{\mu}) \in C^0(\Omega)$ for any $\boldsymbol{\mu} \in \mathcal{P}$. Moreover, we assume to deal with a strongly coercive problem, although all the results of this section can be easily extended to the more general case of weakly coercive problems.

The high-fidelity approximation of (5.23) reads: find $u_h(\boldsymbol{\mu}) \in V_h$ such that

$$a(u_h, v_h; \boldsymbol{\mu}) = f(v_h; \boldsymbol{\mu}) \quad \forall v_h \in V_h, \quad (10.21)$$

with

$$a(u_h, v_h; \boldsymbol{\mu}) = \int_{\Omega} k(\mathbf{x}; \boldsymbol{\mu}) \nabla u_h \cdot \nabla v_h d\Omega, \quad f(v_h; \boldsymbol{\mu}) = \int_{\Omega} s(\mathbf{x}; \boldsymbol{\mu}) v_h d\Omega. \quad (10.22)$$

We assume that the diffusion coefficient $k(\mathbf{x}; \boldsymbol{\mu})$ and the source term $s(\mathbf{x}; \boldsymbol{\mu})$ are non-affine functions of $\boldsymbol{\mu}$, yielding a nonaffine parametric dependence of the linear and bilinear forms. Using the EIM, we replace them by the corresponding interpolants

$$k_M(\mathbf{x}; \boldsymbol{\mu}) = \mathcal{J}_M^{\mathbf{x}} k(\mathbf{x}; \boldsymbol{\mu}) = \sum_{j=1}^{M_k} \gamma_j^k(\boldsymbol{\mu}) \rho_j^k(\mathbf{x}) \quad (10.23)$$

$$s_M(\mathbf{x}; \boldsymbol{\mu}) = \mathcal{J}_M^{\mathbf{x}} s(\mathbf{x}; \boldsymbol{\mu}) = \sum_{j=1}^{M_s} \gamma_j^s(\boldsymbol{\mu}) \rho_j^s(\mathbf{x}). \quad (10.24)$$

Correspondingly, the high-fidelity problem with EIM approximation (to which we refer to as EIM high-fidelity approximation) becomes: find $u_h^M(\boldsymbol{\mu}) \in V_h$ such that

$$a_M(u_h^M, v_h; \boldsymbol{\mu}) = f_M(v_h; \boldsymbol{\mu}) \quad \forall v_h \in V_h, \quad (10.25)$$

where we have set

$$a_M(u_h, v_h; \boldsymbol{\mu}) = \int_{\Omega} k_M(\mathbf{x}; \boldsymbol{\mu}) \nabla u_h \cdot \nabla v_h d\Omega, \quad f_M(v_h; \boldsymbol{\mu}) = \int_{\Omega} s_M(\mathbf{x}; \boldsymbol{\mu}) v_h d\Omega.$$

The linear and the bilinear forms now depend on the number M_k, M_s of terms appearing in the interpolants; we denote the EIM dependence by the subscript M . The associated Galerkin RB problem, referred to as EIM-G-RB problem, thus reads:

find $u_N^M(\boldsymbol{\mu}) \in V_N$ such that

$$a_M(u_N^M, v_N; \boldsymbol{\mu}) = f_M(v_N; \boldsymbol{\mu}) \quad \forall v_N \in V_N. \quad (10.26)$$

We remark that the bilinear form $a_M(\cdot, \cdot; \boldsymbol{\mu})$ does not necessarily preserve the properties of $a(\cdot, \cdot; \boldsymbol{\mu})$. While the symmetry of $a(\cdot, \cdot; \boldsymbol{\mu})$ is automatically inherited by its approximation, this is not the case for the (strong) coercivity. However, by requiring that the high-fidelity problem (10.25) is coercive, the EIM-G-RB problem (10.26) is well-posed too, being a Galerkin projection.

The EIM-G-RB problem can be considered as a *generalized Galerkin method*. The analysis of convergence is based in this case on the following general result:

Theorem 10.3. *Let us suppose that $a_M(\cdot, \cdot; \boldsymbol{\mu})$ is continuous on $V_h \times V_h$ and coercive on V_h for any $\boldsymbol{\mu} \in \mathcal{P}$, that is*

$$\exists \alpha_h^M(\boldsymbol{\mu}) : a_M(v, v; \boldsymbol{\mu}) \geq \alpha_h^M(\boldsymbol{\mu}) \|v\|_V^2 \quad \forall v \in V_h, \forall \boldsymbol{\mu} \in \mathcal{P}.$$

Moreover, let us suppose that $f_M(\cdot; \boldsymbol{\mu})$ is continuous on V_h . Then problem (10.26) admits a unique solution $u_N^M(\boldsymbol{\mu}) \in V_N$ which satisfies

$$\|u_N^M(\boldsymbol{\mu})\|_V \leq \frac{1}{\alpha_N^M(\boldsymbol{\mu})} \|f_M(\cdot; \boldsymbol{\mu})\|_{V_h'},$$

being

$$\alpha_N^M(\boldsymbol{\mu}) = \inf_{v \in V_N} \frac{a_M(v, v; \boldsymbol{\mu})}{\|v\|_V^2} \geq \alpha_h^M(\boldsymbol{\mu})$$

the stability factor of the EIM-G-RB problem, and fulfills the following a priori error estimate

$$\begin{aligned} \|u_h(\boldsymbol{\mu}) - u_N^M(\boldsymbol{\mu})\|_V \leq & \inf_{w \in V_N} \left\{ \left(1 + \frac{\gamma_h(\boldsymbol{\mu})}{\alpha_h^M(\boldsymbol{\mu})} \right) \|u_h(\boldsymbol{\mu}) - w\|_V \right. \\ & + \frac{1}{\alpha_h^M(\boldsymbol{\mu})} \sup_{v \in V_N} \frac{|a(w, v; \boldsymbol{\mu}) - a_M(w, v; \boldsymbol{\mu})|}{\|v\|_V} \Big\} \\ & + \frac{1}{\alpha_h^M(\boldsymbol{\mu})} \sup_{v \in V_N} \frac{|f(v; \boldsymbol{\mu}) - f_M(v; \boldsymbol{\mu})|}{\|v\|_V}. \end{aligned} \quad (10.27)$$

The proof of this result – indeed, of general importance – can be developed by following the proof of the analogous result reported in, e.g., [216, Lemma 10.1]. By observing the right-hand side of the estimate (10.27), we can recognize three different contributions to the error $u_h(\boldsymbol{\mu}) - u_N^M(\boldsymbol{\mu})$: the first is the best approximation error, the second is the error deriving from the approximation of the bilinear form $a(\cdot, \cdot; \boldsymbol{\mu})$ using the form $a_M(\cdot, \cdot; \boldsymbol{\mu})$, and the third is the error arising from the approximation of the linear form $f(\cdot; \boldsymbol{\mu})$ by $f_M(\cdot; \boldsymbol{\mu})$. Note that (10.27) is indeed a generalization of the a priori error estimate (3.46).

It is also possible to derive an a posteriori error estimate by combining the error estimator (3.72) obtained in the case of a generic linear elliptic PDE, and a suitable indicator of the empirical interpolation error.

Let us denote by $e_h^M(\boldsymbol{\mu}) = u_h(\boldsymbol{\mu}) - u_N^M(\boldsymbol{\mu})$ the error between the high-fidelity and the EIM-G-RB solutions and by

$$r_M(v; \boldsymbol{\mu}) = f_M(v; \boldsymbol{\mu}) - a_M(u_N^M(\boldsymbol{\mu}), v; \boldsymbol{\mu}) \quad \forall v \in V$$

the residual of the EIM high-fidelity problem (10.25) computed on the RB solution. We assume that $a(\cdot, \cdot; \boldsymbol{\mu})$ is strongly coercive over $V_h \times V_h$ for any $\boldsymbol{\mu} \in \mathcal{P}$ and denote by $\alpha_h(\boldsymbol{\mu})$ its stability factor; see Exercise 5 for the extension to a weakly coercive problem.

Proposition 10.3. *The following a posteriori error estimate holds*

$$\|u_h(\boldsymbol{\mu}) - u_N^M(\boldsymbol{\mu})\|_V \leq \frac{1}{\alpha_h(\boldsymbol{\mu})} \left(\|r_M(\cdot; \boldsymbol{\mu})\|_{V_h'} + C_f \delta_s(\boldsymbol{\mu}) + C_a \delta_k(\boldsymbol{\mu}) \|u_N^M(\boldsymbol{\mu})\|_V \right) \quad (10.28)$$

where

$$C_f = \sup_{v \in V_h} \frac{\int_{\Omega} v d\Omega}{\|v\|_V}, \quad C_a = \sup_{v \in V_h} \sup_{w \in V_h} \frac{\int_{\Omega} \nabla v \cdot \nabla w d\Omega}{\|v\|_V \|w\|_V}$$

and

$$\delta_k(\boldsymbol{\mu}) = \|k(\cdot; \boldsymbol{\mu}) - k_M(\cdot; \boldsymbol{\mu})\|_{L^\infty(\Omega)}, \quad \delta_s(\boldsymbol{\mu}) = \|s(\cdot; \boldsymbol{\mu}) - s_M(\cdot; \boldsymbol{\mu})\|_{L^\infty(\Omega)}.$$

Proof. From (10.21) we have that, for all $v_h \in V_h$,

$$\begin{aligned} a(u_h(\boldsymbol{\mu}), v_h; \boldsymbol{\mu}) - a(u_N^m(\boldsymbol{\mu}), v_h; \boldsymbol{\mu}) &= f(v_h; \boldsymbol{\mu}) - a(u_N^m(\boldsymbol{\mu}), v_h; \boldsymbol{\mu}) \\ &= f(v_h; \boldsymbol{\mu}) - f_M(v_h; \boldsymbol{\mu}) + f_M(v_h; \boldsymbol{\mu}) \\ &\quad - a_M(u_N^m(\boldsymbol{\mu}), v_h; \boldsymbol{\mu}) + a_M(u_N^m(\boldsymbol{\mu}), v_h; \boldsymbol{\mu}) - a(u_N^m(\boldsymbol{\mu}), v_h; \boldsymbol{\mu}). \end{aligned}$$

By taking $v_h = e_h^M(\boldsymbol{\mu})$, exploiting the continuity of the linear and bilinear forms and the coercivity of $a(\cdot, \cdot; \boldsymbol{\mu})$, and dividing by $\|e_h^M(\boldsymbol{\mu})\|_V$, we find

$$\begin{aligned} \alpha_h(\boldsymbol{\mu}) \|e_h^M(\boldsymbol{\mu})\|_V &\leq \|f(\cdot; \boldsymbol{\mu}) - f_M(\cdot; \boldsymbol{\mu})\|_{V_h'} \\ &\quad + \|a(\cdot, \cdot; \boldsymbol{\mu}) - a_M(\cdot, \cdot; \boldsymbol{\mu})\|_{\mathcal{L}(V_h, V_h')} \|u_N^M(\boldsymbol{\mu})\|_V + \|r_M(\cdot; \boldsymbol{\mu})\|_{V_h'}. \end{aligned}$$

The first term can be bounded as

$$\begin{aligned} \|f(\cdot; \boldsymbol{\mu}) - f_M(\cdot; \boldsymbol{\mu})\|_{V_h'} &\leq \sup_{v \in V_h} \frac{|\int_{\Omega} (s(\cdot; \boldsymbol{\mu}) - s_M(\cdot; \boldsymbol{\mu})) v d\Omega|}{\|v\|_V} \\ &\leq C_f \|s(\cdot; \boldsymbol{\mu}) - s_M(\cdot; \boldsymbol{\mu})\|_{L^\infty(\Omega)}. \end{aligned}$$

Similarly for the second term we find

$$\begin{aligned} \|a(\cdot, \cdot; \boldsymbol{\mu}) - a_M(\cdot, \cdot; \boldsymbol{\mu})\|_{\mathcal{L}(V_h, V_h')} &\leq \sup_{v \in V_h} \sup_{w \in V_h} \frac{|\int_{\Omega} (k(\cdot; \boldsymbol{\mu}) - k_M(\cdot; \boldsymbol{\mu})) \nabla v \cdot \nabla w d\Omega|}{\|v\|_V \|w\|_V} \\ &\leq C_a \|k(\cdot; \boldsymbol{\mu}) - k_M(\cdot; \boldsymbol{\mu})\|_{L^\infty(\Omega)}. \end{aligned}$$

□

Remark 10.5. Surrogates for $\delta_k(\boldsymbol{\mu})$ and $\delta_s(\boldsymbol{\mu})$ could be either the tolerance of the EIM algorithm or the estimator $\hat{\epsilon}_M(\boldsymbol{\mu})$. •

A similar result concerning a posteriori error bounds taking into account the EIM error in the case of nonaffinely linear elliptic PDEs can be found, e.g., in [197]. The evaluation of the norm of the residual in the a posteriori error estimate (10.28) can be efficiently performed according to the offline/online splitting described in Sect. 3.7.1. Instead, for evaluating the stability factor, the only available option is given by the interpolatory radial basis function technique of Sect. 3.7.3: the successive constraint method, relying on the affine parametric dependence of $a(\cdot, \cdot; \boldsymbol{\mu})$, is not applicable for the case at hand.

10.5 Mass Transfer with Parametrized Source: Results

In this section, we exploit the EIM and DEIM to generate a RB approximation of the mass transfer problem of Sect. 8.4. The problem depends on $P = 5$ parameters: $\mu_1 \in [0.01, 0.1]$ is the diffusion coefficient, $\mu_2 \in [0, 2\pi]$ represents the direction of the advection field, while μ_3, μ_4 and μ_5 affects the nonaffine source term

$$s(\mathbf{x}; \boldsymbol{\mu}) = \exp \left(- \frac{(x_1 - \mu_3)^2 + (x_2 - \mu_4)^2}{\mu_5^2} \right).$$

For the high-fidelity approximation we use \mathbb{P}_1 finite elements built over a discretization of the domain made by triangular elements, resulting in 5305 vertices, 10368 triangles and a high-fidelity space V_h of dimension $N_h = 5305$. In order to generate an (approximate) affine decomposition for

$$f(v_h; \boldsymbol{\mu}) = \int_{\Omega} s(\mathbf{x}; \boldsymbol{\mu}) v_h d\Omega \quad \forall v_h \in V_h, \quad (10.29)$$

we apply the (D)EIM to the function $s(\mathbf{x}; \boldsymbol{\mu})$. As the integral in (10.29) is approximated by means of a 4-th order Dunavant quadrature rule [95] (yielding 6 quadrature points per triangle), we have that $N_q = 62208$.

10.5.1 Comparison of EIM and DEIM

As a first test case, we consider $\mu_3 \in [0.2, 0.8]$, $\mu_4 \in [0.15, 0.35]$ with $\mu_5 = 0.25$ being fixed; we recall that μ_1 and μ_2 do not affect $s(\mathbf{x}; \boldsymbol{\mu})$. We first run DEIM using a training sample of $n_s^{DEIM} = 100$ points obtained by latin hypercube sampling and a tolerance $\varepsilon_{DEIM} = 10^{-6}$. POD extracts a basis of dimension $M = 40$, whose singular values are reported in Fig. 10.1. Then, we run EIM using a training sample of size $|\mathcal{Z}_s^{EIM}| = 1000$ and $M_{\max} = 40$. The resulting error and estimate between s and its EIM/DEIM approximations computed over a parameter test sample of 200 random points are reported in Fig. 10.2. We note that both EIM and DEIM provide a very similar accuracy, however the EIM error estimate (10.12) is much sharper than the DEIM estimate (10.19).

We then consider the case $\mu_5 = [0.1, 0.35]$. Using a training sample of size $n_s^{DEIM} = 200$ and a tolerance $\varepsilon_{DEIM} = 10^{-5}$, we end up with $M = 83$ POD basis functions and DEIM indices. We then run EIM with $|\mathcal{Z}_s^{EIM}| = 2000$ and $M_{\max} = 83$, obtaining the interpolation points reported in Fig. 10.4. The errors computed over a parameter test sample of 200 random points are instead reported in Fig. 10.3. We

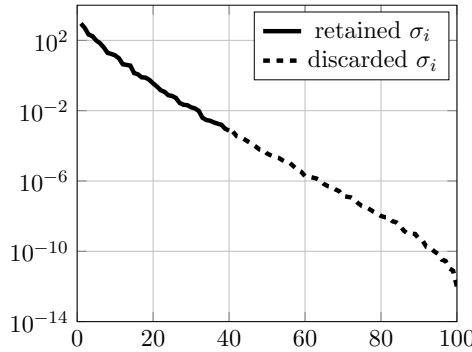


Fig. 10.1 DEIM approximation of $s(\mathbf{x}; \boldsymbol{\mu})$: decay of the singular values of the snapshot matrix

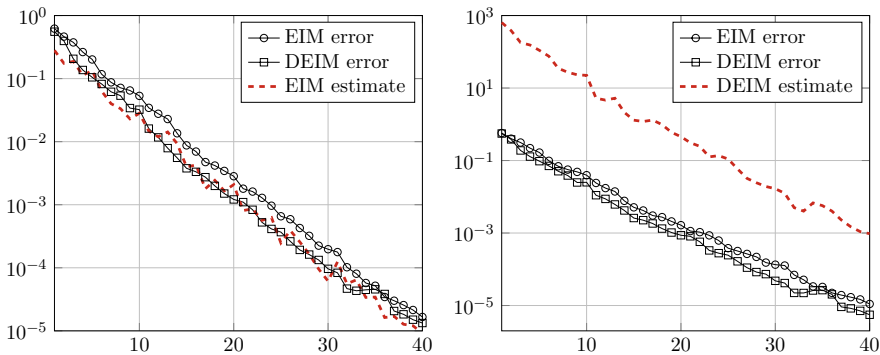


Fig. 10.2 Average relative error (with respect to M) between $s(\mathbf{x}; \boldsymbol{\mu})$ and its EIM and DEIM approximations; we report the ∞ -norm error (*left*) and the 2-norm error on the (*right*)

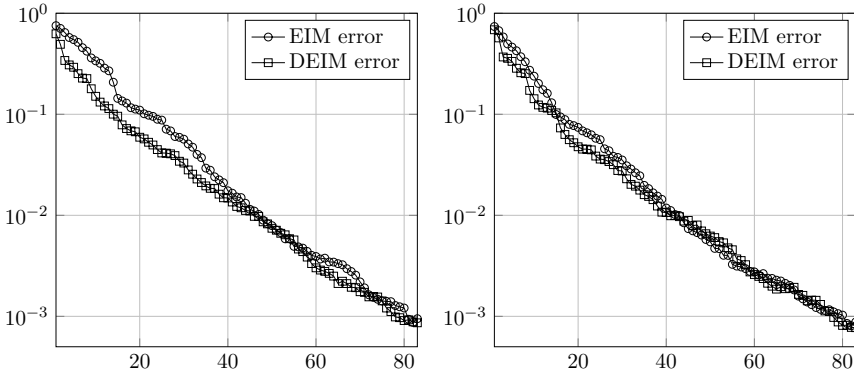


Fig. 10.3 Average relative error between $s(\mathbf{x}; \boldsymbol{\mu})$ and its EIM and DEIM approximations. We report the maximum norm error (*left*) and the 2-norm error (*right*)

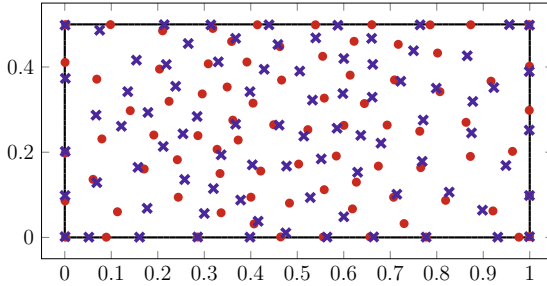


Fig. 10.4 DEIM (*red circles*) and EIM (*blue crosses*) interpolation points

notice in this case that a much larger (about four times) number M of terms has to be considered to achieve the same level of accuracy with respect to the previous test case.

10.5.2 (D)EIM-G-RB Approximation

We now construct an EIM-G-RB approximation to problem (8.32) following the procedure detailed in Sect. 10.4. With respect to the formulation of Sect. 8.4, we consider $\mu_2 \in [0, 2\pi]$, $\mu_3 \in [0.2, 0.8]$, $\mu_4 \in [0.15, 0.35]$, while $\mu_1 = 0.03$ and $\mu_5 = 0.25$ are fixed. In a pre-processing phase, we run the EIM using a training sample of size $|\mathcal{S}_s^{EIM}| = 1000$ and a tolerance $\varepsilon_{EIM} = 10^{-3}$, which yields an affine expansion of the right-hand side featuring $Q_f = 30$ terms. Then, we employ the POD algorithm to construct the RB space: starting from $n_s = 150$ snapshots, we retain the first $N = 85$ POD modes. The online convergence of the error between the EIM-G-RB and the high-fidelity approximation for different values of Q_f , evaluated over a test sample of 100 parameter values, is reported in Fig. 10.5. Some EIM-G-RB solutions of the problem, corresponding to different parameter values, are displayed in Fig. 10.6.

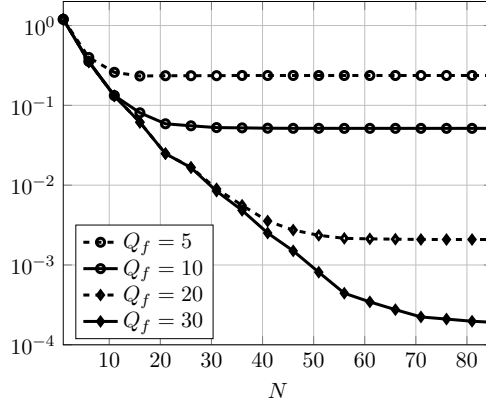


Fig. 10.5 Online convergence (with respect to N) of the error between the EIM-G-RB and the high-fidelity approximation for different values of Q_f , averaged over a test sample of 100 parameter values

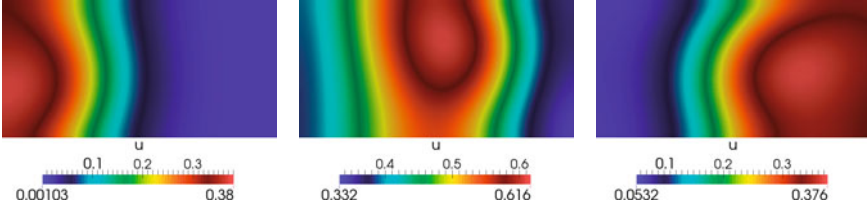


Fig. 10.6 EIM-G-RB solutions of problem (8.32) for different parameter values

We then consider the more challenging case where $\mu_5 = 0.08$, i.e. the Gaussian source term is more localized in space (see Fig. 10.7). Running the DEIM with $n_s^{DEIM} = 250$ and a tolerance $\varepsilon_{DEIM} = 10^{-4}$, we obtain an affine decomposition with $Q_f = 105$ terms. Concerning the construction of the RB space, we run the POD algorithm: starting from $n_s = 200$ snapshots, POD retains the first $N = 157$ modes when a tolerance $\varepsilon_{POD} = 10^{-4}$ is prescribed. As suggested by the error analysis reported in Fig. 10.8, in this case a large number of both DEIM terms and RB functions has to be considered to achieve a good accuracy with respect to the high-fidelity approximation.

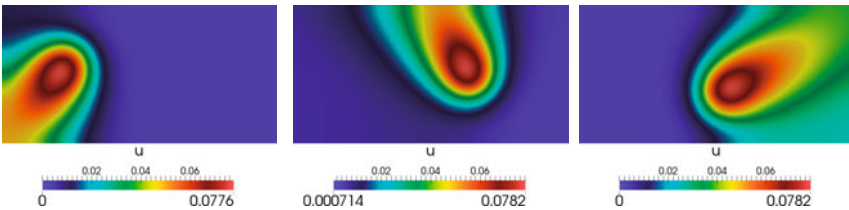


Fig. 10.7 DEIM-G-RB solutions of problem (8.32) for different parameter values

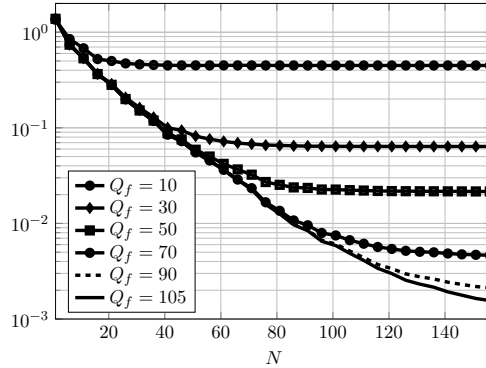


Fig. 10.8 Online convergence (with respect to N) of the error between the DEIM-G-RB and the high-fidelity approximation for different values of Q_f , averaged over a test sample of 150 parameter values

10.6 Mass Transfer in a Parametrized Domain: Results

We consider the mass transfer problem formulated in Sect. 8.5, by considering the following $P = 2$ parameters:

- $\mu_1 \in [0, 0.7]$ is the amplitude of the radial restriction defined by the geometric map (8.36);
- $\mu_2 \in [10^{-3}, 10^{-1}]$ is the diffusion coefficient such that $\tilde{\mathbf{k}} = \mu_2 \mathbf{I}$.

We end up with $\boldsymbol{\mu}_g = \mu_1$ and $\boldsymbol{\mu}_{ph} = \mu_2$. Moreover, we set $\tilde{s} = 0$, $\tilde{a}_0 = 0$, $\tilde{h} = 0$, $\tilde{g} = 8(1 - (x_1^2 + x_2^2))$, while the advection field $\tilde{\mathbf{b}}$ is given by (8.45); the latter is transformed onto the reference domain according to the Piola transformation (8.46).

For the high-fidelity approximation we use \mathbb{P}_1 finite elements built over a discretization of the domain made by tetrahedral elements, resulting in 21 324 vertices, 10 1139 triangles and a high-fidelity space V_h of dimension $N_h = 16 272$. In order to generate an approximate affine expansion for the problem at hand, we run EIM on the parametrized tensors (8.41)–(8.42) accounting for the geometric deformation: using a tolerance $\varepsilon_{\text{EIM}} = 10^{-5}$ we obtain $Q_a = Q_f = 12$.

Then, we employ the POD algorithm to construct the RB space: we retain $N = 50$ POD modes out of $n_s = 100$ snapshots (corresponding to a latin hypercube sampling of the parameter domains). The online evaluation of the resulting EIM-G-RB approximation takes only 22 ms, whereas the high-fidelity FE solution takes 2.5 s, thus yielding a computational speedup of about 110. Some EIM-G-RB solutions of the problem, corresponding to different parameter values, are displayed in Fig. 10.9. The online convergence of the error between the EIM-G-RB and the high-fidelity approximation for different values of Q_f , evaluated over a test sample of 100 parameter values, is reported in Fig. 10.10.

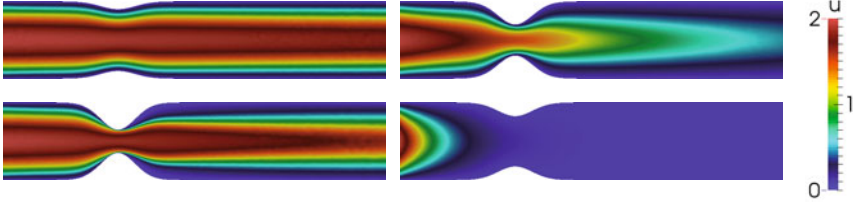


Fig. 10.9 EIM-G-RB solutions of problem (8.32) for different parameter values: $\boldsymbol{\mu} = (0.2, 10^{-3})$ (top left), $\boldsymbol{\mu} = (0.6, 10^{-2})$ (top right), $\boldsymbol{\mu} = (0.7, 10^{-2.7})$ (bottom left), $\boldsymbol{\mu} = (0.35, 10^{-1})$ (bottom right)

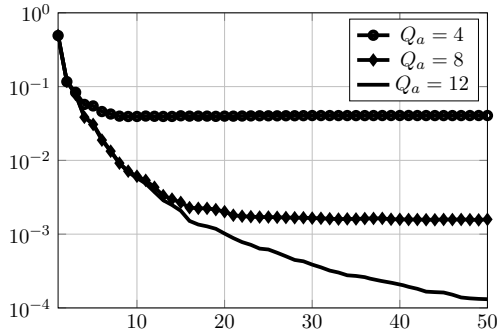


Fig. 10.10 Online convergence (with respect to N) of the error between the EIM-G-RB and the high-fidelity approximation for different values of Q_a , averaged over a test sample of 150 parameter values

10.7 Exercises

1. Suppose that $\dim(X_M) = M$ and that the matrix $\mathbb{B}_M \in \mathbb{R}^{M \times M}$ with entries $(\mathbb{B}_M)_{ij} = \rho_j(\mathbf{x}_i)$ is invertible. Show that $\mathcal{S}_M^{\mathbf{x}} v = v$ for any $v \in X_M$.
2. Show that the set of characteristic functions $\{l_m^M\}_{m=1}^M$ (10.9) is a basis for X_M and that the bases $\{\rho_m, m = 1, \dots, M\}$ and $\{l_m, m = 1, \dots, M\}$ are related by

$$\rho_i(\mathbf{x}) = \sum_{j=1}^M (\mathbb{B}_M)_{ji} l_j(\mathbf{x}), \quad i = 1, \dots, M. \quad (10.30)$$

Equivalently, by exploiting the fact that $l_i(\mathbf{x}_j) = \delta_{ij}$ for any $i, j = 1, \dots, M$, show that

$$l_i(\mathbf{x}) = \sum_{j=1}^M \rho_j(\mathbf{x}) (\mathbb{B}_M^{-1})_{ji}.$$

3. Exploiting (10.30), show that

$$|l_i(\mathbf{x})| \leq 1 + \sum_{k=i+1}^M |l_k(\mathbf{x})|, \quad i = 1, \dots, M-1;$$

then, recalling that $\|\rho_i(\cdot)\|_{L^\infty(\Omega)} \leq 1$ (why?) deduce that $|l_{M+1-i}(\mathbf{x})| \leq 1 + |l_M(\mathbf{x})| + \dots + |l_{M+2-i}(\mathbf{x})| \leq 2^{i-1}$, $i = 2, \dots, M$. Finally, show that

$$\Lambda_M = \sup_{\mathbf{x} \in \Omega} \sum_{i=1}^M |l_i(\mathbf{x})| \leq 2^M - 1.$$

4. Denote by $\varepsilon_M(\boldsymbol{\mu}) = \|g(\cdot; \boldsymbol{\mu}) - \mathcal{J}_M^{\mathbf{x}} g(\cdot; \boldsymbol{\mu})\|_{L^\infty(\Omega)}$ the interpolation error, by $\varepsilon_M^*(\boldsymbol{\mu}) = \inf_{g_M \in X_M} \|g(\cdot; \boldsymbol{\mu}) - g_M\|_{L^\infty(\Omega)} = \|g(\cdot; \boldsymbol{\mu}) - g_M^*(\cdot; \boldsymbol{\mu})\|_{L^\infty(\Omega)}$ the best fit of $g(\cdot; \boldsymbol{\mu})$, being $g_M^*(\cdot; \boldsymbol{\mu}) = \arg \inf_{g_M \in X_M} \|g(\cdot; \boldsymbol{\mu}) - g_M\|_{L^\infty(\Omega)}$.

a. Setting $e_M^*(\mathbf{x}; \boldsymbol{\mu}) = g(\mathbf{x}; \boldsymbol{\mu}) - g_M^*(\mathbf{x}; \boldsymbol{\mu})$, show that

$$\mathcal{J}_M^{\mathbf{x}} g(\cdot; \boldsymbol{\mu}) - g_M^*(\mathbf{x}; \boldsymbol{\mu}) = \sum_{i=1}^M e_M^*(\mathbf{t}^i; \boldsymbol{\mu}) l_i^M(\mathbf{x});$$

b. using the triangular inequality, show that

$$\varepsilon_M(\boldsymbol{\mu}) \leq \varepsilon_M^*(\boldsymbol{\mu}) + \|\mathcal{J}_M^{\mathbf{x}} g(\cdot; \boldsymbol{\mu}) - g_M^*(\cdot; \boldsymbol{\mu})\|_{L^\infty(\Omega)};$$

c. by using the relation at point a, show that

$$\varepsilon_M(\boldsymbol{\mu}) - \varepsilon_M^*(\boldsymbol{\mu}) \leq \max_{\mathbf{t}^i \in T_M} |e_M^*(\mathbf{t}^i; \boldsymbol{\mu})| \Lambda_M$$

and then derive (10.10) by exploiting $|e_M^*(\mathbf{t}^i; \boldsymbol{\mu})| \leq \varepsilon_M^*(\boldsymbol{\mu})$, $i = 1, \dots, M$.

5. Show that the estimate (10.28) holds also in the weakly coercive case, with $\beta_h(\boldsymbol{\mu})$ instead of $\alpha_h(\boldsymbol{\mu})$.

Chapter 11

Extension to Nonlinear Problems

The RB method is extended to the case of parametrized nonlinear PDEs. Examples are discussed concerning the Navier-Stokes equations and an elliptic semilinear equation. Both high-fidelity and RB approximations, as well as their interplay with Newton linearization, are analyzed, before considering in detail the case of Navier-Stokes equations. The underlying RB construction is essentially the same as for the previous linear PDEs, hence the focus is put on the characteristic aspects of the efficient treatment of nonlinear terms.

11.1 Parametrized Nonlinear PDEs

As usual, let $\mathcal{P} \subset \mathbb{R}^P$, $P \geq 1$ be the parameter space, $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$ a (reference) domain, $V = V(\Omega)$ a suitable Hilbert space, V' its dual. Moreover, we denote by $G: V \times \mathcal{P} \rightarrow V'$ a parametrized mapping representing a nonlinear PDE. In abstract form, the problem we focus on reads: given $\boldsymbol{\mu} \in \mathcal{P}$, find $u(\boldsymbol{\mu}) \in V$ such that

$$G(u(\boldsymbol{\mu}); \boldsymbol{\mu}) = 0 \quad \text{in } V'. \quad (11.1)$$

The weak formulation of problem (11.1) reads: find $u(\boldsymbol{\mu}) \in V$ such that

$$g(u(\boldsymbol{\mu}); v; \boldsymbol{\mu}) = 0 \quad \forall v \in V, \quad (11.2)$$

where the parametrized variational form $g(\cdot; \cdot; \boldsymbol{\mu}) : V \times V \rightarrow \mathbb{R}$ is defined as¹

$$g(w; v; \boldsymbol{\mu}) = \langle G(w; \boldsymbol{\mu}), v \rangle \quad \forall w, v \in V.$$

Let us now state some assumptions for the well-posedness of the parametrized nonlinear problem (11.1). We assume the mapping G to be continuously differentiable and denote by $D_u G(z, \boldsymbol{\mu}) : V \rightarrow V'$ and $D_{\boldsymbol{\mu}} G(z, \boldsymbol{\mu}) : \mathcal{P} \rightarrow V'$ its (partial) Fréchet derivatives at $(z, \boldsymbol{\mu}) \in \mathcal{P} \times V$.

¹ By the semicolon we separate the role of the unknown solution u (which shows up nonlinearly) from that of the test function v ; as usual, $\langle \cdot, \cdot \rangle$ denotes the duality pairing between V' and V .

Moreover, we denote by

$$dg[z](w, v; \boldsymbol{\mu}) = \langle D_u G(z; \boldsymbol{\mu}) w, v \rangle \quad \forall w, v \in V,$$

the partial Fréchet derivative of $g(z, \cdot; \boldsymbol{\mu})$ with respect to u at $z \in V$.

The following result holds; its proof is based on a straightforward application of the *Implicit Function Theorem* (see e.g. [267, 66]).

Proposition 11.1. *Let $G : V \times \mathcal{P} \rightarrow V'$ be a C^1 map and suppose that:*

1. $G(u_0, \boldsymbol{\mu}_0) = 0$ for some $\boldsymbol{\mu}_0 \in \mathcal{P}$, $u_0 \in V$;
2. $D_u G(u_0, \boldsymbol{\mu}_0) : V \rightarrow V'$ is bijective or, equivalently, $dg[u_0](\cdot, \cdot; \boldsymbol{\mu}_0)$ is continuous and inf-sup stable.

Then, there exist $r_0, r > 0$ and a unique $u(\boldsymbol{\mu}) \in B_r(u_0) \cap V$ such that

$$G(u(\boldsymbol{\mu}), \boldsymbol{\mu}) = 0 \quad \forall \boldsymbol{\mu} \in B_{r_0}(\boldsymbol{\mu}_0) \cap \mathcal{P}.$$

Furthermore, the map $\boldsymbol{\mu} \mapsto u(\boldsymbol{\mu})$ is continuously differentiable in a suitable open neighborhood of $\boldsymbol{\mu}_0$ and

$$\frac{\partial u(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = -(D_u G(u(\boldsymbol{\mu}), \boldsymbol{\mu}))^{-1} D_{\boldsymbol{\mu}} G(u(\boldsymbol{\mu}), \boldsymbol{\mu}). \quad (11.3)$$

In this case the solution $u(\boldsymbol{\mu})$ to problem (11.1) is called regular.

Proposition 11.1 ensures the existence of a *local branch of nonsingular solutions* to problem (11.1). The first assumption requires that a solution to problem (11.1) indeed exists (Banach Contraction theorem or the Leray-Schauder theorem may be used to prove this – see e.g. [236, Sect. 9.1]). Uniqueness is in general hard to prove, and may require further assumptions to be fulfilled, depending on the problem at hand. We show in this section two examples that fit in this abstract framework, namely (i) the Navier-Stokes equations and (ii) a semilinear elliptic equation: uniqueness is ensured in the former case provided an additional *small data* hypothesis is satisfied, whereas is always automatically fulfilled in the latter case.

We also highlight that the abstract linear parametrized problem (3.1), for which we have already shown a similar property (see Proposition 5.10), fits into this framework provided we define

$$G(u; \boldsymbol{\mu}) = L(\boldsymbol{\mu})u - F(\boldsymbol{\mu}),$$

or, equivalently,

$$g(u; v; \boldsymbol{\mu}) = a(u, v; \boldsymbol{\mu}) - f(v; \boldsymbol{\mu}).$$

In this case, the Fréchet derivative with respect to u is constant, as

$$dg[u](w, v; \boldsymbol{\mu}) = a(w, v; \boldsymbol{\mu}) \quad \forall u, v, w \in V.$$

Thus, Proposition 11.1 allows to extend the results of Sect. 5.3.2 about the differentiability of the solution map. Note that assumption 2 in Proposition 11.1 is the nonlinear analogue of conditions (3.5), (3.7) for linear problems.

11.1.1 Navier-Stokes Equations

We recall from Sects. 2.3.3 and 8.7.2 that the weak formulation of the parametrized Navier-Stokes equations reads: find $(\mathbf{u}(\boldsymbol{\mu}), p(\boldsymbol{\mu})) \in X \times Q$ such that

$$\begin{cases} \tilde{d}(\mathbf{u}(\boldsymbol{\mu}), \mathbf{v}; \boldsymbol{\mu}) + c(\mathbf{u}(\boldsymbol{\mu}), \mathbf{u}(\boldsymbol{\mu}), \mathbf{v}; \boldsymbol{\mu}) + b(\mathbf{v}, p(\boldsymbol{\mu}); \boldsymbol{\mu}) = f_1(\mathbf{v}; \boldsymbol{\mu}) & \forall \mathbf{v} \in X \\ b(\mathbf{u}(\boldsymbol{\mu}), q; \boldsymbol{\mu}) = f_2(q; \boldsymbol{\mu}) & \forall q \in Q, \end{cases} \quad (11.4)$$

where $X = [H_{T_D}^1(\Omega)]^d$ and $Q = L^2(\Omega)$; the bilinear forms $\tilde{d}(\cdot, \cdot; \boldsymbol{\mu})$, $b(\cdot, \cdot; \boldsymbol{\mu})$ and the trilinear form $c(\cdot, \cdot, \cdot; \boldsymbol{\mu})$ have been defined in (8.59)–(8.62); see instead (8.65)–(8.66) for the definition of the functionals $f_1(\cdot; \boldsymbol{\mu})$ and $f_2(\cdot; \boldsymbol{\mu})$, respectively.

In this case we can define $V = X \times Q$ and the parametrized mapping $G: V \times \mathcal{P} \rightarrow V'$ (and the associated variational form $g(\cdot; \cdot; \boldsymbol{\mu})$) as

$$\begin{aligned} \langle G(w; \boldsymbol{\mu}), v \rangle = g(w; v; \boldsymbol{\mu}) &= \tilde{d}(\mathbf{w}, \mathbf{v}; \boldsymbol{\mu}) + c(\mathbf{w}, \mathbf{w}, \mathbf{v}; \boldsymbol{\mu}) + b(\mathbf{w}, r; \boldsymbol{\mu}) \\ &\quad + b(\mathbf{w}, q; \boldsymbol{\mu}) - f_1(\mathbf{v}; \boldsymbol{\mu}) - f_2(q; \boldsymbol{\mu}) \end{aligned} \quad (11.5)$$

for all $w = (\mathbf{w}, r) \in V$ and $v = (\mathbf{v}, q) \in V$, with $V = X \times Q$. Its Fréchet derivative with respect to w at $z = (\mathbf{z}, s) \in V$ is given by

$$\begin{aligned} dg[z](w, v; \boldsymbol{\mu}) &= \tilde{d}(\mathbf{w}, \mathbf{v}; \boldsymbol{\mu}) + c(\mathbf{w}, \mathbf{z}, \mathbf{v}; \boldsymbol{\mu}) + c(\mathbf{z}, \mathbf{w}, \mathbf{v}; \boldsymbol{\mu}) \\ &\quad + b(\mathbf{w}, r; \boldsymbol{\mu}) + b(\mathbf{w}, q; \boldsymbol{\mu}). \end{aligned}$$

Note that $dg[z](w, v; \boldsymbol{\mu})$ is independent from s since the Navier-Stokes equations are linear with respect to the pressure variable. The trilinear form $c(\cdot, \cdot, \cdot; \boldsymbol{\mu})$ is such that

$$c(\mathbf{u}, \mathbf{w}, \mathbf{v}; \boldsymbol{\mu}) \leq \rho^2 M_c(\boldsymbol{\mu}) \|\mathbf{u}\|_X \|\mathbf{w}\|_X \|\mathbf{v}\|_X \quad \forall \boldsymbol{\mu} \in \mathcal{P} \quad (11.6)$$

being $M_c(\boldsymbol{\mu})$ a $\boldsymbol{\mu}$ -dependent factor depending on the affine decomposition of the trilinear form and

$$\rho^2 = \sup_{v \in H^1(\Omega)} \frac{\|v\|_{L^4(\Omega)}^2}{\|v\|_{H^1(\Omega)}^2}. \quad (11.7)$$

Since $H^1(\Omega)$ is embedded into $L^4(\Omega)$ (with continuous embedding, see Definition A.5), ρ represent the Sobolev embedding constant. Moreover $c(\mathbf{u}, \mathbf{u}, \mathbf{u}; \boldsymbol{\mu}) = 0$ provided that Dirichlet conditions are homogeneous and Neumann conditions are imposed on the outflow boundaries. See, e.g., [115, Chap. IV, Theorem 2.3] for the extension to the case of nonhomogeneous Dirichlet conditions.

Existence of the solution to problem (11.4) is ensured thanks to the Leray-Schauder theorem provided the source term is in $L^2(\Omega)$ and Ω is a bounded Lipschitz domain, see e.g. [246, Chap. 2, Theorems 1.2, 1.6]. Under the following *small data* assumption

$$\frac{\rho^2 M_c(\boldsymbol{\mu}) \|f_1(\cdot; \boldsymbol{\mu})\|_{X'}}{(\alpha_d(\boldsymbol{\mu}))^2} < 1 \quad (11.8)$$

the solution to problem (11.4) is also unique. This result holds if homogeneous Dirichlet conditions are imposed, being $\alpha_d(\boldsymbol{\mu})$ the coercivity constant of $d(\cdot, \cdot; \boldsymbol{\mu})$; in this case $\bar{d}(\mathbf{u}(\boldsymbol{\mu}), \mathbf{v}; \boldsymbol{\mu}) = d(\mathbf{u}(\boldsymbol{\mu}), \mathbf{v}; \boldsymbol{\mu})$ and $f_1(\cdot; \boldsymbol{\mu})$ does not include any term related to the lifting function $\mathbf{r}_g \in [H^1(\Omega)]^d$, see Exercise 3. In the more general case of nonhomogeneous Dirichlet conditions, a similar result holds, provided (11.8) is replaced by the following condition

$$\frac{\rho^2 M_c(\boldsymbol{\mu}) \|f_1(\cdot; \boldsymbol{\mu})\|_{X'}}{(\alpha_d(\boldsymbol{\mu}) - \sigma(\mathbf{r}_g))^2} < 1, \quad \sigma(\mathbf{r}_g) = \sup_{\mathbf{v} \in X} \frac{c(\mathbf{v}, \mathbf{r}_g, \mathbf{v}; \boldsymbol{\mu})}{\|\mathbf{v}\|_X^2},$$

see, e.g. [115, Chap. IV, Theorem 2.4] for a detailed proof.

The solution to problem (11.4) is also regular in the sense of Proposition 11.1 by assuming the inf-sup stability of $dg[u(\boldsymbol{\mu})](\cdot, \cdot; \boldsymbol{\mu})$, see e.g. [49]. Moreover, it is possible to show (see Exercise 4) that for any $\boldsymbol{\mu} \in \mathcal{P}$ and for any $\mathbf{u}_1, \mathbf{u}_2 \in X$

$$\begin{aligned} |dg[u_1](w, v; \boldsymbol{\mu}) - dg[u_2](w, v; \boldsymbol{\mu})| &= |c(\mathbf{u}_1 - \mathbf{u}_2, \mathbf{w}, \mathbf{v}; \boldsymbol{\mu}) + c(\mathbf{w}, \mathbf{u}_1 - \mathbf{u}_2, \mathbf{w}, \mathbf{v}; \boldsymbol{\mu})| \\ &\leq 2\rho^2 M_c(\boldsymbol{\mu}) \|\mathbf{u}_1 - \mathbf{u}_2\|_X \|\mathbf{w}\|_X \|\mathbf{v}\|_X \quad \forall \mathbf{v}, \mathbf{w} \in X. \end{aligned} \quad (11.9)$$

11.1.2 A Semilinear Elliptic PDE

We consider the following parametrized semilinear elliptic boundary value problem:

$$\begin{cases} -\mu_1 \Delta u + \mu_2 u^3 = s & \text{in } \Omega \\ u = 0 & \text{on } \Gamma = \partial\Omega, \end{cases} \quad (11.10)$$

with $s \in L^2(\Omega)$ and $\mathcal{P} = [1, a] \times [1, b]$, $a, b > 1$. We define $V = H_0^1(\Omega)$ (endowed with the norm $\|\cdot\|_V = \|\cdot\|_{H^1(\Omega)}$) and the mapping $G : V \times \mathcal{P} \rightarrow V'$,

$$G(w; \boldsymbol{\mu}) = -\mu_1 \Delta w + \mu_2 w^3 - s \quad \forall w \in V,$$

or, equivalently, for all $w, v \in V$

$$\langle G(w; \boldsymbol{\mu}), v \rangle = g(w; v; \boldsymbol{\mu}) = \mu_1 \int_{\Omega} \nabla w \cdot \nabla v \, d\Omega + \mu_2 \int_{\Omega} w^3 v \, d\Omega - \int_{\Omega} s v \, d\Omega.$$

The existence of a solution is ensured thanks to the Leray-Schauder theorem, see e.g. [49, Theorem 3.1] for a complete proof. Uniqueness in this case is easy to prove: if u_1 and u_2 are two solutions to (11.10), then

$$\mu_1 \int_{\Omega} \nabla(u_1 - u_2) \cdot \nabla v \, d\Omega + \mu_2 \int_{\Omega} (u_1^3 - u_2^3) v \, d\Omega = 0 \quad \forall v \in V$$

so that, by choosing $v = u_1 - u_2$,

$$\mu_1 \int_{\Omega} |\nabla(u_1 - u_2)|^2 d\Omega + \mu_2 \int_{\Omega} (u_1 - u_2)^2 \left(\frac{u_1^2 + u_2^2}{2} + \frac{(u_1 + u_2)^2}{2} \right) d\Omega = 0$$

whence $u_1 = u_2$; in this case, for any $\boldsymbol{\mu} \in \mathcal{P}$ the solution is also unique without any further assumption. The Fréchet derivative of $g(w, v; \boldsymbol{\mu})$ with respect to w at $z \in V$ is

$$dg[z](w, v; \boldsymbol{\mu}) = \mu_1 \int_{\Omega} \nabla w \cdot \nabla v d\Omega + 3\mu_2 \int_{\Omega} z^2 w v d\Omega.$$

For any $z \in V$ and $\boldsymbol{\mu} \in \mathcal{P}$, $dg[z](\cdot, \cdot; \boldsymbol{\mu})$ is symmetric and coercive, since

$$dg[z](v, v; \boldsymbol{\mu}) = \mu_1 \int_{\Omega} \nabla v \cdot \nabla v d\Omega + 3\mu_2 \int_{\Omega} z^2 v^2 d\Omega \geq \|v\|_V^2 \quad \forall v \in V.$$

Hence, the solution $u(\boldsymbol{\mu})$ is regular in the sense of Proposition 11.1; this latter thus ensures the existence of a regular branch of nonsingular solutions.

11.2 High-Fidelity Approximation

For the high-fidelity approximation of (11.2) we introduce a suitable finite-dimensional subspace V_h of V and, similarly to Sect. 3.2, we seek $u_h(\boldsymbol{\mu}) \in V_h$ such that

$$g(u_h(\boldsymbol{\mu}); v_h; \boldsymbol{\mu}) = 0 \quad \forall v_h \in V_h. \quad (11.11)$$

Problem (11.11) can be equivalently written as: find $u_h(\boldsymbol{\mu}) \in V_h$ such that

$$\langle G(u_h(\boldsymbol{\mu}); \boldsymbol{\mu}), v_h \rangle = 0 \quad \forall v_h \in V_h.$$

For the discrete problem (11.11) to be well-posed we assume $dg[u_h(\boldsymbol{\mu})](\cdot, \cdot; \boldsymbol{\mu})$ to be inf-sup stable, i.e. that there exists a constant $\beta_{0,h} > 0$ such that

$$\beta_h(\boldsymbol{\mu}) = \inf_{w_h \in V_h} \sup_{v_h \in V_h} \frac{dg[u_h(\boldsymbol{\mu})](w_h, v_h; \boldsymbol{\mu})}{\|w_h\|_V \|v_h\|_V} \geq \beta_{0,h} \quad \forall \boldsymbol{\mu} \in \mathcal{P}, \quad (11.12)$$

and continuous, i.e. that there exists a positive constant $\gamma_{0,h} < \infty$ such that

$$\gamma_h(\boldsymbol{\mu}) = \sup_{v_h \in V_h} \sup_{w_h \in V_h} \frac{dg[u_h(\boldsymbol{\mu})](w_h, v_h; \boldsymbol{\mu})}{\|w_h\|_V \|v_h\|_V} \leq \gamma_{0,h} \quad \forall \boldsymbol{\mu} \in \mathcal{P}. \quad (11.13)$$

A solution $u_h(\boldsymbol{\mu})$ to problem (11.11) is said to be *regular* if it fulfills (11.12)-(11.13).

11.2.1 Newton's Method

Solving problem (11.11) requires nonlinear iterations with a linearized problem being solved at every step. Newton's method is a very natural approach: given $\boldsymbol{\mu} \in \mathcal{P}$ and an initial guess $u_h^0(\boldsymbol{\mu}) \in V_h$, for $k = 0, 1, \dots$ until convergence, we seek $\delta u_h \in V_h$ such that

$$dg[u_h^k(\boldsymbol{\mu})](\delta u_h, v_h; \boldsymbol{\mu}) = -g(u_h^k(\boldsymbol{\mu}); v_h; \boldsymbol{\mu}) \quad \forall v_h \in V_h, \quad (11.14)$$

and then set $u_h^{k+1}(\boldsymbol{\mu}) = u_h^k(\boldsymbol{\mu}) + \delta u_h$.

Newton's method is quadratically convergent provided $dg[u_h(\boldsymbol{\mu})](\cdot, \cdot; \boldsymbol{\mu})$ is locally Lipschitz continuous and u_h^0 is sufficiently close to $u_h(\boldsymbol{\mu})$.

Theorem 11.1. *Let $u_h(\boldsymbol{\mu})$ be a regular solution of (11.11) and $dg[u_h(\boldsymbol{\mu})](\cdot, \cdot; \boldsymbol{\mu})$ locally Lipschitz continuous at $u_h(\boldsymbol{\mu})$, i.e. there exist $\varepsilon(\boldsymbol{\mu}) > 0$ and $K_h(\boldsymbol{\mu}) > 0$ such that*

$$\|dg[u_h(\boldsymbol{\mu})](\cdot, \cdot; \boldsymbol{\mu}) - dg[v_h](\cdot, \cdot; \boldsymbol{\mu})\|_{\mathcal{L}(V_h, V_h')} \leq K_h(\boldsymbol{\mu}) \|u_h(\boldsymbol{\mu}) - v_h\|_V,$$

for all $v_h \in B_{\varepsilon(\boldsymbol{\mu})}(u_h(\boldsymbol{\mu})) = \{w_h \in V_h : \|u_h(\boldsymbol{\mu}) - w_h\|_V \leq \varepsilon(\boldsymbol{\mu})\}$. Then there exists $\delta > 0$ such that if $\|u_h^0 - u_h(\boldsymbol{\mu})\| \leq \delta$, the Newton sequence $\{u_h^k\}$ is well-defined and converges to $u_h(\boldsymbol{\mu})$. Furthermore, for some constant M with $M\delta < 1$, we have the error bound

$$\|u_h^{k+1} - u_h(\boldsymbol{\mu})\|_V \leq M \|u_h^k - u_h(\boldsymbol{\mu})\|_V^2, \quad k \geq 0.$$

This local convergence theorem (see, e.g., [267, Chap. 5] for the proof) regretfully requires the existence of a solution $u_h(\boldsymbol{\mu})$ in advance. The Kantorovich theorem (see, e.g., [267, 66]) overcomes this drawback by establishing the existence of $u_h(\boldsymbol{\mu})$ simply on the basis of the knowledge of $g(u_h^0, \cdot; \boldsymbol{\mu})$ and $dg[u_h^0](\cdot, \cdot; \boldsymbol{\mu})$.

Theorem 11.2 (Kantorovich). *For a fixed $\boldsymbol{\mu} \in \mathcal{P}$, let $u_h^0 \in V_h$ be a given initial guess such that:*

i) $\varepsilon = \|g(u_h^0, \cdot; \boldsymbol{\mu})\|_{V_h'} < \infty$ and $dg[u_h^0](\cdot, \cdot; \boldsymbol{\mu})$ is continuous and inf-sup stable with

$$a^{-1} = \inf_{w_h \in V_h} \sup_{v_h \in V_h} \frac{dg[u_h^0](w_h, v_h; \boldsymbol{\mu})}{\|w_h\|_V \|v_h\|_V} > 0; \quad (11.15)$$

ii) there exists a positive constant K such that

$$\|dg[w_h](\cdot, \cdot; \boldsymbol{\mu}) - dg[v_h](\cdot, \cdot; \boldsymbol{\mu})\|_{\mathcal{L}(V_h, V_h')} \leq K \|w_h - v_h\|_V \quad \forall v_h, w_h \in B_r(u_h^0)$$

with $r = 1/(aK)$ and

$$0 < 2\varepsilon K a^2 \leq 1. \quad (11.16)$$

Define two positive numbers

$$r_- = \frac{1 - \sqrt{1 - 2\varepsilon K a^2}}{aK}, \quad r_+ = \frac{1 + \sqrt{1 - 2\varepsilon K a^2}}{aK}.$$

Then

1. problem (11.11) has a solution $u_h(\boldsymbol{\mu}) \in \overline{B}_{r_-}(u_h^0)$. Moreover, if

$$\|dg[w_h](\cdot, \cdot; \boldsymbol{\mu}) - dg[v_h](\cdot, \cdot; \boldsymbol{\mu})\|_{\mathcal{L}(V_h, V_h')} \leq K \|w_h - v_h\|_V \quad \forall v_h, w_h \in B_{r_+}(u_h^0),$$

this solution is unique in $\overline{B}_{r_+}(u_h^0)$;

2. the Newton's sequence $\{u_h^k(\boldsymbol{\mu})\}$ converges to $u_h(\boldsymbol{\mu})$ and the following error bound holds for each $k \geq 0$

$$\|u_h^k - u_h(\boldsymbol{\mu})\|_V \leq \frac{r}{2^k} \left(\frac{r_-}{r} \right)^{2^k}.$$

11.2.2 Algebraic Formulation

The discrete problem (11.11) is equivalent to the solution of a nonlinear system of N_h equations. Indeed, denoting by $\{\boldsymbol{\varphi}^j\}_{j=1}^{N_h}$ a basis for V_h as in Sect. 2.4.2, we set

$$u_h(\boldsymbol{\mu}) = \sum_{j=1}^{N_h} u_h^{(j)}(\boldsymbol{\mu}) \boldsymbol{\varphi}^j$$

and denote by $\mathbf{u}_h(\boldsymbol{\mu})$ the vector having as components the unknown coefficients $u_h^{(j)}(\boldsymbol{\mu})$. Then, (11.11) is equivalent to: find $\mathbf{u}_h(\boldsymbol{\mu}) \in \mathbb{R}^{N_h}$ such that

$$g\left(\sum_{j=1}^{N_h} u_h^{(j)}(\boldsymbol{\mu}) \boldsymbol{\varphi}^j; \boldsymbol{\varphi}^i; \boldsymbol{\mu}\right) = 0 \quad \forall i = 1, \dots, N_h,$$

that is

$$\mathbf{G}_h(\mathbf{u}_h(\boldsymbol{\mu}); \boldsymbol{\mu}) = \mathbf{0}, \quad (11.17)$$

where the *residual vector* $\mathbf{G}_h(\cdot; \boldsymbol{\mu}) \in \mathbb{R}^{N_h}$ is defined as

$$(\mathbf{G}_h(\mathbf{u}_h(\boldsymbol{\mu}); \boldsymbol{\mu}))_i = g(u_h(\boldsymbol{\mu}); \boldsymbol{\varphi}^i; \boldsymbol{\mu}), \quad i = 1, \dots, N_h.$$

The k -th step (11.14) of the Newton method can be written in matrix form as follows: find $\delta \mathbf{u}_h \in \mathbb{R}^{N_h}$ such that

$$\mathbb{J}_h(\mathbf{u}_h^k; \boldsymbol{\mu}) \delta \mathbf{u}_h = -\mathbf{G}_h(\mathbf{u}_h^k; \boldsymbol{\mu}), \quad (11.18)$$

where the *Jacobian matrix* $\mathbb{J}_h(\mathbf{u}_h^k; \boldsymbol{\mu}) \in \mathbb{R}^{N_h \times N_h}$ is defined as

$$(\mathbb{J}_h(\mathbf{u}_h^k; \boldsymbol{\mu}))_{ij} = dg[u_h^k](\boldsymbol{\varphi}^j, \boldsymbol{\varphi}^i; \boldsymbol{\mu}), \quad i, j = 1, \dots, N_h.$$

Note that the inf-sup condition (11.15) and Kantorovich theorem imply that for every $k \geq 0$ the matrix $\mathbb{J}_h(\mathbf{u}_h^k; \boldsymbol{\mu})$ is nonsingular. The basic steps of the Newton method for the solution of (11.17) are summarized in Algorithm 11.1.

Algorithm 11.1 Newton's method for the solution of the (high-fidelity) algebraic nonlinear problem (11.17)

Input: $\boldsymbol{\mu} \in \mathcal{P}$, tolerance $\delta > 0$, max number of iterations K , $\mathbf{u}_h^0 \in \mathbb{R}^{N_h}$

Output: $\mathbf{u}_h(\boldsymbol{\mu})$

```

1:  $k = 0$ 
2: while  $k \leq K$  and  $\|\mathbf{G}_h(\mathbf{u}_h^k; \boldsymbol{\mu})\| > \delta$ 
3:   solve  $\mathbb{J}_h(\mathbf{u}_h^k; \boldsymbol{\mu}) \delta \mathbf{u}_h = -\mathbf{G}_h(\mathbf{u}_h^k; \boldsymbol{\mu})$ 
4:   set  $\mathbf{u}_h^{k+1} = \mathbf{u}_h^k + \delta \mathbf{u}_h^k$ 
5:    $k = k + 1$ 
6: end while
7:  $\mathbf{u}_h(\boldsymbol{\mu}) = \mathbf{u}_h^k$ 

```

Each evaluation of the map $\boldsymbol{\mu} \mapsto \mathbf{u}_h(\boldsymbol{\mu})$ requires to assemble and solve K linear systems of dimension N_h , thus further motivating the need of a reduced-order approximation.

11.3 Reduced Basis Approximation

The backbone of a RB approximation to nonlinear parametrized problems is the same as that of the linear case (see Sect. 3.3). Indeed, we build the RB problem by means of a suitable projection over a subspace V_N of V_h of reduced dimension $N \ll N_h$, which is generated from a set of snapshots solutions of the high-fidelity problem (11.1) by means of either the POD or the greedy algorithm. However, when dealing with nonlinear problems we have to face two main challenges: (i) the setup of an efficient offline-online computational procedure and (ii) the derivation of suitable estimates for the error between the high-fidelity solution and the RB one.

As in the linear case, we obtain the reduced problem by seeking an approximate solution $u_N(\boldsymbol{\mu}) \in V_N$ and enforcing the orthogonality of the residual of (11.11) to N linearly independent functions of V_h that span a subspace W_N (of dimension N):

find $u_N(\boldsymbol{\mu}) \in V_N$ such that

$$\langle G(u_N(\boldsymbol{\mu}); \boldsymbol{\mu}), w_N \rangle = 0 \quad \forall w_N \in W_N \quad (11.19)$$

or, equivalently,

find $u_N(\boldsymbol{\mu}) \in V_N$ such that

$$g(u_N(\boldsymbol{\mu}); w_N; \boldsymbol{\mu}) = 0 \quad \forall w_N \in W_N. \quad (11.20)$$

For every $\boldsymbol{\mu} \in \mathcal{P}$, the RB problem (11.20) consists of a system of N nonlinear equations, which can be solved by means of Newton's method: given an initial guess $u_N^0(\boldsymbol{\mu}) \in V_N$, for $k = 0, 1, \dots$ until convergence, we seek $\delta u_N \in V_N$ such that

$$dg[u_N^k(\boldsymbol{\mu})](\delta u_N, w_N; \boldsymbol{\mu}) = -g(u_N^k(\boldsymbol{\mu}); w_N; \boldsymbol{\mu}) \quad \forall w_N \in W_N, \quad (11.21)$$

and then set $u_N^{k+1}(\boldsymbol{\mu}) = u_N^k(\boldsymbol{\mu}) + \delta u_N$.

The existence of a unique solution of the reduced problem (11.20) as well as the convergence of Newton's method follow from Kantorovich theorem by assuming that $dg[u_N^0(\boldsymbol{\mu})](\cdot, \cdot; \boldsymbol{\mu})$ is locally Lipschitz continuous and inf-sup stable over $V_N \times W_N$. Before discussing the choice of the subspace W_N , we provide the algebraic formulation of the RB problem (11.20).

11.3.1 Algebraic Formulation

As customarily, we denote by $\{\zeta_m, m = 1, \dots, N\}$ an orthonormal basis for V_N and by $\{\eta_m, m = 1, \dots, N\}$ one for W_N , see (3.22), (3.41), respectively. Since $u_N(\boldsymbol{\mu}) \in V_N$, we can expand it as

$$u_N(\boldsymbol{\mu}) = \sum_{m=1}^N u_N^{(m)}(\boldsymbol{\mu}) \zeta_m \quad (11.22)$$

and denote by $\mathbf{u}_N(\boldsymbol{\mu}) \in \mathbb{R}^N$ the vector whose components are the unknown coefficients $u_N^{(m)}(\boldsymbol{\mu})$. Inserting (11.22) into (11.21) and then choosing $w_N = \eta_n, 1 \leq n \leq N$, we obtain a set of N nonlinear equations for the RB coefficients $u_N^{(m)}(\boldsymbol{\mu})$

$$g\left(\sum_{m=1}^N u_N^{(m)}(\boldsymbol{\mu}) \zeta_m; \eta_n; \boldsymbol{\mu}\right) = 0 \quad \forall n = 1, \dots, N. \quad (11.23)$$

We introduce the transformation matrix $\mathbb{V} \in \mathbb{R}^{N_h \times N}$, see (4.7), and denote by $\mathbf{G}_N(\cdot; \boldsymbol{\mu}) \in \mathbb{R}^N$ the reduced residual vector defined as

$$(\mathbf{G}_N(\mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu}))_n = g\left(\sum_{m=1}^N u_N^{(m)}(\boldsymbol{\mu}) \zeta_m; \eta_n; \boldsymbol{\mu}\right), \quad n = 1, \dots, N.$$

Then, (11.23) is equivalent to the nonlinear system

$$\mathbf{G}_N(\mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu}) = \mathbf{0}. \quad (11.24)$$

By means of the transformation matrix $\mathbb{W} \in \mathbb{R}^{N_h \times N}$, see (4.21), we obtain that

$$\mathbf{G}_N(\mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu}) = \mathbb{W}^T \mathbf{G}_h(\mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu}). \quad (11.25)$$

Indeed,

$$\begin{aligned}
 (\mathbf{G}_N(\mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu}))_n &= g\left(\sum_{m=1}^N u_N^{(m)}(\boldsymbol{\mu}) \zeta_m; \sum_{r=1}^{N_h} \eta_n^{(r)} \varphi^r; \boldsymbol{\mu}\right) \\
 &= \sum_{r=1}^{N_h} \eta_n^{(r)} g\left(\sum_{m=1}^N u_N^{(m)}(\boldsymbol{\mu}) \zeta_m; \varphi^r; \boldsymbol{\mu}\right) \\
 &= \sum_{r=1}^{N_h} \mathbb{W}_{rn} g\left(\sum_{m=1}^N u_N^{(m)}(\boldsymbol{\mu}) \zeta_m; \varphi^r; \boldsymbol{\mu}\right) = (\mathbb{W}^T \mathbf{G}_h(\mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu}))_n.
 \end{aligned}$$

Proceeding in a similar way (see also Sect. 4.1 and Exercise 5), we can show that the Newton step (11.21) is equivalent to the following linear system of dimension N

$$\mathbb{J}_N(\mathbf{u}_N^k; \boldsymbol{\mu}) \delta \mathbf{u}_N = -\mathbf{G}_N(\mathbf{u}_N^k; \boldsymbol{\mu}), \quad (11.26)$$

where the reduced Jacobian matrix $\mathbb{J}_N(\cdot; \boldsymbol{\mu})$ is given by

$$\mathbb{J}_N(\mathbf{u}_N^k; \boldsymbol{\mu}) = \mathbb{W}^T \mathbb{J}_h(\mathbb{V} \mathbf{u}_N^k; \boldsymbol{\mu}) \mathbb{V}. \quad (11.27)$$

11.3.2 Galerkin Projection

The G-RB approximation for linear problems can be extended to the nonlinear case by choosing $\mathbb{W} = \mathbb{V}$, yielding the following linear system to be solved at each Newton step

$$\mathbb{V}^T \mathbb{J}_h(\mathbb{V} \mathbf{u}_N^k(\boldsymbol{\mu}); \boldsymbol{\mu}) \mathbb{V} \delta \mathbf{u}_N(\boldsymbol{\mu}) = -\mathbb{V}^T \mathbf{G}_h(\mathbb{V} \mathbf{u}_N^k(\boldsymbol{\mu}); \boldsymbol{\mu}). \quad (11.28)$$

If, for a given $\mathbf{u}_N^k \in \mathbb{R}^N$, the Jacobian matrix $\mathbb{J}_h(\mathbb{V} \mathbf{u}_N^k; \boldsymbol{\mu})$ is symmetric and positive definite, the increment $\delta \mathbf{u}_N$ satisfies the following residual minimization property (see (4.16))

$$\delta \mathbf{u}_N = \arg \min_{\mathbf{p} \in \mathbb{R}^N} \|\mathbb{J}_h(\mathbb{V} \mathbf{u}_N^k(\boldsymbol{\mu}); \boldsymbol{\mu}) \mathbb{V} \mathbf{p} + \mathbf{G}_h(\mathbb{V} \mathbf{u}_N^k(\boldsymbol{\mu}); \boldsymbol{\mu})\|_{(\mathbb{J}_h(\mathbb{V} \mathbf{u}_N^k(\boldsymbol{\mu}); \boldsymbol{\mu}))^{-1}}^2.$$

Remark 11.1. Requiring the Jacobian matrix \mathbb{J}_h to be symmetric positive definite is equivalent to require $dg[\cdot](\cdot, \cdot; \boldsymbol{\mu})$ to be symmetric and coercive. This is case, for instance, for the semilinear elliptic problem (11.10). •

11.3.3 LS-RB: Newton then Least-Squares

If the high-fidelity Jacobian matrix is nonsymmetric or indefinite, a nonsingular reduced Jacobian (11.27) can be achieved by adopting a least-squares approach, as in Sects. 3.3.2 and 3.4.2.

In this case, at each Newton step we require the solution $\delta \mathbf{u}_N$ of (11.26) to satisfy the following residual minimization property (see (4.17))

$$\delta \mathbf{u}_N = \arg \min_{\mathbf{p} \in \mathbb{R}^N} \|\mathbb{J}_h(\mathbb{V} \mathbf{u}_N^k(\boldsymbol{\mu}); \boldsymbol{\mu}) \mathbb{V} \mathbf{p} + \mathbf{G}_h(\mathbb{V} \mathbf{u}_N^k(\boldsymbol{\mu}); \boldsymbol{\mu})\|_{\mathbb{X}^{-1}}^2. \quad (11.29)$$

The stationarity condition for the quadratic functional in (11.29) yields the following least-squares RB method (omitting the $\boldsymbol{\mu}$ -dependence of \mathbf{u}_N^k)

$$\mathbb{V}^T \mathbb{J}_h^T(\mathbb{V} \mathbf{u}_N^k; \boldsymbol{\mu}) \mathbb{X}^{-1} \mathbb{J}_h(\mathbb{V} \mathbf{u}_N^k; \boldsymbol{\mu}) \mathbb{V} \delta \mathbf{u}_N = -\mathbb{V}^T \mathbb{J}_h^T(\mathbb{V} \mathbf{u}_N^k; \boldsymbol{\mu}) \mathbb{X}^{-1} \mathbf{G}_h(\mathbb{V} \mathbf{u}_N^k; \boldsymbol{\mu}), \quad (11.30)$$

where we recognize that

$$\mathbb{W} = \mathbb{W}_{\boldsymbol{\mu},k} = \mathbb{X}^{-1} \mathbb{J}_h(\mathbb{V} \mathbf{u}_N^k(\boldsymbol{\mu}); \boldsymbol{\mu}) \mathbb{V}. \quad (11.31)$$

Note that $\mathbb{W}_{\boldsymbol{\mu},k}$ not only depends on the parameters as in the linear case (see Sect. 3.3), but also changes from one Newton iteration to another.

11.3.4 LS-RB Revisited: Least-Squares then Gauss-Newton

Following a different approach, we now seek an approximate solution $\mathbb{V} \mathbf{u}_N$ of (11.17) which satisfies the following nonlinear least-squares problem

$$\mathbf{u}_N(\boldsymbol{\mu}) = \arg \min_{\mathbf{x} \in \mathbb{R}^N} \|\mathbf{G}_h(\mathbb{V} \mathbf{x}; \boldsymbol{\mu})\|_{\mathbb{X}^{-1}}^2.$$

First-order optimality conditions yields the following nonlinear problem

$$\mathbb{V}^T \mathbb{J}_h^T(\mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu}) \mathbb{X}^{-1} \mathbf{G}_h(\mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu}) = \mathbf{0},$$

which is equivalent to

$$\mathbb{W}_{\boldsymbol{\mu}}^T \mathbf{G}_h(\mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu}) = \mathbf{0}, \quad (11.32)$$

with

$$\mathbb{W}_{\boldsymbol{\mu}} = \mathbb{X}^{-1} \mathbb{J}_h(\mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu}) \mathbb{V}.$$

Since (11.32) is still a nonlinear system, we may now apply Newton method: at the generic step k we seek $\delta \mathbf{u}_N \in \mathbb{R}^N$ such that (omitting the $\boldsymbol{\mu}$ -dependence of \mathbf{u}_N^k)

$$\underbrace{\left[\mathbb{W}_{\boldsymbol{\mu},k}^T \mathbb{J}_h(\mathbb{V} \mathbf{u}_N^k; \boldsymbol{\mu}) \mathbb{V} + d\mathbb{W}_{\boldsymbol{\mu},k}^T \mathbf{G}_h(\mathbb{V} \mathbf{u}_N^k; \boldsymbol{\mu}) \right]}_{\tilde{\mathbb{J}}_N(\mathbf{u}_N^k; \boldsymbol{\mu})} \delta \mathbf{u}_N = -\mathbb{W}_{\boldsymbol{\mu},k}^T \mathbf{G}_h(\mathbb{V} \mathbf{u}_N^k; \boldsymbol{\mu}), \quad (11.33)$$

where $\mathbb{W}_{\boldsymbol{\mu},k}$ is given by (11.31) and $d\mathbb{W}_{\boldsymbol{\mu},k}$ is a third order tensor involving the Hessian of the residual vector $\mathbf{G}_h(\mathbb{V} \mathbf{u}_N^k(\boldsymbol{\mu}); \boldsymbol{\mu})$. This can be seen as a *least-squares then Newton* approach. As $\mathbf{G}_h(\mathbb{V} \mathbf{u}_N^k(\boldsymbol{\mu}); \boldsymbol{\mu})$ is usually not available and too costly to compute, an alternative approach [54, 57] is to rely on Gauss-Newton method. This latter (see, e.g., [200]) approximates the matrix $\mathbb{J}_N(\cdot; \boldsymbol{\mu})$ by disregarding the

term involving second derivatives of the residual and therefore reduces (11.33) to problem (11.30).

Following one of the approaches described so far is highly problem-dependent, and always entails a trade-off between accuracy and computational complexity. We point out that the reduced problem (11.28) obtained with the G-RB method is not necessarily well-posed, unless the high-fidelity Jacobian matrix \mathbb{J}_h is always symmetric positive definite; indeed, in several cases this assumption is not fulfilled. Nevertheless, provided suitable RB spaces are built, also a Galerkin projection may yield a well-posed reduced problem, see e.g. the Navier-Stokes case in Sect. 11.6.

11.4 Reduction of Computational Complexity

The small dimension $N \ll N_h$ of the linear system (11.26) to be solved at each Newton step does not warrant substantial computational savings, as the assembly of the reduced Jacobian matrix and residual vector still involve computations whose complexity depends on N_h . While in the linear case the assumption of affine parametric dependence proved to be sufficient to deliver computational efficiency, in the nonlinear case this turns out to be only a necessary condition.

Indeed, let us assume that the residual $\mathbf{G}_h(\cdot; \boldsymbol{\mu})$ is affine with respect to the parameters, i.e. it can be expressed as

$$\mathbf{G}_h(\mathbf{w}_h; \boldsymbol{\mu}) = \sum_{q=1}^{Q_g} \theta_g^q(\boldsymbol{\mu}) \mathbf{G}_h^q(\mathbf{w}_h).$$

Considering the case of Galerkin projection, the assembly of the reduced residual

$$\mathbf{G}_N(\mathbf{w}_N; \boldsymbol{\mu}) = \sum_{q=1}^{Q_g} \theta_g^q(\boldsymbol{\mu}) \mathbb{V}^T \mathbf{G}_h^q(\mathbb{V} \mathbf{w}_N)$$

still involves the assembly (and projection) of the high-fidelity one, thus preventing an efficient offline-online decomposition. The latter can only be achieved either in the particular case of low-order (typically 2 or 3) polynomial nonlinearity or by introducing an additional level of reduction. We start by discussing the first case.

If the problem features a low-order polynomial nonlinearity and the parametric dependence is affine, the assembly of both the reduced residual $\mathbf{G}_N(\cdot; \boldsymbol{\mu})$ and the Jacobian $\mathbb{J}_N(\cdot; \boldsymbol{\mu})$ can be efficiently achieved via an offline-online decomposition. For instance, in the case of quadratic nonlinearity (a notable example is provided by the Navier-Stokes equations) we can express the residual as

$$\mathbf{G}_h(\mathbf{w}_h; \boldsymbol{\mu}) = \tilde{\mathbf{G}}_h(\mathbf{w}_h, \mathbf{w}_h; \boldsymbol{\mu}),$$

$\tilde{\mathbf{G}}_h(\cdot; \cdot; \boldsymbol{\mu})$ being linear with respect to the first two arguments. Then, thanks to the affine parametric dependence, we can assume the reduced residual to be expressed

as

$$\mathbf{G}_N(\mathbf{w}_N; \boldsymbol{\mu}) = \sum_{q=1}^{Q_g} \boldsymbol{\theta}_g^q(\boldsymbol{\mu}) \sum_{n,m=1}^N w_N^{(n)} w_N^{(m)} \mathbb{V}^T \tilde{\mathbf{G}}_h^q(\boldsymbol{\zeta}_n, \boldsymbol{\zeta}_m) \quad (11.34)$$

for suitable smooth functions $\boldsymbol{\theta}_g^q : \mathcal{D} \rightarrow \mathbb{R}$ and $\boldsymbol{\mu}$ -independent vectors $\tilde{\mathbf{G}}_h^q(\cdot, \cdot) \in \mathbb{R}^{N_h}$.

Similarly, we can express the reduced Jacobian matrix as

$$\mathbb{J}_N(\mathbf{w}_N; \boldsymbol{\mu}) = \sum_{j=1}^{Q_j} \boldsymbol{\theta}_j^q(\boldsymbol{\mu}) \sum_{n=1}^N w_N^{(n)} \mathbb{V}^T \mathbb{J}_h^q(\boldsymbol{\zeta}_n) \mathbb{V} \quad (11.35)$$

for suitable smooth functions $\boldsymbol{\theta}_j^q : \mathcal{D} \rightarrow \mathbb{R}$ and $\boldsymbol{\mu}$ -independent matrices $\mathbb{J}_q(\boldsymbol{\zeta}_n) \in \mathbb{R}^{N_h \times N_h}$. Note that, thanks to the assumption of quadratic nonlinearity, the Jacobian is linear with respect to its argument. Therefore, the NQ_j reduced matrices $\mathbb{V}^T \mathbb{J}_q(\boldsymbol{\zeta}_n) \mathbb{V}$ and the $N^2 Q_g$ vectors $\mathbb{V}^T \tilde{\mathbf{G}}_q(\boldsymbol{\zeta}_n, \boldsymbol{\zeta}_m)$ can be precomputed offline, so that the RB problem can be assembled and solved online with a number of operations depending on N but not on N_h .

If the problem is quadratically nonlinear but nonaffine with respect to the parameters, approximation techniques such as the empirical interpolation method (EIM) or its discrete variant (DEIM) (see Chap. 10) can be introduced to restore the assumption of affine parametric dependence [53]. We will show an example dealing with the Navier-Stokes equations in a nonaffinely parametrized geometry in Sect. 11.6.

On the other hand, if the problem features a higher (or nonpolynomial) nonlinearity (and possibly a nonaffine parametric dependence), we must introduce a further level of reduction – called *hyper-reduction* or *system approximation* [57] – suitably employing techniques such as EIM [120, 94, 119], DEIM [58, 260], missing point estimation [15] or gappy POD [42, 55, 57] to approximate the nonlinear/nonaffine terms and recover an (approximate) affine structure. All these approaches are aimed at obtaining an affine approximation of the residual of the form

$$\mathbf{G}_N(\mathbf{w}_N; \boldsymbol{\mu}) \approx \sum_{q=1}^{Q_g} \alpha_g^q(\boldsymbol{\mu}; \mathbf{w}_N) \mathbb{V}^T \mathbf{G}_h^q, \quad (11.36)$$

being $\{\mathbf{G}_h^q\}_{q=1}^{Q_g}$ a basis for a suitable subspace of

$$\mathcal{M}_G = \{\mathbf{G}_h(\mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu}) : \boldsymbol{\mu} \in \mathcal{P}\} \subset \mathbb{R}^{N_h}$$

and $\alpha_g^q(\cdot; \cdot)$ some interpolation coefficients to be determined. Similarly, the Jacobian will be approximated by

$$\mathbb{J}_N(\mathbf{w}_N; \boldsymbol{\mu}) \approx \sum_{j=1}^{Q_j} \lambda_j^q(\boldsymbol{\mu}; \mathbf{w}_N) \mathbb{V}^T \mathbb{J}_h^q \mathbb{V}, \quad (11.37)$$

where $\{\mathbb{J}_h^q\}_{q=1}^{Q_j}$ is a basis for a suitable subspace of

$$\mathcal{M}_J = \{\mathbb{J}_h(\mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu}) : \boldsymbol{\mu} \in \mathcal{P}\} \subset \mathbb{R}^{N_h \times N_h}$$

and $\lambda_j^q(\cdot; \cdot)$ are interpolation coefficients to be determined.

We also mention an alternative class of approaches which aim at approximating directly the parametrized reduced operators, rather than the high-fidelity ones. For instance, in [7, 83, 8] a method based on the interpolation on appropriate matrix manifolds have been successfully applied.

11.5 A Posteriori Error Estimation for Nonlinear Problems

We now provide the nonlinear analogue of the error-residual equivalence (3.71) by means of the Brezzi-Rappaz-Raviart theory [39, 115, 49]. We first define the stability factor

$$\beta_h^N(\boldsymbol{\mu}) = \|\mathbb{X}_h^{1/2} \mathbb{J}_h(\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})^{-1} \mathbb{X}_h^{1/2}\|_2^{-1} = \sigma_{\min}(\mathbb{X}_h^{-1/2} \mathbb{J}_h(\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu}) \mathbb{X}_h^{-1/2}),$$

and the continuity factor

$$\gamma_h^N(\boldsymbol{\mu}) = \|\mathbb{X}_h^{1/2} \mathbb{J}_h(\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu}) \mathbb{X}_h^{1/2}\|_2 = \sigma_{\max}(\mathbb{X}_h^{-1/2} \mathbb{J}_h(\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu}) \mathbb{X}_h^{-1/2}).$$

Then, we define

$$R(\boldsymbol{\mu}) = \frac{2\|\mathbf{G}_h(\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})\|_{\mathbb{X}_h^{-1}}}{\beta_h^N(\boldsymbol{\mu})}, \quad \tau_N(\boldsymbol{\mu}) = \frac{4K_h^N(\boldsymbol{\mu})\|\mathbf{G}_h(\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})\|_{\mathbb{X}_h^{-1}}}{(\beta_h^N(\boldsymbol{\mu}))^2}.$$

Finally, we recall the definition (5.36) of the following $(\mathbb{X}_h, \mathbb{X}_h^{-1})$ matrix norm

$$\|\mathbb{B}\|_{\mathbb{X}_h, \mathbb{X}_h^{-1}} = \sup_{\mathbf{v} \in \mathbb{R}^{N_h}} \frac{\|\mathbb{B}\mathbf{v}\|_{\mathbb{X}_h^{-1}}}{\|\mathbf{v}\|_{\mathbb{X}_h}} = \sup_{\mathbf{v} \in \mathbb{R}^{N_h}} \frac{\|\mathbb{X}_h^{-1/2} \mathbb{B} \mathbb{X}_h^{-1/2} \mathbf{v}\|_2}{\|\mathbf{v}\|_2} \quad \forall \mathbb{B} \in \mathbb{R}^{N_h \times N_h}.$$

Proposition 11.2. *Assume that $\mathbf{u}_N(\boldsymbol{\mu})$ is a regular solution of problem (11.24) and that $\mathbb{J}_h(\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})$ is locally Lipschitz continuous at $\mathbf{u}_N(\boldsymbol{\mu})$, i.e. there exists $K_h^N(\boldsymbol{\mu}) > 0$ such that for all $\mathbf{v} \in \bar{B}_{R(\boldsymbol{\mu})}(\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}))$*

$$\|\mathbb{J}_h(\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu}) - \mathbb{J}_h(\mathbf{v}; \boldsymbol{\mu})\|_{\mathbb{X}_h, \mathbb{X}_h^{-1}} \leq K_h^N(\boldsymbol{\mu}) \|\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}) - \mathbf{v}\|_{\mathbb{X}_h}.$$

If $\tau_N(\boldsymbol{\mu}) \leq 1$, there exists a unique solution $\mathbf{u}_h(\boldsymbol{\mu}) \in \bar{B}_{R(\boldsymbol{\mu})}(\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}))$. Moreover,

$$\begin{aligned} \frac{1}{2\gamma_h^N(\boldsymbol{\mu})} \|\mathbf{G}_h(\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})\|_{\mathbb{X}_h^{-1}} &\leq \|\mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{V}\mathbf{u}_N(\boldsymbol{\mu})\|_{\mathbb{X}_h} \\ &\leq \frac{2}{\beta_h^N(\boldsymbol{\mu})} \|\mathbf{G}_h(\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})\|_{\mathbb{X}_h^{-1}}. \end{aligned} \quad (11.38)$$

Proof. For the sake of notation we omit the $\boldsymbol{\mu}$ -dependence. The proof is divided into two steps. First, we have to prove the existence of a unique solution \mathbf{u}_h in the closed ball $\bar{B}_R(\mathbb{V}\mathbf{u}_N)$; to this end, it is sufficient to prove that the map $\mathbb{H} : \mathbb{R}^{N_h} \rightarrow \mathbb{R}^{N_h}$,

$$\mathbb{H}(\mathbf{v}) = \mathbf{v} - \mathbb{J}_h(\mathbb{V}\mathbf{u}_N)^{-1} \mathbf{G}_h(\mathbf{v})$$

is a strict contraction. We first prove that \mathbb{H} maps $\bar{B}_R(\mathbb{V}\mathbf{u}_N)$ into itself; for any $\mathbf{v} \in \bar{B}_R(\mathbb{V}\mathbf{u}_N)$ we can write

$$\begin{aligned} \mathbb{H}(\mathbf{v}) - \mathbb{V}\mathbf{u}_N &= \mathbf{v} - \mathbb{J}_h(\mathbb{V}\mathbf{u}_N)^{-1} \mathbf{G}_h(\mathbf{v}) - \mathbb{V}\mathbf{u}_N \\ &= \mathbb{J}_h(\mathbb{V}\mathbf{u}_N)^{-1} [\mathbb{J}_h(\mathbb{V}\mathbf{u}_N)(\mathbf{v} - \mathbb{V}\mathbf{u}_N) - \mathbf{G}_h(\mathbf{v}) \\ &\quad + \mathbf{G}_h(\mathbb{V}\mathbf{u}_N) - \mathbf{G}_h(\mathbb{V}\mathbf{u}_N)]. \end{aligned} \quad (11.39)$$

By the mean value theorem we have

$$\mathbf{G}_h(\mathbf{v}) - \mathbf{G}_h(\mathbb{V}\mathbf{u}_N) = \int_0^1 \mathbb{J}_h(\mathbb{V}\mathbf{u}_N + s(\mathbf{v} - \mathbb{V}\mathbf{u}_N))(\mathbf{v} - \mathbb{V}\mathbf{u}_N) ds, \quad (11.40)$$

so that

$$\begin{aligned} \mathbb{H}(\mathbf{v}) - \mathbb{V}\mathbf{u}_N &= \mathbb{J}_h(\mathbb{V}\mathbf{u}_N)^{-1} \left(\int_0^1 \mathbb{J}_h(\mathbb{V}\mathbf{u}_N)(\mathbf{v} - \mathbb{V}\mathbf{u}_N) ds \right. \\ &\quad \left. - \int_0^1 \mathbb{J}_h(\mathbb{V}\mathbf{u}_N + s(\mathbf{v} - \mathbb{V}\mathbf{u}_N))(\mathbf{v} - \mathbb{V}\mathbf{u}_N) ds - \mathbf{G}_h(\mathbb{V}\mathbf{u}_N) \right). \end{aligned}$$

Then,

$$\begin{aligned} \|\mathbb{H}(\mathbf{v}) - \mathbb{V}\mathbf{u}_N\|_{\mathbb{X}_h} &\leq \frac{1}{\beta_h^N} \left(\|\mathbf{v} - \mathbb{V}\mathbf{u}_N\|_{\mathbb{X}_h} \int_0^1 \|\mathbb{J}_h(\mathbb{V}\mathbf{u}_N) \right. \\ &\quad \left. - \mathbb{J}_h(\mathbb{V}\mathbf{u}_N + s(\mathbf{v} - \mathbb{V}\mathbf{u}_N))\|_{\mathbb{X}_h, \mathbb{X}_h^{-1}} ds + \|\mathbf{G}_h(\mathbb{V}\mathbf{u}_N)\|_{\mathbb{X}^{-1}} \right) \\ &\leq \frac{1}{\beta_h^N} (K_h^N \|\mathbf{v} - \mathbb{V}\mathbf{u}_N\|_{\mathbb{X}_h}^2 + \|\mathbf{G}_h(\mathbb{V}\mathbf{u}_N)\|_{\mathbb{X}^{-1}}). \end{aligned}$$

Recalling that $\mathbf{v} \in \bar{B}_R(\mathbb{V}\mathbf{u}_N)$ and $\tau_N \leq 1$,

$$\begin{aligned} \|\mathbb{H}(\mathbf{v}) - \mathbb{V}\mathbf{u}_N\|_{\mathbb{X}_h} &\leq \frac{1}{\beta_h^N} \left((\beta_h^N)^{-2} 4K_h^N \|\mathbf{G}_h(\mathbb{V}\mathbf{u}_N)\|_{\mathbb{X}_h^{-1}}^2 + \|\mathbf{G}_h(\mathbb{V}\mathbf{u}_N)\|_{\mathbb{X}_h^{-1}} \right) \\ &\leq \frac{2}{\beta_h^N} \|\mathbf{G}_h(\mathbb{V}\mathbf{u}_N)\|_{\mathbb{X}_h^{-1}} = R, \end{aligned}$$

that is $\mathbb{H}(\mathbf{v}) \in \bar{B}_R(\mathbb{V}\mathbf{u}_N)$.

Let now $\mathbf{v}_1, \mathbf{v}_2$ be in $\bar{B}_R(\mathbb{V}\mathbf{u}_N)$, then

$$\begin{aligned} \mathbb{H}(\mathbf{v}_1) - \mathbb{H}(\mathbf{v}_2) &= \mathbf{v}_1 - \mathbb{J}_h(\mathbb{V}\mathbf{u}_N)^{-1} \mathbf{G}_h(\mathbf{v}_1) - \mathbf{v}_2 + \mathbb{J}_h(\mathbb{V}\mathbf{u}_N)^{-1} \mathbf{G}_h(\mathbf{v}_2) \\ &= \mathbb{J}_h(\mathbb{V}\mathbf{u}_N)^{-1} [\mathbb{J}_h(\mathbb{V}\mathbf{u}_N)(\mathbf{v}_1 - \mathbf{v}_2) - (\mathbf{G}_h(\mathbf{v}_1) - \mathbf{G}_h(\mathbf{v}_2))] \\ &= \mathbb{J}_h(\mathbb{V}\mathbf{u}_N)^{-1} \int_0^1 (\mathbb{J}_h(\mathbb{V}\mathbf{u}_N) - \mathbb{J}_h(\mathbb{V}\mathbf{u}_N + s(\mathbf{v}_1 - \mathbf{v}_2)))(\mathbf{v}_1 - \mathbf{v}_2) ds, \end{aligned}$$

and

$$\|\mathbb{H}(\mathbf{v}_1) - \mathbb{H}(\mathbf{v}_2)\|_{\mathbb{X}_h} \leq \frac{2K_h^N \|\mathbf{G}_h(\mathbb{V}\mathbf{u}_N)\|_{\mathbb{X}_h^{-1}}}{(\beta_h^N)^2} \|\mathbf{v}_1 - \mathbf{v}_2\|_{\mathbb{X}_h} \leq \frac{1}{2} \|\mathbf{v}_1 - \mathbf{v}_2\|_{\mathbb{X}_h}.$$

Thanks to the Banach fixed-point theorem (see, e.g., [66]) there exists a unique $\mathbf{u}_h \in \bar{B}_R(\mathbb{V}\mathbf{u}_N)$ such that $\mathbb{H}(\mathbf{u}_h) = \mathbf{u}_h$, i.e. $\mathbf{G}_h(\mathbf{u}_h) = \mathbf{0}$.

To prove the lower bound in (11.38) we choose $\mathbf{v} = \mathbf{u}_h$ in (11.40) and exploit the fact that $\mathbf{G}_h(\mathbf{u}_h) = \mathbf{0}$, finding

$$\begin{aligned} \mathbf{G}_h(\mathbb{V}\mathbf{u}_N) &= -\mathbb{J}_h(\mathbb{V}\mathbf{u}_N)(\mathbf{u}_h - \mathbb{V}\mathbf{u}_N) \\ &\quad + \int_0^1 [\mathbb{J}_h(\mathbb{V}\mathbf{u}_N) - \mathbb{J}_h(\mathbb{V}\mathbf{u}_N + s(\mathbf{u}_h - \mathbb{V}\mathbf{u}_N))](\mathbf{u}_h - \mathbb{V}\mathbf{u}_N) ds. \end{aligned} \quad (11.41)$$

Then, since $\mathbf{u}_h \in \bar{B}_R(\mathbb{V}\mathbf{u}_N)$ and $\tau_N \leq 1$,

$$\begin{aligned} \|\mathbf{G}_h(\mathbb{V}\mathbf{u}_N; \boldsymbol{\mu})\|_{\mathbb{X}_h^{-1}} &\leq \gamma_h^N \|\mathbf{u}_h - \mathbb{V}\mathbf{u}_N\|_{\mathbb{X}_h} + \frac{K_h^N}{2} \|\mathbf{u}_h - \mathbb{V}\mathbf{u}_N\|_{\mathbb{X}_h}^2 \\ &\leq \gamma_h^N \|\mathbf{u}_h - \mathbb{V}\mathbf{u}_N\|_{\mathbb{X}_h} + \frac{1}{2} \|\mathbf{G}_h(\mathbb{V}\mathbf{u}_N)\|_{\mathbb{X}_h^{-1}}, \end{aligned}$$

which proves the first inequality in (11.38). \square

Proposition 11.2 extends to the nonlinear case the error-residual relation (3.71). As in the linear case, the norm of the error is bounded from below and from above by the dual norm of the (high-fidelity) residual of the reduced solution. Since this latter only involves the high-fidelity problem components and the computed reduced solution $u_N(\boldsymbol{\mu})$, but not $u_h(\boldsymbol{\mu})$, its norm well serves as an a posteriori error estimator.

Remark 11.2. The upper bound in (11.38) can be improved to

$$\|\mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{V}\mathbf{u}_N(\boldsymbol{\mu})\|_{\mathbb{X}_h} \leq \Delta_N(\boldsymbol{\mu}) = \frac{2}{2 - \tau_N(\boldsymbol{\mu})} \frac{\|\mathbf{G}_h(\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})\|_{\mathbb{X}_h^{-1}}}{\beta_h^N(\boldsymbol{\mu})}. \quad (11.42)$$

In fact, denoting by $\mathbf{e}_h(\boldsymbol{\mu}) = \mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{V}\mathbf{u}_N(\boldsymbol{\mu})$ we can rewrite (11.41) as

$$\begin{aligned} \mathbf{e}_h(\boldsymbol{\mu}) &= (\mathbb{J}_h(\mathbb{V}\mathbf{u}_N))^{-1} \left(\int_0^1 [\mathbb{J}_h(\mathbb{V}\mathbf{u}_N) - \mathbb{J}_h(\mathbb{V}\mathbf{u}_N + s(\mathbf{e}_h(\boldsymbol{\mu})))] \mathbf{e}_h(\boldsymbol{\mu}) ds \right. \\ &\quad \left. - \mathbb{J}_h(\mathbb{V}\mathbf{u}_N) \mathbf{e}_h(\boldsymbol{\mu}) \right) \end{aligned}$$

whence, exploiting the definition of $\beta_n^N(\boldsymbol{\mu})$,

$$\begin{aligned}
 \|\mathbf{e}_h(\boldsymbol{\mu})\|_{\mathbb{X}_h} &\leq \frac{1}{\beta_h^N(\boldsymbol{\mu})} \left(\int_0^1 \|\mathbb{J}_h(\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})\|_{\mathbb{X}_h, \mathbb{X}_h^{-1}} ds \|\mathbf{e}_h(\boldsymbol{\mu})\|_{\mathbb{X}_h} + \|\mathbf{G}_h(\mathbb{V}\mathbf{u}_N; \boldsymbol{\mu})\|_{\mathbb{X}^{-1}} \right) \\
 &\leq \frac{1}{\beta_h^N(\boldsymbol{\mu})} \left(2K_h^N(\boldsymbol{\mu}) \frac{\|\mathbf{G}_h(\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})\|_{\mathbb{X}^{-1}}}{\beta_h^N(\boldsymbol{\mu})} \|\mathbf{e}_h(\boldsymbol{\mu})\|_{\mathbb{X}_h} \right. \\
 &\quad \left. + \|\mathbf{G}_h(\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})\|_{\mathbb{X}^{-1}} \right) \\
 &= \frac{1}{2} \tau_N(\boldsymbol{\mu}) \|\mathbf{e}_h(\boldsymbol{\mu})\|_{\mathbb{X}_h} + \frac{\|\mathbf{G}_h(\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})\|_{\mathbb{X}^{-1}}}{\beta_h^N(\boldsymbol{\mu})}.
 \end{aligned}$$

Moreover, the restriction $\tau_N(\boldsymbol{\mu}) \leq 1$ can be slightly relaxed to become

$$\tilde{\tau}_N(\boldsymbol{\mu}) = \frac{1}{2} \tau_N(\boldsymbol{\mu}) < 1, \quad (11.43)$$

which can be regarded as the discrete analogue of condition (11.16) in the Kantorovich theorem 11.2. Actually Proposition 11.2 provides a result similar to this latter for $u_h^0 = u_N(\boldsymbol{\mu})$. Indeed, by identifying

$$u_h^0 \text{ with } \mathbf{u}_N(\boldsymbol{\mu}), \quad \varepsilon \text{ with } \|\mathbf{G}_h(\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})\|_{\mathbb{X}_h^{-1}}, \quad a \text{ with } \frac{1}{\beta_h^N(\boldsymbol{\mu})}, \quad K \text{ with } K_h^N(\boldsymbol{\mu}),$$

at the algebraic level, Kantorovich theorem (for $k = 0$) provides the following bound between the high-fidelity and the RB solutions

$$\|\mathbf{u}_h(\boldsymbol{\mu}) - \mathbb{V}\mathbf{u}_N(\boldsymbol{\mu})\|_{\mathbb{X}_h} \leq r_-(\boldsymbol{\mu}) = \frac{\beta_h^N(\boldsymbol{\mu})}{K_h^N(\boldsymbol{\mu})} \left(1 - \sqrt{1 - \tilde{\tau}_N(\boldsymbol{\mu})} \right). \quad (11.44)$$

Note that $r_-(\boldsymbol{\mu})$ and $\Delta_N(\boldsymbol{\mu})$ behave asymptotically in a similar way when ε tends to zero. Indeed, by Taylor expansion (omitting the $\boldsymbol{\mu}$ dependence) we find

$$\begin{aligned}
 r_- &= \frac{1 - \sqrt{1 - 2\varepsilon K a^2}}{aK} = \frac{1}{aK} \left(1 - 1 + \frac{1}{2} 2\varepsilon K a^2 + \frac{1}{8} 4\varepsilon^2 K^2 a^4 + O(\varepsilon^3) \right) \\
 &= a\varepsilon + \frac{1}{2} \varepsilon^2 K a^2 + O(\varepsilon^3)
 \end{aligned}$$

and

$$\Delta_N = \frac{a\varepsilon}{1 - 2\varepsilon K a^2} = a\varepsilon (1 + 2\varepsilon K a^2 + O(\varepsilon^2)) = a\varepsilon + 2\varepsilon^2 K a^2 + O(\varepsilon^3). \quad \bullet$$

11.6 Application to the Steady Navier-Stokes Equations

Let us now cast the parametrized steady Navier-Stokes equations into the general framework discussed so far, by taking advantage of the formulation addressed in Sects. 8.7–8.8, as well as of the construction of the RB method for the parametrized Stokes equations presented in Sect. 9.3. After the pioneering works by Peterson [210], Ito and Ravindran [146], a general framework for both RB approximation and a posteriori error estimation of parametrized Navier-Stokes equations has been presented by Veroy, Patera [254] and Nguyen [199]. This has been further developed to include both physical parameters [86, 88, 87, 217] and (affine/nonaffine) geometric parametrizations [182]; see also [161].

Similarly to the Stokes case of Sect. 9.3, we denote by $X_h \subset X$ and $Q_h \subset Q$ the velocity and the pressure high-fidelity space, respectively, and set $V_h = X_h \times Q_h$, being $X_h = [Y_h]^d$ and $Y_h \subset H^1(\Omega)$. See, e.g., [104, 124, 222] for further details about the finite element approximation of Navier-Stokes equations.

The Newton method (11.14) in this case reads: given an initial guess $(\mathbf{u}_h^0, p_h^0) \in V_h$, for $k = 0, 1, \dots$ until convergence we seek $(\delta \mathbf{u}_h, \delta p_h) \in V_h$ such that – omitting the $\boldsymbol{\mu}$ -dependence for the sake of notation

$$\begin{cases} \bar{d}(\delta \mathbf{u}_h, \mathbf{v}_h; \boldsymbol{\mu}) + c(\mathbf{u}_h^k, \delta \mathbf{u}_h, \mathbf{v}_h; \boldsymbol{\mu}) + c(\delta \mathbf{u}_h, \mathbf{u}_h^k, \mathbf{v}_h; \boldsymbol{\mu}) + b(\mathbf{v}_h, \delta p_h; \boldsymbol{\mu}) \\ \quad = f_1(\mathbf{v}_h; \boldsymbol{\mu}) - \bar{d}(\mathbf{u}_h^k, \mathbf{v}_h; \boldsymbol{\mu}) - c(\mathbf{u}_h^k, \mathbf{u}_h^k, \mathbf{v}_h; \boldsymbol{\mu}) - b(\mathbf{v}_h, p_h^k; \boldsymbol{\mu}) & \forall \mathbf{v}_h \in X_h, \\ b(\delta \mathbf{u}_h, q; \boldsymbol{\mu}) = f_2(q; \boldsymbol{\mu}) - b(\mathbf{u}_h^k, q; \boldsymbol{\mu}) & \forall q_h \in Q_h, \end{cases} \quad (11.45)$$

and then set $(\mathbf{u}_h^{k+1}, p_h^{k+1}) = (\mathbf{u}_h^k + \delta \mathbf{u}_h, p_h^k + \delta p_h)$. A typical initial guess is the solution of the corresponding Stokes problem.

The Newton iteration is well defined and yields a sequence converging to the solution of (11.4) according to the results of the local convergence theorem 11.1. In fact, similarly to (11.6), it is possible to show that $c(\cdot, \cdot, \cdot; \boldsymbol{\mu})$ is continuous over X_h ,

$$c(\mathbf{u}, \mathbf{w}, \mathbf{v}; \boldsymbol{\mu}) \leq \rho_h^2 M_c(\boldsymbol{\mu}) \|\mathbf{u}\|_X \|\mathbf{w}\|_X \|\mathbf{v}\|_X \quad \forall \mathbf{u}, \mathbf{v}, \mathbf{w} \in X_h, \forall \boldsymbol{\mu} \in \mathcal{P} \quad (11.46)$$

where $\rho_h > 0$ is the discrete Sobolev embedding constant

$$\rho_h = \sup_{\mathbf{v} \in Y_h} \frac{\|\mathbf{v}\|_{L^4(\Omega)}}{\|\mathbf{v}\|_{H^1(\Omega)}}.$$

Thanks to this property, the Fréchet derivative of the Navier-Stokes operator satisfies (11.9) with ρ_h instead of ρ , and thus it is locally Lipschitz-continuous at $(\mathbf{u}_h(\boldsymbol{\mu}), p_h(\boldsymbol{\mu})) \in V_h$, with

$$K_h(\boldsymbol{\mu}) = 2\rho_h^2 M_c(\boldsymbol{\mu}). \quad (11.47)$$

Hence, the Kantorovich theorem 11.2 ensures that the Newton sequence is converging to the solution of the problem, provided the initial guess is sufficiently close to it; typically, this is ensured by taking as initial guess the solution of the corresponding Stokes problem.

The k -th step (11.45) of the Newton method can be written in compact algebraic form as follows: find $\delta \mathbf{U}_h \in \mathbb{R}^{N_h^u + N_h^p}$ such that

$$\mathbb{J}_h(\mathbf{U}_h^k; \boldsymbol{\mu}) \delta \mathbf{U}_h = -\mathbf{G}_h(\mathbf{U}_h^k; \boldsymbol{\mu}), \quad (11.48)$$

then set $\mathbf{U}_h^{k+1} = \mathbf{U}_h^k + \delta \mathbf{U}_h$. The Jacobian matrix $\mathbb{J}_h(\mathbf{U}_h^k; \boldsymbol{\mu}) \in \mathbb{R}^{(N_h^u + N_h^p) \times (N_h^u + N_h^p)}$ is defined as

$$\mathbb{J}_h(\mathbf{U}_h; \boldsymbol{\mu}) = \begin{pmatrix} \mathbb{D}_h(\boldsymbol{\mu}) + \mathbb{C}_{1h}(\mathbf{u}_h; \boldsymbol{\mu}) + \mathbb{C}_{2h}(\mathbf{u}_h; \boldsymbol{\mu}) & \mathbb{B}_h^T(\boldsymbol{\mu}) \\ \mathbb{B}_h(\boldsymbol{\mu}) & 0 \end{pmatrix}, \quad (11.49)$$

being, for $i, j = 1, \dots, N_h$

$$(\mathbb{C}_{1h}(\mathbf{u}_h; \boldsymbol{\mu}))_{ij} = c(\mathbf{u}_h, \boldsymbol{\varphi}^j, \boldsymbol{\varphi}^i; \boldsymbol{\mu}), \quad (\mathbb{C}_{2h}(\mathbf{u}_h; \boldsymbol{\mu}))_{ij} = c(\boldsymbol{\varphi}^j, \mathbf{u}_h, \boldsymbol{\varphi}^i; \boldsymbol{\mu})$$

having denoted by $\{\boldsymbol{\varphi}^i\}_{i=1}^{N_h}$ a basis of the finite element velocity space. The residual is given by

$$\mathbf{G}_h(\mathbf{U}_h; \boldsymbol{\mu}) = \begin{pmatrix} \mathbb{D}_h(\boldsymbol{\mu})\mathbf{u}_h + \mathbf{C}_{1h}(\mathbf{u}_h; \boldsymbol{\mu})\mathbf{u}_h + \mathbb{B}_h^T(\boldsymbol{\mu})\mathbf{p}_h - \mathbf{f}_h(\boldsymbol{\mu}) \\ \mathbb{B}_h(\boldsymbol{\mu})\mathbf{u}_h - \mathbf{g}_h(\boldsymbol{\mu}) \end{pmatrix}, \quad (11.50)$$

and $\mathbf{U}_h = (\mathbf{u}_h; \mathbf{p}_h)$, $\delta \mathbf{U}_h = (\delta \mathbf{u}_h; \delta \mathbf{p}_h) \in \mathbb{R}^{N_h + N_h^p}$ collect both velocity and pressure degrees of freedom. The Stokes matrices $\mathbb{D}_h(\boldsymbol{\mu})$, $\mathbb{B}_h(\boldsymbol{\mu})$, as well the right-hand sides $\mathbf{f}_h(\boldsymbol{\mu})$, $\mathbf{g}_h(\boldsymbol{\mu})$ and the vectors \mathbf{u}_h , \mathbf{p}_h , are defined as in Sect. 9.3.

11.6.1 RB Approximation of the Navier-Stokes Equations

To construct the RB approximation of parametrized Navier-Stokes equations we can exploit the Galerkin RB method choosing the RB spaces similarly to the Stokes case. By considering a greedy algorithm [86, 88, 182], we define

$$Q_N = \text{span}\{p_h(\boldsymbol{\mu}^n), n = 1, \dots, N\}, \quad (11.51)$$

$$X_N = \text{span}\{\mathbf{u}_h(\boldsymbol{\mu}^n), T_p^{\boldsymbol{\mu}^n} p_h(\boldsymbol{\mu}^n), n = 1, \dots, N\} \quad (11.52)$$

being $(\mathbf{u}_h(\boldsymbol{\mu}), p_h(\boldsymbol{\mu}))$ the solution of the high-fidelity Navier-Stokes problem; the pressure supremizer operator $T^{\boldsymbol{\mu}}$ is defined as in (9.4). As in the Stokes case, a further Gram-Schmidt orthonormalization is performed to obtain orthonormal reduced basis functions. A $\boldsymbol{\mu}$ -independent enrichment of the velocity space provides a remarkable computational speedup without affecting the approximation accuracy also in the Navier-Stokes case. A detailed explanation of the (indeed, quite similar) procedure to be followed if the POD algorithm is used can be found, e.g., in [47, 22, 261]. This latter approach however entails the solution of the high-fidelity problem for a larger number $n_s > N$ of snapshots, thus requiring further computational efforts, e.g. in the case of large Reynolds numbers.

The G-RB problem thus reads: find $(\mathbf{u}_N(\boldsymbol{\mu}), p_N(\boldsymbol{\mu})) \in X_N \times Q_N$ such that

$$\begin{cases} \bar{d}(\mathbf{u}_N(\boldsymbol{\mu}), \mathbf{v}_N; \boldsymbol{\mu}) + c(\mathbf{u}_N(\boldsymbol{\mu}), \mathbf{u}_N(\boldsymbol{\mu}), \mathbf{v}_N; \boldsymbol{\mu}) \\ \quad + b(\mathbf{v}_N, p_N(\boldsymbol{\mu}); \boldsymbol{\mu}) = f_1(\mathbf{v}_N; \boldsymbol{\mu}) & \forall \mathbf{v}_N \in X_N \\ b(\mathbf{u}_N(\boldsymbol{\mu}), q_N; \boldsymbol{\mu}) = f_2(q_N; \boldsymbol{\mu}) & \forall q_N \in Q_N. \end{cases} \quad (11.53)$$

For every $\boldsymbol{\mu} \in \mathcal{P}$, (11.53) can be solved by means of Newton method (11.21): given an initial guess $(\mathbf{u}_N^0, p_N^0) \in V_N = X_N \times Q_N$, for $k = 0, 1, \dots$ until convergence, we seek $(\delta \mathbf{u}_N, \delta p_N) \in V_N$ such that

$$\begin{cases} \bar{d}(\delta \mathbf{u}_N, \mathbf{v}_N; \boldsymbol{\mu}) + c(\mathbf{u}_N^k, \delta \mathbf{u}_N, \mathbf{v}_N; \boldsymbol{\mu}) + c(\delta \mathbf{u}_N, \mathbf{u}_N^k, \mathbf{v}_N; \boldsymbol{\mu}) + b(\mathbf{v}_N, \delta p_N; \boldsymbol{\mu}) \\ = f_1(\mathbf{v}_N; \boldsymbol{\mu}) - \bar{d}(\mathbf{u}_N^k, \mathbf{v}_N; \boldsymbol{\mu}) - c(\mathbf{u}_N^k, \mathbf{u}_N^k, \mathbf{v}_N; \boldsymbol{\mu}) - b(\mathbf{v}_N, p_N^k; \boldsymbol{\mu}) & \forall \mathbf{v}_N \in X_N, \\ b(\delta \mathbf{u}_N, q; \boldsymbol{\mu}) = f_2(q; \boldsymbol{\mu}) - b(\mathbf{u}_N^k, q; \boldsymbol{\mu}) & \forall q_N \in Q_N, \end{cases} \quad (11.54)$$

and then set $(\mathbf{u}_N^{k+1}, p_N^{k+1}) = (\mathbf{u}_N^k + \delta \mathbf{u}_N, p_N^k + \delta p_N)$. A typical choice for the initial guess is the solution of the linear Stokes problem. The k -th step (11.54) of the Newton method can be written in compact form as: find $\delta \mathbf{U}_N \in \mathbb{R}^{3N}$ such that

$$\mathbb{J}_N(\mathbf{U}_N^k; \boldsymbol{\mu}) \delta \mathbf{U}_N = -\mathbf{G}_N(\mathbf{U}_N^k; \boldsymbol{\mu}), \quad (11.55)$$

then set $\mathbf{U}_N^{k+1} = \mathbf{U}_N^k + \delta \mathbf{U}_N$. The reduced Jacobian matrix $\mathbb{J}_N(\mathbf{U}_N; \boldsymbol{\mu}) \in \mathbb{R}^{3N \times 3N}$ is

$$\mathbb{J}_N(\mathbf{U}_N; \boldsymbol{\mu}) = \begin{pmatrix} \mathbb{D}_N(\boldsymbol{\mu}) + \mathbb{C}_{1N}(\mathbf{u}_N; \boldsymbol{\mu}) + \mathbb{C}_{2N}(\mathbf{u}_N; \boldsymbol{\mu}) & \mathbb{B}_N^T(\boldsymbol{\mu}) \\ \mathbb{B}_N(\boldsymbol{\mu}) & 0 \end{pmatrix}, \quad (11.56)$$

being

$$\mathbb{C}_{1N}(\mathbf{u}_N; \boldsymbol{\mu}) = \sum_{n=1}^{2N} u_N^{(n)} \nabla_{\mathbf{u}}^T \mathbb{C}_{1h}(\boldsymbol{\xi}_n; \boldsymbol{\mu}) \nabla_{\mathbf{u}}, \quad \mathbb{C}_{2N}(\mathbf{u}_N; \boldsymbol{\mu}) = \sum_{n=1}^{2N} u_N^{(n)} \nabla_{\mathbf{u}}^T \mathbb{C}_{2h}(\boldsymbol{\xi}_n; \boldsymbol{\mu}) \nabla_{\mathbf{u}}$$

whereas the reduced residual is given by

$$\mathbf{G}_N(\mathbf{U}_N; \boldsymbol{\mu}) = \begin{pmatrix} \mathbb{D}_N(\boldsymbol{\mu}) \mathbf{u}_N + \mathbf{C}_{1N}(\mathbf{u}_N; \boldsymbol{\mu}) \mathbf{u}_N + \mathbb{B}_N^T(\boldsymbol{\mu}) \mathbf{p}_N - \mathbf{f}_N(\boldsymbol{\mu}) \\ \mathbb{B}_N(\boldsymbol{\mu}) \mathbf{u}_N - \mathbf{g}_N(\boldsymbol{\mu}) \end{pmatrix}, \quad (11.57)$$

and $\mathbf{U}_N = (\mathbf{u}_N; \mathbf{p}_N)$, $\delta \mathbf{U}_N = (\delta \mathbf{u}_N; \delta \mathbf{p}_N) \in \mathbb{R}^{3N}$ collect the degrees of freedom of the RB approximation of both velocity and pressure. Problem (11.55) corresponds to (11.28) by considering, as in the Stokes case of Sect. 9.3.1,

$$\mathbb{V} = \begin{pmatrix} \mathbb{V}_{\mathbf{u}} & 0 \\ 0 & \mathbb{V}_p \end{pmatrix} \in \mathbb{R}^{(N_h^u + N_h^p) \times 3N}$$

being $\mathbb{V}_{\mathbf{u}}$ defined as in (9.19). Note that in the Navier-Stokes case $Q_j = Q_d + Q_b + 2Q_c$, $Q_g = Q_{f1} + Q_{f2} + Q_j + Q_c$, being Q_d , Q_b , Q_c the number of terms appearing in the affine developments of $d(\cdot, \cdot; \boldsymbol{\mu})$, $b(\cdot, \cdot; \boldsymbol{\mu})$, $c(\cdot, \cdot, \cdot; \boldsymbol{\mu})$, whereas Q_{f1} , Q_{f2} the one of terms in the affine development of $f_1(\cdot; \boldsymbol{\mu})$, $f_2(\cdot; \boldsymbol{\mu})$, respectively.

11.6.2 *A posteriori Error Estimation*

A rigorous *a posteriori* error bound can be obtained for the RB approximation of parametrized Navier-Stokes equations exploiting the general framework of Sect. 11.5. In particular, the bound (11.38) of Proposition 11.2 yields the following error estimate

$$(\|\mathbf{u}_h(\boldsymbol{\mu}) - \mathbf{u}_N(\boldsymbol{\mu})\|_X^2 + \|p_h(\boldsymbol{\mu}) - p_N(\boldsymbol{\mu})\|_Q^2)^{1/2} \leq \frac{2}{\beta_h^N(\boldsymbol{\mu})} \|\mathbf{G}_h(\mathbb{V}\mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})\|_{\mathbb{X}_h^{-1}} \quad (11.58)$$

whereas the bound (11.44) provided by Kantorovich theorem reads

$$(\|\mathbf{u}_h(\boldsymbol{\mu}) - \mathbf{u}_N(\boldsymbol{\mu})\|_X^2 + \|p_h(\boldsymbol{\mu}) - p_N(\boldsymbol{\mu})\|_Q^2)^{1/2} \leq \frac{\beta_h^N(\boldsymbol{\mu})}{K_h^N(\boldsymbol{\mu})} \left(1 - \sqrt{1 - \tilde{\tau}_N(\boldsymbol{\mu})}\right); \quad (11.59)$$

$K_h^N(\boldsymbol{\mu}) = K_h(\boldsymbol{\mu})$ is defined in (11.47), whereas $\tilde{\tau}_N(\boldsymbol{\mu})$ is defined in (11.43). See Sect. 11.5 for the definition of the stability factor $\beta_h^N(\boldsymbol{\mu})$ and (11.50) for the definition of the residual $\mathbf{G}_h(\cdot; \boldsymbol{\mu})$ in the Navier-Stokes case. We underline that (11.58) and (11.59) are valid provided that $\tilde{\tau}_N(\boldsymbol{\mu}) < 1/2$, $\tilde{\tau}_N(\boldsymbol{\mu}) < 1$, respectively. The two error bounds show indeed the same asymptotic behavior, see Exercise 6.

Evaluating *a posteriori* error bounds in this case requires more involved computations than in the linear Stokes case. First of all, the stability factor $\beta_h^N(\boldsymbol{\mu})$ depends on the RB approximation itself. Hence, both error bounds cannot be exploited under the form (11.58)–(11.59) during the greedy procedure, unless the evaluation of $\beta_h^N(\boldsymbol{\mu})$ and the construction of the RB space are performed simultaneously. The evaluation of $\beta_h^N(\boldsymbol{\mu})$ before running the greedy algorithm for the offline construction of the RB space requires the introduction of suitable approximations of the stability factor, see e.g. [183] for further details. In any case, the successive constraint method of Sect. 3.7.2 is very hard to perform in the Navier-Stokes case, entailing severe computational costs; for this reason, introducing a suitable approximation seems unavoidable for the case at hand. Moreover, (11.59) also requires the evaluation of the (indeed, $\boldsymbol{\mu}$ -independent) discrete Sobolev embedding constant: a suitable fixed point algorithm can be devised for this goal, see e.g. [182] for further details.

11.7 Numerical Results: Backward-Facing Step Channel

We now come back to the backward-facing step channel problem we introduced in Sect. 8.7, whose weak formulation over the reference domain Ω is given by (8.57) by taking $\delta = 1$; see Sect. 9.3.3 for the analogous RB approximation in the Stokes case. Here we consider $P = 3$ parameters:

- $\mu_1 \in [20, 50]$, being $\nu = 1/\mu_1$ the viscosity of the fluid;
- $\mu_2 \in [0.5, 1.2]$ is the amplitude of the inlet velocity profile;
- $\mu_3 \in [0.5, 1.5]$ is the step height, see Fig. 8.7.

The Reynolds number is given by $Re = \mu_1 \mu_2 \mu_3 \in [5, 90]$. For the high-fidelity approximation we use $\mathbb{P}_1^b - \mathbb{P}_1$ finite elements built over a discretization of the domain made by linear triangular elements, resulting in 6071 vertices, 11 767 triangles and a high-fidelity space V_h of dimension $N_h = 41\,051$. The problem is affine, featuring $Q_d + Q_b = 6$, $Q_{f1} + Q_{f2} = 6$, $Q_c = 4$ terms in the affine expansions of the operators.

By considering a training sample $\mathcal{E}_{\text{train}}$ of size $n_{\text{train}} = 1000$ obtained by latin hypercube sampling, and a stopping tolerance $\varepsilon_g = 10^{-2}$ on the relative error bound $\Delta_N(\boldsymbol{\mu})/\|u_N(\boldsymbol{\mu})\|_V$, we end up with a reduced space made by $N = 20$ pressure basis functions and $2N = 40$ velocity basis functions in the G-RB case – see (9.18) – thus yielding $3N = 60$ degrees of freedom for the G-RB problem. The offline stage in this case requires a CPU time of about 2 hours.

The online convergence of the error between the G-RB and the high-fidelity approximation, evaluated over a test sample of 100 parameter values, is reported in Fig. 11.1. The online evaluation time is about 0.05 s per Newton iteration in the G-RB case, whereas evaluating the a posteriori error bound, for any parameter value, takes 1.2 s; note that the this latter operation is sensibly more expensive than in the Stokes case, see Tab. 11.1.

Finally, we show in Fig. 11.2 the RB approximation of the solution obtained for different parameter values. We remark that the vortex width becomes larger and larger for increasing Reynolds numbers, coherently with several results in literature; see e.g. [31, 123].

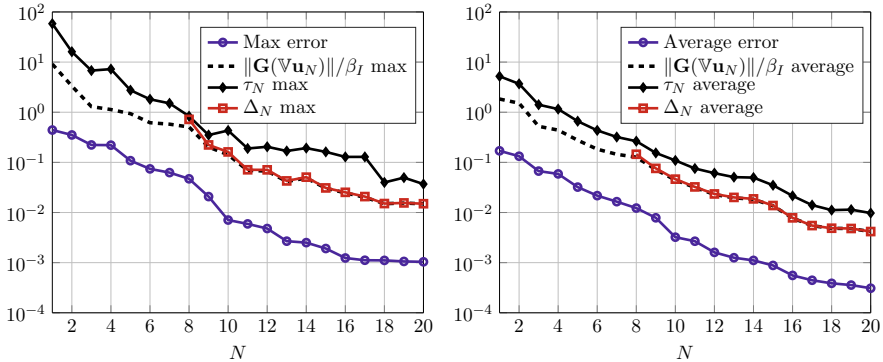


Fig. 11.1 Maximum (*left*) and average (*right*) absolute error and bound computed over a test sample of 100 points

Table 11.1 Computational details for the high-fidelity and reduced-order models for the backward-facing step problem

FE dofs N_h	41051	RB dofs	60
Q_j	18	Dofs reduction	684:1
Q_g	21	Greedy CPU time	2 h
Online CPU time (per Newton iteration)	0.05 s	Estimation CPU time	1.2 s

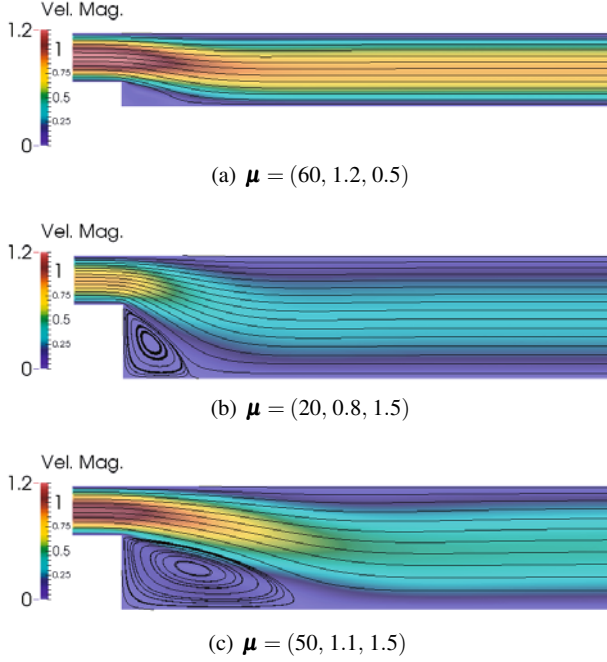


Fig. 11.2 Streamlines and magnitude of the velocity field obtained by solving the RB approximation of problem (8.53) for different parameter values

11.8 Numerical Results: Sudden Expansion Channel

We now solve the problem (8.58) modeling a fluid flow across a channel characterized by a sudden expansion, see Fig. 8.2; here we deal with $P = 2$ parameters:

- $\mu_1 \in [0, 0.65]$, representing the amplitude of the radial restriction, see (8.36);
- $\mu_2 \in [0.5, 2.5]$, giving the maximum amplitude of a fully developed Poiseuille inflow $\tilde{\mathbf{g}}(\mu_2) = [0, 0, \mu_2(1 - (x^2 + y^2)/(0.5^2))]^T$ on $\tilde{\Gamma}_{in}(\mu_1)$.

For the high-fidelity approximation we use $\mathbb{P}_1 - \mathbb{P}_1$ finite elements with pressure stabilization, built over a discretization of the (half of the) domain made by tetrahedral elements, resulting in 9971 vertices, 47088 triangles and a high-fidelity space V_h of dimension $N_h = 29284$.

This problem is nonaffine; applying the empirical interpolation method with a tolerance $\varepsilon_{\text{EIM}} = 10^{-5}$ on the components of the tensors (8.63)–(8.64) that depend on the geometric parameterization, we recover an affine expansion featuring $Q_d + Q_b = 35$, $Q_{f1} + Q_{f2} = 32$, $Q_c = 15$ terms for the linear Stokes operator, the right-hand sides and the nonlinear convective term, respectively.

We consider an EIM-G-RB approximation relying on POD for the construction of RB spaces. Since we deal with a couple of stabilized high-fidelity spaces, an EIM-G-RB approximation without supremizer enrichment yields a stable RB approxi-

mation. Indeed, starting from $n_s = 60$ snapshots randomly selected (by latin hypercube sampling over \mathcal{P}) we retain the first $N = 25$ velocity and pressure modes, see Fig. 11.3, left; the RB system has dimension $2N = 50$. Performing the offline stage (including also the empirical interpolation) in this case requires about 1.5 hours.

The online convergence of the error between the EIM-G-RB and the high-fidelity approximation, evaluated over a test sample of 100 parameter values, is reported in Fig. 11.3, right. The online evaluation time is about 0.3 s in the G-RB case, whereas the solution of the equivalent high-fidelity FE problem takes about 70 s; we thus obtain a computational speedup of about 230 and a reduction $N_h/2N \approx 585$ in the dimension of the system (to be solved at each Newton step). See Tab. 11.2 for further details. Finally, we show in Fig. 11.4 the RB approximation of the solution obtained for different cases in the range $Re < 60$, where $Re = R(1 - \mu_1)V_{max}/\nu$, being V_{max} the maximum velocity at the restriction and $\nu = 0.03$.

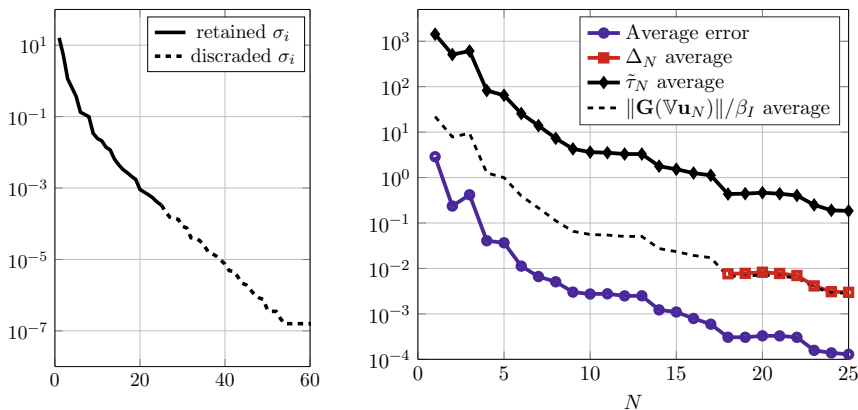


Fig. 11.3 Velocity singular values and online convergence analysis: absolute error

Table 11.2 Computational details for the high-fidelity and reduced-order models for the sudden expansion problem

High-fidelity model		EIM-G-RB (POD)	
FE dofs N_h	29 284	RB dofs	50
Q_j	65	Dofs reduction	585:1
Q_g	112	Offline CPU time	1.5 h
FE solution time	70 s	Online CPU time	0.03 s

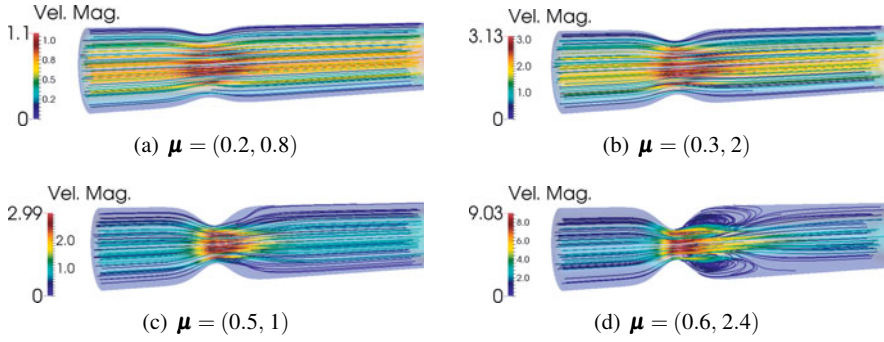


Fig. 11.4 Representative RB solutions of problem (8.58) for different parameter values

11.9 Numerical results: a Simplified Bypass Graft

We close this section by showing the numerical results related to the G-RB approximation of a further fluid flow problem, that represents the state equation of a PDE-constrained optimization that will be addressed in the following chapter (see Sect. 12.3). We consider the flow past a cylinder in presence of two inlets, the circular portion Γ_C on the lateral side (this latter given by the surface $x^2 + y^2 = R^2$, $z \in (0, L)$) and the section Γ_D (on the plane $z = 0$). This can be regarded as a simplified model of flow across a bypass graft (see Fig. 11.5), where the cylinder represents the host artery, Γ_D the (totally or partially) occluded section and the Dirichlet boundary condition on Γ_C represents the flow entering into the artery from the graft; see Fig. 11.6. Further details can be found, e.g., in [159] for a detailed description and references therein.

We impose a homogeneous Neumann (zero stress) condition at the outflow Γ_N , a homogeneous Dirichlet (no slip) condition on the lateral wall Γ_w ,

$$\mathbf{u} = \mathbf{g}_D \quad \text{on } \Gamma_D, \quad \mathbf{u} = (Q_{tot} - Q_D(\mathbf{g}_D))\mathbf{g}_C \quad \text{on } \Gamma_C$$

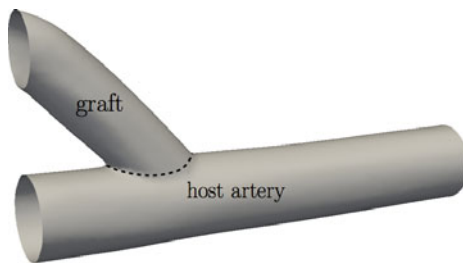


Fig. 11.5 Schematic representation of a bypass graft

where, in order to have a physically meaningful problem, we enforce the total conservation of fluxes between the (partially or totally) occluded branch Γ_D and the graft inlet Γ_C , according to

$$Q_D(\mathbf{g}_D) = \int_D \mathbf{g}_D \cdot \mathbf{n} d\Gamma \quad \int_C \mathbf{g}_C \cdot \mathbf{n} = 1, \quad (11.60)$$

$$Q_{tot} = Q_D(\mathbf{g}_D) + (Q_{tot} - Q_D(\mathbf{g}_D)) \int_C \mathbf{g}_C \cdot \mathbf{n}, \quad (11.61)$$

where Q_{tot} is the known and constant flow rate in the host artery before the bypass. Our goal is to characterize the flow pattern for different graft orientations and residual flows through Γ_D : hence, we parametrize both \mathbf{g}_D and \mathbf{g}_C with respect to

- the angle $\mu_1 = \theta \in [15^\circ, 85^\circ]$ of incidence of the inflow \mathbf{g}_C , representing the graft orientation of the angles;
- the intensity $\mu_2 = \omega \in [0, 3]$ of the residual flow \mathbf{g}_D .

Because of (11.60)–(11.61), $\mathbf{g}_C = \mathbf{g}_C(\theta, \omega)$ is indeed a function of both parameters, see e.g. [159] for further details about the expression of the boundary inflows. Here we take $R = 0.5$ and $L = 5$. For the high-fidelity approximation we use $\mathbb{P}_1 - \mathbb{P}_1$ finite elements with pressure stabilization, built over a discretization of the (half of the) domain made by tetrahedral elements, resulting in 33491 vertices, 160596 triangles and a high-fidelity space V_h of dimension $N_h = 101886$. The problem is affine, and only the linear terms accounting for the boundary conditions are $\boldsymbol{\mu}$ -dependent.

We consider a G-RB approximation relying on the greedy algorithm for the construction of RB spaces. As in the previous case, we deal with a couple of stabilized high-fidelity spaces, so that the supremizer enrichment is not required. By considering a training sample Ξ_{train} of size $n_{\text{train}} = 2000$ obtained by latin hypercube sampling, and a stopping tolerance $\varepsilon_g = 10^{-2}$ on the relative error bound $\Delta_N(\boldsymbol{\mu})/\|u_N(\boldsymbol{\mu})\|_V$, we end up with a reduced space made by $N = 28$ pressure and velocity basis functions, thus yielding $2N = 56$ degrees of freedom for the G-RB problem. The offline stage in this case requires a CPU time of about 2.7 hours.

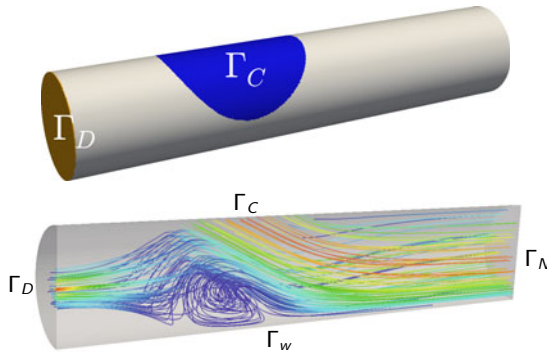


Fig. 11.6 Domain and boundary portions for the bypass model problem and an example of flow in presence of a partial occlusion

The online convergence of the error between the G-RB and the high-fidelity approximation, evaluated over a test sample of 100 parameter values, is reported in Fig. 11.7. The online evaluation time is about 0.1 s in the G-RB case, whereas the solution of the equivalent high-fidelity FE problem takes between 60 s and 250 s depending on the parameter choice; we thus obtain a computational speedup ranging from 600 to 2500 resulting from the reduction $N_h/2N \approx 1886$ in the dimension of the system (to be solved at each Newton step), see Tab. 11.3.

Indeed, the solution of a nonlinear Navier-Stokes problem in a range of Reynolds numbers requires a constant (and very small) CPU time – provided the high-fidelity model has been *trained* on that range during the offline construction of the RB space; this is a further distinguishing feature of our approach.

Finally, we show in Fig. 11.8 the streamlines of the velocity field corresponding to a complete (left) and a partial (right) occlusion. In the former case a larger recirculation is highlighted in the region between the occlusion and the bypass inflow, whereas in the latter this is confined in the proximity of the occlusion.

Several nonlinear problems (other than the steady Navier-Stokes equations here considered) have been solved by reduced basis methods. A non-exhaustive list of references includes the papers [198, 257, 48, 132, 21, 264]. A detailed review including further references can be found in [161].

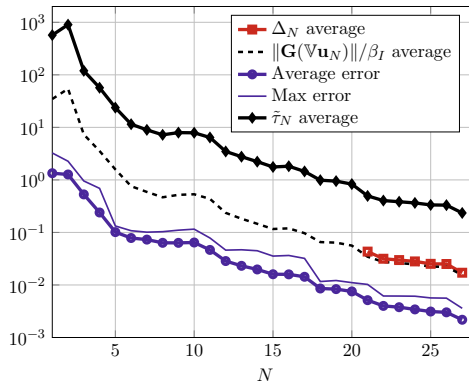


Fig. 11.7 Online convergence analysis: absolute errors and a posteriori error bounds

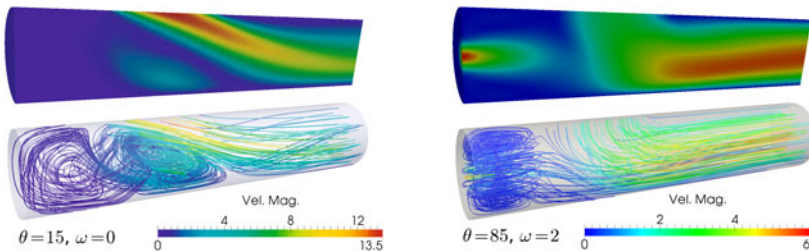


Fig. 11.8 Flow in presence of total (*left*) and partial (*right*) occlusions

Table 11.3 Computational details for the high-fidelity and reduced-order models for the simplified bypass graft problem

High-fidelity model		G-RB (greedy)	
FE dofs N_h	101 886	RB dofs	56
Q_j	3	Dofs reduction	1886:1
Q_g	7	Offline CPU time	2.7 h
FE solution time	60 – 250 s	Online CPU time	0.1 s

11.10 Exercises

1. Show that the trilinear form

$$c(\mathbf{w}, \mathbf{u}, \mathbf{v}; \boldsymbol{\mu}) = \sum_{i,j,k=1}^d \int_{\Omega} w_i \eta_{ji}(\mathbf{x}, \boldsymbol{\mu}) \frac{\partial u_k}{\partial x_j} v_k d\Omega$$

satisfies $c(\mathbf{w}, \mathbf{u}, \mathbf{v}; \boldsymbol{\mu}) \leq \rho^2 M_c(\boldsymbol{\mu}) \|\nabla \mathbf{w}\|_{(L^2(\Omega))^d} \|\nabla \mathbf{u}\|_{(L^2(\Omega))^d} \|\nabla \mathbf{v}\|_{(L^2(\Omega))^d}$ for any $\mathbf{u}, \mathbf{v}, \mathbf{w} \in [H^1(\Omega)]^d$. Here ρ is the Sobolev embedding constant defined by (11.7), whereas $M_c(\boldsymbol{\mu}) > 0$ is a $\boldsymbol{\mu}$ -dependent function to be determined from the affine parametrization of the trilinear form,

$$c(\mathbf{w}, \mathbf{u}, \mathbf{v}; \boldsymbol{\mu}) = \sum_{q=1}^{Q_c} \theta_c^q(\boldsymbol{\mu}) c_q(\mathbf{w}, \mathbf{u}, \mathbf{v})$$

where $\theta_c^q: \mathcal{P} \rightarrow \mathbb{R}$ and $c^q: V \times V \times V \rightarrow \mathbb{R}$, for each $q = 1, \dots, Q_c$.

2. a. Prove that the trilinear form defined over the original domain $\tilde{\Omega}(\boldsymbol{\mu})$

$$\tilde{c}(\tilde{\mathbf{w}}, \mathbf{u}, \mathbf{v}; \boldsymbol{\mu}) = \int_{\tilde{\Omega}(\boldsymbol{\mu})} (\mathbf{w} \cdot \nabla) \mathbf{u} \cdot \mathbf{v} d\Omega \quad (11.62)$$

verifies

$$\tilde{c}(\mathbf{w}, \mathbf{u}, \mathbf{v}; \boldsymbol{\mu}) = - \int_{\tilde{\Omega}(\boldsymbol{\mu})} \operatorname{div} \mathbf{w} \mathbf{u} \cdot \mathbf{v} d\tilde{\Omega} - \tilde{c}(\mathbf{w}, \mathbf{v}, \mathbf{u}; \boldsymbol{\mu}) + \int_{\partial \tilde{\Omega}(\boldsymbol{\mu})} (\mathbf{u} \cdot \mathbf{v}) (\mathbf{w} \cdot \tilde{\mathbf{n}}) d\tilde{\Gamma} \quad (11.63)$$

for any $\mathbf{u}, \mathbf{v}, \mathbf{w} \in (H^1(\tilde{\Omega}))^d$, being $\tilde{\mathbf{n}}$ the unit outward normal on $\partial \tilde{\Omega}$;

- b. provided homogeneous Dirichlet conditions are imposed on $\partial \tilde{\Omega}$, show that

$$\tilde{c}(\mathbf{w}, \mathbf{u}, \mathbf{v}; \boldsymbol{\mu}) = -\tilde{c}(\mathbf{w}, \mathbf{v}, \mathbf{u}; \boldsymbol{\mu}) \quad \forall \mathbf{w} : \operatorname{div}_{\tilde{\mathbf{x}}} \mathbf{w} = 0; \quad (11.64)$$

- c. show that for a mixed problem with homogeneous Dirichlet conditions on \tilde{F}_D and Neumann conditions on $\tilde{F}_N = \partial\tilde{\Omega} \setminus \tilde{F}_D$

$$\tilde{c}(\mathbf{w}, \mathbf{w}, \mathbf{w}; \boldsymbol{\mu}) = \frac{1}{2} \int_{\partial\tilde{\Omega}(\boldsymbol{\mu})} |\mathbf{w}|^2 (\mathbf{w} \cdot \tilde{\mathbf{n}}) d\tilde{\Gamma} \quad \forall \mathbf{w} : \operatorname{div}_{\tilde{\mathbf{x}}} \mathbf{w} = 0; \quad (11.65)$$

- d. consider a componentwise transformation preserving the divergence property of vector fields, thus fulfilling (8.47)–(8.48). By change of variables, show that the trilinear form (11.62) fulfills the equivalent properties (11.63)–(11.65) rewritten on the reference domain Ω .
3. Using the same notation of (11.4), introduce the following (linear) *Oseen problem*: given $\mathbf{w} \in X_0 = \{\mathbf{v} \in X : \operatorname{div} \mathbf{w} = 0\}$, find $(\mathbf{u}, p) \in X \times Q$ such that

$$\begin{cases} \bar{d}(\mathbf{u}, \mathbf{v}; \boldsymbol{\mu}) + c(\mathbf{w}, \mathbf{u}, \mathbf{v}; \boldsymbol{\mu}) + b(\mathbf{v}, p(\boldsymbol{\mu}); \boldsymbol{\mu}) = f_1(\mathbf{v}; \boldsymbol{\mu}) & \forall \mathbf{v} \in X \\ b(\mathbf{u}(\boldsymbol{\mu}), q; \boldsymbol{\mu}) = f_2(q; \boldsymbol{\mu}) & \forall q \in Q. \end{cases} \quad (11.66)$$

- a. For homogeneous Dirichlet conditions on the velocity field, show that problem (11.66) is well-posed and has a unique solution for any $\mathbf{w} \in X_0$;
- b. define $\Lambda(\boldsymbol{\mu}) : X \rightarrow X$ as the operator which for any $\mathbf{w} \in X_0$ gives the solution to (11.66), that is $\Lambda(\mathbf{w}; \boldsymbol{\mu}) = \mathbf{u}(\boldsymbol{\mu})$ and verify that the solution of the Navier-Stokes equations (11.4) is a fixed point of T ;
- c. show that $\|\Lambda(\mathbf{w}_1; \boldsymbol{\mu}) - \Lambda(\mathbf{w}_2; \boldsymbol{\mu})\|_X \leq \rho(\boldsymbol{\mu}) \|\mathbf{w}_1 - \mathbf{w}_2\|_X$ for a suitable $\rho(\boldsymbol{\mu}) > 0$;
- d. using the Banach fixed-point theorem, conclude that the problem (11.4) has a unique solution under the small data assumption (11.8).
4. Show that (11.9) holds.
5. Similarly to the expression of the reduced residual (11.25), show that the reduced Jacobian matrix $\mathbb{J}_N(\cdot; \boldsymbol{\mu})$ is given by

$$\mathbb{J}_N(\mathbf{u}_N^k; \boldsymbol{\mu}) = \mathbb{W}^T \mathbb{J}_h(\nabla \mathbf{u}_N^k; \boldsymbol{\mu}) \mathbb{V}$$

and express the Newton step (11.21) in the following algebraic form

$$\mathbb{J}_N(\mathbf{u}_N^k; \boldsymbol{\mu}) \delta \mathbf{u}_N = -\mathbf{G}_N(\mathbf{u}_N^k; \boldsymbol{\mu}).$$

6. Using a Taylor expansion, show that when $\|\mathbf{G}_h(\nabla \mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})\|_{\mathbb{X}_h^{-1}} \rightarrow 0$

$$\frac{\beta_h^N(\boldsymbol{\mu})}{K_h^N(\boldsymbol{\mu})} \left(1 - \sqrt{1 - \tilde{\tau}_N(\boldsymbol{\mu})}\right) \rightarrow \frac{\|\mathbf{G}_h(\nabla \mathbf{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})\|_{\mathbb{X}_h^{-1}}}{\beta_h^N(\boldsymbol{\mu})}.$$

Chapter 12

Reduction and Control

We exploit RB methods for the efficient solution of parameter-dependent PDE-constrained optimization problems. According to the optimization strategy adopted, these problems feature a very large size, in case a monolithic strategy is chosen, or huge computational cost due to the need of solving a parametrized PDE many times, when preferring an iterative optimization method. We concentrate on (i) *parametric* optimization problems, where control variables are described in terms of a vector of parameters, and (ii) *parametrized* optimal control problems, in which the parameters affect instead the state system and the control variables are functions to be determined. We propose efficient RB strategies to speedup the solution of these problems, pursuing either a (i) *state reduction* in the former case, or (ii) a simultaneous *state and control reduction* in the latter.

12.1 Parameter-Dependent PDE-Constrained Optimization

So far, we have considered the construction of RB methods for PDEs depending on a set of *input variables* expressed in terms of a vector of parameters. Very often, a PDE problem describes a *state system* to be controlled or optimized by varying some input variables. In these cases, the goal is to minimize (or maximize) some quantities of interest related to the underlying state variable, by acting on suitable control or design variables. This is the case of (i) *optimal control* problems, where we act on source/boundary terms or physical coefficients affecting the problem; (ii) *optimal design* problems, where the design variables are related with the geometric configuration of the domain; and (iii) *identification or data assimilation* problems, where some unknown or uncertain features of the state system are estimated by exploiting the measurements of some outputs. We refer to this rather general class of problems as to *PDE-constrained optimization* problems.

After the pioneering works by Ito and Ravindran in the late 90s [146, 147, 224], RB methods have been extensively applied to PDE-constrained optimization problems in the past two decades.

A (far from exhaustive) list includes several works where POD has been exploited to perform reduced-order modeling of optimal control [13, 158, 252, 136, 149] and optimal design [43, 9] problems; more recently, RB methods based on greedy algorithms have been successfully applied to optimal control [79, 248, 196, 195, 150, 90], and optimal design [163, 11, 186, 159, 9, 266] problems casted in a parameter-dependent framework; see also [25]. Another rapidly growing field of interest indeed very close to PDE-constrained optimization is represented by the use of ROM for inverse identification problems, see e.g. [121, 111, 63, 165, 75, 173].

In this chapter we indicate how a RB method can be used to tackle the (indeed, quite demanding) numerical approximation of optimal control problems, even if some ideas can be exploited to solve other PDE-constrained optimization problems. Consider the following state system:

$$\begin{cases} -\operatorname{div}(k\nabla u) + \mathbf{b} \cdot \nabla u = s & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega, \end{cases} \quad (12.1)$$

representing a slightly simplified version of the advection-diffusion equation (1.1), and the following three instances:

1. to control a heating (or cooling) system by acting on a distributed control function s so that a prescribed target temperature u_d is reached in a given subregion $\Omega_{obs} \subseteq \Omega$ of the domain; here $q = s$ plays the role of control variable;
2. to control the same system by seeking the best shape (optimal design) of the domain Ω ; in this case it is the shape of the domain $q = \Omega$ to play the role of control variable;
3. to identify the unknown conductivity field $k = k(\mathbf{x})$ from some measurements u_d of the temperature in Ω_{obs} ; here $q = k$ plays the role of control variable.

In all cases, we look for the minimum of the cost functional

$$J(u(q), q) = \frac{1}{2} \int_{\Omega_{obs}} |u(q) - u_d|^2 d\Omega + \frac{\sigma}{2} P(q) \quad (12.2)$$

where $\sigma \geq 0$ and $P(q)$ is a given, problem-dependent *penalization* or *regularization* term, often required to ensure the well-posedness of the PDE-constrained optimization problem. See e.g. [135, 251] for further details.

Our general notations are as follows: $u \in V$ denotes the state variable, that is, the solution of the PDE problem representing the state system

$$G(u, q) = 0 \quad \text{in } V' \quad (12.3)$$

(problem (12.1) in the example before), $q \in C$ the control variable, being C a suitable functional space called *control space*, $J : V \times C \rightarrow \mathbb{R}$ the cost functional, which has to be minimized in order to reach the objective; we will restrict to the case of quadratic functionals.

In an abstract form, an optimal control problem reads as

$$\text{find } \hat{q} = \arg \min_{q \in C} J(u(q), q) \quad \text{s.t.} \quad G(u, q) = 0. \quad (12.4)$$

The value $\hat{q} \in C$ is called optimal control; the corresponding state $u(\hat{q})$ is the optimal state. (12.4) represents a general constrained optimization problem, the state system playing the role of constraint. Ensuring the existence and uniqueness of a solution to problem (12.4) requires a detailed analysis which goes beyond the scope of this introduction. See e.g. [135, 125, 251, 33].

Among several parametrized PDE-constrained optimization problems, we focus in this chapter on two (indeed, relevant and quite general) classes of problems, depending on the role played by the parameter vector $\mu \in \mathcal{P}$:

1. *parametric optimization* problems; in this case, the control variable is a vector $q = \mu_c$ or a given function $q = q(\mu_c)$ of control parameters $\mu_c \in \mathcal{P}_c \subset \mathbb{R}^{P_c}$; this automatically allows to express a set of control constraints, indeed very simple to deal with. We assume that both G and J might depend also on a set of additional *scenario* parameters $\mu_s \in \mathcal{P}_s \subset \mathbb{R}^{P_s}$, characterizing the system being controlled. Consequently, $\mu = (\mu_s, \mu_c)$ and $\mathcal{P} = \mathcal{P}_s \times \mathcal{P}_c$. Hence, in this case (12.4) can be more precisely formulated as follows (omitting the q - and μ_s -dependence in u): given $\mu_s \in \mathcal{P}_s$, find

$$\hat{\mu}_c = \arg \min_{\mu_c \in \mathcal{P}_c} J(u, q(\mu_c); \mu_s) \quad \text{s.t.} \quad G(u, q(\mu_c); \mu_s) = 0; \quad (12.5)$$

2. *parametrized optimal control* problems; in this case, we only deal with a vector of *scenario* parameters $\mu_s \in \mathcal{P}_s$, while the control variable is not a function of the parameters. Here $\mu = \mu_s$ and $\mathcal{P} = \mathcal{P}_s$, so that in this case (12.4) reads as follows (omitting the q - and μ_s -dependence in u): given $\mu_s \in \mathcal{P}_s$, find

$$\hat{q} = \arg \min_{q \in C} J(u, q; \mu_s) \quad \text{s.t.} \quad G(u, q; \mu_s) = 0. \quad (12.6)$$

For instance, when $q = s$ plays the role of control variable in problem (12.1), we fall in class 1 if the control variable is assumed to be constant over Ω , so that $q = \mu_c$, or it is expressed as a linear combination of given functions $s_1, \dots, s_{P_c} \in L^2(\Omega)$, the control parameters being the weights, i.e.

$$q(\mu_c) = \sum_{i=1}^{P_c} \mu_c^{(i)} s_i(\mathbf{x}).$$

Instead, we fall in class 2 if no a priori assumptions on the expression of s are made, i.e. $q = s \in L^2(\Omega)$.

In the case of parametric optimization problems, a RB method operates a *state reduction* in order to solve the state system in a *reduced state space* for any new parameter vector. On the other hand, a simultaneous *state and control reduction* is instead required in the case of parametrized optimal control problems, where the control variable undergoes the same procedure adopted for achieving a low-dimensional approximation of the state variable. In this case, we need to build both a *reduced state space* and a *reduced control space* at the same time; since state and control variables are coupled through the state equation, suitable (trial and test) reduced spaces must be built.

Both classes of problems feature a *many-query* nature: usually we perform the optimization in (12.5) by relying on numerical iterative optimization algorithms, requiring many evaluations of the state system for different (i.e., updated at each iteration) values of the control parameters, once the state problem has been reduced; this naturally leads to consider a *reduce-then-optimize* approach. On the other hand, we usually perform optimization in (12.6) through the so-called *all-at-once* approach, which features the solution of a monolithic, coupled system obtained from the first-order optimality conditions, whenever no control constraints are imposed. This yields instead an *optimize-then-reduce* approach. In the latter case, the many-query nature is due to the huge number of scenarios in which we aim at controlling the system.

12.2 Parametric Optimization Problems

From now on we will directly work on the discrete, algebraic versions (after space discretization) of problem (12.5). In particular, we denote by \mathbf{u}_h the high-fidelity approximation of the state variable and by \mathbf{q}_h the high-fidelity approximation of the control variable, resulting from the introduction of suitable high-fidelity spaces $V_h \subset V$ and $C_h \subset C$ of dimension N_h and N_h^c , respectively.

Moreover, we denote by $\boldsymbol{\mu}_s \in \mathcal{P}_s$ a vector of scenario parameters and by $\boldsymbol{\mu}_c \in \mathcal{P}_c$ a vector of control parameters yielding a parameter-dependent expression of \mathbf{q}_h , say $\mathbf{q}_h = \mathbf{q}_h(\boldsymbol{\mu}_c)$. Possible control constraints are thus automatically accounted for in the definition of \mathcal{P}_c . As \mathcal{P}_c is a finite-dimensional space, a minimum point of the cost functional always exists provided this latter is at least continuous with respect to $\boldsymbol{\mu}_c$. Under these assumptions, we can cast several problems under the form of the following parametric optimization problem

$$\hat{\boldsymbol{\mu}}_{c,h} = \arg \min_{\boldsymbol{\mu}_c \in \mathcal{P}_c} J_h(\mathbf{u}_h, \mathbf{q}_h(\boldsymbol{\mu}_c); \boldsymbol{\mu}_s) \quad \text{s.t.} \quad \mathbf{G}_h(\mathbf{u}_h, \mathbf{q}_h(\boldsymbol{\mu}_c); \boldsymbol{\mu}_s) = \mathbf{0}. \quad (12.7)$$

Here, $J_h(\mathbf{u}_h, \mathbf{q}_h(\boldsymbol{\mu}_c); \boldsymbol{\mu}_s)$ results from the evaluation of the cost functional J on the high-fidelity approximation of the state variable¹, while $\mathbf{G}_h(\mathbf{u}_h, \mathbf{q}_h(\boldsymbol{\mu}_c); \boldsymbol{\mu}_s) \in \mathbb{R}^{N_h}$ is defined as in Sect. 11.2.2. Assuming to deal with a linear state system whose matrix is \mathbb{A}_h , we might consider:

- optimal control problems (see, e.g., [205, 220, 248]), where the control variable $\mathbf{q}_h(\boldsymbol{\mu}_c)$ is a (either distributed or boundary) function depending solely on $\boldsymbol{\mu}_c$. In this case

$$\mathbf{G}_h(\mathbf{u}_h, \mathbf{q}_h(\boldsymbol{\mu}_c); \boldsymbol{\mu}_s) = \mathbb{A}_h(\boldsymbol{\mu}_s) \mathbf{u}_h - \mathbf{f}_h(\boldsymbol{\mu}_s) - \mathbb{B}_h(\boldsymbol{\mu}_s) \mathbf{q}_h(\boldsymbol{\mu}_c) \quad (12.8)$$

and the optimal value $\hat{\boldsymbol{\mu}}_c \in \mathcal{P}_c$ yields the optimal control $\hat{\mathbf{q}}_h = \mathbf{q}_h(\hat{\boldsymbol{\mu}}_c)$. Here $\mathbb{B}_h \in \mathbb{R}^{N_h \times N_h^k}$ is the high-fidelity approximation of a suitable *control-to-state* operator, and the control variable affects only the right-hand side of the state problem;

- optimal design problems or identification problems, where the control variable is the shape of the domain or a physical coefficient, respectively. In this case

$$\mathbf{G}_h(\mathbf{u}_h, \mathbf{q}_h(\boldsymbol{\mu}_c); \boldsymbol{\mu}_s) = \mathbb{A}_h(\boldsymbol{\mu}_s, \mathbf{q}_h(\boldsymbol{\mu}_c)) \mathbf{u}_h - \mathbf{f}_h(\boldsymbol{\mu}_s, \mathbf{q}_h(\boldsymbol{\mu}_c));$$

both sides of the state problem thus depend on the control parameters, through suitable $\boldsymbol{\mu}_c$ -dependent functions related with either the tensors originated by the geometric parametrization – once the state problem has been rewritten on the reference domain – or the coefficient to be identified.

A crucial assumption in view of setting an efficient RB method is the affine dependence of the high-fidelity arrays with respect to both scenario and control parameters. Relying on the (D)EIM is very often required, for instance in the case of optimal design problems: describing a (possibly large) set of admissible shapes requires indeed complex geometric parametrizations, therefore depending on non-affinely parametrized tensors (with respect to $\boldsymbol{\mu}_c \in \mathcal{P}_c$); see e.g. [186, 159, 160].

Let us assume that $\boldsymbol{\mu}_c \mapsto u_h(\boldsymbol{\mu}_c; \boldsymbol{\mu}_s)$ is a differentiable mapping, and denote by

$$j_h(\boldsymbol{\mu}_c; \boldsymbol{\mu}_s) = J_h(\mathbf{u}_h(\boldsymbol{\mu}_c), \boldsymbol{\mu}_c; \boldsymbol{\mu}_s)$$

the so-called reduced cost functional. The constrained optimization problem (12.7) can be solved in many different ways. In the simplest case, a descent method such as the *projected* gradient method iteratively updates the control parameter $\boldsymbol{\mu}_c$ according to a sequence of descent directions depending on the gradient $\nabla_{\boldsymbol{\mu}_c} j_h(\boldsymbol{\mu}_c; \boldsymbol{\mu}_s)$ of the cost functional: for any $\boldsymbol{\mu}_s$, starting from $\boldsymbol{\mu}_{c,0} \in \mathcal{P}_c$ we define

$$\boldsymbol{\mu}_{c,k+1} = \Pi_{\mathcal{P}_c}(\boldsymbol{\mu}_{c,k} - \tau_k \nabla_{\boldsymbol{\mu}_c} j_h(\boldsymbol{\mu}_{c,k}; \boldsymbol{\mu}_s)), \quad k \geq 0,$$

until, e.g.,

$$\|\boldsymbol{\mu}_{c,k+1} - \boldsymbol{\mu}_{c,k}\|_{\mathbb{R}^{P_c}} \leq \varepsilon_{\text{OPT}}$$

for a prescribed tolerance $\varepsilon_{\text{OPT}} > 0$.

¹ In general $J_h = J$, unless some kind of approximations or quadrature formulas are introduced.

Here we denote by $\tau_k > 0$ a suitable stepsize, for any $k = 0, 1, \dots$, and by $\Pi_{\mathcal{P}_c}(\cdot)$ the projection² over \mathcal{P}_c . Note that, according to (11.3) (see Proposition 11.1) we have

$$\begin{aligned} \nabla_{\mu_c} j_h(\mu_c; \mu_s) &= \nabla_{\mu_c} J_h(\mathbf{u}_h, \mu_c; \mu_s) + \nabla_{\mathbf{u}_h} J_h(\mathbf{u}_h, \mu_c; \mu_s) \frac{\partial \mathbf{u}_h}{\partial \mu_c} \\ &= \nabla_{\mu_c} J_h(\mathbf{u}_h, \mu_c; \mu_s) \\ &\quad - \nabla_{\mathbf{u}_h} J_h(\mathbf{u}_h, \mu_c; \mu_s) (D_{\mathbf{u}_h} \mathbf{G}_h(\mathbf{u}_h, \mu_c; \mu_s))^{-1} D_{\mu_c} \mathbf{G}_h(\mathbf{u}_h, \mu_c; \mu_s), \end{aligned} \quad (12.9)$$

where $\mathbf{u}_h = \mathbf{u}_h(\mu_c)$. Unfortunately, since j_h depends on μ_c through the state variable $\mathbf{u}_h = \mathbf{u}_h(\mu_c)$, evaluating the gradient of the cost functional at each step of the minimization process requires indeed the solution of the state equation plus P_c sensitivity equations. These latter are given by

$$(D_{\mathbf{u}_h} \mathbf{G}_h(\mathbf{u}_h, \mu_c; \mu_s)) \frac{\partial \mathbf{u}_h}{\partial \mu_c^{(j)}} = \frac{\partial \mathbf{G}_h(\mathbf{u}_h, \mu_c; \mu_s)}{\partial \mu_c^{(j)}}, \quad j = 1, \dots, P_c. \quad (12.10)$$

A more computationally sound alternative consists in the introduction of the so-called *adjoint state*: denoting by $\mathbf{p}_h = \mathbf{p}_h(\mu_c) \in \mathbb{R}^{N_h}$ the solution of the following *adjoint problem*

$$(D_{\mathbf{u}_h} \mathbf{G}_h(\mathbf{u}_h, \mu_c; \mu_s))^T \mathbf{p}_h = -\nabla_{\mathbf{u}_h} J_h(\mathbf{u}_h, \mu_c; \mu_s),$$

we can express the gradient (12.9) as

$$\nabla_{\mu_c} j_h(\mu_c; \mu_s) = \nabla_{\mu_c} J_h(\mathbf{u}_h, \mu_c; \mu_s) + (D_{\mu_c} \mathbf{G}_h(\mathbf{u}_h, \mu_c; \mu_s))^T \mathbf{p}_h. \quad (12.11)$$

In this way, evaluating $\nabla_{\mu_c} j_h(\mu_c; \mu_s)$ at each step of the minimization process requires the solution of both the state equation and the adjoint problem. Note that the expression of $D_{\mu_c} \mathbf{G}_h(\mathbf{u}_h, \mu_c; \mu_s)$ just requires the evaluation of partial derivatives of the state operator with respect to μ_c , an inexpensive task under the assumption of affine parametric dependence, see e.g. (5.10). A general strategy to obtain the expression of the adjoint problem, exploiting the Lagrange multipliers method, will be proposed in the following section.

Remark 12.1. In the case of the state equation (12.8), if

$$J_h(\mathbf{u}_h, \mu_c; \mu_s) = \frac{1}{2} (\mathbf{u}_h - \mathbf{u}_{dh})^T \mathbb{F}_h(\mu_s) (\mathbf{u}_h - \mathbf{u}_{dh}) + \frac{\sigma}{2} \mathbf{q}_h^T(\mu_c) \mathbb{G}_h(\mu_s) \mathbf{q}_h(\mu_c) \quad (12.12)$$

where $\mathbb{G}_h(\mu_s) \in \mathbb{R}^{N_h^c \times N_h^c}$ is symmetric positive definite for any $\mu_s \in \mathcal{P}_s$, we have

$$\nabla_{\mu_c} j_h(\mu_c; \mu_s) = \sigma \mathbb{G}_h(\mu_s) (D_{\mu_c} \mathbf{q}_h) \mathbf{q}_h + \mathbb{B}_h^T(\mu_s) (D_{\mu_c} \mathbf{q}_h) \mathbf{p}_h$$

² The projection $\bar{\mu} = P_{\mathcal{P}_c}(\mu)$ of $\mu \in \mathbb{R}^P$ on the box set $\mathcal{P}_c = \prod_{j=1}^P [\mu_{\min}^{(j)}, \mu_{\max}^{(j)}]$ takes the following form: $\bar{\mu}^{(j)} = \min\{\mu_{\max}^{(j)}, \max(\mu_{\min}^{(j)}, \mu^{(j)})\}$, $j = 1, \dots, P$.

being $\mathbf{p}_h = \mathbf{p}_h(\boldsymbol{\mu}_c; \boldsymbol{\mu}_s)$ (called adjoint-state) the solution of the adjoint problem

$$\mathbb{A}_h^T(\boldsymbol{\mu}_s) \mathbf{p}_h = \mathbb{F}_h(\boldsymbol{\mu}_s)(\mathbf{u}_{dh} - \mathbf{u}_h).$$

Note that $\mathbb{F}_h(\boldsymbol{\mu}_s) \in \mathbb{R}^{N_h \times N_h}$ encodes the objective of our optimization; this is, e.g., the mass matrix (restricted to Ω_{obs}) in the case of the cost functional (12.2). •

Evaluating $\nabla_{\boldsymbol{\mu}_c} j_h(\boldsymbol{\mu}_c; \boldsymbol{\mu}_s)$ is thus an intrusive task, requiring the solution of at least a further (adjoint) problem; more general algorithms for nonlinear optimization – such as *sequential quadratic programming* – require also the Hessian of the cost functional, thus entailing even higher costs.

A possible alternative is to rely on suitable approximation, based e.g. on finite differences, of the gradient of the cost functional. This leads to a *black-box* strategy for the solution of the optimization problem, that “simply” requires the solution of the state solution and the evaluation of the cost functional; more ad hoc *derivative-free* methods, could also be considered, see e.g. [200] for further details.

Last, but not least, this allows to treat in the same way any parametric optimization problems – no matter whether stemming from optimal control, optimal design or identification problems – relying on simple and nonintrusive code routines.

Remark 12.2. For the parametrized optimal control problem (12.6) a black box approach would be infeasible. In this case, the (discretized version) of the control variable belongs to a large dimensional space (e.g., \mathbb{R}^{N_h} when dealing with a distributed control) thus making the update of the control variable impossible without any information on the gradient of the cost functional. As we will see in Sect. 12.4, solving an adjoint problem is mandatory in this case. •

12.2.1 Reduction Strategies

Following a black box approach, the solution of a parametric optimization problem requires several evaluations of the cost functional, each of these entailing the solution of the state equation, until convergence. This is even more expensive if the system has to be optimized in many different scenarios, requiring this procedure to be run for each $\boldsymbol{\mu}_s \in \mathcal{P}_s$.

We can take advantage of a RB method, by replacing the high-fidelity solver of the state equation with the corresponding RB approximation. The construction of a RB method for the state problem is performed according to the approaches described in the previous chapters; depending on its nature, either the G-RB or the LS-RB methods can be employed. Similarly to (11.25), the reduced state equation is given by

$$\mathbf{G}_N(\mathbf{u}_N(\boldsymbol{\mu}), \mathbf{q}_h(\boldsymbol{\mu}_c); \boldsymbol{\mu}_s) = \mathbb{W}^T \mathbf{G}_h(\mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}), \mathbf{q}_h(\boldsymbol{\mu}_c); \boldsymbol{\mu}_s) \quad (12.13)$$

expressing the state solution as a linear combination of RB functions $\mathbf{u}_h(\boldsymbol{\mu}_c, \boldsymbol{\mu}_s) \approx \mathbb{V} \mathbf{u}_N(\boldsymbol{\mu}_c, \boldsymbol{\mu}_s)$; here we denote by $\mathbb{V}, \mathbb{W} \in \mathbb{R}^{N_h \times N}$ the corresponding transformations

matrix. The columns of \mathbb{V} can be for instance suitably orthogonalized state solutions computed in correspondence to some selected $\boldsymbol{\mu} = (\boldsymbol{\mu}_c, \boldsymbol{\mu}_s) \in \mathcal{P}_c \times \mathcal{P}_s$; these latter can be chosen by running a (weak) greedy algorithm, according to a suitable a posteriori error bound (see e.g. Sect. 3.6 in the case of a linear state problem, of Sect. 11.5 in the case of a nonlinear state problem).

Note that no reduced basis is needed for the control variable \mathbf{q}_h : the reduction of the high-fidelity arrays containing \mathbf{q}_h can be performed according to the usual off-line/online decomposition, once an affine parametric structure has been recovered, possibly through a (D)EIM pre-processing phase. Moreover, we assume that it is possible to efficiently evaluate the cost functional

$$J_N(\mathbf{u}_N, \boldsymbol{\mu}_c; \boldsymbol{\mu}_s) = J_h(\mathbb{V}\mathbf{u}_N, \boldsymbol{\mu}_c; \boldsymbol{\mu}_s),$$

and consequently

$$j_N(\boldsymbol{\mu}_c; \boldsymbol{\mu}_s) = J_N(\mathbf{u}_N, \boldsymbol{\mu}_c; \boldsymbol{\mu}_s).$$

For instance, in the case of the cost functional (12.12), we have

$$J_h(\mathbb{V}\mathbf{u}_N, \boldsymbol{\mu}_c; \boldsymbol{\mu}_s) = \frac{1}{2} \mathbf{u}_N^T \mathbb{F}_N(\boldsymbol{\mu}_s) \mathbf{u}_N - \mathbf{d}_N(\boldsymbol{\mu}_s)^T \mathbb{V} \mathbf{u}_N + \frac{1}{2} e(\boldsymbol{\mu}_c, \boldsymbol{\mu}_s),$$

where

$$\begin{aligned} \mathbb{F}_N(\boldsymbol{\mu}_s) &= \mathbb{V}^T \mathbb{F}_h(\boldsymbol{\mu}_s) \mathbb{V} \in \mathbb{R}^{N \times N}, & \mathbf{d}_N(\boldsymbol{\mu}_s) &= \mathbf{u}_{dh}^T \mathbb{F}_h(\boldsymbol{\mu}_s) \mathbb{V} \in \mathbb{R}^N, \\ e(\boldsymbol{\mu}_c, \boldsymbol{\mu}_s) &= \mathbf{u}_{dh}^T \mathbb{F}_h(\boldsymbol{\mu}_s) \mathbf{u}_{dh} + \boldsymbol{\sigma} \mathbf{q}_h^T(\boldsymbol{\mu}_c) \mathbb{G}_h(\boldsymbol{\mu}_s) \mathbf{q}_h(\boldsymbol{\mu}_c) \in \mathbb{R} \end{aligned}$$

can be easily evaluated provided the high-fidelity arrays feature an affine expansion.

We end up with the following reduced parametric optimization problem

$$\hat{\boldsymbol{\mu}}_{c,N} = \arg \min_{\boldsymbol{\mu}_c \in \mathcal{P}_c} J_N(\mathbf{u}_N, \boldsymbol{\mu}_c; \boldsymbol{\mu}_s) \quad \text{s.t.} \quad \mathbf{G}_N(\mathbf{u}_N, \mathbf{q}_h(\boldsymbol{\mu}_c); \boldsymbol{\mu}_s) = \mathbf{0}. \quad (12.14)$$

The former is solved by a *reduce-then-optimize* strategy: first, during the offline stage we perform the reduction of the state equation, by taking into account both control and scenario parameters. Then, for any new scenario characterized by $\boldsymbol{\mu}_s \in \mathcal{P}_s$, we perform online the numerical optimization according to a black-box strategy, see the block representation of Fig. 12.1.

Alternatively, a descent method could be adopted for the optimization at the on-line stage, too. However, this requires the construction of a RB approximation for the adjoint problem, thus entailing additional costs; see e.g. [184] for further details.

In the same way, the evaluation of an a posteriori estimate for the error on the cost functional $|j_h(\boldsymbol{\mu}_c; \boldsymbol{\mu}_s) - j_N(\boldsymbol{\mu}_c; \boldsymbol{\mu}_s)|$ requires the reduction of the adjoint problem; a similar bound for the gradient $\|\nabla_{\boldsymbol{\mu}_c} j_h(\boldsymbol{\mu}_c; \boldsymbol{\mu}_s) - \nabla_{\boldsymbol{\mu}_c} j_N(\boldsymbol{\mu}_c; \boldsymbol{\mu}_s)\|_{\mathbb{R}^{p_c}}$ requires the reduced approximation of the sensitivity equations (12.10). The interested reader can refer, e.g., to [91, 90, 150, 184] for further details.

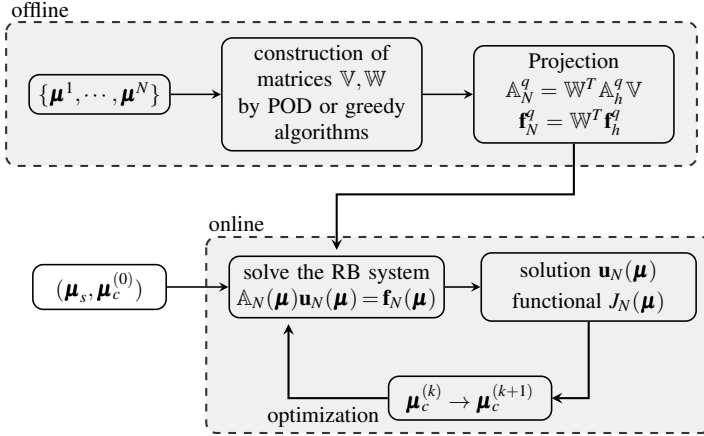


Fig. 12.1 Reduce-then-optimize approach for parametric optimization problems in the case of a linear state system

12.3 Application to an Optimal Flow Control Problem

We apply the method derived in the previous section to solve an optimal flow control problem for the vorticity minimization through boundary control, in presence of different scenarios. This problem may occur, e.g., when designing the shape of a bypass graft to restore blood perfusion downfield an occluded coronary artery. Here we aim at minimizing the flow vorticity downstream the anastomosis – that is, in proximity of the junction between the graft and the host occluded artery.

Many works (see, e.g., [186, 159, 160] and references therein) have focused on the optimal shape design of end-to-side anastomoses, typically by acting on the wall shape near the anastomosis by local shape variations. Here we consider an extremely simplified version of this problem: following [27, 127], instead of acting on the shape of the graft we solve an optimal control problem for which the boundary control represents the flow entering into the artery from the graft.

We consider³ the setting of Sect. 11.9, where the fluid flow has been parametrized with respect to the angle $\theta \in [15^\circ, 85^\circ]$ of incidence of the inflow through the bypass, and the intensity $\omega \in [0, 3]$ of the residual flow through the (totally, or partially) occluded artery. Since one of the most significant design variables in end-to-side anastomoses is the angle between the graft and the host artery, we take $\mu_c = \theta$ as *control parameter*; moreover, in order to find the optimal bypass angle within a range of different residual flows, we select $\mu_s = \omega$ as *scenario parameter*.

³ Modeling a blood flow with a steady equation is of course a great simplification, because of flow pulsatility. The main purpose here however is to illustrate the way RB methods play in the context of an optimal control problem, rather than improving the real life significance of the mathematical model.



Fig. 12.2 Domain and observation region for the optimal flow control problem; see Fig. 11.6 for the definition of the boundary portions

The geometric properties of the bypass graft are resumed into the velocity profile $\mathbf{g}_C = \mathbf{g}_C(\mu_c, \mu_s)$ imposed on Γ_C , which has to be controlled in order to minimize the following cost functional, representing the magnitude of the vorticity of the velocity field on the observation region Ω_{obs} (see Fig. 12.2)

$$J(\mathbf{u}) = \frac{1}{2} \int_{\Omega_{obs}} |\nabla \times \mathbf{u}|^2 d\Omega \quad (12.15)$$

where $\mathbf{u} = \mathbf{u}(\mu_c, \mu_s)$ is the fluid velocity. This latter is the solution of the state (Navier-Stokes) problem on a frozen, fixed domain the one given by the cylinder Ω representing the occluded artery, see Fig. 11.6. Hence, the problem is recast under the form (12.5) provided we identify J with the functional (12.15) and G as the weak formulation (11.4) of the Navier-Stokes equations; here $q = \mathbf{g}_C$ is the control function, defined in terms of the control parameter μ_c .

We solve the problem following the *reduce-then-optimize* approach, by taking advantage of the RB approximation of the state Navier-Stokes problem described in Sect. 11.9. In particular, the RB velocity and pressure spaces have dimension $N = 28$; the RB approximation of the state problem takes 0.1s, requiring a nonlinear system of dimension 56×56 to be solved.

We perform a *black-box* optimization using the Matlab `fmincon` function, implementing a sequential quadratic programming method; we do not provide any information about the gradient of the cost functional, which is indeed evaluated by a finite difference approximation.

For any fixed value of $\mu_s = \omega$, the online optimization requires about 15 evaluations of the cost functional $J_N(\mathbf{u}_N(\mu_c, \mu_s))$ and takes about 2s. We thus manage to compute the optimal angle $\hat{\mu}_c$ of the bypass in a very short amount of time, for any different scenario, thus spanning a wide range of different flow conditions, where the velocity pattern (and, correspondingly, the vorticity pattern) might vary significantly; see Figs. 12.3–12.4.

In particular, we obtain an optimal angle $\hat{\mu}_c = 41.8$ in the case of total occlusion ($\omega = 0$) and $\hat{\mu}_c = 33.9$ in the case of partial occlusion ($\omega = 3$). The plot of the cost functional with respect to the anastomosis angle θ , as well as the optimal angles in several scenario, are reported in Fig. 12.5.

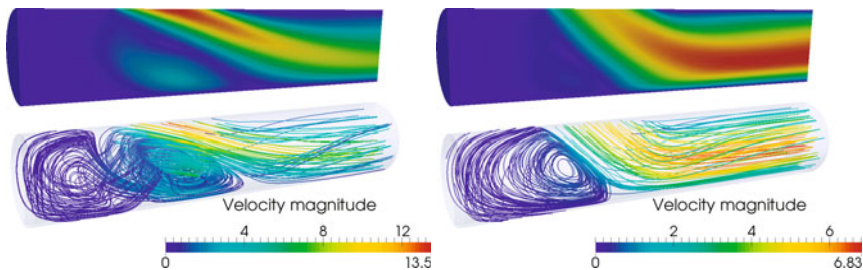


Fig. 12.3 Velocity field (magnitude and streamlines) corresponding to the initial value $\mu_{c,0}$ (left) and to the optimal value $\hat{\mu}_c = 41.8$ of the control parameter (right), in the case $\mu_s = \omega = 0$ (total occlusion)

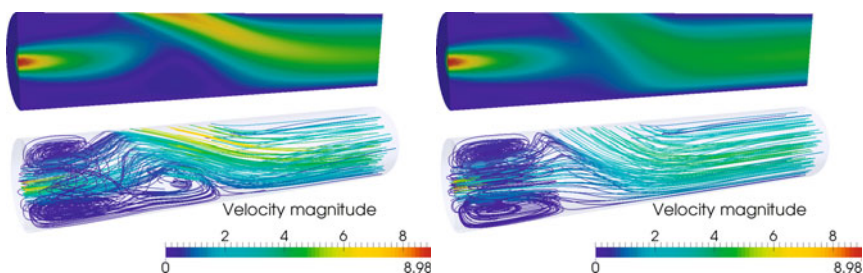


Fig. 12.4 Velocity field (magnitude and streamlines) corresponding to the initial value $\mu_{c,0}$ (left) and to the optimal value $\hat{\mu}_c = 33.9$ of the control parameter (right), in the case $\mu_s = \omega = 3$ (partial occlusion)

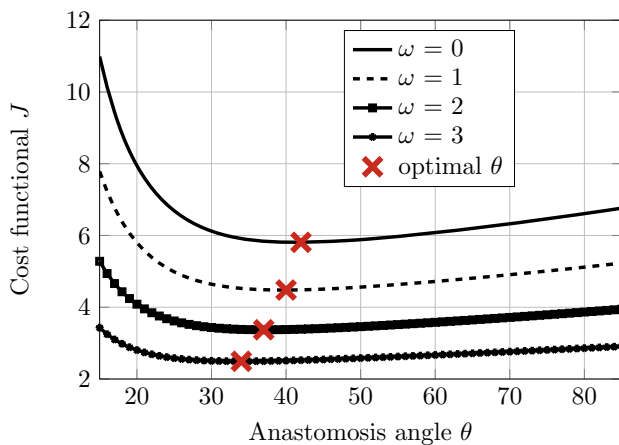


Fig. 12.5 Cost functional and optimal angles in presence of different degrees of occlusion; the optimal angle decreases for increasing magnitude of the residual flow

Other examples of parametric optimization problems can be found, e.g., in [43, 163, 266] aimed at the optimal design of an airfoil, concerning the optimal control of thermal flows [232], for the optimal design for a Stokes flow [186] and robust optimal design for a Navier-Stokes flow in presence of uncertain parameters [159].

12.4 Parametrized Optimal Control Problems

We now turn to a second class of problems, namely parametrized optimal control problems. Let us denote by $\boldsymbol{\mu}_s \in \mathcal{P}_s$ a vector of scenario parameters representing either physical or geometric quantities; these are the only parameters we deal with, since in this case control functions are not parametrized.

Also in this case we deal with a purely algebraic formulation, considering for the sake of simplicity a linear state system; see e.g. [195] for an extension to fluid flows. Let us denote by $\mathbf{u}_h \in \mathbb{R}^{N_h}$ the high-fidelity approximation of the state variable, being $V_h \subset V$ the discrete state space of dimension $N_h = \dim(V_h)$, and by $\mathbf{q}_h \in \mathbb{R}^{N_h^c}$ the high-fidelity approximation of the control variable, resulting from the introduction of a suitable high-fidelity space $C_h \subset C$ of dimension N_h^c . In particular we assume that C is a Hilbert space and we do not consider control constraints, see e.g. [80].

From a general standpoint, omitting the \mathbf{q}_h -dependence in \mathbf{u}_h , a parametrized optimal control problem reads

$$\hat{\mathbf{q}}_h = \arg \min_{\mathbf{q}_h \in \mathbb{R}^{N_h^c}} J_h(\mathbf{u}_h, \mathbf{q}_h; \boldsymbol{\mu}_s) \quad \text{s.t.} \quad \mathbf{G}_h(\mathbf{u}_h, \mathbf{q}_h; \boldsymbol{\mu}_s) = \mathbf{0}, \quad (12.16)$$

where $J_h(\mathbf{u}_h, \mathbf{q}_h; \boldsymbol{\mu}_s)$ results from the evaluation of the cost functional on the high-fidelity approximation of the state variable. In particular, we deal with a linear state equation, defining

$$\mathbf{G}_h(\mathbf{u}_h, \mathbf{q}_h; \boldsymbol{\mu}_s) = \mathbb{A}_h(\boldsymbol{\mu}_s)\mathbf{u}_h - \mathbf{f}_h(\boldsymbol{\mu}_s) - \mathbb{B}_h(\boldsymbol{\mu}_s)\mathbf{q}_h, \quad (12.17)$$

although the whole framework can be adapted to the case of nonlinear state equations, too. Here $\mathbb{B}_h \in \mathbb{R}^{N_h \times N_h^c}$ is the high-fidelity approximation of a suitable *control-to-state* operator, and the control variable affects only the right-hand side of the state problem. Moreover, we assume the matrix $\mathbb{A}_h(\boldsymbol{\mu}_s)$ to be positive definite, i.e. we consider a strongly coercive state problem. The quadratic cost functional J_h has the following general form,

$$J_h(\mathbf{u}_h, \mathbf{q}_h; \boldsymbol{\mu}_s) = \frac{1}{2}(\mathbf{u}_h - \mathbf{u}_{dh}(\boldsymbol{\mu}_s))^T \mathbb{F}_h(\boldsymbol{\mu}_s)(\mathbf{u}_h - \mathbf{u}_{dh}(\boldsymbol{\mu}_s)) + \frac{\sigma}{2} \mathbf{q}_h^T \mathbb{G}_h(\boldsymbol{\mu}_s) \mathbf{q}_h, \quad (12.18)$$

where $\mathbf{u}_{dh}(\boldsymbol{\mu}_s)$ is (the high-fidelity discretization of) a given desired state, $\mathbb{F}_h(\boldsymbol{\mu}_s)$ is a positive semidefinite matrix defining the objective of the minimization, depending

a priori on $\boldsymbol{\mu}_s$, $\sigma > 0$ is a given penalization constant, while $\mathbb{G}_h(\boldsymbol{\mu}_s)$ is a symmetric positive definite matrix.

We can now derive a system of first-order necessary optimality conditions in order to characterize the solution to problem (12.16); see, e.g. [135, 251] for a detailed description. By introducing a Lagrange multiplier $\mathbf{p}_h \in \mathbb{R}^{N_h}$, we can define a Lagrangian functional that incorporates the PDE constraint,

$$\mathcal{L}_h(\mathbf{u}_h, \mathbf{q}_h, \mathbf{p}_h; \boldsymbol{\mu}_s) = J_h(\mathbf{u}_h, \mathbf{q}_h; \boldsymbol{\mu}_s) + \mathbf{p}_h^T (\mathbb{A}_h(\boldsymbol{\mu}_s)\mathbf{u}_h - \mathbb{B}_h(\boldsymbol{\mu}_s)\mathbf{q}_h - \mathbf{f}_h(\boldsymbol{\mu}_s)).$$

A constrained minimum for the cost functional is also a solution for the unconstrained minimization problem for the Lagrangian functional. Hence, the necessary optimality conditions associated with (12.16) can be obtained by imposing the stationarity of \mathcal{L}_h ; this yields:

$$\begin{cases} \nabla_{\mathbf{u}_h} \mathcal{L}_h(\mathbf{u}_h, \mathbf{q}_h, \mathbf{p}_h; \boldsymbol{\mu}_s) = \mathbf{0} & \Leftrightarrow \mathbb{F}_h(\boldsymbol{\mu}_s)(\mathbf{u}_h - \mathbf{u}_{dh}(\boldsymbol{\mu}_s)) + \mathbb{A}_h^T(\boldsymbol{\mu}_s)\mathbf{u}_h = \mathbf{0} \\ \nabla_{\mathbf{q}_h} \mathcal{L}_h(\mathbf{u}_h, \mathbf{q}_h, \mathbf{p}_h; \boldsymbol{\mu}_s) = \mathbf{0} & \Leftrightarrow \sigma \mathbb{G}_h(\boldsymbol{\mu}_s)\mathbf{q}_h - \mathbb{B}_h^T(\boldsymbol{\mu}_s)\mathbf{q}_h = \mathbf{0} \\ \nabla_{\mathbf{p}_h} \mathcal{L}_h(\mathbf{u}_h, \mathbf{q}_h, \mathbf{p}_h; \boldsymbol{\mu}_s) = \mathbf{0} & \Leftrightarrow \mathbb{A}_h(\boldsymbol{\mu}_s)\mathbf{u}_h - \mathbb{B}_h(\boldsymbol{\mu}_s)\mathbf{q}_h - \mathbf{f}_h(\boldsymbol{\mu}_s) = \mathbf{0}. \end{cases} \quad (12.19)$$

In the case at hand – quadratic functionals and linear state equations – conditions (12.19) are also sufficient for the optimality. In a more compact form, we can rewrite system (12.19) yielding the following system of equations:

$$\begin{pmatrix} \mathbb{F}_h(\boldsymbol{\mu}_s) & 0 & \mathbb{A}_h^T(\boldsymbol{\mu}_s) \\ 0 & \sigma \mathbb{G}_h(\boldsymbol{\mu}_s) - \mathbb{B}_h^T(\boldsymbol{\mu}_s) \\ \mathbb{A}_h(\boldsymbol{\mu}_s) - \mathbb{B}_h(\boldsymbol{\mu}_s) & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_h \\ \mathbf{q}_h \\ \mathbf{p}_h \end{pmatrix} = \begin{pmatrix} \mathbb{F}_h(\boldsymbol{\mu}_s)\mathbf{u}_{dh}(\boldsymbol{\mu}_s) \\ \mathbf{0} \\ \mathbf{f}_h(\boldsymbol{\mu}_s) \end{pmatrix}. \quad (12.20)$$

Equivalently,

$$\mathbb{K}_h(\boldsymbol{\mu}_s)\mathbf{U}_h = \mathbf{F}_h(\boldsymbol{\mu}_s) \quad (12.21)$$

where $\mathbf{U}_h = (\mathbf{u}_h, \mathbf{q}_h, \mathbf{p}_h) \in \mathbb{R}^{2N_h + N_h^c}$ and $\mathbb{K}_h(\boldsymbol{\mu}_s) \in \mathbb{R}^{(2N_h + N_h^c) \times (2N_h + N_h^c)}$.

12.4.1 Reduction Strategies

To approximate the optimal solutions $(\mathbf{u}_h(\boldsymbol{\mu}_s), \mathbf{q}_h(\boldsymbol{\mu}_s))$ we directly build a RB approximation of the high-fidelity optimality system (12.20). Following the general methodology described in Chap. 3, we approximate the state, control and adjoint variables as

$$\mathbf{u}_h \approx \mathbb{V}_u \mathbf{u}_N, \quad \mathbf{q}_h \approx \mathbb{V}_q \mathbf{q}_N, \quad \mathbf{p}_h \approx \mathbb{V}_p \mathbf{p}_N,$$

where $\mathbb{V}_u, \mathbb{V}_p \in \mathbb{R}^{N_h \times N}$, $\mathbb{V}_q \in \mathbb{R}^{N_h^c \times N}$ denote the transformation matrices for the state, the adjoint and the control variables, respectively.

The columns of \mathbb{V}_u (respectively $\mathbb{V}_q, \mathbb{V}_p$) can be for instance suitably orthogonalized state (resp. control, adjoints) solutions of (12.16) computed for some parameters $\boldsymbol{\mu}_s \in \mathcal{P}_s$; these latter can be selected using a (weak) greedy algorithm, according to a suitable a posteriori error bound (see Sect. 12.4.2). We also introduce suitable test bases $\mathbb{W}_u, \mathbb{W}_p \in \mathbb{R}^{N_h \times N}$ for the state and adjoint variables.

By enforcing the orthogonality of the residual of (12.20) to the product basis $\mathbb{W}_p \times \mathbb{V}_q \times \mathbb{W}_u$ we obtain the reduced optimality system (omitting $\boldsymbol{\mu}_s$):

$$\begin{pmatrix} \mathbb{W}_p^T \mathbb{F}_h \mathbb{V}_u & 0 & \mathbb{W}_p^T \mathbb{A}_h^T \mathbb{V}_p \\ 0 & \sigma \mathbb{V}_q^T \mathbb{G}_h \mathbb{V}_q & -\mathbb{V}_q^T \mathbb{B}_h \mathbb{V}_p \\ \mathbb{W}_u^T \mathbb{A}_h \mathbb{V}_u & -\mathbb{W}_u^T \mathbb{B}_h \mathbb{V}_q & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_N \\ \mathbf{q}_N \\ \mathbf{p}_N \end{pmatrix} = \begin{pmatrix} \mathbb{W}_p^T \mathbb{F}_h \mathbf{u}_d \\ \mathbf{0} \\ \mathbb{W}_u^T \mathbf{f}_h \end{pmatrix}. \quad (12.22)$$

The choice of the test bases \mathbb{W}_u and \mathbb{W}_p is crucial to ensure the stability of the reduced approximation and the symmetry of the reduced matrix, in order to guarantee that (12.22) corresponds to the stationarity condition for a suitable reduced Lagrangian. A possible strategy is to define an aggregated trial space [79, 196, 150] made of both state and adjoint snapshots, and to employ a G-RB method to find the reduced approximation. Hence, we define

$$\mathbb{V}_{up} = [\mathbb{V}_u \ \mathbb{V}_p] \in \mathbb{R}^{N_h \times 2N}$$

and we set

$$\mathbb{V} = \mathbb{W} = \begin{pmatrix} \mathbb{V}_{up} & 0 & 0 \\ 0 & \mathbb{V}_q & 0 \\ 0 & 0 & \mathbb{V}_{up} \end{pmatrix}.$$

The G-RB approximation of problem (12.20) then seeks $\mathbf{U}_N(\boldsymbol{\mu}_s) \in \mathbb{R}^{5N}$ satisfying

$$\mathbb{V}^T \mathbb{K}_h(\boldsymbol{\mu}_s) \mathbb{V} \mathbf{U}_N(\boldsymbol{\mu}_s) = \mathbb{V}^T \mathbf{F}_h(\boldsymbol{\mu}_s), \quad (12.23)$$

that is,

$$\mathbb{K}_N(\boldsymbol{\mu}_s) \mathbf{U}_N(\boldsymbol{\mu}_s) = \mathbf{F}_N(\boldsymbol{\mu}_s), \quad (12.24)$$

being

$$\mathbb{K}_N(\boldsymbol{\mu}_s) = \mathbb{V}^T \mathbb{K}_h(\boldsymbol{\mu}_s) \mathbb{V}, \quad \mathbf{F}_N(\boldsymbol{\mu}_s) = \mathbb{V}^T \mathbf{F}_h(\boldsymbol{\mu}_s).$$

Note that these latter express nothing but the relations (3.59) in the case of the G-RB approximation of the problem (12.21).

The G-RB approximation (12.24) of the optimality system is *consistent*, meaning that full-order state, control and adjoint solutions are recovered by the RB method if the corresponding snapshots are contained in the reduced spaces. Moreover, the reduced system (12.24) can be obtained by differentiating the reduced Lagrangian

$$\mathcal{L}_N(\mathbf{u}_N, \mathbf{q}_N, \mathbf{p}_N; \boldsymbol{\mu}_s) = \mathcal{L}_h(\mathbb{V} \mathbf{u}_N, \mathbb{V} \mathbf{q}_N, \mathbb{V} \mathbf{p}_N; \boldsymbol{\mu}_s).$$

Indeed, a Galerkin projection over aggregated spaces enables to obtain a symmetric optimality system which preserves the Lagrangian structure.

Once the projection bases have been defined, the whole RB methodology discussed so far still applies. In particular, the selection of the RB functions can be performed via either the POD or greedy algorithm, and the affine parametric dependence of $\mathbb{K}_h(\boldsymbol{\mu}_s)$ and $\mathbf{F}_h(\boldsymbol{\mu}_s)$ enables the usual offline/online decomposition.

Remark 12.3. In principle, a *reduce-then-optimize* paradigm could be adopted also in the case of parametrized optimal control problems, thus seeking an approximate solution to (12.16) of the form $\mathbf{u}_h \approx \mathbb{V}_u \mathbf{u}_N$, $\mathbf{q}_h \approx \mathbb{V}_q \mathbf{q}_N$, where $\mathbb{V}_u \in \mathbb{R}^{N_h \times N}$, $\mathbb{V}_q \in \mathbb{R}^{N_c \times N}$ encode the state and the control trial bases, respectively. This would yield the following reduced parametrized optimal control problem

$$\hat{\mathbf{q}}_N = \arg \min_{\mathbf{q}_N \in \mathbb{R}^N} J_N(\mathbf{u}_N, \mathbf{q}_N; \boldsymbol{\mu}_s) \quad \text{s.t.} \quad \mathbf{G}_N(\mathbf{u}_N, \mathbf{q}_N; \boldsymbol{\mu}_s) = \mathbf{0}. \quad (12.25)$$

When considering the quadratic cost functional (12.18) – with $\mathbf{u}_{dh} = \mathbf{0}$ for simplicity – and the linear state equation (12.17), we obtain

$$J_N(\mathbf{u}_N, \mathbf{q}_N; \boldsymbol{\mu}_s) = \frac{1}{2} \mathbf{u}_N^T \mathbb{F}_N(\boldsymbol{\mu}_s) \mathbf{u}_N + \frac{\sigma}{2} \mathbf{q}_N^T \mathbb{G}_N(\boldsymbol{\mu}_s) \mathbf{q}_N$$

$$\mathbf{G}_N(\mathbf{u}_N, \mathbf{q}_N; \boldsymbol{\mu}_s) = \mathbb{A}_N(\boldsymbol{\mu}_s) \mathbf{u}_N - \mathbf{f}_N(\boldsymbol{\mu}_s) - \mathbb{B}_N(\boldsymbol{\mu}_s) \mathbf{q}_N$$

where \mathbb{W}_u encodes the state test basis, $\mathbf{f}_N = \mathbb{W}_u^T \mathbf{f}_h$ and

$$\mathbb{A}_N = \mathbb{W}_u^T \mathbb{A}_h \mathbb{V}_u, \quad \mathbb{B}_N = \mathbb{W}_u^T \mathbb{B}_h \mathbb{V}_q, \quad \mathbb{F}_N = \mathbb{V}_u^T \mathbb{F}_h \mathbb{V}_u, \quad \mathbb{G}_N = \mathbb{V}_q^T \mathbb{G}_h \mathbb{V}_q.$$

By deriving the reduced Lagrangian

$$\begin{aligned} \mathcal{L}_N(\mathbf{u}_N, \mathbf{q}_N, \mathbf{p}_N; \boldsymbol{\mu}_s) &= \frac{1}{2} \mathbf{u}_N^T \mathbb{F}_N(\boldsymbol{\mu}_s) \mathbf{u}_N + \frac{\sigma}{2} \mathbf{q}_N^T \mathbb{G}_N(\boldsymbol{\mu}_s) \mathbf{q}_N \\ &\quad + \mathbf{p}_N^T (\mathbb{A}_N(\boldsymbol{\mu}_s) \mathbf{u}_N - \mathbb{B}_N(\boldsymbol{\mu}_s) \mathbf{q}_N - \mathbf{f}_N(\boldsymbol{\mu}_s)), \end{aligned}$$

we obtain (omitting $\boldsymbol{\mu}_s$) the following reduced optimality system associated to (12.25)

$$\begin{pmatrix} \mathbb{F}_N & 0 & \mathbb{A}_N^T \\ 0 & \sigma \mathbb{G}_N & -\mathbb{B}_N^T \\ \mathbb{A}_N & -\mathbb{B}_N & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_N \\ \mathbf{q}_N \\ \mathbf{p}_N \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{f}_N \end{pmatrix}. \quad (12.26)$$

In this case, the approximation of the adjoint variable is determined by the choice of the test space used in the projection of the state equation, i.e. we seek for an approximate adjoint variable $\mathbf{p}_h \approx \mathbb{W}_u \mathbf{p}_N$. Furthermore, we remark that we are testing the adjoint equation onto the reduced state space, indeed the adjoint equation reads

$$\mathbb{V}_u^T \mathbb{A}_h^T(\boldsymbol{\mu}_s) \mathbb{W}_u \mathbf{p}_N = -\mathbb{V}_u^T \mathbb{F}_h(\boldsymbol{\mu}_s) \mathbb{V}_u \mathbf{u}_N.$$

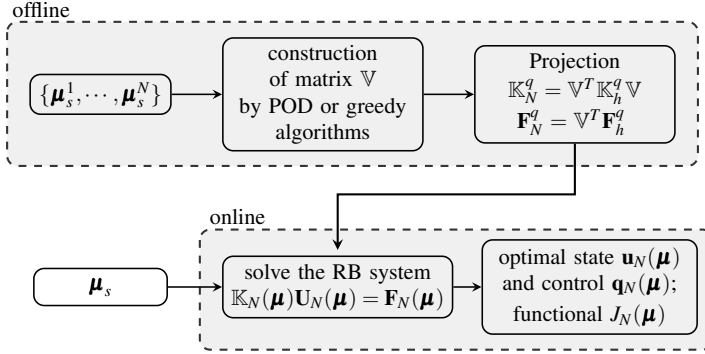


Fig. 12.6 Optimize-then-reduce approach for parametrized optimal control problems

If the reduction of the state system is obtained via a Galerkin projection with $\mathbb{W}_u = \mathbb{V}_u$, the adjoint variable turns out to be approximated by the same RB space used for the solution of the state system. This reduction however is somehow *inconsistent*: in fact, the reduced scheme is not guaranteed to recover any full-order adjoint solution, since the adjoint snapshots are not contained in \mathbb{V}_u . The only way to recover a consistent RB approximation is to define aggregated spaces for state and adjoint variables; this yields a G-RB method equivalent to the one obtained following the *optimize-then-reduce* approach. •

12.4.2 A Posteriori Error Estimation

Following the general framework of Sect. 3.6, an estimate for the error on the three variables involved in the optimality conditions can be derived. Indeed, as problem (12.21) falls into the class of weakly coercive problems, the a posteriori error bound (3.76) holds: for any $\boldsymbol{\mu}_s \in \mathcal{P}_s$,

$$\|\mathbf{U}_h(\boldsymbol{\mu}_s) - \mathbb{V}\mathbf{U}_N(\boldsymbol{\mu}_s)\|_{\mathbb{X}_h} \leq \Delta_N(\boldsymbol{\mu}_s) = \frac{1}{\beta_h(\boldsymbol{\mu}_s)} \|\mathbf{R}_h(\mathbb{V}\mathbf{U}_N; \boldsymbol{\mu}_s)\|_{\mathbb{X}_h^{-1}}, \quad (12.27)$$

where $\mathbf{R}_h(\mathbf{W}; \boldsymbol{\mu}_s) = \mathbb{K}_h(\boldsymbol{\mu}_s)\mathbf{W} - \mathbf{F}_h(\boldsymbol{\mu}_s)$ is the residual of the optimality system (12.21) and $\beta_h(\boldsymbol{\mu}_s) = \sigma_{\min}(\mathbb{X}_h^{-1/2} \mathbb{K}_h(\boldsymbol{\mu}_s) \mathbb{X}_h^{-1/2})$ represents its stability factor, see Sect. 3.7.

Moreover, we have the following bound for the error on the cost functional:

Proposition 12.1. *For any $\boldsymbol{\mu}_s \in \mathcal{P}_s$,*

$$|J_h(\mathbf{u}_h, \mathbf{q}_h; \boldsymbol{\mu}_s) - J_N(\mathbf{u}_N, \mathbf{q}_N; \boldsymbol{\mu}_s)| \leq \Delta_N^J(\boldsymbol{\mu}_s) = \frac{1}{2\beta_h(\boldsymbol{\mu}_s)} \|\mathbf{R}_h(\nabla \mathbf{U}_N; \boldsymbol{\mu}_s)\|_{\mathbb{X}_h}^2. \quad (12.28)$$

Proof. We first note that, thanks to the Lagrangian preserving property,

$$\begin{aligned} J_h(\mathbf{u}_h, \mathbf{q}_h; \boldsymbol{\mu}_s) - J_N(\mathbf{u}_N, \mathbf{q}_N; \boldsymbol{\mu}_s) &= J_h(\mathbf{u}_h, \mathbf{q}_h; \boldsymbol{\mu}_s) - J_h(\nabla \mathbf{u}_N, \nabla \mathbf{q}_N; \boldsymbol{\mu}_s) \\ &= \mathcal{L}_h(\mathbf{U}_h; \boldsymbol{\mu}_s) - \mathcal{L}_h(\nabla \mathbf{U}_N; \boldsymbol{\mu}_s). \end{aligned}$$

By applying the mean value theorem we then obtain

$$\mathcal{L}_h(\mathbf{U}_h; \boldsymbol{\mu}_s) - \mathcal{L}_h(\nabla \mathbf{U}_N; \boldsymbol{\mu}_s) = (\mathbf{U}_h - \nabla \mathbf{U}_N)^T \int_0^1 \nabla \mathcal{L}_h(\mathbf{U}_h + s(\nabla \mathbf{U}_N - \mathbf{U}_h); \boldsymbol{\mu}_s) ds.$$

By approximating the integral using the trapezoidal rule and exploiting the linearity of $\nabla \mathcal{L}_h$, we obtain

$$\mathcal{L}_h(\mathbf{U}_h; \boldsymbol{\mu}_s) - \mathcal{L}_h(\nabla \mathbf{U}_N; \boldsymbol{\mu}_s) = \frac{1}{2}(\mathbf{U}_h - \nabla \mathbf{U}_N)^T (\nabla \mathcal{L}_h(\mathbf{U}_h; \boldsymbol{\mu}_s) + \nabla \mathcal{L}_h(\nabla \mathbf{U}_N; \boldsymbol{\mu}_s)).$$

Then, since $\nabla \mathcal{L}_h(\mathbf{U}_h; \boldsymbol{\mu}_s) = \mathbf{R}_h(\mathbf{U}_h; \boldsymbol{\mu}_s) = 0$, and using (12.21) we have

$$\begin{aligned} |\mathcal{L}_h(\mathbf{U}_h; \boldsymbol{\mu}_s) - \mathcal{L}_h(\nabla \mathbf{U}_N; \boldsymbol{\mu}_s)| &= \frac{1}{2}(\mathbf{U}_h - \nabla \mathbf{U}_N)^T \mathbf{R}_h(\nabla \mathbf{U}_N; \boldsymbol{\mu}_s) \\ &= \frac{1}{2}(\mathbf{U}_h - \nabla \mathbf{U}_N)^T \mathbb{X}^{1/2} \mathbb{X}^{-1/2} \mathbf{R}_h(\nabla \mathbf{U}_N; \boldsymbol{\mu}_s) \\ &\leq \frac{1}{2} \|\mathbf{U}_h - \nabla \mathbf{U}_N\|_{\mathbb{X}} \|\mathbf{R}_h(\nabla \mathbf{U}_N; \boldsymbol{\mu}_s)\|_{\mathbb{X}^{-1}} \leq \frac{\|\mathbf{R}_h(\nabla \mathbf{U}_N; \boldsymbol{\mu}_s)\|_{\mathbb{X}^{-1}}^2}{2\beta_h(\boldsymbol{\mu}_s)}. \quad \square \end{aligned}$$

12.5 Application to an Optimal Heat Transfer Problem

We consider (once more) the heat transfer problem introduced in Sect. 3.8. Now our goal is to regulate the temperature $q = q(\mathbf{x})$ imposed on the three baffles of the heat exchanger (see Fig. 12.7) in such a way that the temperature distribution $u = u(\mathbf{x})$ approaches as much as possible a desired temperature u_d in the outflow region Ω_{obs} of the domain Ω . To this end, we consider the following cost functional,

$$J(u, q; \boldsymbol{\mu}_s) = \frac{1}{2} \int_{\Omega_{\text{obs}}} (u - u_d)^2 d\Omega + \frac{\sigma}{2} \int_{\Gamma_C} (|\nabla_{\Gamma} q|^2 + q^2) d\Gamma, \quad (12.29)$$

where the first terms penalizes the misfit between the desired and the predicted temperature, while the second one penalizes rapid variations and high values of the control variable q .

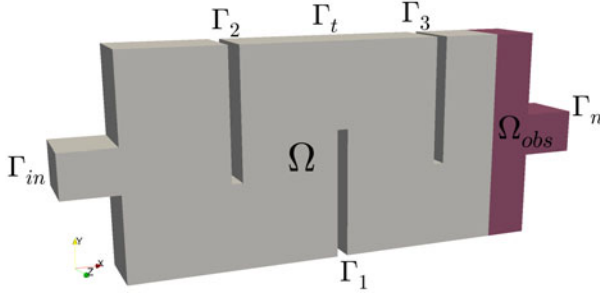


Fig. 12.7 Computational domain Ω and boundaries; the dark red portion of the domain identifies the observation subdomain Ω_{obs} . The control boundary Γ_C is given by the union of the three baffles Γ_1 , Γ_2 and Γ_3

The state and control variables are linked together by the following state problem,

$$\left\{ \begin{array}{ll} -\alpha \Delta u + \mathbf{v} \cdot \nabla u = 0 & \text{in } \Omega \\ u = q & \text{on } \Gamma_C \\ u = 0 & \text{on } \Gamma_w \cup \Gamma_{in} \\ \alpha \nabla u \cdot \mathbf{n} = h & \text{on } \Gamma_t \\ \alpha \nabla u \cdot \mathbf{n} = 0 & \text{on } \Gamma_n, \end{array} \right. \quad (12.30)$$

where the control variable q acts as a Dirichlet datum and we impose a (given) non-zero heat flux h on the top wall Γ_t (see Fig. 12.7). We consider $P_s = 4$ scenario parameters: $\mu_{s1} = u_d \in [2, 12]$ is the desired temperature, $\mu_{s2} = 1/\sigma \in [5, 50]$ is the penalization constant in the cost functional, $\mu_{s3} = h \in [0, 0.5]$ is the heat flux imposed on Γ_t , while $\mu_{s4} \in [1, 500]$ is the Péclet number. The problem is then discretized by piecewise linear finite elements for the state, control and adjoint variables, leading to a discretized optimality system (12.21) of dimension 90408 which admits an affine decomposition with $Q_k = 3$ and $Q_f = 4$ terms (for the matrix and right-hand side, respectively).

For the construction of the RB spaces we employ the greedy procedure. The algorithm selects $N = 37$ sample points with a fixed tolerance $\varepsilon_{\text{tol}} = 10^{-3}$ so that $\Delta_N(\mu_s) \leq \varepsilon_{\text{tol}} \forall \mu_s \in \mathcal{E}_{\text{train}}$, where $\mathcal{E}_{\text{train}}$ is a training set of $5 \cdot 10^3$ random points. In Fig. 12.9 we compare the error estimate $\Delta_N(\mu_s)$ with the true error between the high-fidelity and RB approximations, as well as the error estimate $\Delta_N^J(\mu_s)$ with the error on the cost functional $|J_h(\mu_s) - J_N(\mu_s)|$. In Fig. 12.8 some representative RB solutions are shown, while the computational details are reported in Table 12.1.

The example above is certainly among the simplest optimal control problems that can be addressed. Still, the remarkable speedup that we have obtained should help the reader appreciating the great potential that reduced basis methods possess in dampening the huge computational costs of standing-alone high-fidelity approximations of this class of problems. In this regards, we warn the reader to keep abreast on the fast developments of this exciting field.

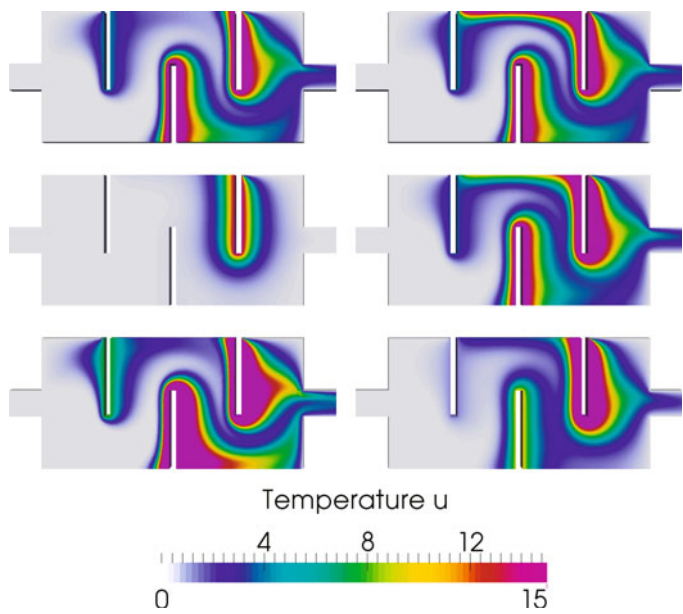


Fig. 12.8 RB state solutions of (12.29)-(12.30) for different parameter values (with $\mu_1 = 8$ fixed)

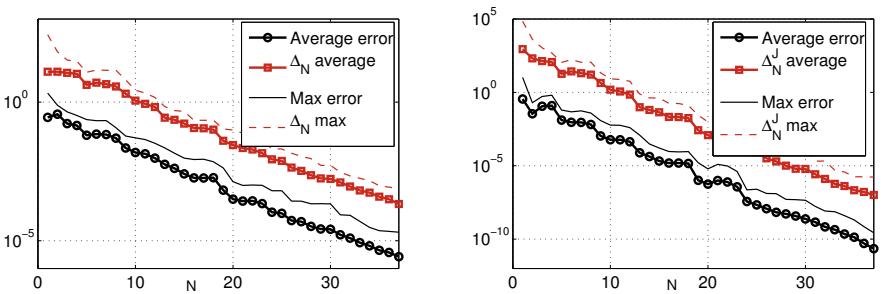


Fig. 12.9 *Left*: average and max computed errors and estimate between the high-fidelity and RB approximations. *Right*: average and max error and estimate between J_h and J_N . Computations have been performed over a test sample set of 200 random points

Table 12.1 Computational details for the high-fidelity and RB approximations of (12.29)-(12.30)

High-fidelity model		Reduced-order model	
Number of FE dofs	90408	Number of RB dofs	37 · 5
Number of parameters P	4	Dofs reduction	488:1
Affine components $Q_a + Q_f$	3 + 4	Offline greedy time	1317 s
Error tolerance greedy ε_{tol}	10^{-3}	RB online solution	3 ms
FE solution time (assembly + sol) ≈ 10 s		RB online estimation	7 ms

Appendix A

Basic Theoretical Tools

This chapter collects basic notions of functional analysis and numerical analysis, together with tools that are extensively used throughout the book. For a more in-depth reading, we refer to e.g. [265, 36, 166, 66, 1, 236].

A.1 Linear Maps, Functionals and Bilinear Forms

Let V and W be two normed vector spaces over \mathbb{R} . A map (or operator) $L : V \rightarrow W$ is said to be *linear* if for any two vectors $x, y \in V$ and any scalar $\alpha \in \mathbb{R}$

$$L(x+y) = L(x) + L(y), \quad L(\alpha x) = \alpha L(x),$$

that is, if it preserves addition and multiplication by a scalar. A linear transformation always maps linear subspaces onto linear subspaces – these latter possibly of lower dimension. A linear operator is *bounded* if there exists a constant $M > 0$ such that

$$\|Lv\|_W \leq M\|v\|_V \quad \forall v \in V.$$

A linear bounded operator is also *continuous*. The vector space $\mathcal{L}(V, W)$ formed by all linear continuous operators from V into W can be endowed with the norm

$$\|L\|_{\mathcal{L}(V, W)} = \sup_{v \in V \setminus \{0\}} \frac{\|Lv\|_W}{\|v\|_V}. \quad (\text{A.1})$$

Let us introduce a relevant class of maps, the so-called *affine* maps, which are extensively exploited throughout the book. An affine map (or affine transformation) $L : V \rightarrow W$ is a function of the form

$$v \mapsto Mv + b,$$

where $M : V \rightarrow W$ is a linear map on V and b is an element of W . Unlike a (purely) linear map, an affine map does not map the zero point in itself. Every linear transformation is affine, but not every affine transformation is linear.

An affine map preserves collinearity (that is, all points lying on a line initially still lie on a line after transformation) and ratios of distances. Moreover, sets of parallel lines remain parallel under an affine transformation. In the Euclidean space $V = \mathbb{R}^d$, examples of affine transformations include translations, rotations, scalings, homotheties, similarity transformations, reflections, and their compositions. We point out that although an affine transformation preserves proportions on lines, it does not preserve angles or lengths. All triangles are affine, that is, any triangle can be transformed into any other by an affine transformation.

Let us define two important (linear) subspaces of V and W , respectively. Given a linear map $L : V \rightarrow W$, we define the *kernel* and the *image* or *range* of L by

$$\text{Ker}(L) = \{x \in V : L(x) = 0\}, \quad \text{Range}(L) = \{w \in W : w = L(x), x \in V\}.$$

The dimensions of these two subspaces are the *rank* and the *nullity* of L , respectively,

$$\dim(\text{Range}(L)) = \text{rank}(L), \quad \dim(\text{Ker}(L)) = \text{nullity}(L).$$

These two subspaces are related through the so-called *rank-nullity* theorem

$$\dim(\text{Ker}(L)) + \dim(\text{Range}(L)) = \text{nullity}(L) + \text{rank}(L) = \dim(V).$$

Remark A.1. Matrices yield examples of linear maps over finite-dimensional vector spaces: if $\mathbb{A} \in \mathbb{R}^{m \times n}$ then $L(\mathbf{x}) = \mathbb{A}\mathbf{x}$, $\mathbf{x} \in \mathbb{R}^n$, describes a linear map between \mathbb{R}^n and \mathbb{R}^m . In this case, the rank and nullity of L coincide with the well-known notions of rank and nullity of the matrix \mathbb{A} , respectively.

A remarkable property exploited throughout the book is the following one. Let us denote by $\{\varphi_j\}_{j=1}^n$ a basis for the space V , where $n = \dim(V)$. Each element of V is uniquely determined by a linear combination of the basis functions, under the form $c_1\varphi_1 + \dots + c_n\varphi_n$, with $c_1, \dots, c_n \in \mathbb{R}$. Thus, if $L : V \rightarrow W$ is a linear map,

$$L(c_1\varphi_1 + \dots + c_n\varphi_n) = c_1L(\varphi_1) + \dots + c_nL(\varphi_n),$$

that is, the map L is entirely determined by the vectors $L(\varphi_1), \dots, L(\varphi_n)$. By denoting $\{\zeta_i\}_{i=1}^m$ a basis of W , we can represent each element $L(\varphi_j) \in W$ as

$$L(\varphi_j) = a_{1j}\zeta_1 + \dots + a_{mj}\zeta_m.$$

In this way, the map is entirely determined by the matrix $\mathbb{A} \in \mathbb{R}^{m \times n}$, where $\mathbb{A}_{ij} = a_{ij}$, $i = 1, \dots, m$, $j = 1, \dots, n$, so that for any $v \in V$ we can evaluate its image through L as $L(v) = \sum_{j=1}^n a_{ij}v_j$. •

In the special case $W = \mathbb{R}$, linear maps are called *functionals*.

Definition A.1. Given a vector space V , we call *functional* on V an operator $F : V \mapsto \mathbb{R}$ associating a real number to each element of V . ◊

A functional F is often denoted by means of the *duality* $F(v) = \langle F, v \rangle$. A functional is said to be *linear* if it is linear with respect to its argument, that is if

$$F(\lambda v + \mu w) = \lambda F(v) + \mu F(w) \quad \forall \lambda, \mu \in \mathbb{R}, v, w \in V.$$

A linear functional is *bounded* if there is a constant $C > 0$ such that

$$|F(v)| \leq C\|v\|_V \quad \forall v \in V. \quad (\text{A.2})$$

A linear and bounded functional on a Banach space (i.e. a normed and complete space) is also continuous. We can define the space V' , called *dual* of V , as the set of linear and bounded functionals on V , that is

$$V' = \{F : V \mapsto \mathbb{R} \text{ such that } F \text{ is linear and bounded} \}$$

and we equip it with the norm $\|\cdot\|_{V'}$ defined as

$$\|F\|_{V'} = \sup_{v \in V \setminus \{0\}} \frac{|F(v)|}{\|v\|_V}. \quad (\text{A.3})$$

The constant C appearing in (A.2) is greater or equal to $\|F\|_{V'}$.

A linear operator between two spaces for which the inverse also exists is called *isomorphism*, according to the following

Definition A.2. A linear and bounded (hence continuous) operator T between two functional spaces V and W is an *isomorphism* if it maps bijectively the elements of the spaces V and W and its inverse T^{-1} exists. If also $V \subset W$ holds, such isomorphism is called *canonical*. \diamond

We further introduce the definition of another fundamental ingredient of abstract variational problems, bilinear forms.

Definition A.3. Given a normed functional space V we call *form* an application which associates to each pair of elements of V a real number $a : V \times V \mapsto \mathbb{R}$. \diamond

A form is called:

1. *bilinear* if it is linear with respect to both its arguments, i.e. if

$$\begin{aligned} a(\lambda u + \mu w, v) &= \lambda a(u, v) + \mu a(w, v) & \forall \lambda, \mu \in \mathbb{R}, \forall u, v, w \in V, \\ a(u, \lambda w + \mu v) &= \lambda a(u, w) + \mu a(u, v) & \forall \lambda, \mu \in \mathbb{R}, \forall u, v, w \in V; \end{aligned}$$

2. *continuous* if there exists a constant $M > 0$ such that

$$|a(u, v)| \leq M\|u\|_V\|v\|_V \quad \forall u, v \in V; \quad (\text{A.4})$$

3. *symmetric* if

$$a(u, v) = a(v, u) \quad \forall u, v \in V; \quad (\text{A.5})$$

4. *positive* (or positive definite) if

$$a(v, v) > 0 \quad \forall v \in V; \quad (\text{A.6})$$

5. *coercive* if there exists a constant $\alpha > 0$ such that

$$a(v, v) \geq \alpha\|v\|_V^2 \quad \forall v \in V. \quad (\text{A.7})$$

A.2 Hilbert Spaces

Hilbert spaces represent the ideal setting to formulate the most common boundary value problems. To this aim, let us give the following

Definition A.4. A map $(\cdot, \cdot) : V \times V \rightarrow \mathbb{R}$ is an *inner* or *scalar product* in V if it satisfies the following properties: for all $x, y, z \in V$, $\alpha, \beta \in \mathbb{R}$,

1. positivity: $(x, x) \geq 0$ and $(x, x) = 0$ if and only if $x = 0$;
2. symmetry: $(x, y) = (y, x)$;
3. bilinearity: $(\alpha x + \beta z, y) = \alpha(x, y) + \beta(z, y)$. \diamond

Any scalar product yields an *induced norm* over V , defined as

$$\|x\| = \sqrt{(x, x)} \quad \forall x \in V. \quad (\text{A.8})$$

An inner product space – that is, a vector space with an inner product – is also a normed vector space. Moreover, the following *Cauchy-Schwarz* inequality holds:

$$|(x, y)| \leq \|x\| \|y\| \quad \forall x, y \in V; \quad (\text{A.9})$$

equality holds in (A.9) if and only if $y = \alpha x$, for some $\alpha \in \mathbb{R}$.

If V is a (linear) space endowed with an inner product – that is, an inner product space – we say that V is a *Hilbert space* if it is complete with respect to the induced norm (A.8). We will often denote the corresponding scalar product by $(\cdot, \cdot)_V$.

Two vectors $x, y \in V$ are said to be *orthogonal* with respect to the V -scalar product and write $x \perp y$ if $(x, y)_V = 0$. A vector x is said to be *orthogonal* to a set W (written $x \perp W$) if $x \perp w$ for each $w \in W$. We now consider the following problem: given a vector $x \in V$ and a closed subspace W in V , find (if it exists) the vector $w \in W$ closest to x , in the sense that it minimizes $\|x - w\|_V$.

The following theorem gives an answer to this problem, as well as a useful characterization of its solution in terms of orthogonality properties.

Theorem A.1 (Orthogonal projections). *Let V be a Hilbert space and W a closed subspace of V . Corresponding to any vector $x \in V$, there is a unique vector $w^* \in W$ (called the projection of x onto W) such that*

$$\|x - w^*\|_V = \inf_{w \in W} \|x - w\|_V.$$

Furthermore, a necessary and sufficient condition for $w^ \in W$ to be the unique minimizing vector is that $x - w^*$ be orthogonal to W , i.e.*

$$(x - w^*, w)_V = 0 \quad \forall w \in W.$$

The *orthogonal complement* of a subspace W of a Hilbert vector space V , denoted by W^\perp , is the set of all vectors in V that are orthogonal to every vector in W

$$W^\perp = \{v \in V \mid (v, w)_V = 0 \quad \forall w \in W\}. \quad (\text{A.10})$$

The orthogonal complement of a subspace W of H is a subspace of V , too, and it is always closed (in the topology induced by the metric defined over V). If W is a closed (linear) subspace of V , then $W \oplus W^\perp = V$, that is, each element $v \in V$ can be uniquely expressed as a sum of an element $w \in W$ and an element $w^\perp \in W^\perp$. Finally, $\dim W + \dim W^\perp = \dim V$.

The following theorem, called identification or representation theorem (see e.g. [236] or [265] for the proof), holds.

Theorem A.2 (Riesz representation theorem). *Let V be a Hilbert space. For each linear and bounded functional f on V there exists a unique element $x_f \in V$ such that*

$$f(y) = (y, x_f)_V \quad \forall y \in V, \quad \text{and} \quad \|f\|_{V'} = \|x_f\|_V. \quad (\text{A.11})$$

Conversely, each element $x \in V$ identifies a linear and bounded functional f_x on V such that

$$f_x(y) = (y, x)_V \quad \forall y \in V \quad \text{and} \quad \|f_x\|_{V'} = \|x\|_V. \quad (\text{A.12})$$

If V is a Hilbert space, its dual space V' of linear and bounded functionals on V is a Hilbert space too. Moreover, thanks to Theorem A.2, there exists a bijective and isometric (i.e. norm-preserving) transformation $f \leftrightarrow x_f$ between V' and V thanks to which V' and V can be identified. We can denote this transformation – called *Riesz (isometric) isomorphism* – as follows:

$$\begin{aligned} R_V : V &\rightarrow V', & x &\rightarrow f_x = R_V x, \\ R_V^{-1} : V' &\rightarrow V, & f &\rightarrow x_f = R_V^{-1} x. \end{aligned} \quad (\text{A.13})$$

Thanks to the Riesz isomorphism, we can define an inner product over V' as follows:

$$(F, G)_{V'} = (R_V^{-1} F, R_V^{-1} G)_V \quad \forall F, G \in V'.$$

As a consequence,

$$\|F\|_{V'} = \|R_V^{-1} F\|_V \quad \forall F \in V'. \quad (\text{A.14})$$

Definition A.5. Let X and Y be two Hilbert spaces. We say that X is embedded in Y with continuous embedding if there exists a constant C such that $\|w\|_Y \leq C\|w\|_X \quad \forall w \in X$. Moreover X is *dense* in Y if each element belonging to Y can be obtained as the limit, in the $\|\cdot\|_Y$ norm, of a sequence of elements of X . \diamond

A.3 Adjoint Operators

Riesz's Theorem enables the definition of *adjoint operator* of L , which extends the notion of transpose \mathbb{A}^T of a matrix $\mathbb{A} \in \mathbb{R}^{m \times n}$, that is

$$(\mathbb{A}\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbb{A}^T \mathbf{y}) \quad \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m.$$

Actually, the notion of adjoint operator can be defined for any given $L \in \mathcal{L}(V, W)$, being V and W two Banach spaces. To this end, let us consider the real-valued map $T_y : x \mapsto_{W'} \langle y, Lx \rangle_W$. For any prescribed $y \in W'$, T_y defines an element of V' .

In fact,

$$|T_y x| = |_{W'} \langle y, Lx \rangle_W| \leq \|Lx\|_W \|y\|_{W'} \leq \|L\|_{\mathcal{L}(V,W)} \|x\|_V \|y\|_{W'},$$

so that $\|T_y\| \leq \|L\|_{\mathcal{L}(V,W)} \|y\|_{W'}$. The operator $L^* : W' \rightarrow V'$ defined by the map $W' \ni y \mapsto T_y \in V'$ is called the *adjoint* of L . More precisely:

Definition A.6. The operator $L^* : W' \rightarrow V'$ defined by the identity

$$_{V'} \langle L^* y, x \rangle_{V'} = _{W'} \langle y, Lx \rangle_W \quad \forall x \in V, y \in W', \quad (\text{A.15})$$

is called the adjoint of L . L^* is a linear and bounded operator between W' and V' , that is $L^* \in \mathcal{L}(W', V')$, moreover $\|L^*\|_{\mathcal{L}(W', V')} = \|L\|_{\mathcal{L}(V, W)}$. \diamond

In the case where V and W are two Hilbert spaces, an additional adjoint operator, $L^* : W \rightarrow V$, called *transpose* or *Hilbert space adjoint* or simply *adjoint* of L , can be introduced. It is defined by

$$(L^* y, x)_V = (y, Lx)_W \quad \forall x \in V, y \in W. \quad (\text{A.16})$$

Here, $(\cdot, \cdot)_V$ denotes the scalar product of V , while $(\cdot, \cdot)_W$ denotes the scalar product of W . The above definition can be explained as follows: for any given element $y \in W$, the real-valued function $x \rightarrow (y, Lx)_W$ is linear and continuous, hence it defines an element of V' . By Riesz's theorem (Theorem A.2) there exists an element x of V , which we name $L^* y$, that satisfies (A.16). Such operator belongs to $\mathcal{L}(W, V)$ (that is, it is linear and bounded from Y to X), moreover

$$\|L^*\|_{\mathcal{L}(W,V)} = \|L\|_{\mathcal{L}(V,W)}. \quad (\text{A.17})$$

We say that L is *selfadjoint* (or *Hermitian*) if $V = W$ and $L^* = L$. Then, (A.16) in this case reduces to $(Lx, y)_V = (x, Ly)_V$. In particular, symmetric matrices are special cases of selfadjoint operators.

In the case where V and W are two Hilbert spaces, we thus have two notions of adjoint operator, L^* and L^* , which are linked through the following relationship:

$$R_V L^* = L^* R_W, \quad (\text{A.18})$$

R_V and R_W being Riesz's canonical isomorphisms from V to V' and from W to W' , respectively (see (A.13)). Indeed, $\forall x \in V, y \in W$,

$$_{V'} \langle R_V L^* y, x \rangle_{V'} = (L^* y, x)_V = (y, Lx)_W = _{W'} \langle R_W y, Lx \rangle_W = _{V'} \langle L^* R_W y, x \rangle_{V'}.$$

A.4 Compact Operators

In this section, we briefly recall the definition and some remarkable properties of compact operators. For a further analysis of this topic, see, e.g., [226, 72, 66]. In the following V and W will denote two Hilbert spaces.

Definition A.7. An operator $L \in \mathcal{L}(V, W)$ is *compact* if L maps bounded sets into precompact sets, i.e. $L(E)$ is compact in W for every bounded $E \subset V$. \diamond

Definition A.8. $L \in \mathcal{L}(V, W)$ is said to have *finite rank* if $\text{Range}(L) \subset W$ is finite dimensional. \diamond

If $L \in \mathcal{L}(V, W)$ is a finite rank operator, then L is compact. In particular, if either $\dim(V) < \infty$ or $\dim(W) < \infty$, then any $L \in \mathcal{L}(V, W)$ has finite rank and thus is compact. Moreover, if L is compact, then so is L^* .

Theorem A.3 (Spectral theorem for compact self-adjoint operators). *Let $L \in \mathcal{L}(V, V)$ be a compact and self-adjoint linear operator on a Hilbert space V . Then, there exist a finite or at most countable sequence of eigenvalues $\lambda_n \in \mathbb{R}$ of L and a sequence of corresponding eigenvectors $\zeta_n \in V$ such that*

$$L\zeta_n = \lambda_n \zeta_n, \quad n \geq 1, \quad (\text{A.19})$$

$|\lambda_1| \geq |\lambda_2| \geq \dots |\lambda_n| \geq \dots$ and $(\zeta_i, \zeta_j)_V = \delta_{ij}$ for all $i, j \geq 1$.

Theorem A.4 (Spectral theorem for compact operators). *Let $L \in \mathcal{L}(V, W)$ be a compact linear operator. Then, there exist a finite or at most countable of singular values $\sigma_n \in \mathbb{R}$ of L and two orthonormal sets $\zeta_n \in V$, $\psi_n \in W$ such that*

$$L\zeta_n = \sigma_n \psi_n, \quad L^* \psi_n = \sigma_n \zeta_n, \quad n \geq 1. \quad (\text{A.20})$$

Moreover, $\sigma_n(L) = (\lambda_n(L^*L))^{1/2} = (\lambda_n(LL^*))^{1/2}$ and $\sigma_n(L) = \sigma_n(L^*)$.

Definition A.9. Suppose V and W are separable Hilbert spaces¹ and that $L \in \mathcal{L}(V, W)$. We say that L is a *Hilbert-Schmidt operator* if there exists an orthonormal basis $\{e_n\}_{n=1}^\infty$ of V such that

$$\sum_{n=1}^\infty \|Le_n\|_W^2 < \infty. \quad \diamond$$

It can be proved that $L \in \mathcal{L}(V, W)$ is a Hilbert-Schmidt operator if and only if its adjoint L^* is also Hilbert-Schmidt. Moreover, Hilbert-Schmidt operators are compact. If $\{e_n\}_{n=1}^\infty$ is an orthonormal basis for V and $L \in \mathcal{L}(V, W)$, we define the Hilbert-Schmidt (HS) norm of L as

$$\|L\|_{HS} = \left(\sum_{n=1}^\infty \|Le_n\|_W^2 \right)^{1/2}. \quad (\text{A.21})$$

By Theorem A.4, the HS norm of a compact operator $L \in \mathcal{L}(V, W)$ is given by

$$\|L\|_{HS} = \left(\sum_{n=1}^\infty \sigma_n(L)^2 \right)^{1/2},$$

and is indeed similar to the Frobenius norm for matrices.

¹ A Hilbert space is separable if and only if it admits a countable orthonormal basis.

A.5 Differentiation in Linear Spaces

In this section, we briefly recall the notions of differentiability and differentiation for operators between Banach spaces; see e.g. [155] for a further analysis of this topic. Let us first consider the notion of *strong* (or *Fréchet*) *differential*:

Definition A.10. Let V and W be two normed linear spaces and $F : E \subseteq V \rightarrow W$. F is said to be *differentiable* at $x \in E$ if there exists a linear and bounded operator $L_x : V \rightarrow W$ such that $\forall \varepsilon > 0, \exists \delta > 0$ such that

$$\|F(x+h) - F(x) - L_x h\|_W \leq \varepsilon \|h\|_V \quad \forall h \in V \text{ with } \|h\|_V < \delta.$$

We denote the expression $L_x h$ (or $L_x[h]$), which generates an element in W for each $h \in V$, *strong* (or *Fréchet*) *differential* of the application F at $x \in E$; L_x is called *strong* (or *Fréchet*) *derivative* of F at x and will be denoted as $DF(x)$. \diamond

From the above definition, we deduce that a differentiable application in x is also continuous in x . Here we recall some properties following from Definition A.10:

- if $F(x) = \text{constant}$, then $DF(x)$ is the null operator, that is $L_x[h] = 0 \forall h \in V$;
- the strong derivative of a continuous linear application $F(x)$ is the application itself, that is $DF(x) = F(x)$;
- given two continuous applications F and G of V in W , if these are differentiable at x , so are the applications $F + G$ and αF , for all $\alpha \in \mathbb{R}$, and we have:

$$D(F + G)(x) = DF(x) + DG(x), \quad D(\alpha F)(x) = \alpha DF(x).$$

A second, relevant definition concerns the *weak* (or *Gâteaux*) *differential*.

Definition A.11. Let F be a mapping from V to W ; we call *Gâteaux differential* (or *directional derivative*) of the application F at x the limit:

$$\delta F(x, h) = \lim_{t \rightarrow 0} \frac{F(x + th) - F(x)}{t} \quad \forall h \in V,$$

where $t \in \mathbb{R}$ and the convergence of the limit must be intended with respect to the norm of the space W . If the weak differential $\delta F(x, h)$ is linear (in general it is not), it can be expressed as $DF(x, h) = F'_G(x)h$ for each $h \in V$. The linear bounded operator $F'_G(x)$ is called *weak* (or *Gâteaux*) *derivative* of F . \diamond

Moreover, we have

$$F(x + th) - F(x) = tF'_G(x)h + o(t) \quad \forall h \in V,$$

which implies

$$\|F(x + th) - F(x) - tF'_G(x)h\| = o(t) \quad \forall h \in V.$$

Remark A.2. If an application F has a strong derivative, then it also admits a weak derivative, which coincide with the strong one; the converse instead is not generally

true. However, if there exists a weak derivative $F'_G(x)$ of the application F on a neighbourhood $U(x)$ of x , and $F'_G(x)$ is continuous at x , then the strong derivative $DF(x)$ at x exists, too, and $DF(x) = F'_G(x)$, that is it coincides with the weak one. •

A.6 Sobolev Spaces

In this section we recall the main definitions regarding Sobolev spaces, necessary to the abstract formulation of PDEs. For further details see, e.g., the monographs [36], [1] and [166]. Let Ω be an open set of \mathbb{R}^d and $f : \Omega \mapsto \mathbb{R}$.

Definition A.12. By *support* of a function f we mean the closure of the set where the function itself takes values different from zero, i. e. $\text{supp } f = \overline{\{\mathbf{x} : f(\mathbf{x}) \neq 0\}}$. ◇

A function $f : \Omega \mapsto \mathbb{R}$ is said to have a *compact support* in Ω if there exists a compact set ² $K \subset \Omega$ such that $\text{supp } f \subset K$. We can provide the following key definition:

Definition A.13. $\mathcal{D}(\Omega)$ is the space of infinitely differentiable functions with compact support in Ω , that is

$$\mathcal{D}(\Omega) = \{f \in C^\infty(\Omega) : \exists K \subset \Omega, \text{ compact} : \text{supp } f \subset K\}. \quad \diamond$$

We are now able to define the space of distributions on Ω :

Definition A.14. We call *distribution* on Ω any linear and continuous transformation T from $\mathcal{D}(\Omega)$ into \mathbb{R} . The space of distributions on Ω is therefore given by the dual space $\mathcal{D}'(\Omega)$ of $\mathcal{D}(\Omega)$. ◇

The action of a distribution $T \in \mathcal{D}'(\Omega)$ on a function $\phi \in \mathcal{D}(\Omega)$ will always be denoted via the duality pairing $\langle T, \phi \rangle$. Differentiation in the sense of distributions is a fundamental concept in view of the definition of Sobolev spaces.

Definition A.15. Let $\Omega \subset \mathbb{R}^d$ and $T \in \mathcal{D}'(\Omega)$. Its derivatives $\frac{\partial T}{\partial x_i}$ in the *sense of distributions* are distributions defined in the following way

$$\left\langle \frac{\partial T}{\partial x_i} \phi \right\rangle = - \left\langle T \frac{\partial \phi}{\partial x_i} \right\rangle \quad \forall \phi \in \mathcal{D}(\Omega), \quad i = 1, \dots, n. \quad \diamond$$

In a similar way, we define derivatives of arbitrary order. We finally remark that differentiation in the sense of distributions is an extension of the classical differentiation of functions. Indeed, if a function f is differentiable with continuity (in classical sense) on Ω , then the derivative of the distribution T_f corresponding to f coincides with the distribution $T_{f'}$ corresponding to the classical derivative f' of f .

² With $\Omega \subset \mathbb{R}^d$, a compact set is a closed and bounded set.

A.6.1 Square-Integrable Functions

Let us now introduce the space of square-integrable functions on $\Omega \subset \mathbb{R}^d$,

$$L^2(\Omega) = \left\{ f : \Omega \mapsto \mathbb{R} : f \text{ is Lebesgue measurable, } \int_{\Omega} (f(\mathbf{x}))^2 d\Omega < +\infty \right\}.$$

More precisely, $L^2(\Omega)$ is a space of *equivalence classes* of Lebesgue-measurable functions, the equivalence relation to be intended as follows: v is equivalent to w if and only if v and w are equal almost everywhere in Ω – in short, a.e. in Ω – i.e. they differ at most on a subset of Ω with zero measure.

The space $L^2(\Omega)$ is a Hilbert space whose scalar product and the corresponding induced norm are given respectively by

$$(f, g)_{L^2(\Omega)} = \int_{\Omega} f(\mathbf{x})g(\mathbf{x}), \quad \|f\|_{L^2(\Omega)} = \sqrt{(f, f)_{L^2(\Omega)}}.$$

To each function $f \in L^2(\Omega)$ we associate a distribution $T_f \in \mathcal{D}'(\Omega)$ such that

$$\langle T_f, \phi \rangle = \int_{\Omega} f(\mathbf{x})\phi(\mathbf{x}) d\Omega \quad \forall \phi \in \mathcal{D}(\Omega).$$

In particular, we can identify $L^2(\Omega)$ with a subset of $\mathcal{D}'(\Omega)$, writing $L^2(\Omega) \subset \mathcal{D}'(\Omega)$. Thus, we can identify a function f of $L^2(\Omega)$ with the corresponding distribution T_f of $\mathcal{D}'(\Omega)$, by writing f in place of T_f . Similarly, when we talk about derivatives, we will always refer to the latter in the sense of distributions – in this way we can define derivatives of functions in $L^2(\Omega)$ which could present some discontinuities, and for which derivatives in the classical sense cannot be defined.

A.6.2 The Spaces $H^1(\Omega)$ and $H_0^1(\Omega)$

Although the functions of $L^2(\Omega)$ are particular instances of distributions, this does not guarantee that their derivatives are still functions of $L^2(\Omega)$. Nevertheless, this property is very relevant when dealing with the abstract variational formulation of a second-order PDE. It is therefore useful to introduce the following spaces:

Definition A.16. Let Ω be an open set of \mathbb{R}^d and k be a positive integer. We call *Sobolev space* of order k on Ω the space formed by the totality of functions of $L^2(\Omega)$ whose (distributional) derivatives up to order k belong to $L^2(\Omega)$:

$$H^k(\Omega) = \{f \in L^2(\Omega) : D^{\alpha} f \in L^2(\Omega) \quad \forall \alpha : |\alpha| \leq k\}. \quad \diamond$$

Here we denote by $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$ an n -tuple of non-negative integers (called *multi-index*), so that $D^\alpha f(\mathbf{x}) = \frac{\partial^{|\alpha|} f(\mathbf{x})}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \dots \partial x_n^{\alpha_n}}$, $|\alpha| = \alpha_1 + \alpha_2 + \dots + \alpha_n$ being the length of the multi-index coinciding with the order of differentiation of f . It follows that $H^{k+1}(\Omega) \subset H^k(\Omega)$ for each $k \geq 0$ and this inclusion is continuous. The space $L^2(\Omega)$ is sometimes denoted by $H^0(\Omega)$.

The Sobolev space $H^k(\Omega)$ is a Hilbert space with respect to the scalar product

$$(f, g)_k = \sum_{|\alpha| \leq k} \int_{\Omega} (D^\alpha f)(D^\alpha g) d\Omega,$$

whose associated norm is

$$\|f\|_k = \|f\|_{H^k(\Omega)} = \sqrt{(f, f)_k} = \sqrt{\sum_{|\alpha| \leq k} \int_{\Omega} (D^\alpha f)^2 d\Omega} = \sqrt{\sum_{m=0}^k |f|_{H^m(\Omega)}^2}$$

having defined the seminorm

$$|f|_k = |f|_{H^k(\Omega)} = \sqrt{\sum_{|\alpha|=k} \int_{\Omega} (D^\alpha f)^2 d\Omega}.$$

A very important Sobolev space arising in the abstract formulation of PDEs is

$$H^1(\Omega) = \{f \in L^2(\Omega) : \nabla f \in (L^2(\Omega))^d\};$$

here $\nabla f \in (L^2(\Omega))^d$ if and only if $\partial_{x_i} f \in L^2(\Omega)$ for any $i = 1, \dots, n$ and

$$\|\nabla f\|_{(L^2(\Omega))^n} = \left(\sum_{i=1}^n \|\partial_{x_i} f\|_{L^2(\Omega)}^2 \right)^{1/2}.$$

For the sake of notation, often we denote by $\|\nabla f\|_{L^2(\Omega)}$ the L^2 norm of a vector function when the meaning is clear from the context. Then

$$(f, g)_1 = (f, g)_{H^1(\Omega)} = \int_{\Omega} f g d\Omega + \int_{\Omega} \nabla f \cdot \nabla g d\Omega,$$

whence

$$\|f\|_1 = \|f\|_{H^1(\Omega)} = \sqrt{\|f\|_{L^2(\Omega)}^2 + \|\nabla f\|_{(L^2(\Omega))^n}^2}, \quad |f|_1 = |f|_{H^1(\Omega)} = \|\nabla f\|_{(L^2(\Omega))^n}.$$

We also recall that if Ω is a (sufficiently regular) open set of \mathbb{R}^d , $n \geq 1$, then $H^k(\Omega) \subset C^m(\overline{\Omega})$ if $k > m + n/2$. In particular, in one spatial dimension ($n = 1$), the functions of $H^1(\Omega)$ are continuous (they are indeed *absolutely continuous*, see [236] and [36], while in two or three dimensions they are not necessarily so. Instead, the functions of $H^2(\Omega)$ are always continuous for $n = 1, 2, 3$.

Another Sobolev space, ubiquitous in the variational formulation of PDEs, whenever we impose Dirichlet boundary conditions, is $H_0^1(\Omega)$, that is, the space of functions $v \in H^1(\Omega)$ vanishing on the boundary $\partial\Omega$. We can thus define

$$H_0^1(\Omega) = \{v \in H^1(\Omega) : v|_{\partial\Omega} = 0\}$$

if we require that $v = 0$ over the entire boundary $\partial\Omega$, or

$$H_{\Gamma_D}^1(\Omega) = \{v \in H^1(\Omega) : v|_{\Gamma_D} = 0, \Gamma_D \subset \partial\Omega\}$$

if we require that $v = 0$ over the portion $\Gamma_D \subset \partial\Omega$. To give a precise meaning to the value $v|_{\partial\Omega}$ of $v \in H^1(\Omega)$ on $\partial\Omega$ – the so-called *trace* of v on $\partial\Omega$ – we need to introduce a *trace* operator, which associates to each function $v \in L^2(\Omega)$, with gradient in $L^2(\Omega)$, a function $v|_{\partial\Omega}$ representing its values on $\partial\Omega$:

Theorem A.5. *Let Ω be a domain³ of \mathbb{R}^d provided with a sufficiently regular boundary $\partial\Omega$, and let $k \geq 1$. There exists one and only one linear and continuous application $\gamma_0 : H^k(\Omega) \mapsto L^2(\partial\Omega)$ such that $\gamma_0 v = v|_{\partial\Omega}$ for any $v \in H^k \cap C^0(\overline{\Omega})$; $\gamma_0 v$ is called trace of v on $\partial\Omega$. In particular, there exists a constant $C > 0$ such that*

$$\|\gamma_0 v\|_{L^2(\Gamma)} \leq C \|v\|_{H^k(\Omega)}.$$

The result still holds if we consider the trace operator $\gamma_{\Gamma_D} : H^k(\Omega) \mapsto L^2(\Gamma_D)$ where Γ_D is a sufficiently regular portion of the boundary of Ω with positive measure.

Owing to this result, Dirichlet boundary conditions make sense when seeking solutions v in $H^k(\Omega)$, with $k \geq 1$, provided we interpret the boundary value in the sense of the trace. The trace operators allow for an interesting characterization of the previously defined space $H_0^1(\Omega)$. Indeed, we have the following property:

Proposition A.1. *Let Ω be a domain of \mathbb{R}^d provided with a sufficiently regular boundary $\partial\Omega$ and let γ_0 be the trace operator from $H^1(\Omega)$ in $L^2(\partial\Omega)$. Then*

$$H_0^1(\Omega) = \text{Ker}(\gamma_0) = \{v \in H^1(\Omega) : \gamma_0 v = 0\}.$$

In other words, $H_0^1(\Omega)$ is formed by the functions of $H^1(\Omega)$ having null trace on the boundary. The functions of $H_0^1(\Omega)$ and, more in general, those of $H_{\Gamma_D}^1(\Omega)$, for every $\Gamma_D \subseteq \partial\Omega$, $\text{meas}(\Gamma_D) > 0$, enjoy the following relevant properties:

Proposition A.2 (Poincaré inequality). *Let Ω be a bounded set in \mathbb{R}^d ; then there exists a constant C_Ω such that*

$$\|v\|_{L^2(\Omega)} \leq C_\Omega \|v\|_{H^1(\Omega)} \quad \forall v \in H_0^1(\Omega). \quad (\text{A.22})$$

Proposition A.3. *The seminorm $|v|_{H^1(\Omega)}$ is a norm on the space $H_0^1(\Omega)$ that turns out to be equivalent to the norm $\|v\|_{H^1(\Omega)}$.*

³ A domain Ω of \mathbb{R}^d is a bounded connected open subset of \mathbb{R}^d with a Lipschitz continuous boundary $\partial\Omega$, see, e.g., [66, Chap. 1].

A.7 Bochner Spaces

When considering parametrized functions $v(\mathbf{x}, \boldsymbol{\mu})$, $\boldsymbol{\mu} \in \mathcal{P} \subset \mathbb{R}^P$, $P > 0$, it is natural to introduce the definition of L^p -spaces of $W^{k,q}(\Omega)$ -valued functions. Let us denote by S a compact subset \mathcal{P} of \mathbb{R}^P . For $1 \leq p, q < \infty$ and a positive integer k , the space $L^p(S; W^{k,q}(\Omega))$ consists of all measurable functions $v : S \rightarrow W^{k,q}(\Omega)$ such that

$$\int_S \|v(s)\|_{W^{k,q}(\Omega)}^p ds < \infty$$

endowed with the norm

$$\|v\|_{L^p(S; W^{k,q}(\Omega))} = \left(\int_S \|v(s)\|_{W^{k,q}(\Omega)}^p ds \right)^{1/p}. \quad (\text{A.23})$$

For every $s \in S$ we have used the shorthand notation $v(s)$ to indicate the function

$$v(s) : \Omega \rightarrow \mathbb{R}, \quad v(s)(\mathbf{x}) = v(\mathbf{x}, s) \quad \forall \mathbf{x} \in \Omega. \quad (\text{A.24})$$

The spaces $L^\infty(S; W^{k,q}(\Omega))$ and $C^0(S; W^{k,q}(\Omega))$ are defined in a similar way. For further details see, e.g., [106, 265].

A.8 Polynomial Interpolation and Orthogonal Polynomials

Let us recall some basic notions about Lagrange polynomial interpolation and orthogonal polynomials in approximation theory. According to a celebrated Weierstrass theorem, every continuous function on a bounded interval can be approximated to arbitrary accuracy by polynomials.

Given a function $f \in C^0(I)$, $n+1$ distinct points x_0, x_1, \dots, x_n (or interpolation points) and the corresponding pairs $(x_i, f(x_i))$, $i = 0, \dots, n$, there exists a unique (interpolatory) polynomial $\mathcal{J}_n f(x) \in \mathbb{P}_n$ such that

$$\mathcal{J}_n f(x_i) = f(x_i), \quad i = 0, \dots, n.$$

The (characteristic) polynomials $l_i \in \mathbb{P}_n$, defined by

$$l_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x - x_j}{x_i - x_j}, \quad i = 0, \dots, n \quad (\text{A.25})$$

form a basis for \mathbb{P}_n , so that the interpolating polynomial can be expressed under the following Lagrange form

$$\mathcal{J}_n f(x) = \sum_{i=0}^n f(x_i) l_i(x).$$

If we denote by $p^* = \arg \min_{p \in \mathbb{P}_n} \|f - p\| \in \mathbb{P}_n$ the best approximation of f among the polynomials of degree n , the interpolation error is bounded by

$$\|E_n\|_\infty = \|f - \mathcal{I}_n f\|_\infty \leq (1 + \Lambda_n(T)) \|f - p^*\|_\infty \quad (\text{A.26})$$

where $\Lambda_n(T)$ denotes the Lebesgue constant related to the set $T = \{x_0, \dots, x_n\}$ and can be computed in terms of the Lagrange polynomials as

$$\Lambda_n(T) = \max_{x \in [a, b]} \lambda_n(x) \quad \text{where} \quad \lambda_n(x) = \sum_{j=0}^n |l_j(x)|.$$

Here $\lambda_n(x)$ denotes the *Lebesgue function* for the given set of interpolation points. Equivalently, we can think the Lebesgue constant as the maximum norm of the linear mapping from data to interpolant. Indeed, if we denote by $\mathcal{I}_n \tilde{f}$ the interpolating polynomial on the set of values $\{\tilde{f}(x_i)\}$, we have

$$\|\mathcal{I}_n f - \mathcal{I}_n \tilde{f}\|_\infty \leq \Lambda_n(T) \max_{i=0, \dots, n} |f(x_i) - \tilde{f}(x_i)|.$$

Hence, small changes on the data give rise to small changes on the interpolating polynomial only if Λ_n is small; the Lebesgue constant plays the role of the condition number for the interpolation problem.

The Lebesgue constant $\Lambda_n(T)$ for degree $n \geq 0$ polynomial interpolation in any set of $n + 1$ distinct points in $[-1, 1]$ is such that

$$\Lambda_n(T) \geq \frac{2}{\pi} \log(n+1) + \frac{2}{\pi} \left(\gamma + \log \frac{4}{\pi} \right), \quad \gamma \approx 0.577.$$

In particular, for equispaced points ($n \geq 1$)

$$\Lambda_n(T) > \frac{2^{n-2}}{n^2} \quad \text{and} \quad \lambda_n(T) \sim \frac{2^{n+1}}{en \log n}, \quad n \rightarrow \infty$$

so that for large n Lagrange interpolation shall become unstable. This falls under the name of *Runge phenomenon*: the interpolant $\mathcal{I}_n f$ might show oscillations close to the extrema of the interval nearly 2^n times larger than f , even if f is analytic. See, e.g., [221, Sect. 8.1].

Interpolation based on Chebyshev (or Legendre) polynomials allows to overcome the Runge phenomenon. Chebyshev and Legendre polynomials are two examples of families of orthogonal polynomials, which provide a general tool in approximation theory (see, e.g., [51, 221, 249]).

Let $w = w(x)$ be a positive, integrable function on $(-1, 1)$ and denote by $\{p_k \in \mathbb{P}_k, k = 0, 1, \dots\}$ a system of algebraic polynomials which are mutually orthogonal on $(-1, 1)$ with respect to w , that is,

$$\int_{-1}^1 p_k(x) p_m(x) w(x) dx = 0 \quad \text{if } k \neq m.$$

Let us denote by $(\cdot, \cdot)_w$ the inner product defined by

$$(f, g)_w = \int_{-1}^1 f(x)g(x)w(x)dx$$

and $\|f\|_w = (f, f)_w^{1/2}$; $(\cdot, \cdot)_w$ and $\|\cdot\|_w$ are respectively the scalar product and the norm for the weighted $L_w^2(-1, 1)$ space.

Legendre and Chebyshev polynomials over $[-1, 1]$ correspond to the following two cases, for which we can also provide the expression of Gauss points and coefficients – respectively, Gauss-Lobatto points and coefficients, in the case we also include the extrema of the interval among the set of points:

- *Chebyshev weight* $w(x) = (1 - x^2)^{-1/2}$, resulting in the Gauss points and weights

$$x_j = -\cos \frac{(2j+1)\pi}{2(n+1)}, \quad w_j = \frac{\pi}{n+1}, \quad 0 \leq j \leq n;$$

the corresponding Gauss-Lobatto points and weights are

$$\bar{x}_j = -\cos \frac{\pi j}{n}, \quad \bar{w}_j = \frac{\pi}{d_j n}, \quad 0 \leq j \leq n, n \geq 1 \quad (\text{A.27})$$

being $d_0 = d_n = 2$, $D_j = 1$ for $j = 1, \dots, n-1$. The Gauss points are, for a fixed $n \geq 0$, the zeros of the Chebyshev polynomial $T_{n+1} \in \mathbb{P}_{n+1}$, being

$$T_k(x) = \cos k\theta, \quad \theta = \arccos x, \quad k = 0, 1, \dots$$

whereas, for $n \geq 1$, the internal Gauss-Lobatto points are the zeros of T'_n ;

- *Legendre weight* $w(x) = 1$, resulting in the Gauss points and weights

$$x_j \text{ zeros of } L_{n+1}(x), \quad w_j = \frac{2}{(1 - x_j^2)(L'_{n+1}(x_j))^2}, \quad 0 \leq j \leq n;$$

the corresponding Gauss-Lobatto points and weights are

$$\bar{x}_0 = -1, \bar{x}_n = 1, \quad \bar{x}_j \text{ zeros of } L'_n(x), \quad 1 \leq j \leq n-1 \quad (\text{A.28})$$

$$\bar{w}_j = \frac{2}{n(n+1)} \frac{1}{(L_n(x_j))^2}, \quad 0 \leq j \leq n$$

where

$$L_k(x) = \frac{1}{2^k} \sum_{l=0}^{[k/2]} (-1)^l \binom{k}{l} \binom{2(k-l)}{k} x^{k-2l} \quad k = 0, 1, \dots$$

is the k -th Legendre polynomial.

The Chebyshev interpolant of f is the polynomial $\mathcal{J}_{n,w}^{GL}f$ of degree n that interpolates f at the Gauss-Lobatto points (A.27), and can be expressed as

$$\mathcal{J}_{n,w}^{GL}f(x) = \sum_{i=0}^n f(x_i) l_i(x) \quad (\text{A.29})$$

where $l_i \in \mathbb{P}_n$ is the i -th characteristic Lagrange polynomial defined by (A.25), such that $l_i(x_j) = \delta_{ij}$ for any $i, j = 0, \dots, n$. In the same way we can obtain the Legendre interpolant of f , by replacing the points (A.27) with those defined in (A.28). A more efficient (and stable) interpolation is obtained by relying on the so-called *barycentric formulae*, see e.g. [249] for further details.

The clustering of Chebyshev points close to the extrema of the interval is indeed a key feature – Legendre points have a similar distribution and share the same good behavior. From a quantitative standpoint, the Lebesgue constant grows only logarithmically if Chebyshev points are used, since

$$\Lambda_n(T) \leq \frac{2}{\pi} \log(n+1) + 1 \quad \text{and} \quad \lambda_n(T) \sim \frac{2}{\pi} \log n, \quad n \rightarrow \infty. \quad (\text{A.30})$$

This result makes Chebyshev points a better choice for polynomial interpolation.

Thanks to this result, we can also state that Chebyshev interpolants are *near-best*: putting together (A.26) and (A.30), we have that the maximum norm accuracy difference between Chebyshev interpolants and the best approximant can never be large. In fact, if $\mathcal{J}_{n,w}^{GL}f$ is the Chebyshev interpolant of f at the $n+1$ Gauss-Lobatto points, then

$$\|f - \mathcal{J}_{n,w}^{GL}f\|_{\infty} \leq \left(2 + \frac{2}{\pi} \log(n+1)\right) \|f - p_n^*\|_{\infty}$$

where the constant appearing at the right-hand side is of order 10^2 for $n > 10^{66}$.

Finally, another relevant feature of Chebyshev (or Legendre) interpolation is that the smoother the function being interpolated, the faster the decay of the error with respect to n . In particular, the interpolation error can be bounded as

$$\|f - \mathcal{J}_{n,w}^{GL}f\|_w \leq Cn^{-s} \|f\|_{s,w}, \quad s \geq 1$$

in the case of both Chebyshev and Legendre interpolation, provided for some $s \geq 1$ $f^{(k)} \in L_w^2(-1, 1)$ for any $k = 0, \dots, s$ and $\|f\|_{s,w} = (\sum_{k=0}^s \|f^{(k)}\|_w^2)^{1/2}$. We underline that the convergence depends on the degree of regularity of f in addition to the number of interpolation points. We also obtain an exponential convergence result in the case of analytic functions, $s \rightarrow \infty$. Moreover, for any continuous function f ,

$$\|f - \mathcal{J}_{n,w}^{GL}f\|_{\infty} \leq Cn^{1/2-s} \|f\|_{s,w}, \quad s \geq 1.$$

References

1. Adams, R.A.: Sobolev Spaces. Academic Press, New York (1975)
2. Afanasiev, K., Hinze, M.: Adaptive control of a wake flow using proper orthogonal decomposition. In: J. Cagnol, M. Polis, J.P. Zolesio (eds.) Shape optimization and optimal design, *Lecture Notes in Pure and Appl. Math.*, vol. 216, pp. 317–332. Dekker, New York (1999)
3. Ainsworth, M., Oden, J.T.: A posteriori error estimation in finite element analysis, vol. 37. John Wiley & Sons, New York (2011)
4. Alleborn, N., Nandakumar, K., Raszillier, H., Durst, F.: Further contributions on the two-dimensional flow in a sudden expansion. *J. Fluid Mech.* **330**, 169–188 (1997)
5. Almroth, B.O., Stehlin, P., Brogan, F.A.: Use of global functions for improvement in efficiency of nonlinear analysis. AIAA paper 81-0575 pp. 286–292 (1981)
6. Almroth, B.O., Stern, P., Brogan, F.A.: Automatic choice of global shape functions in structural analysis. *AIAA J.* **16**(5), 525–528 (1978)
7. Amsallem, D., Cortial, J., Carlberg, K., Farhat, C.: A method for interpolating on manifolds structural dynamics reduced-order models. *Int. J. Numer. Meth. Engng.* **80**(9), 1241–1258 (2009)
8. Amsallem, D., Farhat, C.: An online method for interpolating linear parametric reduced-order models. *SIAM J. Sci. Comput.* **33**(5), 2169–2198 (2011)
9. Amsallem, D., Zahr, M.J., Choi, Y., Farhat, C.: Design optimization using hyper-reduced-order models. *Struct. Multidisc. Optim.* pp. 1–22 (2014). DOI 10.1007/s00158-014-1183-y
10. Amsallem, D., Zahr, M.J., Farhat, C.: Nonlinear model order reduction based on local reduced-order bases. *Int. J. Numer. Methods Engr.* **92**(10), 891–916 (2012)
11. Antil, H., Heinkenschloss, M., Hoppe, R.W., Linsenmann, C., Wixforth, A.: Reduced order modeling based shape optimization of surface acoustic wave driven microfluidic biochips. *Math. Comput. Simul.* **82**(10), 1986 – 2003 (2012)
12. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. Society for Industrial and Applied Mathematics, Philadelphia (2005)
13. Arian, E., Fahl, M., Sachs, E.W.: Trust-region proper orthogonal decomposition for flow control. Tech. Rep. 25, Institute for Computer Applications in Science and Engineering, NASA Langley Research Center (2000)
14. Arnold, D.N., Brezzi, F., Cockburn, B., Marini, L.: Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.* **39**(5), 1749–1779 (2002)
15. Astrid, P., Weiland, S., Willcox, K., Backx, T.: Missing point estimation in models described by proper orthogonal decomposition. *IEEE Trans. Automat. Control* **53**, 2237–2251 (2008)
16. Aubry, N.: On the hidden beauty of the proper orthogonal decomposition. *Theor. Comp. Fluid. Dyn.* **2**, 339–352 (1991)
17. Aubry, N., Lian, W.Y., Titi, E.S.: Preserving symmetries in the proper orthogonal decomposition. *SIAM J. Sci. Comput.* **14**(2), 483–505 (1993)

18. Audouze, C., De Vuyst, F., Nair, P.B.: Reduced-order modeling of parameterized PDEs using time-space parameter principal component analysis. *Int. J. Numer. Methods Engrg.* **80**(10), 1025–1057 (2009)
19. Babuška, I.: Error-bounds for finite element method. *Numer. Math.* **16**, 322–333 (1971)
20. Babuška, I., Nobile, F., Tempone, R.: A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM Review* **52**(2), 317–355 (2010)
21. Baiges, J., Codina, R., Idelsohn, S.: Explicit reduced-order models for the stabilized finite element approximation of the incompressible Navier-Stokes equations. *Int. J. Numer. Meth. Fluids* **72**(12), 1219–1243 (2013)
22. Ballarin, F., Manzoni, A., Quarteroni, A., Rozza, G.: Supremizer stabilization of POD-Galerkin approximation of parametrized steady incompressible Navier-Stokes equations. *Int. J. Numer. Methods Engrg.* **102**(5), 1136–1161 (2015)
23. Barrault, M., Maday, Y., Nguyen, N.C., Patera, A.T.: 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)
24. Benner, P., Mehrmann, V., Sorensen, D.C. (eds.): *Dimension Reduction of Large-Scale Systems, Lecture Notes in Computational Science and Engineering*, vol. 45. Springer, Berlin Heidelberg (2005)
25. Benner, P., Sachs, E.W., Volkwein, S.: Model order reduction for PDE constrained optimization. In: G. Leugering, P. Benner, S. Engell, A. Griewank, H. Harbrecht, M. Hinze, R. Rannacher, S. Ulbrich (eds.) *Trends in PDE Constrained Optimization, International Series of Numerical Mathematics*, vol. 165, pp. 303–326. Springer International Publishing, Basel (2014)
26. Benzi, M., Golub, G.H., Liesen, J.: Numerical solution of saddle point problems. *Acta Numerica* **14**, 1–137 (2005)
27. Berggren, M.: Numerical solution of a flow-control problem: vorticity reduction by dynamic boundary action. *SIAM J. Sci. Comput.* **19**(3), 829–860 (1998)
28. Berkooz, G., Holmes, P., Lumley, J.L.: The proper orthogonal decomposition in the analysis of turbulent flows. *Annu. Rev. Fluid Mech.* **25**(1), 539–575 (1993)
29. Bernardi, C., Maday, Y.: *Approximations spectrales de problèmes aux limites elliptiques*. Springer, Berlin-Heidelberg (1992)
30. Binev, P., Cohen, A., Dahmen, W., DeVore, R., Petrova, G., Wojtaszczyk, P.: Convergence rates for greedy algorithms in reduced basis methods. *SIAM J. Math. Anal.* **43**(3), 1457–1472 (2011)
31. Biswas, G., Breuer, M., Durst, F.: Backward-facing step flows for various expansion ratios at low and moderate Reynolds numbers. *J. Fluids Eng.* **126**(3), 362–374 (2004)
32. Boffi, D., Brezzi, F., Fortin, M.: *Mixed Finite Elements and Applications*. Springer-Verlag, Berlin-Heidelberg (2013)
33. Borzi, A., Schulz, V.: *Computational Optimization of Systems Governed by Partial Differential Equations*. Society for Industrial and Applied Mathematics, Philadelphia (2011)
34. Bramble, J., Pasciak, J.: A new approximation technique for div-curl systems. *Math. Comp.* **73**(248), 1739–1762 (2004)
35. Brenner, S.C., Scott, L.R.: *The Mathematical Theory of Finite Element Methods*, third edn. Springer-Verlag, New York (2008)
36. Brezis, H.: *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer-Verlag, New York (2011)
37. Brezzi, F.: On the existence, uniqueness, and approximation of saddle point problems arising from Lagrangian multipliers. *R.A.I.R.O., Anal. Numér.* **2**, 129–151 (1974)
38. Brezzi, F., Bathe, K.J.: A discourse on the stability conditions for mixed finite element formulations. *Comput. Meth. Appl. Mech. Engrg.* **82**(1–3), 27–57 (1990)
39. Brezzi, F., Rappaz, J., Raviart, P.A.: Finite dimensional approximation of nonlinear problems. Part I: Branches of nonsingular solutions. *Numer. Math.* **36**, 1–25 (1980)
40. Buffa, A., Maday, Y., Patera, A.T., Prud’homme, C., Turinici, G.: A priori convergence of the greedy algorithm for the parametrized reduced basis method. *ESAIM Math. Modelling Numer. Anal.* **46**(3), 595–603 (2012)

41. Buhmann, M.D.: Radial Basis Functions: Theory and Implementations. Cambridge University Press, Cambridge (2003)
42. Bui-Thanh, T., Damodaran, M., Willcox, K.: Proper orthogonal decomposition extensions for parametric applications in transonic aerodynamics (AIAA Paper 2003-4213). In: Proceedings of the 15th AIAA Computational Fluid Dynamics Conference (2003)
43. Bui-Thanh, T., Damodaran, M., Willcox, K.: Aerodynamic data reconstruction and inverse design using proper orthogonal decomposition. *AIAA Journal* **42**(8), 1505–1516 (2004)
44. Bui-Thanh, T., Willcox, K., Ghattas, O.: Model reduction for large-scale systems with high-dimensional parametric input space. *SIAM J. Sci. Comput.* **30**(6), 3270–3288 (2008)
45. Bui-Thanh, T., Willcox, K., Ghattas, O.: Parametric reduced-order models for probabilistic analysis of unsteady aerodynamics applications. *AIAA Journal* **46**(10) (2008)
46. Bungartz, H.J., Griebel, M.: Sparse grids. *Acta Numerica* **13**, 147–269 (2004)
47. Burkardt, J., Gunzburger, M.D., Lee, H.C.: POD and CVT-based reduced-order modeling of Navier-Stokes flows. *Comput. Meth. Appl. Mech. Engrg.* **196**(1-3), 337–355 (2006)
48. Caiazzo, A., Iliescu, T., John, V., Schyschlowa, S.: A numerical investigation of velocity-pressure reduced order models for incompressible flows. *J. Comput. Phys.* **259**, 598–616 (2014)
49. Caloz, G., Rappaz, J.: Numerical analysis for nonlinear and bifurcation problems. In: P.G. Ciarlet, J. Lions (eds.) *Handbook of Numerical Analysis*, vol. V, pp. 487–637. North-Holland, Amsterdam (1997)
50. Canuto, C., Hussaini, M., Quarteroni, A., Zang, T.A.J.: *Spectral Methods in Fluid Dynamics*. Springer-Verlag, New York (1987)
51. Canuto, C., Hussaini, M., Quarteroni, A., Zang, T.A.J.: *Spectral methods: Fundamentals in Single Domains*. Springer-Verlag, Berlin (2006)
52. Canuto, C., Hussaini, M., Quarteroni, A., Zang, T.A.J.: *Spectral methods: Evolution to Complex Geometries and Applications to Fluid Dynamics*. Springer-Verlag, Berlin (2007)
53. Canuto, C., Tonn, T., Urban, K.: A posteriori error analysis of the reduced basis method for non-affine parameterized nonlinear PDEs. *SIAM J. Numer. Anal.* **47**(3), 2001–2022 (2009)
54. Carlberg, K.: Model reduction of nonlinear mechanical systems via optimal projection and tensor approximation. Ph.D. thesis, Stanford University, Stanford (2011)
55. Carlberg, K., Bou-Mosleh, C., Farhat, C.: Efficient non-linear model reduction via a least-squares Petrov-Galerkin projection and compressive tensor approximations. *Int. J. Numer. Meth. Engrg.* **86**(2), 155–181 (2011)
56. Carlberg, K., Farhat, C.: A low-cost, goal-oriented compact proper orthogonal decomposition basis for model reduction of static systems. *Int. J. Numer. Meth. Engrg.* **86**(3), 381–402 (2011)
57. Carlberg, K., Farhat, C., Cortial, J., Amsallem, D.: The GNAT method for nonlinear model reduction: Effective implementation and application to computational fluid dynamics and turbulent flows. *J. Comput. Phys.* **242**, 623–647 (2013)
58. Chaturantabut, S., Sorensen, D.C.: Nonlinear model reduction via discrete empirical interpolation. *SIAM J. Sci. Comput.* **32**(5), 2737–2764 (2010)
59. Chen, Y., Hesthaven, J.S., Maday, Y., Rodriguez, J.: A monotonic evaluation of lower bounds for inf-sup stability constants in the frame of reduced basis approximations. *C. R. Acad. Sci. Paris, Ser. I* **346**, 1295–1300 (2008)
60. Chen, Y., Hesthaven, J.S., Maday, Y., Rodriguez, J.: Improved successive constraint method based a posteriori error estimate for reduced basis approximation of 2D Maxwell's problem. *ESAIM Math. Modelling Numer. Anal.* **43**, 1099–1116 (2009)
61. Chen, Y., Hesthaven, J.S., Maday, Y., Rodríguez, J.: Certified reduced basis methods and output bounds for the harmonic Maxwell's equations. *SIAM J. Sci. Comput.* **32**(2), 970–996 (2010)
62. Chen, Y., Hesthaven, J.S., Maday, Y., Rodriguez, J., Zhu, X.: Certified reduced basis method for electromagnetic scattering and radar cross section estimation. *Comput. Meth. Appl. Mech. Engrg.* **233-236**, 92–108 (2012)

63. Chevreuril, M., Nouy, A.: Model order reduction based on proper generalized decomposition for the propagation of uncertainties in structural dynamics. *Int. J. Numer. Meth. Engng* **89**(2), 241–268 (2012)
64. Christensen, E.A., Brøns, M., Sørensen, J.N.: Evaluation of proper orthogonal decomposition–based decomposition techniques applied to parameter-dependent nonturbulent flows. *SIAM J. Sci. Comput.* **21**, 1419–1434 (1999)
65. Ciarlet, P.G.: *The Finite Element Method for Elliptic Problems*. Classics in Applied Mathematics, 40. Society for Industrial and Applied Mathematics, Philadelphia (2002)
66. Ciarlet, P.G.: *Linear and nonlinear functional analysis with applications*. Society for Industrial and Applied Mathematics, Philadelphia (2014)
67. Clenshaw, C., Curtis, A.: A method for numerical integration on an automatic computer. *Numer. Math.* **2**(1), 197–205 (1960)
68. Cline, A.K., Dhillon, I.S.: Computation of the Singular Value Decomposition. In: L. Hogben (ed.) *Handbook of Linear Algebra*, first edn. CRC Press (2006)
69. Cochran, W.G.: *Sampling techniques*. John Wiley & Sons, Chichester (2007)
70. Cohen, A., DeVore, R.: Approximation of high-dimensional parametric PDEs. *ArXiv e-prints* 1502.06797 (2015)
71. Cohen, A., DeVore, R., Schwab, C.: Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDEs. *Anal. Appl.* **9**(1), 11–47 (2011)
72. Conway, J.B.: *A Course in Operator Theory*. American Mathematical Soc., Providence (2000)
73. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: *Introduction to Algorithms*, third edn. The MIT Press, Cambridge (2009)
74. Cottrell, J.A., Hughes, T.J.R., Bazilevs, Y.: *Isogeometric analysis: toward integration of CAD and FEA*. John Wiley & Sons, Chichester (2009)
75. Cui, T., Marzouk, Y., Willcox, K.: Data-driven model reduction for the Bayesian solution of inverse problems. *Int. J. Numer. Methods Engrg.* **102**(5), 966–990 (2015)
76. Dahmen, W., Plesken, C., Welper, G.: Double greedy algorithms: reduced basis methods for transport dominated problems. *ESAIM Math. Modelling Numer. Anal.* **48**(3), 623–663 (2014)
77. Dautray, R., Lions, J.L.: *Mathematical Analysis and Numerical Methods for Science and Technology*. Springer-Verlag, Berlin Heidelberg (2000)
78. Deane, A.E., Kevrekidis, I.G., Karniadakis, G.E., Orszag, S.A.: Low-dimensional models for complex geometry flows: Application to grooved channels and circular cylinders. *Phys. Fluids A* **3**, 2337 (1991)
79. Dedè, L.: Reduced basis method and a posteriori error estimation for parametrized linear-quadratic optimal control problems. *SIAM J. Sci. Comput.* **32**(2), 997–1019 (2010)
80. Dedè, L.: Reduced basis method and error estimation for parametrized optimal control problems with control constraints. *J. Sci. Comput.* **50**(2), 287–305 (2012)
81. Dedner, A., Klöforn, R., Nolte, M., Ohlberger, M.: A generic interface for parallel and adaptive scientific computing: abstraction principles and the DUNE-FEM module. *Computing* **90**(3–4), 165–196 (2010)
82. Dedner, A., Klöforn, R., Nolte, M., Ohlberger, M.: DUNE-FEM Web page. URL: <http://dune.mathematik.uni-freiburg.de> (2011)
83. Degroote, J., Vierendeels, J., Willcox, K.: Interpolation among reduced-order matrices to obtain parameterized models for design, optimization and probabilistic analysis. *Int. J. Numer. Methods Fluids* **63**(2), 207–230 (2010)
84. Demkowicz, L.: Babuška = Brezzi? Tech. Rep. 08-06, ICE, Univ. of Texas, Austin (2006)
85. Demkowicz, L., Gopalakrishnan, J.: A class of discontinuous Petrov–Galerkin methods. II. optimal test functions. *Numer. Methods Partial Differential Equations* **27**(1), 70–105 (2011)
86. Deparis, S.: Reduced basis error bound computation of parameter-dependent Navier-Stokes equations by the natural norm approach. *SIAM J. Num. Anal.* **46**(4), 2039–2067 (2008)
87. Deparis, S., Løvgren, A.E.: Stabilized reduced basis approximation of incompressible three-dimensional Navier-Stokes equations in parametrized deformed domains. *J. Sci. Comput.* **50**(1), 198–212 (2012)

88. Deparis, S., Rozza, G.: Reduced basis method for multi-parameter-dependent steady Navier-Stokes equations: Applications to natural convection in a cavity. *J. Comp. Phys.* **228**(12), 4359–4378 (2009)
89. Dihlmann, M., Drohmann, M., Haasdonk, B., Ohlberger, M., Schaefer, M.: RB-matlab. URL: <http://www.ians.uni-stuttgart.de/MoRePaS/software/rbmatlab/0.11.04/doc/>
90. Dihlmann, M., Haasdonk, B.: Certified nonlinear parameter optimization with reduced basis surrogate models. *Proc. Appl. Math. Mech* **13**(1), 3–6 (2013)
91. Dihlmann, M., Haasdonk, B.: Certified PDE-constrained parameter optimization using reduced basis surrogate models for evolution problems. *Comput. Optim. Appl.* **60**(3), 753–787 (2015)
92. Drikakis, D.: Bifurcation phenomena in incompressible sudden expansion flows. *Phys. Fluids* **9**(76) (1997). DOI 10.1063/1.869174
93. Drohmann, M., Haasdonk, B., Kaulmann, S., Ohlberger, M.: A software framework for reduced basis methods using Dune-RB and RBmatlab. In: A. Dedner, B. Flemisch, R. Klöforn (eds.) *Advances in DUNE*, pp. 77–88. Springer-Verlag, Berlin Heidelberg (2012)
94. Drohmann, M., Haasdonk, B., Ohlberger, M.: Reduced basis approximation for nonlinear parametrized evolution equations based on empirical operator interpolation. *SIAM J. Sci. Comput.* **34**(2), A937–A969 (2012)
95. Dunavant, D.A.: High degree efficient symmetrical gaussian quadrature rules for the triangle. *Int. J. Num. Meth. Engrg.* **21**(6), 1129–1148 (1985)
96. Eckart, C., Young, G.: The approximation of one matrix by another of lower rank. *Psychometrika* **1**(3), 211–218 (1936)
97. Edmonds, J.: Matroids and the greedy algorithm. *Math. Program.* **1**(1), 127–136 (1971)
98. Eftang, J.L., Grepl, M.A., Patera, A.T.: A posteriori error bounds for the empirical interpolation method. *C.R. Math. Acad. Sci. Paris, Série I* **348**(9–10), 575–579 (2010)
99. Eftang, J.L., Huynh, D.B.P., Knezevic, D.J., Patera, A.T.: A two-step certified reduced basis method. *J. Sci. Comput.* **51**(1), 28–58 (2012)
100. Eftang, J.L., Knezevic, D.J., Patera, A.T.: An “hp” certified reduced basis method for parametrized parabolic partial differential equations. *Math. Comput. Model. Dynam. Syst.* **17**(4), 395–422 (2011)
101. Eftang, J.L., Patera, A.T., Rønquist, E.M.: An “hp” certified reduced basis method for parametrized elliptic partial differential equations. *SIAM J. Sci. Comput.* **32**(6), 3170–3200 (2010)
102. Eftang, J.L., Stamm, B.: Parameter multi-domain hp empirical interpolation. *Int. J. Numer. Methods Engrg.* **90**, 412–428 (2012)
103. Elman, H.C., Liao, Q.: Reduced basis collocation methods for partial differential equations with random coefficients. *SIAM/ASA J. Uncertain. Quantif.* **1**, 192–217 (2013)
104. Elman, H.C., Silvester, D.J., Wathen, A.: *Finite Elements and Fast Iterative Solvers with Applications in Incompressible Fluid Dynamics*. Oxford University Press, New York (2004)
105. Ern, A., Guermond, J.L.: *Theory and practice of finite elements*. Springer-Verlag, New York (2004)
106. Evans, L.C.: *Partial Differential Equations*. American Mathematical Society (1998)
107. Fares, M., Hesthaven, J.S., Maday, Y., Stamm, B.: The reduced basis method for the electric field integral equation. *J. Comput. Phys.* **230**, 5532–5555 (2011)
108. Fink, J.P., Rheinboldt, W.C.: On the discretization error of parametrized nonlinear equations. *SIAM J. Numer. Anal.* **20**(4), 732–746 (1983)
109. Fink, J.P., Rheinboldt, W.C.: On the error behavior of the reduced basis technique for nonlinear finite element approximations. *Z. Angew. Math. Mech.* **63**(1), 21–28 (1983)
110. Fink, J.P., Rheinboldt, W.C.: Solution manifolds and submanifolds of parametrized equations and their discretization errors. *Num. Math.* **45**(3), 323–343 (1984)
111. Galbally, D., Fidkowski, K., Willcox, K., Ghattas, O.: Nonlinear model reduction for uncertainty quantification in large-scale inverse problems. *Int. J. Numer. Methods Engrg.* **81**(12), 1581–1608 (2010)

112. Gerner, A.L., Veroy, K.: Reduced basis a posteriori error bounds for the Stokes equations in parametrized domains: a penalty approach. *Math. Models Meth. Appl. Sci.* **21**(10), 2103–2134 (2010)
113. Gerner, A.L., Veroy, K.: Certified reduced basis methods for parametrized saddle point problems. *SIAM J. Sci. Comput.* **34**(5), A2812–A2836 (2012)
114. Gerstner, T., Griebel, M.: Numerical integration using sparse grids. *Numerical algorithms* **18**(3–4), 209–232 (1998)
115. Girault, V., Raviart, P.A.: Finite element methods for Navier-Stokes equations: Theory and algorithms. Springer-Verlag, Berlin (1986)
116. Gohberg, I., Goldberg, S., Krupnik, N.: Traces and determinants of linear operators. Birkhäuser, Basel (2000)
117. Golub, G.H., Van Loan, C.F.: Matrix Computations, fourth edn. The John Hopkins University Press, Baltimore (2013)
118. Grätsch, T., Bathe, K.J.: A posteriori error estimation techniques in practical finite element analysis. *Comput. and Struct.* **83**, 235–265 (2005)
119. Grepl, M.A.: Certified reduced basis methods for nonaffine linear time-varying and nonlinear parabolic partial differential equations. *Math. Models Methods Appl. Sci.* **22**(3), 1150,015 (2012)
120. Grepl, M.A., Maday, Y., Nguyen, N.C., Patera, A.T.: Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations. *ESAIM Math. Modelling Numer. Anal.* **41**(3), 575–605 (2007)
121. Grepl, M.A., Nguyen, N.C., Veroy, K., Patera, A.T., Liu, G.R.: Certified rapid solution of partial differential equations for real-time parameter estimation and optimization. In: L. Biegler, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Van Bloemen Waanders (eds.) *Real-time PDE-Constrained Optimization*, pp. 197–215. Society for Industrial and Applied Mathematics, Philadelphia (2007)
122. Grepl, M.A., Patera, A.T.: A posteriori error bounds for reduced-basis approximations of parametrized parabolic partial differential equations. *ESAIM Math. Modelling Numer. Anal.* **39**(1), 157–181 (2005)
123. Guevel, Y., Boutyou, H., Cadou, J.M.: Automatic detection and branch switching methods for steady bifurcation in fluid mechanics. *J. Comput. Phys.* **230**, 3614–3629 (2011)
124. Gunzburger, M.D.: Finite Element Methods for Viscous Incompressible Flows. Academic Press, San Diego (1989)
125. Gunzburger, M.D.: Perspectives in Flow Control and Optimization. Advances in Design and Control. Society for Industrial and Applied Mathematics, Philadelphia (2003)
126. Gunzburger, M.D., Bochev, P.: Least-Squares Finite Element Methods. Springer-Verlag, New York (2009)
127. Gunzburger, M.D., Hou, L., Svobodny, T.P.: Boundary velocity control of incompressible flow with an application to viscous drag reduction. *SIAM J. Control Optim.* **30**, 167 (1992)
128. Haasdonk, B., Dihlmann, M., Ohlberger, M.: A training set and multiple bases generation approach for parameterized model reduction based on adaptive grids in parameter space. *Math. Comput. Model. Dyn. Syst.* **17**(4), 423–442 (2011)
129. Haasdonk, B., Ohlberger, M.: Reduced basis method for finite volume approximations of parametrized linear evolution equations. *ESAIM Math. Modelling Numer. Anal.* **42**, 277–302 (2008)
130. Hay, A., Borggaard, J.T., Akhtar, I., Pelletier, D.: Reduced-order models for parameter dependent geometries based on shape sensitivity analysis. *J. Comp. Phys.* **229**(4), 1327–1352 (2010)
131. Hay, A., Borggaard, J.T., Pelletier, D.: Local improvements to reduced-order models using sensitivity analysis of the proper orthogonal decomposition. *J. Fluid Mech.* **629**, 41–72 (2009)
132. Herrero, H., Maday, Y., Pla, F.: RB (reduced basis) for RB (Rayleigh-Bénard). *Comput. Meth. Appl. Mech. Engrg.* **261–262**, 132–141 (2013)

133. Hesthaven, J.S., Stamm, B., Zhang, S.: Efficient greedy algorithms for high-dimensional parameter spaces with applications to empirical interpolation and reduced basis methods. *ESAIM Math. Modelling Numer. Anal.* **48**, 259–283 (2014)
134. Hesthaven, J.S., Warburton, T.: *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications*. Springer-Verlag, New York (2008)
135. Hinze, M., Pinnau, R., Ulbrich, M., Ulbrich, S.: *Optimization with PDE Constraints*. Springer, Netherlands (2009)
136. Hinze, M., Volkwein, S.: Proper orthogonal decomposition surrogate models for nonlinear dynamical systems: Error estimates and suboptimal control. In: P. Benner, D.C. Sorensen, V. Mehrmann (eds.) *Dimension Reduction of Large-Scale Systems, Lecture Notes in Computational Science and Engineering*, vol. 45, pp. 261–306. Springer, Berlin Heidelberg (2005)
137. Holmes, P.J., Lumley, J.L., Berkooz, G.: *Turbulence, coherent structures, dynamical systems and symmetry*. Cambridge University Press, Cambridge (1998)
138. Holmes, P.J., Lumley, J.L., Berkooz, G., Mattingly, J.C., Wittenberg, R.W.: Low-dimensional models of coherent structures in turbulence. *Physics Reports* **287**(4), 337–384 (1997)
139. Hotelling, H.: Simplified calculation of principal components. *Psychometrika* **1**, 27–35 (1936)
140. Huynh, D., Rozza, G., Sen, S., Patera, A.T.: A successive constraint linear optimization method for lower bounds of parametric coercivity and inf-sup stability constants. *C. R. Acad. Sci. Paris. Sér. I Math.* **345**, 473–478 (2007)
141. Huynh, D.B.P., Knezevic, D., Patera, A.T.: A static condensation reduced basis element method : approximation and a posteriori error estimation. *ESAIM Math. Modelling Numer. Anal.* **47**, 213–251 (2013)
142. Huynh, D.B.P., Knezevic, D., Patera, A.T.: A static condensation reduced basis element method: Complex problems. *Comput. Meth. Appl. Mech. Engrg.* **259**, 197–216 (2013)
143. Huynh, D.B.P., Knezevic, D.J., Chen, Y., Hesthaven, J.S., Patera, A.T.: A natural-norm successive constraint method for inf-sup lower bounds. *Comput. Meth. Appl. Mech. Engrg.* **199**(29–32), 1963–1975 (2010)
144. Huynh, D.B.P., Nguyen, N.C., Patera, A.T., Rozza, G.: rbMIT, Copyright MIT. URL: <http://augustine.mit.edu> (2006–2007)
145. Iapichino, L., Quarteroni, A., Rozza, G.: A reduced basis hybrid method for the coupling of parametrized domains represented by fluidic networks. *Comput. Methods Appl. Mech. Engrg.* **221–222**, 63–82 (2012)
146. Ito, K., Ravindran, S.S.: A reduced basis method for control problems governed by PDEs. In: W. Desch, F. Kappel, K. Kunisch (eds.) *Control and Estimation of Distributed Parameter System, International Series of Numerical Mathematics*, vol. 126, pp. 153–168. Birkhäuser, Basel (1998)
147. Ito, K., Ravindran, S.S.: A reduced order method for simulation and control of fluid flows. *J. Comput. Phys.* **143**(2), 403–425 (1998)
148. Kahlbacher, M., Volkwein, S.: Galerkin proper orthogonal decomposition methods for parameter dependent elliptic systems. *Discuss. Math., Differ. Incl. Control Optim.* **27**(1), 95–117 (2007)
149. Kahlbacher, M., Volkwein, S.: POD a-posteriori error based inexact SQP method for bilinear elliptic optimal control problems. *ESAIM: Math. Model. Numer. Anal.* **46**(02), 491–511 (2012)
150. Kärcher, M., Grepl, M.A.: A certified reduced basis method for parametrized elliptic optimal control problems. *ESAIM Control Optim. Calc. Var.* **20**(2), 416–441 (2014)
151. Kerschen, G., Golinval, J.C., Vakakis, A.F., Bergman, L.A.: The method of proper orthogonal decomposition for dynamical characterization and order reduction of mechanical systems: an overview. *Nonlin. Dyn.* **41**, 147–169 (2005)
152. Kirk, B.S., Peterson, J.W., Stogner, R.H., Carey, G.F.: *libMesh: A C++ Library for Parallel Adaptive Mesh Refinement/Coarsening Simulations*. *Eng. Comput.* **22**(3–4), 237–254 (2006)
153. Knezevic, D.J., Peterson, J.W.: A high-performance parallel implementation of the certified reduced basis method. *Comput. Meth. Appl. Mech. Engrg.* **200**(13), 1455–1466 (2011)

154. Kolmogorov, A.N.: Ber die beste annäherung von funktionen einer gegebenen funktionenklasse. *Ann. of Math.* **37**, 107–110 (1936)
155. Kolmogorov, A.N., Fomin, S.V.: *Elements of the Theory of Functions and Functional Analysis*. V.M. Tikhominov, Nauka - Moscow (1989)
156. Kunisch, K., Volkwein, S.: Galerkin proper orthogonal decomposition methods for parabolic problems. *Num. Math.* **90**, 117–148 (2001)
157. Kunisch, K., Volkwein, S.: Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics. *SIAM J. Numer. Anal.* **40**(2), 492–515 (2003)
158. Kunisch, K., Volkwein, S.: Proper orthogonal decomposition for optimality systems. *ESAIM Math. Modelling Numer. Anal.* **42**(1), 1–23 (2008)
159. Lassila, T., Manzoni, A., Quarteroni, A., Rozza, G.: Boundary control and shape optimization for the robust design of bypass anastomoses under uncertainty. *ESAIM Math. Modelling Numer. Anal.* **47**(4), 1107–1131 (2013)
160. Lassila, T., Manzoni, A., Quarteroni, A., Rozza, G.: A reduced computational and geometrical framework for inverse problems in haemodynamics. *Int. J. Numer. Methods Biomed. Engng.* **29**(7), 741–776 (2013)
161. Lassila, T., Manzoni, A., Quarteroni, A., Rozza, G.: Model order reduction in fluid dynamics: challenges and perspectives. In: A. Quarteroni, G. Rozza (eds.) *Reduced Order Methods for Modeling and Computational Reduction, Modeling, Simulation and Applications (MS&A)*, vol. 9, pp. 235–274. Springer International Publishing, Switzerland (2014)
162. Lassila, T., Manzoni, A., Rozza, G.: On the approximation of stability factors for general parametrized partial differential equations with a two-level affine decomposition. *ESAIM Math. Modelling Numer. Anal.* **46**(6), 1555–1576 (2012)
163. Lassila, T., Rozza, G.: Parametric free-form shape design with PDE models and reduced basis method. *Comput. Methods Appl. Mech. Engrg.* **199**(23–24), 1583–1592 (2010)
164. Lee, J.: Introduction to smooth manifolds, *Graduate Texts in Mathematics*, vol. 218. Springer, New York (2012)
165. Lieberman, C., Willcox, K., Ghattas, O.: Parameter and state model reduction for large-scale statistical inverse problems. *SIAM J. Sci. Comput.* **32**(5), 2523–2542 (2010)
166. Lions, J.L., Magenes, E.: *Quelques Méthodes des Résolution des Problèmes aux Limites non Linéaires*. Dunod, Paris (1968)
167. Lohr, S.L.: *Sampling: Design and Analysis*, second edn. Cengage Learning, Boston (2010)
168. Løvgrén, A.E., Maday, Y., Rønquist, E.M.: A reduced basis element method for the steady Stokes problem. *ESAIM Math. Modelling Numer. Anal.* **40**(3), 529–552 (2006)
169. Løvgrén, A.E., Maday, Y., Rønquist, E.M.: The reduced basis element method: offline-online decomposition in the nonconforming, nonaffine case. In: J.S. Hesthaven, E. Rønquist (eds.) *Spectral and High Order Methods for Partial Differential Equations. Selected papers from the ICOSAHOM '09 conference, June 22–26, Trondheim, Norway, Lecture Notes in Computational Science and Engineering*, vol. 76, pp. 247–254. Springer, Berlin Heidelberg (2011)
170. Lumley, J.: The structure of inhomogeneous turbulent flows. In: A.M. Yaglom, V.I. Takarski (eds.) *Atmospheric Turbulence and Radio Wave Propagation*, pp. 166–178. Nauka, Moscow (1967)
171. Machiels, L., Maday, Y., Patera, A.T.: Output bounds for reduced-order approximations of elliptic partial differential equations. *Comput. Meth. Appl. Mech. Engrg.* **190**(26–27), 3413–3426 (2001)
172. Maday, Y., Nguyen, N.C., Patera, A.T., Pau, G.S.H.: A general multipurpose interpolation procedure: the magic points. *Commun. Pure Appl. Anal.* **8**(1), 383–404 (2009)
173. Maday, Y., Patera, A., Penn, J., Yano, M.: A parametrized-background data-weak approach to variational data assimilation: formulation, analysis, and application to acoustics. *Int. J. Numer. Methods Engrg.* **102**, 933–965 (2015)
174. Maday, Y., Patera, A.T., Turinici, G.: *A Priori* convergence theory for reduced-basis approximations of single-parameter elliptic partial differential equations. *J. Sci. Comput.* **17**(1–4), 437–446 (2002)

175. Maday, Y., Patera, A.T., Turinici, G.: Global *a priori* convergence theory for reduced-basis approximation of single-parameter symmetric coercive elliptic partial differential equations. *C. R. Acad. Sci. Paris, Série I* **335**(3), 289–294 (2002)
176. Maday, Y., Quarteroni, A.: Legendre and Chebyshev spectral approximations of Burgers' equation. *Numer. Math.* **37**, 321–332 (1981)
177. Maday, Y., Rønquist, E.M.: A reduced-basis element method. *J. Sci. Comput.* **17**, 447–459 (2002)
178. Maday, Y., Rønquist, E.M.: The reduced basis element method: Application to a thermal fin problem. *SIAM J. Sci. Comput.* **26**(1), 240–258 (2004)
179. Maday, Y., Stamm, B.: Locally adaptive greedy approximations for anisotropic parameter reduced basis spaces. *SIAM J. Sci. Comput.* **35**(6), A2417A2441 (2013)
180. Maday, Y., Tadmor, E.: Analysis of the spectral vanishing viscosity method for periodic conservation laws. *SIAM J. Numer. Anal.* **26**(4), 854–870 (1989)
181. Manzoni, A.: Reduced models for optimal control, shape optimization and inverse problems in haemodynamics. Ph.D. thesis, École Polytechnique Fédérale de Lausanne (2012)
182. Manzoni, A.: An efficient computational framework for reduced basis approximation and a posteriori error estimation of parametrized Navier-Stokes flows. *ESAIM Math. Modelling Numer. Anal.* **48**, 1199–1226 (2014)
183. Manzoni, A., Negri, F.: Heuristic strategies for the approximation of stability factors in quadratically nonlinear parametrized PDEs. *Adv. Comput. Math.* (2015). DOI 10.1007/s10444-015-9413-4. In press
184. Manzoni, A., Pagani, S.: A certified reduced basis method for PDE-constrained parametric optimization problems by an adjoint-based approach. Tech. Rep. 15-2015, MATHICSE Report, Ecole Polytechnique Fédérale de Lausanne (2015). Submitted
185. Manzoni, A., Quarteroni, A., Rozza, G.: Model reduction techniques for fast blood flow simulation in parametrized geometries. *Int. J. Numer. Methods Biomed. Engng.* **28**(6–7), 604–625 (2012)
186. Manzoni, A., Quarteroni, A., Rozza, G.: Shape optimization of cardiovascular geometries by reduced basis methods and free-form deformation techniques. *Int. J. Numer. Methods Fluids* **70**(5), 646–670 (2012)
187. Manzoni, A., Salmoiraghi, F., Heltai, L.: Reduced basis isogeometric methods (RB-IGA) for the real-time simulation of potential flows about parametrized NACA airfoils. *Comput. Meth. Appl. Mech. Engrg.* **284**, 1147–1180 (2015)
188. Melenk, J.M.: On n -widths for elliptic problems. *J. Math. Anal. Appl.* **247**, 272–289 (2000)
189. Meyer, M., Matthies, H.G.: Efficient model reduction in non-linear dynamics using the Karhunen-Loève expansion and dual-weighted-residual methods. *Comput. Mech.* **31**(1–2), 179–191 (2003)
190. Milk, R., Rave, S., Schindler, F.: pyMOR. URL: <http://pymor.org/>
191. Mirsky, L.: Symmetric gauge functions and unitarily invariant norms. *Quart. J. Math.* **11**(1), 50–59 (1960)
192. Morin, P., Nochetto, R.H., Siebert, K.G.: Convergence of adaptive finite element methods. *SIAM Review* **44**(4), 631–658 (2002)
193. Nagy, D.A.: Modal representation of geometrically nonlinear behaviour by the finite element method. *Comput. Struct.* **10**, 683–688 (1979)
194. Necas, J.: *Les Methodes Directes en Theorie des Equations Elliptiques*. Masson, Paris (1967)
195. Negri, F., Manzoni, A., Rozza, G.: Reduced basis approximation of parametrized optimal flow control problems for the Stokes equations. *Comput. & Math. with Appl.* **69**(4), 319–336 (2015)
196. Negri, F., Rozza, G., Manzoni, A., Quarteroni, A.: Reduced basis method for parametrized elliptic optimal control problems. *SIAM J. Sci. Comput.* **35**(5), A2316–A2340 (2013)
197. Nguyen, N.C.: A posteriori error estimation and basis adaptivity for reduced-basis approximation of nonaffine-parametrized linear elliptic partial differential equations. *J. Comp. Phys.* **227**, 983–1006 (2007)
198. Nguyen, N.C., Rozza, G., Patera, A.T.: Reduced basis approximation and a posteriori error estimation for the time-dependent viscous Burgers' equation. *Calcolo* **46**(3), 157–185 (2009)

199. Nguyen, N.C., Veroy, K., Patera, A.T.: Certified real-time solution of parametrized partial differential equations. In: S. Yip (ed.) *Handbook of Materials Modeling*, pp. 1523–1558. Springer, The Netherlands (2005)
200. Nocedal, J., Wright, S.J.: *Numerical Optimization*, second edn. Springer Verlag, New York (2006)
201. Noor, A.: Recent advances in reduction methods for nonlinear problems. *Comput. Struct.* **13**, 31–44 (1981)
202. Noor, A.K.: On making large nonlinear problems small. *Comput. Meth. Appl. Mech. Engrg.* **34**, 955–985 (1982)
203. Noor, A.K., Peters, J.M.: Reduced basis technique for nonlinear analysis of structures. *AIAA J.* **18**, 455–462 (1980)
204. Noor, A.K., Peters, J.M.: Bifurcation and post-buckling analysis of laminated composite plates via reduced basis techniques. *Comput. Meth. Appl. Mech. Engrg.* **29**, 271–295 (1981)
205. Oliveira, I.B., Patera, A.T.: Reduced-basis techniques for rapid reliable optimization of systems described by affinely parametrized coercive elliptic partial differential equations. *Optim. Eng.* **8**, 43–65 (2007)
206. Pacciarini, P., Rozza, G.: Stabilized reduced basis method for parametrized advection-diffusion PDEs. *Comput. Meth. Appl. Mech. Engrg.* **274**(1), 1–18 (2014)
207. Patera, A.: A spectral element method for fluid dynamics: laminar flow in a channel expansion. *J. Comput. Phys.* **54**, 468–488 (1984)
208. Patera, A.T., Rozza, G.: Reduced Basis Approximation and A Posteriori Error Estimation for Parametrized Partial Differential Equations. To appear in MIT Pappalardo Graduate Monographs in Mechanical Engineering (2007). Version 1.0, ©Massachusetts Institute of Technology
209. Pearson, K.: On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**, 559–572 (1901)
210. Peterson, J.S.: The reduced basis method for incompressible viscous flow calculations. *SIAM J. Sci. Stat. Comput.* **10**(4), 777–786 (1989)
211. Pinkus, A.: *n-Widths in Approximation Theory*. Springer-Verlag, Ergebnisse (1985)
212. Pinnau, R.: Model reduction via proper orthogonal decomposition. In: W.H.A. Schilders, H.A. van der Vorst, J. Rommes (eds.) *Model Order Reduction: Theory, Research Aspects and Applications, Mathematics in Industry*, vol. 13, pp. 96–109. Springer, Berlin Heidelberg (2008)
213. Porsching, T.A.: Estimation of the error in the reduced basis method solution of nonlinear equations. *Math. Comput.* **45**(172), 487–496 (1985)
214. Prud'homme, C., Rovas, D.V., Veroy, K., Machiels, L., Maday, Y., Patera, A.T., Turinici, G.: Reliable real-time solution of parametrized partial differential equations: Reduced-basis output bound methods. *J. Fluid Eng.* **124**(1), 70–80 (2002)
215. Prud'homme, C., Rovas, D.V., Veroy, K., Patera, A.T.: A mathematical and computational framework for reliable real-time solution of parametrized partial differential equations. *ESAIM Math. Modelling Numer. Anal.* **36**(5), 747–771 (2002)
216. Quarteroni, A.: Numerical Models for Differential Problems, *Modeling, Simulation and Applications (MS&A)*, vol. 8, second edn. Springer-Verlag Italia, Milano (2014)
217. Quarteroni, A., Rozza, G.: Numerical solution of parametrized Navier-Stokes equations by reduced basis methods. *Numer. Meth. Partial Differential Equations* **23**(4), 923–948 (2007)
218. Quarteroni, A., Rozza, G. (eds.): *Reduced Order Methods for Modeling and Computational Reduction, Modeling, Simulation and Applications (MS&A)*, vol. 9. Springer International Publishing, Switzerland (2014)
219. Quarteroni, A., Rozza, G., Manzoni, A.: Certified reduced basis approximation for parametrized partial differential equations in industrial applications. *J. Math. Ind.* **1**(3) (2011)
220. Quarteroni, A., Rozza, G., Quaini, A.: Reduced basis methods for optimal control of advection-diffusion problem. In: W. Fitzgibbon, R.W. Hoppe, J. Periaux, O. Pironneau, Y. Vassilevski (eds.) *Advances in Numerical Mathematics*, pp. 193–216 (2007)
221. Quarteroni, A., Sacco, R., Saleri, F.: *Numerical Mathematics*, second edn. Springer, New York (2007)

222. Quarteroni, A., Valli, A.: Numerical Approximation of Partial Differential Equations, first edn. Springer-Verlag, Berlin-Heidelberg (1994)
223. Rathinam, M., Petzold, L.R.: A new look at proper orthogonal decomposition. *SIAM J. Numer. Anal.* **41**(5), 1893–1925 (2003)
224. Ravindran, S.S.: A reduced-order approach for optimal control of fluids using proper orthogonal decomposition. *Int. J. Numer. Meth. Fluids* **34**, 425–448 (2000)
225. Renardy, M., Rogers, R.C.: An Introduction to Partial Differential Equations, vol. 13, second edn. Springer-Verlag, New York (2004)
226. Retherford, J.R.: Hilbert space: Compact Operators and the Trace Theorem. Cambridge University Press, Cambridge (1993)
227. Rivière, B.: Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations. Society for Industrial and Applied Mathematics, Philadelphia (2008)
228. Rovas, D.: Reduced-basis output bound methods for parametrized partial differential equations. Ph.D. thesis, Massachusetts Institute of Technology (2003)
229. Rozza, G.: Reduced basis methods for Stokes equations in domains with non-affine parameter dependence. *Comput. Vis. Sci.* **12**(1), 23–35 (2009)
230. Rozza, G., Huynh, D.B.P., Manzoni, A.: Reduced basis approximation and a posteriori error estimation for Stokes flows in parametrized geometries: roles of the inf-sup stability constants. *Numer. Math.* **125**(1), 115–152 (2013)
231. Rozza, G., Huynh, D.B.P., Patera, A.T.: Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations. *Arch. Comput. Methods Engrg.* **15**, 229–275 (2008)
232. Rozza, G., Lassila, T., Manzoni, A.: Reduced basis approximation for shape optimization in thermal flows with a parametrized polynomial geometric map. In: J.S. Hesthaven, E. Rønquist (eds.) *Spectral and High Order Methods for Partial Differential Equations. Selected papers from the ICOSAHOM '09 conference, June 22–26, Trondheim, Norway, Lecture Notes in Computational Science and Engineering*, vol. 76, pp. 307–315. Springer, Berlin Heidelberg (2011)
233. Rozza, G., Veroy, K.: On the stability of reduced basis methods for Stokes equations in parametrized domains. *Comput. Meth. Appl. Mech. Engrg.* **196**(7), 1244–1260 (2007)
234. Rudin, W.: Principles of Mathematical Analysis, third edn. International Series in Pure and Applied Mathematics. McGraw-Hill, Inc. (1976)
235. Saad, Y.: Iterative Methods for Sparse Linear Systems, second edn. Society for Industrial and Applied Mathematics, Philadelphia (2003)
236. Salsa, S.: Partial Differential Equations in Action, *Unitext*, vol. 86, second edn. Springer-Verlag Italia, Milano (2015)
237. Santner, T.J., Williams, B.J., Notz, W.: The design and Analysis of Computer Experiments. Springer-Verlag, New York (2003)
238. Schilders, W.H.A.: Introduction to model order reduction. In: W.H.A. Schilders, H.A. van der Vorst, J. Rommes (eds.) *Model Order Reduction: Theory, Research Aspects and Applications*, *Mathematics in Industry*, vol. 13, pp. 3–32. Springer, Berlin Heidelberg (2008)
239. Schmidt, E.: Zur theorie der linearen und nichtlinearen integralgleichungen. *Math. Ann.* **63**(4), 433–476 (1907)
240. Sen, S.: Reduced basis approximation and a posteriori error estimation for non-coercive elliptic problems: application to acoustics. Ph.D. thesis, Massachusetts Institute of Technology (2007)
241. Sen, S., Veroy, K., Huynh, D.B.P., Deparis, S., Nguyen, N.C., Patera, A.T.: “Natural norm” a posteriori error estimators for reduced basis approximations. *J. Comp. Phys.* **217**(1), 37–62 (2006)
242. Sirovich, L.: Turbulence and the dynamics of coherent structures, part i: Coherent structures. *Quart. Appl. Math.* **45**(3), 561–571 (1987)
243. Sirovich, L.: Analysis of turbulent flows by means of the empirical eigenfunctions. *Fluid Dynam. Res.* **8**(1–4), 85–100 (1991)
244. Stewart, G.W.: On the early history of the singular value decomposition. *SIAM review* **35**(4), 551–566 (1993)

245. Stewart, G.W.: Fredholm, Hilbert, Schmidt: three fundamental papers on integral equations. Translated with commentary (2011). URL: www.cs.umd.edu/stewart/FHS.pdf
246. Temam, R.: Navier-Stokes Equations. AMS Chelsea, Providence, Rhode Island (2001)
247. Tonn, T., Urban, K., Volkwein, S.: Optimal control of parameter-dependent convection-diffusion problems around rigid bodies. *SIAM J. Sci. Comput.* **32**(3), 1237–1260 (2010)
248. Tonn, T., Urban, K., Volkwein, S.: Comparison of the reduced basis and POD a posteriori error estimators for an elliptic linear-quadratic optimal control problem. *Math. Comput. Model. Dynam. Syst.* **17**(4), 355–369 (2011)
249. Trefethen, L.N.: Approximation Theory and Approximation Practice. Society for Industrial and Applied Mathematics, Philadelphia (2013)
250. Trefethen, L.N., Bau, D.: Numerical Linear Algebra. Society for Industrial and Applied Mathematics, Philadelphia (1997)
251. Tröltzsch, F.: Optimal control of partial differential equations: theory, methods and applications, *Graduate Studies in Mathematics*, vol. 112. American Mathematical Society, Providence (2010)
252. Tröltzsch, F., Volkwein, S.: POD a-posteriori error estimates for linear-quadratic optimal control problems. *Comput. Optim. Appl.* **44**, 83–115 (2009)
253. Verfürth, R.: A posteriori Error Estimation Techniques for Finite Element Methods. Oxford University Press, Oxford (2013)
254. Veroy, K., Patera, A.T.: 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**, 773–788 (2005)
255. Veroy, K., Prud’homme, C., Rovas, D.V., Patera, A.T.: A posteriori error bounds for reduced basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations. In: Proceedings of the 16th AIAA Computational Fluid Dynamics Conference (2003). Paper 2003-3847
256. Volkwein, S.: Proper orthogonal decomposition: Theory and reduced-order modelling. URL: www.math.uni-konstanz.de/numerik/personen/volkwein/teaching/POD-book.pdf (2013). Lecture Notes, University of Konstanz
257. Wang, Z., Akhtar, I., Borggaard, J.T., Iliescu, T.: Proper orthogonal decomposition closure models for turbulent flows: A numerical comparison. *Comput. Meth. Appl. Mech. Engrg.* **237–240**, 10–26 (2012)
258. Weyl, H.: Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Math. Ann.* **71**(4), 441–479 (1912)
259. Wirtz, D.: KerMor. URL: <http://www.ians.uni-stuttgart.de/MoRePaS/software/kermor/>
260. Wirtz, D., Sorensen, D.C., Haasdonk, B.: A Posteriori Error Estimation for DEIM Reduced Nonlinear Dynamical Systems. *SIAM J. Sci. Comput.* **36**(2), A311–A338 (2014)
261. Xiao, D., Fang, F., Buchan, A.G., Pain, C.C., Navon, I.M., Du, J., Hu, G.: Non-linear model reduction for the Navier-Stokes equations using the residual DEIM method. *J. Comp. Phys.* **263**, 1–18 (2014)
262. Xiu, D., Hesthaven, J.S.: High-order collocation methods for differential equations with random inputs. *SIAM J. Sci. Comput.* **27**(3), 1118–1139 (2005)
263. Xu, J., Xikatanov, L.: Some observation on Babuška and Brezzi theories. *Numer. Math.* **94**(1), 195–202 (2003)
264. Yano, M.: A space-time Petrov-Galerkin certified reduced basis method: Application to the Boussinesq equations. *SIAM J. Sci. Comput.* **36**(1), A232–A266 (2014)
265. Yosida, K.: Functional Analysis. Springer-Verlag, Berlin (1974)
266. Zahr, M.J., Farhat, C.: Progressive construction of a parametric reduced-order model for PDE-constrained optimization. *Int. J. Numer. Methods Engrg.* **102**(5), 1111–1135 (2015)
267. Zeidler, E.: Nonlinear Functional Analysis and its Applications, vol. I: Fixed-Point Theorems. Springer-Verlag, New York (1985)
268. Zeidler, E.: Nonlinear Functional Analysis and its Applications, vol. II/A: Linear Monotone Operators. Springer-Verlag, New York (1990)

Index

A

affine

- expansion 163
- parametric dependence 7, 8, 52
- transformation 157, 162, 175

approximation

- full-order 1
- Galerkin 24, 31
- high-fidelity 1, 41
- low-rank 117
- Petrov-Galerkin 27

B

- best approximation 96, 118
- branch of nonsingular solutions 216

C

- compliance 47
- computer aided design 179
- condition
 - discrete inf-sup 27, 32, 187
 - inf-sup 16, 18
 - uniform ellipticity 20
- constant
 - coercivity 14
 - continuity 14, 228
 - discrete coercivity 26
 - discrete continuity 31
 - discrete inf-sup 30
 - Lebesgue 278
 - Sobolev embedding 217, 232, 242
- control variable 3
- convergence 25

D

- derivative
 - Fréchet 215, 217, 219, 272

Gâteaux 272

in the sense of distributions 273

deviation 96

direct sum 79

discrepancy 96

distribution 273

domain 276

original 40, 155

reference 40, 155, 161

domain decomposition 179

E

equation

- advection-diffusion-reaction 12, 20, 159, 181

equations

- linear elasticity 172
- linear elasticity 12, 22, 184
- Navier-Stokes 13, 23, 173, 217, 232
- Oseen 243
- Stokes 13, 22, 173, 186

error

- a priori estimate 34
- approximation 24

error estimate

- optimal 24, 27, 32

error estimator

- a posteriori 56
- effectivity 59

F

factor

- coercivity 41
- continuity 41
- stability 41, 46

finite elements

- Lagrangian basis 34

form 267

- affine 52
- bilinear 14, 267
- coercive 14, 41, 267
- continuous 14, 40, 267
- inf-sup stable 16, 41
- nonaffine 167, 205
- parametric 52
- positive 267
- strongly coercive 14
- symmetric 15, 267
- trilinear 23, 217
- weakly coercive 16

free-form deformation 179

full-order model *see* high-fidelity model

function

- compact support 273

functional 266

- bounded 267
- continuous 14, 265
- cost 246
- linear 14, 266
- norm 267
- objective 246
- quadratic 15, 246

G

Galerkin orthogonality 24, 46

Gram-Schmidt orthonormalization 109, 143

greedy algorithm 8, 142, 143

H

high-fidelity model 1, 41

hp reduced basis 109

hyper-reduction 227

I

image 266

inequality

- Gårding 17, 21
- Korn 35
- Poincaré 21, 276

interpolation

- Chebyshev 279
- Lagrange 102, 194, 277
- Legendre 279

isogeometric analysis 179

isomorphism 267

K

Karhunen-Loève decomposition 123

kernel 266

Kolmogorov n -width 96, 97, 149, 199

L

Lagrange

- multiplier 250, 257

Lagrangian

- characteristic polynomials 277

lemma

- Céa 25
- Lax-Milgram 15
- Strang 206

lifting function 20, 22, 165

linear program 64

M

magic points 195

map

- affine 265
- linear 265

matrix

- correlation 124
- inverse
 - generalized 117
- Jacobian 221, 224, 233, 234
- Moore-Penrose generalized inverse 117, 140
- stiffness 25, 42
- transformation 6, 54, 75

method

- collocation 5, 103, 194
- descent 249
- discrete empirical interpolation 193, 203, 227
- empirical interpolation 193, 195, 227
- finite element 33
- Galerkin 24, 76
 - convergence 25
- Gauss-Newton 225
- generalized Galerkin 206
- gradient 249
- least-squares 77
- Newton 220, 221, 225, 232
- Petrov-Galerkin 78
- projection-based 6
- reduced basis 43
 - EIM-G-RB 205
 - element 179
 - G-RB 76, 224
 - LS-RB 77, 224, 225
 - nonlinear LS-RB 225
 - PG-RB 78
- reduced basis (RB)
 - EIM-G-RB 206
- stochastic collocation 194
- strongly consistent 24
- successive constraint 62

model order reduction 1

N

nodes 34

nonaffine

parametrization 167

problems 167

norm

energy 15, 24

Frobenius 116, 117

Hilbert-Schmidt 271

normal equations 77

nullity 266

number

Péclet 68

Reynolds 174

O

offline/online strategy 7

operator 265

adjoint 270

bounded 265

compact 271

continuous 265

discrete supremizer 30

finite rank 271

Hilbert space adjoint 270

Hilbert-Schmidt 129, 271

linear 265

projection 79

selfadjoint 270

semilinear elliptic 218

supremizer 28, 49, 188, 233

trace 276

transpose 270

optimal test space 28

optimality system 257, 258

orthogonal complement 268

P

parameter

set 39

parameters

control 247

geometric 39

physical 39

scenario 247

parametric complexity 89, 104

parametrized PDE 1

Piola transformation 171

points

equispaced 278

Gauss 279

Gauss-Lobatto 279

polynomials

Chebyshev 102, 278

Legendre 102, 278

orthogonal 104, 278

Principal Component Analysis 123

principal component analysis 121

problem

reduced basis (RB) 6

abstract variational 14

adjoint 250, 251, 259

constrained minimization 19

data assimilation 245

forward 2

heat transfer 3

high-fidelity 1, 41

identification 3, 245

inverse 3

many-query 4

minimization 15

mixed variational 17

optimal control 3, 245

optimal design 3, 245

parametric optimization 247, 248

parametrized 1, 40

parametrized optimal control 247, 256

PDE-constrained optimization 245

saddle point 17

strongly coercive 14

variational 15

weakly coercive 16

projection 6

projection-based

method 5

projector 79

oblique 81

orthogonal 81

proper orthogonal decomposition 8, 123

basis 124

gappy 204

R

radial basis function 66, 179

range 266

rank 266

reduced basis

error estimator 59

functions 44

Galerkin 45, 76

Hermite 110

Lagrange 109

least-squares 48, 77

method 2

Petrov-Galerkin 45, 78

- solution 6, 44
- space 44
- Taylor 110
- reduced-order model 2, 45
- reduced-order modeling 1
- residual 6, 58, 76
- Riesz representative 30, 48, 269
- Runge phenomenon 278

S

- saddle point 19
- sample
 - test 69
 - training 69, 143
- sampling
 - full factorial 135
 - latin hypercube 135
 - random 135
 - sparse grid 135
 - tensorial 135
- scalar product
 - discrete 73
- seminorm 20, 276
- sensitivity equations 94, 250
- sequential quadratic programming 251
- shape optimization 3, 246
- singular value decomposition 115
- singular values 115
- singular vectors 115
- snapshots 2, 5, 43, 124, 143
- solution
 - manifold 87
 - map 87
 - regular 219, 228
 - set 87

space

- Banach 267
- dual 267
- Hilbert 268
- of distributions 273
- Sobolev 274
- sparse grid 195
- stability estimate 24, 41
- stability factor 26, 62, 65, 188, 228
- interpolant 65
- lower bound 62
- support
 - compact 273
 - of a function 273
- system approximation 227

T

- theorem
 - Babuška 27
 - Banach fixed-point 230, 243
 - Brezzi 18, 32
 - Implicit Function 93, 216
 - Kantorovich 220, 231
 - Leray-Schauder 217, 218
 - Nečas 16
 - orthogonal projection 268
 - Riesz 269
 - Schmidt 130
 - Schmidt-Eckart-Young 118
 - spectral 271
 - trace 276
- transformation
 - Piola 158, 178

V

- vertices 34